Cell Reports

Best of 2015

Year 4 in Review

CellPress
www.cell.com
CELLBOREP2015

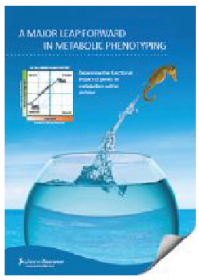
A MAJOR LEAP FORWARD IN PHENOTYPING

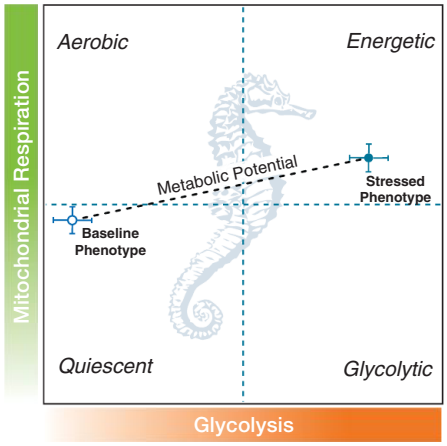
Determine the functional impact of genes on metabolism in an hour

XFp Cell Energy Phenotype Test Kit

Using as few as 15,000 live cells and an XFp Analyzer, the XF Cell Energy Phenotype Test identifies your cells' metabolic phenotype as well as their metabolic potential - the utilization of mitochondrial respiration and glycolysis.

Request a FREE poster and energy phenotype analysis of your cells at www.seahorsebio.com/pheno





Seahorse Bioscience

Focused on antibodies for 40 years.
Not on advertising.

BETHYL
LABORATORIES, INC

Bethyl Laboratories, Inc. has been dedicated to supporting scientific discovery through its qualified antibody products and custom antibody services since its founding in 1972.

Every antibody that Bethyl sells has been manufactured to exacting standards at its sole location in Montgomery, Texas, and has been validated in-house by Bethyl's team of scientists. Antibodies are tested across a range of applications including western blot, immunoprecipitation, immunohistochemistry, immunocytochemistry, ChIP, proximity ligation assay and ELISA.

Currently, Bethyl's portfolio consists of over 7,150 catalog products; offering close to 5,700 primary antibodies targeting over 2,700 proteins and 1,450 secondary antibodies raised against immunoglobulins from over 25 species. Trial sizes are available for over 4,000 antibodies targeting more than 2,350 protein targets. They are conveniently priced at $50 and serve as an opportunity to discover for yourself why our antibodies are really good.

For really good antibodies, visit **bethyl.com/trialsize**

Terms & Conditions: $50 pricing for US customers only; international customers please contact your distributor for details. Trial sizes (catalog # ending in "-T") are not available for all antibodies and existing promotions or discounts do not apply.

BETHYL
LABORATORIES, INC

SAY GOOD-BYE TO THE DARKROOM

Introducing the ZOE™ Fluorescent Cell Imager.
No darkroom, no training, no overwhelming user interface.

Combining brightfield capabilities with multichannel fluorescence, this cell imager is both affordable and easy to use — your perfect solution for routine cell culture and imaging applications.

Learn more at **bio-rad.com/info/newzoe**

BIO-RAD

Foreword

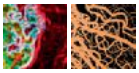
We are so pleased to present the *Best of Cell Reports 2015*. In year 4 we saw some major changes, including moving to weekly publication. From tales of very long-lived whales to an analysis of what might have improved last year's influenza vaccine, we published top-notch content from across the life sciences. That made it all the more difficult to put together this issue, and there were some truly tough choices to make. With the final content in this supplement, we try to represent the breadth of biology covered by *Cell Reports* and have based our choices on reader downloads, altmetric scores, and citation information. We would urge you (once you have perused the *Best of* of course!) to browse our table of contents to explore the full range of Open Access papers we publish in your field and beyond.

We are really looking forward to 2016. In the last year, *Cell Reports* has been globetrotting. We organized the *Ubiquitous Ubiquitin Signaling* Lablinks in Cambridge, UK and the *RNA in the Nervous System* meeting in New York, USA. In November 2015, we cohosted a major symposium in Singapore on human genomics and organized a day-long cancer and metabolism meeting in Philadelphia. We plan to be out and about in 2016 and will hopefully visit your neighborhood, too!

Last, but certainly not least, we want to thank the researchers who are the authors, reviewers, readers, and editorial and advisory board members that make *Cell Reports* possible. With an expanding editorial and production team internally and editorial board advisors externally, we are so excited about 2016. But for now, kick back, grab a coffee, tea, or beverage, and enjoy the following selection of papers.

Finally, we are grateful for the generosity of our sponsors, who helped to make this reprint collection possible.

Cell Reports



For information for the Best of Series, please contact:
Jonathan Christison
Program Director, Best of Cell Press
e: jchristison@cell.com
p: 617-397-2893
t: @cellPressBiz

MojoSort™

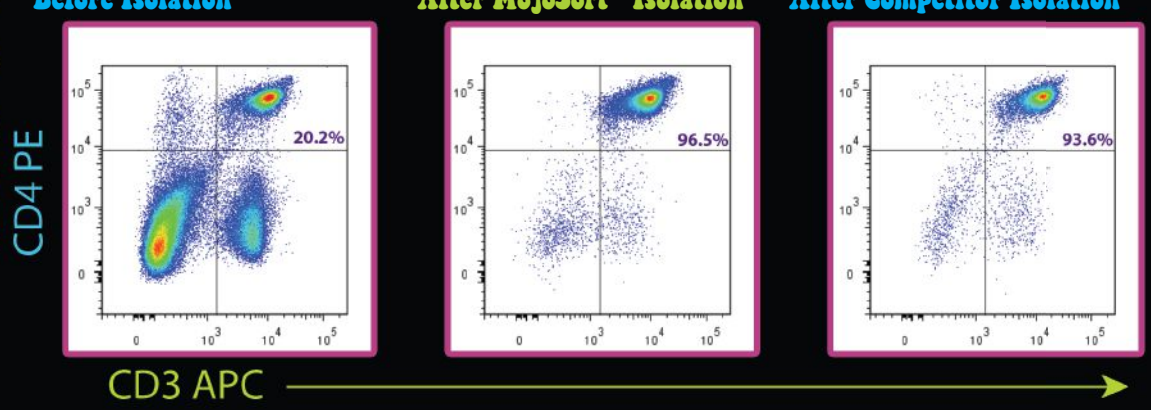
Magnetic Cell Separation System

The MojoSort™ Magnetic Cell Separation System is designed for the separation of target populations using positive or negative selection. MojoSort™ nanoparticles deliver excellent purity and yield at an unmatched, affordable price. Magnetically sorted cells can be used for multiple downstream applications.

MojoSort™ advantages:

- **Small and large test size formats to meet research needs**
- **Robust performance**
- **Preserved cell functionality after sorting**
- **Excellent price**

Add some Mojo to your experiment and explore the possibilities!



A suspension of single cells from pooled C57BL/6 mouse spleen and lymph nodes was prepared to isolate CD4+ T cells using the MojoSort™ Mouse CD4 T Cell Isolation Kit. Cells were stained with PE anti-mouse CD4 (clone RM4-4), APC anti-mouse CD3ε (145-2C11), and 7-AAD. Grateful Dead cells were excluded from the analysis.

To learn more, visit: biolegend.com/mojosort

BioLegend®

BioLegend is ISO 9001:2008 and ISO 13485:2003 Certified

Toll-Free Tel: (US & Canada): 1.877.BIOLEGEND (246.5343)
Tel: 858.768.5800
biolegend.com

08-0052-23

World-Class Quality | Superior Customer Support | Outstanding Value

Cell Reports

Best of 2015

Reports

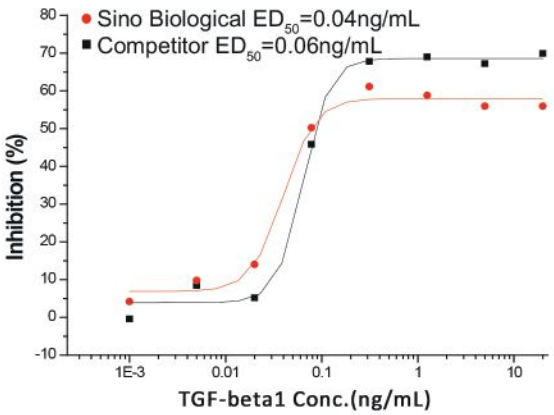
(continued)

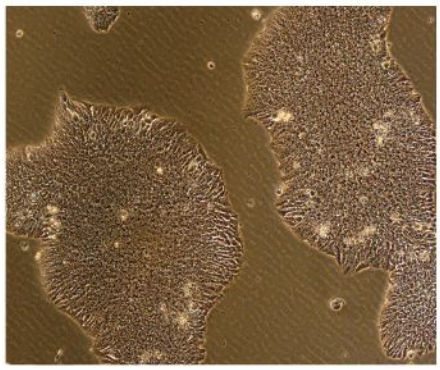
Articles

Resources

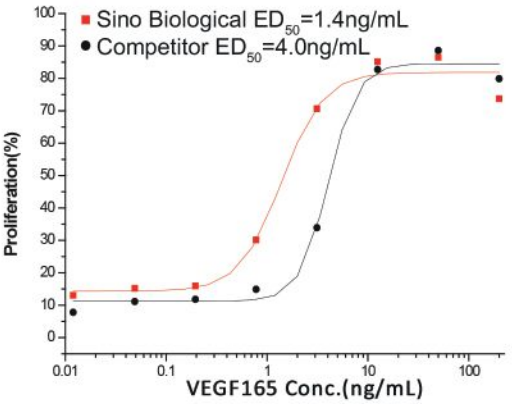
Cytokines & Growth Factors in Cell Culture

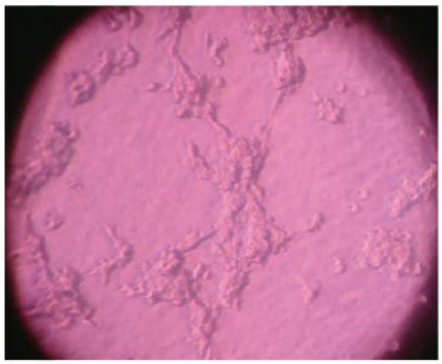
>> Produced in house
>> Bulks in stock
>> Multiple species
>> Save up to 60%





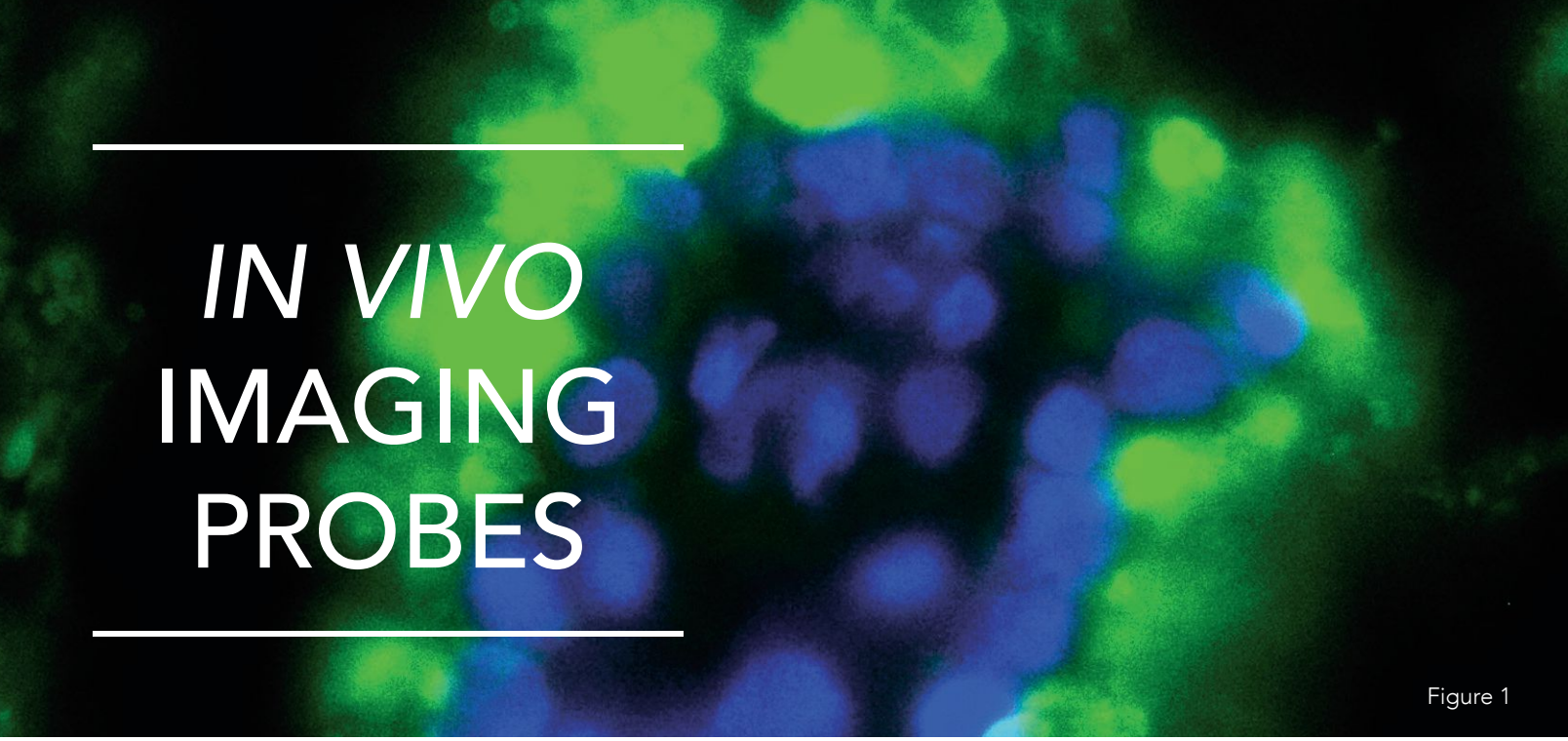
hESC cultured with TGF-β1+bFGF





Measured in a cell proliferation assay using HUVEC

Sino Biological Inc.
Biological Solution Specialist
www.sinobiological.com



IN VIVO IMAGING PROBES

Figure 1

Detect apoptosis & caspase activity in cancer, embryonic development & ischemic conditions.

In vivo imaging probes bring greater insight to your preclinical research. CAS-MAP™ imaging probes allow you to distinguish between apoptotic and healthy cells *in vivo*.

In vivo imaging probes are injected IV. The probes are cell-permeant and bind to active caspases while unbound probe is removed via the circulatory system. Visualize using live animal imaging or harvest tissue for *ex vivo* analysis by microscopy or FACS.

Translate from cell to animal using *in vivo* imaging probes.

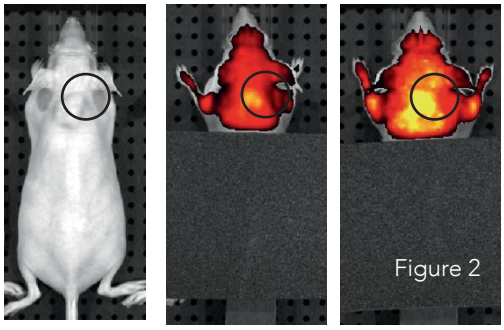
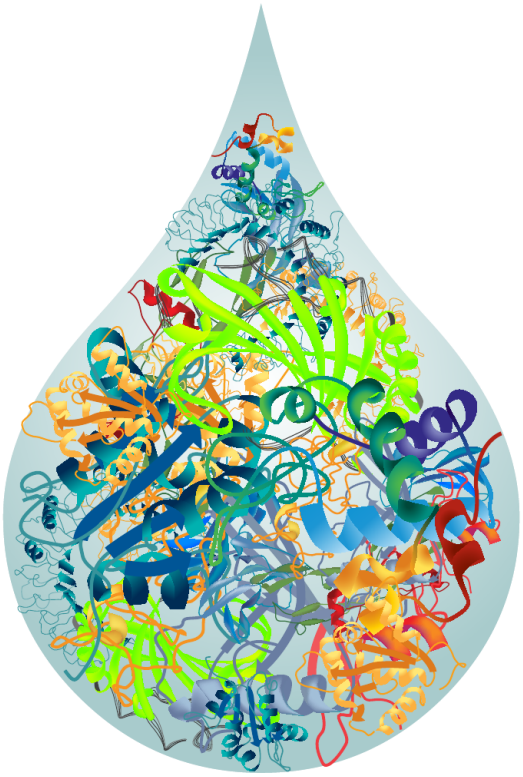


Figure 1: Apoptotic cells are labeled green using CAS-MAP™ probes in an in vivo study of postnatal development. Figure 2: Apoptotic COLO205 tumor cells are labeled in response to treatment.

Download application note:
vergentbio.com/invivo
844.803.0346

VERGENT BIOSCIENCE

R&D SYSTEMS
a bio-techne brand



What's in Your Sample?

Find Your Answers!

rndsystems.com/
Immunoassays

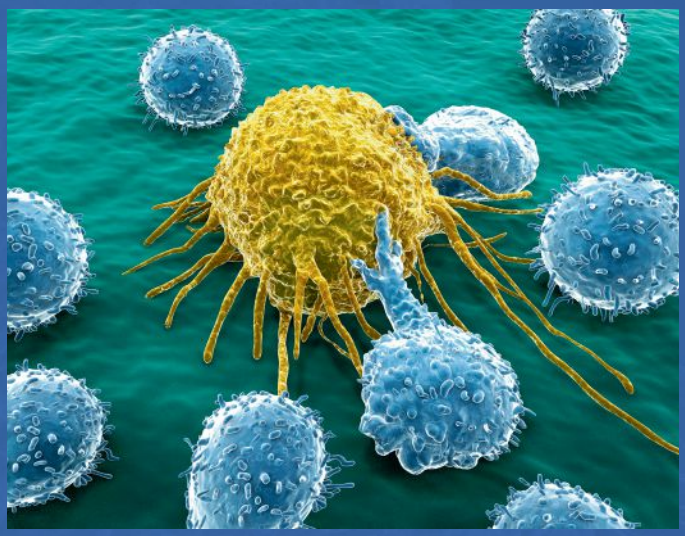
MBL International Corporation

Data You Can Trust

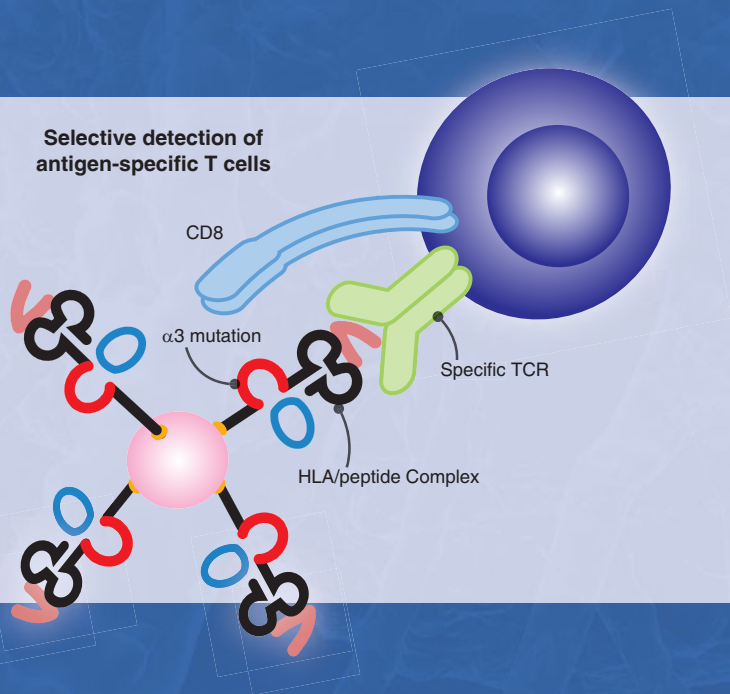
MHC Class I & II Tetramers • MHC Monomers • CD1d Tetramers

MBL has been providing tetramer reagents for over a decade and has recently extended its portfolio through the acquisition of the Beckman Coulter iTAg™ MHC Tetramer product line. MBL International is your source for quality reagents for detection of antigen-specific T cells and NKT cells.

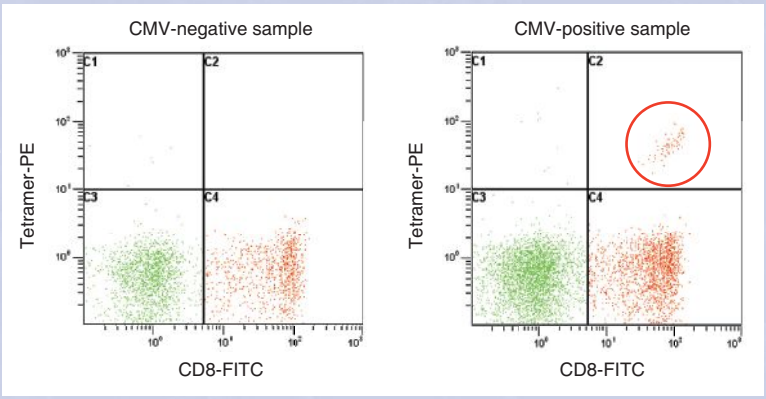
- Increased specificity of proprietary α3 mutation
- Wide selection of alleles: Human, Mouse, Rhesus Macaque, Chicken
- Identification of Ag-specific $CD8^+$ and $CD4^+$ T cells
- Assay flexibility using biotinylated MHC monomers
- Detection of NKT cells using CD1d tetramers



Selective detection of antigen-specific T cells



HLA-A*02:01 CMV pp65 Tetramer



Contact MBL International today to learn more about our products and how we can help you discover ways to make your science more efficient and effective.

mblintl.com • 800-200-5459 • 781-939-6963

life.science.discovery.™

MC-RUO-020

Dharmacon™
RNAi, Gene Expression, and Gene Editing

Specific, functional, and scalable

CRISPR-Cas9 gene editing

Simplify CRISPR-Cas9 gene editing with Dharmacon **predesigned CRISPR guide RNA reagents**. Selected by the **proprietary Dharmacon CRISPR RNA algorithm, these genome-wide products** are designed with unparalleled specificity checking, plus selection criteria trained and validated on functional knockout data. Now you can easily order specific, functional, predesigned CRISPR guide RNAs – without any time-consuming design steps or tedious cloning – for editing one gene or thousands.

Optimized tools for confident CRISPR-Cas9 genome engineering

gelifesciences.com/dharmacon

The BMG LABTECH All Stars

Innovative, high-performance microplate readers for all assay needs

CLARIOstar®

The most sensitive monochromator-based microplate reader. Equipped with revolutionary LVF monochromators™, it is the perfect reader for flexible assay development.

PHERAstar® *FSX*

The new gold standard for High Throughput Screening. The PHERAstar *FSX* is the new reference multi-mode plate reader, combining highest sensitivity with the fastest read times.

Omega series

Upgradeable single to multi-mode filter-based microplate readers. The Omega series offers a combination of flexibility and performance for any life science application.

SPECTROstar® *Nano*

Microplate reader with ultra-fast detection of UV/Vis absorbance. Spectra from 220 - 1000 nm are detected in less than one second per sample in microplates and cuvettes.

Find all our microplate readers on www.bmglabtech.com

BMG LABTECH

The Microplate Reader Company

ThawSTAR™

Automated Cell Thawing System

biocision

the solution:

the problem:

Solve the water bath problem.

- Standardized cell thawing
- High recovery, viability and reproducibility
- Ideal for use in the hood

2015 Design Award Gold IBO

Winner of IBO Gold Award

Visit: www.biocision.com/thawstar/cell

EXQUISITE SENSITIVITY.

HARNESS THE POWER OF ADVANCED SENSITIVITY AND RESOLUTION.

The CytoFLEX Flow Cytometer has the dynamic range to see dim and bright populations in the same sample. And with 15 parameters, including 13 channels for fluorescent detection, heterogeneous populations can be defined with exquisite clarity.

Learn more at goo.gl/Ew8KYw

BECKMAN COULTER
Life Sciences

Are you still injecting?

Focus on your research instead, and let ALZET® Osmotic Pumps do the dosing for you.

ALZET pumps are a superior alternative to repetitive injections and other dosing methods that require frequent animal handling. These fully implantable pumps provide continuous and precise administration, for up to 6 weeks with a single pump, to unrestrained lab animals as small as mice. ALZET pumps are economical and easy to use by research personnel. Connection to a catheter enables direct delivery to vessels, cerebral ventricles, and other target sites. Learn more at alzet.com.

Now available: iPRECIO Pumps

- Programmable
- Refillable
- Implantable
- Small size for mice and rats

Learn more at www.alzet.com/iprecio

alzet® OSMOTIC PUMPS

DURECT

BAKER

Environments For Science™

From the moment when
the pieces fell into place,
and you saw the big picture,

through late nights reviewing literature,
designing procedures,
testing hypotheses —

nothing fell through the cracks,
because too much was at stake.

The work you did then
advanced your field,
and your career.

We are proud to have been there,
protecting your work, and you,
with the very best in air containment,
contamination control, and
controlled environment technology.

Now, as your work gives rise
to new discoveries —

We can't wait to see what you do next.

bakerco.com

Biological Safety Cabinets • Clean Benches • Fume Hoods • CO_2 Incubators • Hypoxia & Anaerobic Workstations

89 NORTH®

ADVANCED FLUORESCENCE IMAGING SYSTEMS

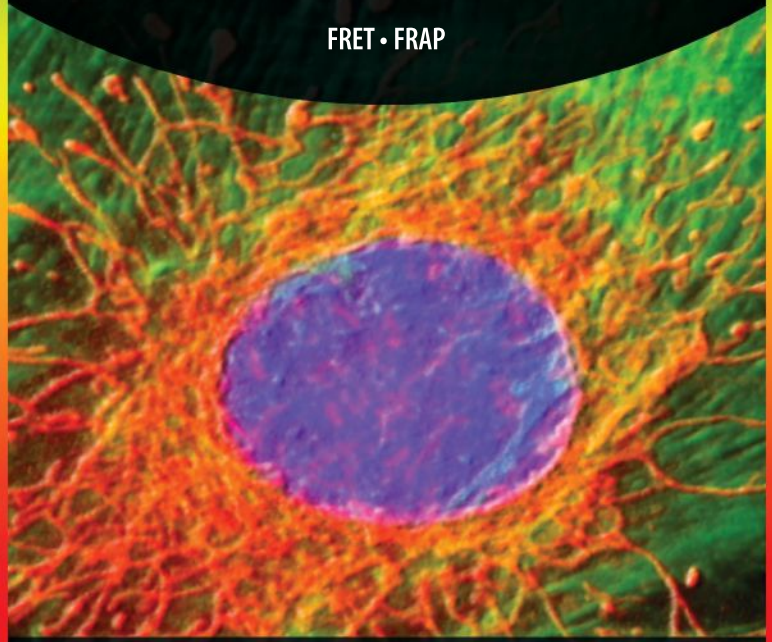
Imaging products from 89 North, Cairn Research and CrestOptics create light re-engineered.

Confocal Imaging • Super-Resolution Imaging

High-Speed Ratiometric Imaging • Optogenetics • Photoactivation

Photoconversion • Simultaneous Multichannel Imaging

FRET • FRAP



PHONE +1-802-881-0302
1-877-417-8313

WEB sales@89north.com
89north.com

CHROMA®

DURABLE HIGH TRANSMISSION SPUTTERED FILTERS FROM UV TO NIR

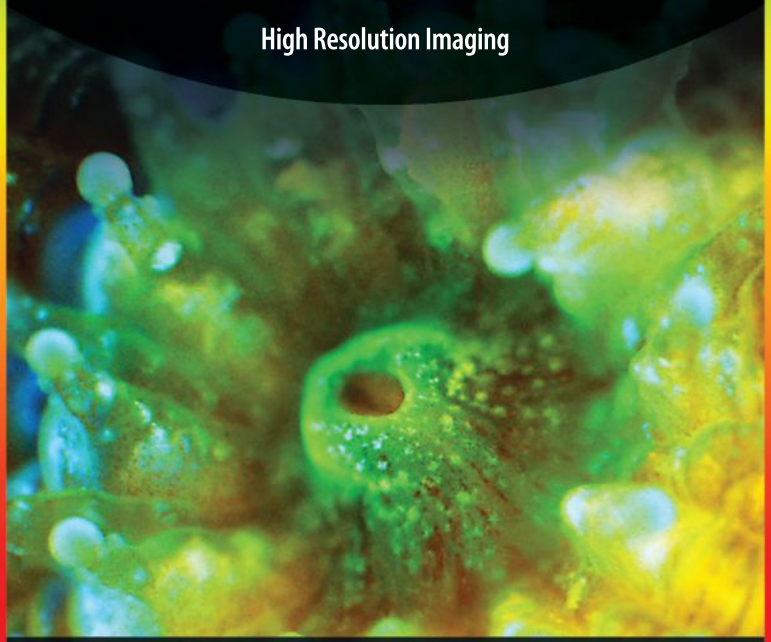
From super-resolution to Raman spectroscopy, Chroma filters give you more options.

Flow Cytometry • Machine Vision

Point-of-Care • Laser Applications

Raman • Fluorescence • Astronomy

High Resolution Imaging



PHONE +1-802-428-2500
1-800-824-7662

WEB sales@chroma.com
chroma.com

HUMANIZED NSG™ MICE

SOPHISTICATED. INNOVATIVE. VERSATILE.

JAX humanized NSG™ is the platform of choice for studying immuno-oncology, viral host-pathogen interactions, or the development of novel drug therapies.

START YOUR STUDY TODAY!

Expect delivery of your order within 7-10 business days.

The Jackson Laboratory

1-800-422-6423 (US, Canada & Puerto Rico)
1-207-288-5845 (from any location)

jax.org/humanized-mice

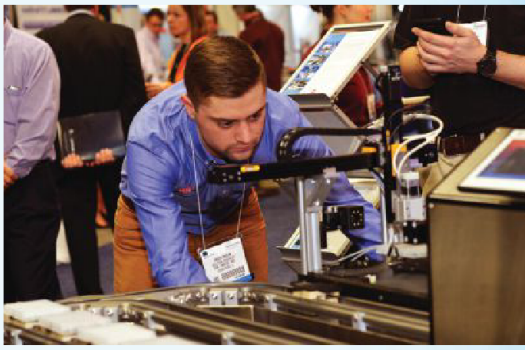
slas 2016

5th annual INTERNATIONAL CONFERENCE & EXHIBITION

JANUARY 23-27 | SAN DIEGO CONVENTION CENTER SAN DIEGO :: CALIFORNIA

INFORMATION. INNOVATION. INSPIRATION.

VISIT SLAS2016.ORG FOR FULL PROGRAM INFORMATION.



SAN DIEGO CONVENTION CENTER

JANUARY 23-27, 2016

slas Come Transform Research

SLAS2016.ORG

METTLER TOLEDO

Recycling plastic is fantastic!

Rainin TerraRack reduces tip rack waste.

Request your free sample today!

▸ **mt.com/rainin-tr**

RAININ

Because the future is in your hands™

TALK TO HENRIK

PrecisA Monoclonals are precise, accurate and targeted. The stringent production process and characterization procedure ensures premium performance in approved applications, with defined specificity, multiplexing opportunities, secured continuity and stable supply.

If you're in need of an antibody to target a protein we don't list in our online catalog you might be interested in our antibody co-development program. Learn more today at atlasantibodies.com/talktohenrikj

Henrik is our R&D Manager

MADE IN SWEDEN

ATLAS ANTIBODIES
Totally human

Cell Reports
Report

Leukemia-Associated Somatic Mutations Drive Distinct Patterns of Age-Related Clonal Hemopoiesis

Thomas McKerrell,[1,13] Naomi Park,[2,13] Thaidy Moreno,[3] Carolyn S. Grove,[1] Hannes Ponstingl,[1] Jonathan Stephens,[4,5] Understanding Society Scientific Group,[6] Charles Crawley,[7] Jenny Craig,[7] Mike A. Scott,[7] Clare Hodkinson,[4,8] Joanna Baxter,[4,8] Roland Rad,[9,10] Duncan R. Forsyth,[11] Michael A. Quail,[2] Eleftheria Zeggini,[12] Willem Ouwehand,[4,5,12] Ignacio Varela,[3] and George S. Vassiliou[1,4,7,*]

[1]Haematological Cancer Genetics, Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK
[2]Sequencing Research Group, Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK
[3]Instituto de Biomedicina y Biotecnología de Cantabria (CSIC-UC-Sodercan), Departamento de Biología Molecular, Universidad de Cantabria, 39011 Santander, Spain
[4]Department of Haematology, Cambridge Biomedical Campus, University of Cambridge, Cambridge CB2 0XY, UK
[5]NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge CB2 0PT, UK
[6]Institute for Social and Economic Research, University of Essex, Colchester CO4 3SQ, UK
[7]Department of Haematology, Cambridge University Hospitals NHS Trust, Cambridge CB2 0QQ, UK
[8]Cambridge Blood and Stem Cell Biobank, Department of Haematology, University of Cambridge, Cambridge CB2 0XY, UK
[9]Department of Medicine II, Klinikum Rechts der Isar, Technische Universität München, 81675 München, Germany
[10]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[11]Department of Medicine for the Elderly, Cambridge University Hospitals NHS Trust, Cambridge CB2 0QQ, UK
[12]Human Genetics, Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK
[13]Co-first author
*Correspondence: gsv20@sanger.ac.uk
http://dx.doi.org/10.1016/j.celrep.2015.02.005

SUMMARY

Clonal hemopoiesis driven by leukemia-associated gene mutations can occur without evidence of a blood disorder. To investigate this phenomenon, we interrogated 15 mutation hot spots in blood DNA from 4,219 individuals using ultra-deep sequencing. Using only the hot spots studied, we identified clonal hemopoiesis in 0.8% of individuals under 60, rising to 19.5% of those ≥90 years, thus predicting that clonal hemopoiesis is much more prevalent than previously realized. *DNMT3A*-R882 mutations were most common and, although their prevalence increased with age, were found in individuals as young as 25 years. By contrast, mutations affecting spliceosome genes *SF3B1* and *SRSF2*, closely associated with the myelodysplastic syndromes, were identified only in those aged >70 years, with several individuals harboring more than one such mutation. This indicates that spliceosome gene mutations drive clonal expansion under selection pressures particular to the aging hemopoietic system and explains the high incidence of clonal disorders associated with these mutations in advanced old age.

INTRODUCTION

Cancers develop through the combined action of multiple mutations that are acquired over time (Nowell, 1976). This paradigm is well established in hematological malignancies, whose clonal history can be traced back for several years or even decades (Ford et al., 1998; Kyle et al., 2002). It is also clear from studies of paired diagnostic-relapsed leukemia samples that recurrent disease can harbor some, but not always all, mutations present at diagnosis, providing evidence for the presence of a clone of ancestral pre-leukemic stem cells that escape therapy and give rise to relapse through the acquisition of new mutations (Ding et al., 2012; Krönke et al., 2013). Studies of such phenomena have defined a hierarchical structure among particular leukemia mutations, with some, such as those affecting the gene *DNMT3A*, displaying the characteristics of leukemia-initiating lesions and driving the expansion of hemopoietic cell clones prior to the onset of leukemia (Ding et al., 2012; Shlush et al., 2014).

These observations suggest that individuals without overt features of a hematological disorder may harbor hemopoietic cell clones carrying leukemia-associated mutations. In fact, such mutations, ranging from large chromosomal changes (Jacobs et al., 2012; Laurie et al., 2012) to nucleotide substitutions (Busque et al., 2012), have been found to drive clonal hemopoiesis in some individuals. Recent reanalyses of large exome-sequencing data sets of blood DNA showed that clonal hemopoiesis is more common than previously realized and increases with age to affect up to 11% of those over 80 and 18.4% of those over 90 years (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014). The presence of such clones was associated with an increased risk of developing hematological or other cancers and a higher all-cause mortality, probably due to an increased risk of cardiovascular disease (Genovese et al., 2014; Jaiswal et al., 2014).

Table 1. Mutation Hot Spots Interrogated in This Study

Gene	Target Codon
DNMT3A	R882
JAK2	V617
NPM1	L287
SRSF2	P95
SF3B1	K666
SF3B1	K700
IDH1	R132
IDH2	R140
IDH2	R172
KRAS	G12
NRAS	G12
NRAS	Q61
KIT	D816
FLT3	D835
FLT3	N676

Also see Table S1 for detailed information about numbers of samples screened for each mutation.

The important findings of these studies were based on analysis of exome-sequencing data sets that were generated for the study of constitutional genomes, thus trading genome-wide coverage for reduced sensitivity for detecting small subclonal events. We used the different approach of targeted re-sequencing of selected leukemia-associated mutation hot spots in blood DNA from more than 4,000 individuals unselected for blood disorders. In addition to increasing the sensitivity for detecting subclonal mutations, this approach enabled us to prospectively select and study a large number of elderly individuals. Our results show that clonal hemopoiesis is significantly more common than anticipated, give new insights into the distinct age-distribution and biological behavior of clonal hemopoiesis driven by different mutations, and help explain the increased incidence of myelodysplastic syndromes (MDSs) with advancing age.

RESULTS

To investigate the incidence, target genes, and age distribution of age-related clonal hemopoiesis (ARCH), we performed targeted re-sequencing for hot spot mutations at 15 gene loci recurrently mutated in myeloid malignancies (Table 1) using blood DNA from 3,067 blood donors aged 17–70 (Wellcome Trust Case Control Consortium [WTCCC]) and 1,152 unselected individuals aged 60–98 years (United Kingdom Household Longitudinal Study [UKHLS]; see Figure S1 for detailed age distributions). To do this, we developed and validated a robust methodology, employing barcoded multiplex PCR of mutational hot spots followed by next-generation sequencing (MiSeq) and bioinformatic analysis, to extract read counts and allelic fractions for reference and non-reference nucleotides. This reliably detected mutation-associated circulating blood cell clones with a variant allele fraction (VAF) $\geq$ 0.008 (0.8%; see Supplemental Experimental Procedures and Figure S2).

We obtained adequate coverage ($\geq$1,000 reads at all studied hot spots) from 4,067 blood DNA samples and identified mutation-bearing clones in 105 of these. Of note, not all hot spots were studied in all samples and the derived incidence of mutations in our population as a whole was 3.24% (Table S1). However, the incidence rose significantly with age from 0.2% in the 17–29 to 19.5% in the 90–98 years age group (Figure 1A). We found one or more samples with mutations at 9 of the 15 hot spot codons studied, with VAFs varying widely within and between mutation groups (Table 2).

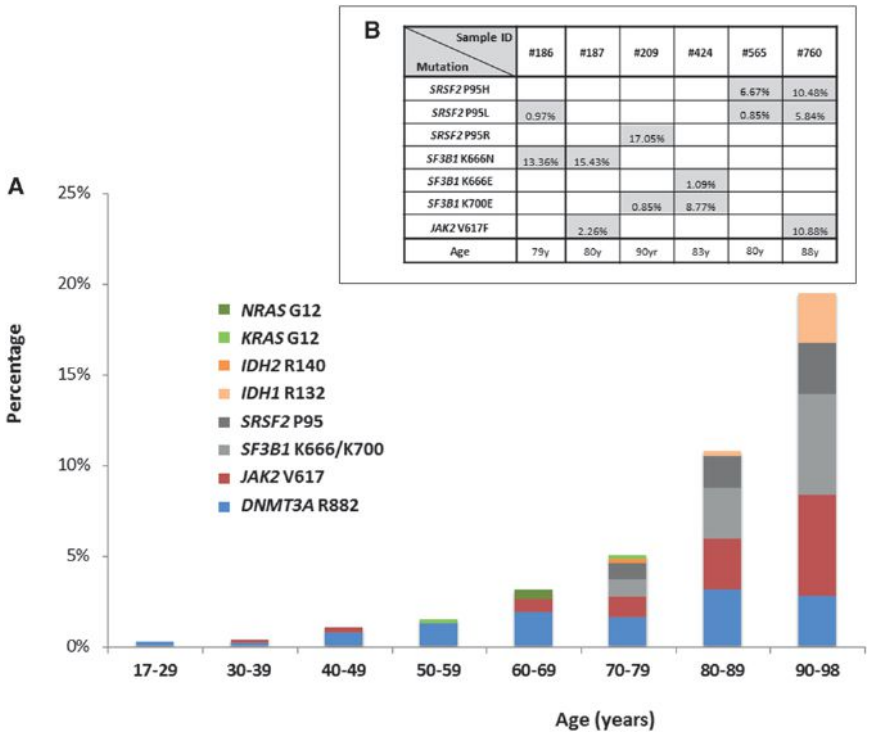
The most-common mutations were those affecting *DNMT3A* R882, whose incidence rose with age from 0.2% (1/489) in the 17–25 to a peak of 3.1% (11/355) in the 80–89 age group. A similar pattern was observed with *JAK2* V617F mutations (Figure 1A). By contrast, spliceosome gene mutations at *SRSF2* P95, *SF3B1* K666, and *SF3B1* K700 were exclusively observed in people aged over 70 years, rising sharply from 1.8% in those aged 70–79 to 8.3% in the 90–98 years age group. Among all samples, we identified only six individuals with more than one mutation; significantly, five of them had two independent spliceosome gene mutations of different VAFs (Figure 1B). Unfortunately, in each of three cases with two mutations at the same or nearby positions, neighboring SNPs were not informative and the variants could not be phased (see Supplemental Experimental Procedures).Occasional mutations in the genes *IDH1*, *IDH2*, *NRAS*, and *KRAS* were also seen. Except for three samples with IDH1/2 mutations, hemoglobin concentrations did not differ significantly between individuals with and without hot spot mutations (Figure S3A). For samples with full blood count results available, *JAK2* V617F mutant cases had a higher platelet count (albeit within the normal range) than ''no mutation cases,'' whereas other results did not differ (Figure S3B). No hot spot mutations were found in the few cord blood (n = 18) and post-transplantation (n = 32) samples studied.

Finally, despite using a very sensitive method and a mutation-calling script written specifically for this purpose, no samples with *NPM1* mutations of VAF $\geq$ 0.008 were identified. In fact, variant reads reporting a canonical *NPM1* mutation (mutation A; TCTG duplication) were detected in only 1 of 4,067 samples at a VAF of 0.0012 (4/3,466 reads).

DISCUSSION

Hematological malignancies develop through the serial acquisition of somatic mutations in a process that can take many years or even decades (Ford et al., 1998; Kyle et al., 2002). Also, it is clear that the presence of hemopoietic cells carrying leukemia-associated mutations is only followed by the onset of hematological malignancies in a minority of cases (Busque et al., 2012; Genovese et al., 2014; Jacobs et al., 2012; Jaiswal et al., 2014; Laurie et al., 2012; Xie et al., 2014). In order to understand the incidence and clonal dynamics of pre-leukemic clonal hemopoiesis, we interrogated 15 leukemia-associated mutation hot spots using a highly sensitive methodology able to detect small clones with mutations.

We show that clonal hemopoiesis is rare in the young but becomes common with advancing age. In particular, we observed that ARCH driven by the mutations studied here doubled in



B

Mutation \ Sample ID	#186	#187	#209	#424	#565	#760
SRSF2 P95H					6.67%	10.48%
SRSF2 P95L	0.97%				0.85%	5.84%
SRSF2 P95R			17.05%			
SF3B1 K666N	13.36%	15.43%				
SF3B1 K666E				1.09%		
SF3B1 K700E			0.85%	8.77%		
JAK2 V617F		2.26%				10.88%
Age	79y	80y	90yr	83y	80y	88y

Figure 1. Prevalence and Age Distribution of Hot Spot Mutations Driving Clonal Hemopoiesis

(A) Prevalence of mutations driving clonal hemopoiesis by age.

(B) Samples with more than one mutation, variant allele fraction (VAF) of each mutation present, and age of participant.

Also see Figure S1 for age distribution of all participants.

frequency in successive decades after the age of 50, rising from 1.5% in those aged 50–59 to 19.5% in those aged 90–98 (Figure 1). Of note, 61 of 112 clones identified had a VAF $\leq$ 3% (Table 2), and it is likely that most of these would not have been detected by conventional exome sequencing, which gives lower than 10-fold average coverage compared to the current study (see Table S2 for comparison to such studies), with some recurrently mutated regions giving particularly low coverage (Genovese et al., 2014). Notably, our study did not search for non-hot-spot mutations associated with ARCH such as those affecting genes *TET2* and *ASXL1* or *DNMT3A* codons other than R882 (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014). Assuming that the incidence of small clones is similar for such mutations as for the hot spot mutations we studied here, the mean projected true incidence of ARCH driven by leukemia-associated mutations in those older than 90 years is greater than 70% (Figure S4). This makes clonal hemopoiesis an almost inevitable consequence of advanced aging.

Another significant finding of our study is the disparate age distribution of ARCH associated with different mutation types. In particular, we found that, although *DNMT3A* R882 and *JAK2* V617F mutations become more common with age, they were also found in younger individuals. This is in keeping with the increasing cumulative likelihood of their stochastic acquisition with the passage of time. In contrast, spliceosome gene mutations were found exclusively in those aged 70 years or older, replicating the sharp rise beyond this age in the incidence of MDSs driven by these mutations and the fact that, among unselected MDS patients, those with spliceosome mutations are significantly older than those without (Haferlach et al., 2014; Lin et al., 2014; Papaemmanuil et al., 2013; Wu et al., 2012). Exome-sequencing studies describe a much-lower rate of spliceosome mutations (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014), but this is again likely to reflect their lower sensitivity for detecting small clones, which was a particular limitation at spliceosome mutation hot spots as these were captured/sequenced at lower-than-average depths (Table S2). In our study, 19/33 *SF3B1*- or *SRSF2*-associated clones had a VAF $\leq$ 5%, with 13 of these at VAFs $\leq$ 3% (Table 2), the majority of which would not have been detected by low-coverage sequencing. The identification of ARCH driven by spliceosome gene mutations is in keeping with the fact that these are founding mutations in the clonal evolution of MDS and related hematological malignancies (Cazzola et al., 2013; Haferlach et al., 2014; Papaemmanuil et al., 2013).

We propose that the exclusive identification of spliceosome gene mutations in those aged $\geq$70 years can be explained by differences in the prevailing pressures on clonal selection at different ages, which can in turn explain how different gene mutations can generate detectable clonal expansions at different ages (Figure 2). The alternatives are that spliceosome mutations are associated with slower rates of clonal expansion or that they are detected later because they contribute less to circulating leukocytes. Both of these scenarios are less plausible, given the complete absence of such mutations even at low VAFs in younger age groups. For any somatic mutation imparting a clonal advantage to a stem/progenitor cell and leading to the generation of a steadily expanding clone, one would expect such a clone to be detectable at a smaller size at earlier and a larger size at later time points, as is the case for *DNMT3A* R882 and *JAK2* V617 mutations. Instead, clones (of any size) driven by mutant *SRSF2* and *SF3B1* were observed exclusively in individuals aged 70 years or older, suggesting that these only begin to expand later in life. Furthermore, considerable support for the presence of a different selection milieu comes from the observation that five of six patients with multiple mutations harbored two independent spliceosome gene mutations, indicative of convergent evolution, i.e., evolution to overcome a shared selective pressure or to exploit a shared environment (Greaves and Maley, 2012; Rossi et al., 2008).

It is tempting to consider the nature of age-related changes in normal hemopoiesis that make it permissive to the outgrowth of

Table 2. Amino Acid Consequences and VAFs of the 112 Clonal Mutations Identified in This Study

Mutation Hot Spot	Codon	VAF (%)	Age	Mutation Hot Spot	Codon	VAF (%)	Age	Mutation Hot Spot	Codon	VAF (%)	Age
DNMT3A R882	p.R882H	4.14	25		p.R882H	32.02	81	*IDH1 R132*	p.R132H	42.13	84
	p.R882C	2.33	35		p.R882H	1.14	81		p.R132C	0.92	92
	p.R882H	3.80	42		p.R882H	3.06	81	*IDH2 R140*	p.R140Q	6.67	76
	p.R882H	4.00	42		p.R882H	2.17	81	*SRSF2 P95*	p.P95R	4.46	70
	p.R882H	1.25	43		p.R882H	1.13	82		p.P95L	3.35	72
	p.R882H	19.00	48		p.R882H	1.46	82		p.P95H	0.86	73
	p.R882H	1.18	49		p.R882C	2.62	82		p.P95H	0.84	77
	p.R882S	1.74	49		p.R882C	6.15	89		p.P95L	0.97	79†
	p.R882H	9.87	50		p.R882C	2.00	94		p.P95L	0.85	80††
	p.R882H	0.83	51	*JAK2V617F*	p.V617F	1.56	34		p.P95H	6.67	80††
	p.R882C	1.10	51		p.V617F	4.91	42		p.P95L	0.96	81
	p.R882C	12.50	52		p.V617F	7.72	45		p.P95H	6.40	82
	p.R882C	1.28	53		p.V617F	0.85	62		p.P95L	2.74	85
	p.R882C	2.47	54		p.V617F	25.44	64		p.P95R	7.52	87
	p.R882H	1.95	55		p.V617F	7.41	65		p.P95L	5.84	88**
	p.R882C	30.22	55		p.V617F	1.03	67		p.P95H	10.48	88**
	p.R882C	1.22	56		p.V617F	0.88	71		p.P95R	2.71	88
	p.R882H	0.91	58		p.V617F	3.75	71		p.P95R	17.05	90‡
	p.R882H	4.17	60		p.V617F	1.16	75	*SF3B1* K700	p.K700E	1.04	76
	p.R882H	5.90	60		p.V617F	2.30	77		p.K700E	6.63	81
	p.R882H	9.60	60		p.V617F	1.92	78		p.K700E	0.79	82
	p.R882H	2.73	60		p.V617F	2.26	80*		p.K700E	12.59	83
	p.R882C	9.33	60		p.V617F	4.25	80		p.K700E	8.77	83‡‡
	p.R882H	7.03	61		p.V617F	1.92	80		p.K700E	1.02	84
	p.R882C	1.21	61		p.V617F	3.71	80		p.K700E	0.85	90‡
	p.R882H	0.86	63		p.V617F	15.48	81		p.K700E	1.37	90
	p.R882H	2.54	64		p.V617F	1.21	82	*SF3B1* K666	p.K666N	1.33	70
	p.R882H	3.19	67		p.V617F	1.62	85		p.K666N	5.01	79
	p.R882H	2.74	70		p.V617F	0.83	85		p.K666N	13.36	79†
	p.R882H	4.27	74		p.V617F	1.98	86		p.K666N	15.43	80*
	p.R882H	0.85	74		p.V617F	25.94	88		p.K666N	4.60	81
	p.R882H	0.85	75		p.V617F	10.88	88**		p.K666E	1.09	83‡‡
	p.R882C	1.12	77		p.V617F	2.94	90		p.K666N	35.11	86
	p.R882C	1.15	78		p.V617F	1.23	90		p.K666N	19.70	86
	p.R882H	1.26	79	*KRAS G12*	p.G12 R	0.94	55		p.K666N	16.55	86
	p.R882H	16.66	80		p.G12S	2.78	78		p.K666E	3.34	95
	p.R882C	4.28	80	*NRAS G12*	p.G12S	1.50	61				
	p.R882C	3.66	80		p.G12D	0.96	62				

Mutations identified in the same sample are highlighted with the same symbol (*, **, †, ††, ‡, and ‡‡).

clones driven by spliceosome mutations. HSCs do not operate in isolation; instead, their normal survival and behavior are closely dependent on interactions with the hemopoietic microenvironment (Calvi et al., 2003; Rossi et al., 2008; Zhang et al., 2003). Therefore, both cell-intrinsic and microenvironmental factors influence hemopoietic aging (Rossi et al., 2008; Woolthuis et al., 2011). For example, there is good evidence for age-related changes in cell-intrinsic properties of HSCs in both mice (Chambers et al., 2007; Rossi et al., 2005) and humans (Rübe et al., 2011; Taraldsrud et al., 2009), and it is also clear that aging has a profound effect on the hemopoietic niche, reducing its ability to sustain polyclonal hemopoiesis, favoring oligo- or monoclonality instead (Vas et al., 2012). These and many other observations provide strong evidence that changes in the hemopoietic system subject HSCs to changing pressures during normal aging, driving clonal selection (Rossi et al., 2008).

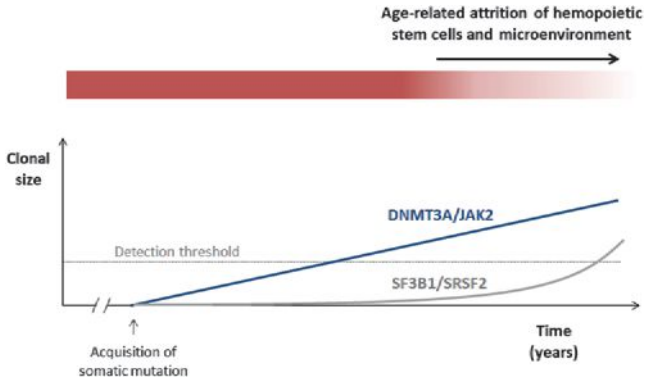


Figure 2. Proposed Kinetics of Hemopoietic Clones Driven by Different Gene Mutations

Mutations such as *DNMT3A* R882H/C or *JAK2* V617F drive a slow but inexorable clonal expansion, leading to the outgrowth of a detectable clone after a certain latency. By contrast, mutations affecting spliceosome genes, such as *SF3B1* and *SRSF2*, and acquired at the same age for the purposes of this model give no proliferative advantage initially but do so later in the context of an aging hemopoietic compartment. Their effects may operate by prolonging stem cell survival and repopulating fitness beyond that of normal stem cells or by exploiting cell-extrinsic changes in the aging microenvironment.

A striking example of such selection was described in a 115-year-old woman whose peripheral white blood cells were shown to be primarily the offspring of only two related HSC clones, whose cargo of approximately 450 somatic mutations did not include known leukemogenic mutations (Holstege et al., 2014). In the absence of somatic driver mutations, it is probable that such selection is driven by well-demonstrated epigenetic differences between individual HSCs (Fraga et al., 2005) or by stochastic events. Furthermore, clonal hemopoiesis in the absence of a known leukemia-driver mutation was also well documented recently (Genovese et al., 2014), and whereas unknown or undetected drivers may be responsible for many cases of this phenomenon, it is also highly plausible that a stochastic process of clonal selection or loss may operate in others. Our study provides evidence that spliceosome gene mutations offer a means to exploit age-related changes in hemopoiesis to drive clonal hemopoiesis in advanced old age, an observation that blurs the boundary between "driver" and "passenger" mutations. Such a context dependency is not a surprising attribute for the effects of spliceosome mutations, which have not, so far, been shown to impart a primary proliferative advantage to normal hemopoietic stem and progenitor cells (Matsunawa et al., 2014; Visconte et al., 2012).

A final important finding of our study was the almost complete absence of canonical *NPM1* mutations in our collection of more than 4,000 people, despite the use of a highly sensitive assay for their detection, designed specifically for this study. Among more than 10 million mapped reads covering this mutation hot spot, we identified only four reads in a single sample reporting a canonical mutation (mutation A; TCTG duplication). Given their frequency in myeloid leukemia (Cancer Genome Atlas Research Network, 2013) and the fact that they are not late mutations (Krönke et al., 2013; Shlush et al., 2014), this observation frames *NPM1* mutations as "gatekeepers" of leukemogenesis, i.e., their acquisition appears to be closely associated with the development of frank leukemia. In this light, the frequent co-occurrence of *DNMT3A* and *NPM1* mutations suggests that the former behave as "rafts" that enable *NPM1* mutant clones to be founded and expanded, thus facilitating onward evolution toward acute myeloid leukemia.

We used a highly sensitive method to search for evidence of clonal hemopoiesis driven by 15 recurrent leukemogenic mutations in more than 4,000 individuals. Our results demonstrate that the incidence of clonal hemopoiesis is much higher than suggested by exome-sequencing studies, that spliceosome gene mutations drive clonal outgrowth primarily in the context of an aging hemopoietic compartment, and that *NPM1* mutations do not drive ARCH, indicating that their acquisition is closely associated with frank leukemia.

EXPERIMENTAL PROCEDURES

Patient Samples

Samples were obtained with written informed consent and in accordance with the Declaration of Helsinki and appropriate ethics committee approvals from all participants (approval reference numbers 10/H0604/02, 07/MRE05/44, and 05/Q0106/74). Maternal consent was obtained for the use of cord blood samples. Samples were obtained from 3,067 blood donors aged 17–70 years (WTCCC; UK Blood Services 1 [UKBS1] and UKBS2 common controls), 1,152 unselected individuals aged 60–98 years (UKHLS; https://www.understandingsociety.ac.uk/), 32 patients that had undergone a hemopoietic stem cell transplant (12 autologous and 20 allogeneic; Tables S3 and S4) 1 month to 14 years previously, and 18 cord blood samples. Age distribution of the WTCCC and UKHLS cohorts/samples is shown in Figure S1. Hemoglobin concentrations were available for a total of 3,587 of the 4,067 samples from which adequate sequencing data were obtained for analysis, including 102 of 105 samples with mutations. Full blood count results were available for 2,952 WTCCC samples. The average blood donation frequency for WTCCC donors was 1.6 donations of one unit per year. Details of donations by individual participants were not available.

Targeted Sequencing

Genomic DNA was used to simultaneously amplify several gene loci using multiplex PCR, in order to capture and analyze 15 mutational hot spots enriched for, but not exclusive to, targets of mutations thought to arise early in leukemogenesis (Table 1). We used three multiplex primer combinations (Plex1-3), guided by our findings, to capture the targeted mutational hot spots (Table S1). Primers were designed using the Hi-Plex PCR-MPS (massively parallel sequencing) strategy (Nguyen-Dumont et al., 2013), except for *JAK2* V617 and "Plex2" primers, which were designed using MPRIMER (Shen et al., 2010). These and additional primer sequences used in each Plex and details of PCR- and DNA-sequencing protocols are detailed in Supplemental Experimental Procedures. Methodological validation experiments are shown in Figure S2.

Bioinformatic Analysis

Sequencing data were aligned to the human reference genome (hg19) using BWA. Subsequently, the SAMTOOLS pileup command was used to generate pileup files from the generated bam files (version 0.1.8; http://samtools.sourceforge.net; Li et al., 2009). A flexible in-house Perl script generated by our group, MIDAS (Conte et al., 2013), was modified in order to interrogate only the hot spot nucleotide positions of interest (those with reported mutations in the COSMIC database; Forbes et al., 2015) on the pileup file, considering only those reads with a sequence quality higher than 25 and a mapping quality higher than 15. For each sample, the numbers of reads reporting the reference and variant alleles at each position were extracted. VAFs were derived by dividing the number of reads reporting the most-frequent variant nucleotide to the total. In order to detect NPM1 mutations with high sensitivity,

we wrote a bespoke Perl script described in Supplemental Experimental Procedures.

Statistical Analyses and Mutation-Calling Threshold

We chose a threshold VAF of $\geq$0.008 (0.8%) to "call" clones with a heterozygous mutation representing $\geq$1.6% of blood leukocytes. From validation experiments and data analysis (see Supplemental Experimental Procedures and Figure S2D), we determined that the maximum false-positive error rate for calling a mutation (VAF $\geq$ 0.008) due to variant allele counts that are solely due to PCR-MiSeq error was negligible ($p < 10^{-5}$). For comparisons of blood cell counts and hemoglobin concentrations, we used non-paired t tests. For summary statistics of read coverage (Table S2) and for the purposes of deriving an estimate of the overall incidence of clonal hemopoiesis (Figure S4), we used published tables of all mutations reported by three recent studies that employed whole-exome-sequencing analyses to identify individuals with clonal hemopoiesis (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014).

ACCESSION NUMBERS

The European Genome-Phenome Archive (EGA) accession number for the sequencing data reported in this paper is EGAS00001000814.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2015.02.005.

AUTHOR CONTRIBUTIONS

G.S.V. conceived and designed the study. G.S.V. and T. McKerrell supervised the study, analyzed data, and wrote the manuscript. N.P. and T. McKerrell performed experimental procedures. I.V. and T. Moreno wrote scripts and performed bioinformatics analysis. H.P., T. McKerrell, and G.S.V. performed statistical analyses. E.Z., C.S.G., M.A.Q., and R.R. contributed to study strategy and to technical and analytical aspects. U.S.S.G., E.Z., W.O, J.C., C.C., J.B., J.S., C.H., M.A.S., and D.R.F. contributed to sample acquisition and subject recruitment.

ACKNOWLEDGMENTS

This project was funded by a Wellcome Trust Clinician Scientist Fellowship (100678/Z/12/Z; to T. McKerrell) and by the Wellcome Trust Sanger Institute (grant number WT098051). G.S.V. is funded by a Wellcome Trust Senior Fellowship in Clinical Science (WT095663MA), and work in his laboratory is also funded by Leukaemia Lymphoma Research and the Kay Kendal Leukaemia Fund. I.V. is funded by Spanish Ministerio de Economía y Competitividad subprograma Ramón y Cajal. C.S.G. is funded by a Leukaemia Lymphoma Research Clinical Research Training Fellowship. We thank Servicio Santander Supercomputación for their support. We acknowledge use of DNA from The UK Blood Services Collection of Common Controls (UKBS collection), funded by the Wellcome Trust grant 076113/C/04/Z, by the Juvenile Diabetes Research Foundation grant WT061858, and by the National Institute of Health Research of England. The collection was established as part of the Wellcome Trust Case-Control Consortium. We also gratefully acknowledge use of blood DNA samples and data from participants of the UK Household Longitudinal Study (https://www.understandingsociety.ac.uk/), collected by NatCen and the Institute for Social and Economic Research, University of Essex, and funded by the Economic and Social Research Council, UK. We thank the Cambridge Blood and Stem Cell Biobank and the Cancer Molecular Diagnosis Laboratory, Cambridge Biomedical Research Centre (National Institute for Health Research, UK) for help with sample collection and processing. Finally, we thank Nathalie Smerdon, Richard Rance, Lucy Hildyard, Ben Softly, and Britt Killian for help with sample management, DNA sequencing, and data processing. G.S.V. is a consultant for KYMAB and receives an educational grant from Celgene.

Received: December 14, 2014
Revised: January 19, 2015
Accepted: January 29, 2015
Published: February 26, 2015

REFERENCES

Busque, L., Patel, J.P., Figueroa, M.E., Vasanthakumar, A., Provost, S., Hamilou, Z., Mollica, L., Li, J., Viale, A., Heguy, A., et al. (2012). Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. Nat. Genet. *44*, 1179–1181.

Calvi, L.M., Adams, G.B., Weibrecht, K.W., Weber, J.M., Olson, D.P., Knight, M.C., Martin, R.P., Schipani, E., Divieti, P., Bringhurst, F.R., et al. (2003). Osteoblastic cells regulate the haematopoietic stem cell niche. Nature *425*, 841–846.

Cancer Genome Atlas Research Network (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N. Engl. J. Med. *368*, 2059–2074.

Cazzola, M., Della Porta, M.G., and Malcovati, L. (2013). The genetic basis of myelodysplasia and its clinical relevance. Blood *122*, 4021–4034.

Chambers, S.M., Shaw, C.A., Gatza, C., Fisk, C.J., Donehower, L.A., and Goodell, M.A. (2007). Aging hematopoietic stem cells decline in function and exhibit epigenetic dysregulation. PLoS Biol. *5*, e201.

Conte, N., Varela, I., Grove, C., Manes, N., Yusa, K., Moreno, T., Segonds-Pichon, A., Bench, A., Gudgin, E., Herman, B., et al. (2013). Detailed molecular characterisation of acute myeloid leukaemia with a normal karyotype using targeted DNA capture. Leukemia *27*, 1820–1825.

Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature *481*, 506–510.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. *43*, D805–D811.

Ford, A.M., Bennett, C.A., Price, C.M., Bruin, M.C., Van Wering, E.R., and Greaves, M. (1998). Fetal origins of the TEL-AML1 fusion gene in identical twins with leukemia. Proc. Natl. Acad. Sci. USA *95*, 4584–4588.

Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suñer, D., Cigudosa, J.C., Urioste, M., Benitez, J., et al. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. Proc. Natl. Acad. Sci. USA *102*, 10604–10609.

Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M., et al. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. *371*, 2477–2487.

Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306–313.

Haferlach, T., Nagata, Y., Grossmann, V., Okuno, Y., Bacher, U., Nagae, G., Schnittger, S., Sanada, M., Kon, A., Alpermann, T., et al. (2014). Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. Leukemia *28*, 241–247.

Holstege, H., Pfeiffer, W., Sie, D., Hulsman, M., Nicholas, T.J., Lee, C.C., Ross, T., Lin, J., Miller, M.A., Ylstra, B., et al. (2014). Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. Genome Res. *24*, 733–742.

Jacobs, K.B., Yeager, M., Zhou, W., Wacholder, S., Wang, Z., Rodriguez-Santiago, B., Hutchinson, A., Deng, X., Liu, C., Horner, M.J., et al. (2012). Detectable clonal mosaicism and its relationship to aging and cancer. Nat. Genet. *44*, 651–658.

Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., et al. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. *371*, 2488–2498.

Krönke, J., Bullinger, L., Teleanu, V., Tschürtz, F., Gaidzik, V.I., Kühn, M.W., Rücker, F.G., Holzmann, K., Paschka, P., Kapp-Schwörer, S., et al. (2013). Clonal evolution in relapsed NPM1-mutated acute myeloid leukemia. Blood *122*, 100–108.

Kyle, R.A., Therneau, T.M., Rajkumar, S.V., Offord, J.R., Larson, D.R., Plevak, M.F., and Melton, L.J., 3rd. (2002). A long-term study of prognosis in monoclonal gammopathy of undetermined significance. N. Engl. J. Med. *346*, 564–569.

Laurie, C.C., Laurie, C.A., Rice, K., Doheny, K.F., Zelnick, L.R., McHugh, C.P., Ling, H., Hetrick, K.N., Pugh, E.W., Amos, C., et al. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat. Genet. *44*, 642–650.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lin, C.C., Hou, H.A., Chou, W.C., Kuo, Y.Y., Wu, S.J., Liu, C.Y., Chen, C.Y., Tseng, M.H., Huang, C.F., Lee, F.Y., et al. (2014). SF3B1 mutations in patients with myelodysplastic syndromes: the mutation is stable during disease evolution. Am. J. Hematol. *89*, E109–E115.

Matsunawa, M., Yamamoto, R., Sanada, M., Sato-Otsubo, A., Shiozawa, Y., Yoshida, K., Otsu, M., Shiraishi, Y., Miyano, S., Isono, K., et al. (2014). Haploinsufficiency of Sf3b1 leads to compromised stem cell function but not to myelodysplasia. Leukemia *28*, 1844–1850.

Nguyen-Dumont, T., Pope, B.J., Hammet, F., Southey, M.C., and Park, D.J. (2013). A high-plex PCR approach for massively parallel sequencing. Biotechniques *55*, 69–74.

Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23–28.

Papaemmanuil, E., Gerstung, M., Malcovati, L., Tauro, S., Gundem, G., Van Loo, P., Yoon, C.J., Ellis, P., Wedge, D.C., Pellagatti, A., et al.; Chronic Myeloid Disorders Working Group of the International Cancer Genome Consortium (2013). Clinical and biological implications of driver mutations in myelodysplastic syndromes. Blood *122*, 3616–3627, quiz 3699.

Rossi, D.J., Bryder, D., Zahn, J.M., Ahlenius, H., Sonu, R., Wagers, A.J., and Weissman, I.L. (2005). Cell intrinsic alterations underlie hematopoietic stem cell aging. Proc. Natl. Acad. Sci. USA *102*, 9194–9199.

Rossi, D.J., Jamieson, C.H., and Weissman, I.L. (2008). Stems cells and the pathways to aging and cancer. Cell *132*, 681–696.

Rübe, C.E., Fricke, A., Widmann, T.A., Fürst, T., Madry, H., Pfreundschuh, M., and Rübe, C. (2011). Accumulation of DNA damage in hematopoietic stem and progenitor cells during human aging. PLoS ONE *6*, e17487.

Shen, Z., Qu, W., Wang, W., Lu, Y., Wu, Y., Li, Z., Hang, X., Wang, X., Zhao, D., and Zhang, C. (2010). MPprimer: a program for reliable multiplex PCR primer design. BMC Bioinformatics *11*, 143.

Shlush, L.I., Zandi, S., Mitchell, A., Chen, W.C., Brandwein, J.M., Gupta, V., Kennedy, J.A., Schimmer, A.D., Schuh, A.C., Yee, K.W., et al.; HALT Pan-Leukemia Gene Panel Consortium (2014). Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. Nature *506*, 328–333.

Taraldsrud, E., Grøgaard, H.K., Solheim, S., Lunde, K., Fløisand, Y., Arnesen, H., Seljeflot, I., and Egeland, T. (2009). Age and stress related phenotypical changes in bone marrow CD34+ cells. Scand. J. Clin. Lab. Invest. *69*, 79–84.

Vas, V., Senger, K., Dörr, K., Niebel, A., and Geiger, H. (2012). Aging of the microenvironment influences clonality in hematopoiesis. PLoS ONE *7*, e42080.

Visconte, V., Rogers, H.J., Singh, J., Barnard, J., Bupathi, M., Traina, F., McMahon, J., Makishima, H., Szpurka, H., Jankowska, A., et al. (2012). SF3B1 haploinsufficiency leads to formation of ring sideroblasts in myelodysplastic syndromes. Blood *120*, 3173–3186.

Woolthuis, C.M., de Haan, G., and Huls, G. (2011). Aging of hematopoietic stem cells: Intrinsic changes or micro-environmental effects? Curr. Opin. Immunol. *23*, 512–517.

Wu, S.J., Kuo, Y.Y., Hou, H.A., Li, L.Y., Tseng, M.H., Huang, C.F., Lee, F.Y., Liu, M.C., Liu, C.W., Lin, C.T., et al. (2012). The clinical implication of SRSF2 mutation in patients with myelodysplastic syndrome and its stability during disease evolution. Blood *120*, 3106–3111.

Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., Wendl, M.C., McMichael, J.F., Schmidt, H.K., Yellapantula, V., Miller, C.A., et al. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat. Med. *20*, 1472–1478.

Zhang, J., Niu, C., Ye, L., Huang, H., He, X., Tong, W.G., Ross, J., Haug, J., Johnson, T., Feng, J.Q., et al. (2003). Identification of the haematopoietic stem cell niche and control of the niche size. Nature *425*, 836–841.

OPEN ACCESS
CellPress

Cell Reports
Report

Accelerating Novel Candidate Gene Discovery in Neurogenetic Disorders via Whole-Exome Sequencing of Prescreened Multiplex Consanguineous Families

Anas M. Alazami,[1,23] Nisha Patel,[1,23] Hanan E. Shamseldin,[1,23] Shamsa Anazi,[1] Mohammed S. Al-Dosari,[2] Fatema Alzahrani,[1] Hadia Hijazi,[1] Muneera Alshammari,[3] Mohammed A. Aldahmesh,[1] Mustafa A. Salih,[3] Eissa Faqeih,[4] Amal Alhashem,[5,6] Fahad A. Bashiri,[3] Mohammed Al-Owain,[5,7] Amal Y. Kentab,[3] Sameera Sogaty,[8] Saeed Al Tala,[9] Mohamad-Hani Temsah,[3] Maha Tulbah,[10] Rasha F. Aljelaify,[11] Saad A. Alshahwan,[6] Mohammed Zain Seidahmed,[12] Adnan A. Alhadid,[3] Hesham Aldhalaan,[13] Fatema AlQallaf,[13] Wesam Kurdi,[10] Majid Alfadhel,[14] Zainab Babay,[15] Mohammad Alsogheer,[16] Namik Kaya,[1] Zuhair N. Al-Hassnan,[5,7] Ghada M.H. Abdel-Salam,[17] Nouriya Al-Sannaa,[18] Fuad Al Mutairi,[14] Heba Y. El Khashab,[3,19] Saeed Bohlega,[13] Xiaofei Jia,[20] Henry C. Nguyen,[20] Rakad Hammami,[1] Nouran Adly,[1] Jawahir Y. Mohamed,[1] Firdous Abdulwahab,[1] Niema Ibrahim,[1] Ewa A. Naim,[1,21] Banan Al-Younes,[1,21] Brian F. Meyer,[1,21] Mais Hashem,[1] Ranad Shaheen,[1] Yong Xiong,[20] Mohamed Abouelhoda,[1,21] Abdulrahman A. Aldeeri,[1,22] Dorota M. Monies,[1,21] and Fowzan S. Alkuraya[1,5,21,*]

[1]Department of Genetics, King Faisal Specialist Hospital and Research Center, Riyadh 11211, Saudi Arabia
[2]Department of Pharmacognosy, College of Pharmacy, King Saud University, Riyadh 11451, Saudi Arabia
[3]Department of Pediatrics, King Khalid University Hospital and College of Medicine, King Saud University, Riyadh 11451, Saudi Arabia
[4]Department of Pediatrics, King Fahad Medical City, Riyadh 11525, Saudi Arabia
[5]Department of Anatomy and Cell Biology, College of Medicine, Alfaisal University, Riyadh 11533, Saudi Arabia
[6]Department of Pediatrics, Prince Sultan Military Medical City, Riyadh 11159, Saudi Arabia
[7]Department of Medical Genetics, King Faisal Specialist Hospital and Research Center, Riyadh 11211, Saudi Arabia
[8]Department of Pediatrics, King Fahad General Hospital, Jeddah 23325, Saudi Arabia
[9]Department of Pediatrics, Armed Forces Hospital, Khamis Mushayt 62413, Saudi Arabia
[10]Department of Obstetrics & Gynecology, King Faisal Specialist Hospital, Riyadh 11211, Saudi Arabia
[11]Center of Excellence for Genomics, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia
[12]Department of Pediatrics, Security Forces Hospital, Riyadh 12625, Saudi Arabia
[13]Department of Neurosciences, King Faisal Specialist Hospital and Research Center, Riyadh 11211, Saudi Arabia
[14]Division of Genetics, Department of Pediatrics, King Saud bin Abdulaziz University for Health Sciences, King Abdulaziz Medical City, Riyadh 14611, Saudi Arabia
[15]Department of Obstetrics and Gynecology, College of Medicine, King Saud University, Riyadh 11451, Saudi Arabia
[16]Department of Psychiatry, College of Medicine, King Saud University, Riyadh 11451, Saudi Arabia
[17]Department of Clinical Genetics, Human Genetics and Genome Research Division, National Research Centre, Cairo 12345, Egypt
[18]Department of Pediatrics, Johns Hopkins Aramco Healthcare, Dhahran 34465, Saudi Arabia
[19]Department of Pediatrics, Children's Hospital, Ain Shams University, Cairo 01234, Egypt
[20]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA
[21]Saudi Human Genome Program, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia
[22]Department of Internal Medicine, College of Medicine, King Saud University, Riyadh 11451, Saudi Arabia
[23]Co-first author
*Correspondence: falkuraya@kfshrc.edu.sa
http://dx.doi.org/10.1016/j.celrep.2014.12.015

SUMMARY

Our knowledge of disease genes in neurological disorders is incomplete. With the aim of closing this gap, we performed whole-exome sequencing on 143 multiplex consanguineous families in whom known disease genes had been excluded by autozygosity mapping and candidate gene analysis. This prescreening step led to the identification of 69 recessive genes not previously associated with disease, of which 33 are here described (*SPDL1*, *TUBA3E*, *INO80*, *NID1*, *TSEN15*, *DMBX1*, *CLHC1*, *C12orf4*, *WDR93*, *ST7*, *MATN4*, *SEC24D*, *PCDHB4*, *PTPN23*, *TAF6*, *TBCK*, *FAM177A1*, *KIAA1109*, *MTSS1L*, *XIRP1*, *KCTD3*, *CHAF1B*, *ARV1*, *ISCA2*, *PTRH2*, *GEMIN4*, *MYOCD*, *PDPR*, *DPH1*, *NUP107*, *TMEM92*, *EPB41L4A*, and *FAM120AOS*). We also encountered instances in which the phenotype departed significantly from the established clinical presentation of a known disease gene. Overall, a likely causal mutation was identified in >73% of our cases. This study contributes to the global effort toward a full compendium of disease genes affecting brain function.

INTRODUCTION

Neurogenetic disorders represent the largest category of Mendelian diseases in humans. They encompass a wide array of

clinical presentations that range from the common e.g., intellectual disability (>1%) to the very rare, e.g., neurodegeneration with brain iron accumulation (one to three per 10^6) (Kalman et al., 2012; Maulik et al., 2011). The highly prevalent involvement of the nervous system in many Mendelian disorders coincides with the observation that >80% of all human genes are expressed at some stage of brain development (Hawrylycz et al., 2012) and suggests that the brain is one of the most vulnerable organs to genetic perturbation. In fact high-resolution microarray analysis of the human genome reveals that intellectual disability is the common phenotypic denominator of genomic disorders that involve losses or gains of genes (Coe et al., 2012).

Variances in clinical presentation are a major obstacle in establishing a working molecular classification of neurological disease, because even where the clinical presentation is highly specific, genetic heterogeneity is the rule. In the setting of autosomal recessive neurogenetic disorders where parents are related, a homozygosity scan can serve as a guide to the underlying genetic cause even when the phenotype is atypical (Alkuraya, 2010). Another major challenge in assigning a molecular classification is that many neurological disease genes have not been identified yet.

Novel disease gene discovery in this field has been tremendously abetted by next-generation sequencing, a tool with the capacity to, theoretically, unravel the genetic cause of all neurological diseases. This full theoretical potential has not yet been reached unfortunately, although the technology continues to evolve. For example, two large studies on the genetics of intellectual disability using whole-exome sequencing (WES) provided a yield of 16%–55%, and even though the collective sample size was >150, only seven novel genes were identified (de Ligt et al., 2012; Rauch et al., 2012). In these studies, samples could not be enriched for novel gene discovery, and simultaneously the anticipated mutations were heterozygous, detection of which poses a challenge for the currently available sequencing technology (especially regarding insertions/deletions) (Harismendy et al., 2009). The presence of these two obstacles likely hindered the authors' ability to obtain a higher yield. Thus, alternative/complementary approaches are required to facilitate the discovery of novel neurogenetic disease genes. In this study, we show that the analysis of the entire set of autozygous intervals per individual (the autozygome) in multiplex consanguineous patients, as a prescreen, can markedly increase the yield of WES to identify candidate genes not previously associated with disease. Even when known genes were identified using this approach, the phenotype was often sufficiently different to explain why the gene had been missed by the autozygosity filter. The 33 candidate disease genes we showcase in this study will augment the global hunt for the genetics of brain development and function.

RESULTS

Clinical Report

In total, 143 multiplex families met our inclusion criteria (a neurogenetic diagnosis, positive family history, consanguineous parents, and no candidates identified by autozygosity mapping). Intellectual disability was the most common clinical feature. Other phenotypes that were also enriched included global developmental delay, autism, epilepsy, primary microcephaly, ataxia, and neurodegeneration. Table S1 summarizes the clinical features of the entire cohort.

Autozygosity Mapping Is a Powerful Enrichment Tool for Novel Candidate Gene Discovery

As a prescreening step for each case, the regions of homozygosity (ROH), the telltale sign of shared ancestral haplotypes, were interrogated for disease genes that matched the patients' phenotype. These genes were prioritized for Sanger sequencing. If negative, or when no compelling disease genes were evident in the ROH, patient DNA was subjected to WES under the assumption that this will reveal a novel disease gene. This assumption fails, however, to account for certain scenarios. When disease genes within the ROH are examined for a likely candidate, it is possible that the clinical picture may sufficiently deviate from the classical phenotype ascribed to a particular gene such that the gene falls outside our consideration. Specifically, we list in Table 1 cases in which there was sufficient discrepancy between the classical and observed phenotypes that the respective genes were missed in the autozygosity mapping stage. Some phenotypes can even be considered unique rather than an expansion of a known phenotype.

A second reason why WES may reveal a known disease gene is that the causal mutation may lie within an ROH but fail to be detected, which is one of the known "pitfalls" of autozygosity mapping (Alkuraya, 2012). Figure S1 captures the homozygosity pattern of a number of patients for whom the candidate ROH was missed, because it did not meet our ROH size cutoff, or it appeared to be shared by an unaffected member of the family, or the overlap of the gene with the ROH was not clearly discernible. One case (11DG0165) evaded detection due to the presence of a deep intronic mutation, which was later uncovered using RT-PCR. A third scenario for why WES may expose a known disease gene is that the gene was simply overlooked when searching for plausible candidates within an ROH. This occurred in 13DG1803 where *TUSC3*, a known intellectual disability gene, was not noticed during the prescreening stage. Further scenarios include incomplete clinical information at the time of analysis (as occurred with 09DG-00774 and 12DG0926) and poor description in the literature (discussed below). We emphasize, however, that cases of failed autozygosity map prescreening are the exception rather than the rule because they represent only ~7.7% (11/143) of all cases and have reduced the hypothetical yield of WES in revealing candidate genes not associated with disease, or known genes with unique phenotypes, by only 2.2%.

WES Is a Powerful Novel Candidate Gene Discovery Tool in Neurogenetic Disorders

Consistent with our prescreening enrichment step, WES revealed 69 genes that were not known to the authors at the time of exome capture. Of these, 36 have since been published by either us or others (detailed in Table S1), leaving 33 candidate genes that are here reported. The phenotypes associated with these are described in Table 2 and include intellectual disability, autism, progressive cerebellar atrophy, primary microcephaly, brain atrophy and other malformations, and myopathy. Of these

Table 1. Cases with Mutations in Known Disease Genes following WES, Where the Patient Phenotype Diverged from the Established Literature

ID	Gene	Mutation	Published Phenotype	Observed Phenotype	Reference
09DG0057	*GM2A*	NM_000405.4:c.164C > T:p.P55L	GM2-gangliosidosis, AB variant	progressive neurodegeneration with onset at 8 years, no organomegaly and normal retina	this study
12DG0096	*SPG20*	NM_001142294:c.1450_1451insA:p.T484fs	spastic paraplegia	speech and motor delay, tremor, microcephaly, and strabismus	this study
12DG1571	*CYP27A1*	NM_000784:c.1342C > T:p.R448C	cerebrotendinous xanthomatosis	severe choreoathetosis, no cataract, and normal cholestanol	this study
10DG0672	*NPC2*	NM_006432:c.88G > A: p.V30M	Nieman-Pick disease	microcephaly, static encephalopathy, no organomegaly, and normal retina	this study
11DG1951	*ARFGEF2*	NM_006420:c.656_657insC:p.P219fs	periventricular heterotopia with microcephaly	global developmental delay, epilepsy, and hydrocephalus	this study
11DG1510	*PNKP*	NM_007254:c.1250_1251insAACGGGTCGCCATCGAC:p.R418Tfs*55	progressive microcephaly, infantile-onset seizures, and developmental delay	primary microcephaly, global developmental delay, no seizures	this study
12DG0975	*RYR1*	NM_000540:c.6617C > T:p.T2206M	minicore myopathy	ptosis, no motor delay, and sacral agenesis	this study
12DG0104	*FBN2*	NM_001999:c.1064G > A:p.G355D	congenital contractural arachnodactyly	fetal akinesia with brain ischemia and neonatal death	this study
08DG00385	*BRCA2*	NM_000059.3:c.9152 delC:p.P3051Hfs*11	Fanconi anemia	primordial dwarfism	Shaheen et al. (2014)
10DG1721	*EVC2*	NM_147127.4:c.3870_3893 dup:p.K1293_K1300 dup	Ellis-van-Creveld syndrome	Meckel-Gruber syndrome	Shaheen et al. (2013)
13DG0583	*DDHD2*	NM_015214.2:c.1249_1891 del:p.A417Mfs*23 (large scale deletion)	spastic paraplegia	isolated cerebellar atrophy	this study
13DG0010	*WDR81*	NM_001163809:c.845G > A:p.G282E	cerebellar ataxia, mental retardation, and dysequilibrium syndrome-2	neonatal death due to severe brain malformation (hydranencephaly and severe cerebellar hypoplasia)	this study
12DG0685	*ZNF526*	NM_133444:c.479A > C:p.K160T	nonsyndromic intellectual disability	intellectual disability, Noonan-like facies, and pulmonary stenosis	this study
09DG00930	*ADCK3*	NM_020247:c.1744 dup:p.S582Kfs*148	cerebellar ataxia, seizure and cerebellar atrophy	isolated cerebellar hypoplasia	this study
09DG0301	*SBF1*	NM_002972:c.1327G > A:p.D443N	Charcot-Marie-Tooth disease type 4B3	Charcot-Marie-Tooth with microcephaly, ophthalmoplegia and syndactyly	Alazami et al. (2014)
11DG1767	*AP4M1*	NM_004722:c.C952T:p.R318*	spastic paraplegia, severe ID, poor speech development	microcephaly, speech delay, spasticity; brain MRI: hypomyelination, hypoplastic corpus callosum, and brain atrophy.	this study

mutations, ten are truncating or located at splice sites. For the missense changes, we determined pathogenicity based on the in silico prediction of at least two established algorithms, as well as 3D modeling of wild-type and mutant residues whenever structure information on homologous proteins was available (Figure S2). Our minimum threshold for assigning candidacy to a gene was that the variant had to be the only one to survive the stringent pipeline illustrated in Figure 2. A total of 37 cases remain "unsolved," some because more than one variant survived our filters. Table S1 lists these, and it is important to note that some of these variants may indeed represent bona fide novel disease genes.

In addition, WES revealed what appears to be phenotypes that have not been described for the respective genes in 16 cases (Table 1). One striking example is case 10DG0672 in which we identified a previously reported homozygous mutation in *NPC2*, a known gene for Nieman-Pick disease. Neither the index nor his sibling had the typical presentation of progressive neurodegeneration or hepatosplenomegaly, thus making the diagnosis of Nieman-Pick nearly impossible on clinical grounds. Finally, WES also revealed mutations in known genes for cases with classical phenotypes, where the gene was missed either due to a pitfall in autozygosity mapping or incomplete phenotyping (nine cases, discussed above) or because the phenotype was not well described in the literature as in two instances. The first such instance is *MGAT2*-related dysmorphism, which had only been documented by a single photograph (Cormier-Daire et al., 2000). The second instance involves *CTSD*, which was described as causing congenital neuronal ceroid lipofuscinosis when, in fact, the detailed description of the case was severe microlissencephaly and hyperekplexia, which is identical to the phenotype of our case 09DG00288 (Fritchie et al., 2009). Overall, the yield of WES in our cohort was 105 out of 143 (73.4%).

Although we used autozygosity mapping coupled with WES, an alternative strategy is to simply use the mapping information provided by WES analysis. Although most autozygous regions can indeed be inferred by this second strategy, we had previously shown that the use of high-throughput genotyping results in cleaner ROH data with sharper overall resolution (Carr et al., 2013).

3D Modeling of Missense Variants to Support Pathogenicity

3D modeling was performed for selected gene products using the web-based homology modeling engine Phyre2 (Kelley and Sternberg, 2009). Four of the models were built with very high confidence and suggested a pathogenic nature for the identified mutations. In the case of TUBA3E, the model was built based on the structure of tubulin alpha-3E (PDB ID: 3EDL)(Tan et al., 2008) that shares 97% sequence identity with the *TUBA3E* gene product (Figure S2A). The R215C mutation changes the charge distribution on the surface of the protein. Although not participating in microtubule formation (Tan et al., 2008), this site might be involved in the binding of microtubule to other proteins or cofactors. The structure of the gene product of *TSEN15*, human tRNA splicing endonuclease, has been solved by nuclear magnetic resonance (Song and Markley, 2007) (PDB ID: 2GW6). The conserved W76 is embedded inside the protein and is critical in protein folding (Figure S2B). Mutation of W76G would leave a void in the core of the protein and very likely lead to misfolded proteins. *PTRH2* encodes peptidyl-tRNA hydrolase 2. The crystal structure of a domain of PTRH2 has been solved (PDB ID: 1QSR). The observed mutation at Q85 is located in the middle of a helix and forms a hydrogen bond with the side chain of T157 (Figure S2C). Mutation of Q85P would destabilize PTRH2 as it would disrupt the hydrogen bond. Furthermore, the mutation to a proline may cause a kink in the helix and distort the overall fold of the structure. Iron-sulfur cluster assembly 2, or ISCA2, was also modeled with the structure of its homolog of IscA from Thermosynechococcus elongates (PDB ID: 1X0G), which is a $\alpha 2\beta 2$ heterotetramer. ISCA2 has modest sequence identity (27% to PDB ID: 1X0G) to the α chain of the IscA. However, Phyre2 predicted the same structural fold with 100% confidence. The protein is involved in the maturation of iron-sulfur proteins. According to the model, the G77S mutation occurs in a loop that is directly involved in iron-sulfur cluster binding by providing a chelating cysteine, C79 (Figure S2D). We speculate that such a mutation might affect the stability or flexibility of the loop and therefore interfere with its efficiency of binding to the iron-sulfur cluster.

DISCUSSION

Several attempts have been made in the recent past to accelerate the discovery of novel neurological disease genes. One of the earliest attempts was the high-throughput Sanger sequencing of all coding exons on the X chromosome in a large cohort of >200 families with suspected X-linked intellectual disability (Tarpey et al., 2009). In addition to the laborious nature of this approach, the yield was somewhat modest (three novel genes) partly because enrichment for novel gene discovery was not feasible, and also partly due to a large proportion of X-linked disease genes having already been established (de Brouwer et al., 2007). High-resolution molecular karyotyping is a powerful tool to identify a large number of DNA gains and losses that are associated with various neurological phenotypes, but the yield is typically <15%, and it rarely identifies single genes due to the nature of the assay (Miller et al., 2010). Morrow et al. used autozygosity mapping in nearly 90 consanguineous families with autism, followed by Sanger sequencing of candidate genes within the linked ROH, to identify five novel autism genes (Morrow et al., 2008). The lower yield of that study likely originates from the use of conventional sequencing methods, coupled with the potentially non-Mendelian behavior of autism genes.

The advent of next-generation sequencing has revolutionized the search for Mendelian neurocognitive genes. Rauch et al. and de Ligt et al. studied >150 cases of intellectual disability using a trio-exome design, and, although that approach is compatible with identifying recessive disease genes, they only identified heterozygous de novo mutations, including seven novel genes (de Ligt et al., 2012; Rauch et al., 2012). The known bias of current WES against heterozygous mutations, especially insertions and deletions, as well as the inability of the investigators to enrich their cohort for novel genes are likely explanations for the lower

Table 2. Cases with Mutations in Candidate Genes, along with Available Evidence from the Literature to Support Candidacy

ID	Phenotype	Gene	Mutation	Mutation Type	Gene Description	Supporting Evidence	Reference
12DG1528	primary microcephaly and neonatal death	*CCDC99 (SPDL1)*	NM_017785:c.1724_1747 del:p.S575_T582 del	in-frame deletion	spindle apparatus coiled-coil protein 1	CCDC99 has been demonstrated to control poleward movement of chromosomes along the mitotic spindles. Many primary microcephaly genes encode proteins that are involved in mitotic spindle regulation, e.g., ASPM and WDR62. Only surviving variant and segregates in family.	PMID: 20427577 and 24875059
11DG0443	microlissencephaly and global developmental delay	*TUBA3E*	NM_207312:c.643C > T:p.R215C	missense	tubulin, alpha 3e	Mutations in several tubulins have been linked to lissencephaly and other cortical malformations (TUBA1 A, TUBA8, TUBB2B, TUBB3, TUBB5, and TUBG1). Only surviving variant and segregates in family.	PMID: 24860126
10DG1705	primary microcephaly and global developmental delay	*INO80*	INO80:NM_017553: c.1501T > C:p.S501P, INO80:NM_017553: c.3737G > A:p.R1246Q	missense	INO80 complex subunit E	INO80 has been shown as necessary for DNA damage repair. Abnormal DNA damage repair underlies multiple forms of microcephaly, e.g., PHC1 and PNKP. Supported by a single linkage peak. Only surviving variant and segregates in family.	PMID: 21947284, 19829069, 24029917, and 20118933
08DG00041	hydrocephalus, muscle weakness and global developmental delay	*NID1*	NM_002508.2: c.3385+1G > A	splicing (in-frame insertion confirmed by RT-PCR)	Nidogen 1	Mice deficient of NID1 exhibit neurologic deficits including seizure-like symptoms and loss of muscle control in the hind legs and show altered basement membrane morphology in selected locations including brain capillaries and the lens capsule. Additionally, a variant of unknown significance has been reported in a family with brain malformations. Only surviving variant and segregates in family.	PMID: 12480912 and 23674478
13DG0167	primary microcephaly and global developmental delay	*TSEN15*	NM_001127394: c.226T > G:p.W76G	missense	tRNA splicing endonuclease 15	Mutations in other family members, e.g., TSEN54, have been linked to pontocerebellar hypoplasia. Only surviving variant and segregates in family.	PMID: 18711368

(*Continued on next page*)

Table 2. ***Continued***

ID	Phenotype	Gene	Mutation	Mutation Type	Gene Description	Supporting Evidence	Reference
12DG0929	global developmental delay, epilepsy and poor weight gain	*DMBX1*	NM_147192:c.367C > T:p.R123W	missense	diencephalon/ mesencephalon homeobox 1	Mouse model displays hyperactivity and hypophagia. Gene is highly expressed in developing brain. Only surviving variant and segregates in family.	PMID: 15314164 and 17873059
09DG0405	myopathy	*C2orf63 (CLHC1)*	NM_001135598: c.779G > A:p.R260Q	missense	clathrin heavy-chain linker domain containing 1	Mutation in BICD2 which interacts with CLHC1 causes spinal muscular atrophy. Only surviving variant and segregates in family.	PMID: 23664116
09DG00102	global developmental delay	*C12orf4*	NM_020374:exon6: c.637_638insAAAC: p.K213fs	frameshift	chromosome 12 open reading frame 4	Only surviving variant and segregates in family.	
11DG1513	autism spectrum disorder	*WDR93*	NM_020212:c.280T > C:p.Y94H	missense	WD repeat domain 93	Only surviving variant and segregates in family.	
11DG2479	global developmental delay and brain atrophy	*ST7*	NM_021908:c.489T > G:p.Y163X	stopgain	suppression of tumorigenicity 7	Disruption of ST7 has been reported in one ASD patient with a translocation (t(7;13) (q31.3;q21)). Only surviving variant and segregates in family.	PMID: 10889047
12DG1901	holoprosencephaly	*MATN4*	NM_030590:c.515G > C:p.G172A	missense	Matrilin 4	Gene is highly expressed in developing brain. Only surviving variant and segregates in family.	PMID: 11549321
12DG2051	intellectual disability and epilepsy	*SEC24D*	NM_014822:c.697G > C:p.G233R	missense	SEC24 family, member D (*S. cerevisiae*)	Mouse model displays early embryonic lethality, but the mouse model of its paralog SEC24B displays abnormal neural tube development. Only surviving variant and segregates in family.	PMID: 23596517
10DG1069	primary microcephaly and global developmental delay	*PCDHB4*	NM_018938.2:c.915 del:p.K305Nfs*12	frameshift	protocadherin beta 4	PDCHB4 has been associated with autism. Other protocadherin members have been linked to epilepsy, cognitive impairment as well as autistic features. Only surviving variant and segregates in family.	PMID: 22495309 and 22765916
08DG-00322	brain atrophy and global developmental delay	*PTPN23*	NM_015466:c.3995G > T:p.R1332L	missense	protein tyrosine phosphatase, nonreceptor type 23	Mouse model displays early embryonic lethality. Gene is highly expressed in developing brain. Only surviving variant and segregates in family.	PMID: 19378249

(*Continued on next page*)

Table 2. *Continued*

ID	Phenotype	Gene	Mutation	Mutation Type	Gene Description	Supporting Evidence	Reference
11DG0932	global developmental delay and dysmorphism	*TAF6*	NM_005641.3:c.212T > C:p.I71T	missense	TAF6 RNA polymerase II, TATA box binding protein (TBP)-associated factor	Independently identified by another group (B. Yuan, D. Pehlivan, E. Karaca, N.P., W.-L. Charng, T. Gambin, C. Gonzaga-Jauregui, V.R. Sutton, G. Yesil, S.T. Bozdogan, T. Tos, A. Koparir, E. Koparir, C.R. Beck, S. Gu, H. Aslan, O.O. Yuregir, K. Al Rubeaan, D. Alnaqeb, M.J.A., Y. Bayram, M.M. Atik, H. Aydin, B. Geckinli, M. Seven, H. Ulucan, E. Fenercioglu, M. Ozen, S. Jhangiani, D.M. Muzny, E. Boerwinkle, Baylor-Hopkins Center for Mendelian Genomics, B. Tuysuz, F.S.A., R.A. Gibbs, and J.R. Lupski, unpublished data). Only surviving variant and segregates in family.	
10DG1670	global developmental delay, epilepsy, dysmorphism, hypotonia, and VSD	*TBCK*	NM_033115: c.1708+1G > A	splicing (frameshift)	TBC1 domain containing kinase	TBCK knockdown significantly suppresses mTOR signaling, which plays a critical role in multiple neurological disorders. Only surviving variant and segregates in family.	PMID: 23977024, 19963289
13DG1472	macrocephaly, ID, dolichocephaly, and mild obesity	*FAM177A1*	NM_001079519:c.297_298insA:p.L99fs	frameshift	family with sequence similarity 177, member A1	Only surviving variant and segregates in family.	
13DG1900	Dandy-Walker malformation, hydrocephalus, flexed deformity, club feet, micrognathia, and pleural effusion	*KIAA1109*	NM_015312.3:c.1557T > A:p.Y519*	stopgain	KIAA1109	Deletion in Drosophila results in lethality, and the rare homozygous flies that reach adulthood exhibit severe neurological signs: seizures and inability to walk or stand for prolonged period. Only surviving variant and segregates in family.	PMID: 19640479
10DG0264	neurodegeneration and brain iron accumulation	*MTSS1L*	NM_138383:c.1790C > T:p.T597M	missense	metastasis suppressor 1-like	This mutation affects a poorly characterized isoform of *VAC14* deficiency of which in mouse results in severe neurodegeneration. Only surviving variant and segregates in family.	PMID: 17956977

(*Continued on next page*)

Table 2. ***Continued***

ID	Phenotype	Gene	Mutation	Mutation Type	Gene Description	Supporting Evidence	Reference
13DG1935	primary microcephaly	*XIRP1*	NM_194293:c.4495G > A:p.E1499K	missense	xin actin-binding repeat containing 1	Gene is highly expressed in the brain in response to oxidative stress. Only surviving variant and segregates in family.	PMID: 22366181
13DG2274	severe psychomotor retardation, seizure, and cerebellar hypoplasia	*KCTD3*	NM_016121:c.1036_1073 del:p.P346Tfs*4	frameshift	potassium channel tetramerization domain containing 3	Gene is enriched in copy number changes associated with intellectual disability. Supported by a single linkage peak. Only surviving variant and segregates in family.	PMID: 19623214
13DG0832	global developmental delay and ADHD	*CHAF1B*	NM_005441:c.496A > G:p.I166V	missense	chromatin assembly factor 1, subunit B	CHAF1B is part of the chromatin assembly complex deficiency of which results in loss of asymmetry during nervous system development in *C. elegans*. Only surviving variant and segregates in family.	PMID: 22177093
08DGRC00077	neurodegenerative disease	*ARV1*	NM_022786:c.565G > A:p.G189R	missense	ARV1 homolog (*S. cerevisiae*)	ARV1 is required for normal sphingolipid metabolism, a process known to be defective in other neurodegenerative diseases such as Nieman-Pick disease. Only surviving variant and segregates in family.	PMID: 12145310
14DG0152	neurodegeneration with marked white matter changes with high lactate peak in the brain, consistent with mitochondrial encephalopathy	*ISCA2*	NM_194279:c.229G > A:p.G77S	missense	iron-sulfur cluster assembly 2 homolog	Maps to the only shared haplotype in the extended family. ISCA2 is a member of mitochondrial iron-sulfur cluster (ISC) assembly machinery. Depletion of ISCA2 results in massively swollen mitochondria that are devoid of cristae membranes indicating that it is required for normal mitochondrial biogenesis. Same mutation was identified in other families with identical presentation (Z.N.A.-H., M. Al-Dosary, M. Alfadhel, E.F., M. Alsagob, R. Kenana, R. Almas, O.S. Al-Harazi, H. Al-Hindi, O.I. Malibari, F.B. Almutari, T. Al-Sheddi, S. Tulbah, F. Alhadeq, R. Alamro, A, AlAsmari, M. Almuntashri, H. Alshaalan, F.A. Al-Mohanna, D. Colak, N.K., unpublished data). Only surviving variant and segregates in family.	PMID: 22323289

(*Continued on next page*)

Table 2. *Continued*

ID	Phenotype	Gene	Mutation	Mutation Type	Gene Description	Supporting Evidence	Reference
13DG0215	global developmental delay, hearing loss, and ataxia	*PTRH2*	NM_016077.3:c.254A > C:p.Q85P	missense	peptidyl-tRNA hydrolase 2	Null mice show ataxia and weakness; tissue examination revealed delayed development. Only surviving variant and segregates in family.	PMID: 18218778
08DG0048510DG070313DG1542	(08DG00485) global developmental delay, severe dystonia, and congenital cataract (10DG0703) global developmental delay and congenital cataract(13DG1542) global developmental delay, congenital cataract, tubulopathy, and severe osteopenia	*GEMIN4*	NM_015721:c.2452T > C:p.W818R	missense	gem (nuclear organelle) associated protein 4	Maps to the only shared haplotype in the three families. GEMIN4 is part of the SMN complex, and reduced SMN protein results in spinal muscular atrophy. Supported by a single linkage peak. Only surviving variant and segregates in family.	PMID: 11914277
13DG1549	intellectual disability and epilepsy	*MYOCD*	NM_001146312: c.1252A > G:p.I418V	missense	myocardin	Only surviving variant and segregates in family.	PMID: 12867591
13DG0274	global developmental delay, typical Joubert syndrome, MRI findings	*PDPR*	NM_017990:c.1360G > T:p.G454C	missense	pyruvate dehydrogenase phosphatase regulatory subunit	Only surviving variant and segregates in family.	
10DG0934	cerebellar vermis hypoplasia, Dandy-Walker malformation, hydrocephalus, developmental delay	*DPH1*	NM_001383.3:c.701T > C:p.L234P	missense	diphthamide biosynthesis 1	Diphthamide is a unique posttranslationally modified histidine found only in translation elongation factor-2 (eEF2), which is linked to spinocerebellar ataxia. Diphthamide modification of eEF2 is essential for normal mouse development. Only surviving variant and segregates in family.	PMID: 18765564
11DG0417	global developmental delay, light complexion, early onset focal segmental glomerulosclerosis	*NUP107*	NM_020401:c.303G > A:p.M101I	splice site	nucleoporin 107 kDa	Only surviving variant and segregates in family.	
14DG0221	cerebellar atrophy, hydrocephalus, and global developmental (cognitive, speech, and motor) delay	*TMEM92*	NM_001168215: c.95+3A > G	splice site	transmembrane protein 92	Only surviving variant and segregates in family.	

(*Continued on next page*)

Table 2. *Continued*

ID	Phenotype	Gene	Mutation	Mutation Type	Gene Description	Supporting Evidence	Reference
10DG0840	spastic paraplegia, failure to thrive.	*EPB41L4A*	NM_022140:c.1298C > T:p.S433L	missense	erythrocyte membrane protein band 4.1 like 4A	Only surviving variant and segregates in family.	
12DG0321	coarse facial features (open mouth, bilateral ptosis, and hypertelorism), scoliosis, pectus excavatum, skin laxity, hypotonia, GERD, chronic lung disease, undescended testicles.	*FAM120AOS*	NM_198841:c.743C > T:p.T248I	missense	family with sequence similarity 120A opposite strand	Only surviving variant and segregates in family.	

yield. In 2011, Najmabadi et al. reported the identification of 50 novel candidate genes (Najmabadi et al., 2011). Although that study, similar to ours, combines autozygosity mapping with next-generation sequencing of candidate ROH, their cohort was not enriched for novel gene discovery. Consequently, there were numerous cases that, upon exome sequencing, identified a gene that was classically linked to the clinical presentation (as per their Table 1). More importantly, the pipeline used in that study did not strictly require that the candidate variant be the only one to survive filtering in that family; hence, 24 of their candidate genes (48%) were not the sole surviving variant. Of the remainder, one variant (*HIST3H3*) is present at sufficiently high frequency in our collection of 485 exomes to be excluded (allele frequency 0.0082).

Beyond our analytic pipeline, the candidacy of many of our candidate genes can be corroborated through other lines of evidence. In the case of *INO80*, *KCTD3*, *GEMIN4*, and *ISCA2*, each of these genes maps to the only shared haplotype genome-wide across multiple or extended multiplex families (Figure S3). TUBA3E belongs to a family of proteins that are known to be involved in brain malformation syndromes including lissencephaly, which is present in the affected patient (Figure S4) (Jaglin et al., 2009; Keays et al., 2007; Poirier et al., 2010). As well, TSEN15 belongs to a family of proteins (t-RNA splicing endonucleases) that have been found mutated in patients with pontocerebellar hypoplasia and microcephaly, which is what we observed in our patient (Budde et al., 2008; Cassandrini et al., 2010; Namavar et al., 2011). SPDL1 controls poleward movement of chromosomes along the mitotic spindles, analogous to many of the known primary microcephaly genes that are involved in mitotic spindle regulation (Barisic et al., 2010; Chen et al., 2014); our patient exhibits an extreme reduction in overall brain volume resulting in an almost empty skull (Figure S4). ISCA2 is involved in mitochondrial protein maturation, and depletion of the protein results in massively swollen mitochondria that are devoid of cristae membranes (Sheftel et al., 2012). This is consistent with the mitochondrial encephalopathy detected in our patient. Table 2 summarizes what is known of the 33 candidate genes and available evidence that supports candidacy.

We were unable to identify a causal mutation in 25.9% of cases. Many of these have two or more variants that remained after filtering (Table S1), leaving open the possibility that one of these surviving variants may indeed be causal. Some of these are strong candidates. For instance, case 11DG0375 with brain atrophy has the first reported null mutation in *SNCG*, which encodes synuclein-γ, and, although a role in neurodegeneration has been suspected, such a role has remained elusive (Gretenharrison et al., 2010). Similarly, TRIM4 is a member of a family of proteins that has been implicated in a number of neurological disorders (TRIM2 in Charcot-Marie-Tooth, TRIM32 in Limb-Girdle Muscular Dystrophy, and TRIM18 in Opitz-GBBB (a syndromic form of intellectual disability) (Balastik et al., 2008; Frosk et al., 2002; Quaderi et al., 1997). Thus, the homozygous *TRIM4* truncation we identified in 12DG2083 may prove causal if additional mutations in this gene are identified in the future.

Cases in which no variants survived our filters may harbor classes of mutations that either the sequencing technology or our analysis pipeline are biased against e.g., compound

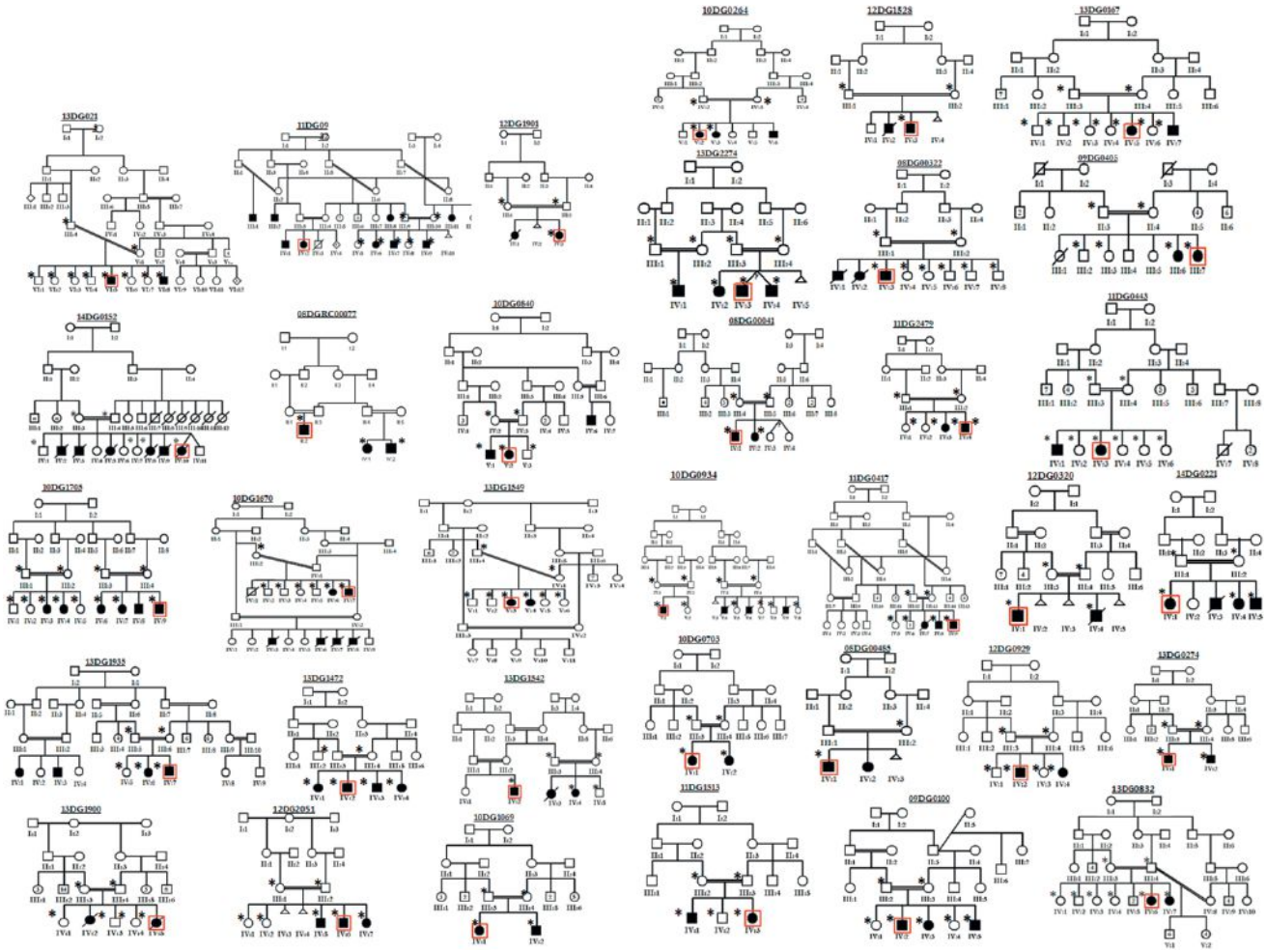


Figure 1. Pedigrees of Families with Novel Candidate Genes following WES Analysis

The family ID is presented above each pedigree. A red box indicates the affected family member who was submitted for WES, and asterisks denote all the family members we had access to for confirming segregation.

heterozygous mutations, X-linked mutations, or mutations involving introns, UTRs, regulatory elements, and repeats. Nonetheless, our study clearly shows the utility of our approach when applied in the right population. Our pipeline as mentioned here has been very successful historically. Many unpublished genes we highlighted as novel and disease causing at the time of data analysis were subsequently published and verified by others, and this lends weight to the candidacy of the genes we report here. Indeed, such reports appeared as recently as a few months from this submission, e.g., *DIAPH1*, *PIGQ*, and *WWOX* (Ercan-Sencicek et al., 2014; Mallaret et al., 2014; Martin et al., 2014). This is also true for certain variants in "unsolved" cases where more than one variant remained, e.g., *SLC13A5* (Thevenon et al., 2014). Our aim in publishing these 33 candidate genes is to accelerate the discovery of other independent mutations, thereby confirming pathogenicity and assisting in genetic diagnosis. Scaling of our study is feasible given the availability of the appropriate patient samples and the dropping cost of sequencing technology, with the promise that all autosomal recessive neurogenetic disease genes can be mapped within the timeframe required by the global brain initiatives.

EXPERIMENTAL PROCEDURES

Human Subjects

Consanguineous families with history of a neurological disorder were clinically evaluated and recruited for this study using a King Faisal Specialist Hospital and Research Center institutional-review-board-approved protocol (RAC# 2121053) with informed consent. All families were multiplex, and all parents were confirmed to be healthy. Pedigrees of families with the 33 candidate genes are given in Figure 1. For each family, blood was collected in EDTA tubes from all available members, and families were only included if at least one affected individual was available for sampling. DNA was extracted from whole blood using standard protocols.

Autozygome-Guided Mutation Analysis

Genome-wide SNP genotyping and homozygosity mapping was performed using the AxiomGWH SNP Chip platform (Affymetrix). See Supplemental Information for more details. Genes known to cause a neurological disorder compatible with the patients' phenotype, and present within the autozygome,

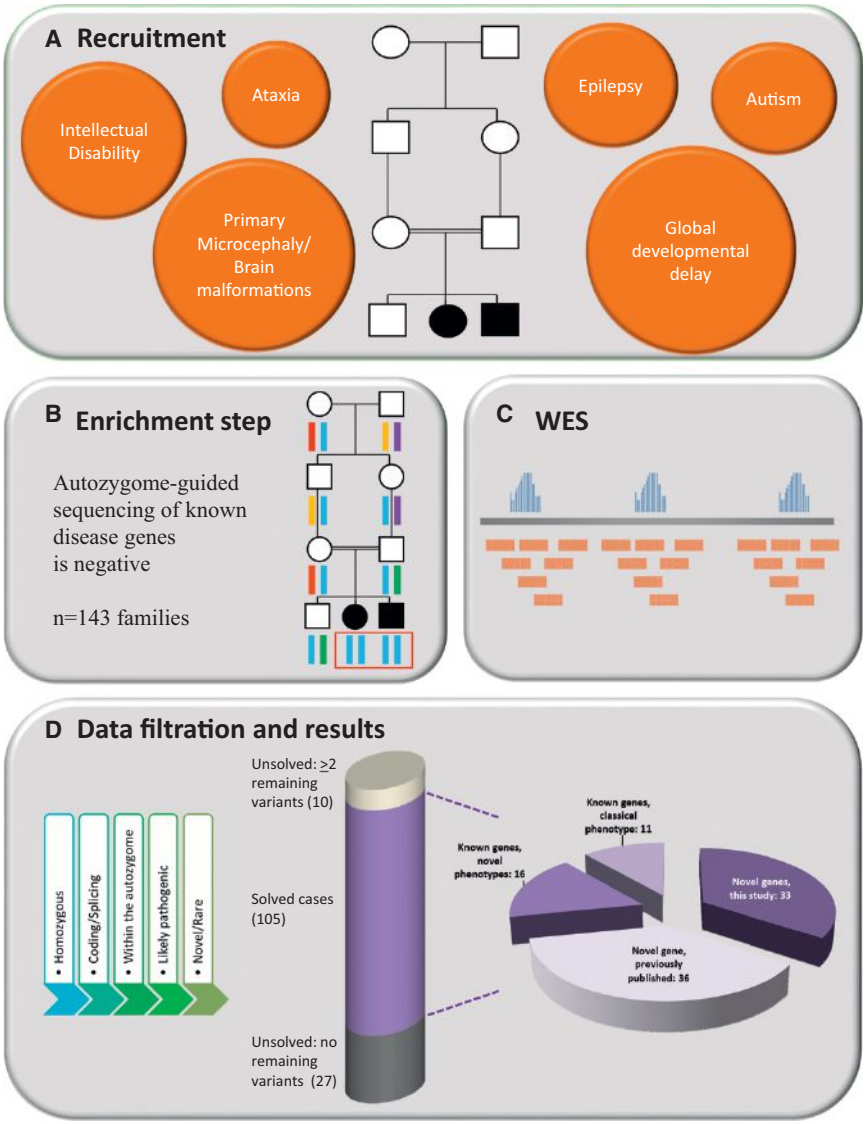


Figure 2. Schematic of the Experimental Pipeline for This Study

A total of 143 multiplex consanguineous families with history of a neurological disorder were found negative for known disease genes following autozygome-guided analysis and were recruited for this study. The bar illustrates the breakdown based on the number of cases, whereas the pie chart is based on the number of distinct genes identified in the solved cases.

were screened by PCR and Sanger sequencing. If no such genes existed, or if they did exist but were excluded by sequencing, exome capture was performed. In total, individuals from 143 families were exome sequenced, and these families formed the cohort for this study.

Exome Sequencing

Exome capture was performed using the TruSeq Exome Enrichment kit (Illumina). See Supplemental Information for more details. A summary of the quality control data for exome sequencing is provided in Table S2.

Analysis of Exomic Variants

Exome-derived data were filtered according to the schematic in Figure 2. See Supplemental Information for full details. Segregation was assessed for all surviving variants, using all family members we had access to (Figure 1).

ACCESSION NUMBERS

All variants within the 143 exomes in this study can be accessed through the following link (part of the Saudi Variome Database): http://shgp.kfshrc.edu.sa/bioinf/db/variants/dg/index.html. The likely pathogenic variants reported in this study have also been uploaded to ClinVar, and this section will be updated with the corresponding accession number.

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.12.015.

AUTHOR CONTRIBUTIONS

A.M.A. collected and analyzed data and wrote the manuscript, N.P. collected and analyzed data and wrote the manuscript, H.E.S. collected and analyzed data and wrote the manuscript.

ACKNOWLEDGMENTS

We thank all the families for their enthusiastic participation. This work was funded in part by KACST 13-BIO1113-20 (F.S.A.). We acknowledge the Saudi

Human Genome Project for infrastructure and informatics support relating to work presented in this manuscript. We are also grateful to the sequencing core facility, and to Salma Wakil and the genotyping core facility, at KFSH&RC for their invaluable assistance. M.A.S. was supported by the Deanship of Scientific Research, King Saud University, Riyadh, Saudi Arabia through Research Group no. RGP-VPP-301.

Received: August 21, 2014
Revised: November 19, 2014
Accepted: December 8, 2014
Published: December 31, 2014

REFERENCES

Alazami, A.M., Alzahrani, F., Bohlega, S., and Alkuraya, F.S. (2014). SET binding factor 1 (SBF1) mutation causes Charcot-Marie-tooth disease type 4B3. Neurology *82*, 1665–1666.

Alkuraya, F.S. (2010). Homozygosity mapping: one more tool in the clinical geneticist's toolbox. Genet. Med. *12*, 236–239.

Alkuraya, F.S. (2012). Discovery of rare homozygous mutations from studies of consanguineous pedigrees. Curr. Protoc. Hum. Genet. *6*, 16.12.

Balastik, M., Ferraguti, F., Pires-da Silva, A., Lee, T.H., Alvarez-Bolado, G., Lu, K.P., and Gruss, P. (2008). Deficiency in ubiquitin ligase TRIM2 causes accumulation of neurofilament light chain and neurodegeneration. Proc. Natl. Acad. Sci. USA *105*, 12016–12021.

Barisic, M., Sohm, B., Mikolcevic, P., Wandke, C., Rauch, V., Ringer, T., Hess, M., Bonn, G., and Geley, S. (2010). Spindly/CCDC99 is required for efficient chromosome congression and mitotic checkpoint regulation. Mol. Biol. Cell *21*, 1968–1981.

Budde, B.S., Namavar, Y., Barth, P.G., Poll-The, B.T., Nürnberg, G., Becker, C., van Ruissen, F., Weterman, M.A., Fluiter, K., te Beek, E.T., et al. (2008). tRNA splicing endonuclease mutations cause pontocerebellar hypoplasia. Nat. Genet. *40*, 1113–1118.

Carr, I.M., Bhaskar, S., O'Sullivan, J., Aldahmesh, M.A., Shamseldin, H.E., Markham, A.F., Bonthron, D.T., Black, G., and Alkuraya, F.S. (2013). Autozygosity mapping with exome sequence data. Hum. Mutat. *34*, 50–56.

Cassandrini, D., Biancheri, R., Tessa, A., Di Rocco, M., Di Capua, M., Bruno, C., Denora, P.S., Sartori, S., Rossi, A., Nozza, P., et al. (2010). Pontocerebellar hypoplasia: clinical, pathologic, and genetic studies. Neurology *75*, 1459–1464.

Chen, J.-F., Zhang, Y., Wilde, J., Hansen, K.C., Lai, F., and Niswander, L. (2014). Microcephaly disease gene Wdr62 regulates mitotic progression of embryonic neural stem cells and brain size. Nat. Commun. *5*, 3885.

Coe, B.P., Girirajan, S., and Eichler, E.E. (2012). A genetic model for neurodevelopmental disease. Curr. Opin. Neurobiol. *22*, 829–836.

Cormier-Daire, V., Amiel, J., Vuillaumier-Barrot, S., Tan, J., Durand, G., Munnich, A., Le Merrer, M., and Seta, N. (2000). Congenital disorders of glycosylation IIa cause growth retardation, mental retardation, and facial dysmorphism. J. Med. Genet. *37*, 875–877.

de Brouwer, A.P., Yntema, H.G., Kleefstra, T., Lugtenberg, D., Oudakker, A.R., de Vries, B.B., van Bokhoven, H., Van Esch, H., Frints, S.G., Froyen, G., et al. (2007). Mutation frequencies of X-linked mental retardation genes in families from the EuroMRX consortium. Hum. Mutat. *28*, 207–208.

de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N. Engl. J. Med. *367*, 1921–1929.

Ercan-Sencicek, A.G., Jambi, S., Franjic, D., Nishimura, S., Li, M., El-Fishawy, P., Morgan, T.M., Sanders, S.J., Bilguvar, K., Suri, M., et al. (2014). Homozygous loss of DIAPH1 is a novel cause of microcephaly in humans. Eur. J. Hum. Genet.

Fritchie, K., Siintola, E., Armao, D., Lehesjoki, A.-E., Marino, T., Powell, C., Tennison, M., Booker, J.M., Koch, S., Partanen, S., et al. (2009). Novel mutation and the first prenatal screening of cathepsin D deficiency (CLN10). Acta Neuropathol. *117*, 201–208.

Frosk, P., Weiler, T., Nylen, E., Sudha, T., Greenberg, C.R., Morgan, K., Fujiwara, T.M., and Wrogemann, K. (2002). Limb-girdle muscular dystrophy type 2H associated with mutation in TRIM32, a putative E3-ubiquitin-ligase gene. Am. J. Hum. Genet. *70*, 663–672.

Greten-Harrison, B., Polydoro, M., Morimoto-Tomita, M., Diao, L., Williams, A.M., Nie, E.H., Makani, S., Tian, N., Castillo, P.E., Buchman, V.L., and Chandra, S.S. (2010). αβγ-Synuclein triple knockout mice reveal age-dependent neuronal dysfunction. Proc. Natl. Acad. Sci. USA *107*, 19573–19578.

Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S., and Frazer, K.A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. *10*, R32.

Hawrylycz, M.J., Lein, E.S., Guillozet-Bongaarts, A.L., Shen, E.H., Ng, L., Miller, J.A., van de Lagemaat, L.N., Smith, K.A., Ebbert, A., Riley, Z.L., et al. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. Nature *489*, 391–399.

Jaglin, X.H., Poirier, K., Saillour, Y., Buhler, E., Tian, G., Bahi-Buisson, N., Fallet-Bianco, C., Phan-Dinh-Tuy, F., Kong, X.P., Bomont, P., et al. (2009). Mutations in the β-tubulin gene TUBB2B result in asymmetrical polymicrogyria. Nat. Genet. *41*, 746–752.

Kalman, B., Lautenschlaeger, R., Kohlmayer, F., Büchner, B., Kmiec, T., Klopstock, T., and Kuhn, K.A. (2012). An international registry for neurodegeneration with brain iron accumulation. Orphanet J. Rare Dis. *7*, 66.

Keays, D.A., Tian, G., Poirier, K., Huang, G.-J., Siebold, C., Cleak, J., Oliver, P.L., Fray, M., Harvey, R.J., Molnár, Z., et al. (2007). Mutations in α-tubulin cause abnormal neuronal migration in mice and lissencephaly in humans. Cell *128*, 45–57.

Kelley, L.A., and Sternberg, M.J. (2009). Protein structure prediction on the Web: a case study using the Phyre server. Nat. Protoc. *4*, 363–371.

Mallaret, M., Synofzik, M., Lee, J., Sagum, C.A., Mahajnah, M., Sharkia, R., Drouot, N., Renaud, M., Klein, F.A., and Anheim, M. (2014). The tumour suppressor gene WWOX is mutated in autosomal recessive cerebellar ataxia with epilepsy and mental retardation. Brain *137*, 411–419.

Martin, H.C., Kim, G.E., Pagnamenta, A.T., Murakami, Y., Carvill, G.L., Meyer, E., Copley, R.R., Rimmer, A., Barcia, G., Fleming, M.R., et al.; WGS500 Consortium (2014). Clinical whole-genome sequencing in severe early-onset epilepsy reveals new genes and improves molecular diagnosis. Hum. Mol. Genet. *23*, 3200–3211.

Maulik, P.K., Mascarenhas, M.N., Mathers, C.D., Dua, T., and Saxena, S. (2011). Prevalence of intellectual disability: a meta-analysis of population-based studies. Res. Dev. Disabil. *32*, 419–436.

Miller, D.T., Adam, M.P., Aradhya, S., Biesecker, L.G., Brothman, A.R., Carter, N.P., Church, D.M., Crolla, J.A., Eichler, E.E., Epstein, C.J., et al. (2010). Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am. J. Hum. Genet. *86*, 749–764.

Morrow, E.M., Yoo, S.-Y., Flavell, S.W., Kim, T.-K., Lin, Y., Hill, R.S., Mukaddes, N.M., Balkhy, S., Gascon, G., Hashmi, A., et al. (2008). Identifying autism loci and genes by tracing recent shared ancestry. Science *321*, 218–223.

Najmabadi, H., Hu, H., Garshasbi, M., Zemojtel, T., Abedini, S.S., Chen, W., Hosseini, M., Behjati, F., Haas, S., Jamali, P., et al. (2011). Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature *478*, 57–63.

Namavar, Y., Barth, P.G., Kasher, P.R., van Ruissen, F., Brockmann, K., Bernert, G., Writzl, K., Ventura, K., Cheng, E.Y., Ferriero, D.M., et al.; PCH Consortium (2011). Clinical, neuroradiological and genetic findings in pontocerebellar hypoplasia. Brain *134*, 143–156.

Poirier, K., Saillour, Y., Bahi-Buisson, N., Jaglin, X.H., Fallet-Bianco, C., Nabbout, R., Castelnau-Ptakhine, L., Roubertie, A., Attie-Bitach, T., Desguerre, I., et al. (2010). Mutations in the neuronal ß-tubulin subunit TUBB3 result in malformation of cortical development and neuronal migration defects. Hum. Mol. Genet. *19*, 4462–4473.

Quaderi, N.A., Schweiger, S., Gaudenz, K., Franco, B., Rugarli, E.I., Berger, W., Feldman, G.J., Volta, M., Andolfi, G., Gilgenkrantz, S., et al. (1997). Opitz G/BBB syndrome, a defect of midline development, is due to mutations in a new RING finger gene on Xp22. Nat. Genet. *17*, 285–291.

Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet *380*, 1674–1682.

Shaheen, R., Faqeih, E., Alshammari, M.J., Swaid, A., Al-Gazali, L., Mardawi, E., Ansari, S., Sogaty, S., Seidahmed, M.Z., AlMotairi, M.I., et al. (2013). Genomic analysis of Meckel-Gruber syndrome in Arabs reveals marked genetic heterogeneity and novel candidate genes. Eur. J. Hum. Genet. *21*, 762–768.

Shaheen, R., Faqeih, E., Ansari, S., Abdel-Salam, G., Al-Hassnan, Z.N., Al-Shidi, T., Alomar, R., Sogaty, S., and Alkuraya, F.S. (2014). Genomic analysis of primordial dwarfism reveals novel disease genes. Genome Res. *24*, 291–299.

Sheftel, A.D., Wilbrecht, C., Stehling, O., Niggemeyer, B., Elsässer, H.-P., Mühlenhoff, U., and Lill, R. (2012). The human mitochondrial ISCA1, ISCA2, and IBA57 proteins are required for [4Fe-4S] protein maturation. Mol. Biol. Cell *23*, 1157–1166.

Song, J., and Markley, J.L. (2007). Three-dimensional structure determined for a subunit of human tRNA splicing endonuclease (Sen15) reveals a novel dimeric fold. J. Mol. Biol. *366*, 155–164.

Tan, D., Rice, W.J., and Sosa, H. (2008). Structure of the kinesin13-microtubule ring complex. Structure *16*, 1732–1739.

Tarpey, P.S., Smith, R., Pleasance, E., Whibley, A., Edkins, S., Hardy, C., O'Meara, S., Latimer, C., Dicks, E., Menzies, A., et al. (2009). A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. Nat. Genet. *41*, 535–543.

Thevenon, J., Milh, M., Feillet, F., St-Onge, J., Duffourd, Y., Jugé, C., Roubertie, A., Héron, D., Mignot, C., Raffo, E., et al. (2014). Mutations in SLC13A5 cause autosomal-recessive epileptic encephalopathy with seizure onset in the first days of life. Am. J. Hum. Genet. *95*, 113–120.

OPEN ACCESS
CellPress

Cell Reports
Report

Analysis of Intron Sequences Reveals Hallmarks of Circular RNA Biogenesis in Animals

Andranik Ivanov,[1] Sebastian Memczak,[1,5] Emanuel Wyler,[2,5] Francesca Torti,[1,5] Hagit T. Porath,[3] Marta R. Orejuela,[1] Michael Piechotta,[4] Erez Y. Levanon,[3] Markus Landthaler,[2] Christoph Dieterich,[4] and Nikolaus Rajewsky[1,*]

[1]Laboratory for Systems Biology of Gene Regulatory Elements, Berlin Institute for Medical Systems Biology, Max Delbruck Center for Molecular Medicine, Robert Roessle Straße 10, 13125 Berlin-Buch, Germany

[2]Laboratory for RNA Biology and Posttranscriptional Regulation, Berlin Institute for Medical Systems Biology, Max Delbruck Center for Molecular Medicine, Robert Roessle Straße 10, 13125 Berlin-Buch, Germany

[3]The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan 5290002, Israel

[4]Max Planck Institute for Biology of Ageing, Cologne, Joseph Stelzmann Straße 9B, 50931 Köln, Germany

[5]These authors contributed equally to this work

*Correspondence: rajewsky@mdc-berlin.de

http://dx.doi.org/10.1016/j.celrep.2014.12.019

SUMMARY

Circular RNAs (circRNAs) are a large class of animal RNAs. To investigate possible circRNA functions, it is important to understand circRNA biogenesis. Besides human ALU repeats, sequence features that promote exon circularization are largely unknown. We experimentally identified circRNAs in *C. elegans*. Reverse complementary sequences between introns bracketing circRNAs were significantly enriched in comparison to linear controls. By scoring the presence of reverse complementary sequences in human introns, we predicted and experimentally validated circRNAs. We show that introns bracketing circRNAs are highly enriched in RNA editing or hyperediting events. Knockdown of the double-strand RNA-editing enzyme ADAR1 significantly and specifically upregulated circRNA expression. Together, our data support a model of animal circRNA biogenesis in which competing RNA-RNA interactions of introns form larger structures that promote circularization of embedded exons, whereas ADAR1 antagonizes circRNA expression by melting stems within these interactions.

INTRODUCTION

Recently, several studies have revealed that the transcriptome of animals contains many single-stranded exonic circular RNAs (circRNAs) (Jeck et al., 2013; Jeck and Sharpless, 2014; Memczak et al., 2013; Salzman et al., 2012; Wang et al., 2014). Although circRNAs have tissue- and stage-specific expression (Memczak et al., 2013), the function of circRNAs is altogether unknown. The human circRNA *CDR1as* (Hansen et al., 2011;) can act as a miRNA sponge (Hansen et al., 2013; Memczak et al., 2013). However, we and others have proposed that generally circRNAs might function in assembly of complexes, in transport, in *trans* (Memczak et al., 2013), by competing with linear splicing, or as regulators of the local concentration of RNA-binding proteins (Ashwal-Fluss et al., 2014). In vitro studies (Braun et al., 1996; Pasman et al., 1996) and a recent in vivo study (Ashwal-Fluss et al., 2014) provided evidence that circRNAs are often generated cotranscriptionally by "head-to-tail" splicing. In humans, circularized exons are typically bracketed by unusually long introns (Jeck et al., 2013). Moreover, the circRNA *SRY* (Capel et al., 1993) is bracketed by very long (~15,000 nt), almost perfectly complementary intronic sequences, and these reverse complementary matches (RCMs) are required for *SRY* circularization (Dubin et al., 1995). Therefore, it has been proposed that RCMs promote hairpin formation of the transcript. This would explain how the 5′ and 3′ ends of an exon can be in spatial proximity, perhaps thereby inducing "head-to-tail" splicing (Figure 1A). Jeck et al. (2013) reported that in humans, introns bracketing circRNAs are highly enriched in ALU repeats. The fact that ALU repeats contain RCMs supports this model; however, ALU repeats are specific to a small branch of vertebrates, and thus the widespread existence of circRNAs in other animals remains to be explained.

Caenorhabditis elegans, a well-annotated genome that is not rich in repeats, offers the possibility of identifying conserved features of circRNA biogenesis outside of vertebrates. We first sequenced RNA from several life stages of *C. elegans* and were able to boost the number of annotated exonic circRNAs from ~300 (Memczak et al., 2013) to ~1,100. Computational analysis of the bracketing introns revealed that these circRNAs are significantly enriched for RCMs. We developed a simple model for scoring circRNA biogenesis from intronic sequence analysis and asked whether our model could predict novel human circRNAs. We successfully validated the predicted human circRNAs.

The RNA-editing factor ADAR binds double-stranded RNA. Thus, the model of circRNA biogenesis predicts that circRNAs should be flanked by intronic sequences that are enriched in adenosine to inosine (A-to-I) editing. Moreover, it is known that ALU elements are edited, often by ADAR1 (Athanasiadis et al., 2004; Levanon et al., 2004; Osenberg et al., 2010; Ramaswami

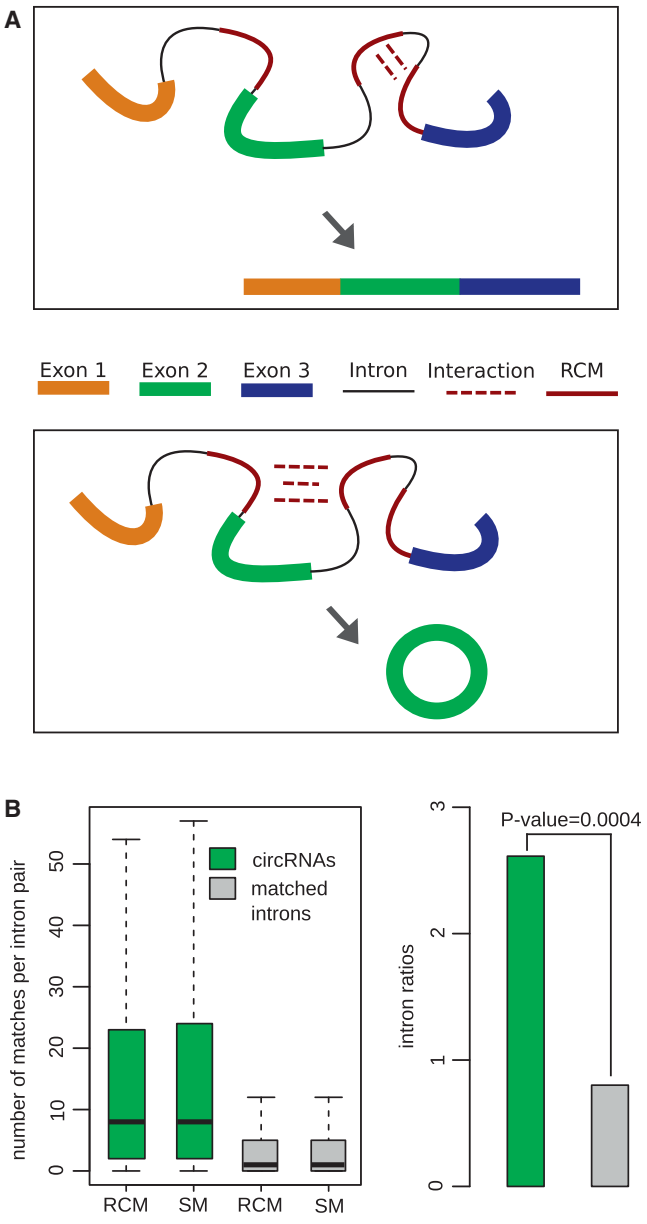


Figure 1. RCMs between or within Introns and circRNA Biogenesis

(A) Model for the possible influence of RCMs on circRNA biogenesis. RCMs between different introns (lower panel) in competition with intron internal RCMs (upper panel) may promote hairpin formation and circularization of the embedded exon.

(B) Enrichment of RCMs and SMs in introns bracketing *C. elegans* circRNAs. Left: distribution of RCM and SM counts per intron pair that flanks circRNAs (green) and length-matched controls (gray). Right: ratio of the number of intron pairs bracketing circRNAs that contain only RCMs but no SMs to the number of intron pairs bracketing circRNAs with only SMs but no RCMs (p value: Fisher's exact test).

See also Figures S1, S4, and S5.

et al., 2012). Also in *C. elegans*, long inverse repeats are enriched in A-to-I editing (Morse et al., 2002). Indeed, we discovered significant A-to-I editing, including hyperediting (Carmi et al., 2011), in introns bracketing circRNAs. To test whether ADAR1 and ADAR2 knockdown in human cells affects circRNA expression, we performed RNA sequencing (RNA-seq) and quantitative RT-PCR (qRT-PCR). We observed a significant and specific upregulation of most circRNAs, whereas their linear host transcripts were less perturbed. Thus, our data reveal a surprising function for ADAR as an antagonist of circRNA production. We discuss the implications of these findings.

RESULTS

Identification and Characterization of circRNAs in *C. elegans*

Recently, hundreds of circRNAs were reported in samples from very early developmental stages of *C. elegans* (one/two-cell embryo, oocytes, and sperm) (Memczak et al., 2013). To include circRNAs expressed in later stages, we sequenced ribosomal-depleted RNA from major life stages, including adulthood (see Experimental Procedures). We computationally identified circRNA candidates by applying our pipeline (Glažar et al., 2014; Memczak et al., 2013; http://www.circbase.org) and ensuring that the head-to-tail junctions precisely overlapped the annotated canonical splice sites (Experimental Procedures). This extended the published set of exonic circRNAs from ~300 to 1,111. Analogously to human circRNAs (Jeck et al., 2013), the circRNA flanking introns were much longer (median ~10-fold) than all of the *C. elegans* introns (Figure S1A). Therefore, we asked whether intron length alone is sufficient for circularization or additional sequence features are needed. More precisely, in *C. elegans* we tested the idea that RCMs between introns bracketing circRNAs may induce larger hairpin structures that promote the circularization of embedded exons (Figure 1A).

In *C. elegans*, Introns Bracketing circRNAs Are Highly Enriched for RCMs

For each intron pair that flanked a circRNA, we aligned the respective introns using Basic Local Alignment Search Tool (BLAST; Experimental Procedures). RCMs were strongly enriched compared with length-matched introns (Figure 1B), with a median of eight RCMs per intron pair (control pairs: one RCM). We observed the same number and significance of matches on the same strand ("sense matches" [SMs]) of intron pairs (Figure 1B). However, the exclusive occurrence of RCMs on intron pairs bracketing circRNAs was significantly enriched compared with the exclusive occurrence of SMs (Fisher's exact test, p value < 0.0004; Figures 1B and S1C). This result suggests that exon circularization is promoted by RCMs that can induce basepairing between flanking introns.

Predicting circRNAs from RCMs

We asked whether the occurrence of RCMs between circRNA flanking introns is significant enough to predict circularized exons. Therefore, we developed a simple probabilistic score *H* to capture the likelihood that introns will basepair via RCMs (Supplemental Experimental Procedures) and therefore promote hairpin formation. Competition between RCMs is naturally taken into account by this model. For each intron pair in a transcript that contained one of the 1,111 circRNAs, we computed *H*.

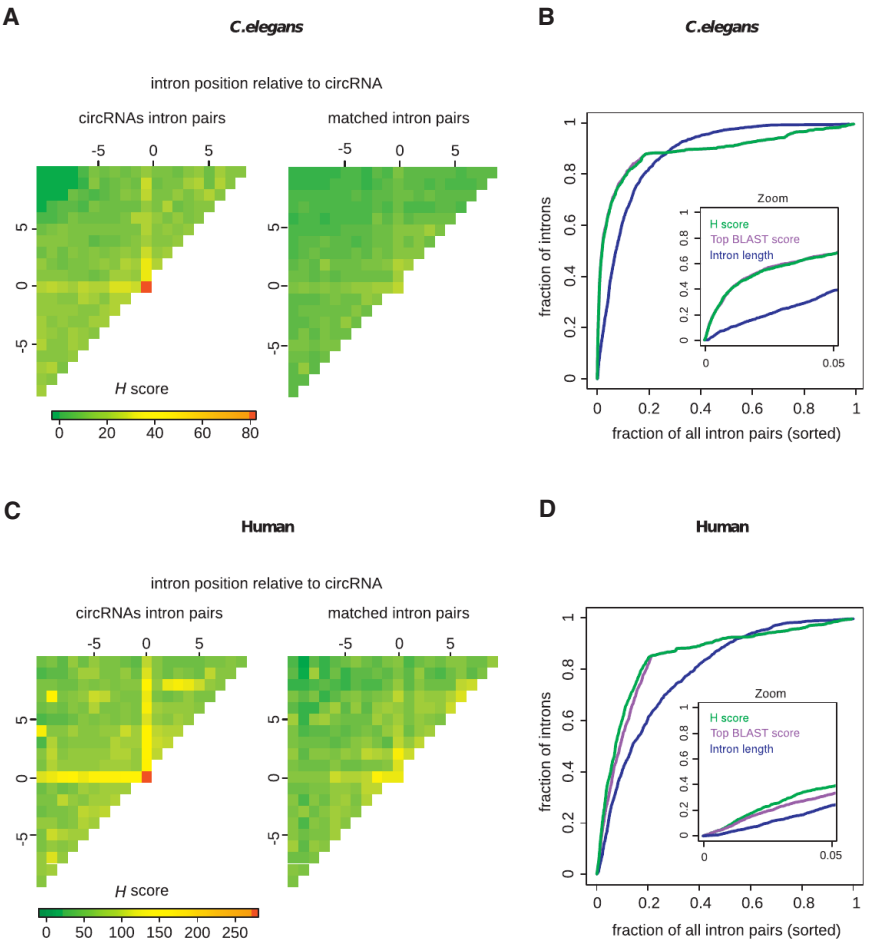


Figure 2. Prediction of circRNAs Based on Sequence Analysis of Introns

(A) *C. elegans*. The matrix represents the average circularization scores for intron pairs of circRNA producing genes. (0,0) coordinates correspond to circRNA flanking introns. Upstream and downstream introns are enumerated with decreasing and increasing numbers, respectively (with 0 being directly adjacent, −1/+1 being the preceding/next intron, etc.; see Supplemental Experimental Procedures).

(B) *C. elegans*. Different prediction methods are used. The x axis is the fraction of all introns pairs per gene sorted by *H* scores. Small (high) x values denote intron pairs with high (low) scores for bracketing a circRNA. The y axis is the fraction of 1,111 exonic circRNAs recovered by *H* score.

(C and D) Analyses for human circRNAs (Memczak et al., 2013), analogous to (A) and (B).

See also Figures S4 and S5, and Tables S1, S2, and S3.

The resulting symmetric matrix (Figure 2A; Supplemental Experimental Procedures) shows that the intron pairs flanking circularized exons stood out from all intron pairs. This was not the case for the intron length-matched control transcripts (Figure 2A). We then ranked, genome wide, all intron pairs of all transcripts annotated in *C. elegans* by *H* and calculated the cumulative fraction of known circRNAs as a function of intron-pair rank (Figure 2B). The top 4,200 (top ~1%) of *C. elegans* intron pairs precisely bracketed 430 (~38%) of the already annotated circRNAs, a highly statistically significant enrichment (hypergeometric test, p value $< 2.2 \times 10^{-16}$). Together, these results suggest that RCMs provide highly significantly improved accuracy in predicting circRNAs compared with intron length. However, we note that (for *C. elegans*) the top BLAST score of RCMs in a pair of introns yielded similar results.

To determine whether the presence of RCMs is a conserved feature of circRNA formation, we next asked whether we could predict circRNAs in human by intron sequence analysis.

The Presence of RCMs in Introns Flanking circRNAs Is a Conserved Feature of circRNA Biogenesis

For human circRNAs, we used 1,067 exons overlapping circRNAs (Memczak et al., 2013). As in *C. elegans*, RCMs were highly significantly enriched (Fisher's exact test, p value $< 4 \times 10^{-6}$; Figures S1B and S1C). Again, the number of intron pairs flanking circularized exons that contained only SM elements was much lower compared with intron pairs containing only RCMs (Figure S1C). The median length of human RCMs was 20-fold higher than that of *C. elegans* (Figures S1D and S1E). We found that 88% of the top-scoring RCMs overlapped with ALU elements, 4.5% of RCMs overlapped with L1/L2 repeats, and the remaining RCMs did not overlap any annotated repetitive element and had the highest sequence conservation (Figure S1F).

Further, we analyzed circRNA biogenesis conservation between mouse and human. We defined 71 circRNAs that are circularized in human and mouse ("conserved circular expression," Figure S1G). Since repetitive sequences such as ALU elements are rapidly evolving, it is difficult to align them between species. To circumvent this problem, we computed *H* scores by independently scoring the human and mouse introns (Figure S1G). As controls, we defined (1) circRNAs that are circularized in humans but have not been annotated as circular in mouse ("nonconserved circRNAs"), and (2) randomly selected exons that have a similar bracketing intron length as human circRNAs, and have positive *H* scores. Mouse *H* scores for conserved circRNAs were significantly higher (p value $< 2.2 \times 10^{-16}$) compared with the scores for exons that circularize only in human or the length-matched controls (Figure S1G). Thus, the set of circRNAs that are conserved between mouse and human are also conserved in their biogenesis as quantified by *H*.

Human circRNAs Can Be Predicted Based on the Sequence Composition of Their Flanking Introns

To predict circRNAs on a genome-wide level, we computed *H* for all possible intron pairs in the UCSC RefSeq database

(Experimental Procedures). As a control, we used transcripts with introns matched in length (Figure 2C; Supplemental Experimental Procedures) or predictions using the intron length or simply the top BLAST score between intron pairs (Figure 2D). For high-ranking intron pairs, the enrichment in circRNAs was highest when the *H* score was used, followed by BLAST and intron length. The top 20,000 (~1% of all) predictions comprised 610 (~9%) exonic circRNAs cataloged in the Memczak et al. (2013), Jeck et al. (2013), and Zhang et al. (2014) data sets, a highly significant enrichment (hypergeometric test, p value $< 2.2 \times 10^{-16}$). Predictions based on sequences associated with ALU repeats yielded similar results (data not shown), showing that in humans, ALU elements have likely contributed dominantly to circRNA formation (Experimental Procedures). We note that the number of false positives is very difficult to estimate because many human circRNAs have not yet been reported. For example, when we restricted the circRNA predictions to well-expressed mRNAs in HEK293 cells, the success rate of our predictions was much higher (Figure S1H). Therefore, experimental validation of predictions seems to be a better way to estimate false-positive rates.

Predicted circRNAs Are Experimentally Validated in HEK293 Cells

Since the circRNA predictions were not informed by expression data, we expected that we could only validate circRNAs that are isoforms of expressed transcripts. To test our predictions, we considered the top 1% (*H* ranked) predicted circRNAs and grouped them into three bins based on the expression of the linear host transcripts in HEK293 cells (top 1% expressed, medium, and bottom 50%). For experimental validation, we randomly selected six, ten, and five circRNAs from the high-, medium-, and low-expression bins, respectively. As negative controls, we used three well-expressed linear mRNAs and two exons that are well expressed in HEK293 cells and have flanking introns but an *H* score of 0. The linear controls were not circularized (RNase R negative) and the two exons with an *H* score of 0 could not be amplified with divergent primers. However, 12 out of 16 circRNAs with top *H* scores and high/medium expression of the host transcripts were (1) resistant to RNase R and (2) had the predicted head-to-tail junctions, as validated by Sanger sequencing (Experimental Procedures; Figure 3), suggesting that these 12 circRNAs exist in circularized form in HEK293 cells. Five candidates with decent *H* scores that we could not confirm were, as expected, from the third, low-expression bin. The predicted head-to-tail splicing of ten (out of 12) of the RNase R-resistant circRNAs was validated by Sanger sequencing (Experimental Procedures). We found that two circRNA candidates were circularized (by RNase R assay) but had an additional exon incorporated (as observed in Sanger sequencing; Figure 3A). We note that five of the tested circRNAs had reasonable expression (1%–10% of *VCL* expression; Experimental Procedures).

RNA Editing of Introns Flanking circRNAs

ADAR is a highly conserved RNA-editing enzyme that binds double-stranded RNA (Nishikura, 2010) and deaminates adenosine bases to inosine. In humans, ADAR1 and ADAR2 interact with double-stranded ALU repeats (Athanasiadis et al., 2004; Levanon et al., 2004; Nishikura, 2010; Ramaswami et al., 2012). We compared A-to-I conversions (Ramaswami and Li, 2014) in 1,500 bp regions flanking circRNA splice sites with (1) other splice sites in transcripts that produce circRNAs, and (2) length-matched introns (Figure 4A). A-to-I conversions nearby circRNA splice sites were enriched compared with controls. In general, ALU repeats in circRNAs flanking introns were edited significantly higher compared with expression- and length-matched controls (Figures S2A and S2B). Since A-to-I editing is a hallmark of basepaired RNA, we asked whether RCMs are preferentially located at sites of editing. We compared the position of the RCMs (defined as the nearest RCMs that match between the pair of introns bracketing circRNA) with the same controls as before (Figure 4B). These results suggest that indeed A-to-I editing preferentially occurs at regions that are basepaired and proximal (upstream and downstream 200–600 nt) to the splice sites of circularized exons.

To test whether ADAR proteins are involved in circRNA biogenesis, we codepleted ADAR1 and ADAR2 in HEK293 cells using RNAi (Supplemental Experimental Procedures). We used two controls: untreated total RNA and codepletion of three proteins of the APOBEC family (APOBEC3B, APOBEC3C, and APOBEC3F), which are known to bind mRNAs (Baltz et al., 2012). APOBEC enzymes are known to edit single-stranded DNA or RNA (Vasudevan et al., 2013), but not double-stranded RNA. The efficacy of the different knockdowns (KDs) was validated by western blotting and qRT-PCR (Figure S2C). Total RNA extracted from the different experiments was depleted from rRNAs and sequenced (Supplemental Experimental Procedures). We reproducibly observed that ADAR depletion resulted in significantly higher (p value $< 2.2 \times 10^{-16}$) circRNA expression compared with controls (Figure S2D), whereas the linear host transcripts were less strongly affected (Figure 4C). For example, 84 circRNAs and 11 linear hosts were upregulated more than 2-fold (Fisher's exact test, p value 5.7×10^{-13}). This effect was seen in independent biological replicates, as well as in a comparison of ADAR1/2 KD with APOBEC3 KD (Figures S2E and S2F).

We set out to validate these observations in independently carried out ADAR1 and ADAR2 KD experiments (using the previous and an independent small interfering RNAs [siRNAs] against ADAR1) followed by qRT-PCR assays (Experimental Procedures; Figures 4D, 4E, and S2G). CircRNA candidates were selected based on the increased fold changes observed in sequencing data sets. circRNA expression was compared with the expression of the respective linear host gene. Four out of eight circRNAs (*UBAC2*, *SMARCA*, *SPECC1*, and *HIPK2*) were upregulated upon ADAR1 depletion, whereas the linear host RNAs did not show consistent expression changes. *PUM1* and *CREBBP* circRNA were upregulated to the same level as their linear host transcripts. Consistent with sequencing data, *CDR1as* was expressed 4-fold higher in ADAR1 KD samples compared with control. Two circRNAs (*GAPVD1* and *PDS5B*) did not show the expression changes observed in the sequencing data (Figures 4E and S2G). Based on these findings, we conclude that ADAR1 depletion can induce upregulation of circRNAs independently of the expression level of the linear host circRNA.

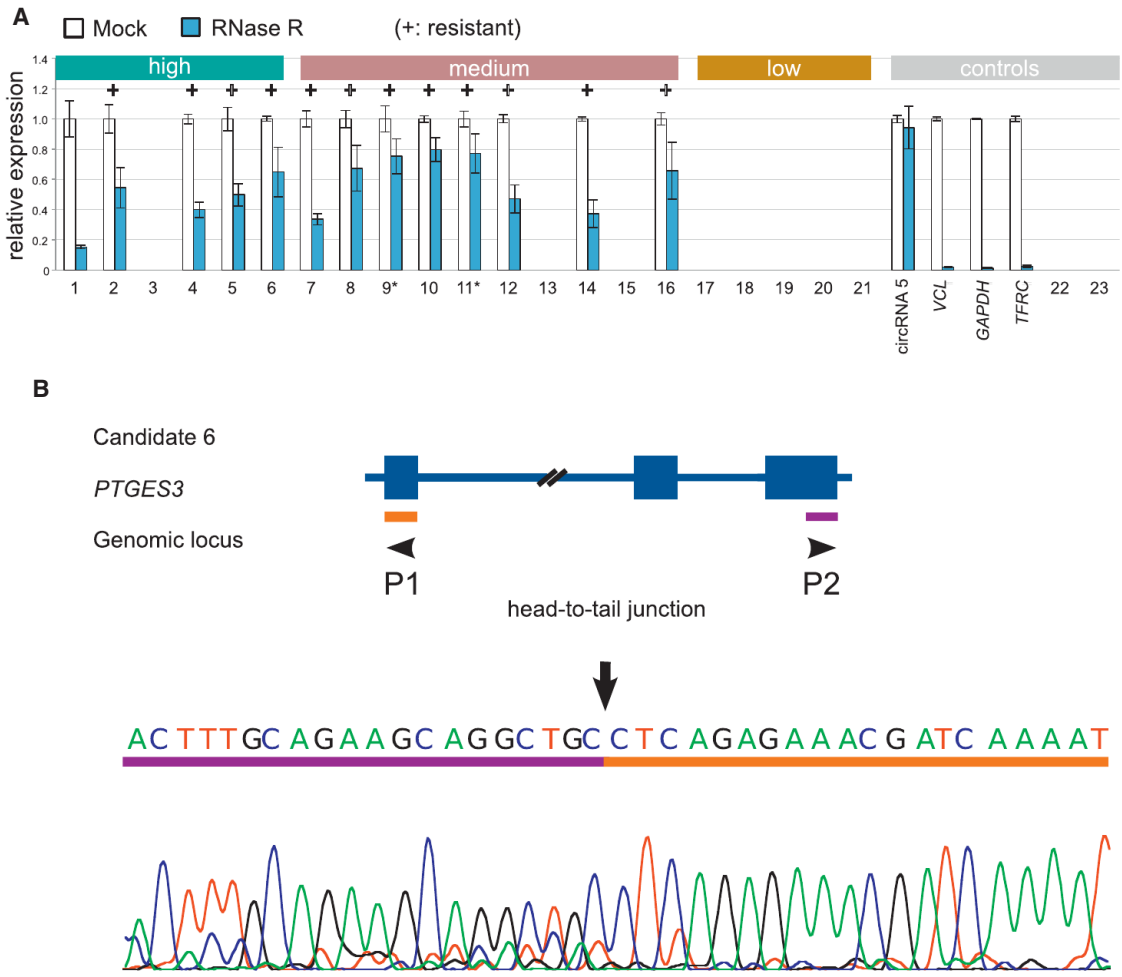


Figure 3. Experimental Validation of Human circRNAs Predicted by RCM Analyses

(A) CircRNA candidates from high-, medium-, and low-expression sets were assayed by qPCR with divergent primers and RNase R treatment. Linear control: VCL, GAPDH, TFRC; positive control: a known circRNA (Memczak et al., 2013). Sanger sequencing of amplicons confirmed in all tested cases the predicted head-to-tail junctions (candidates 9 and 11 contained an additional exon: marked with *). CircRNA candidates that were >10-fold resistant to RNase R treatment compared with Vinculin were counted as positive (+). As negative controls, we selected exons from highly expressed genes with *H* score = 0. Error bars, SEM; n = 4.

(B) The chromatogram of a Sanger sequencing experiment confirms the presence of the predicted head-to-tail junction of the candidate circRNA from the *PTGES3* gene locus.

See also Figure S5 and Tables S1, S2, and S3.

To explore whether circRNA flanking introns undergo extensive hyperediting (Carmi et al., 2011), we applied a computational pipeline that detects hyperediting events in the RNA sequencing data sets used by Memczak et al. (2013) to predict circRNAs, as well as in ADAR KD and control data sets. This analysis (for details, see Porath et al. 2014) identified ~165,000 unique hyperediting sites in the human genome (~72.000, ~49.000, ~45.000, and ~11.000 in Memczak et al. [2013] and two control and ADAR KD samples, respectively). Notably, the number of hyperedited sites identified in the ADAR KD sample was 4- to 5-fold smaller compared with controls. We found that 25% (19%) of circRNA upstream (downstream) introns had at least one conversion, whereas only 6% of other introns from the same genes had conversions. The number of conversions per base was also found to be elevated for the upstream (downstream) introns (Figures S3A–S3C). A similar highly significant enrichment of hyperediting events was detected for *C. elegans* circRNA upstream introns (Supplemental Experimental Procedures; Figure S3D). Together, our data suggest that A-to-I editing by ADAR is a "universal" hallmark of circRNA biogenesis in animals.

DISCUSSION

Our analyses support a model in which RCMs between introns that bracket an exon promote the circularization of that exon. We note that this model is powerful enough to enable us to successfully predict and experimentally validate circRNAs. Our data suggest that this model of circRNA biogenesis is conserved across animals.

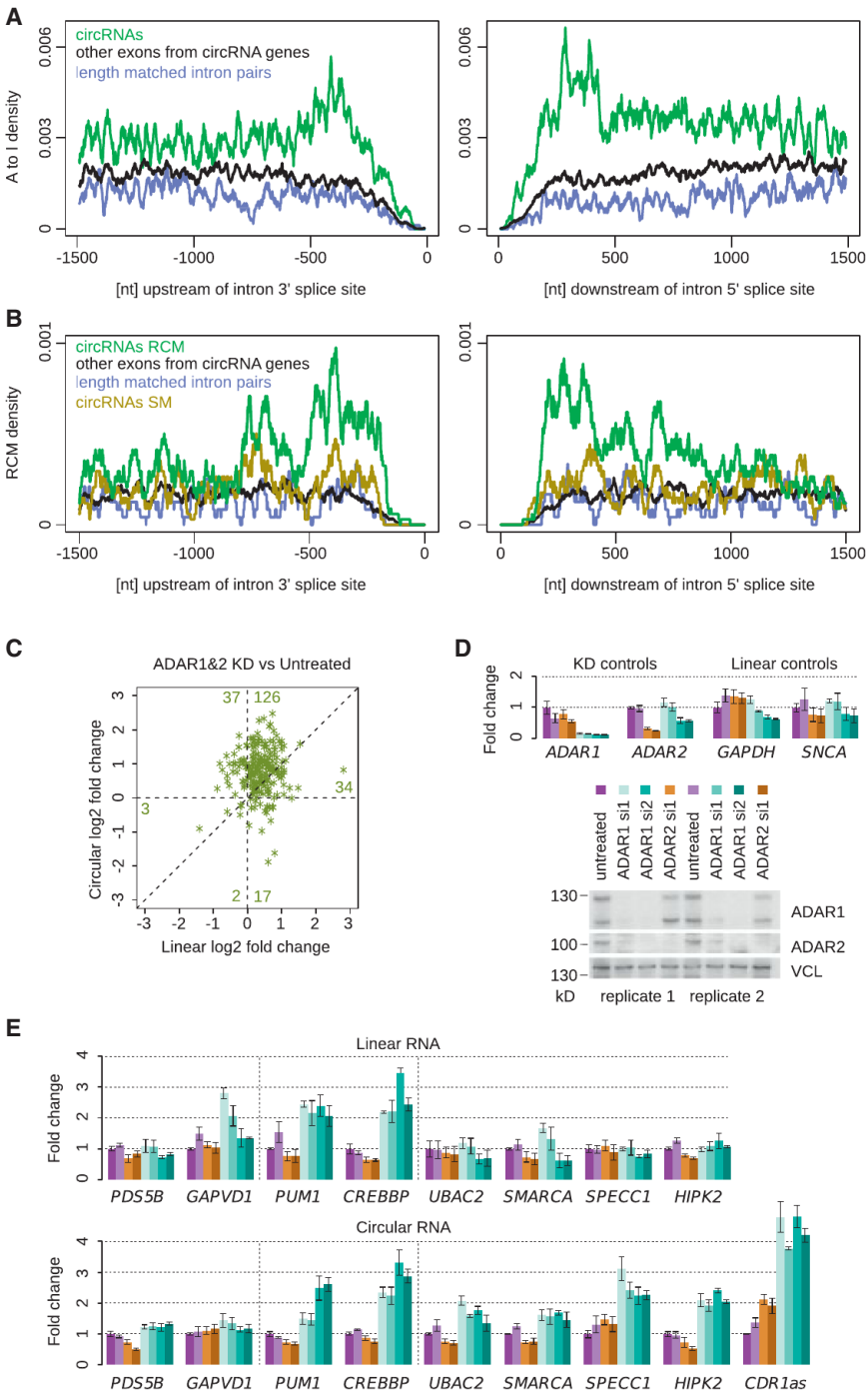


Figure 4. ADAR Antagonizes circRNA Expression

(A) Normalized distribution of A-to-I conversions upstream/downstream of the head/tail splice sites of circRNAs (Memczak et al., 2013). As controls, we selected length-matched introns and introns from the same genes that produce circRNAs. Introns smaller than 1.5 kb were removed from the analysis. The curves were smoothed within ±10 bp at each position and normalized to the total number of analyzed introns.

(B) Position of the nearest RCMs (or SMs) around circRNA splice sites. For each intron pair, we selected the top-scoring RCMs within the 1.5 kb region around the splice site. The curves were smoothed within a ±20 bp window and normalized to the total number of analyzed intron pairs.

(C) Comparison of the differential expression (ADAR1 and ADAR2 codepletion against untreated control) of linear RNA and circRNA in two independent replicates (merged). The y axis is the log2 fold change of the reads (+5) supporting the head-to-tail splice sites. The x axis is log2 fold change of the reads supporting linear splice sites of circRNA host genes. For the analysis, we selected only circRNAs with at least ten reads in ADAR1/2 KD or control samples. The numbers tally the count of circRNAs/corresponding linear hosts in each segment of the graph.

(D) Depletion of ADAR1 by RNAi. ADAR1 was depleted from HEK293 cells for 96 hr using two different siRNAs and compared with untreated and ADAR2-depleted cells. Upper panel: qRT-PCR of *ADAR1* and *ADAR2* transcripts, as well as *GAPDH* and *SNCA* as controls. Lower panel: western blot using ADAR1- and ADAR2-specific antibodies. Error bars, SD; n = 3.

(E) Quantification of circRNAs and host transcript levels. Using circRNA-specific primers (upper panel) or exon-spanning primers that were not part of the circRNA (lower panel), we quantified RNA abundances relative to untreated cells by qRT-PCR (Delta-Delta Ct values were normalized to *C. elegans* spike-in). Error bars, SD; n = 3.

See also Figures S2, S3, and S5, and Tables S1, S2, and S3.

An immediate follow-up question is, do RCMs have sequence specificity? In general, our data do not support sequence specificity in *C. elegans*, since the most prevalent motif in RCMs between introns bracketing circRNAs was clearly present in only 11% of all such RCMs. The equivalent motif analysis in humans yielded (perhaps as expected) the ALU element. For mouse circRNAs, we also found an ALU-like element. Sequence alignment between these motifs (Figure S4) suggests that there may be an ancient RNA sequence that promotes circularization. This motif awaits experimental testing.

RCMs can be generated by simple random matches under neutral evolution. One possible scenario is that relatively quickly evolving RCMs constantly generate a pool of circRNAs that can subsequently be selected for and fixed. Finding circRNAs that are under negative selection seems possible because, as we have shown, RCMs between introns can be analyzed across species. Thus, we think that cross-species comparisons of competing RCMs will provide a way to find circRNAs that may be functionally important. Indeed, our results show that our *H* score is already able to link conserved circular expression to conserved biogenesis.

In humans, 88% of circRNAs have ALU repeats in their flanking introns, which, as we have shown, likely promote circularization by RCMs. Therefore, it is interesting to speculate about the possible functions of circRNAs, since ALU repeats have expanded relatively recently in vertebrate evolution. We noticed that in fly, genes with neuronal functions often have long introns and express circRNAs at relatively high levels (Ashwal-Fluss et al., 2014). This also holds true for human brain tissues, where dozens of circRNAs appear to be more highly expressed than their (well-expressed) linear hosts (N.R., unpublished data). For human circRNAs, we found that A-to-I editing events had a tendency to occur at intronic positions that were proximal (upstream and downstream 200–600 nt) to the splice sites of circularized exons. These results link A-to-I editing to circRNA biogenesis and predict that A-to-I editing events "melt" the stems that are formed across introns that bracket a circRNA. Our KD experiments showed that indeed ADAR1 antagonizes circRNA biogenesis. The upregulation of circRNAs was stronger compared with upregulation of their linear host transcripts, suggesting that ADAR1 has a specific effect on circRNA biogenesis. However, we cannot rule out that indirect effects explain upregulation of circRNAs upon ADAR1 KD.

It is very interesting to think about the implications of these findings. As a class, circRNAs may become more important in systems where ADAR1 expression is temporarily low. For example, ADAR1 expression decreases in human embryonic stem cells that are differentiating into the neuronal lineage (Osenberg et al., 2010). Since circRNAs are unusually stable, this might be a mechanism to generate a long-term "memory" of past states. However, one should not forget that other factors, such as muscleblind, have been shown to promote the biogenesis of circRNAs (Ashwal-Fluss et al., 2014). Therefore, circRNAs may also be well expressed in systems in which ADAR1 expression is high. Finally, it has been shown that circular splicing and linear splicing can compete with each other (Ashwal-Fluss et al., 2014). Thus, it is possible that regulation of circRNA biogenesis by ADAR1 serves as a mechanism to regulate expression of the linear isoforms.

Note: while this paper was under review, two studies were published describing RCM-dependent circularization in human cells (Liang and Wilusz, 2014; Zhang et al., 2014). We also acknowledge Starke et al. (2014), which presents related content in this issue of *Cell Reports*.

CONCLUSIONS

In summary, in this work, we explored a specific model of circRNA biogenesis and showed that this model seems to be conserved across animals and is powerful enough to successfully predict circRNAs. Our data also link RNA editing to circRNA biogenesis and suggest a function for ADAR1. These results will enable the future detection and understanding of possible circRNA functions, particularly in neuronal tissues.

EXPERIMENTAL PROCEDURES

Identification of *C. elegans* circRNAs

We first mapped worm sequencing data to rRNA to reduce the ribosomal reads. The remaining reads were analyzed as described in Memczak et al. (2013) with the additional filtering step of requiring unique alignments for both of the read anchors prior to extension. Thereafter, we retained circRNAs that overlap annotated internal splice sites.

Intron Alignments

We carried out intron alignments using BLAST (Altschul et al., 1990) with the parameters "-task blastn -word_size 6" for *C. elegans* and "-task blastn -word_size 11" for human. For further analysis, we considered only alignments that exceeded BLAST score cutoffs of 20 and 100 for *C. elegans* and human, respectively. For each intron pair, we calculated circularization *H* as a top BLAST score multiplied by the probability of forming at least one stem around the exon (Supplemental Experimental Procedures).

Nonrepetitive RCMs were defined as BLAST matches that did not overlap with any sequence present in UCSC RepeatMasker. Putative circRNAs with nonrepetitive RCMs are shown in Table S3.

Conservation Analysis

To determine homologous exon groups, we used UCSC liftOver with minMatch = 0.1 option. We found that 71 out of 1,067 human circRNAs (Memczak et al., 2013) overlapped annotated splice sites and were present in circular form in mouse. We compared *H* score distributions using the Mann-Whitney two-sided test.

qRT-PCR

We performed qPCR using the Maxima SYBR-Green/ROX qPCR master mix (Thermo Scientific) and a StepOnePlus PCR system (Applied Biosystems). To detect putative head-to-tail junctions, we designed divergent primers for each circRNA candidate. Ct values for mock/RNase R-treated circRNAs were normalized to *C. elegans* spike-in RNA (for standard curves, see Figures S5A–S5C and Table S2). Amplicons were gel or bead purified (Zymoclean gel DNA recovery kit [Zymo Research]; Agencourt AMPure XP [Beckman Coulter]) and subjected to Sanger sequencing by LGC Genomics. Confirmed head-to-tail junctions are available in Figure S5D. A list of the oligos is given in Table S1.

For more details, see the Supplemental Experimental Procedures.

ACCESSION NUMBERS

The data have been deposited in the NCBI Gene Expression Omnibus and are available under accession numbers GSE63823 and SRP050149. circRNAs identified in this study, as well as *H* scores, are available at http://www.circbase.org.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.12.019.

AUTHOR CONTRIBUTIONS

A.I. performed all computational analyses except hyperediting. S.M. contributed experimental validation assays. E.W. designed and carried out ADAR RNAi, RNA-seq, and qPCR experiments and initial analysis of the RNA-seq data. H.P. and E.L. performed hyperediting detection. F.T. collected and prepared worm samples, carried out initial ADAR KDs, and helped in the design of the validation experiments. M.O. annotated worm circRNAs. M.P. performed an initial analysis of the RNA-seq data and was supervised by C.D. A.I. and N.R. analyzed the data and wrote the paper with input from E.L.

ACKNOWLEDGMENTS

A.I. and N.R. thank Sebastian Kadener (Hebrew University), Albrecht Bindereif (University of Giessen), and Marvin Jens (N.R. lab) for helpful discussions. We thank all members of the N.R. lab for discussions and support. This work was

supported by German-Israeli-Foundation for Scientific Research and Development (G.I.F) and German Ministry for Education and Research (SatNet program).

Received: September 16, 2014
Revised: November 25, 2014
Accepted: December 9, 2014
Published: December 31, 2014

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Ashwal-Fluss, R., Meyer, M., Pamudurti, N.R., Ivanov, A., Bartok, O., Hanan, M., Evantal, N., Memczak, S., Rajewsky, N., and Kadener, S. (2014). circRNA biogenesis competes with pre-mRNA splicing. Mol. Cell *56*, 55–66.

Athanasiadis, A., Rich, A., and Maas, S. (2004). Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol. *2*, e391.

Baltz, A.G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., et al. (2012). The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Mol. Cell *46*, 674–690.

Braun, S., Domdey, H., and Wiebauer, K. (1996). Inverse splicing of a discontinuous pre-mRNA intron generates a circular exon in a HeLa cell nuclear extract. Nucleic Acids Res. *24*, 4152–4157.

Capel, B., Swain, A., Nicolis, S., Hacker, A., Walter, M., Koopman, P., Goodfellow, P., and Lovell-Badge, R. (1993). Circular transcripts of the testis-determining gene Sry in adult mouse testis. Cell *73*, 1019–1030.

Carmi, S., Borukhov, I., and Levanon, E.Y. (2011). Identification of widespread ultra-edited human RNAs. PLoS Genet. *7*, e1002317.

Dubin, R.A., Kazmi, M.A., and Ostrer, H. (1995). Inverted repeats are necessary for circularization of the mouse testis Sry transcript. Gene *167*, 245–248.

Glažar, P., Papavasileiou, P., and Rajewsky, N. (2014). circBase: a database for circular RNAs. RNA *20*, 1666–1670.

Hansen, T.B., Wiklund, E.D., Bramsen, J.B., Villadsen, S.B., Statham, A.L., Clark, S.J., and Kjems, J. (2011). miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. EMBO J *30*, 4414–4422.

Hansen, T.B., Jensen, T.I., Clausen, B.H., Bramsen, J.B., Finsen, B., Damgaard, C.K., and Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. Nature *495*, 384–388.

Jeck, W.R., and Sharpless, N.E. (2014). Detecting and characterizing circular RNAs. Nat. Biotechnol. *32*, 453–461.

Jeck, W.R., Sorrentino, J.A., Wang, K., Slevin, M.K., Burd, C.E., Liu, J., Marzluff, W.F., and Sharpless, N.E. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA *19*, 141–157.

Levanon, E.Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z.Y., Shoshan, A., Pollock, S.R., Sztybel, D., et al. (2004). Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat. Biotechnol. *22*, 1001–1005.

Liang, D., and Wilusz, J.E. (2014). Short intronic repeat sequences facilitate circular RNA production. Genes Dev. *28*, 2233–2247.

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., Munschauer, M., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. Nature *495*, 333–338.

Morse, D.P., Aruscavage, P.J., and Bass, B.L. (2002). RNA hairpins in noncoding regions of human brain and Caenorhabditis elegans mRNA are edited by adenosine deaminases that act on RNA, *Volume 99* (USA: Proc. Natl. Acad. Sci), pp. 7906–7911.

Nishikura, K. (2010). Functions and regulation of RNA editing by ADAR deaminases. Annu. Rev. Biochem. *79*, 321–349.

Osenberg, S., Paz Yaacov, N., Safran, M., Moshkovitz, S., Shtrichman, R., Sherf, O., Jacob-Hirsch, J., Keshet, G., Amariglio, N., Itskovitz-Eldor, J., and Rechavi, G. (2010). Alu sequences in undifferentiated human embryonic stem cells display high levels of A-to-I RNA editing. PLoS ONE *5*, e11173.

Pasman, Z., Been, M.D., and Garcia-Blanco, M.A. (1996). Exon circularization in mammalian nuclear extracts. RNA *2*, 603–610.

Porath, H.T., Carmi, S., and Levanon, E.Y. (2014). A genome-wide map of hyper-edited RNA reveals numerous new sites. Nat. Commun. *5*, 4726.

Ramaswami, G., and Li, J.B. (2014). RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res. *42*, D109–D113.

Ramaswami, G., Lin, W., Piskol, R., Tan, M.H., Davis, C., and Li, J.B. (2012). Accurate identification of human Alu and non-Alu RNA editing sites. Nat. Methods *9*, 579–581.

Salzman, J., Gawad, C., Wang, P.L., Lacayo, N., and Brown, P.O. (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. PLoS ONE *7*, e30733.

Starke, S., Jost, I., Rossbach, O., Schneider, T., Schreiner, S., Hung, L.-H., and Bindereif, A. (2014). Exon circularization requires canonical splice signals. Cel Rep *10*, this issue.

Vasudevan, A.A.J., Smits, S.H.J., Höppner, A., Häussinger, D., Koenig, B.W., and Münk, C. (2013). Structural features of antiviral DNA cytidine deaminases. Biol. Chem. *394*, 1357–1370.

Wang, P.L., Bao, Y., Yee, M.-C., Barrett, S.P., Hogan, G.J., Olsen, M.N., Dinneny, J.R., Brown, P.O., and Salzman, J. (2014). Circular RNA is expressed across the eukaryotic tree of life. PLoS ONE *9*, e90859.

Zhang, X.-O., Wang, H.-B., Zhang, Y., Lu, X., Chen, L.-L., and Yang, L. (2014). Complementary sequence-mediated exon circularization. Cell *159*, 134–147.

OPEN ACCESS
CellPress

Cell Reports
Report

ThermoMouse: An In Vivo Model to Identify Modulators of UCP1 Expression in Brown Adipose Tissue

Andrea Galmozzi,[2,3] Si B. Sonne,[1,3,4] Svetlana Altshuler-Keylin,[1] Yutaka Hasegawa,[1] Kosaku Shinoda,[1] Ineke H.N. Luijten,[1,5] Jae Won Chang,[2] Louis Z. Sharp,[1] Benjamin F. Cravatt,[2] Enrique Saez,[2,*] and Shingo Kajimura[1,*]

[1]UCSF Diabetes Center, Department of Cell and Tissue Biology, University of California, San Francisco, 35 Medical Center Way, San Francisco, CA 94143, USA

[2]Department of Chemical Physiology and The Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

[3]Co-first author

[4]Present address: Department of Biology, University of Copenhagen, Copenhagen 2200, Denmark

[5]Present address: The Wenner-Gren Institute, The Arrhenius Laboratories, Stockholm University, Stockholm 106 91, Sweden

*Correspondence: esaez@scripps.edu (E.S.), skajimura@diabetes.ucsf.edu (S.K.)

http://dx.doi.org/10.1016/j.celrep.2014.10.066

SUMMARY

Obesity develops when energy intake chronically exceeds energy expenditure. Because brown adipose tissue (BAT) dissipates energy in the form of heat, increasing energy expenditure by augmenting BAT-mediated thermogenesis may represent an approach to counter obesity and its complications. The ability of BAT to dissipate energy is dependent on expression of mitochondrial uncoupling protein 1 (UCP1). To facilitate the identification of pharmacological modulators of BAT UCP1 levels, which may have potential as antiobesity medications, we developed a transgenic model in which luciferase activity faithfully mimics endogenous UCP1 expression and its response to physiologic stimuli. Phenotypic screening of a library using cells derived from this model yielded a small molecule that increases UCP1 expression in brown fat cells and mice. Upon adrenergic stimulation, compound-treated mice showed increased energy expenditure. These tools offer an opportunity to identify pharmacologic modulators of UCP1 expression and uncover regulatory pathways that impact BAT-mediated thermogenesis.

INTRODUCTION

The epidemic of obesity poses a dire public health problem, for obesity is a major risk factor for development of insulin resistance, type 2 diabetes, cardiovascular disease, and cancer. Obesity is the result of a sustained energy imbalance in which intake exceeds expenditure. Current antiobesity drugs work by limiting energy intake, either through suppression of appetite or inhibition of intestinal lipid absorption (Kim et al., 2014). These medications are effective, but side effects often associated with long-term use, such as depression or steatorrhea, limit patient compliance. The discovery of brown adipose tissue (BAT) in adult humans and its correlation with body mass index (Cypess et al., 2009; Saito et al., 2009; van Marken Lichtenbelt et al., 2009; Virtanen et al., 2009) suggests that active BAT may protect from obesity. BAT dissipates energy in the form of heat, thus increasing energy expenditure. It is thought that pharmacological activation of BAT thermogenesis may be an alternative approach to alter energy balance, one complementary to existing obesity medications (Nedergaard and Cannon, 2010; Kajimura and Saito, 2014).

The ability of BAT to produce heat is dependent on expression of the BAT-specific uncoupling protein 1 (UCP1). In response to exposure to cold or a high-fat diet, UCP1 reduces the mitochondrial membrane potential and uncouples cellular respiration from ATP synthesis, thereby generating heat. Because other UCP proteins (e.g., UCP2 and UCP3) do not contribute to adaptive thermogenesis (Golozoubova et al., 2001), UCP1 is thought to be solely responsible for adaptive nonshivering thermogenesis. UCP1-null mice are intolerant to cold (Enerbäck et al., 1997) and develop obesity at thermoneutral conditions (Feldmann et al., 2009). In contrast, transgenic expression of UCP1 in fat increases oxygen consumption in BAT and epididymal white adipose tissue (WAT) and reduces body weight gain (Kopecky et al., 1995). Contrary to the mechanism of action of small-molecule mitochondrial uncouplers such as 2,4-dinitrophenol that proved too toxic as weight loss agents (Grundlingh et al., 2011), UCP1-mediated uncoupling is a highly regulated process that requires direct binding of long-chain free fatty acids to UCP1 in response to physiologic cyclic AMP (cAMP) signaling (Fedorenko et al., 2012). A pharmacological approach to increase UCP1 expression and activity in adipose tissue is thus likely to constitute a safer avenue to enhance whole-body thermogenic capacity and energy expenditure. To test this concept, it is of great interest to identify small molecules that can stimulate UCP1 expression in fat tissue.

Phenotypic screens with adipocytes have proven to be a powerful method to isolate small molecules that ameliorate the symptoms of metabolic syndrome through novel mechanisms

of action (Waki et al., 2007; Dominguez et al., 2014). To facilitate the identification of pharmacologic agents to modulate UCP1 expression in adipocytes, we have generated a transgenic reporter mouse, ThermoMouse, in which luciferase activity recapitulates the pattern of expression of UCP1 in vivo and allows real-time visualization and quantification of UCP1 expression in live animals. A chemical screen using brown adipocytes derived from this model yielded a compound that can induce UCP1 expression in cells and enhance UCP1 expression in vivo. In response to adrenergic stimulation, mice treated with this compound show increased energy expenditure. These results demonstrate the utility of these models to identify pharmacological modulators of UCP1 levels. Discovery of compounds with this ability is an important stride toward the goal of enhancing BAT function in obese individuals with drug-like molecules.

RESULTS

Development of a Transgenic Model to Image UCP1 Expression In Vivo

To develop an in vivo reporter system to monitor endogenous UCP1 expression in a noninvasive manner, we generated transgenic mice that express luciferase2 under the control of the *Ucp1* genetic locus. A luciferase2-T2A-tdTomato cassette was inserted at the initiation codon of the *Ucp1* gene in a 98.6 kb bacterial artificial chromosome (BAC) containing the entire *Ucp1* gene locus (Figure S1A) and proper targeting confirmed (Figure S1B). Next, we used the IVIS Spectrum Imaging System to monitor luciferase activity in transgenic mice generated using this BAC construct. 3D imaging detected robust luciferase signals in interscapular BAT, perirenal BAT, and inguinal WAT (Figure 1A). No signals were detected in adipose depots of wild-type mice (Figure S1C). To assess if luciferase activity recapitulated endogenous UCP1 protein levels, we measured luciferase activity and UCP1 protein in BAT, WAT, liver, and muscle. A tight correlation was found between luciferase activity and endogenous UCP1 protein expression (Figures 1B and 1C), indicating that the reporter model mirrors the tissue distribution of UCP1.

To examine whether reporter mice responded to physiologic stimuli known to induce UCP1 expression and BAT activity, we monitored changes in luciferase activity during cold adaptation (Figure 1D). Luciferase signal in interscapular BAT was low in mice maintained at 28°C (Figure S1D) but increased robustly in mice kept at 9°C for 24 hr (Figures 1D and 1E). Luciferase induction correlated with increases in UCP1 mRNA (Figure 1F) and protein levels (Figure 1G). Similarly, subchronic (4 days) or acute (1 day) treatment of reporter mice kept at 28°C with a specific β_3 adrenergic receptor agonist (CL-316,243; 1 mg/kg) dramatically enhanced the luciferase signal and endogenous UCP1 protein levels in BAT (Figures 1H and 1I). To evaluate the response of beige cells in inguinal WAT, imaging was performed from a side angle. Strong luciferase activity was detected in inguinal WAT in response to chronic CL-316,243 treatment (Figure 1J). The increase in signal was paralleled by robust induction of UCP1 protein in this depot (Figure 1K). These results show that our *Ucp1* luciferase reporter accurately indicates changes in UCP1 elicited by physiological responses and is a useful tool to quantify changes in UCP1 expression in vivo.

A Cell-Based Screening Platform to Monitor UCP1 Protein Expression

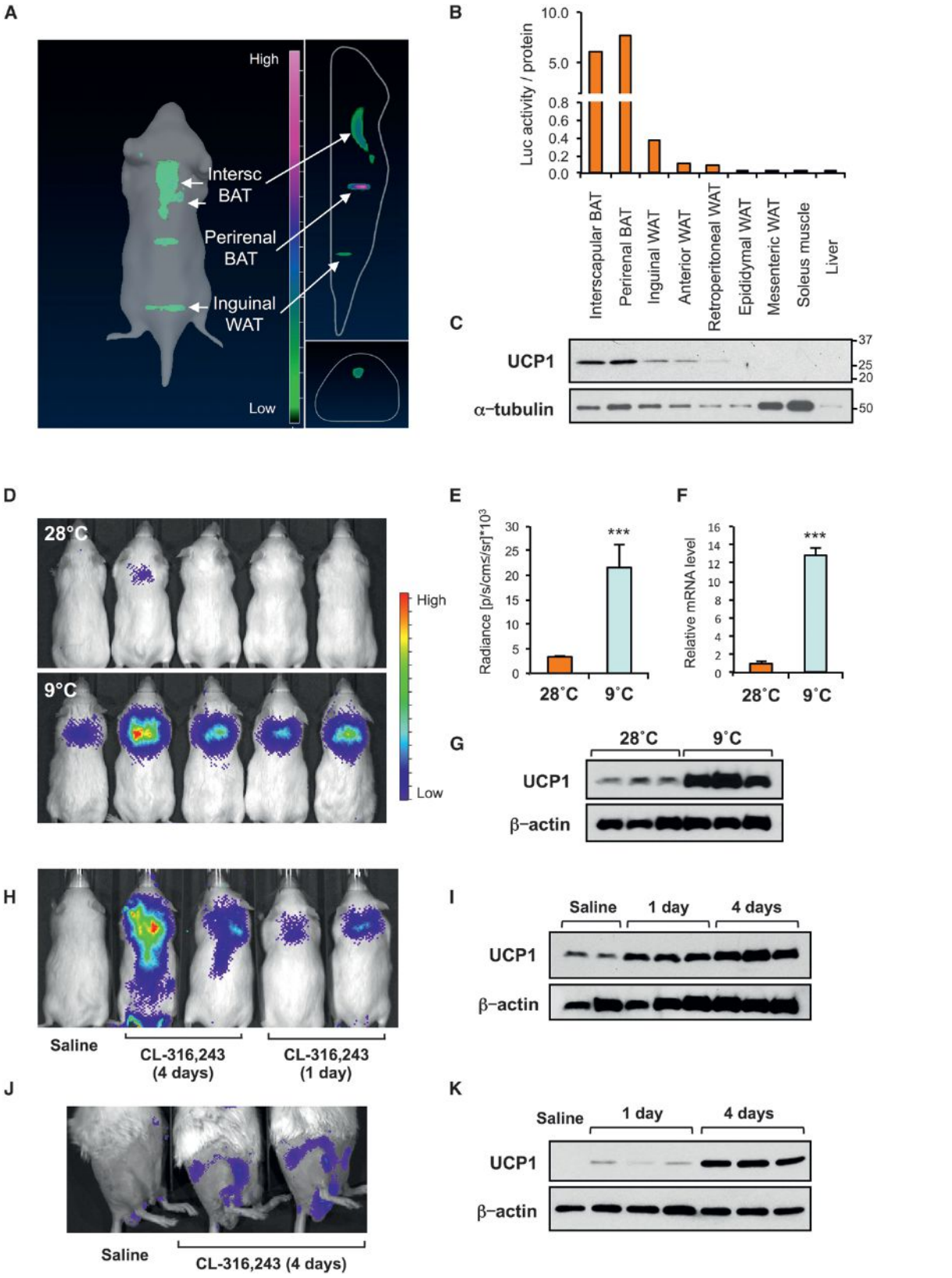
Next, to establish a cell-based system for quantifying UCP1 expression in fat cells suited for screening of small-molecule or genomic libraries, we generated immortalized preadipocyte lines from *Ucp1* luciferase transgenics. From six cell lines derived from interscapular BAT, we selected one that showed high adipogenic capacity and significant levels of UCP1 and luciferase expression. Because PPARγ ligands can stimulate UCP1 expression (Sears et al., 1996), we tested the response of this line to rosiglitazone. Luciferase activity was induced in a dose-dependent manner and correlated tightly with endogenous UCP1 protein levels (Figures 2A and 2B). Treatment with the cAMP-signaling activator forskolin also increased luciferase activity and UCP1 protein expression (Figures 2C and 2D).

To confirm that the *Ucp1* luciferase reporter cell line preserved the response to BAT activators, and that it could report sequential changes in UCP1 expression, nude mice were implanted with *Ucp1* luciferase preadipocytes subcutaneously and treated 6 days posttransplantation with saline or rosiglitazone (10 mg/kg per day) for 7 days. Luciferase activity in transplants of rosiglitazone-treated mice was increased significantly at 4 days of treatment and thereafter (Figures 2E and 2F). Transplanted preadipocytes in mice treated with rosiglitazone formed discrete adipose tissue containing multilocular adipocytes (Figure 2G, upper left). These adipocytes were positive for UCP1 protein (Figure 2G, upper right) and GFP (i.e., tdTomato; Figure 2G, lower left), indicating that transplanted cells retained brown adipogenic capacity in vivo. These observations indicate that this cell-based system is a robust surrogate to assess endogenous UCP1 expression.

A Small-Molecule Screen Identifies a Regulator of UCP1 Expression

To test the ability of our monitoring systems to identify regulators of UCP1 expression, we performed a phenotypic screen using a modestly sized library of small molecules, primarily carbamates and triazole ureas (Adibekian et al., 2011; Bachovchin et al., 2010). *Ucp1* luciferase brown preadipocytes were seeded and differentiated in 96-well plates. Mature adipocytes (day 8) were incubated with compounds and luciferase activity quantified 16 hr later (Figure 3A). To evaluate the power of the assay to identify compounds that modulate UCP1 levels in either direction, four hits with differing properties (activators and inhibitors) were selected for study (Figure 3B). As it is often the case with reporter-based screens, two hits could not be validated and may represent, for example, nonspecific stabilizers of luciferase signal. Of those that confirmed, compound 4 robustly increased luciferase activity (Figure 3C). In contrast, compound 3 significantly reduced luciferase signal. Importantly, compound 4 increased endogenous UCP1 protein expression to a similar level to that induced by rosiglitazone (Figure 3D). None of the compounds had any effect on brown preadipocytes (Figure S2).

Compound 4 (WWL113 in original nomenclature) was also a hit in a different, image-based screen we recently described, in which white preadipocytes were treated chronically (8 days) during differentiation (Dominguez et al., 2014). WWL113 inhibits



(legend on next page)

two serine hydrolases expressed in adipocytes, carboxylesterase 3 (Ces3 or Ces1d) and Ces1f (CesML1). However, the ability of WWL113 to induce UCP1 expression in brown adipocytes does not appear to be mediated by Ces3/1f inhibition, for a structurally distinct Ces3 inhibitor (WWL229) failed to have the same effect on UCP1 expression, whereas the urea version of WWL113 (WWL113U), which does not inhibit Ces3, retained the ability to induce UCP1 (Figure S3A). Nonetheless, we concluded that WWL113 could be a valuable chemical probe to further validate our reporter systems.

Induction of UCP1 Expression by WWL113 Relies on PPARα Signaling

To characterize the effect of WWL113 on endogenous UCP1 expression, we treated cultured brown adipocytes with several doses of WWL113. At a dose of 1 μM and higher, WWL113 significantly increased UCP1 protein expression (Figure 4A). The effect of WWL113 on UCP1 protein levels was largely due to activation of *Ucp1* transcription, because WWL113 powerfully increased expression of *Ucp1* mRNA (Figure 4B). WWL113 treatment also stimulated expression of mRNAs for other thermogenic genes, such as *Cidea*, *Pgc1a*, and *Cox7a1* (Figure 4B). WWL113 treatment had no effect on expression of the adipogenic marker *Adiponectin*. WWL113 treatment for 24 hr was sufficient to activate endogenous *Ucp1* mRNA expression without affecting expression of multiple adipogenic markers (Figures 4C and S3B), indicating that WWL113 enhanced the BAT-selective thermogenic gene program in a cell-autonomous manner without affecting adipogenesis per se. To test the functional consequences of WWL113 treatment, we examined the extent to which WWL113 could sensitize brown adipocytes to physiologic activators of BAT such as norepinephrine. WWL113 pretreatment enhanced the increase in *Ucp1* mRNA expression normally induced by norepinephrine (Figure 4D). More importantly, WWL113-treated cells showed greater total and uncoupled respiration in response to norepinephrine and greater uncoupled basal respiration (Figures 4E and S3C).

Next, we explored the mechanism by which WWL113 increased *Ucp1* transcription. Because PPARα is a critical regulator of thermogenic gene expression in BAT (Barbera et al., 2001), we hypothesized that the action of WWL113 in BAT could be mediated via PPARα. To test this notion, primary brown adipocytes were treated with a selective PPARα antagonist (GW6471) and/or a PPARγ antagonist (GW9662) in the presence or absence of WWL113. The capacity of WWL113 to increase *Ucp1* mRNA expression was largely blunted by treatment with GW6471, but not with GW9662 (Figure 4F). The inhibitory effect of GW6471 on WWL113-induced *Ucp1* mRNA expression was not altered when cells were cotreated with GW9662. These results indicate that the ability of WWL113 to enhance *Ucp1* expression is principally dependent on PPARα, but not PPARγ, activity. Because WWL113 is not a direct PPARα activator (Dominguez et al., 2014), we tested the extent to which cotreatment of differentiated brown adipocytes with WWL113 and a PPARα agonist (GW9578) would further boost *Ucp1* mRNA levels. Cotreatment with WWL113 and GW9578 modestly but significantly increased expression of *Ucp1* relative to treatment with either compound alone (Figure 4G), suggesting that WWL113 cooperates with the PPARα pathway to enhance UCP1 expression.

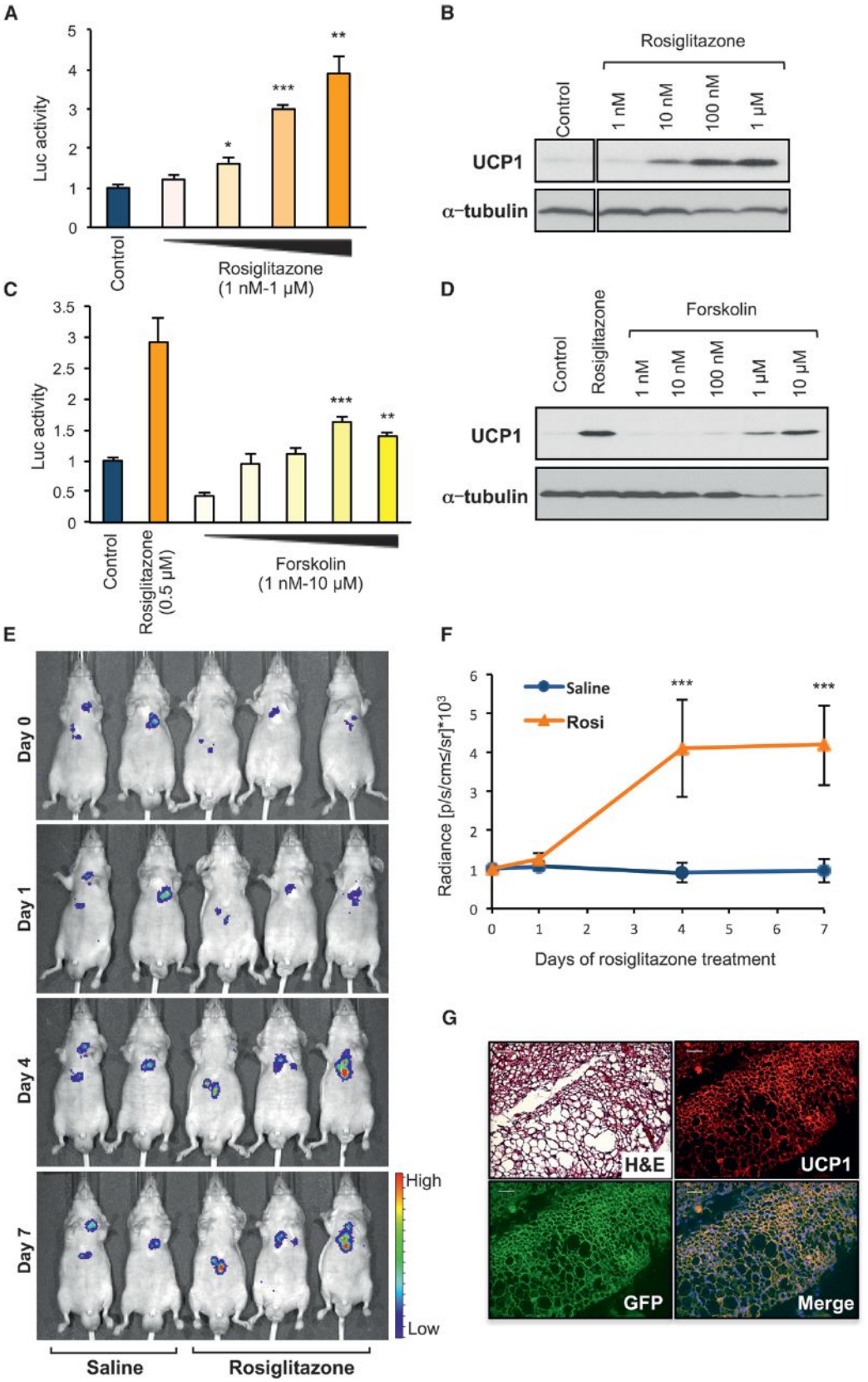
WWL113 Increases UCP1 Expression and the Thermogenic Response In Vivo

We next tested if WWL113 could enhance UCP1 expression in vivo. *Ucp1* luciferase mice were treated with vehicle or WWL113 (50 mg/kg once daily) for 5 days. This dose has been shown to be effective in mice (Dominguez et al., 2014). A robust and significant increase (5-fold) in *Ucp1*-driven luciferase expression was detected in the interscapular BAT depots of transgenic mice treated with WWL113 (Figures 5A and 5B). To confirm these findings, we examined the ability of WWL113 to enhance thermogenic gene expression in C57BL/6J mice treated with the compound for 5 days. WWL113 treatment induced significant increases in mRNA expression of *Ucp1* and thermogenic genes, such as *Pgc1a* and *Dio2*, in the BAT of wild-type mice (Figure 5C). No change in the general marker *Pparγ* was seen. Importantly, UCP1 protein expression was highly induced in vivo by WWL113 (Figure 5D). In contrast, no difference in mRNA expression of *Ucp1*, *Pgc1a*, *Dio2*, and *Pparγ* was observed in inguinal WAT of WWL113-treated mice (Figure S5A), indicating that the effects of WWL113 on the thermogenic gene program may be specific to brown fat.

Finally, we examined the effects of WWL113 on whole-body energy expenditure. Mice treated with WWL113 for 7 days showed no differences in basal energy expenditure, but upon adrenergic stimulation (CL-316,243 injection), they responded with a more robust increase in energy expenditure than controls (Figure 5E). WWL113 did not affect locomotor activity (Figure 5F), food intake (Figure 5G), or heart rate (Figure 5H). These data indicate that WWL113 increased UCP1 levels in vivo to a functionally meaningful degree, increasing the adaptive thermogenic capacity of treated mice.

Figure 1. Luciferase Imaging of UCP1 Expression In Vivo

(A) 3D reconstruction of luciferase signals in *Ucp1* luciferase reporter mice. Adipose depots with specific luciferase signal are indicated.
(B) Quantification of luciferase in adipose depots, skeletal muscle, and liver of mice kept at room temperature. Values normalized to protein content.
(C) UCP1 protein expression in tissue lysates from the mouse in (B).
(D) Luciferase activity in *Ucp1* luciferase reporter mice kept at 28°C and subsequently kept at 9°C for 24 hr. Representative mice are shown.
(E) Quantification of luciferase signal in interscapular BAT of *Ucp1* luciferase reporter mice shown in (D) (n = 9). ***$p < 0.001$. Data are expressed as mean ± SEM.
(F) *Ucp1* mRNA expression in interscapular BAT of *Ucp1* luciferase reporter mice kept at 28°C and 9°C for 24 hr. ***$p < 0.001$. Data are expressed as mean ± SEM.
(G) UCP1 protein expression in interscapular BAT of transgenic mice analyzed in (F).
(H) Representative image of luciferase signal in *Ucp1* luciferase reporter mice treated with saline or CL-316,243 for 1 day (acute) or 4 days (subchronic; n = 3).
(I) UCP1 protein expression in BAT of transgenic mice analyzed in (H).
(J) Representative images of luciferase signal in inguinal WAT depots of *Ucp1* luciferase reporter mice treated with saline or CL-316,243 for 4 days.
(K) UCP1 protein expression in inguinal WAT depots of *Ucp1* luciferase mice treated with saline or CL-316,243 for 1 day or 4 days.



(legend on next page)

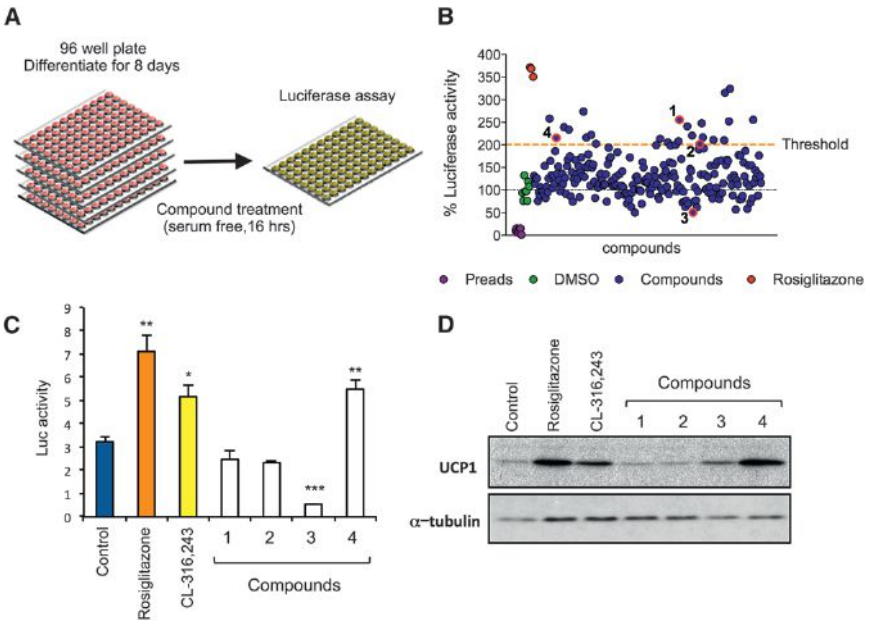


Figure 3. Small-Molecule Screen to Identify Regulators of UCP1 Expression

(A) Screen scheme. *Ucp1* luciferase brown preadipocytes were differentiated in 96-well plates and treated at day 8 with compounds for 16 hr (n = 3).

(B) Screen performance. Mean luciferase activity of compounds (blue circles) plotted relative to the value of DMSO-treated *Ucp1* luciferase adipocytes (100%; green circles). Rosiglitazone served as positive control (red circles). *Ucp1* luciferase preadipocytes (purple circles) were used to determine background and induction of signal upon differentiation. Hits selected for evaluation (blue circles, red outline) are numbered according to the scheme used in validation experiments.

(C) Luciferase activity in differentiated *Ucp1* luciferase brown adipocytes treated with compounds for 5 days. Rosiglitazone (0.5 μM) and CL-316,243 (10 nM) served as positive controls (n = 3). *p < 0.05; **p < 0.01; ***p < 0.001 versus control. Data are expressed as mean + SD.

(D) UCP1 protein expression in cells from (C).

DISCUSSION

We have developed an in vivo monitoring system that allows to quantitatively track sequential changes in UCP1 expression within the same individual. Because UCP1 expression shows a high level of interindividual variation (Boeuf et al., 2002) and changes dynamically during circadian oscillation (Gerhart-Hines et al., 2013), seasonal changes (Au-Yong et al., 2009), and aging (Rogers et al., 2012), this model may prove of wide utility. ^{18}fluoro-labeled 2-deoxy-glucose positron emission tomography (^{18}FDG-PET) scanning has been applied to assess BAT activity in rodents and humans (Cypess et al., 2009; Saito et al., 2009; van Marken Lichtenbelt et al., 2009; Virtanen et al., 2009), but detection of ^{18}FDG-PET signals in BAT depends entirely on glucose uptake. In contrast, the level of UCP1 expression in BAT is a more-direct measure of the thermogenic capacity of this tissue (Nedergaard and Cannon, 2010). Hence, the transgenic UCP1 reporter we describe provides an opportunity to identify signaling pathways and transcriptional events that control thermogenic capacity in brown adipocytes in vivo. Unlike prior UCP1 models (Cassard-Doulcier et al., 1993), it also enables characterization of modulators of BAT function in real time in live animals.

*Ucp*1 luciferase brown adipocyte lines derived from this transgenic retain the characteristics of bona fide brown fat cells. A phenotypic screen using these cells identified a compound, WWL113, that can increase UCP1 expression in vitro and in vivo. It is important to note that UCP1's uncoupling activity is dependent on sympathetic nerve activation and increased intracellular cAMP levels. As BAT activity is stimulated by cold exposure, long-chain free fatty acids supplied from cAMP-induced lipolysis and from the circulation directly bind UCP1 and serve as a substrate to transport protons into the mitochondrial matrix (Fedorenko et al., 2012). Thus, small molecules that solely stimulate UCP1 expression are unlikely to induce substantial thermogenesis. In agreement with this notion, we found that the effect of WWL113 treatment on cellular respiration was greater when cells were stimulated with norepinephrine. In that setting, this chemical probe significantly increased total and uncoupled cellular respiration. More importantly, mice treated with WWL113 showed considerably enhanced energy expenditure, but this increase required adrenergic stimulation. WWL113-treated mice showed no

Figure 2. Cell-Based System to Monitor UCP1 Expression

(A) Luciferase activity in immortalized *Ucp1* luciferase brown adipocytes. Differentiated adipocytes were treated with DMSO (control) or rosiglitazone for 5 days (n = 3).

(B) UCP1 protein expression in cells analyzed in (A).

(C) Luciferase activity in differentiated *Ucp1* luciferase brown adipocytes treated with DMSO (control), forskolin, or rosiglitazone (0.5 μM) for 5 days (n = 3).

(D) UCP1 protein expression in cells analyzed in (C).

(E) Luciferase activity monitored at days 0, 1, 4, and 7 after the start of rosiglitazone treatment in mice implanted with *Ucp1* luciferase immortalized preadipocytes. Saline (n = 4) or rosiglitazone (n = 6; 10 mg/kg) treatment started 6 days postimplantation. Representative mice are shown.

(F) Sequential changes in luciferase activity measured in fat transplants from mice in (E).

(G) Hematoxylin and eosin and immunofluorescent stains for UCP1 or GFP (i.e., tdTomato) in transplants from mice treated with rosiglitazone. Merged UCP1 and GFP image counterstained with DAPI. The scale bar represents 50 μm. H&E, hematoxylin and eosin staining.

*p < 0.05; **p < 0.01; ***p < 0.001. Data are expressed as mean ± SEM.

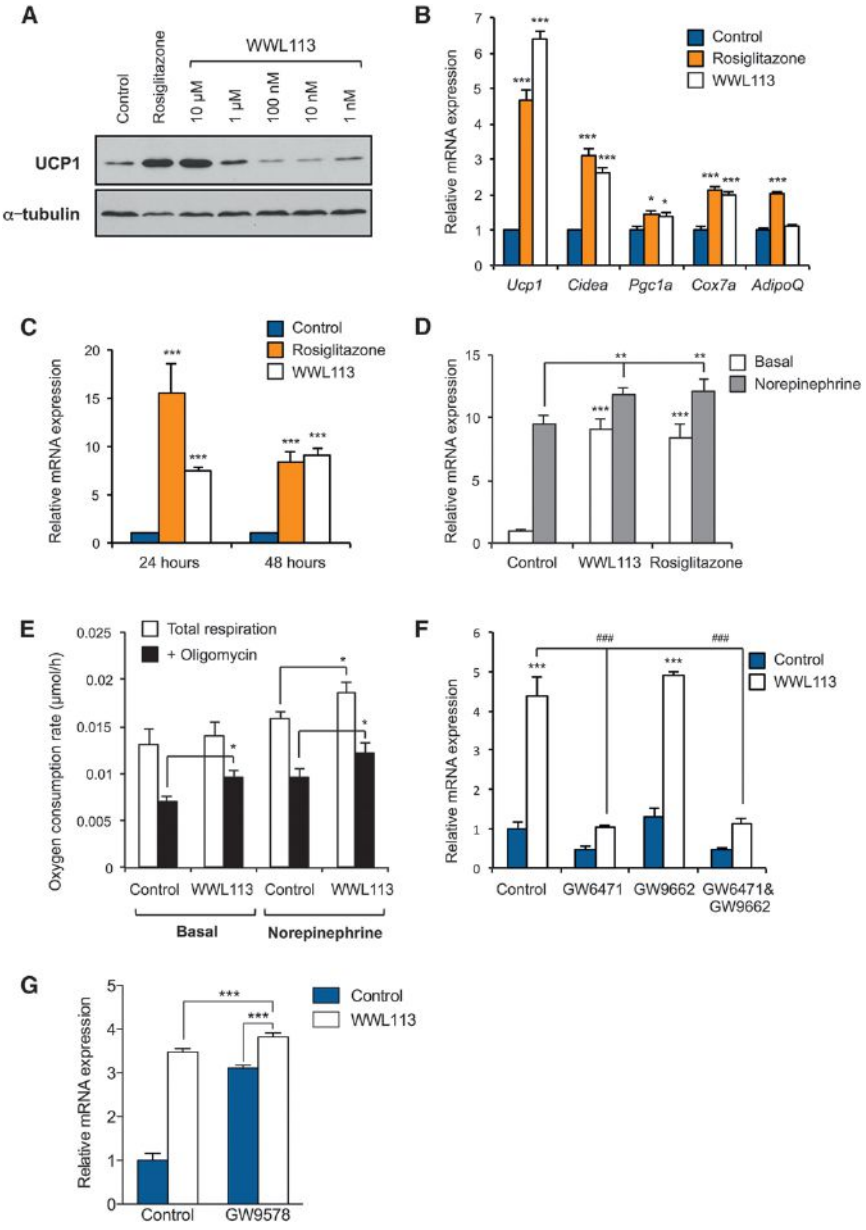


Figure 4. Effect of WWL113 on UCP1 Expression and Cellular Respiration

(A) UCP1 protein expression in differentiated *Ucp1* luciferase brown adipocytes treated with WWL113 for 5 days. Rosiglitazone (0.5 μM) served as positive control.

(B) Expression of thermogenic genes in differentiated *Ucp1* luciferase brown adipocytes treated with WWL113 (10 μM) for 5 days. Rosiglitazone (0.5 μM) served as positive control (n = 3). *p < 0.05; ***p < 0.001 versus control.

(C) *Ucp1* mRNA expression in primary differentiated brown adipocytes treated with WWL113 (10 μM) or rosiglitazone (5 μM) for 24 or 48 hr (n = 3). ***p < 0.001 versus control.

(D) *Ucp1* mRNA expression in primary differentiated brown adipocytes treated with WWL113 (10 μM) or rosiglitazone (5 μM) for 48 hr. Norepinephrine (0.1 μM) was added 8 hr prior to harvest (n = 3). **p < 0.01; ***p < 0.001 versus control.

(E) Total and uncoupled (oligomycin-insensitive) respiration of differentiated brown adipocytes (5×10^5 cells/sample) treated with WWL113 (10 μM) in the presence or absence of norepinephrine (0.1 μM; n = 3 to 4). *p < 0.05 versus control.

(F) *Ucp1* mRNA expression in differentiated brown adipocytes treated with vehicle or WWL113 (10 μM) for 24 hr with or without 30 min pretreatment with a PPARα-selective antagonist (GW6471; 3 μM) and/or a PPARγ-selective antagonist (GW9662; 10 μM; n = 3). ***p < 0.001 versus control. ###p < 0.001 versus WWL113-treated cells.

(G) *Ucp1* mRNA expression in differentiated brown adipocytes treated with vehicle, WWL113 (10 μM), a PPARα-selective agonist (GW9578), or the combination for 24 hr (n = 4). *p < 0.05; ***p < 0.001.

Data are expressed as mean + SD.

differences in locomotor activity, food intake, or heartbeat, indicating that the compound does not enhance basal sympathetic drive. Together with the finding that WWL113 increases UCP1 expression in brown fat cells in a cell-autonomous manner, these results suggest that WWL113 boosts energy expenditure primarily by increasing the content of UCP1 in BAT that can be activated by physiologic stimuli such as cold.

We previously reported that chronic administration (2 months) of WWL113 to obese-diabetic mice reduced body weight gain, improved systemic glucose and lipid homeostasis, and cleared hepatic steatosis (Dominguez et al., 2014). We ascribed the effects of WWL113 to inhibition of Ces3 in WAT and liver. Using a distinct phenotypic screen as a starting point, in this study, we have found that WWL113 can have Ces3-independent effects in BAT. WWL113 is not a direct activator of PPARα (Dominguez et al., 2014), but its effects on UCP1 expression in brown fat cells depend to a large extent, though not completely, on PPARα signaling. Although the molecular target(s) for WWL113's action in BAT remains to be clarified, this tool compound has nonetheless demonstrated the utility of our UCP1-monitoring systems to identify pharmacologic activators of UCP1 expression.

BAT is the major adipose depot that contains UCP1-positive adipocytes (classical brown adipocytes), but rodents and humans also possess an inducible type of thermogenic fat cells, termed beige or brite adipocytes (Sharp et al., 2012; Wu et al., 2012; Cypess et al., 2013; Lidell et al., 2013). UCP1-positive beige adipocytes emerge within WAT in response to external cues, such as sustained cold exposure or exercise. Beige adipocytes are considered promising reservoirs for enhancing energy expenditure, but current technologies (e.g., ^{18}FDG-PET and MRI scans) do not possess enough resolution to detect beige cells in vivo. Our data indicate that our *Ucp1* luciferase mouse may

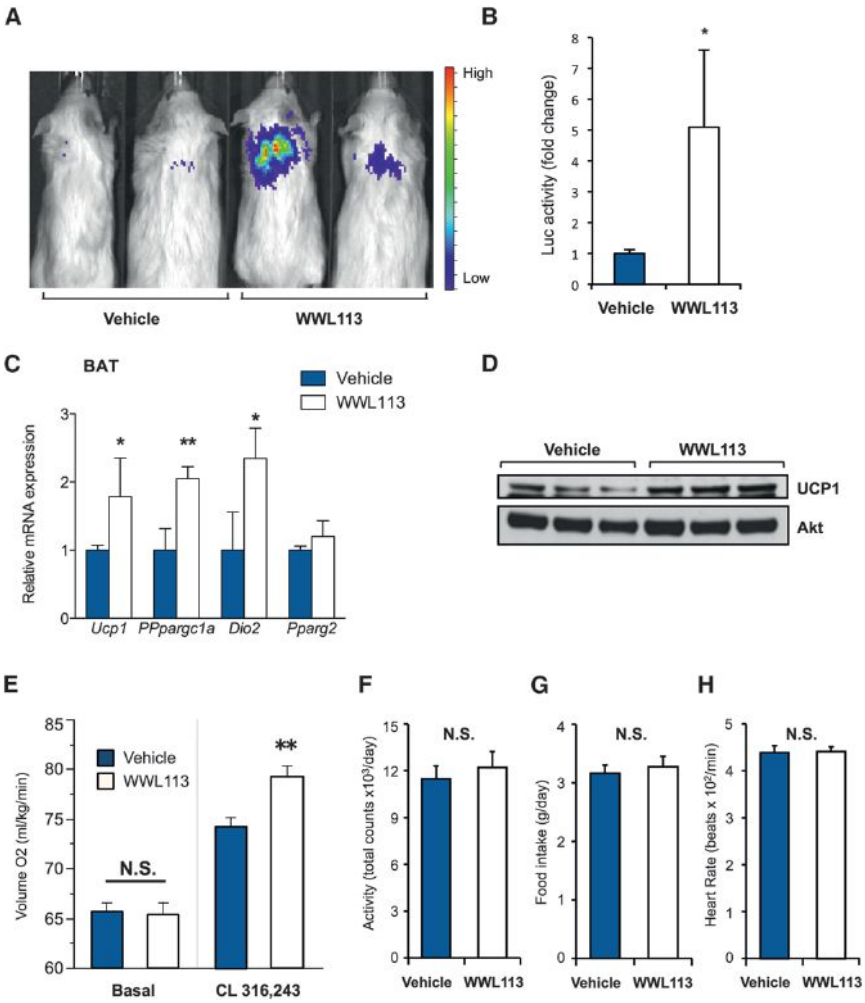


Figure 5. WWL113 Increases UCP1 Expression in Mice

(A) Luciferase activity in *Ucp1* luciferase reporter mice treated daily with WWL113 (50 mg/kg) or vehicle for 5 days (n = 5). Representative mice are shown.

(B) Quantification of luciferase signal in interscapular BAT of mice treated as in (A). Values normalized to protein content and shown as fold change relative to vehicle.

(C) Expression of thermogenic genes in interscapular BAT of C57BL/6 mice treated daily with WWL113 (50 mg/kg) or vehicle for 5 days (n = 5). *p < 0.05; **p < 0.01

(D) UCP1 protein expression in interscapular BAT of C57BL/6 mice analyzed in (C).

(E) VO_2 of wild-type mice treated daily with WWL113 (50 mg/kg) or vehicle for 7 days (n = 6). CL-316,243 (1 mg/kg) was injected to examine the response to adrenergic stimulation. **p < 0.01. N.S., not significant.

(F) Locomotor activity of mice in (E).

(G) Food intake of mice in (E).

(H) Heart rate of mice in (E).

Data are expressed as means ± SEM.

serve as a tool to monitor the effects of compounds and biological factors on beige cells.

Although they share the thermogenic ability of brown adipocytes, beige adipocytes have a distinct, heterogeneous developmental origin that is not fully understood. Beige adipocytes in multiple WAT depots can arise from *Myf5*-positive and negative cells (Sanchez-Gurmaches and Guertin, 2014), and a subset of inguinal beige adipocytes originates from a smooth-muscle lineage (Long et al., 2014). Prior work has shown that regulation of *Ucp1* is distinct in the two types of thermogenic adipocytes (Guerra et al., 1998; Koza et al., 2000; Xue et al., 2007). We found that WWL113 could activate UCP1 expression in BAT, but not in inguinal WAT, implying that WWL113-initiated signaling events that regulate UCP1 expression may be unique to brown adipocytes.

BAT and liver are the major organs responsible for triglyceride uptake (Bartelt et al., 2011), and emerging evidence points to a close connection between BAT activity, liver function, and systemic lipid homeostasis. For example, defects in thermogenesis caused by BAT-specific knockout of the enzyme euchromatic histone-lysine N-methyltransferase 1 resulted in hepatic steatosis and insulin resistance, even when weight-matched mice were compared (Ohno et al., 2013). Conversely, activation of BAT thermogenesis, for example by overexpression of PTEN, reduced hepatic lipid accumulation and enhanced systemic insulin sensitivity (Ortega-Molina et al., 2012). Thus, the therapeutic benefits of increased BAT thermogenesis may not be limited to effects on weight but could include improvements in lipid homeostasis and whole-body insulin sensitivity. The set of tools we have developed should aid the development of pharmacologically tractable approaches to activate BAT function as a therapy against obesity and its complications.

EXPERIMENTAL PROCEDURES

Animals

Experiments were approved by the Institutional Animal Care and Use Committees of the University of California, San Francisco, and The Scripps Research Institute. Unless stated, mice were kept at room temperature. WWL113 was administered either intraperitoneally once daily (50 mg/kg in a 4:1 PEG300:Tween 80 vehicle solution) or orally (50 mg/kg in 0.5% hydroxypropylmethylcellulose). At the conclusion of treatment, BAT, WAT, skeletal muscle, and liver were snap frozen for luciferase assays and RNA and protein analysis. Energy balance studies were performed over the course of 7 days using a Comprehensive Lab Animal Monitoring System (Columbus Instruments) as described (Ohno et al., 2013). Data were normalized to body weight, as there were no differences between groups. Heart rate in the conscious state was measured by the indirect tail cuff method with a SC1000 MSP (Hatteras Instruments).

Generation of Ucp1 Luciferase Reporter Mice

A 98.6 kb BAC (bMQ353d13; Source BioScience) containing the entire *Ucp1* gene locus was obtained and a luciferase2-T2A-tdTomato reporter cassette (Addgene 32904) inserted at the initiation codon of the *Ucp1*-coding sequence located in exon 1 using BAC recombineering techniques (Warming et al.,

2005). *Ucp1 luciferase* BAC DNA was microinjected into single-cell FVB embryos and transgenic founders and their offspring identified by PCR (primers provided in Table S1). Segregation patterns indicate that the transgene inserted into the Y chromosome. Transgenics display no decrease in fertility or any other abnormalities.

In Vivo Luciferase Imaging

Luciferase activity was monitored using an IVIS Spectrum Instrument (Caliper Life Sciences). For 3D reconstruction, six images (exposure time: 180 s; binning: L; F/Stop:1; emission filters: 560–660 nm; field of view: C) were collected starting 8 min after injection of 150 mg/kg luciferin (Goldbio). In other experiments, one image per scan (exposure time: 300 s; binning: M; F/Stop:1; emission filters: open; field of view: D) was acquired 15 min after luciferin injection. 3D reconstructions and luciferase activity were calculated using Living Image Software (Caliper Life Sciences).

Immortalized Ucp1 Luciferase Adipocyte Lines

The stromal vascular fraction from interscapular BAT of 3-week-old male *UCP1 luciferase* transgenic mice was isolated (Ohno et al., 2012) and cells infected with a retrovirus expressing large T antigen (pBabe SV40 Large T antigen; Addgene) and selected in puromycin (2 μg/ml). Immortalized preadipocytes were cultured in Dulbecco's modified Eagle's medium (DMEM), 10% fetal bovine serum (FBS), penicillin, and streptomycin. Upon 100% confluence (day 0), differentiation was induced with medium containing 10% FBS, 5 μg/ml insulin, 1 nM T3, 0.125 mM indomethacin, 2 μg/ml dexamethasone, and 0.125 mM 3-isobutyl-1-methylxanthine for 2 days. From day 2 on, cells were cultured only in the presence of insulin and T3.

Phenotypic Screen

A library of approximately 300 compounds (Adibekian et al., 2011; Bachovchin et al., 2010) was screened. Immortalized *Ucp1* luciferase brown preadipocytes were seeded in 96-well plates and differentiated as described above. At day 8, cells were exposed to compounds (5 μM for carbamates, 2.5 μM for triazole ureas) in serum-free DMEM. Rosiglitazone (0.5 μM) served as positive control. After 16 hr, media was replaced with Glo Lysis Buffer (Promega) and luciferase activity quantified in a PHERAstar reader (BMG Labtech). Activity was normalized relative to the signal in DMSO-treated cells in each plate. Assays were performed in triplicate. Compounds inducing >50% increase or decrease in luciferase activity were selected for follow-up.

Luciferase Assays

Cells or tissues were lysed in Cell Culture Lysis Reagent (Promega) and luciferase activity quantified using the Luciferase Assay System (Promega). Activity was measured in an Optocomp I reader (MGM Instruments) and normalized to total protein content.

Gene Expression

Total RNA was isolated using RiboZol reagents (AMRESCO) and reversed transcribed using an iScript cDNA synthesis kit (Bio-Rad) and quantitative real-time PCR performed using SYBR green and an ABI ViiA7 machine. Relative mRNA expression was determined by the ΔΔ-Ct method with TATA-binding protein as a normalization control. Primer sequences are provided in Table S1.

Metabolic Studies

Whole-body energy expenditure was measured at ambient temperature using a Comprehensive Lab Animal Monitoring System (Columbus Instruments) at the University of California, San Francisco (UCSF) animal metabolic core facility. Wild-type mice were treated daily with WWL113 (50 mg/kg) or vehicle for 7 days (n = 6). The mice were injected intraperitoneally with a β3-adrenergic receptor-specific agonist CL-316,243 at a dose of 1 mg/kg to examine the response to adrenergic stimulation. VO_2 was normalized by body weight.

Western Analysis

Proteomes were separated by SDS-PAGE. Antibodies used: UCP1 (Abcam; ab10983); Akt (Cell Signaling Technology; 9272); α-tubulin (Sigma; T8203); and β-actin (Sigma-Aldrich; AC-15).

Cellular Respiration

Oxygen consumption rate was measured in a MT200A Cell Respirometer (Strathkelvin), as previously described (Kajimura et al., 2009). Briefly, differentiated brown adipocytes treated with DMSO or WWL113 (10 μM) for 48 hr were trypsinized and incubated in serum-free medium in the presence or absence of norepinephrine. Uncoupled and nonmitochondrial cellular respiration were measured using oligomycin (1 μM) and antimycin A (1 μM).

Fat Transplantation

Preadipocytes were implanted as described (Kajimura et al., 2009). Briefly, cultured immortalized *Ucp1* luciferase preadipocytes were trypsinized, washed, and resuspended in PBS. Preadipocytes in a volume of 300 μl (~4 × 10^7 cells) were injected subcutaneously into NCr nude mice (Taconic). Six days after transplantation, mice were injected with either saline (n = 4) or rosiglitazone (n = 6; 10 mg/kg) twice daily for 7 days. Luciferase activity was monitored on days 0, 1, 4, and 7.

Histology

Tissues were fixed in 4% paraformaldehyde and embedded in paraffin. Sections (7 μm) were analyzed with hematoxylin and eosin staining and immunofluorescence to detect UCP1 (ab10983 1:1,000 in 5% goat serum) or GFP (GFP-1020 1:500 in 5% goat serum) as described (Sharp et al., 2012).

Statistics

Statistical significance was defined as $p < 0.05$ and determined by two-tailed Student t tests, Wilcoxon, or ANOVA, with Dunne's multiple comparison post hoc analysis.

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures and one table and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.10.066.

AUTHOR CONTRIBUTIONS

A.G., S.B.S., E.S., and S.K. designed experiments. J.W.C. and B.F.C. provided libraries and synthesized compounds. S.K. and S.B.S. generated transgenic model and cell lines. A.G. performed screen and follow-up. A.G., S.B.S., S.A.-K., Y.H., K.S., and S.K. performed in vivo and vitro experiments. L.Z.S. and I.H.N.L. provided technical help. A.G., S.B.S., E.S., and S.K. wrote the manuscript.

ACKNOWLEDGMENTS

We thank Haemin Hong, Kathleen Jay, Christophe Paillart, and Denis Glenn for assistance. Work was supported by NIH grants DK087853 and DK97441 to S.K. and DK099810 and CA179489 to E.S. S.K. acknowledges support from the DERC center grant (DK63720), UCSF PBBR program, the Pew Charitable Trust, and PRESTO from Japan Science and Technology Agency. S.B.S. was supported by a fellowship from the Alfred Benzon Foundation and A.G. by fellowship 14POST18200019 from the American Heart Association.

Received: April 2, 2014
Revised: October 13, 2014
Accepted: October 24, 2014
Published: November 26, 2014

REFERENCES

Adibekian, A., Martin, B.R., Wang, C., Hsu, K.L., Bachovchin, D.A., Niessen, S., Hoover, H., and Cravatt, B.F. (2011). Click-generated triazole ureas as ultrapotent in vivo-active serine hydrolase inhibitors. Nat. Chem. Biol. *7*, 469–478.

Au-Yong, I.T., Thorn, N., Ganatra, R., Perkins, A.C., and Symonds, M.E. (2009). Brown adipose tissue and seasonal variation in humans. Diabetes *58*, 2583–2587.

Bachovchin, D.A., Ji, T., Li, W., Simon, G.M., Blankman, J.L., Adibekian, A., Hoover, H., Niessen, S., and Cravatt, B.F. (2010). Superfamily-wide portrait of serine hydrolase inhibition achieved by library-versus-library screening. Proc. Natl. Acad. Sci. USA *107*, 20941–20946.

Barbera, M.J., Schluter, A., Pedraza, N., Iglesias, R., Villarroya, F., and Giralt, M. (2001). Peroxisome proliferator-activated receptor alpha activates transcription of the brown fat uncoupling protein-1 gene. A link between regulation of the thermogenic and lipid oxidation pathways in the brown fat cell. J. Biol. Chem. *276*, 1486–1493.

Bartelt, A., Bruns, O.T., Reimer, R., Hohenberg, H., Ittrich, H., Peldschus, K., Kaul, M.G., Tromsdorf, U.I., Weller, H., Waurisch, C., et al. (2011). Brown adipose tissue activity controls triglyceride clearance. Nat. Med. *17*, 200–205.

Boeuf, S., Keijer, J., Franssen-Van Hal, N.L., and Klaus, S. (2002). Individual variation of adipose gene expression and identification of covariated genes by cDNA microarrays. Physiol. Genomics *11*, 31–36.

Cassard-Doulcier, A.M., Gelly, C., Fox, N., Schrementi, J., Raimbault, S., Klaus, S., Forest, C., Bouillaud, F., and Ricquier, D. (1993). Tissue-specific and β-adrenergic regulation of the mitochondrial uncoupling protein gene: control by *cis*-acting elements in the 5′-flanking region. Mol. Endocrinol. *7*, 497–506.

Cypess, A.M., Lehman, S., Williams, G., Tal, I., Rodman, D., Goldfine, A.B., Kuo, F.C., Palmer, E.L., Tseng, Y.H., Doria, A., et al. (2009). Identification and importance of brown adipose tissue in adult humans. N. Engl. J. Med. *360*, 1509–1517.

Cypess, A.M., White, A.P., Vernochet, C., Schulz, T.J., Xue, R., Sass, C.A., Huang, T.L., Roberts-Toler, C., Weiner, L.S., Sze, C., et al. (2013). Anatomical localization, gene expression profiling and functional characterization of adult human neck brown fat. Nat. Med. *19*, 635–639.

Dominguez, E., Galmozzi, A., Chang, J.W., Hsu, K.L., Pawlak, J., Li, W., Godio, C., Thomas, J., Partida, D., Niessen, S., et al. (2014). Integrated phenotypic and activity-based profiling links Ces3 to obesity and diabetes. Nat. Chem. Biol. *10*, 113–121.

Enerbäck, S., Jacobsson, A., Simpson, E.M., Guerra, C., Yamashita, H., Harper, M.E., and Kozak, L.P. (1997). Mice lacking mitochondrial uncoupling protein are cold-sensitive but not obese. Nature *387*, 90–94.

Fedorenko, A., Lishko, P.V., and Kirichok, Y. (2012). Mechanism of fatty-acid-dependent UCP1 uncoupling in brown fat mitochondria. Cell *151*, 400–413.

Feldmann, H.M., Golozoubova, V., Cannon, B., and Nedergaard, J. (2009). UCP1 ablation induces obesity and abolishes diet-induced thermogenesis in mice exempt from thermal stress by living at thermoneutrality. Cell Metab. *9*, 203–209.

Gerhart-Hines, Z., Feng, D., Emmett, M.J., Everett, L.J., Loro, E., Briggs, E.R., Bugge, A., Hou, C., Ferrara, C., Seale, P., et al. (2013). The nuclear receptor Rev-erbα controls circadian thermogenic plasticity. Nature *503*, 410–413.

Golozoubova, V., Hohtola, E., Matthias, A., Jacobsson, A., Cannon, B., and Nedergaard, J. (2001). Only UCP1 can mediate adaptive nonshivering thermogenesis in the cold. FASEB J. *15*, 2048–2050.

Grundlingh, J., Dargan, P.I., El-Zanfaly, M., and Wood, D.M. (2011). 2,4-dinitrophenol (DNP): a weight loss agent with significant acute toxicity and risk of death. J. Med. Toxicol. *7*, 205–212.

Guerra, C., Koza, R.A., Yamashita, H., Walsh, K., and Kozak, L.P. (1998). Emergence of brown adipocytes in white fat in mice is under genetic control. Effects on body weight and adiposity. J. Clin. Invest. *102*, 412–420.

Kajimura, S., and Saito, M. (2014). A new era in brown adipose tissue biology: molecular control of brown fat development and energy homeostasis. Annu. Rev. Physiol. *76*, 225–249.

Kajimura, S., Seale, P., Kubota, K., Lunsford, E., Frangioni, J.V., Gygi, S.P., and Spiegelman, B.M. (2009). Initiation of myoblast to brown fat switch by a PRDM16-C/EBP-beta transcriptional complex. Nature *460*, 1154–1158.

Kim, G.W., Lin, J.E., Blomain, E.S., and Waldman, S.A. (2014). Antiobesity pharmacotherapy: new drugs and emerging targets. Clin. Pharmacol. Ther. *95*, 53–66.

Kopecky, J., Clarke, G., Enerbäck, S., Spiegelman, B., and Kozak, L.P. (1995). Expression of the mitochondrial uncoupling protein gene from the aP2 gene promoter prevents genetic obesity. J. Clin. Invest. *96*, 2914–2923.

Koza, R.A., Hohmann, S.M., Guerra, C., Rossmeisl, M., and Kozak, L.P. (2000). Synergistic gene interactions control the induction of the mitochondrial uncoupling protein (Ucp1) gene in white fat tissue. J. Biol. Chem. *275*, 34486–34492.

Lidell, M.E., Betz, M.J., Dahlqvist Leinhard, O., Heglind, M., Elander, L., Slawik, M., Mussack, T., Nilsson, D., Romu, T., Nuutila, P., et al. (2013). Evidence for two types of brown adipose tissue in humans. Nat. Med. *19*, 631–634.

Long, J.Z., Svensson, K.J., Tsai, L., Zeng, X., Roh, H.C., Kong, X., Rao, R.R., Lou, J., Lokurkar, I., Baur, W., et al. (2014). A smooth muscle-like origin for beige adipocytes. Cell Metab. *19*, 810–820.

Nedergaard, J., and Cannon, B. (2010). The changed metabolic world with human brown adipose tissue: therapeutic visions. Cell Metab. *11*, 268–272.

Ohno, H., Shinoda, K., Spiegelman, B.M., and Kajimura, S. (2012). PPARγ agonists induce a white-to-brown fat conversion through stabilization of PRDM16 protein. Cell Metab. *15*, 395–404.

Ohno, H., Shinoda, K., Ohyama, K., Sharp, L.Z., and Kajimura, S. (2013). EHMT1 controls brown adipose cell fate and thermogenesis through the PRDM16 complex. Nature *504*, 163–167.

Ortega-Molina, A., Efeyan, A., Lopez-Guadamillas, E., Muñoz-Martin, M., Gómez-López, G., Cañamero, M., Mulero, F., Pastor, J., Martinez, S., Romanos, E., et al. (2012). Pten positively regulates brown adipose function, energy expenditure, and longevity. Cell Metab. *15*, 382–394.

Rogers, N.H., Landa, A., Park, S., and Smith, R.G. (2012). Aging leads to a programmed loss of brown adipocytes in murine subcutaneous white adipose tissue. Aging Cell *11*, 1074–1083.

Saito, M., Okamatsu-Ogura, Y., Matsushita, M., Watanabe, K., Yoneshiro, T., Nio-Kobayashi, J., Iwanaga, T., Miyagawa, M., Kameya, T., Nakada, K., et al. (2009). High incidence of metabolically active brown adipose tissue in healthy adult humans: effects of cold exposure and adiposity. Diabetes *58*, 1526–1531.

Sanchez-Gurmaches, J., and Guertin, D.A. (2014). Adipocyte lineages: tracing back the origins of fat. Biochim. Biophys. Acta *1842*, 340–351.

Sears, I.B., MacGinnitie, M.A., Kovacs, L.G., and Graves, R.A. (1996). Differentiation-dependent expression of the brown adipocyte uncoupling protein gene: regulation by peroxisome proliferator-activated receptor gamma. Mol. Cell. Biol. *16*, 3410–3419.

Sharp, L.Z., Shinoda, K., Ohno, H., Scheel, D.W., Tomoda, E., Ruiz, L., Hu, H., Wang, L., Pavlova, Z., Gilsanz, V., and Kajimura, S. (2012). Human BAT possesses molecular signatures that resemble beige/brite cells. PLoS ONE *7*, e49452.

van Marken Lichtenbelt, W.D., Vanhommerig, J.W., Smulders, N.M., Drossaerts, J.M., Kemerink, G.J., Bouvy, N.D., Schrauwen, P., and Teule, G.J. (2009). Cold-activated brown adipose tissue in healthy men. N. Engl. J. Med. *360*, 1500–1508.

Virtanen, K.A., Lidell, M.E., Orava, J., Heglind, M., Westergren, R., Niemi, T., Taittonen, M., Laine, J., Savisto, N.J., Enerbäck, S., and Nuutila, P. (2009). Functional brown adipose tissue in healthy adults. N. Engl. J. Med. *360*, 1518–1525.

Waki, H., Park, K.W., Mitro, N., Pei, L., Damoiseaux, R., Wilpitz, D.C., Reue, K., Saez, E., and Tontonoz, P. (2007). The small molecule harmine is an antidiabetic cell-type-specific regulator of PPARgamma expression. Cell Metab. *5*, 357–370.

Warming, S., Costantino, N., Court, D.L., Jenkins, N.A., and Copeland, N.G. (2005). Simple and highly efficient BAC recombineering using galK selection. Nucleic Acids Res. *33*, e36.

Wu, J., Boström, P., Sparks, L.M., Ye, L., Choi, J.H., Giang, A.H., Khandekar, M., Virtanen, K.A., Nuutila, P., Schaart, G., et al. (2012). Beige adipocytes are a distinct type of thermogenic fat cell in mouse and human. Cell *150*, 366–376.

Xue, B., Rim, J.S., Hogan, J.C., Coulter, A.A., Koza, R.A., and Kozak, L.P. (2007). Genetic variability affects the development of brown adipocytes in white fat but not in interscapular brown fat. J. Lipid Res. *48*, 41–51.

OPEN ACCESS
CellPress

Cell Reports
Report

TRIM28 Represses Transcription of Endogenous Retroviruses in Neural Progenitor Cells

Liana Fasching,[1] Adamandia Kapopoulou,[2] Rohit Sachdeva,[1] Rebecca Petri,[1] Marie E. Jönsson,[1] Christian Männe,[1] Priscilla Turelli,[2] Patric Jern,[3] Florence Cammas,[4] Didier Trono,[2] and Johan Jakobsson[1,*]

[1]Laboratory of Molecular Neurogenetics, Department of Experimental Medical Science, Wallenberg Neuroscience Center and Lund Stem Cell Center, Lund University, 221 84 Lund, Sweden

[2]School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland

[3]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, 751 23 Uppsala, Sweden

[4]IRCM, Institut de Recherche en Cancérologie de Montpellier, INSERM, U896, Université Montpellier; Institut Régional du Cancer Montpellier, Montpellier 34298, France

*Correspondence: johan.jakobsson@med.lu.se

http://dx.doi.org/10.1016/j.celrep.2014.12.004

SUMMARY

TRIM28 is a corepressor that mediates transcriptional silencing by establishing local heterochromatin. Here, we show that deletion of TRIM28 in neural progenitor cells (NPCs) results in high-level expression of two groups of endogenous retroviruses (ERVs): *IAP1* and *MMERVK10C*. We find that NPCs use TRIM28-mediated histone modifications to dynamically regulate transcription and silencing of ERVs, which is in contrast to other somatic cell types using DNA methylation. We also show that derepression of ERVs influences transcriptional dynamics in NPCs through the activation of nearby genes and the expression of long noncoding RNAs. These findings demonstrate a unique dynamic transcriptional regulation of ERVs in NPCs. Our results warrant future studies on the role of ERVs in the healthy and diseased brain.

INTRODUCTION

The mammalian brain is an extremely complex organ harboring more than a thousand different types of neurons that serve a wide variety of functions. How this complexity is achieved remains largely unknown. However, epigenetic mechanisms such as DNA methylation, histone modification, and noncoding RNAs are thought to be important in establishing a high diversity of gene expression from the same template, leading to a spatial pattern of transcription. How distinct transcriptional programs are established in different neuronal populations remains poorly understood, but one interesting recently proposed hypothesis suggests transposable elements (TEs) to be involved in this process (Muotri et al., 2007; Reilly et al., 2013). TEs are repetitive mobile genetic elements that were originally considered to be parasitic DNA without any function, popularly termed "junk DNA." Today, it is becoming increasingly clear that TEs can act as gene regulatory elements by serving as hubs for chromatin modifications and by acting as transcriptional start sites for noncoding RNAs. Consequently, TEs are very well suited to influence gene expression and may play an important role in controlling and fine-tuning gene networks in the brain (Jern and Coffin, 2008; Cowley and Oakey, 2013).

Retroviruses are found in most vertebrates and can transform their genetic material and integrate into the host genome as proviruses to produce new viruses. Occasionally, retroviruses infect germline cells allowing the integrated proviruses to be passed on to the offspring as an endogenous retrovirus (ERV). Around 8%–10% of the human and mouse genome are composed of this type of TE, and, despite up to millions of years since their integration in host germline, many ERVs contain sequences that can serve as transcriptional start sites or as *cis*-acting regulatory elements in the host genomes (Jern and Coffin, 2008). The large amount of ERVs in mammalian genomes suggest that they play important roles in the host organisms, for instance, by influencing gene regulatory networks (Kunarso et al., 2010; Feschotte, 2008), but ERVs have also been linked to diseases. In humans, aberrant expression of ERVs has been found in both cancer and brain disorders, although causality remains to be established (Jern and Coffin, 2008; Douville et al., 2011). Thus, ERVs may contribute both beneficial and detrimental effects, which have been balanced throughout evolution, to the host organism.

ERVs are silenced during the first few days of embryogenesis by TRIM28 (tripartite motif-containing protein 28, also known as KAP1 or TIF1beta), a transcriptional corepressor essential for early mouse development (Cammas et al., 2000; Rowe et al., 2010). During the extensive genome reprogramming that takes place at this period, TRIM28 is recruited to ERVs via sequence-specific Krüppel-associated box zinc-finger proteins (KRAB-ZFPs), a family of transcription factors that has undergone a rapid expansion in mammalian genomes in parallel with the expansion of ERVs (Wolf and Goff, 2009; Thomas and Schneider, 2011). TRIM28 then induces repressive histone modifications by recruiting multiprotein complexes including the H3K9 methyltransferase SETDB1 (also known as ESET), the histone deacetylase-containing NuRD complex, and heterochromatin protein 1 (HP1) (Schultz et al., 2002; Sripathy et al.,

2006). Deletion of *Trim28* or *Setdb1* in ESCs leads to loss of the H3K9me3-mark at ERVs, resulting in transcriptional activation of these elements (Matsui et al., 2010; Rowe et al., 2010).

However, KRAB-ZFP/TRIM28 histone-based repression of ERVs rapidly gives place to a more permanent silencing mechanism, as the TRIM28-mediated recruitment of de novo DNA methyltransferases leads to cytosine methylation at CpG dinucleotides (Ellis et al., 2007; Wiznerowicz et al., 2007; Rowe and Trono, 2011). The maintenance DNA methyltransferase complex then ensures that DNA methylation is maintained, alleviating the need for sequence-specific KRAB-ZFPs and TRIM28. In mouse embryonic fibroblasts as well as in all adult tissues examined so far, TRIM28 depletion has no impact on ERV silencing, which is instead released by drugs such as 5-azacytidine or by deletion of DNA methyltransferases (Jackson-Grusby et al., 2001; Hutnick et al., 2010).

DNA methylation has long been considered as a stable epigenetic mark resulting in maintenance of DNA-methylation patterns throughout the lifespan of an organism. However, several recent studies demonstrate a unique dynamic regulation of DNA-methylation patterns in the brain (Sweatt, 2013). There is also evidence that retroelements and transposons are highly active during brain development and in neural progenitor cells (NPCs) (Muotri et al., 2005, 2010; Baillie et al., 2011; Evrony et al., 2012; Li et al., 2013; Perrat et al., 2013). For example, *LINE-1* elements have been found to be transcriptionally active and to retrotranspose in NPCs (Muotri et al., 2005, 2010; Coufal et al., 2009). In addition, we have previously found that deletion of TRIM28 in postmitotic forebrain neurons results in complex behavioral alterations, including vulnerability to stress (Jakobsson et al., 2008). In the present work, we demonstrate that NPCs use TRIM28-mediated histone modifications to dynamically regulate the transcription and silencing of ERVs, rather than the DNA methylation at play in other somatic tissues. We also unveil that derepression of ERVs influences transcriptional dynamics in NPCs, by activating nearby genes and the expression of long noncoding RNAs (lncRNAs).

RESULTS

TRIM28-Deficient NPCs Express High Levels of ERVs

To investigate if TRIM28 contributes to ERV silencing in NPCs we established *Trim28*-deficient NPC cultures. We crossed transgenic *NestinCre* mice (Tronche et al., 1999) with mice carrying floxed *Trim28*-alleles ($Trim28^{fl/fl}$) (Weber et al., 2002), resulting in excision of *Trim28* in neural progenitors at the time when Nestin-expression is initiated, starting around embryonic day 10 (E10). At E13.5, we collected embryos, dissected the forebrain, and established NPC cultures from individual embryos (Figures 1A and 1B). We confirmed the deletion of *Trim28* by genotyping for the excised allele and by verifying the absence of TRIM28 protein (Figures 1C and 1D). We collected RNA from $Trim28^{-/-}$ NPCs and wild-type controls and performed RNA extraction followed by deep sequencing (RNA-seq). The resulting reads were mapped against reference sequences from Repbase, a database containing consensus sequences for known repetitive elements (Jurka et al., 2005). We found that several ERVs were highly upregulated in $Trim28^{-/-}$ NPCs, including, e.g., *Mus musculus ERV using* $tRNA^{Lys}$ *type 10C* (*MMERVK10C)* and *intracisternal A-particles class 1* (*IAP1*) (Figure 1E; Tables S1 and S2). Other retroelements such as *MusD* and *LINE-1* were modestly upregulated, whereas several other types of common repetitive elements were unaffected (Figure 1E; Tables S1 and S2).

We confirmed increased transcription of *MMERVK10C* and *IAP1* elements using quantitative RT-PCR (qRT-PCR) (Figure 1F). In contrast, when we used primer pairs designed to recognize the consensus sequence of the entire *IAP*-family, including more ancient *IAP* elements, we detected only a modest upregulation (Figure 1F). This finding is in line with the results of the RNA-seq, which indicated that only certain types of IAP elements were upregulated in $Trim28^{-/-}$ NPCs. Also in agreement with the RNA-seq, qRT-PCR analyses indicated that deletion of *Trim28* in NPCs only modestly increased the expression of other retroelements such as *LINE-1* or *MusD* (Figure 1F). We confirmed these results in cultures derived from two separate embryos (data not shown).

$Trim28^{-/-}$ NPCs proliferated at a similar rate compared to cells generated from wild-type and heterozygous siblings and could be expanded for more than 60 passages. However, we observed that $Trim28^{-/-}$ NPCs were growing in dense cluster-like formations, which seemed to attach less to the flask surface compared to the wild-type control. $Trim28^{-/-}$ NPCs could also be differentiated to both neurons and astrocytes suggesting that TRIM28 has no major influence on the self-renewal and differentiation of NPCs (Figures 1G and 1H).

MMERVK10C Elements Are Controlled by TRIM28

The RNA-seq analysis indicated that *MMERVK10C* elements were among the most upregulated ERVs following *Trim28*-deletion in NPCs. *MMERVK10C* is a beta-like ERV similar to *HERVK* (*HML2*), one of the most recent ERVs to invade the human genome (Reichmann et al., 2012) (Belshaw et al., 2005). *MMERVK10C* sequences flanked by *RLTR10C* make up putative proviral sequences of around 8.4 kb. In the mouse genome, *MMERVK10C* is present in a few complete provirus loci (~20) and more than 1,000 incomplete loci (Reichmann et al., 2012). We performed sequence analysis of the *MMERVK10C* provirus for the presence of retroviral features using the RetroTector software (Sperber et al., 2007). Based on this analysis, we designed primers recognizing the *LTRs*, *gag*, *pol*, and *env* of the *MMERVK10C* provirus and investigated expression levels in $Trim28^{-/-}$ NPCs (schematics in Figure 2A). We found that transcripts over the entire region of the provirus were increased, including a massive expression of *env* sequences when compared to wild-type controls (170-fold; Figure 2B).

Ascertaining that the ERV induction observed in NPCs isolated from $Trim28^{-/-}$ animals was not secondary to more general developmental anomalies, knocking down TRIM28 in wild-type NPCs by lentivector-mediated RNA interference led to a marked upregulation of these retroelements (Figure 2C). Furthermore, increased ERV expression was detected in forebrain tissue from $Trim28^{-/-}$ embryos (Figure 2D).

In ESCs, TRIM28 controls ERV expression via histone modifications including H3K9 trimethylation (Rowe et al., 2010), whereas it is DNA methylation that instead prevails in somatic

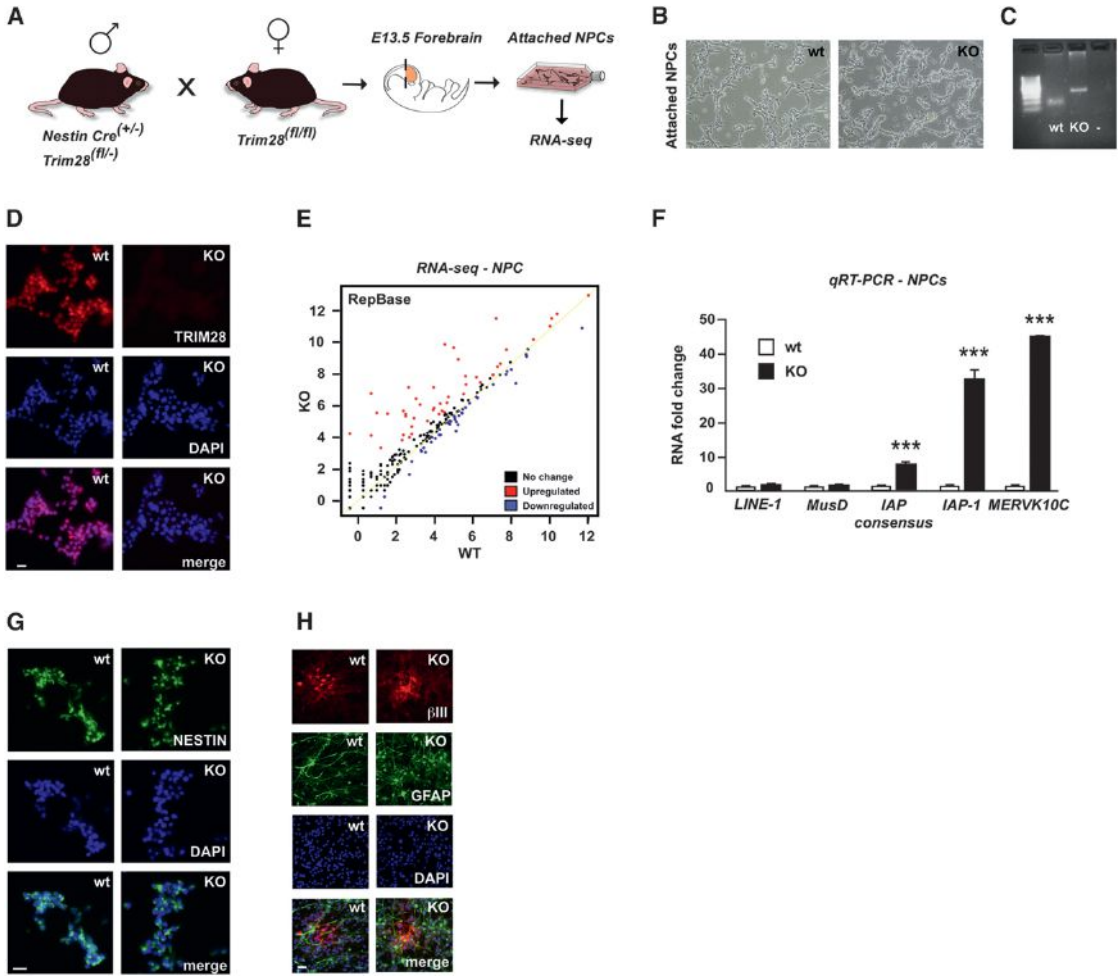


Figure 1. Establishment of *Trim28*-Deficient Neural Progenitor Cultures

(A) Illustration of the experimental approach.

(B) Representative images of early passage *Trim28*$^{-/-}$ NPCs.

(C) PCR analysis of genomic DNA from wild-type and *Trim28*$^{-/-}$ NPCs demonstrates the presence of the 152 and 290 bp products corresponding to *loxP*-flanked or excised *Trim28*, respectively.

(D) Verification of a complete lack of TRIM28 protein via immunocytochemistry.

(E) RNA-seq analysis. The graph shows KO samples plotted versus wild-type samples, where each dot represents a Repbase sequence.

(F) qRT-PCR of RNA isolated from wild-type and *Trim28*$^{-/-}$ NPCs.

(G) *Trim28*-deficient NPCs display a homogenous expression of NESTIN.

(H) Immunofluorescent analysis of differentiated NPCs.

Data are presented as mean of relative values ± SEM. **p < 0.01, ***p < 0.001, Student's t test. Scale bars, 200 (A) and 50 (B) μm. See also Tables S1 and S2.

tissues. In NPCs, we found that the *MMERVK10C* provirus was enriched in H3K9me3, and that this repressive mark was markedly reduced in *Trim28*$^{-/-}$ NPCs (Figure 2E).

Because *MMERVK10C* appeared to be under TRIM28 control in NPCs, we hypothesized that at least a proportion of these retroelements escaped DNA methylation in these cells. To probe this issue, we examined the DNA methylation status of full-length *MMERVK10C*, which were among the most highly upregulated retroelements in *Trim28*$^{-/-}$ NPCs. Bisulfite sequencing of a CpG-island located in the 3′ region of *MMERVK10C* revealed several clones with some unmethylated CpGs (17% unmethylated CpGs, Figure 2F) in NPCs, whereas this region was almost fully methylated in DNA extracted from mouse tail (7% unmethylated CpGs, Figure 2F, Fisher's exact test one-sided $p < 0.05$). Moreover, we found no difference in the level of CpG methylation between wild-type and *Trim28*$^{-/-}$ NPCs. In summary, these data suggest that a proportion of the *MMERVK10C* elements are spared from undergoing DNA methylation specifically in NPCs during early development.

Increased Expression of *IAP1* Results in ERV-Derived Protein Expression

IAP1 elements, which lose H3K9me3 marks and were also highly upregulated in *Trim28*$^{-/-}$ NPCs (Figures 3A and 3B), are

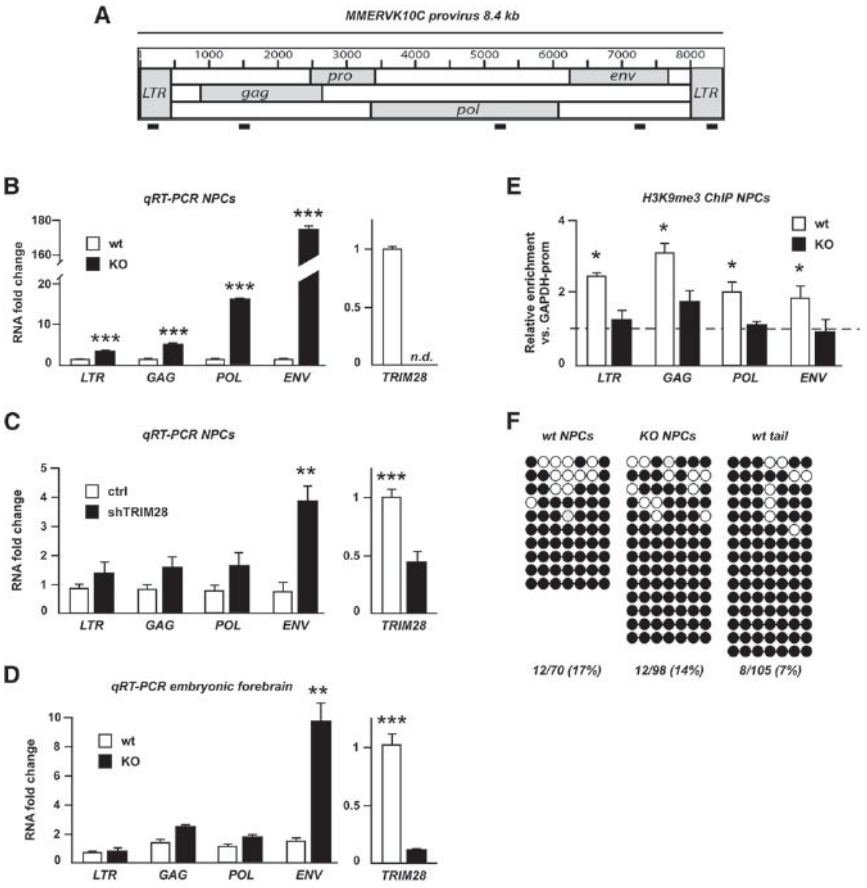


Figure 2. Analysis of the Putative *MMERVK10C* Provirus

(A) Schematic drawing of the *MMERVK10C* provirus and approximate primer positions.

(B) Quantitative analysis of transcript levels of different regions of the *MMERVK10C* provirus in *Trim28*$^{-/-}$ and wild-type NPCs.

(C) qRT-PCR analysis of *MMERVK10C* following TRIM28-shRNA knockdown.

(D) qRT-PCR analysis of E13.5 forebrain dissected from intercrosses of *NestinCre Trim28*^{floxed} mice.

(E) ChIP for H3K9me3 in *Trim28*$^{-/-}$ and wild-type NPCs.

(F) Bisulfite sequencing analysis of the 3′ end region of *MMERVK10C*. Empty and full circles represent unmethylated and methylated CpGs, respectively.

Data are presented as mean of relative values ± SEM. *p < 0.05, Student's t test.

internalized *env*-lacking mouse ERVs that demonstrate a large degree of polymorphism among different mouse strains and maintain the capacity to retrotranspose. Using immunocytochemistry with an IAP-specific antibody, we found a uniform, high-level IAP-gag expression located to the cytoplasm in *Trim28*$^{-/-}$ NPCs (Figure 3C).

Taken together, these data demonstrate that deletion of TRIM28 in NPCs results in a massive transcriptional increase of ERVs, including *MMERVK10C* and *IAP1*. NPCs thus appear to constitute a cellular environment distinct from that of other somatic cells studied so far, with the TRIM28-induced histone-based repressive mechanism playing a role in ERV control.

Activation of ERVs Correlates with Increased Transcription of Nearby Genes

The ability of ERVs to attract transcription factors and silencing complexes has led to a reassessment of their role in the host genome. ERVs are now considered to be important transcriptional regulatory elements that shape and influence gene expression during early development (Isbel and Whitelaw, 2012). For example, we have recently found that TRIM28 controls the expression of developmental genes by repressing ERV-associated enhancers in pluripotent cells (Rowe et al., 2013). Twenty-six *MMERVK10C* proviruses and 361 *IAP* proviruses that were upregulated in *Trim28*$^{-/-}$ NPCs were mapped to precise genomic locations (Figure S1). Out of these 387 proviruses, 90 were situated close to genes (<50 kb). We found that 25 of those genes (28%) demonstrated significantly increased expression, whereas expression of only six of them was decreased (7%) (Figure 4A). We also found that those 90 genes located close to upregulated ERVs (ERV-up genes) were on average 3-fold upregulated in *Trim28*$^{-/-}$ cells (Figure 4B). In contrast, a random selection of ERVs that was not upregulated in *Trim28*$^{-/-}$ cells (n = 129, *MMERVK10C* and *IAP1* elements) did not affect nearby genes (ERVs-ctrl genes, n = 50, Figure 4B). Interestingly, we also found that ERV-up genes were expressed at low levels in wild-type cells (Figure 4C), which is in agreement with a model where ERVs mediate repressive regulation of nearby genes caused by the attraction of the TRIM28 silencing complex to ERV sequences. We validated the increased expression of five ERV-up genes in *Trim28*$^{-/-}$ cells using qRT-PCR (Figure 4D).

ERVs Produce Long Noncoding RNAs

We looked in detail at *BC048671*, which is a protein-coding transcript that is induced in *Trim28*$^{-/-}$ NPCs but completely absent in wild-type NPCs. *BC048671* is located 5 kb downstream of an *IAP* element, which is also highly upregulated in *Trim28*$^{-/-}$ NPCs. The RNA-seq data show that transcriptional initiation at the *IAP* element results in the formation of a long transcript (>10 kb) that extends into the coding sequence of *BC048671* (Figure 4E). The presence of high levels of this long transcript was verified using qRT-PCR primers located both upstream and within the coding sequence of *BC048671* (Figure 4F). Thus, readthrough of an ERV-derived transcript into another locus is likely to be one of several mechanisms by which nearby gene expression can be affected (see also Figure S2). This finding supports the notion that a general feature of ERVs might

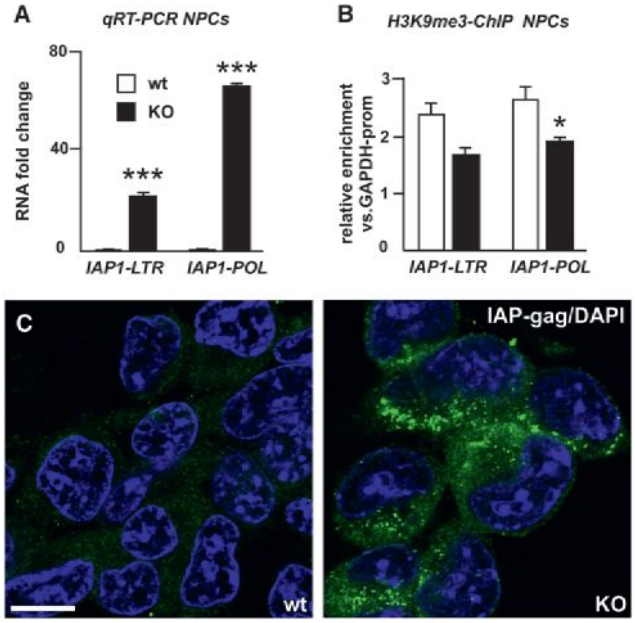


Figure 3. Analysis of *IAP1* Expression

(A) Quantitative analysis of transcript levels of different regions of *IAP1* provirus in *Trim28*$^{-/-}$ and wild-type NPCs.

(B) ChIP for H3K9me3 in *Trim28*$^{-/-}$ and wild-type NPCs.

(C) Confocal analysis of immunofluorescence staining for IAP-gag on *Trim28*$^{-/-}$ and wild-type NPCs. Scale bar, 10 μm.

Data are presented as mean of relative values ± SEM. *p < 0.05, Student's t test.

be to act as transcriptional start sites for long noncoding RNAs (lncRNAs). Indeed, when we scrutinized ERV elements located in gene free regions, we found that both *IAP* and *MMERVK10C* elements serve as start sites for lncRNAs (Figures 4G and 4I). Using qRT-PCR, we confirmed high-level expression of two ERV-derived lncRNAs in *Trim28*$^{-/-}$ NPCs (Figures 4H and 4J). The length of the ERV-derived lncRNAs did in many cases extend 25 kb (Figure 4K). These data demonstrate that derepression of ERVs in NPCs can result in the expression of multiple lncRNAs. The functional role of lncRNAs in NPCs remains largely unexplored, but they are thought to play important regulatory roles and have been implicated as scaffolds for nuclear protein complexes and as antisense transcripts in the control of epigenetic pathways (Guttman and Rinn, 2012).

DISCUSSION

In pluripotent stem cells, TRIM28 is a master corepressor of retroelements including ERVs (Matsui et al., 2010; Rowe et al., 2010). When these cells differentiate into various somatic cell types, DNA methylation is instated on ERV sequences, which ultimately results in stable silencing that is no longer dependent on TRIM28 (Wiznerowicz et al., 2007; Rowe et al., 2013). Thus, when TRIM28 is deleted from various somatic cell types such as fibroblasts, hepatocytes, and white blood cells, no increased ERV expression is detected (Rowe et al., 2010; Bojkowska et al., 2012; Santoni de Sio et al., 2012a, 2012b). Here, we describe an exception to this rule. When TRIM28 is deleted in NPCs, several ERVs become highly expressed. This finding unravels a unique transcriptional regulation of ERVs in NPCs.

ERVs regulated by TRIM28 in NPCs are recent invaders of the mouse genome. *IAP1* is the most recent member of the well-studied IAP ERVs (Qin et al., 2010). IAPs are ERVs that have lost the *env* gene and adopted an intracellular life cycle (Ribet et al., 2008). *IAP1* has been shown to retrotranspose and has distinct integration patterns in different strains of laboratory mice (Li et al., 2012). *MMERVK10C*, another ERV massively upregulated in *Trim28*$^{-/-}$ NPCs, is poorly characterized, and it is unclear if it is still endowed with retrotransposition potential, whether on its own or with the support of factors provided in *trans*. A previous study that analyzed the structure of *MMERVK10C* elements in the mouse genome found that the majority of these elements have 3′ deletions removing the start of the *gag* open reading frame as well as the major part of *env* (Reichmann et al., 2012). Our data demonstrate that, in NPCs, TRIM28 controls the rare copies of *env*-containing *MMERVK10C* elements, which are most likely to be the youngest ones, raising the possibility that these recent invaders of the mouse genome contain *cis*-acting genomic elements that allow them to escape DNA methylation in NPCs.

The classic view of repetitive mobile genetic elements as parasitic DNA without beneficial function to the host is challenged in many ways. There are a number of recent studies indicating that transposable elements (TEs) play important roles in establishing and rewiring gene networks (Kunarso et al., 2010; Chuong et al., 2013). TEs have been shown to act as enhancers, repressors, and alternative promoters. In addition, TEs can affect splicing patterns and produce peptides with important functional roles (Jern and Coffin, 2008). In this study, we demonstrate that activated ERVs can influence gene expression of nearby genes, such as *BC048671*, and serve as start sites for lncRNAs. Taken together, our findings indicate that ERVs participate in the control of gene networks in the brain.

We have previously demonstrated that deletion of *Trim28* in postmitotic forebrain neurons results in complex behavioral changes (Jakobsson et al., 2008). In addition, heterozygous germline deletion of *Trim28* has been described to result in abnormal behavioral phenotypes (Whitelaw et al., 2010). In this study, we found that deletion of *Trim28* during brain development is lethal (Figure S3). In addition, we also noted that heterozygous deletion of *Trim28* during brain development resulted in behavioral changes characterized by hyperactivity (Figure S3). Together, these findings demonstrate that disruption of TRIM28 levels in the mouse brain results in behavioral changes that are similar to impairments found in humans with certain psychiatric disorders. With this in mind, it is noteworthy that increased levels of ERV transcripts have been detected in patients with several neurological and psychiatric disorders (Jeong et al., 2010; Douville et al., 2011; Li et al., 2012; Karlsson et al., 2001). The significance of these findings has been questioned because the human genome does not appear to harbor ERVs with known retrotransposing capacity (Jern and Coffin, 2008). However, the increasing evidence that derepression of

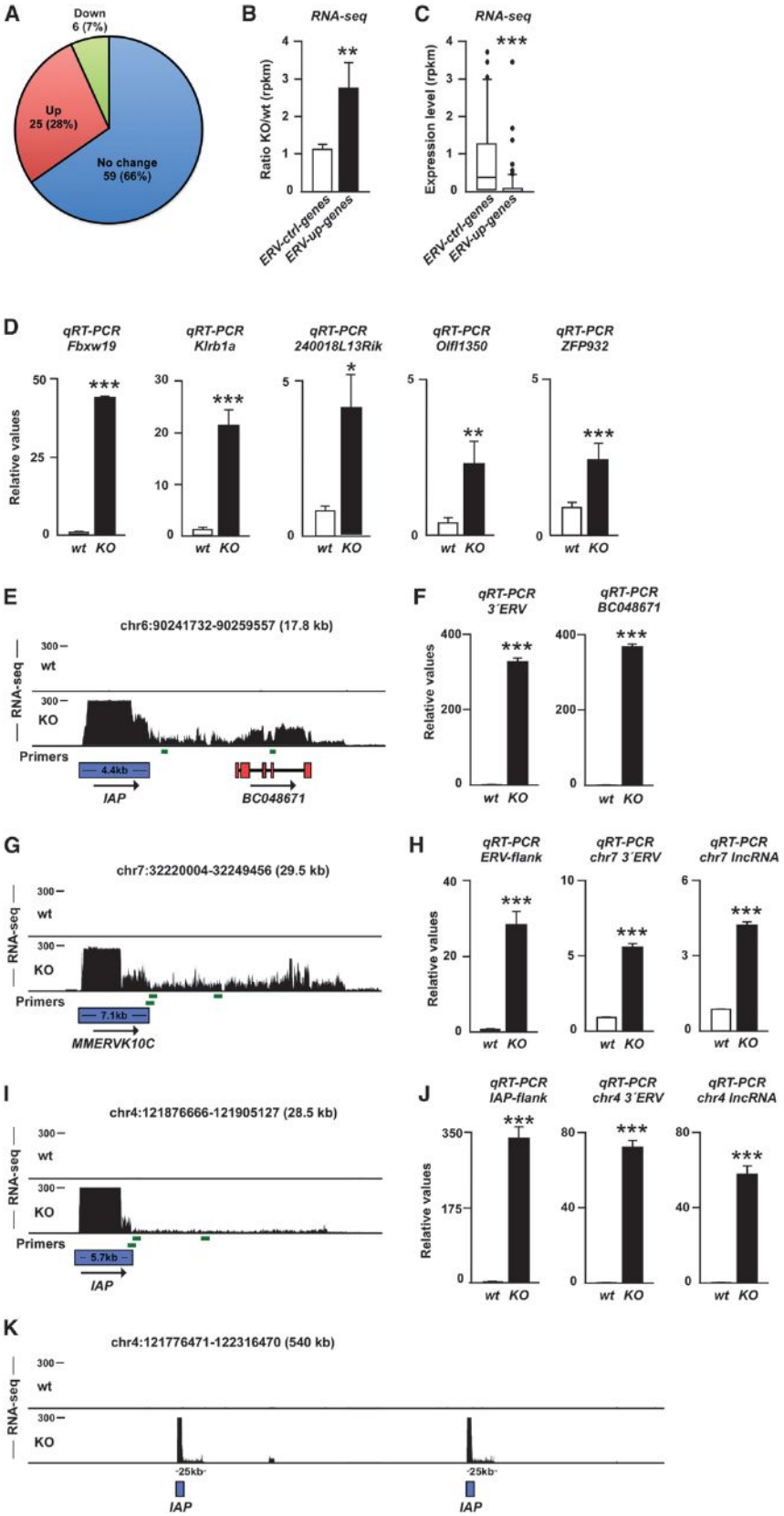


Figure 4. Activation of ERVs Influences Expression of Nearby Genes and Results in the Expression of lncRNAs

(A) Transcriptional change of genes located close (<50 kb) to ERVs in *Trim28*$^{-/-}$ NPCs.

(B) Mean transcriptional change of genes located to ERVs with increased transcription (ERV-up genes) and genes located close to unchanged ERVs (ERV-ctrl genes) in *Trim28*$^{-/-}$ NPCs.

(C) Absolute expression level of ERV-up genes and ERV-ctrl genes in wild-type NPCs.

(D) qRT-PCR of RNA isolated from wild-type and *Trim28*$^{-/-}$ NPCs.

(E) Screen shot from the USCS genome browser (mm9) showing induced transcription of *BC048671 in Trim28*$^{-/-}$ NPCs.

(G, I, and K) Activation of ERVs results in the expression of lncRNAs. Screen shot from the USCS genome browser (mm9).

(F, H, and J) qRT-PCR of RNA isolated from wild-type and *Trim28*$^{-/-}$ NPCs. Primers are indicated as green bars and include primers over the ERV junction as well as close and more distant from the 3′ end of the ERVs.

Data are presented as mean of relative values ± SEM. *p < 0.05, Student's t test. See also Figures S1 and S2.

ERVs influence gene networks, including the findings presented here, provides a potential mechanistic explanation for these observations.

In summary, our data suggest that ERVs may be involved in the regulation of gene expression in NPCs and may hereby offer a link between ERVs and brain disorders. It seems unlikely that behavioral phenotypes would arise from the derepression of a single ERV-induced gene. Instead, the presence of ERVs in multiple copies scattered throughout the genome allows for a powerful network-like control of gene expression, where dysregulation could result in widespread consequences. However, due to the large numbers of ERVs present in the mouse and human genome and their sequence variation, it is currently unfeasible to demonstrate a causal role for ERVs in controlling complex behavior or brain disorders using loss-of-function approaches, such as gene targeting and small hairpin RNA (shRNA) knockdown. Instead, improving our knowledge of critical host factors and networks controlling ERVs is essential to appreciate their impact on the genome and pathologies that may stem from their dysregulation. The demonstration that there is an ongoing dynamic TRIM28-mediated regulation of ERVs in NPCs is a step in this direction and warrants future studies of epigenetic and posttranscriptional regulation of ERVs in the healthy and diseased brain.

EXPERIMENTAL PROCEDURES

Detailed experimental procedures can be found in the Supplemental Experimental Procedures.

Procedures

Transgenic Animals

All animal-related procedures were approved by and conducted in accordance with the committee for use of laboratory animals at Lund University. NestinCre and floxed *Trim28* mice have been described previously (Weber et al., 2002; Tronche et al., 1999).

Cell Culture

NPC was established from embryonic day 13.5 (E13.5) forebrain and cultured as previously described (Conti et al., 2005).

Immunofluorescence

Immunofluorescence was performed as previously described (Thompson et al., 2005; Sachdeva et al., 2010).

RNA Studies

RNA-seq and qRT-PCR was performed as previously described (Rowe et al., 2010). The 50-base-paired end reads were mapped onto the RepBase version 16.08 (Jurka et al., 2005) and to the mouse genome (mm9) assembly. Mapping was done using the bowtie short read aligner (Langmead et al., 2009).

Chromatin Immunoprecipitation

Chromatin immunoprecipitation was performed with iDeal chromatin immunoprecipitation sequencing (ChIP-seq) kit (Diagenode) according to supplier's recommendations.

DNA-Methylation Analysis

Bisulfite sequencing was performed with the EpiTect bisulfite kit (QIAGEN) according to the supplier's recommendations. Sequence data were analyzed with the QUantification tool for Methylation Analysis (Kumaki et al., 2008).

Statistical Analysis

An unpaired t test was performed in order to test for statistical significance. Data are presented as mean ± SEM.

ACCESSION NUMBERS

The RNA-seq data were deposited in the NCBI Gene Expression Omnibus and are available under accession number GSE45930.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, three figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.12.004.

AUTHOR CONTRIBUTIONS

L.F., A.K., R.S., R.P., M.E.J., and C.M. designed and performed research and analyzed data. P.J., P.T., and D.T. designed research and analyzed data. F.C. contributed reagents. J.J. designed and coordinated the project and analyzed data. L.F. and J.J. wrote the paper, and all authors reviewed the manuscript.

ACKNOWLEDGMENTS

We are grateful to A. Björklund, S. Quenneville, and all members of the J.J. and Parmar laboratories for stimulating discussions. We thank U. Jarl, A. Josefsson, C. Isaksson, I. Nilsson, A.-K. Oldén, E. Ling, S. Smiljanic, M. Sparrenius, and E. Tjon for technical assistance. This study was supported by grants from Swedish Research Council (J.J.), Formas (P.J.), the Swedish Cancer Foundation (J.J.), the Lundqvist, Jeansson, and Crafoord foundations (J.J.), the Swedish Government Initiative for Strategic Research Areas *MultiPark* (J.J.), the French government, CNRS and INSERM (F.C.), and the French Agence Nationale pour la Recherche (F.C.).

Received: June 10, 2014
Revised: October 28, 2014
Accepted: December 1, 2014
Published: December 24, 2014

REFERENCES

Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P.M., Rizzu, P., Smith, S., Fell, M., et al. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. Nature *479*, 534–537.

Belshaw, R., Dawson, A.L., Woolven-Allen, J., Redding, J., Burt, A., and Tristem, M. (2005). Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. J. Virol. *79*, 12507–12514.

Bojkowska, K., Aloisio, F., Cassano, M., Kapopoulou, A., Santoni de Sio, F., Zangger, N., Offner, S., Cartoni, C., Thomas, C., Quenneville, S., et al. (2012). Liver-specific ablation of Krüppel-associated box-associated protein 1 in mice leads to male-predominant hepatosteatosis and development of liver adenoma. Hepatology *56*, 1279–1290.

Cammas, F., Mark, M., Dollé, P., Dierich, A., Chambon, P., and Losson, R. (2000). Mice lacking the transcriptional corepressor TIF1beta are defective in early postimplantation development. Development *127*, 2955–2963.

Chuong, E.B., Rumi, M.A., Soares, M.J., and Baker, J.C. (2013). Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat. Genet. *45*, 325–329.

Conti, L., Pollard, S.M., Gorba, T., Reitano, E., Toselli, M., Biella, G., Sun, Y., Sanzone, S., Ying, Q.L., Cattaneo, E., and Smith, A. (2005). Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. PLoS Biol. *3*, e283.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J.V., and Gage, F.H. (2009). L1 retrotransposition in human neural progenitor cells. Nature *460*, 1127–1131.

Cowley, M., and Oakey, R.J. (2013). Transposable elements re-wire and fine-tune the transcriptome. PLoS Genet. *9*, e1003234.

Douville, R., Liu, J., Rothstein, J., and Nath, A. (2011). Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis. Ann. Neurol. *69*, 141–151.

Ellis, J., Hotta, A., and Rastegar, M. (2007). Retrovirus silencing by an epigenetic TRIM. Cell *131*, 13–14.

Evrony, G.D., Cai, X., Lee, E., Hills, L.B., Elhosary, P.C., Lehmann, H.S., Parker, J.J., Atabay, K.D., Gilmore, E.C., Poduri, A., et al. (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell *151*, 483–496.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. Nat. Rev. Genet. *9*, 397–405.

Guttman, M., and Rinn, J.L. (2012). Modular regulatory principles of large non-coding RNAs. Nature *482*, 339–346.

Hutnick, L.K., Huang, X., Loo, T.C., Ma, Z., and Fan, G. (2010). Repression of retrotransposal elements in mouse embryonic stem cells is primarily mediated by a DNA methylation-independent mechanism. J. Biol. Chem. *285*, 21082–21091.

Isbel, L., and Whitelaw, E. (2012). Endogenous retroviruses in mammals: an emerging picture of how ERVs modify expression of adjacent genes. Bioessays *34*, 734–738.

Jackson-Grusby, L., Beard, C., Possemato, R., Tudor, M., Fambrough, D., Csankovszki, G., Dausman, J., Lee, P., Wilson, C., Lander, E., and Jaenisch, R. (2001). Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. Nat. Genet. *27*, 31–39.

Jakobsson, J., Cordero, M.I., Bisaz, R., Groner, A.C., Busskamp, V., Bensadoun, J.C., Cammas, F., Losson, R., Mansuy, I.M., Sandi, C., and Trono, D. (2008). KAP1-mediated epigenetic repression in the forebrain modulates behavioral vulnerability to stress. Neuron *60*, 818–831.

Jeong, B.H., Lee, Y.J., Carp, R.I., and Kim, Y.S. (2010). The prevalence of human endogenous retroviruses in cerebrospinal fluids from patients with sporadic Creutzfeldt-Jakob disease. J. Clin. Virol. *47*, 136–142.

Jern, P., and Coffin, J.M. (2008). Effects of retroviruses on host genome function. Annu. Rev. Genet. *42*, 709–732.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462–467.

Karlsson, H., Bachmann, S., Schröder, J., McArthur, J., Torrey, E.F., and Yolken, R.H. (2001). Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia. Proc. Natl. Acad. Sci. USA *98*, 4634–4639.

Kumaki, Y., Oda, M., and Okano, M. (2008). QUMA: quantification tool for methylation analysis. Nucleic Acids Res. *36* (Web Server issue), W170-5.

Kunarso, G., Chia, N.Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.S., Ng, H.H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat. Genet. *42*, 631–634.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Li, J., Akagi, K., Hu, Y., Trivett, A.L., Hlynialuk, C.J., Swing, D.A., Volfovsky, N., Morgan, T.C., Golubeva, Y., Stephens, R.M., et al. (2012). Mouse endogenous retroviruses can trigger premature transcriptional termination at a distance. Genome Res. *22*, 870–884.

Li, W., Prazak, L., Chatterjee, N., Grüninger, S., Krug, L., Theodorou, D., and Dubnau, J. (2013). Activation of transposable elements during aging and neuronal decline in Drosophila. Nat. Neurosci. *16*, 529–531.

Matsui, T., Leung, D., Miyashita, H., Maksakova, I.A., Miyachi, H., Kimura, H., Tachibana, M., Lorincz, M.C., and Shinkai, Y. (2010). Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. Nature *464*, 927–931.

Muotri, A.R., Chu, V.T., Marchetto, M.C., Deng, W., Moran, J.V., and Gage, F.H. (2005). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature *435*, 903–910.

Muotri, A.R., Marchetto, M.C., Coufal, N.G., and Gage, F.H. (2007). The necessary junk: new functions for transposable elements. Hum. Mol. Genet. *16 Spec No. 2*, R159–R167.

Muotri, A.R., Marchetto, M.C., Coufal, N.G., Oefner, R., Yeo, G., Nakashima, K., and Gage, F.H. (2010). L1 retrotransposition in neurons is modulated by MeCP2. Nature *468*, 443–446.

Perrat, P.N., DasGupta, S., Wang, J., Theurkauf, W., Weng, Z., Rosbash, M., and Waddell, S. (2013). Transposition-driven genomic heterogeneity in the Drosophila brain. Science *340*, 91–95.

Qin, C., Wang, Z., Shang, J., Bekkari, K., Liu, R., Pacchione, S., McNulty, K.A., Ng, A., Barnum, J.E., and Storer, R.D. (2010). Intracisternal A particle genes: Distribution in the mouse genome, active subtypes, and potential roles as species-specific mediators of susceptibility to cancer. Mol. Carcinog. *49*, 54–67.

Reichmann, J., Crichton, J.H., Madej, M.J., Taggart, M., Gautier, P., Garcia-Perez, J.L., Meehan, R.R., and Adams, I.R. (2012). Microarray analysis of LTR retrotransposon silencing identifies Hdac1 as a regulator of retrotransposon expression in mouse embryonic stem cells. PLoS Comput. Biol. *8*, e1002486.

Reilly, M.T., Faulkner, G.J., Dubnau, J., Ponomarev, I., and Gage, F.H. (2013). The role of transposable elements in health and diseases of the central nervous system. J. Neurosci. *33*, 17577–17586.

Ribet, D., Harper, F., Dupressoir, A., Dewannieux, M., Pierron, G., and Heidmann, T. (2008). An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus. Genome Res. *18*, 597–609.

Rowe, H.M., and Trono, D. (2011). Dynamic control of endogenous retroviruses during development. Virology *411*, 273–287.

Rowe, H.M., Jakobsson, J., Mesnard, D., Rougemont, J., Reynard, S., Aktas, T., Maillard, P.V., Layard-Liesching, H., Verp, S., Marquis, J., et al. (2010). KAP1 controls endogenous retroviruses in embryonic stem cells. Nature *463*, 237–240.

Rowe, H.M., Kapopoulou, A., Corsinotti, A., Fasching, L., Macfarlan, T.S., Tarabay, Y., Viville, S., Jakobsson, J., Pfaff, S.L., and Trono, D. (2013). TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. Genome Res. *23*, 452–461.

Sachdeva, R., Jönsson, M.E., Nelander, J., Kirkeby, A., Guibentif, C., Gentner, B., Naldini, L., Björklund, A., Parmar, M., and Jakobsson, J. (2010). Tracking differentiating neural progenitors in pluripotent cultures using microRNA-regulated lentiviral vectors. Proc. Natl. Acad. Sci. USA *107*, 11602–11607.

Santoni de Sio, F.R., Barde, I., Offner, S., Kapopoulou, A., Corsinotti, A., Bojkowska, K., Genolet, R., Thomas, J.H., Luescher, I.F., Pinschewer, D., et al. (2012a). KAP1 regulates gene networks controlling T-cell development and responsiveness. FASEB J. *26*, 4561–4575.

Santoni de Sio, F.R., Massacand, J., Barde, I., Offner, S., Corsinotti, A., Kapopoulou, A., Bojkowska, K., Dagklis, A., Fernandez, M., Ghia, P., et al. (2012b). KAP1 regulates gene networks controlling mouse B-lymphoid cell differentiation and function. Blood *119*, 4675–4685.

Schultz, D.C., Ayyanathan, K., Negorev, D., Maul, G.G., and Rauscher, F.J., 3rd. (2002). SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. Genes Dev. *16*, 919–932.

Sperber, G.O., Airola, T., Jern, P., and Blomberg, J. (2007). Automated recognition of retroviral sequences in genomic data—RetroTector. Nucleic Acids Res. *35*, 4964–4976.

Sripathy, S.P., Stevens, J., and Schultz, D.C. (2006). The KAP1 corepressor functions to coordinate the assembly of de novo HP1-demarcated microenvironments of heterochromatin required for KRAB zinc finger protein-mediated transcriptional repression. Mol. Cell. Biol. *26*, 8623–8638.

Sweatt, J.D. (2013). The emerging field of neuroepigenetics. Neuron *80*, 624–632.

Thomas, J.H., and Schneider, S. (2011). Coevolution of retroelements and tandem zinc finger genes. Genome Res. *21*, 1800–1812.

Thompson, L., Barraud, P., Andersson, E., Kirik, D., and Björklund, A. (2005). Identification of dopaminergic neurons of nigral and ventral tegmental area

subtypes in grafts of fetal ventral mesencephalon based on cell morphology, protein expression, and efferent projections. J. Neurosci. *25*, 6467–6477.

Tronche, F., Kellendonk, C., Kretz, O., Gass, P., Anlag, K., Orban, P.C., Bock, R., Klein, R., and Schütz, G. (1999). Disruption of the glucocorticoid receptor gene in the nervous system results in reduced anxiety. Nat. Genet. *23*, 99–103.

Weber, P., Cammas, F., Gerard, C., Metzger, D., Chambon, P., Losson, R., and Mark, M. (2002). Germ cell expression of the transcriptional co-repressor TIF1beta is required for the maintenance of spermatogenesis in the mouse. Development *129*, 2329–2337.

Whitelaw, N.C., Chong, S., Morgan, D.K., Nestor, C., Bruxner, T.J., Ashe, A., Lambley, E., Meehan, R., and Whitelaw, E. (2010). Reduced levels of two modifiers of epigenetic gene silencing, Dnmt3a and Trim28, cause increased phenotypic noise. Genome Biol. *11*, R111.

Wiznerowicz, M., Jakobsson, J., Szulc, J., Liao, S., Quazzola, A., Beermann, F., Aebischer, P., and Trono, D. (2007). The Kruppel-associated box repressor domain can trigger de novo promoter methylation during mouse early embryogenesis. J. Biol. Chem. *282*, 34535–34541.

Wolf, D., and Goff, S.P. (2009). Embryonic stem cells use ZFP809 to silence retroviral DNAs. Nature *458*, 1201–1204.

Cell Reports
Article

Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture

Matteo Vietri Rudan,[1] Christopher Barrington,[1] Stephen Henderson,[1] Christina Ernst,[2] Duncan T. Odom,[2] Amos Tanay,[3] and Suzana Hadjur[1,*]
[1]Research Department of Cancer Biology, Cancer Institute, University College London, 72 Huntley Street, London WC1E 6BT, UK
[2]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK
[3]Department of Computer Science and Applied Mathematics, Department of Biological Regulation, Weizmann Institute, Rehovot 76100, Israel
*Correspondence: s.hadjur@ucl.ac.uk
http://dx.doi.org/10.1016/j.celrep.2015.02.004

SUMMARY

Topological domains are key architectural building blocks of chromosomes, but their functional importance and evolutionary dynamics are not well defined. We performed comparative high-throughput chromosome conformation capture (Hi-C) in four mammals and characterized the conservation and divergence of chromosomal contact insulation and the resulting domain architectures within distantly related genomes. We show that the modular organization of chromosomes is robustly conserved in syntenic regions and that this is compatible with conservation of the binding landscape of the insulator protein CTCF. Specifically, conserved CTCF sites are co-localized with cohesin, are enriched at strong topological domain borders, and bind to DNA motifs with orientations that define the directionality of CTCF's long-range interactions. Conversely, divergent CTCF binding between species is correlated with divergence of internal domain structure, likely driven by local CTCF binding sequence changes, demonstrating how genome evolution can be linked to a continuous flux of local conformation changes. We also show that large-scale domains are reorganized during genome evolution as intact modules.

INTRODUCTION

The discovery of a topological-domain-like three-dimensional organization in metazoan chromosomes (Sexton et al., 2012; Dixon et al., 2012; Nora et al., 2012; Hou et al., 2012) is re-shaping our understanding of genome structure and function. This new layer of large-scale genome organization provides insights into the way by which sparsely embedded regulatory elements could interact to drive long-range transcriptional regulation. However, the extent by which the multi-scale domain architecture facilitates long-range regulation or is implied by it, as well as the precise mechanisms organizing chromosomes into domains, is not truly understood.

Currently, the best-characterized mechanism for domain organization involves long-range interactions between insulator proteins (CCCTC-binding factor [CTCF] in mammals) and the cohesin complex (Phillips-Cremins et al., 2013; Sofueva et al., 2013; Zuin et al., 2014). CTCF is a DNA-binding protein that engages its 11 zinc fingers to bind to DNA at a large, information-rich consensus motif (Kim et al., 2007). CTCF is a critical transcriptional regulator, originally described as a repressor of the myc oncogene (Filippova et al., 1996) and subsequently shown to function as an enhancer blocker and an insulator element (Bell et al., 1999). The insulator activity of CTCF depends on cohesin (Parelho et al., 2008; Wendt et al., 2008), an essential protein complex required for sister chromatid cohesion during mitosis (Michaelis et al., 1997; Guacci et al., 1997), which also functions in gene regulation (Rollins et al., 1999; Pauli et al., 2008). Together, CTCF and cohesin exert their effects on gene regulation primarily through the formation or stabilization of long-range chromatin loops (Hadjur et al., 2009; Mishiro et al., 2009; Nativio et al., 2009; Seitan et al., 2011). Such CTCF/cohesin-anchored loops are distributed throughout the genome, creating a network of long-range contacts spanning multiple scales, including not only loops that define the borders of strongly demarcated topological domains but also loops within such domains (Phillips-Cremins et al., 2013; Seitan et al., 2013; Sofueva et al., 2013; Zuin et al., 2014). While CTCF binding specificity depends to a large extent on specific DNA sequence elements, the specificity and directionality of CTCF/cohesin long-range contacts (Sofueva et al., 2013) and the way by which specific sites are assembled to define topological domains are not fully understood.

The dependency of CTCF recruitment on DNA sequence elements and the role for this insulator in mediating long-range chromosomal organization suggest that CTCF may function as a key link between genome sequence and the evolution of chromosomal domain organization. Indeed, some conservation of chromosomal domain structures has been reported between human and mouse through both linear epigenomic analysis (Yaffe et al., 2010) and high-throughput chromosome conformation capture (Hi-C) comparisons (Dixon et al., 2012). Moreover, a comparative analysis of CTCF binding in several mammalian

genomes suggests its evolutionary dynamics are context dependent, and conservation can be interrupted by mobile element activity (Schmidt et al., 2012). Despite these observations, a link between the evolutionary dynamics of CTCF binding and the evolution of chromosomal domain organization is yet to be explored.

Studies that have tracked the evolution of different transcription factor (TF) binding patterns have shown that sequence evolution alone is incapable of fully explaining the evolutionary dynamics of TF binding landscapes (Dermitzakis and Clark, 2001; Birney et al., 2007; Borneman et al., 2007; Schmidt et al., 2010). TF binding landscapes and large-scale chromosomal organization may function cooperatively to drive the evolution of genome regulation. These observations highlight the importance of multi-species comparative chromosomal structure analysis and its integration with insulator binding profiles across evolution. If the binding patterns of trans-factors such as CTCF are indeed strong drivers of domain organization, then their evolutionary dynamics should drive evolutionary conservation and divergence of chromosome domains.

With this in mind, we performed comparative Hi-C in non-cycling primary liver cells and analyzed the data together with CTCF binding profiles from the same species and tissue. Analysis of four mammalian Hi-C maps allowed us to explore how the evolution of CTCF binding profiles correlates, and in some cases likely drives, the evolution of chromosomal topologies. We find that the large-scale chromosomal domain structure is highly conserved between species, in a way that is correlated with the conservation of both the CTCF binding site and the orientation of its motif, resulting in directional long-range interactions that demarcate conserved domains. On the other hand, internal domain structure is observed to be more dynamic, and we discover remarkable correlation between evolutionary dynamics of CTCF sites and divergence of local insulation structure. Since the evolution of CTCF binding profiles is strongly driven at the nucleotide level within *cis* elements, our data suggest that internal domain structure can be modulated flexibly through local sequence evolution. Conversely, we show that interruption of large-scale domain structure is rare, and we suggest that instead of local sequence divergence, evolutionary manipulation of global chromosomal topologies is driven by processes involving duplications or rearrangements such as inversions, insertions/deletions, and translocations. We demonstrate this by charting cases of evolutionary domain shuffling in mouse and dog.

RESULTS

Sequence-Driven Evolution of CTCF Binding Profiles

CTCF binding is strongly correlated with the topological architecture of mammalian chromosomes and participates in long-range chromatin loops, thereby underlying global contact insulation. We analyzed mouse (*Mus musculus* [Mmus]), dog (*Canis familiaris* [Cfam]), and macaque (*Macaca mulatta* [Mmul]) CTCF chromatin immunoprecipitation sequencing (ChIP-seq) profiles from primary liver cells (Schmidt et al., 2012), aiming to define how conservation and divergence of the insulator binding landscape co-evolve with chromosomal topology. Pairwise CTCF ChIP-seq analysis identified conserved or divergent CTCF binding sites within syntenic chromosomal regions (Figures 1A, 1B, and S1). Sites with the strongest CTCF binding intensities were highly conserved (77% of the top 0.1 percentile), while lower-intensity CTCF binding sites were enriched for divergent binding (57% of mouse-divergent sites) (Figure 1B). We computed the sequence affinity of the different classes of binding sites to the canonical CTCF consensus motif in mouse and found that the levels of motif affinity for conserved sites were overall higher than the level for the mouse-divergent sites (Figure S1).

To understand the relationship between sequence affinity and CTCF binding at conserved or divergent sites, we correlated changes in CTCF binding with changes in CTCF sequence motif affinity among species. For this analysis, we used the canonical consensus motif from mouse that is the same in the other species (Schmidt et al., 2012). Remarkably, we found a direct association between sequence divergence and CTCF binding divergence. Conserved CTCF binding sites showed overall high motif affinities and a high degree of affinity conservation. Conversely, motifs underlying the divergent sites were evolutionarily dynamic and diverged in strong correlation to divergent binding intensity (Figures 1C and S1). The data show that when strong motifs in CTCF binding sites diverge, CTCF binding itself is concomitantly gained or lost. Interestingly, 65% of the sites that were conserved between mouse and dog were also conserved in macaque, while macaque-specific and dog-specific sites constituted another two populations of 775 and 891 sites, respectively, with weaker, more evolutionarily plastic motifs. Together, these data suggest that the CTCF insulator landscape is evolving under two regimes: the first involves a tight conservation of both sequence and binding landscape, and the second shows a dynamic interplay between divergence of specific *cis* elements and consequential evolution of the CTCF binding trait. The relatively direct influence of motif divergence on CTCF binding forms a potential link between sequence evolution and large-scale genome evolution.

CTCF Binding Site Evolution Is Correlated with the Mouse Hi-C Domain Structure

To investigate how the different classes of CTCF binding conservation correlate with chromosomal structure, we prepared Hi-C datasets on mouse liver cells (Figure S2). Filtering and normalization of the Hi-C ligation products was performed as before (Sofueva et al., 2013), revealing the characteristic chromosomal domain structure in these cells. Visualization of the CTCF occupancy groups with the Hi-C contact maps suggested that conserved CTCF binding sites were found at the borders of large-scale Hi-C domains while species-specific CTCF sites are located internal to domains (Figure 1D). This observation was supported with a genome-wide analysis whereby the relative position of conserved and divergent CTCF sites was determined with respect to all domains in the mouse genome (Figure 1E).

To further characterize chromosomal contacts around conserved and divergent CTCF sites, we analyzed the average contact distribution around these sites globally, measuring ''contact insulation'' by quantifying the decrease in contact probability between multiple elements separated by a CTCF site (Sofueva et al., 2013). Analysis of the composite contact insulation at

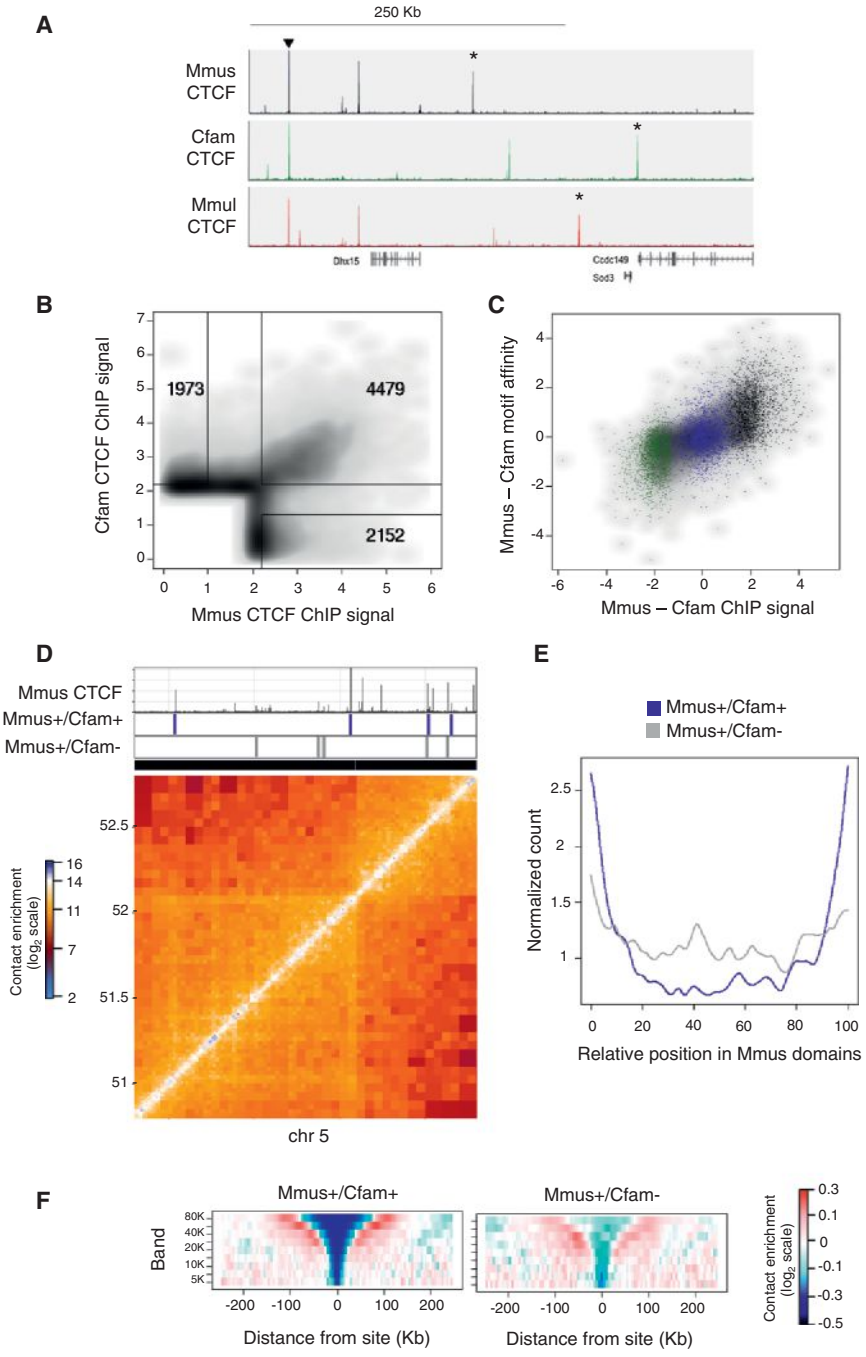


Figure 1. Evolution of CTCF Binding Correlates with Hi-C Domain Structure

(A) Representative CTCF ChIP-seq tracks on mouse (Mmus) chromosome 5; dog (Cfam) and macaque (Mmul) tracks are shown after liftOver to the mouse genome (mm10). Sites conserved across all three species (arrowhead) or specific to a single species (asterisks) are indicated as examples.

(B) A pairwise comparison of mouse CTCF ChIP and dog CTCF ChIP (liftOver track), identifying conserved or divergent binding sites (see Experimental Procedures).

(C) A comparison of the interspecies difference in CTCF ChIP signal against the difference in CTCF motif affinity in mouse versus dog. Scatterplots are highlighted for conserved (Mmus+/Cfam+, blue), mouse-divergent (Mmus+/Cfam−, black), and dog-divergent sites (Mmus−/Cfam+, green).

(D) A representative 2-Mb region from chromosome 5 of the mouse Hi-C contact maps. Also shown are the mouse (Mmus) CTCF ChIP track and conserved (Mmus+/Cfam+, blue) and divergent (Mmus+/Cfam−, gray) CTCF sites.

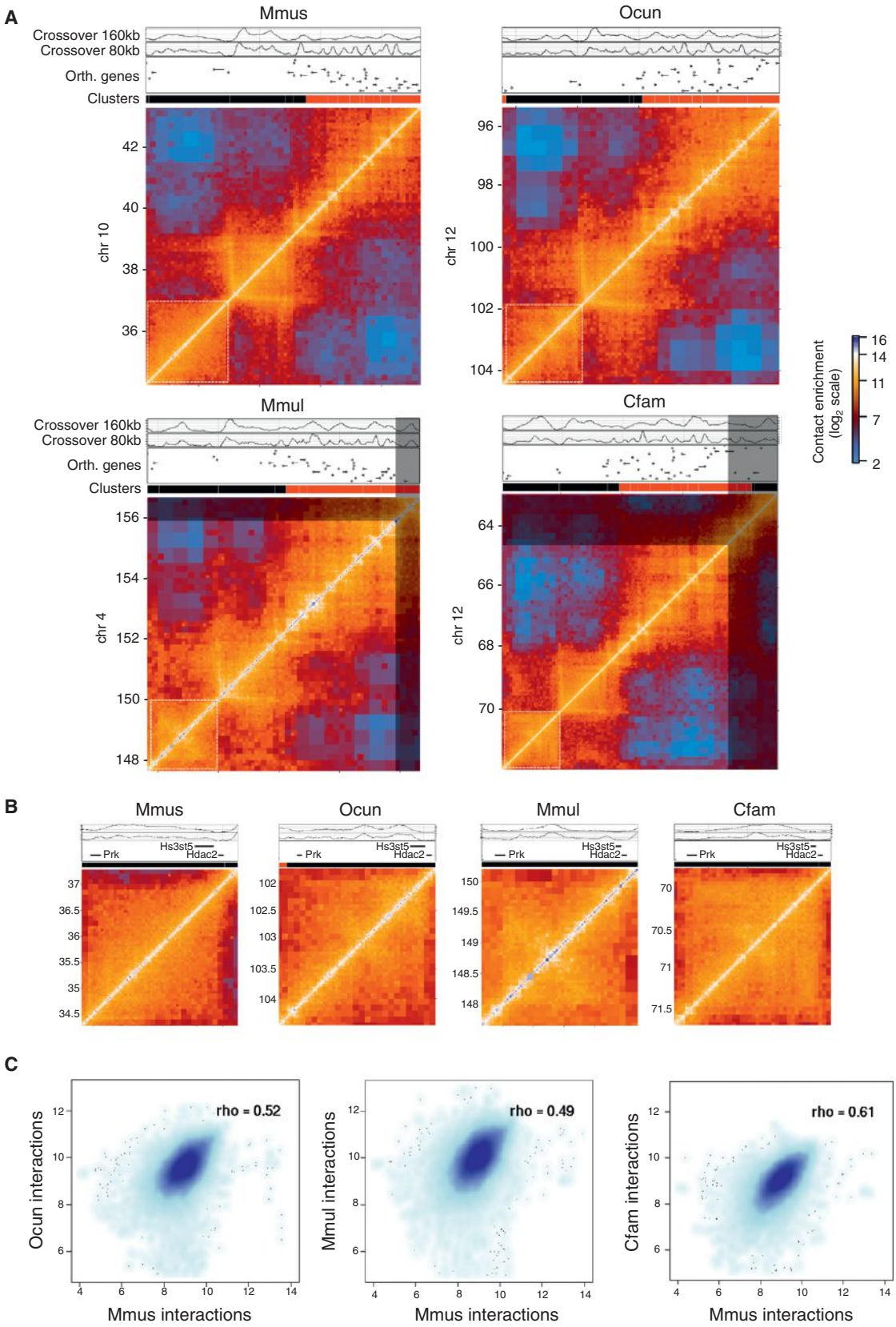
(E) Genome-wide relative position of conserved and divergent CTCF sites within mouse Hi-C domains.

(F) Contact insulation analysis for conserved (Mmus+/Cfam+) and mouse-divergent (Mmus+/Cfam−) CTCF sites in the mouse Hi-C data. See also Figure S1.

multiple distance ranges, indicated strong insulation profiles for conserved CTCF sites, further supporting the idea that these conserved, high-intensity CTCF sites were co-occurring with the borders of large-scale domain (reminiscent of topological chromosomal domains) (Figure 1F, left panel). In comparison, the lower-intensity mouse-divergent sites showed a significantly weaker, more localized insulation profile (Figure 1F, right panel). Similar trends were also observed when classifying CTCF sites according to their conservation in macaque (Figure S1). In summary, we found strong correlation between the evolutionary dynamics of CTCF binding sites and mouse chromosome topology, indicating the possibility of a direct link between insulator site divergence and the evolution of topological domain structure.

Comparative Hi-C Reveals the Evolution of Chromosome Topologies

We used comparative Hi-C to examine conservation and divergence of chromosome topology and to test how evolution of CTCF binding sites might underpin this. We collected liver cells from macaque, rabbit (*Oryctolagus cuniculus* [Ocun]), and dog and processed them using the same approach applied to mouse, yielding chromosomal contact maps for each of these species (Figures S2 and S3). Evaluation of the overall topological structure within the three newly profiled species first indicated the integrity of their reference genome structures and generated a resource for future refinement of such assemblies. More importantly, the data showed that the chromosome topologies in macaque, dog, and rabbit are characterized by a chromosomal domain structure that is similar to the one inferred before for human and mouse (Dixon et al., 2012). For example, comparison of a 9-Mb syntenic region highlighted the extensive conservation of chromosomal structure



(legend on next page)

across all species (Figure 2A). The maps also revealed evidence of intra-domain differences between species (Figure 2B). We quantified the extent of structural conservation genome-wide using a computational approach that allowed us to comprehensively describe domain structure at multiple scales. This pairwise approach revealed extensive genome-wide interspecies conservation of chromosome structure (Figures 2C and S3). A systematic analysis of paired domains in mouse and dog revealed that conserved domains are smaller in size compared to other domains and are classified as both active and passive clusters (Figure S4). Together, these data facilitated extensive analysis of the evolution of chromosomal topologies within regions that did not go through substantial genome rearrangement, allowing examination of the evolution of both large-scale domain borders and the insulation structure within domains.

Divergent CTCF Binding Drives Local Structural Change within Domains

Hi-C maps from liver cells of different species allowed us to ask how the evolutionary dynamics of CTCF correlate with conservation or divergence of domain structure. Analysis of specific loci showed that conserved CTCF sites were typically located at the borders of large-scale chromosomal domains that were themselves conserved between mouse and dog (Figure 3A). To test these observations globally, we computed the contact insulation profiles from either the mouse or dog Hi-C maps around conserved CTCF sites, showing that these sites indeed globally served as conserved insulation points (Figure 3B). Similar results were derived using a comparison of mouse and macaque (Figure S5). We also observed that conserved CTCF sites were strongly enriched for Rad21 in mouse (79% of conserved sites compared to 51% mouse-divergent sites co-localize with Rad21) and that CTCF/cohesin co-occupied sites exhibited strong contact insulation in all three species (data not shown).

In contrast to these highly stable sites, our data showed that divergent CTCF sites were located primarily within domains and exhibited local contact insulation. Comparative analysis of contact insulation at divergent CTCF sites revealed that indeed these sites correlated with divergent contact insulation profiles. For example, dog-divergent CTCF sites (Mmus−/Cfam+) exhibited local contact insulation specifically in the dog genome, whereas these same sites exhibited background levels of contact insulation when examined in the mouse Hi-C data (Figures 3C and S5). Importantly, the change in insulation following CTCF binding site evolution was stronger at the local (20-kb) scale, but was not significant at the higher (80-kb) scale (Figure 3D), suggesting that large-scale domain changes either are not affected by CTCF evolution or are under strong negative selection and are therefore not observed. These observations were further strengthened when we examined CTCF binding sites that were "partially" conserved. CTCF sites that were bound in mouse and dog, but lost in macaque, were associated with reduced contact insulation in the macaque genome. Thus, the data demonstrate a relationship between CTCF binding divergence and divergence of local insulation structure and therefore point to a role for CTCF in driving structural change in the genome.

The continuous evolutionary dynamics of intra-domain looping can play a key role in tuning promoter-enhancer contacts within domains. Consistent with this, we observe long-range contacts between divergent CTCF sites and enhancers or transcription start sites (TSSs) (Figure S6). Furthermore, analysis of transcription data from mouse, dog, and macaque liver reveals that divergent CTCF sites are contacting differentially expressed genes with a greater frequency than non-differentially expressed genes (Kolmogorov-Smirnov test, $p < 0.05$) (Figure S6). Together, these data support the hypothesis that emergence of divergent CTCF binding sites can contribute to changes in gene expression.

Conserved CTCF Sites Are Directional and Interact with Other Conserved Sites

While it is known that CTCF binding specificity greatly depends on its specific DNA consensus sequence, the specificity and directionality of CTCF/cohesin long-range contacts (Sofueva et al., 2013) and the way by which specific sites are assembled to define topological domains are not fully understood. As our data indicated that conserved CTCF binding sites have conserved motif affinities (Figure 1C), and because it is known that the CTCF consensus motif is non-symmetric, we asked whether conserved sites could also be conserved for the *orientation* of the CTCF motif. Indeed, 94% (3,265/3,483) of CTCF binding sites that are conserved between mouse and dog are also conserved in their orientation. To explore this further, we profiled contact insulation around conserved CTCF binding sites grouped according to the strand that the consensus motif was found on. We observed an asymmetric insulation behavior that was mirrored when the orientation of the motif was reversed (Figure 4A). This analysis uncoupled "insulation" (blue) from "preferential contacts" (red) and revealed that preferential contacts are made on one side of the oriented CTCF binding site, indicating that the orientation of the motif likely contributes to directionality

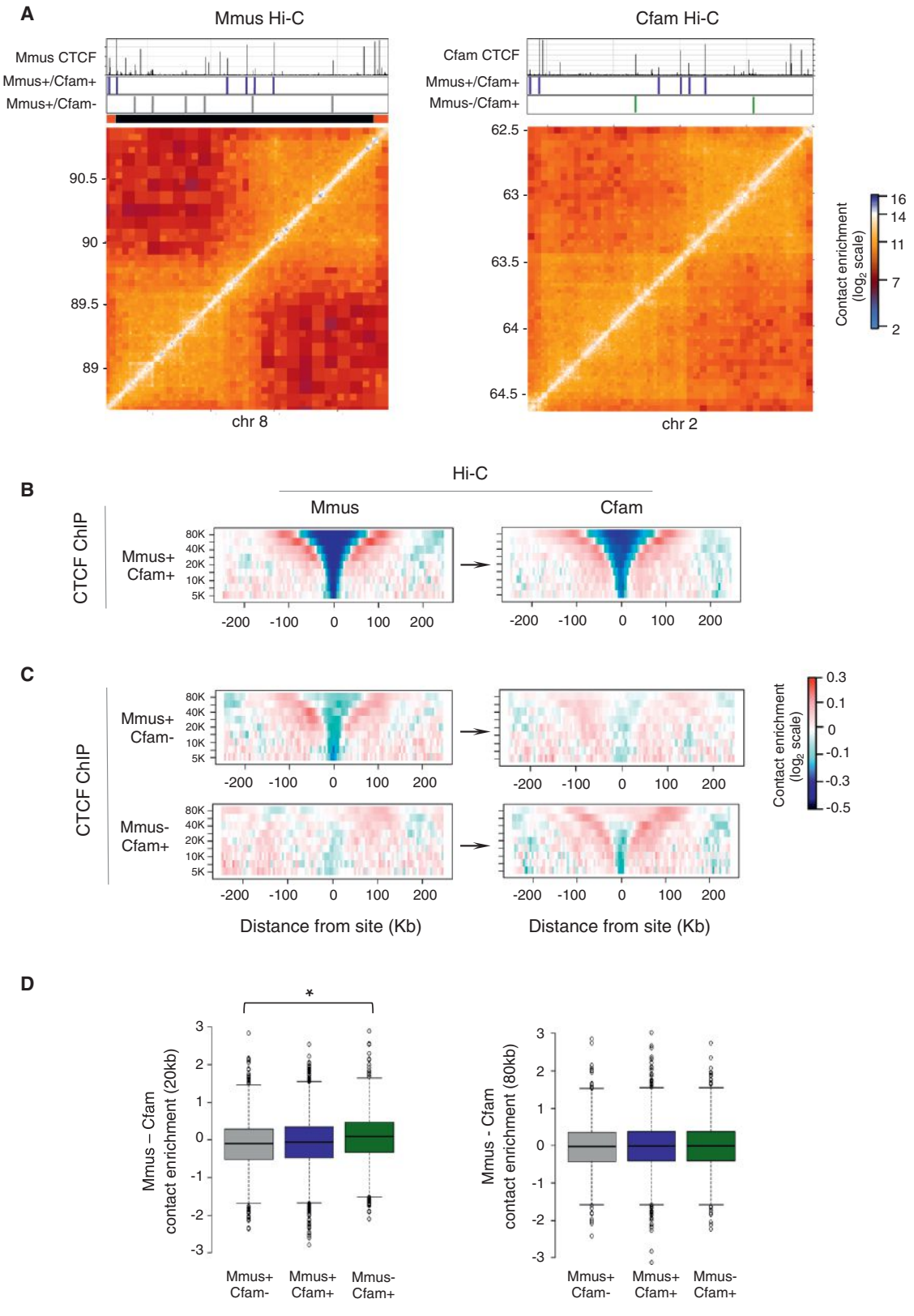
Figure 2. Chromosomal Domain Structure Is Robustly Conserved in Mammals

(A) Representative Hi-C contact maps of a 9-Mb syntenic region from mouse (Mmus), rabbit (Ocun), macaque (Mmul), and dog (Cfam). Maps are colored according to technically corrected contact enrichment. For scale purposes, the same size region is shown for all species, but in the macaque and dog genome, the syntenic region is reduced in size. The portion of the map that is outside of the synteny boundary is shaded out. Shown above each map is the quantification of contacts for distance bands of 160–480 kb (crossover 160 kb) and 80–240 kb (crossover 80 kb), as well as a track of orthologous genes, highlighting their conserved distribution in chromosomal domains.

(B) A zoom in of the domain highlighted with a white dashed box in (A) reveals differences in its internal organization across species.

(C) Global quantification of the correlation between the maps. Genome-wide contact enrichments between elements separated by 160–480 kb in rabbit, macaque, and dog were compared to the mouse genome after liftOver. Spearman correlation values are shown inside the plots. Axes units are contact enrichments [$\log_2$(observed/expected)].

See also Figures S2 and S3.



(legend on next page)

of CTCFs long-range interactions. Consistent with this, we profiled the genome-wide relative position within chromosomal domains of Mmus+/Cfam+ conserved CTCF sites grouped according to the orientation of their binding motif as above. We observed that the conserved CTCF binding sites that are enriched at the edges of conserved domains (Figure 1E) have a specific orientation of their motif relative to chromosomal domains (Figure 4B). These observations were replicated when we compared mouse and macaque.

To characterize the contact relationship between evolutionarily stable or flexible CTCF sites and to further understand how they contribute to the evolution of chromosome domain structure, we performed a high-resolution high-throughput circular chromosome conformation capture (4C-seq) study. We designed four 4C-seq viewpoints to a series of neighboring conserved CTCF binding sites bordering conserved domains in the mouse and dog as well as to a mouse-specific site. The results showed that each conserved CTCF site engages in very strong and directional interactions with neighboring conserved CTCF sites (Figure 4C). Remarkably, the specific interactions mediated by conserved sites in the mouse genome were themselves precisely conserved in the dog genome (Figure 4D) and define the underlying domain structure. In each case, the long-range interaction was anchored by a pair of conserved CTCF sites whereby one CTCF site had an orientation on the "+" strand and the other on the "−" strand and could provide the basis for the observed directionality of CTCF-mediated interactions. Moreover, a viewpoint designed to a mouse-divergent site exhibited weak interactions within the mouse domain, analogous to the local insulation behavior observed in Figure 3B (Figure 4C). Importantly, the mouse-divergent viewpoint had no prominent interactions in the dog genome, confirming the specificity of its interaction network.

Global analysis of Hi-C contacts between pairs of CTCF binding sites stratified according to genomic distances in *cis* (Sofueva et al., 2013) confirmed the 4C-seq observation systematically (Figure 4E). Consistent with the high-resolution 4C-seq profiles, Hi-C trends showed that conserved CTCF sites strongly contacted one another within the same domain. Divergent CTCF sites engaged in significantly weaker contacts with other divergent sites, even when stratifying thoroughly for genomic distances. Importantly, little to no contact was observed in the mouse genome between dog-divergent sites. These results show that evolutionarily stable CTCF sites are engaged in strong contacts with one another and suggest that in so doing, they create an interaction network that may support the conservation of domain structure. On the other hand, divergent CTCF sites are involved in weaker interactions, perhaps reflecting the evolutionary flexibility of the binding sites themselves.

Domains Maintain Their Integrity during Chromosomal Rearrangements

Our data suggested that large-scale domain re-organization does not typically occur following insulator divergence. How then can it still be observed? Our interspecies comparative Hi-C data allowed us to ask what happens to the integrity of conserved chromosomal domains when genomes are challenged by structural rearrangements. If chromosomal domains act as modular units (e.g., to regulate gene expression), then large-scale rearrangements would be expected to occur at domain borders, so as to maintain the integrity of these structures. We scanned the mouse and dog genomes for differences in the distance between contiguous orthologous genes in the two species. Our analysis uncovered a number of complex rearrangements between the mouse and dog genomes involving insertions, inversions, and duplications. In each case, we discovered that the rearrangement occurred at the border between two chromosomal domains. This is exemplified in the Hi-C map from chromosome 15 in dog (Figure 5). Here, we found two domains, one containing the *Slc5a9* gene and the other containing the *Trabd2b* gene (highlighted by red dots). Comparison of this region to the mouse genome revealed that a 2-Mb insertion occurred in the mouse genome that contains the *Skint* gene cluster, which is rapidly evolving and unique to the mouse lineage (Boyden et al., 2008). Remarkably, the insertion occurred directly between two neighboring dog domains in such a way as to perfectly maintain their integrity. A similar rearrangement event occurred at the *Mrgpr* gene cluster in the mouse genome (Dong et al., 2001), again preserving the structure of the neighboring domains (Figure S7). In another example, we observed a large-scale 5.5-Mb insertion in the dog genome containing multiple domains, and again, the domains on either side of the insertion have been maintained intact (Figure S8). These examples suggest that domains function as modular units and are selected against breakage during genome rearrangements.

DISCUSSION

In this study, we examined Hi-C contact maps and CTCF binding profiles from four mammalian species to understand the relationship between the evolution of CTCF binding sites and chromosome structures. Our data reveal that CTCF binding sites have

Figure 3. Conserved and Divergent CTCF Sites Show Differential Contact Insulation Behavior

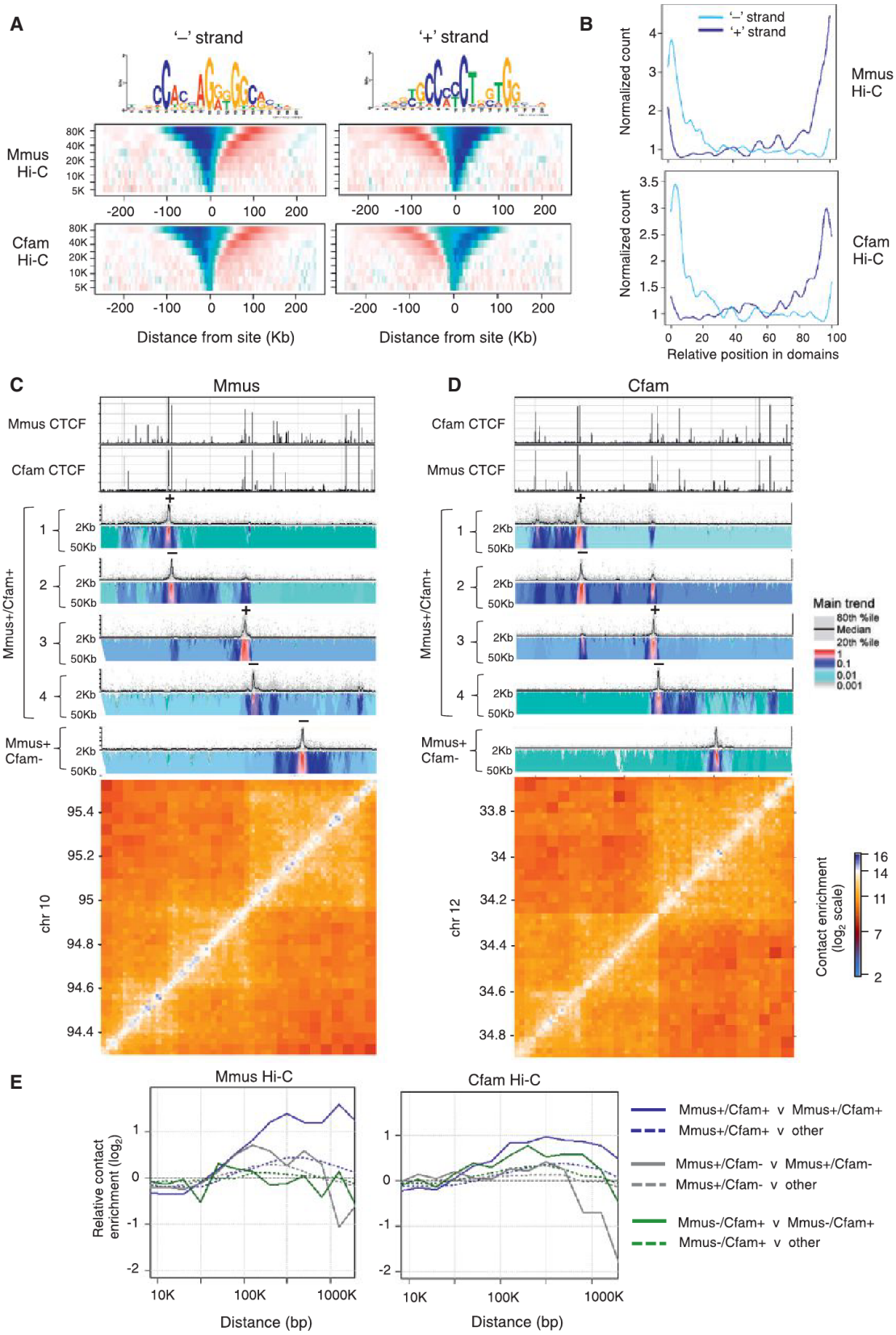
(A) Representative Hi-C contact maps for a 2-Mb syntenic region in mouse (left panel) and dog (right panel). Also shown are the CTCF ChIP-seq tracks in each species as well as the conserved (Mmus+/Cfam+, blue) and divergent (Mmus+/Cfam−, gray or Mmus−/Cfam+, green) CTCF sites.

(B) Average contact insulation analysis at conserved (Mmus+/Cfam+) CTCF sites in mouse (leftmost panel) and dog (rightmost panel) Hi-C datasets. Arrows indicate liftOver of sites.

(C) Same as for (B), but for mouse-divergent (Mmus+/Cfam−) and dog-divergent (Mmus−/Cfam+) sites in both genomes. Divergent sites appear to mediate weaker, shorter-range insulation that disappears in the species where the site is not bound.

(D) Distribution of the difference in contact insulation between mouse and dog at conserved or divergent CTCF sites at 20–60 kb (left panel, 20-kb band) or 80–240 kb (right panel, 80-kb band) scales. Divergent sites exhibit a significant (Kolmogorov-Smirnov test, $p < 1 \times 10^{-12}$, marked by an asterisk) shift in their distributions in the 20-kb band, but not in the 80-kb band, compatible with them mediating insulation at a local scale, but not at higher ranges.

See also Figure S5.



(legend on next page)

evolved under two regimes, whereby some CTCF elements are constrained both at the level of DNA sequence as well as in their binding while other CTCF elements exhibit significantly more flexibility. While both groups can mediate contact insulation, conserved CTCF elements are enriched at large-scale domain borders that tend to be themselves conserved. Meanwhile, evolutionarily flexible CTCF sites tend to be located internal of large-scale domains and mediate local structural change uniquely in that lineage. Our data thereby point to a strong correlation between the evolution of CTCF binding and chromosomal structure and extend on our current understanding of context-dependent CTCF binding sites and their specific roles in chromosomal domain architecture (Dixon et al., 2012; Dowen et al., 2014). Importantly, since CTCF binding information is encoded in high specificity *cis* elements, the intra-domain insulator dynamics we observe directly link local sequence evolution with chromosomal architectures. This direct linkage has strong implications for the study of CTCF and genome function and for our understanding of the evolutionary dynamics in complex genomes.

A central causal role for CTCF/cohesin in establishing domain structure is widely hypothesized, but direct experimental evidence has proven difficult to attain. Previous studies have observed a correlation between insulator binding and domain borders (Sexton et al., 2012; Dixon et al., 2012; Nora et al., 2012; Hou et al., 2012), and knockout experiments have suggested a quantitative link between loss of chromosomal looping structure and loss of the CTCF/cohesin binding landscapes (Sofueva et al., 2013; Zuin et al., 2014; Seitan et al., 2013). Given the pervasive impact of CTCF/cohesin on nuclear organization and gene regulation, it is difficult to identify the mechanisms of their action through classic genetic perturbation. Instead, the evolutionary comparison used here offers us thousands of naturally occurring genomic perturbations that can be identified and characterized at both the sequence and chromosomal topology levels. This strategy has yielded strong evidence of a direct link between the gain/loss of CTCF binding sites and a corresponding gain/loss of local domain insulation. Our comparative Hi-C analysis therefore strongly supports the idea that CTCF is causally connected to chromosomal looping structures.

The comparative chromosomal domain analysis described here has revealed a spectrum of evolutionary consequences, ranging from the conservation of essential large-scale chromosomal domains to the flexibility of continuous genomic adaptation. CTCF and cohesin complexes are deeply evolutionarily conserved, and the data here show that their role in mediating chromosome topologies and, even more remarkably, the large-scale building blocks of such topologies are also highly conserved. Our data suggest that the orientation of the CTCF motif may underlie the observed directionality of CTCF/cohesin-mediated long-range contacts and provide a rationale by which specific sites are assembled to define topological domains. Given that CTCF binding is strongly influenced by its consensus sequence, our data suggest that the assembly of domain structure is "hardwired" in the genome. This also has implications for further understanding the nature of the relationship between CTCF and cohesin, since biochemical studies have revealed that cohesin subunits interact with CTCF primarily through its C-terminal tail (Xiao et al., 2011), placing cohesin on a particular side of the chromosomal domain.

Interestingly, while we were able to observe cases whereby local sequence evolution perturbed CTCF binding and disrupted chromosomal looping, the structures that were affected due to this insulator divergence were primarily local loops. Cases of large-scale topological domains that were split or fused due to insulator divergence were not observed. We hypothesize that this stability is achieved by a combination of both local purifying selection on key CTCF binding sites and by buffering of major topological loops by additional factors. Strikingly, the cases of large-scale domain divergence that we were able to characterize were all linked with evolutionary genome rearrangements and revealed a mechanism that can reshuffle whole domains such that the rearranged chromosomal modules are aligned with existing domain borders. It is still, however, formally possible that rearrangements take place between CTCF sites that are mediating strong interactions.

In addition to the importance of topological domain and insulator conservation described here, the evolutionary dynamics that couple intra-domain CTCF divergence with changes in the local domain structure emerge as potentially fundamental for genome regulation. Loops contained within domains link enhancers (and their bound trans-factors) to target gene promoters. While it is still unclear how such targeting is regulated and how evolution can manipulate it, based on our data, we hypothesize that flexible CTCF binding sites within domains can influence looping from the promoter or enhancer as well by

Figure 4. Conserved CTCF Sites Engage in Strong, Directional Interactions with Other Conserved Sites

(A) Average contact insulation analysis in both Mmus and Cfam genomes for conserved (Mmus+/Cfam+) CTCF binding sites grouped according to the orientation of their binding motif. The consensus motif shown was generated from mouse binding sites.

(B) Genome-wide relative position within chromosomal domains of Mmus+/Cfam+ conserved CTCF sites grouped according to the orientation of their binding motif as above.

(C and D) (C) 4C-seq analysis and Hi-C maps of a 1.2-Mb syntenic region in the mouse and (D) dog genomes. Shown are 4C-seq viewpoints designed to four conserved (Mmus+/Cfam+) CTCF binding sites proximal to Hi-C domain borders and one 4C-seq viewpoint located at a mouse-divergent (Mmus+/Cfam−) CTCF site. The symbol above each 4C-seq bait indicates the strand (and orientation) of each viewpoint. Each 4C-seq experiment is represented by the median normalized 4C-seq coverage in a sliding window of 5 kb (top) and a multi-scale domainogram indicating normalized mean coverage in windows ranging between 2 kb and 50 kb.

(E) Relative intra-domain contact enrichment between CTCF sites (solid lines) as a function of distance when the two sites are <5 kb away from a CTCF binding site (solid lines) or where only one site is <5 kb away from a CTCF site (dotted lines). Shown are the relative contact enrichments within repressive domains for conserved (Mmus+/Cfam+, blue), mouse-divergent (Mmus+/Cfam−, gray) and dog-divergent (Mmus−/Cfam+, green) CTCF sites in mouse Hi-C (left) and dog Hi-C (right).

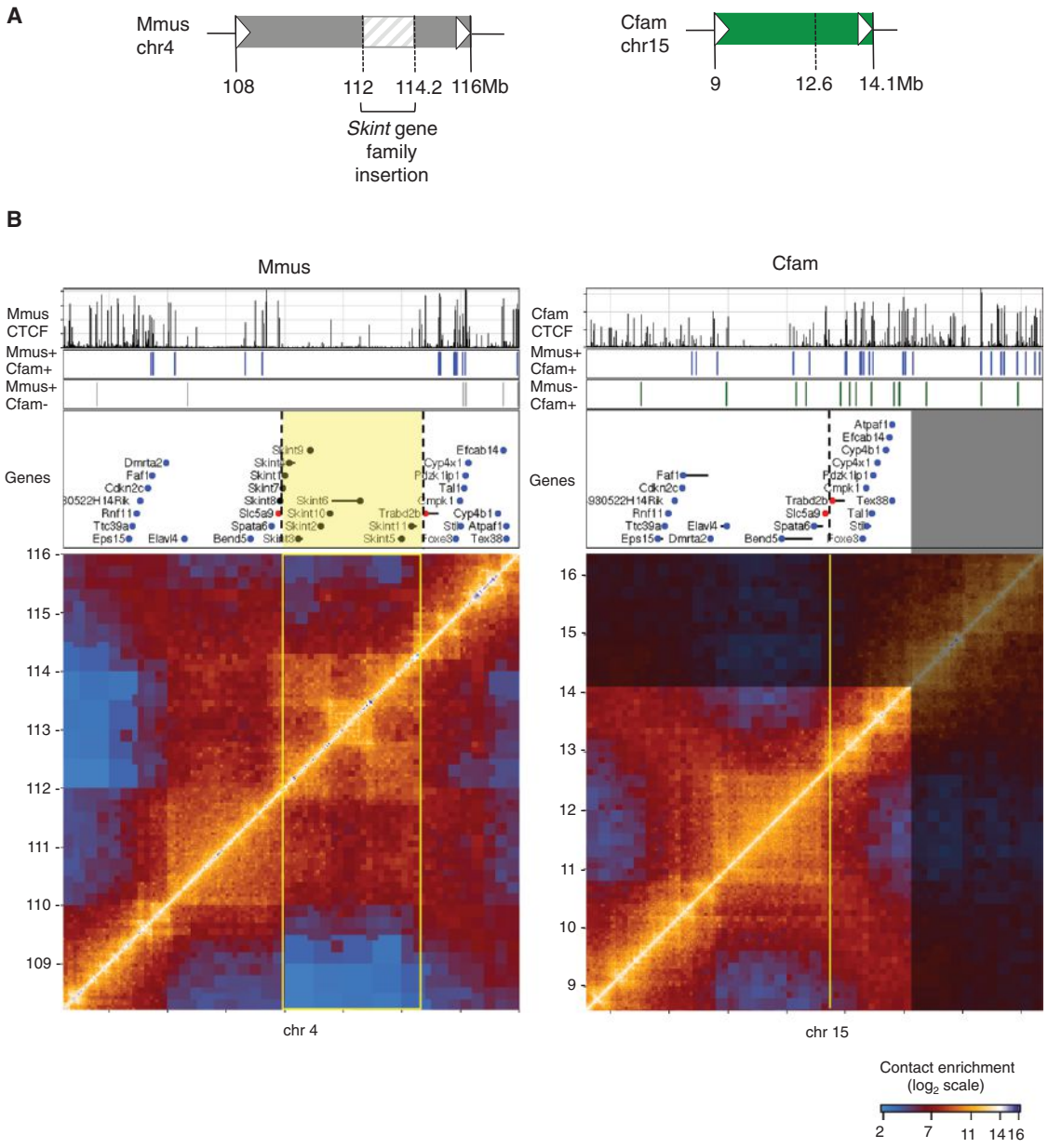


Figure 5. Chromosomal Domains Evolve as Modular Units

(A) A schematic of the rearrangement between mouse (gray) and dog (green).

(B) Hi-C maps from a syntenic region in mouse (left) and dog (right). Shown also are CTCF ChIP tracks, conserved and divergent sites, and genes in the region. A cluster of non-orthologous *Skint* genes (black dots) has been inserted in mouse between the orthologous genes *Slc5a9* and *Trabd2b* (highlighted as red dots). The inserted region (bordered in yellow) forms its own nested chromosomal domain structure, probably as a result of gene-duplication events. Highlighted in blue are other orthologous genes in the region.

See also Figures S7 and S8.

demarcating the implicated functional elements. As CTCF sites are sufficiently sequence-specific to be directly tunable by local nucleotide substitutions, it is intriguing to speculate that the intra-domain looping structure is a key and evolvable feature affecting gene regulation. Such a trait, if indeed quantitatively important, should be further studied between and within populations and species.

EXPERIMENTAL PROCEDURES

Liver Homogenization and Fixation

Fresh or frozen liver from mouse, rabbit, macaque, and dog were processed for Hi-C or 4C-seq libraries. With the exception of mouse, the samples used for the Hi-C libraries were the same as the material used for CTCF ChIP-seq (Schmidt et al., 2012). Livers were fixed in 10% formalin for 20 min, and ~1 g was cut and processed with a Dounce homogenizer (ten strokes with a

Table 1. 4C-Seq Primers Used in This Study

Viewpoint	CTCF Peak	Reading Primer	Non-reading Primer
Mouse 4C-seq primers (mm10)			
Mmus+ Cfam+ 1 (Figure 4)	chr10:94609250	5′-CCATCTGTTTGAACAAGATC-3′	5′-CAAGAGAGAGTGGAAACAGG-3′
Mmus+ Cfam+ 2 (Figure 4)	chr10:94623583	5′-AGTCAGATGGAATGCAGATC-3′	5′-CTAGATACAGCAATCAGCCC-3′
Mmus+ Cfam+ 3 (Figure 4)	chr10:94958324	5′-ATTGCTTTCTCTGGTTGATC-3′	5′-AGTCACTCCTGCTCCTGTAA-3′
Mmus+ Cfam+ 4 (Figure 4)	chr10:94991353	5′-GTTTCTGTTGGTTCACGATC-3′	5′-AAGCATTGTCCTACGTGATT-3′
Mmus+ Cfam− (Figure 4)	chr10:95218005	5′-CTACTCTGGCTTCTATGATC-3′	5′-CCCTTCCCTTCTATGTTTCT-3′
Dog 4C-seq primers (canFam3)			
Mmus+ Cfam+ 1 (Figure 4)	chr15:34606369	5′-GCTCTTGCTCTAAACTGATC-3′	5′-TGGACCTCACCTCTCCTA-3′
Mmus+ Cfam+ 2 (Figure 4)	chr15:34596229	5′-TGAGGTCCAGCAGAGATC-3′	5′-GTCGCATCACTTACTGGG-3′
Mmus+ Cfam+ 3 (Figure 4)	chr15:34269944	5′-CTCCACTGAGCATTAAGATC-3′	5′-GCGGGATAGTTCTTTTCTCT-3′
Mmus+ Cfam+ 4 (Figure 4)	chr15:34244336	5′-CTTATGTGCTCCTCCAGATC-3′	5′-AATCATATGCCTCCTCCTCT-3′
Mmus+ Cfam− (Figure 4)	chr15:33989305	5′-AAAGTAATCCCACCCAGATC-3′	5′-CTGAAGGAAACAACAATGTCA-3′

loose pestle followed by ten strokes with the tight pestle). After filtration through a 70-μm nylon cell strainer, the sample was washed twice with PBS, spinning down at 852 rcf for 5 min at 4°C to collect the cells between washes. 1–5 × 10^7 liver cells were then fixed for a second time in fixation buffer (1% formaldehyde, 750 μg/ml BSA in DMEM/Ham's F12 [Invitrogen]) for 10–30 min at room temperature. The fixation reaction was quenched using 0.125 M glycine for 5 min at room temperature. Samples were washed twice with 10 ml PBS, pelleted into 1 × 10^7 cells aliquots, and stored at −80°C. Mouse Hi-C libraries were prepared from fresh liver samples of biological replicates (9-week-old C57/BL6 mouse and the pooled livers from 2- to 4-week-old outbred mice. The libraries for the other three organisms were technical replicates.

Propidium Iodide Staining of Hepatocytes

Formaldehyde-fixed liver cells were lysed on ice in a hypotonic buffer (10 mM Tris-HCl [pH 8], 10 mM NaCl, 0.2% Igepal CA-640, EDTA-free protease inhibitors) for 30 min. Nuclei were stained with a propidium iodide (PI) staining buffer (100 μg/ml PI, 50 μg/ml RNase A, 0.05% Triton X-100) for 60 min on ice. Samples were analyzed on a MoFlo cell sorter (Beckman Coulter).

High-Throughput Mapping of Chromatin Interactions via Hi-C

The Hi-C method previously used (Sofueva et al., 2013) was modified to accommodate primary liver samples. Hepatocytes were lysed in Hi-C lysis buffer (10 mM Tris-HCl [pH 8], 10 mM NaCl, 0.2% Igepal CA-640, EDTA-free protease inhibitors) for 30 min on ice. The sample was transferred to Protein LoBind tubes (Eppendorf) and the nuclei were permeabilized by incubation with 0.1%–0.6% SDS for 1 hr at 37°C with 800 rpm shaking. The reaction was quenched with 0.67%–4% Triton X-100, 1 hr at 37°C, 800 rpm shaking. Nuclei were digested in 500 μl 1X NEBuffer 2 with 1500 U HindIII (New England Biolabs) and monitored for maximal digestion of the chromatin template, thus digestion times ranged from 24-72 hr. All other parts of the Hi-C protocol, including library preparation were performed as previously described. 75 bp paired-end sequencing was performed for each library according to manufacturers conditions using the Illumina Hi-seq platform.

Hi-C Interaction Matrix Generation and Domain Calling

Sequencing reads were aligned to the mouse (mm10), rabbit (oryCun2), macaque (rheMac2), and dog (canFam3) genome assemblies using Bowtie 0.12.8 (Langmead et al., 2009). The parameters used for the alignment allowed a maximum of three mismatches and strictly one alignment per read. Processing of the aligned reads and normalization of the interaction matrices were performed as previously described (Yaffe and Tanay, 2011; Sofueva et al., 2013). The pipeline produced normalized matrices of interactions binning the genome at different resolutions. Interaction matrices for each library were generated displaying seven different resolutions simultaneously (12,500, 25,000, 50,000, 100,000, 250,000, 500,000, and 1,000,000 bp). Domains were identified and clustered as described (Sexton et al., 2012) with the modification that scaling factors were inferred using fends 100–400 kb apart, to account for the lower resolution of the mouse map compared to the *Drosophila* map. Domain borders were called using the 95% percentile of the scaling track. A domain-level map was partitioned into two clusters, and clusters were assigned as passive/active according to Lamin B mouse embryonic fibroblast (MEF) data, as before. For the rabbit, macaque, and dog genome, the Lamin B MEF track for mouse was lifted over to the corresponding genome to label domain clusters. Domain calls in mouse and dog are available in Table S1.

ChIP-Seq Analysis

We used previously published ChIP-seq data for CTCF from mouse, macaque, and dog livers (Schmidt et al., 2012) and for Rad21 for mouse liver (Faure et al., 2012). Rad21 ChIP-seq data for macaque and dog was prepared as for CTCF. Mouse, macaque, and dog ChIP-seq reads were mapped using bowtie. Alignment was followed by extension of sequenced tags to 300-bp fragments and pileup into 50-bp bins. We normalized ChIP-seq coverage by computing the distribution of pile-up coverage on 50-bp bins and transforming each coverage value v into −log10 (1-quantile(v)). To define binding sites, we used a simple threshold on the sum of values from two biological replicates for each CTCF dataset and for the macaque Rad21 data. Rad21 ChIP data from mouse and dog were done in single, and the data were thresholded. Thresholds used were as follows: mouse CTCF = 2.2, macaque CTCF = 2.4, dog CTCF = 2.2, mouse Rad21 = 2.3, macaque Rad21 = 2.5, dog Rad21 = 3. Different thresholds did not change the results. Binding site width was standardized at 200 bp, and the ChIP-seq intensity for each site was calculated as the maximum value across the 200 bp. The relative distribution of CTCF within topological domains (Figures 1E and S5) was calculated as the distance of each CTCF site from the center of its domain. Half the size of the domain was added to convert it to a measure of distance from the edge of the domain, and this number was then divided by the size of the domain.

Interspecies Comparison of CTCF Sites

Macaque and dog CTCF ChIP-seq libraries were converted to mouse genome coordinates using the liftOver tool from UCSC. To reduce the chance of inaccurate liftOver, a number of filters were implemented: sites within low-mappability regions, repeats, or windows of 100 kb with insufficient synteny were excluded. To estimate mappability, each genome was broken into 50-bp reads and the whole-genome sequence was split into artificial reads and then mapped back to the genome. For each 50-bp bin, the mappability score was then defined to be the portion of artificial reads mapped uniquely to that bin. To estimate the level of synteny in the 100 kb around a CTCF site, the mappability tracks for macaque and dog were converted to the mouse genome using liftOver and all bins for which liftOver was not possible were converted to zeroes. The converted tracks were subsequently smoothed over 100 kb, and CTCF

sites falling in regions below the top quartile of such smoothed tracks were excluded from all subsequent analysis. Divergent CTCF sites in mouse and dog are available in Table S1.

CTCF Binding Energy Function

A CTCF DNA-binding energy function from the Cortex CTCF binding sites (ENCODE Cortex CTCF mouse, GSM769019; Shen et al., 2012) was used to profile all genomes for their similarity to the CTCF consensus motif. The consensus motif is very highly conserved across all species (Schmidt et al., 2012). Given a set of genomic sites, we compute for each site the maximal energy value within a 200-bp window centered on the point.

Motif Orientation Analysis

Orientation of the motifs underneath conserved CTCF peaks was obtained using MEME (http://meme.nbcr.net/meme/), (Bailey and Elkan, 1994) with the parameters -revcomp -dna -nmotifs 1 -w 20 -mod zoops -maxsize 100,000.

Crossover Analysis

Crossover analysis was performed as described previously (Sofueva et al., 2013). The bands used were 5–7.5, 7.5–11.25, 10–15, 15–22.5, 20–30, 30–45, 40–60, 60–90, and 80–120 kb.

Distal Contact Analysis

To calculate the average interaction profiles for a group of genomic landmarks, HindIII fragment ends were grouped into classes by associating each end with a genomic element located within 5 kb and then grouping all fragment ends associated with an element of the same class. For the mouse, macaque, and dog genomes, three classes of CTCF sites (conserved, divergent present, divergent absent) and TSS sites were defined. These classes were further divided to sites within active or passive Hi-C domains. The remaining fragment end (not classified given other landmarks) was defined as the background.

4C-Seq

Preparation of 4C-seq samples, libraries, sequencing analysis, and normalization were all performed as previously described (Sofueva et al., 2013). Primer sequences were chosen to viewpoint sites that were as close as possible to CTCF ChIP-seq peaks (Table 1). Mouse primers were designed according to the genome-wide 4C-seq primer database from (van de Werken et al., 2012). For dog primers, a similar database was generated for the regions of interest.

ACCESSION NUMBERS

The data analyzed in this study have been deposited in the GEO database with the accession number GSE65126.

SUPPLEMENTAL INFORMATION

Supplemental Information includes eight figures and one table and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2015.02.004.

AUTHOR CONTRIBUTIONS

M.V.R., D.T.O., and S.H. initiated the project. M.V.R. performed the Hi-C and 4C-Seq experiments, including library preparations. D.T.O. and C.E. provided the liver samples for all species (except mouse) and sequenced the Hi-C libraries. M.V.R., C.B., and A.T. processed and statistically analyzed the data. M.V.R., A.T., and S.H. wrote the manuscript, with contributions from all authors.

ACKNOWLEDGMENTS

The authors wish to thank Sevil Sofueva for help with Hi-C and 4C-seq library preparations; Wen-Ching Chan for analysis; Christine Feig for Rad21 ChIP-seq data in macaque and dog; Bianca Schmidt and the Cancer Research UK CI Genomic Core facility for technical assistance with Hi-C sequencing and Pedro Olivares for advice on analysis and data manipulation. We would also like to acknowledge all members of the Hadjur group for discussions. This work was supported by the Medical Research Council UK (G0900491/1 and G1001649) (S.H.), the EPIGENESYS EU NoE (S.H. and A.T.), and Cancer Research UK (studentship to M.V.R.). D.T.O. is supported by Cancer Research UK.

Received: September 27, 2014
Revised: December 9, 2014
Accepted: January 29, 2015
Published: February 26, 2015

REFERENCES

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. *2*, 28–36.

Bell, A.C., West, A.G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell *98*, 387–396.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Stamatoyannopoulos, J.A., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816. http://dx.doi.org/10.1038/nature05874.

Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R., Wang, L.Y., Gerstein, M., and Snyder, M. (2007). Divergence of transcription factor binding sites across related yeast species. Science *317*, 815–819. http://dx.doi.org/10.1126/science.1140748.

Boyden, L.M., Lewis, J.M., Barbee, S.D., Bas, A., Girardi, M., Hayday, A.C., Tigelaar, R.E., and Lifton, R.P. (2008). Skint1, the prototype of a newly identified immunoglobulin superfamily gene cluster, positively selects epidermal gammadelta T cells. Nat. Genet. *40*, 656–662. http://dx.doi.org/10.1038/ng.108.

Dermitzakis, E.T., and Clark, A.G. (2001). Differential selection after duplication in mammalian developmental genes. Mol. Biol. Evol. *18*, 557–562.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in Mamm. Genomes identified by analysis of chromatin interactions. Nature *485*, 376–380. http://dx.doi.org/10.1038/nature11082.

Dong, X., Han, S., Zylka, M.J., Simon, M.I., and Anderson, D.J. (2001). A diverse family of GPCRs expressed in specific subsets of nociceptive sensory neurons. Cell *106*, 619–632.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., and Young, R.A. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374–387. http://dx.doi.org/10.1016/j.cell.2014.09.030.

Faure, A.J., Schmidt, D., Watt, S., Schwalie, P.C., Wilson, M.D., Xu, H., Ramsay, R.G., Odom, D.T., and Flicek, P. (2012). Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. Genome Res. *22*, 2163–2175. http://dx.doi.org/10.1101/gr.136507.111.

Filippova, G.N., Fagerlie, S., Klenova, E.M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P.E., Collins, S.J., and Lobanenkov, V.V. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. Mol. Cell. Biol. *16*, 2802–2813.

Guacci, V., Koshland, D., and Strunnikov, A. (1997). A direct link between sister chromatid cohesion and chromosome condensation revealed through the analysis of MCD1 in S. cerevisiae. Cell *91*, 47–57.

Hadjur, S., Williams, L.M., Ryan, N.K., Cobb, B.S., Sexton, T., Fraser, P., Fisher, A.G., and Merkenschlager, M. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. Nature *460*, 410–413. http://dx.doi.org/10.1038/nature08079.

Hou, C., Li, L., Qin, Z.S., and Corces, V.G. (2012). Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into

physical domains. Mol. Cell *48*, 471–484. http://dx.doi.org/10.1016/j.molcel.2012.08.031.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245. http://dx.doi.org/10.1016/j.cell.2006.12.048.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25. http://dx.doi.org/10.1186/gb-2009-10-3-r25.

Michaelis, C., Ciosk, R., and Nasmyth, K. (1997). Cohesins: chromosomal proteins that prevent premature separation of sister chromatids. Cell *91*, 35–45.

Mishiro, T., Ishihara, K., Hino, S., Tsutsumi, S., Aburatani, H., Shirahige, K., Kinoshita, Y., and Nakao, M. (2009). Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. EMBO J. *28*, 1234–1245. http://dx.doi.org/10.1038/emboj.2009.81.

Nativio, R., Wendt, K.S., Ito, Y., Huddleston, J.E., Uribe-Lewis, S., Woodfine, K., Krueger, C., Reik, W., Peters, J.-M., and Murrell, A. (2009). Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. PLoS Genet. *5*, e1000739. http://dx.doi.org/10.1371/journal.pgen.1000739.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385. http://dx.doi.org/10.1038/nature11049.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., et al. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell *132*, 422–433. http://dx.doi.org/10.1016/j.cell.2008.01.011.

Pauli, A., Althoff, F., Oliveira, R.A., Heidmann, S., Schuldiner, O., Lehner, C.F., Dickson, B.J., and Nasmyth, K. (2008). Cell-type-specific TEV protease cleavage reveals cohesin functions in Drosophila neurons. Dev. Cell *14*, 239–251. http://dx.doi.org/10.1016/j.devcel.2007.12.009.

Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.-T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281–1295. http://dx.doi.org/10.1016/j.cell.2013.04.053.

Rollins, R.A., Morcillo, P., and Dorsett, D. (1999). Nipped-B, a Drosophila homologue of chromosomal adherins, participates in activation by remote enhancers in the cut and Ultrabithorax genes. Genetics *152*, 577–593.

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., et al. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science *328*, 1036–1040. http://dx.doi.org/10.1126/science.1186176.

Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Gonçalves, A., Kutter, C., Brown, G.D., Marshall, A., Flicek, P., and Odom, D.T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell *148*, 335–348. http://dx.doi.org/10.1016/j.cell.2011.11.058.

Seitan, V.C., Hao, B., Tachibana-Konwalski, K., Lavagnolli, T., Mira-Bontenbal, H., Brown, K.E., Teng, G., Carroll, T., Terry, A., Horan, K., et al. (2011). A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. Nature *476*, 467–471. http://dx.doi.org/10.1038/nature10312.

Seitan, V.C., Faure, A.J., Zhan, Y., McCord, R.P., Lajoie, B.R., Ing-Simmons, E., Lenhard, B., Giorgetti, L., Heard, E., Fisher, A.G., et al. (2013). Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. Genome Res. *23*, 2066–2077. http://dx.doi.org/10.1101/gr.161620.113.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472. http://dx.doi.org/10.1016/j.cell.2012.01.010.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V., and Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120. http://dx.doi.org/10.1038/nature11243.

Sofueva, S., Yaffe, E., Chan, W.-C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., Pollard, S.M., Schroth, G.P., Tanay, A., and Hadjur, S. (2013). Cohesin-mediated interactions organize chromosomal domain architecture. EMBO J. *32*, 3119–3129. http://dx.doi.org/10.1038/emboj.2013.237.

van de Werken, H.J.G., Landan, G., Holwerda, S.J.B., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Öz, Y., Bouwman, B.A.M., et al. (2012). Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat. Methods *9*, 969–972. http://dx.doi.org/10.1038/nmeth.2173.

Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., et al. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature *451*, 796–801. http://dx.doi.org/10.1038/nature06634.

Xiao, T., Wallace, J., and Felsenfeld, G. (2011). Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity. Mol. Cell. Biol. *31*, 2174–2183. http://dx.doi.org/10.1128/MCB.05093-11.

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat. Genet. *43*, 1059–1065. http://dx.doi.org/10.1038/ng.947.

Yaffe, E., Farkash-Amar, S., Polten, A., Yakhini, Z., Tanay, A., and Simon, I. (2010). Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. PLoS Genet. *6*, e1001011. http://dx.doi.org/10.1371/journal.pgen.1001011.s016.

Zuin, J., Dixon, J.R., van der Reijden, M.I.J.A., Ye, Z., Kolovos, P., Brouwer, R.W.W., van de Corput, M.P.C., van de Werken, H.J.G., Knoch, T.A., van IJcken, W.F.J., et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc. Natl. Acad. Sci. USA *111*, 996–1001. http://dx.doi.org/10.1073/pnas.1317788111.

OPEN ACCESS
CellPress

Cell Reports
Article

Heterogeneities in *Nanog* Expression Drive Stable Commitment to Pluripotency in the Mouse Blastocyst

Panagiotis Xenopoulos,[1,5] Minjung Kang,[1,2,5] Alberto Puliafito,[3] Stefano Di Talia,[4] and Anna-Katerina Hadjantonakis[1,*]
[1]Developmental Biology Program, Sloan Kettering Institute, New York, NY 10065, USA
[2]Biochemistry, Cell and Molecular Biology Program, Weill Graduate School of Medical Sciences of Cornell University, New York, NY 10065, USA
[3]Laboratory of Cell Migration, Candiolo Cancer Institute - FPO, IRCCS, Candiolo, Torino 10060, Italy
[4]Department of Cell Biology, Duke University Medical Center, Durham, NC 27710, USA
[5]Co-first author
*Correspondence: hadj@mskcc.org
http://dx.doi.org/10.1016/j.celrep.2015.02.010

SUMMARY

The pluripotent epiblast (EPI) is the founder tissue of almost all somatic cells. EPI and primitive endoderm (PrE) progenitors arise from the inner cell mass (ICM) of the blastocyst-stage embryo. The EPI lineage is distinctly identified by its expression of pluripotency-associated factors. Many of these factors have been reported to exhibit dynamic fluctuations of expression in embryonic stem cell cultures. Whether these fluctuations correlating with ICM fate choice occur in vivo remains an open question. Using single-cell resolution quantitative imaging of a *Nanog* transcriptional reporter, we noted an irreversible commitment to EPI/PrE lineages in vivo. A period of apoptosis occurred concomitantly with ICM cell-fate choice, followed by a burst of EPI-specific cell proliferation. Transitions were occasionally observed from PrE-to-EPI, but not vice versa, suggesting that they might be regulated and not stochastic. We propose that the rapid timescale of early mammalian embryonic development prevents fluctuations in cell fate.

INTRODUCTION

Pluripotency is defined as the ability of a cell to differentiate and give rise to all somatic and germ cells (Nichols and Smith, 2012). Although pluripotency can be induced in differentiated cells (Takahashi and Yamanaka, 2006), how a pluripotent population emerges in its native context, within the early mammalian embryo, remains an open question. Insight will come from elucidating the dynamic cell behaviors and molecular mechanisms underlying the development of the mammalian blastocyst, the embryonic stage at which a bona fide pluripotent population—the epiblast (EPI)—is established.

The EPI is molecularly distinct and spatially segregated from the two extra-embryonic lineages, the primitive endoderm (PrE) and trophectoderm (TE) of the mouse blastocyst. The specification of these lineages occurs as two sequential binary cell-fate decisions. The first involves specification and segregation of TE from inner cell mass (ICM), while the second occurs within the ICM and involves the specification of EPI and PrE precursors, and their eventual segregation into adjacent tissue layers (reviewed in Schrode et al., 2013). By late blastocyst stage, the EPI and PrE lineages are defined both by their position within the embryo and expression of lineage-specific transcription factors, such as NANOG in the EPI, and GATA6 and GATA4 in the PrE (Xenopoulos et al., 2012). Recent studies have illustrated that EPI/PrE allocation occurs in at least three successive steps (Chazaud et al., 2006; Frankenberg et al., 2011; Plusa et al., 2008). Initially, lineage-specific transcription factors, such as NANOG and GATA6, are co-expressed by all ICM cells, suggesting a multi-lineage priming state. Thereafter, NANOG and PrE lineage-specific transcription factors exhibit mutually exclusive expression, as lineage progenitors emerge in a salt-and-pepper distribution within the ICM. At this stage, GATA4 becomes activated in PrE progenitors, concomitant with NANOG downregulation. Finally, lineage segregation is achieved with the localization of PrE cells to the surface of the ICM. At this time, other pluripotency-associated factors become restricted to EPI cells, which have become positioned internally within the ICM. Notably, NANOG is one of the first markers to be restricted within the EPI, whereas OCT4 and SOX2 become subsequently downregulated in PrE progenitors and restricted to EPI progenitors.

The initial specification of EPI and PrE progenitors appears to occur in a spatially random manner (Schrode et al., 2014) and could be achieved if a stochastic process were to underlie this second fate decision. Indeed, an analysis of transcriptomes of single ICM cells revealed that gene expression is highly heterogeneous at earlier stages, exhibiting no apparent lineage specificity and a hierarchical relationship of marker expression only appearing in the late blastocyst (Guo et al., 2010; Kurimoto et al., 2006; Ohnishi et al., 2014).

A degree of heterogeneity has been observed at both protein and mRNA level for various pluripotency-associated factors in embryonic stem cell (ESC) cultures. Many studies have focused

on *Nanog*, a central component of the core pluripotency transcriptional network (Chambers et al., 2007; Kalmar et al., 2009). Experiments in ESCs have suggested that *Nanog* expression displays dynamic fluctuations that may correlate with a cell's fate choice between self-renewal and differentiation. However, it is unclear whether fluctuations in gene expression take place in vivo in embryos where cell differentiation occurs on a shorter timescale, nor whether they predict fate choice or fate reversion. Notably, understanding how pluripotent cells behave in embryos may provide information that can be reconciled with observations made in ESCs (Smith, 2013).

To determine how the EPI emerges within the mouse blastocyst, we generated a reporter of *Nanog* transcription (*Nanog:H2B-GFP*). Derivation of ESCs from reporter-expressing embryos revealed heterogeneous gene expression as an adaptation to ESC propagation. Using live imaging, we quantified the dynamics of *Nanog* expression in individual cells of live blastocysts, establishing how *Nanog* expression influences the fate of ICM cells. By contrast to ESCs maintained in culture, fluctuations in *Nanog* expression between distinct developmental states did not, generally, occur in vivo. However, we noted rare cases of cells unidirectionally switching their fate, from a PrE to EPI identity. Since ICM fate changes were only observed toward the EPI, and not toward PrE, we concluded that this change was not stochastic. Our analyses also revealed events of selective apoptosis at the onset of ICM lineage differentiation, followed by a burst of cell proliferation in EPI-committed cells. Collectively, these data suggest that, although it may be dynamic in ESCs, the emergence of a pluripotent identity is sequential and linear in vivo and not accommodating reversibility in fate. In this way, after its allocation, the pluripotent cell population might be protected, thereby ensuring the development of somatic lineages.

RESULTS

BAC-Based *Nanog* Transcriptional Reporters Mark the Pluripotent State in ESCs and Embryos

To probe the dynamics of the pluripotent state, we developed a BAC-based *Nanog* transgenic transcriptional reporter, based on a previous design used as a readout of cellular reprogramming during iPS cell generation (Okita et al., 2007) (Figure S1H). For single-cell resolution readouts of *Nanog* expression, we generated nuclear-localized human histone H2B fusion versions of the reporter (Figures 1A, 1C, and S1E).

To validate transgene activity, we analyzed reporter expression in transgenic ESCs under various culture conditions. These conditions included the presence or absence of leukemia inhibitory factor (LIF), and 2i+LIF, which promote the self-renewal of ESCs, induce differentiation, or ground state pluripotency, respectively (Ying et al., 2008). Immunostaining of ESCs in 2i+LIF or serum-LIF conditions revealed markedly increased or decreased expression, respectively, of both reporter and NANOG protein. Heterogeneous but correlated GFP and NANOG expression was observed in *Nanog:GFP*$^{Tg/+}$ and *Nanog:H2B-GFP*$^{Tg/+}$ ESCs maintained in serum+LIF conditions (Figures 1A, 1B, S1D, and S1E). Single-cell quantitative image analyses of immunostained *Nanog:H2B-GFP*$^{Tg/+}$ ESCs maintained under different conditions further validated reporter efficacy (Figures 1B, S1E, and S1F). Moreover, we observed an increased correlation of reporter activity with NANOG protein for the *Nanog:H2B-GFP* transgene, compared to the targeted *Nanog* transcriptional reporter in the heterozygous TNGA ESCs (Chambers et al., 2007) (Figures S1E and S1F). These data suggest that the BAC transgenic reporters we constructed faithfully marked the pluripotent state in ESC cultures and could be used to probe *Nanog* expression dynamics at single-cell resolution.

Next, we generated *Nanog:H2B-GFP* transgenic mice for single-cell resolution quantitative visualization of the EPI lineage in vivo. We used live imaging to analyze the distribution and validate the reporter in embryos (Figure 1C). We first observed reporter activity in embryos at the 8–16 cell stage (Plusa et al., 2008). Cells displaying high and low GFP levels were first observed at the mid blastocyst stage (70–100 cells), as a salt-and-pepper distribution of EPI and PrE precursors was established within the ICM (Chazaud et al., 2006; Plusa et al., 2008). In late blastocysts (>100 cells), where the EPI and PrE lineages have sorted into distinct layers, cells within the ICM forming the EPI exhibited significantly elevated levels of GFP compared to PrE cells located on the surface of the ICM. Thereafter, GFP was detected within the EPI, albeit at reduced levels in implanting E4.5 and post-implantation E5.5 embryos consistent with the downregulation of NANOG observed at periimplantation (Chambers et al., 2003). Notably, the reporter was strongly expressed in TE cells of early and mid blastocysts (corresponding to ~32–90 cell stage), possibly resulting from robust NANOG localization in ICM and TE cells at these stages (Figure S1I) (Dietrich and Hiiragi, 2007; Messerschmidt and Kemler, 2010; Morgani et al., 2013). However, GFP expression in the TE was significantly reduced in late (>100 cells) and in implanting blastocysts (Figures 1C and S1G). TE localization was also noted for other *Nanog*-based reporters analyzed at comparable embryonic stages (Figure S1G). Collectively, these data lead us to conclude that the *Nanog:H2B-GFP* reporter faithfully marks the pluripotent state both in vitro in ESC cultures and in vivo in the emerging EPI lineage of the mouse blastocyst.

Derivation of ESCs from Mouse Blastocysts Results in Highly Variable Pluripotency-Associated Gene Expression

The EPI lineage of the blastocyst represents the in vivo counterpart to ESC cultures propagated in vitro (Boroviak et al., 2014). We therefore sought to investigate the profile of *Nanog:H2B-GFP* reporter activity in transgenic embryo-derived ESCs. We used an ESC derivation protocol in which an ICM outgrowth emerges in conditions promoting ground state pluripotency in serum-free medium containing 2i (Czechanski et al., 2014). Indeed, under these conditions GFP was strongly expressed by all cells of outgrowths from *Nanog:H2B-GFP*$^{Tg/+}$ blastocysts (Figure 1D). However, both reporter and NANOG expression became heterogeneous when the derived ESCs were propagated under standard serum+LIF conditions, either in the presence or absence of mouse embryonic feeders (MEFs). Even in the presence of MEFs, GFP-low/NANOG-low cells were observed within ESC colonies (Figure 1E). Fluorescence-activated cell sorting (FACS) and quantitative immunofluorescence analyses confirmed the presence of a highly heterogeneous

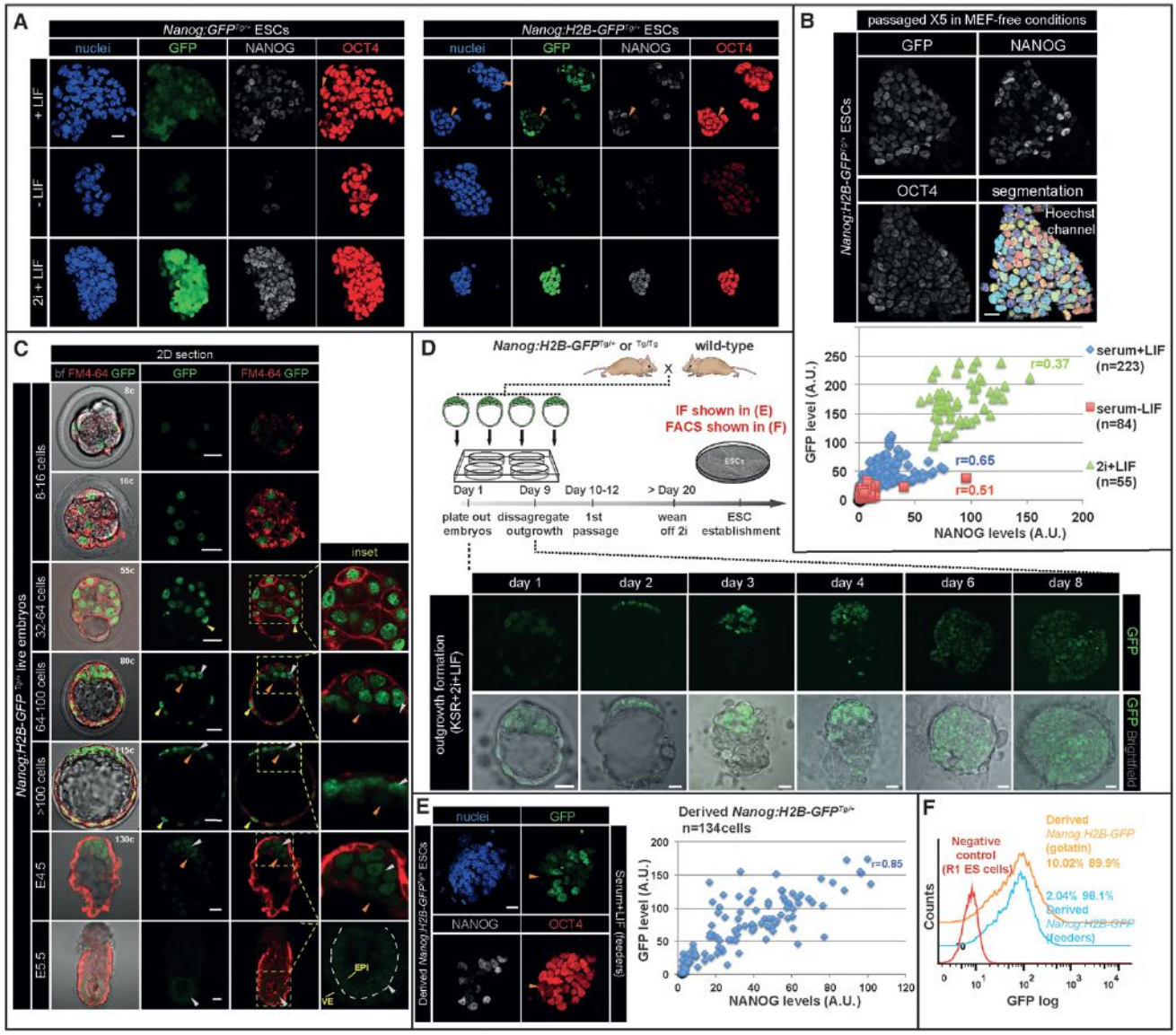


Figure 1. BAC-Based *Nanog* Transcriptional Reporters Faithfully Mark the Pluripotent State in ESCs and Embryos

(A) Immunofluorescence images of *Nanog:GFP*$^{Tg/+}$ and *Nanog:H2B-GFP*$^{Tg/+}$ ESCs grown in MEF-free conditions including serum+LIF, 2i+LIF, and serum-LIF for three passages. Orange arrowheads identify GFP-low and NANOG-low cells within ESC colonies.

(B) Quantitative immunofluorescence analysis after nuclear segmentation of *Nanog:H2B-GFP*$^{Tg/+}$ ESCs. GFP (y axis) and NANOG (x axis) fluorescence values plotted for individual ESCs propagated for five passages in MEF-free serum+LIF conditions then grown for 4 days in various culture conditions.

(C) Reporter expression in live embryos stained with membrane marker FM4-64. GFP-hi cells, white arrowheads; GFP-low cells, orange arrowheads; TE cells expressing GFP, yellow arrowheads. Dashed line depicts boundary between epiblast (EPI) and visceral endoderm (VE) layers of an E5.5 embryo. Cell number was determined by staining with Hoechst.

(D) Schematic of ESC derivation. After 20 days, *Nanog:H2B-GFP*$^{Tg/+}$ ESCs were established in the presence of MEF feeders in serum+LIF conditions and then propagated in the presence or absence of MEFs.

(E) Immunostaining and analysis of derived *Nanog:H2B-GFP*$^{Tg/+}$ ESCs. Orange arrowheads mark GFP-low/NANOG-low/OCT4^{+} cells.

(F) FACS analysis of derived *Nanog:H2B-GFP*$^{Tg/+}$ ESCs grown in serum+LIF conditions in the presence or absence of MEFs. Numbers in FACS histograms indicate percentage of GFP^{-} (left) and GFP^{+} (right) populations. r = Pearson correlation coefficient. Scale bar represents 20 μm.

population, where both reporter and protein expression where highly correlated (r = 0.85) (Figures 1E and 1F). From these observations, we conclude that gene expression heterogeneities of the pluripotent state likely arise in ESCs as a result of their in vitro propagation.

Quantitative Single-Cell Analysis of *Nanog:H2B-GFP* Expression in the Emerging Pluripotent Cells of the Blastocyst

To determine whether the transgenic reporter allowed a quantitative evaluation of *Nanog* expression in vivo, we analyzed the

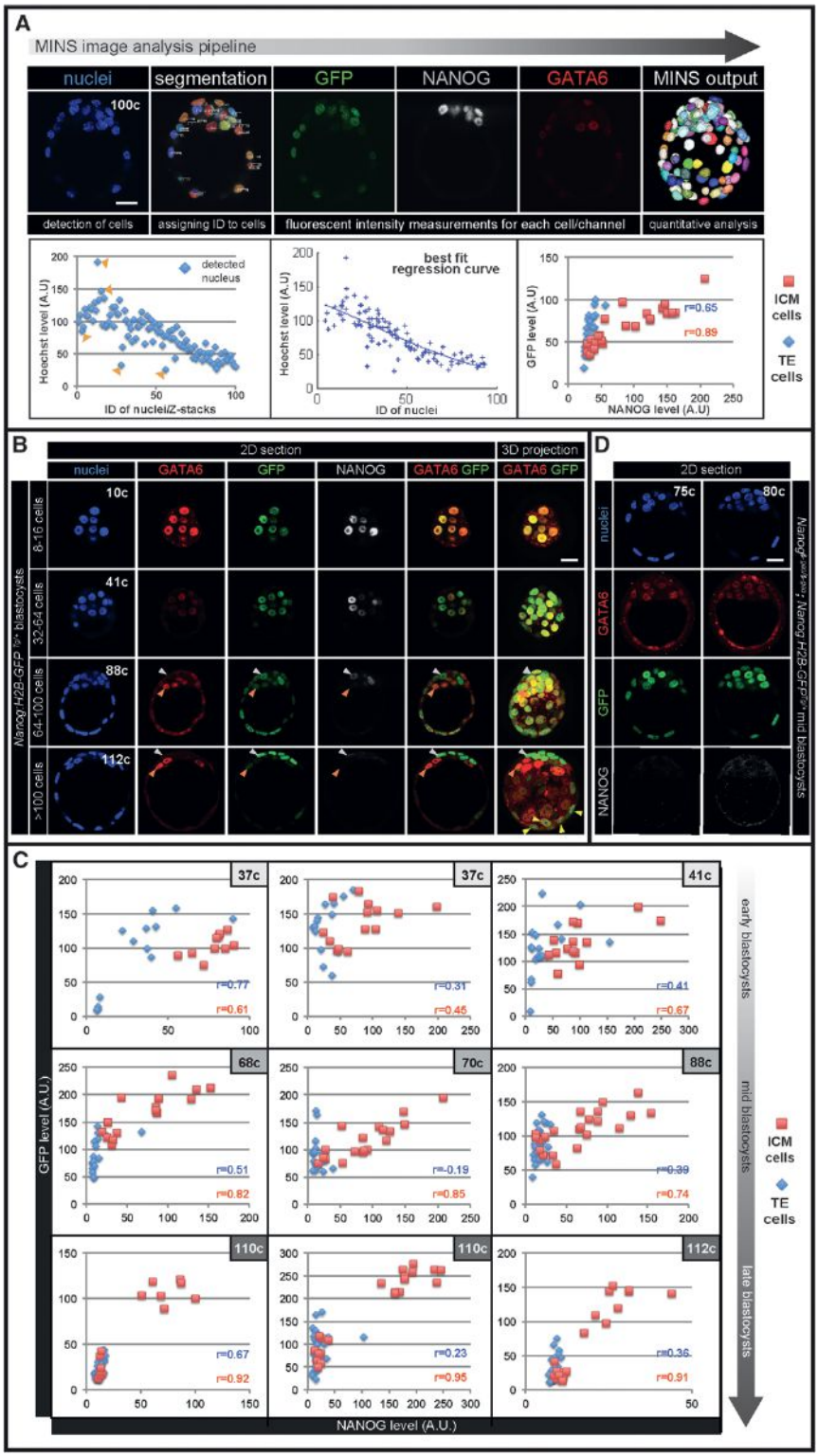


Figure 2. Differential Expression of *Nanog:H2B-GFP* Reporter in Segregating EPI and PrE Cells

(A) MINS image analysis pipeline (see Experimental Procedures). GFP and NANOG expression plotted for individual cells after automated nuclear segmentation and fluorescence intensity normalization (best-fit regression curve).

(B) Immunofluorescence images of fixed *Nanog:H2B-GFP*$^{Tg/+}$ embryos. EPI cells, white arrowheads; PrE cells, orange arrowheads; TE cells expressing GFP, yellow arrowheads.

(C) Quantitative immunofluorescence analyses.

(D) Reporter, GATA6, and NANOG expression in *Nanog*$^{\beta\text{-}geo/\beta\text{-}geo}$ blastocysts carrying the *Nanog:H2B-GFP* reporter. NANOG staining was absent in *Nanog* mutant transgenic embryos. r = Pearson correlation coefficient. Scale bar represents 20 μm.

pattern of reporter expression in transgenic embryos that had been fixed and stained for lineage-specific markers such as GATA6/PrE and NANOG/EPI using single-cell quantitative image analysis (Lou et al., 2014) (Figure 2A). In morulae (8-16 cells) and early blastocyst (32–64 cells) stages, GFP was observed throughout the embryo, reflecting a double (GATA6$^+$ NANOG$^+$) -positive state. The differential levels of GFP expression were evident in mid blastocysts (~70–100 cells), as embryos established a mutually exclusive distribution of GATA6$^+$ PrE and NANOG$^+$ EPI progenitors (Figures 2B and S2A). In late blastocysts (>100 cells), where the EPI/PrE sorting had occurred, reporter expression was markedly elevated in NANOG$^+$ EPI progenitor cells and diminished in GATA6$^+$ PrE progenitor cells (Figure 2B). Quantitative fluorescence analysis of immunostained *Nanog:H2B-GFP*$^{Tg/+}$ blastocysts at different stages confirmed a strong correlation between GFP and NANOG levels in ICM cells, as seen in ESCs, thus indicating that the reporter could be used to quantitatively infer levels of expression of *Nanog* (Figures 2C and S2B). Notably, reporter expression in TE cells was prominent in early and mid blastocysts but displayed reduced correlation with NANOG protein compared to the ICM and hence was not investigated further (Figures 2C and S2B). The H2B-GFP reporter responds to changes in *Nanog* expression with a characteristic time that is determined by the stability of the H2B-GFP fusion protein. Our control experiments and computational analysis indicate that we can measure changes in *Nanog* expression higher than 2-fold with about 1 hr resolution (see the Supplemental Experimental Procedures). Such temporal resolution is

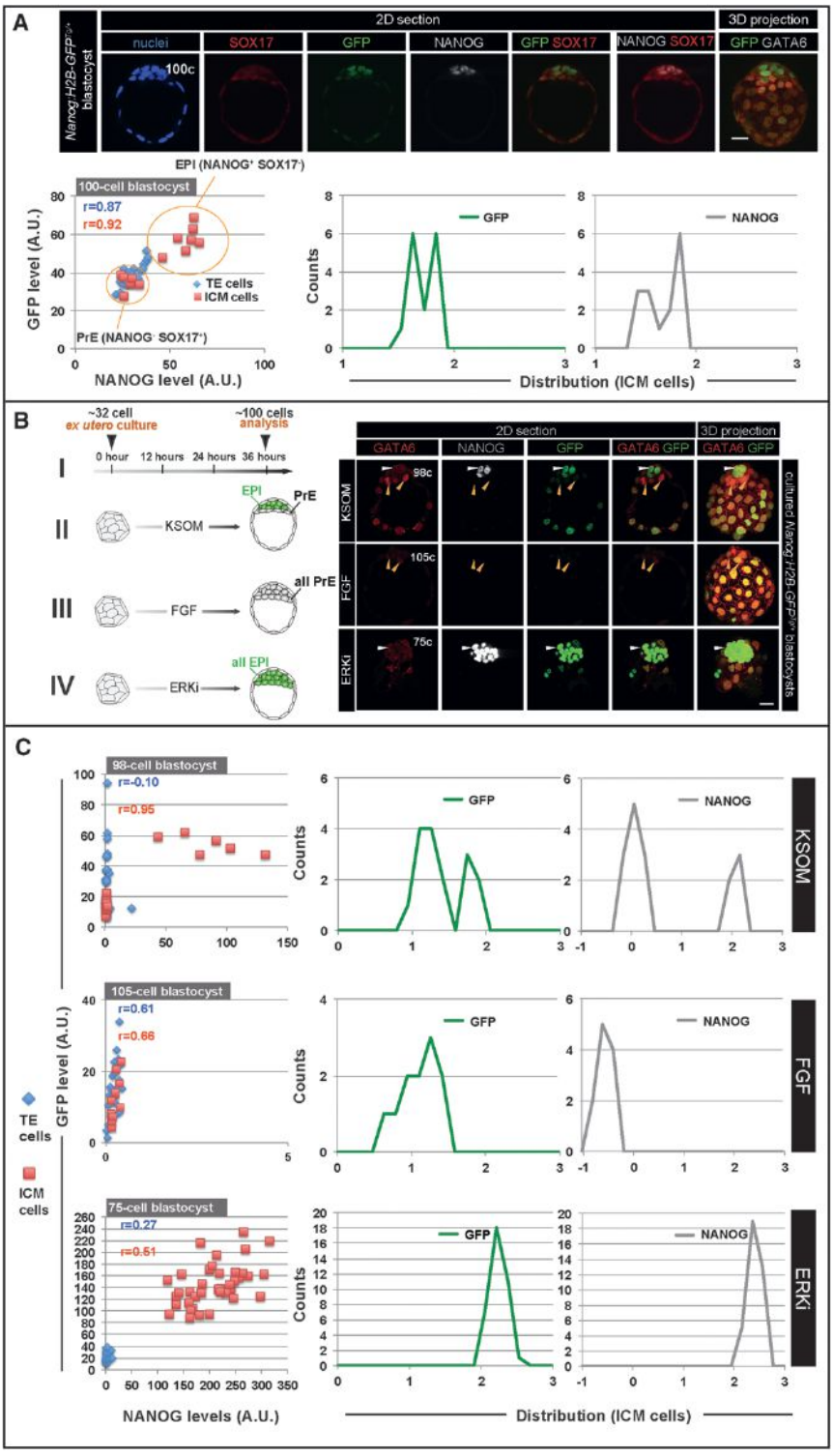


Figure 3. *Nanog* Expression in the ICM Is Altered by Modulation of FGF Signaling

(A) Immunofluorescence images of late blastocyst stage *Nanog:H2B-GFP*$^{Tg/+}$ embryo. Fluorescence values for GFP and NANOG plotted for individual cells. Values for GFP and NANOG of ICM cells also plotted as frequency distributions obtained by binning fluorescence values in 20 logarithmically spaced categories, as described previously (Muñoz Descalzo et al., 2012).

(B) Regimen used for exogenous FGF and ERK1/2 inhibitor (ERKi) treatment experiments. *Nanog:H2B-GFP*$^{Tg/+}$ blastocysts recovered at E2.75 and cultured for 36 hr in KSOM medium (I), KSOM + FGF2 (II), and ERKi (III), followed by staining for Hoechst, NANOG, and SOX17. White arrowheads identify NANOG^{+};SOX17^{-} ICM cells exhibiting high levels of GFP. Orange arrowheads identify SOX17^{+};NANOG^{-};GFP-low ICM cells.

(C) Reporter and NANOG distribution analysis in cells of cultured embryos shown in (B).

Note that x and y axis scales for scatterplots vary due to changes in reporter and NANOG expression between embryos cultured under different conditions. r = Pearson correlation coefficient. Scale bar represents 20 μm.

much smaller than the timescales of cell differentiation. Finally, by analyzing *Nanog* mutant embryos carrying the reporter, we confirmed that no functional NANOG protein was produced from the BAC transgene (Figure 2D). We therefore conclude that the *Nanog:H2B-GFP* reporter faithfully marks pluripotent cells emerging within the ICM and provides an accurate quantitative readout of *Nanog* expression.

Distribution of NANOG Expression in the Blastocyst Is Altered by Modulation of Fibroblast Growth Factor Signaling

Next, we investigated the distribution of the *Nanog:H2B-GFP* transcriptional reporter, as well as NANOG protein, as the EPI compartment emerges within the ICM of late blastocysts. We performed quantitative immunofluorescence analysis on mid-to-late blastocysts (90–110 cells), stained with Hoechst, NANOG, and the PrE marker SOX17 (Figure 3A). We noted a bimodal distribution of reporter expression within the ICM, which corresponded to bimodality in NANOG distribution within prospective EPI and PrE cells (Figures 3A and S3A–S3D).

We investigated how these distributions could be modulated by perturbation of fibroblast growth factor (FGF) signaling, which plays a critical role in ICM lineage choice. Pathway inhibition results in an ICM composed exclusively of EPI precursors, while incubation of embryos in exogenous FGF results in an all-PrE ICM (Chazaud et al., 2006; Kang et al., 2013; Nichols et al., 2009; Yamanaka et al., 2010). Consistent with these studies, we observed that the bimodal distribution of GFP and NANOG levels within the

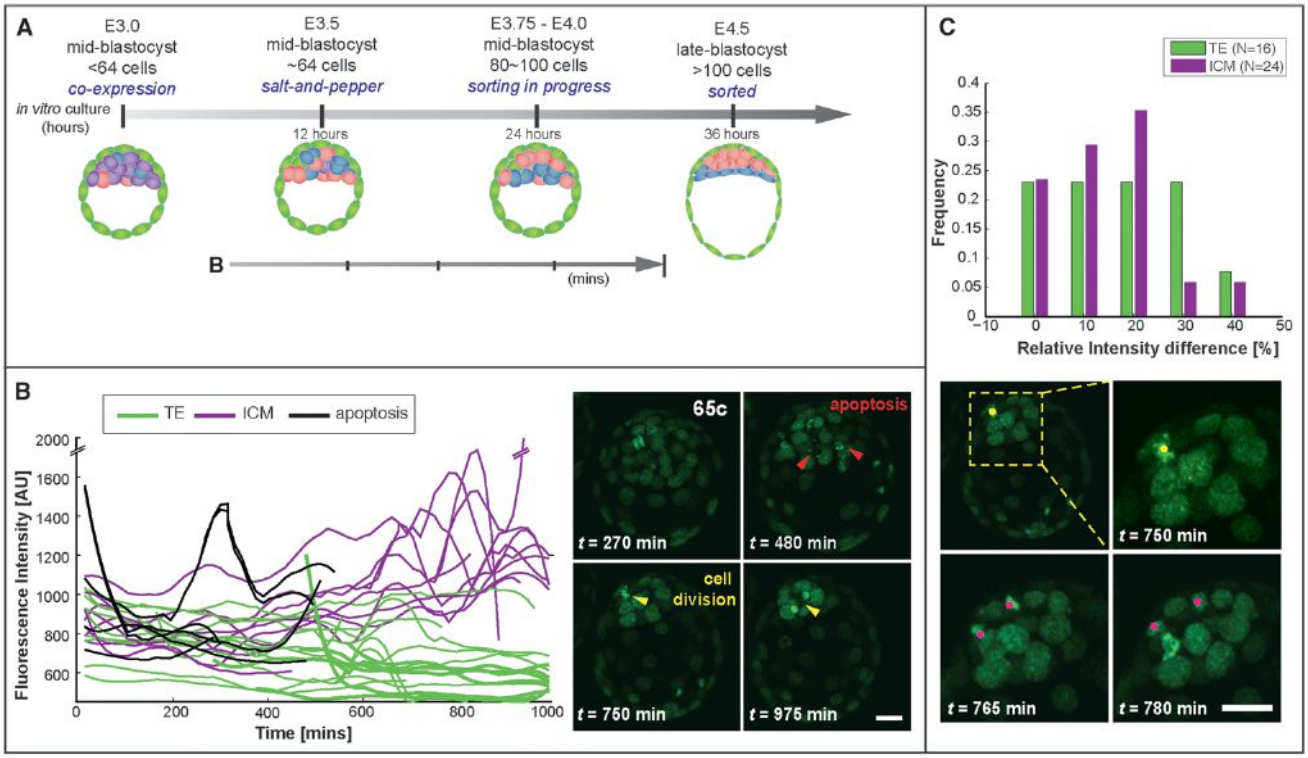


Figure 4. EPI Cells Emerge In Vivo with an Accompanying Increase in *Nanog* Expression

(A) Schematic of mouse blastocyst stage embryo development.

(B) Quantification of GFP intensity in individual nuclei of a living *Nanog:H2B-GFP*$^{Tg/+}$ blastocyst recovered at E3.5 (65 cells) and corresponding images of single time points. Single-cell intensity traces were obtained by tracking single cells in time-lapse experiments. Mitotic and apoptotic cells were also tracked, resulting in abrupt peaks in fluorescence intensity. Cells belonging to the ICM are depicted in purple, while TE cells are depicted in green; cells having undergone apoptosis are depicted as black lines. The developmental timing and the time spanned by tracks analyzed in this panel are illustrated by an arrow in (A).

(C) Top panel, quantification of relative GFP intensity in daughters/sisters versus mother cells upon division in cells of TE (green bars) or ICM (purple bars) origin. Daughter cells retained expression levels after mitosis, with most changes within 20%.

Data were obtained from the analysis of time-lapse movies of five or more embryos. Bottom panel, cell divisions of EPI progenitors occurring at late blastocyst stages shown in (B). Red arrowheads identify apoptotic events, and yellow arrowheads mark cell divisions. Scale bar represents 20 μm.

ICM was lost when FGF signaling activity was modulated (Figures 3B, 3C, and S3E–S3J). We therefore conclude that a heterogeneous distribution of NANOG expression exists in vivo. This is represented as a stable bimodal distribution within the ICM and reflects cell fate specification toward PrE and EPI lineages, and that this distribution is altered by modulation of FGF signaling.

Highly Variable *Nanog* Dynamics Are Initially Observed Followed by the Establishment of Differential Expression at the Onset of ICM Lineage Specification

A bimodal distribution of *Nanog* expression is established as the pluripotent EPI emerges within the ICM. However, we currently do not know whether fluctuating expression of pluripotency-associated factors, such as *Nanog*, occurs during EPI specification, nor, if they do exist, whether such fluctuations correlate or predict state reversions (Smith, 2013). To address this question, we investigated the behavior of individual ICM cells in *Nanog:H2B-GFP*$^{Tg/+}$ blastocysts using time-lapse imaging coupled with quantitative image analyses performed in an accurate and automated fashion (Supplemental Experimental Procedures).

Using this methodology, we first observed that in early blastocysts (until ~60 cell stage) all cells (ICM and TE) displayed highly heterogeneous reporter activity (0–400 min in Figure 4B), consistent with previous observations (Dietrich and Hiiragi, 2007; Ohnishi et al., 2014; Plusa et al., 2008). Following this phase, TE cells downregulated reporter expression, whereas ICM cells retained or increased their expression, presumably coinciding with the establishment of a lineage bias (400 min to end of movie; Figure 4B). We also noted apoptotic events, occurring during the process of lineage specification in randomly positioned GFP-hi or GFP-low ICM cells (red arrowheads; Figure 4B). As described previously, apoptotic events occur during blastocyst development and are not a consequence of in vitro culture or phototoxicity (Artus et al., 2013; Plusa et al., 2008). Furthermore, we noted several cell divisions occurring in GFP-hi ICM cells (yellow arrowheads; Figure 4B). Heritability of *Nanog:H2B-GFP* levels, and likely cell fate, was observed in all ICM cells after division (Figure 4C). Of note, an increase in GFP fluorescence was routinely observed during mitosis, due to chromosome condensation; thus any resulting rapid oscillations (of the order of 150 min) in reporter activity were excluded from the analysis, as they were not

considered as reflecting changes in gene expression. By using a computational approach, we confirmed that GFP expression was not diluted in dividing cells, suggesting that daughter cells inherited the lineage identity of their parental cell (Figure S4). Finally, cells within the TE displayed low, somewhat variable and continually decreasing levels of reporter activity (Figure 4B).

Next, we focused our analyses on subsequent phases of development, initiating from early-to-mid blastocyst stages (at around 50–60 cell) until late blastocyst stages (at around 100 cells), where the EPI and PrE cells have sorted to their final positions (Figures 5A and 5B; Movie S1). A subpopulation of ICM cells is specified to the EPI-lineage, with GFP expression maintained or increased. By contrast, cells biased toward the PrE lineage extinguished the reporter during differentiation (Figure 5A).

Infrequent Cell-State Reversals Occur toward, Not Away from, a Pluripotent Identity

Notably, we observed a few instances where cells in mid-stage blastocysts, exhibiting low levels of reporter activity, would rapidly increase reporter activity, and concomitantly become segregated with the EPI (Figure 5A). These could, in principle, represent rare PrE-to-EPI conversions. Importantly, we never observed events of EPI-to-PrE progenitor switching associated with downregulation of *Nanog* expression. To obtain an independent confirmation of the absence of EPI-to-PrE transitions, we analyzed time-lapse data from a PrE-specific single-cell resolution reporter (*Pdgfra*$^{H2B\text{-}GFP/+}$, Plusa et al., 2008). In contrast to *Nanog:H2B-GFP*, expression of the *Pdgfra*$^{H2B\text{-}GFP}$ reporter is activated only in PrE-biased cells after cell-fate specification. Therefore, PrE-to-EPI conversion would be detected as downregulation of the reporter. On the other hand, EPI-to-PrE conversion would give rise to significantly delayed activation of the reporter in cells lacking expression. We imaged embryos expressing the *Pdgfra*$^{H2B\text{-}GFP}$ reporter starting at the time of cell differentiation, i.e., the time when expression of the reporter becomes reliably detectable. We found that all positive cells at the onset of differentiation retained their expression and no new cells initiated expression with a significant delay, corresponding to an EPI-to-PrE conversion (Figure 5B). Collectively, our results with the *Nanog:H2B-GFP* and *Pdgfra*$^{H2B\text{-}GFP/+}$ reporters imply that cell-fate reversals are extremely rare and preferentially happen from the PrE to the EPI state.

The PrE-to-EPI unidirectionality would suggest that fate transitions are regulated and do not result from purely stochastic fluctuations in gene expression. To rule out that the inability to detect EPI-to-PrE transition in *Nanog:H2B-GFP* embryos was due to perdurance of GFP reporter, we performed the following analysis. First, we quantified the dynamics of downregulation of *Nanog:H2B-GFP* in PrE cells (Figure S6) and estimated the lifetime of H2B-GFP to be shorter than 4 hr, which is short enough to detect cell-fate reversal events (details in the Supplemental Experimental Procedures). Second, we analyzed time-lapse movies of the *Nanog:H2B-GFP* reporter in embryos treated with exogenous FGF (all ICM cells would downregulate NANOG) and showed that a clear downregulation can be observed (Figure S5). In addition, we calculated the half-life of H2B-GFP from time-lapse movies of *Nanog:H2B-GFP* in exogenous FGF or cycloheximide (CHX). The decay rates of H2B-GFP in ICM cells in the embryos under CHX or FGF treatments were around 6.5 and 5.5 hr, respectively (Figure S6). Collectively, these experiments allow us to conclude that the half-life of H2B-GFP is shorter than 6 hr and thus not dissimilar to the half-life of NANOG, which has been calculated to be approximately 4 hr (Abranches et al., 2013). Altogether, these results show that the *Nanog:H2B-GFP* reporter allows measuring *Nanog* transcriptional dynamics on timescales longer than 1 hr (see the Supplemental Experimental Procedures).

At Late Blastocyst Stages, EPI and PrE Do Not Exhibit Fluctuations

Next, we examined whether fluctuations in *Nanog:H2B-GFP*, and thus *Nanog*, levels were observed after ICM lineage specification. At these late blastocyst stages (>90–100 cells), the active sorting of EPI and PrE populations to adjacent tissue layers is evident. Our analyses revealed that the EPI cells exhibited increasing levels of reporter activity, whereas a decrease in GFP levels was observed in cells forming the emergent PrE epithelial layer (0 min to end of movie; Figure 5C; Movie S1). At the end of the specification period, apoptotic events were observed (red arrowhead; 0–100 min; Figure 5C). Thereafter, we did not observe any apoptosis. Instead, after ICM lineage specification, several cell divisions were observed in EPI progenitors (yellow arrowheads; Figure 5C). Collectively, these observations suggest a wave of temporarily restricted apoptotic events, occurring around the period of lineage specification, followed by a burst of EPI lineage-specific cell proliferation. *Nanog* levels remained stable after ICM cell-fate divergence suggesting that fluctuations between EPI and PrE states do not occur. Based on these data, we conclude that within the ICM the majority of pluripotent EPI and extra-embryonic PrE progenitor cells do not change their fate after specification. Furthermore, our data lead us to propose that a burst of cell proliferation following cell differentiation ensures EPI lineage-specific expansion.

Quantitative Analyses of *Nanog:H2B-GFP* Time-Lapse Movies Revealed a Correlation between Cell Behaviors and Cell-Fate Choice within the ICM

Time-lapse imaging in combination with cell tracking and quantitative analysis allowed us to determine whether fate reversals correlate with the spatial position of individual cells within the ICM. We simultaneously analyzed the position of a cell relative to the blastocyst cavity, and its *Nanog* reporter expression levels as a function of time. Analysis of the spatial distribution of cells converting from PrE-to-EPI indicated that the cells migrated toward the inner region of the ICM (Figure 6A, blue cell). The cell undergoing fate switching exhibited similar rate of GFP increase as in cells of the embryo under conditions of ERK inhibition (Figure 6B). This observation suggests that the fate switching might be a result of PrE-biased cells that stop responding to ERK signaling (due to their localization and fates of their neighbors) and, as a consequence, increase *Nanog* expression and change fate. Consistently, under conditions of ERK inhibition, all ICM cells committed to an EPI fate and displayed increasing reporter activity, whereas GFP expression in TE cells remained unaffected, consistent with our previous observations in fixed embryos (Figures 3B, 3C, and S3E–S3J). Notably, we did not

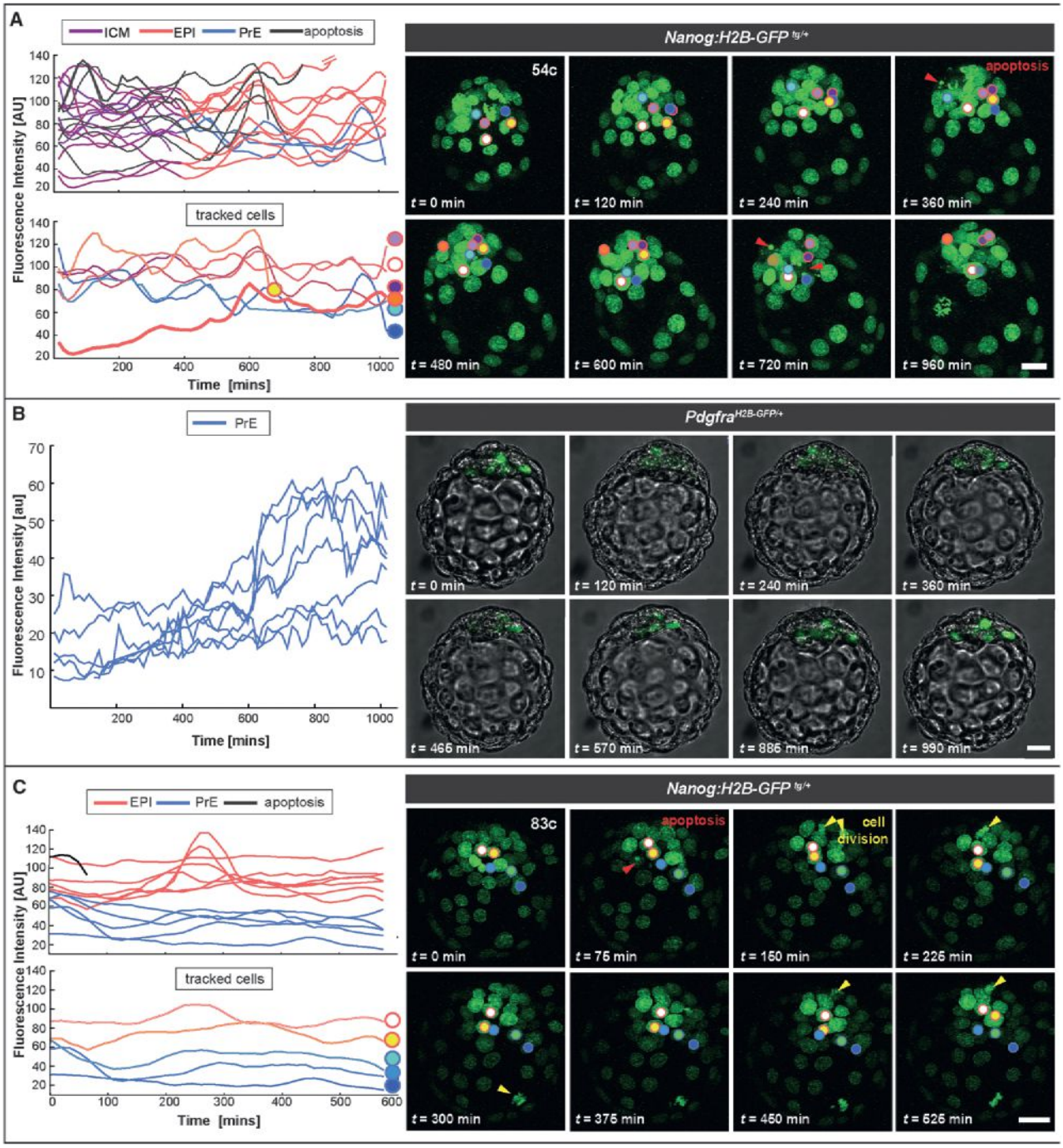


Figure 5. State Reversals Can Occur toward a Pluripotent Identity during Lineage Specification in Mid Blastocysts, but Cells Do Not Change Fate in Late Embryos

(A) Quantification of GFP intensity in single nuclei of a living *Nanog:H2B-GFP*$^{Tg/+}$ at E3.5 (54 cells) and corresponding snapshots from a time-lapse movie. A subset of tracks are detailed in the lower plot, corresponding to cells highlighted in the images in the panel on the right. Cell highlighted with an orange dot and red outline represents a GFP-low cell that upregulated reporter expression and contributed to EPI.

(B) Quantification of GFP intensity in single nuclei of a living *Pdgfra*$^{H2B\text{-}GFP/+}$ at E3.5 (~60 cells) and corresponding snapshot of fluorescence channel overlapped with bright-field images from a time-lapse movie. Only PrE-biased cells express H2B-GFP.

(C) The same analysis applied in (A) was repeated in a later *Nanog:H2B-GFP*$^{Tg/+}$ embryo at E3.75 (85 cells). ICM cells segregated toward EPI (red) and PrE (blue) lineages. Cells that underwent apoptosis depicted with black lines. Red arrowheads mark apoptotic events and yellow arrowheads mark cell divisions. Tracked nuclei are highlighted by dots; the outline of each dot depicts the lineage choice; red outline depicts EPI progenitor cells; blue outline depicts PrE progenitor cells. Scale bar represents 20 μm.

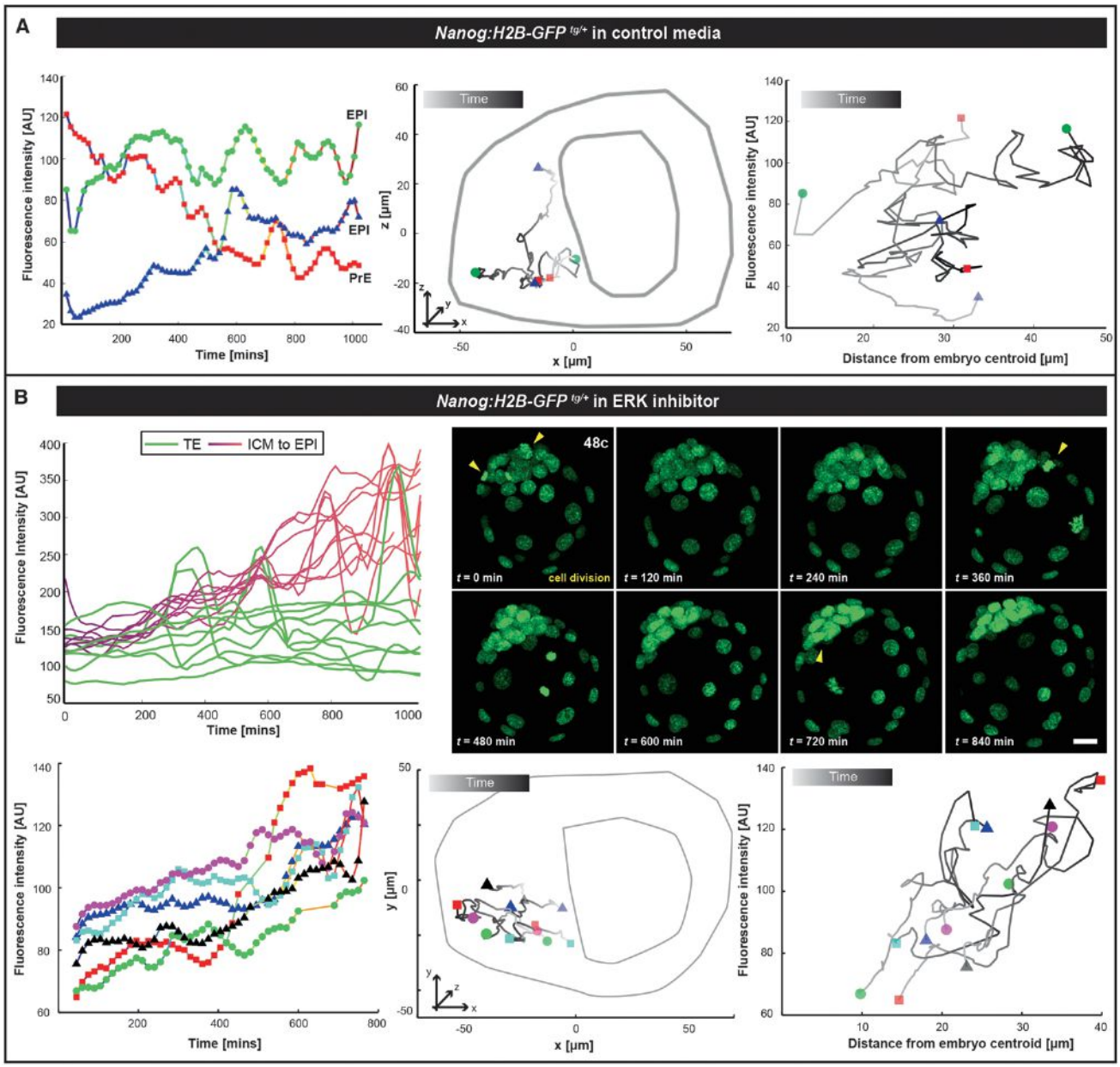


Figure 6. Cell-State Reversals toward a Pluripotent Identity Are Associated with Changes in Position within ICM

Fluorescence levels of single ICM cells (A, left, and B, lower-left panels) and two-dimensional projections of representative space trajectories (A, central, and B, lower-central panels) are plotted for the duration of time-lapse movies of *Nanog:H2B-GFP*$^{Tg/+}$ embryo (A) corresponding to embryo in Figure 5A and (B) in the presence of ERKi. The distance of each cell from the barycenter of the embryo is plotted versus GFP intensity (A, right, and B, lower right).

(A) EPI cells, green and blue; PrE cell, red trajectories. Blue trajectory represents a cell that switched state acquiring an EPI identity, moving from an initial position on the surface of ICM inward while increasing *Nanog* expression.

(B) Top, quantification of GFP fluorescence intensities and corresponding snapshots. Bottom, all six trajectories depict the behavior of ICM cells that acquire an EPI fate increasing reporter activity. All cells change their absolute position due to embryo growth, but there are no events of abrupt spatial change in position of individual cells within the cohort. Lines represent cell trajectories. Displayed trajectories for mitotic cells were excluded. Unique colors identify individual cells. The outline of the embryo is drawn for plotting relative position of nuclei. Shades of gray in the trajectory depict time: earlier (light) to later (darker).

observe any apoptotic events during ERK inhibition, whereas a burst of proliferation was evident in the all-EPI ICM (yellow arrowheads in Figure 6B and Movie S2). These findings support our previous observations, suggesting that apoptosis serves as a selection mechanism within the ICM, while the proliferation burst in EPI-committed cells ensures the rapid expansion of

the pluripotent lineage after its specification. In addition, we observed that there was no abrupt spatial position change of EPI cells within the ICM of ERK inhibitor-treated embryos; instead, cell motility was restrained, and only passive cell movement was observed as these embryos developed (Figure 6B).

DISCUSSION

Pluripotency-associated factors such as *Nanog* have been reported to exhibit dynamic fluctuations of expression in ESC cultures. Whether fluctuations in gene expression correlating with cell-state transitions occur in vivo remains an open question. Here, we have investigated *Nanog* expression heterogeneities using single-cell resolution *Nanog* transcriptional reporters coupled with 3D time-lapse imaging and high-resolution automated quantitative image analyses. Our data suggest that BAC-based *Nanog:H2B-GFP* transgenic reporters are expressed at physiological levels, exhibit minimal reporter perdurance, and serve as a faithful readout of *Nanog* expression both in vitro in ESCs, and in vivo in mouse embryos.

For both ESCs rendered transgenic through introduction of a *Nanog:H2B-GFP* construct, as well as transgenic embryo-derived ESCs, heterogeneities in the levels of reporter, as well as NANOG protein, were noted. A GFP-low cell population, which also displayed low levels of NANOG protein, was evident in cells that were propagated in MEF-free conditions, which promote a differentiation bias; this population displayed OCT4 expression (Figure 1A, 1B, and S1E), suggesting that it might consist of pluripotent cells that were actively differentiating and/or primed for differentiation. Importantly, this variability, and the presence of a GFP-low/NANOG-low population, in ESC cultures have also been recently reported with another *Nanog* transcriptional reporter (Abranches et al., 2013). These observations therefore call for caution in the choice of culture conditions for ESC propagation; hence, it was recently shown that random monoallelic gene expression could occur stochastically as ESCs differentiated, resulting in the acquisition of heterogeneities during the adaptation of cells to in vitro culture (Eckersley-Maslin et al., 2014; Gendrel et al., 2014).

Furthermore, it has been suggested that reporters targeted to the *Nanog* locus, which concomitantly ablate *Nanog* activity, might exhibit behaviors resulting from *Nanog* heterozygosity (Faddah et al., 2013; Filipczyk et al., 2013). We noted that GFP expression from the knockin/knockout TNGA reporter was elevated compared to a BAC-based *Nanog:GFP*$^{Tg/+}$ reporter (Figures S1A–S1C). We therefore investigated whether *Nanog* allele heterozygosity per se might result in increased NANOG expression, consistent with its reported auto-repressive activity (MacArthur et al., 2012; Navarro et al., 2012). However, we failed to observe any noticeable difference in reporter activity between *Nanog:H2B-GFP*$^{Tg/+}$ ESCs that harbored a two (wild-type) or one functional *Nanog* alleles (Figures S1J and S1K). Thus, our data suggest that the elevated GFP expression observed in the TNGA ESCs is not the result of *Nanog* heterozygosity, and could result from allele design. The development of a faithful EPI lineage-specific reporter providing a single-cell resolution quantitative readout of *Nanog* expression, coupled with high-resolution image data analyses allowed us to address a central open question pertaining the behavior of EPI cells in vivo. Critical for this type of analysis was a single-cell resolution live imaging reporter that was sufficiently bright for time-lapse image acquisition, but expressed at physiological levels, so that reporter perdurance would not mask downregulation in cells. Notably, our experience with destabilized fluorescent protein reporters reveals reduced levels of fluorescence, not amenable to the image analyses methodologies used in this study.

Our data reveal a range of cell behaviors before and after the pluripotent EPI population has been specified in vivo within the ICM (Figure 7). Prior to EPI versus PrE specification has occurred, apoptosis serves as a selective mechanism to ensure proper segregation of lineage progenitors. Rare fate reversal events likely occur whereby GFP-low/PrE progenitor cells convert to a GFP-hi/EPI progenitor state as cells migrate inward in the ICM. At this time, cell migration correlates with fate choice and is linked to the presence of a heterogeneous population of EPI and PrE progenitors. Progenitors could be sorting toward a niche comprising cells with a similar lineage bias. By contrast, in the presence of ERK inhibition, lineage choice is forced toward one direction (all-EPI) resulting in a homogenous ICM population and is accompanied by a lack of cell movement (Figure 6B). Finally, after lineage specification has occurred within the ICM, fluctuations between EPI and PrE progenitors were not observed. However, a burst of cell proliferation was observed in EPI progenitors, as they sorted to the interior of the ICM.

Rarely cells changed their state toward, but never away from, a pluripotent identity. By using a single-cell resolution reporter for the PrE lineage, we previously showed that PrE progenitor cells could downregulate reporter expression but could not confirm the fate reversal to EPI, as cells could not be followed after loss of the reporter (Plusa et al., 2008). Here, by using a single-cell resolution reporter of the EPI lineage, we directly visualized these transitions. Perhaps cells converting from PrE-to-EPI might cease responding to an FGF signal, as in the presence of an ERK-inhibitor, resulting in *Nanog* upregulation and establishment of pluripotency. These data agree with recent studies suggesting that pulsatile FGF signaling induces differential *Nanog* expression within ICM cells and drives *Nanog* mRNA degradation for rapid post-transcriptional control of pluripotency (Torres-Padilla and Chambers, 2014; Tan and Elowitz, 2014).

Importantly, we observed no fate reversals between EPI and PrE subsequent to their specification. This is in agreement with a recent mathematical model accounting for the dynamics of the regulatory network that controls ICM differentiation; simulations indicated that after specification cells would not change identity, and thus EPI and PrE states are not interchangeable (Bessonnard et al., 2014). In addition, we observed that after specification a burst of cell proliferation in the pluripotent compartment was evident. This might ensure adequate numbers of EPI progenitors available for subsequent development. Furthermore, our observations suggest that PrE progenitors exhibit increased plasticity, compared to EPI progenitors, consistent with recent studies suggesting that PrE progenitors have a broader developmental potential than their EPI counterparts (Grabarek et al., 2012), and observations reporting that

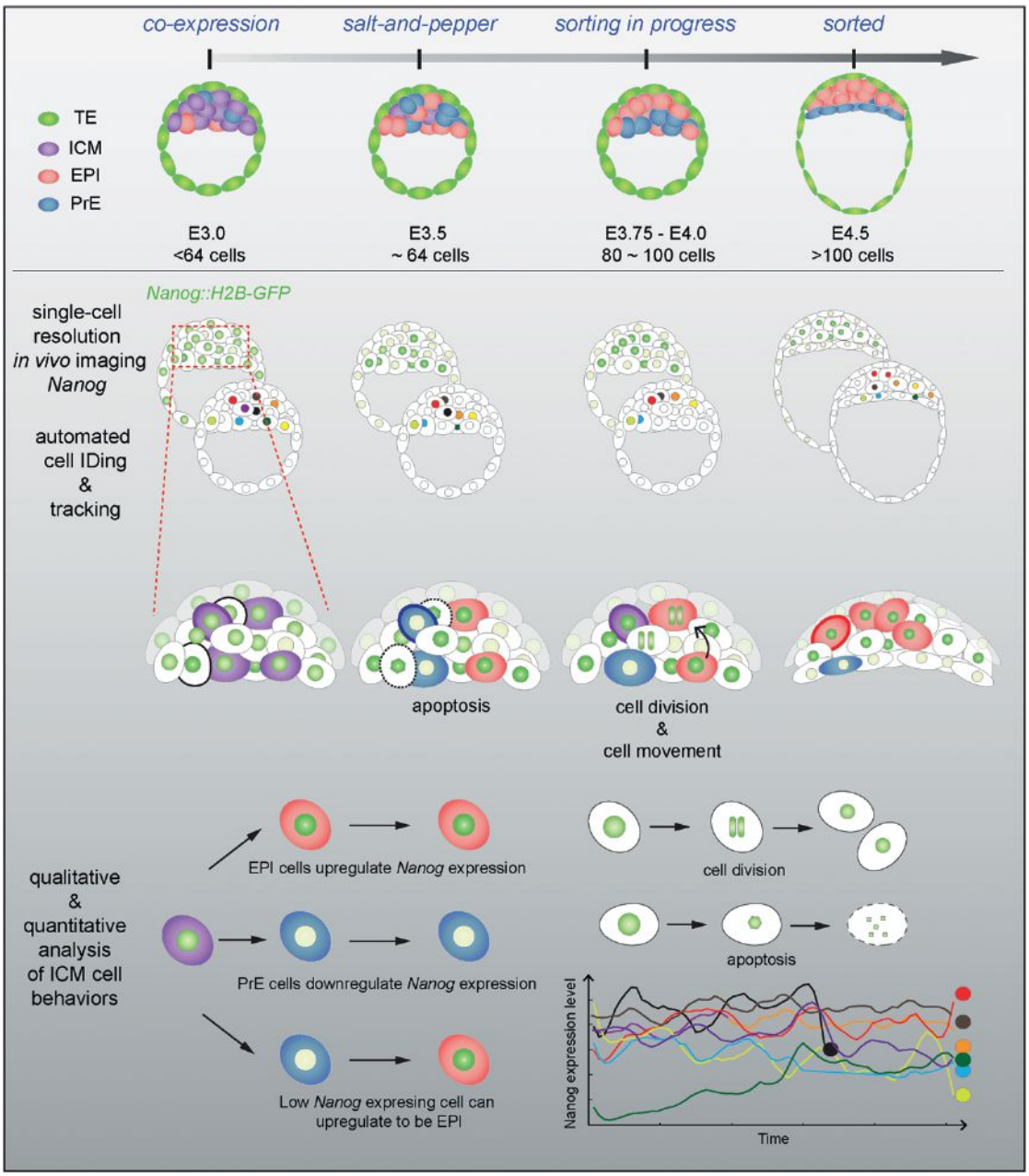


Figure 7. Cell Behaviors during the Emergence of the Pluripotent EPI Lineage

Schematic representation of embryo development from mid-to-late blastocyst. Prior to lineage specification (around 64–80 cells): (1) apoptosis occurs randomly in EPI or PrE progenitors as they segregate to their appropriate layers, (2) segregation is linked to spatial movements within the ICM, and (3) few GFP-low cells might acquire a pluripotent identity and migrate inward the ICM. After lineage specification (>80–90 cells): (1) cells do not fluctuate between EPI and PrE states, and (2) cell divisions occur in EPI-specified population while relocating to the interior of the ICM.

ESCs are primed toward endoderm co-express embryonic and extra-embryonic markers (Morgani et al., 2013).

Our data suggest that once specified in wild-type embryos, pluripotent cells do not, in general, change their fate. However, a very limited number of state reversals may occur at this time, but only toward a pluripotent identity. During development, sufficient numbers of lineage progenitors must be generated, and there is a time window during which cell-fate reversals can occur. Moreover, mechanisms involving apoptotic events and symmetric cell divisions may ensure that a pluripotent identity is protected and maintained in vivo. These observations within the embryonic environment appear contrasting with studies in ESCs. A possible explanation for this apparent disparity could be that ESCs in culture do not receive the appropriate inputs

from a niche, namely, neighboring cells and extra-cellular components. Another could be due to the differential timescales. ICM cells commit to PrE and EPI fates in less than 24 hr, and, since development proceeds unidirectionally, even though there may be non-differentiating (or symmetric) cell divisions, neither cell population arising within the ICM self-renews. By contrast, ESCs can be maintained indefinitely in vitro under conditions of self-renewal, a timescale that presents an unrestricted period for conversion between alternative states.

EXPERIMENTAL PROCEDURES

Mouse Husbandry

All animal experiments were approved by the Institutional Animal Care and Use Committee at the Memorial Sloan Kettering Cancer Center. Mice were maintained under a 12-hr light-dark cycle. Mouse lines used in this study were *Nanog:H2B-GFP*$^{Tg/+}$, *Nanog*$^{\beta\text{-}geo/+}$ (Mitsui et al., 2003), and *Nanog*$^{GFP/+}$ (Hatano et al., 2005). Alleles are schematized in Figure S1H.

Live Embryo Imaging

For live imaging, embryos were cultured in glass-bottomed dishes (MatTek) in an environmental chamber as done previously (Kang et al., 2013). Live imaging conditions used were compatible with normal development as shown previously (Plusa et al., 2008). For incubation experiments, an ERK1/ERK2 inhibitor, 1 μM PD0325901 (StemGent), was added to medium 2–3 hr prior to initiation of 3D time-lapse imaging. GFP was excited using a 488-nm Argon laser. Live image data were acquired using four laser scanning confocal imaging systems: Zeiss LSM510META, LSM710, LSM780, and Leica SP8. Images were acquired using 20×/0.75, 40×/1.3, or 63×/1.4 objectives. 20–30 xy planes separated by 2 μm were acquired per z stack, every 15 min. Movies of 3D time-lapse sequences were compiled and annotated using QuickTime Pro (Apple).

Immunostaining of ESCs and Embryos

Immunostaining of ESCs and embryos was performed as previously (Kalmar et al., 2009; Kang et al., 2013; Muñoz Descalzo et al., 2012). Primary antibodies used were CDX2 (1:100, Biogenex), GATA6 (1:100, R&D Systems), NANOG (1:500, Cosmo Bio), OCT4 (1:100, Santa Cruz Biotechnology), and SOX17 (1:100, R&D Systems). Secondary Alexa Fluor (Invitrogen) conjugated antibodies were used at 1:500. DNA was visualized with Hoechst 33342 (5 μg/ml; Invitrogen).

Quantitative Fluorescence Image Analysis

Quantitative fluorescence measurements from images of fixed ESCs and fixed or live embryos were performed using an automated image processing workflow comprising the segmentation of multiple nuclei in 3D data (Figure 2A). The front-end software, MINS, is a MATLAB (MathWorks)-based graphic user interface described previously (Lou et al., 2014; http://katlab-tools.org). We noted that for normalizing fluorescence values of each channel the Hoechst channel per cell was not optimal to compensate for loss of fluorescence intensity throughout the sample depth (differences in Hoechst fluorescence intensity throughout z stack indicated by orange arrowheads in Figure 2A). We thus developed an algorithm to generate a regression curve across values of the Hoechst channel for individual cells (details in the Supplemental Experimental Procedures). Fluorescence values for other channels were normalized using this curve (Figure 2A). To compare fluorescence values between ICM and TE cells, TE cells being analyzed needed to be in the same focal plane as ICM cells. Thus, TE cells positioned at the beginning and end of the z stack were excluded from the analysis. Only accurately segmented nuclei were included in analyses.

Nuclear Segmentation and Cell Tracking

Using the segmentation output of the MINS software, we developed an algorithm for cell tracking. To determine the position of any given cell at a subsequent time point, the algorithm first seeks to identify a clear, well-defined minimum in the distance between the cell centroid and the centroids of all cells in the subsequent frame. If such minimum does not exist, the algorithm uses cell segmentation to look for maximum overlap between cells at the subsequent time point and the cell being tracked. As additional criteria, changes in nuclear shape and overall distance are required to lie within control ranges. Such criteria generate unambiguous cell assignments and provide reliable tracking of a large fraction of cells (60%–70% in this study). The algorithm was validated by visual inspection, and each track was obtained by matching forward (prospective) and reverse (retrospective) tracking data (details provided in the Supplemental Experimental Procedures).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and two movies and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2015.02.010.

AUTHOR CONTRIBUTIONS

P.X. and A.-K.H. conceived the project and designed the experiments. P.X. generated the *Nanog:H2B-GFP* mouse line and performed ESC and fixed embryo imaging experiments. M.K. performed time-lapse embryo imaging experiments. A.P. and S.D.T. carried out quantitative analyses of experimental data sets. P.X., M.K., and A.-K.H. wrote the manuscript with input from A.P. and S.D.T.

ACKNOWLEDGMENTS

We thank X. Lou for developing the algorithm used to calculate regression curves for fluorescence intensity normalizations; J. Nichols for the *Nanog*$^{\beta\text{-}geo/+}$ mouse strain; A. Martinez-Arias for TNGA ESCs; S. Nowotschin, N. Saiz, and N. Schrode for discussions and comments on the manuscript; the Memorial Sloan Kettering Molecular Cytology and Rockefeller University Bio-Imaging Core Facilities for use of their instruments for live embryo imaging. Work in A.-K.H's. laboratory is supported by the NIH (R01-HD052115 and R01-DK084391) and NYSTEM (N13G-236). S.D.T. is supported by NIH (R00-HD074670). A.P. is supported by Finalized Research and Founding for Investments in Basic Research (RBAP11BYNP-Newton).

Received: July 14, 2014
Revised: January 4, 2015
Accepted: January 31, 2015
Published: March 5, 2015

REFERENCES

Abranches, E., Bekman, E., and Henrique, D. (2013). Generation and characterization of a novel mouse embryonic stem cell line with a dynamic reporter of Nanog expression. PLoS ONE *8*, e59928.

Artus, J., Kang, M., Cohen-Tannoudji, M., and Hadjantonakis, A.K. (2013). PDGF signaling is required for primitive endoderm cell survival in the inner cell mass of the mouse blastocyst. Stem Cells *31*, 1932–1941.

Bessonnard, S., De Mot, L., Gonze, D., Barriol, M., Dennis, C., Goldbeter, A., Dupont, G., and Chazaud, C. (2014). Gata6, Nanog and Erk signaling control cell fate in the inner cell mass through a tristable regulatory network. Development *141*, 3637–3648.

Boroviak, T., Loos, R., Bertone, P., Smith, A., and Nichols, J. (2014). The ability of inner-cell-mass cells to self-renew as embryonic stem cells is acquired following epiblast specification. Nat. Cell Biol. *16*, 516–528.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell *113*, 643–655.

Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., and Smith, A. (2007). Nanog safeguards pluripotency and mediates germline development. Nature *450*, 1230–1234.

Chazaud, C., Yamanaka, Y., Pawson, T., and Rossant, J. (2006). Early lineage segregation between epiblast and primitive endoderm in mouse blastocysts through the Grb2-MAPK pathway. Dev. Cell *10*, 615–624.

Czechanski, A., Byers, C., Greenstein, I., Schrode, N., Donahue, L.R., Hadjantonakis, A.K., and Reinholdt, L.G. (2014). Derivation and characterization of mouse embryonic stem cells from permissive and nonpermissive strains. Nat. Protoc. *9*, 559–574.

Dietrich, J.E., and Hiiragi, T. (2007). Stochastic patterning in the mouse preimplantation embryo. Development *134*, 4219–4231.

Eckersley-Maslin, M.A., Thybert, D., Bergmann, J.H., Marioni, J.C., Flicek, P., and Spector, D.L. (2014). Random monoallelic gene expression increases upon embryonic stem cell differentiation. Dev. Cell *28*, 351–365.

Faddah, D.A., Wang, H., Cheng, A.W., Katz, Y., Buganim, Y., and Jaenisch, R. (2013). Single-cell analysis reveals that expression of nanog is biallelic and equally variable as that of other pluripotency factors in mouse ESCs. Cell Stem Cell *13*, 23–29.

Filipczyk, A., Gkatzis, K., Fu, J., Hoppe, P.S., Lickert, H., Anastassiadis, K., and Schroeder, T. (2013). Biallelic expression of nanog protein in mouse embryonic stem cells. Cell Stem Cell *13*, 12–13.

Frankenberg, S., Gerbe, F., Bessonnard, S., Belville, C., Pouchin, P., Bardot, O., and Chazaud, C. (2011). Primitive endoderm differentiates via a three-step mechanism involving Nanog and RTK signaling. Dev. Cell *21*, 1005–1013.

Gendrel, A.V., Attia, M., Chen, C.J., Diabangouaya, P., Servant, N., Barillot, E., and Heard, E. (2014). Developmental dynamics and disease potential of random monoallelic gene expression. Dev. Cell *28*, 366–380.

Grabarek, J.B., Zyzyńska, K., Saiz, N., Piliszek, A., Frankenberg, S., Nichols, J., Hadjantonakis, A.K., and Plusa, B. (2012). Differential plasticity of epiblast and primitive endoderm precursors within the ICM of the early mouse embryo. Development *139*, 129–139.

Guo, G., Huss, M., Tong, G.Q., Wang, C., Li Sun, L., Clarke, N.D., and Robson, P. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev. Cell *18*, 675–685.

Hatano, S.Y., Tada, M., Kimura, H., Yamaguchi, S., Kono, T., Nakano, T., Suemori, H., Nakatsuji, N., and Tada, T. (2005). Pluripotential competence of cells associated with Nanog activity. Mech. Dev. *122*, 67–79.

Kalmar, T., Lim, C., Hayward, P., Muñoz-Descalzo, S., Nichols, J., Garcia-Ojalvo, J., and Martinez Arias, A. (2009). Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. PLoS Biol. *7*, e1000149.

Kang, M., Piliszek, A., Artus, J., and Hadjantonakis, A.K. (2013). FGF4 is required for lineage restriction and salt-and-pepper distribution of primitive endoderm factors but not their initial expression in the mouse. Development *140*, 267–279.

Kurimoto, K., Yabuta, Y., Ohinata, Y., Ono, Y., Uno, K.D., Yamada, R.G., Ueda, H.R., and Saitou, M. (2006). An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. Nucleic Acids Res. *34*, e42.

Lou, X., Kang, M., Xenopoulos, P., Muñoz-Descalzo, S., and Hadjantonakis, A.K. (2014). A rapid and efficient 2D/3D nuclear segmentation method for analysis of early mouse embryo and stem cell image data. Stem Cell Reports *2*, 382–397.

MacArthur, B.D., Sevilla, A., Lenz, M., Müller, F.J., Schuldt, B.M., Schuppert, A.A., Ridden, S.J., Stumpf, P.S., Fidalgo, M., Ma'ayan, A., et al. (2012). Nanog-dependent feedback loops regulate murine embryonic stem cell heterogeneity. Nat. Cell Biol. *14*, 1139–1147.

Messerschmidt, D.M., and Kemler, R. (2010). Nanog is required for primitive endoderm formation through a non-cell autonomous mechanism. Dev. Biol. *344*, 129–137.

Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003). The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. Cell *113*, 631–642.

Morgani, S.M., Canham, M.A., Nichols, J., Sharov, A.A., Migueles, R.P., Ko, M.S., and Brickman, J.M. (2013). Totipotent embryonic stem cells arise in ground-state culture conditions. Cell Rep. *3*, 1945–1957.

Muñoz Descalzo, S., Rué, P., Garcia-Ojalvo, J., and Martinez Arias, A. (2012). Correlations between the levels of Oct4 and Nanog as a signature for naïve pluripotency in mouse embryonic stem cells. Stem Cells *30*, 2683–2691.

Navarro, P., Festuccia, N., Colby, D., Gagliardi, A., Mullin, N.P., Zhang, W., Karwacki-Neisius, V., Osorno, R., Kelly, D., Robertson, M., and Chambers, I. (2012). OCT4/SOX2-independent Nanog autorepression modulates heterogeneous Nanog gene expression in mouse ES cells. EMBO J. *31*, 4547–4562.

Nichols, J., and Smith, A. (2012). Pluripotency in the embryo and in culture. Cold Spring Harb. Perspect. Biol. *4*, a008128.

Nichols, J., Silva, J., Roode, M., and Smith, A. (2009). Suppression of Erk signalling promotes ground state pluripotency in the mouse embryo. Development *136*, 3215–3222.

Ohnishi, Y., Huber, W., Tsumura, A., Kang, M., Xenopoulos, P., Kurimoto, K., Oleś, A.K., Araúzo-Bravo, M.J., Saitou, M., Hadjantonakis, A.K., and Hiiragi, T. (2014). Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. Nat. Cell Biol. *16*, 27–37.

Okita, K., Ichisaka, T., and Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. Nature *448*, 313–317.

Plusa, B., Piliszek, A., Frankenberg, S., Artus, J., and Hadjantonakis, A.K. (2008). Distinct sequential cell behaviours direct primitive endoderm formation in the mouse blastocyst. Development *135*, 3081–3091.

Schrode, N., Xenopoulos, P., Piliszek, A., Frankenberg, S., Plusa, B., and Hadjantonakis, A.K. (2013). Anatomy of a blastocyst: cell behaviors driving cell fate choice and morphogenesis in the early mouse embryo. Genesis *51*, 219–233.

Schrode, N., Saiz, N., Di Talia, S., and Hadjantonakis, A.K. (2014). GATA6 levels modulate primitive endoderm cell fate choice and timing in the mouse blastocyst. Dev. Cell *29*, 454–467.

Smith, A. (2013). Nanog heterogeneity: tilting at windmills? Cell Stem Cell *13*, 6–7.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.

Tan, F.E., and Elowitz, M.B. (2014). Brf1 posttranscriptionally regulates pluripotency and differentiation responses downstream of Erk MAP kinase. Proc. Natl. Acad. Sci. USA *111*, E1740–E1748.

Torres-Padilla, M.E., and Chambers, I. (2014). Transcription factor heterogeneity in pluripotent stem cells: a stochastic advantage. Development *141*, 2173–2181.

Xenopoulos, P., Kang, M., and Hadjantonakis, A.K. (2012). Cell lineage allocation within the inner cell mass of the mouse blastocyst. Results Probl. Cell Differ. *55*, 185–202.

Yamanaka, Y., Lanner, F., and Rossant, J. (2010). FGF signal-dependent segregation of primitive endoderm and epiblast in the mouse blastocyst. Development *137*, 715–724.

Ying, Q.L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519–523.

Inspire your peers
with Cell Reports

Breaking new ground? From molecular genetics to developmental neurobiology, *Cell Reports* publishes thought-provoking, cutting edge research spanning the spectrum of life sciences.

Cell Reports is an open access journal that offers the quality, rigor, and visibility you've come to expect from Cell Press.

Do you have a new biological insight? Submit your paper to *Cell Reports*.

For more information visit www.cell.com/cell-reports

An open access journal with impact.
From Cell Press.

Cell Reports

OPEN ACCESS
CellPress

Cell Reports
Article

Direct Activation of STING in the Tumor Microenvironment Leads to Potent and Systemic Tumor Regression and Immunity

Leticia Corrales,[1,3] Laura Hix Glickman,[2,3] Sarah M. McWhirter,[2] David B. Kanne,[2] Kelsey E. Sivick,[2] George E. Katibah,[2] Seng-Ryong Woo,[1] Edward Lemmens,[2] Tamara Banda,[2] Justin J. Leong,[2] Ken Metchette,[2] Thomas W. Dubensky, Jr.,[2,4,*] and Thomas F. Gajewski[1,4,*]

[1]Department of Pathology, The University of Chicago, 929 E57th Street GCIS 3H, Chicago, IL 60637, USA
[2]Aduro BioTech, Inc., 626 Bancroft Way, 3C, Berkeley, CA 94710, USA
[3]Co-first author
[4]Co-senior author
*Correspondence: tdubensky@aduro.com (T.W.D.), tgajewsk@medicine.bsd.uchicago.edu (T.F.G.)
http://dx.doi.org/10.1016/j.celrep.2015.04.031

SUMMARY

Spontaneous tumor-initiated T cell priming is dependent on IFN-β production by tumor-resident dendritic cells. On the basis of recent observations indicating that IFN-β expression was dependent upon activation of the host STING pathway, we hypothesized that direct engagement of STING through intratumoral (IT) administration of specific agonists would result in effective anti-tumor therapy. After proof-of-principle studies using the mouse STING agonist DMXAA showed a potent therapeutic effect, we generated synthetic cyclic dinucleotide (CDN) derivatives that activated all human STING alleles as well as murine STING. IT injection of STING agonists induced profound regression of established tumors in mice and generated substantial systemic immune responses capable of rejecting distant metastases and providing long-lived immunologic memory. Synthetic CDNs have high translational potential as a cancer therapeutic.

INTRODUCTION

The responsiveness of tumors to immunotherapy depends, at least in part, on the immunophenotype of the tumor microenvironment (TME) (Gajewski et al., 2013). Substantial evidence indicates that tumor-infiltrating lymphocytes (TILs) are correlated with favorable prognosis in diverse malignancies (Galon et al., 2012) and predicts a positive clinical outcome in response to several immunotherapy strategies (Postow et al., 2012; Wolchok et al., 2013). Understanding the underlying mechanisms that promote spontaneous T cell infiltration is critical toward developing new therapeutic strategies that can be used to effectively promote an immune-responsive TME.

Innate immune sensing in the TME is a critical step in promoting spontaneous tumor-initiated T cell priming and subsequent TIL infiltration (Fuertes et al., 2011). Transcriptional profiling analyses of melanoma patients have revealed that tumors containing infiltrating activated T cells are characterized by a type I IFN transcriptional signature (Harlin et al., 2009). Studies in mice have demonstrated that type I IFN signaling plays a critical role in tumor-initiated T cell priming (Diamond et al., 2011; Fuertes et al., 2011). Mice lacking the IFN-α/β receptor in DCs cannot reject immunogenic tumors, and CD8α$^+$ DCs from these mice are defective in antigen cross-presentation to CD8$^+$ T cells. Furthermore, *Baft3*$^{-/-}$ mice that lack the CD8α$^+$ DC lineage lose the capacity to spontaneously prime tumor-specific CD8$^+$ T cells (Fuertes et al., 2011; Hildner et al., 2008). These findings in humans and in mice indicate that the tumor-resident antigen-presenting cell (APC) compartment is defective in non-T-cell-inflamed tumors. Thus, strategies to induce type I IFN signaling and APC activation in the TME to bridge the innate and adaptive immune responses may have therapeutic utility.

Recent work has demonstrated that activation of the STING pathway in tumor-resident host APCs is required for induction of a spontaneous CD8$^+$ T cell response against tumor-derived antigens in vivo (Woo et al., 2014). In addition, activation of this pathway and the subsequent production of IFN-β contributes to the anti-tumor effect of radiation (Deng et al., 2014), which can be potentiated with co-administration of a natural STING agonist. STING (stimulator of interferon genes, also known as TMEM173, MITA, ERIS, and MPYS) is a transmembrane protein localized to the ER that undergoes a conformational change in response to direct binding of cyclic dinucleotides (CDNs), resulting in a downstream signaling cascade involving TBK1 activation, IRF-3 phosphorylation, and production of IFN-β and other cytokines (Burdette et al., 2011; Burdette and Vance, 2013; Ishikawa and Barber, 2008). IFN-β is the signature cytokine induced in response to activating STING, by either exogenous CDNs produced by bacterial infection or through binding of a structurally distinct endogenous CDN produced by a host cyclic GMP-AMP synthetase (cGAS) in response to sensing cytosolic double-stranded DNA (dsDNA) (Ablasser et al., 2013; Diner et al., 2013; McWhirter et al., 2009; Sun et al., 2013; Woodward et al., 2010; Zhang et al., 2013). These observations suggested

that direct activation of the STING pathway in the TME by intratumoral (IT) injection of specific agonists might be an effective therapeutic strategy to promote broad tumor-initiated T cell priming against an individual's tumor antigen repertoire.

To test this therapeutic approach, we began with 5,6-dimethylxanthenone-4-acetic acid (DMXAA), a defined flavonoid compound known as a vascular disrupting agent that was shown to have anti-tumor activity in mouse models (Baguley and Ching, 1997). This drug ultimately failed in humans when combined with standard-of-care chemotherapy in a phase 3 efficacy trial in non-small-cell lung cancer (Lara et al., 2011). Interestingly, recent structure-function studies of mouse STING (mSTING) and human STING (hSTING) demonstrated that DMXAA is a direct ligand for mSTING (Conlon et al., 2013; Gao et al., 2013a; Kim et al., 2013; Prantner et al., 2012). However, detailed analysis revealed that polymorphisms in hSTING rendered it unable to bind DMXAA, therefore abrogating its activity in human cells. These findings provide a mechanistic insight for the lack of DMXAA efficacy in humans as well as the rationale for the development of new pharmacologic compounds that potently activate hSTING. Whereas STING agonists are being developed as vaccine adjuvants (Dubensky et al., 2013; Ebensen et al., 2011; Gray et al., 2012), whether STING agonists could have direct anti-tumor therapeutic effects has been under-explored and the lack of defined agonists that could activate all known hSTING alleles has been lacking.

In the current report, we confirm that DMXAA is a strong agonist of the mSTING pathway in vitro and in vivo. We show that IT injection of DMXAA effectively primes $CD8^+$ T cell responses to promote rejection of established tumors in a STING-dependent fashion. Based on these proof-of-concept results, we synthesized a large panel of CDNs and selected compounds capable of activating all known hSTING alleles. Unlike DMXAA, selected compounds indeed stimulated human PBMCs to produce IFN-β. Like DMXAA, these STING agonists exhibit significant anti-tumor efficacy in several mouse tumor models, without significant local or systemic toxicity. Strikingly, direct IT injection of selected CDNs into established B16 melanoma, CT26 colon, and 4T1 breast carcinomas resulted in rapid and profound tumor regression and promoted lasting systemic antigen-specific T cell immunity. These effects were entirely STING-dependent and resulted in regression of non-injected tumors in the same hosts. We selected dithio-(R_P, R_P)-[cyclic[A(2′,5′)pA(3′,5′)p]], (ML RR-S2 CDA) as the lead molecule for continued development. This agent has high translational potential as a therapeutic intervention strategy to induce activation of the TME in multiple tumor types, with the mechanistic goal of generating effective tumor-initiated $CD8^+$ T cell priming and lasting anti-tumor efficacy.

RESULTS

DMXAA Stimulates the STING Pathway In Vitro

We first confirmed that DMXAA was a functional agonist of the STING pathway using mouse macrophages in vitro. STING aggregation was assessed using $STING^{-/-}$ macrophages expressing mSTING-HA. Control macrophages presented a diffuse pattern of STING in the cytoplasm, but after 1 hr of incubation with DMXAA, approximately 60% of cells displayed aggregates of STING in perinuclear sites (Figure 1A). Downstream phosphorylation of TBK1 and IRF3 was observed, which was abolished in $STING^{-/-}$ cells (Figure 1B). This correlated with an increase in the apparent molecular weight of STING, which has been reported to be due to its phosphorylation (Konno et al., 2013). $STING^{-/-}$ macrophages reconstituted with mSTING-HA showed restored phosphorylation of TBK1 and IRF3. IFN-β secretion was detected from wild-type (WT), but not from $STING^{-/-}$, macrophages in response to DMXAA (Figure 1C). Similar results were observed with bone-marrow-derived DCs (BM-DC) from WT versus $STING^{-/-}$ mice (Figures S1A and S1B). We also used BM-DCs cells to study the expression of additional immunoregulatory molecules. IFN-β, IFN-α, TNF-α, IL-1β, IL-6, and IL12p40 were induced after stimulation with DMXAA in WT cells, but not $STING^{-/-}$ BM-DCs (Figure S1C). Whereas LPS induced expression of CD40, PD-L1, CD86, and MHC class II in both WT and STING-deficient DCs, induction with DMXAA was observed only in WT cells (Figures S1D and S1E). Together, these data along with previous studies (Conlon et al., 2013; Gao et al., 2013a; Kim et al., 2013; Prantner et al., 2012) confirm that DMXAA is a strong agonist of mSTING, resulting in the production of IFN-β and other innate cytokines and activation of DCs.

DMXAA Induces Strong Anti-tumor Immunity In Vivo

In order to evaluate whether stimulation of STING could augment anti-tumor immunity in vivo, we chose an IT route of administration to focus activation on those APCs acquiring tumor antigens. To assess an antigen-specific immune response, we utilized the B16 melanoma cell line transduced to express the model antigen SIYRYYGL (B16.SIY) (Blank et al., 2004). B16.SIY tumor cells were inoculated into the flank of WT or STING-deficient mice and injected IT with DMXAA at day 7. The dose of 500 μg of DMXAA was chosen after examining single doses ranging from 150 to 625 μg, with the highest dose of 625 μg showing unacceptable toxicity (data not shown). In WT animals, the selected dosage induced potent tumor regression and complete tumor rejection in the majority of mice; however, no reduction in tumor growth was observed in response to DMXAA in the absence of host STING (Figure 2A). Analysis of splenocytes 5 days after treatment showed a marked increase in the frequency of SIY-specific IFN-γ-producing T cells in WT mice, but not in $STING^{-/-}$ mice (Figure 2B). Similarly, treatment with DMXAA caused a high frequency of SIY-specific $CD8^+$ T cells detected by SIY/K^b pentamer staining in WT animals, but not in STING-deficient mice (Figure 2C). Next, we examined whether DMXAA treatment had any effect in animals deficient in the type I IFN receptor (IFNAR). A significant portion of the anti-tumor effect of DMXAA was lost in IFNAR-deficient mice, and none of the deficient animals showed complete tumor rejection (Figure 2D).

To determine whether immunologic memory was induced, WT mice that had rejected B16.SIY tumors were re-challenged 60 days after the initial inoculation with the same tumor cells. None of the re-challenged animals developed tumors (Figure 2E). We then investigated whether the anti-tumor immune response induced following DMXAA administration could be potent enough to reject non-injected secondary tumors. B16.SIY cells were injected in both flanks of mice, but only one tumor was

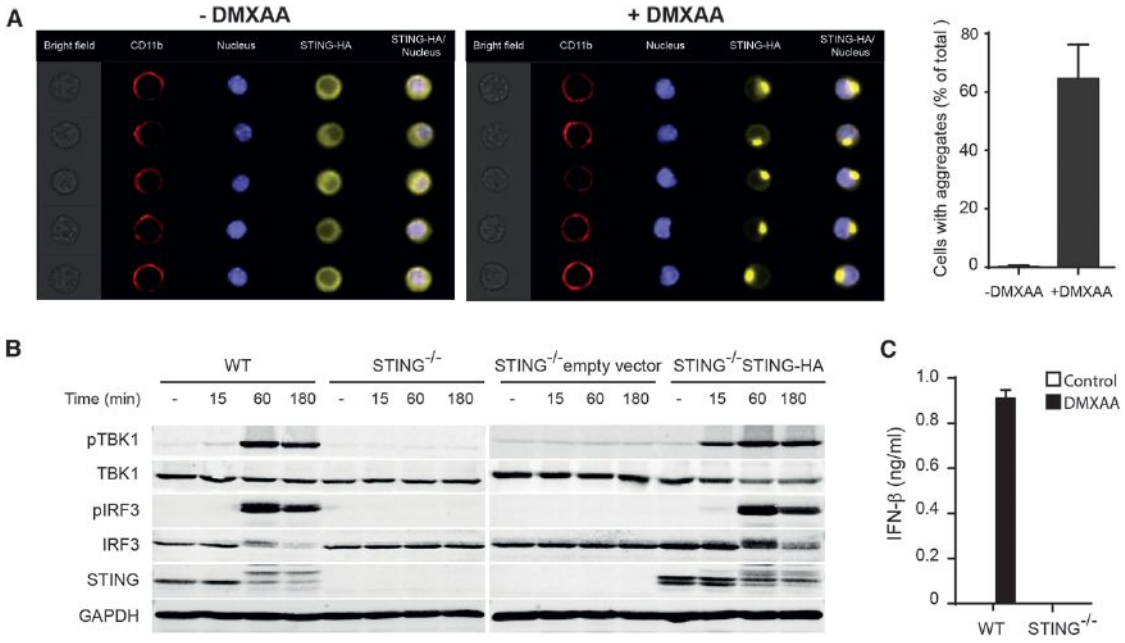


Figure 1. DMXAA Activates the STING Pathway and Promotes the Activation of APCs

(A) STING$^{-/-}$ mouse bone-marrow-derived macrophages (BMM) transduced to express STING-HA tag were stimulated for 1 hr with 50 μg/ml DMXAA and stained with specific antibodies against HA-tag, CD11b, and DAPI. Single cell images were acquired in the ImageStream, and data were analyzed with the IDEAS software (Amnis; Millipore). The data in the graph represent average of percentage of cells with aggregates from three independent experiments.

(B) WT or STING$^{-/-}$ BMM were stimulated with 50 μg/ml of DMXAA for the indicated time points. The amount of pTBK1, total TBK1, pIRF3, total IRF3, STING, and GAPDH was measured by western blot.

(C) WT or STING$^{-/-}$ BMM were stimulated with 50 μg/ml of DMXAA for 12 hr. The amount of secreted IFN-β was measured by ELISA.

treated with DMXAA. Tumor regression was observed in both sites (Figure 2F), suggesting that IT DMXAA administration can have a therapeutic effect on distant tumors. This effect was unlikely secondary to systemic distribution of the drug, because deliberate systemic administration of DMXAA via intraperitoneal administration had an inferior therapeutic effect (data not shown). These results demonstrate that a STING agonist can activate tumor-specific immune response capable of eliminating distal tumors and protecting from tumor challenge.

To assess whether the potent anti-tumor efficacy resulting from IT administration of DMXAA could be broadly applied, we tested several additional syngeneic tumor models. Treatment with DMXAA significantly reduced the growth of B16.F10 (without expression of SIY) and TRAMPC2 tumors in C57BL/6 mice, 4T1 tumors in BALB/c mice, and Ag104L tumors in C3H mice, indicating that the therapeutic effect of DMXAA is not restricted to a specific tumor histology or mouse genetic background (Figures S2A–S2D).

To determine whether the adaptive immune response was required for tumor control, B16.SIY cells were inoculated into RAG2$^{-/-}$ mice that lack mature T and B cells. DMXAA treatment lost most of its therapeutic effect in RAG2$^{-/-}$ hosts, although there was a partial control of tumor growth (Figure S2E). A similar loss of therapeutic effect was observed in TCR$\alpha^{-/-}$ mice (Figure S2F) and in mice depleted of CD8$^+$ T cells (Figure S2G), but not in mice depleted of CD4$^+$ T cells or NK cells (Figures S2H and S2I). These results indicate that a significant component of the therapeutic effect of DMXAA is mediated by CD8$^+$ T cells.

Identification of Synthetic hSTING-Activating Molecules

Having shown that the STING pathway could be harnessed to promote tumor antigen-specific CD8$^+$ T cell priming, leading to significant therapeutic efficacy, we sought to identify compounds that could potently activate hSTING and therefore be considered for clinical translation. CDNs have been studied as small-molecule second messengers synthesized by bacteria, which regulate diverse processes including motility and formation of biofilms (Römling et al., 2013). The immunogenicity of recombinant protein antigens can be augmented with CDNs used as an adjuvant, giving CDNs a potential application toward vaccine development (Dubensky et al., 2013; Ebensen et al., 2007a, 2007b, 2011; Gray et al., 2012). We sought to develop synthetic CDN compounds with increased activity in human cells as well as the ability to engage all known polymorphic hSTING molecules. The availability of CDN-STING crystal structures, along with recent results describing hSTING allele/CDN-dependent signaling relationships, facilitated structure-based studies to design CDN compounds with increased activity. We synthesized compounds that varied in purine nucleotide base, structure of the phosphate bridge linkage, and substitution of the non-bridging oxygen atoms at the phosphate bridge with sulfur atoms. Native CDN molecules are sensitive to degradation by phosphodiesterases that are present in host cells or in the

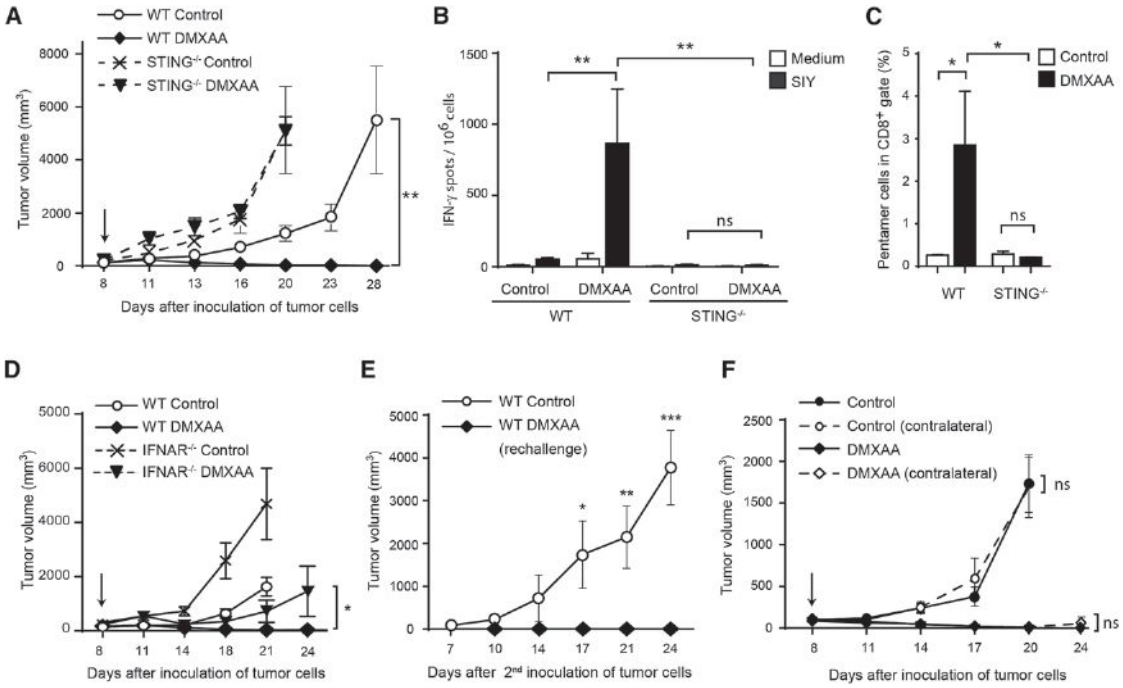


Figure 2. Rejection of Tumors in Response to DMXAA Is STING Dependent

(A) WT or STING$^{-/-}$ C57BL/6 mice were inoculated with 10^6 B16.SIY cells in the left flank. When tumor volumes were 100–200 mm^3, they received a single IT dose of 500 μg of DMXAA or saline. Tumor volume was measured at the indicated time points (n = 5).

(B and C) WT or STING$^{-/-}$ C57BL/6 mice (n = 5) were treated as in (A), and 5 days later, splenocytes were harvested and re-stimulated in vitro in the presence of culture medium or soluble SIY peptide for 16 hr. The frequency of tumor-specific IFN-γ-producing cells was assessed by ELISPOT (B), and the percentage of SIY-specific CD8$^+$ T cells was assessed by staining splenocytes with antibodies against TCRβ, CD4, CD8, and SIY pentamer (C). Cells were acquired in the LSRII-Blue cytometer and analyzed with FlowJo software. Results are shown as mean ± SEM. *p < 0.05; **p < 0.01.

(D) WT or IFNAR$^{-/-}$ C57BL/6 mice were inoculated with 10^6 B16.SIY cells in the left flank (n = 5). When tumor volumes were 100–200 mm^3, they received a single IT dose of 500 μg of DMXAA or saline. Tumor volume was measured at the indicated time points.

(E) WT mice that had rejected B16.SIY tumors were re-challenged with 10^6 B16.SIY in the contralateral flank. Naive mice were used as controls. Tumor size was measured at the indicated time points.

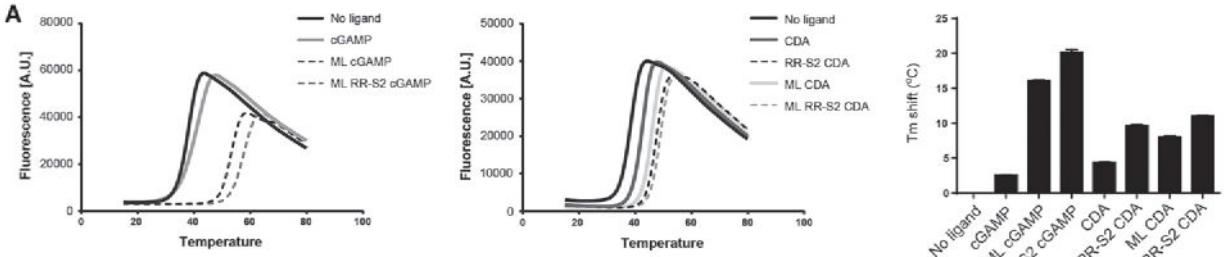
(F) WT mice were inoculated with 10^6 B16.SIY cells in the left and the right flanks (n = 5). When tumor volumes were 100–200 mm^3, 500 μg of DMXAA or saline was injected IT in the right flank only and tumor volume was measured at the indicated time points. Data are representative of at least three independent experiments or two independent experiments for the contralateral tumor model. Results are shown as mean tumor volume ± SEM. *p < 0.5; **p < 0.01; ***p < 0.001. ns, not significant.

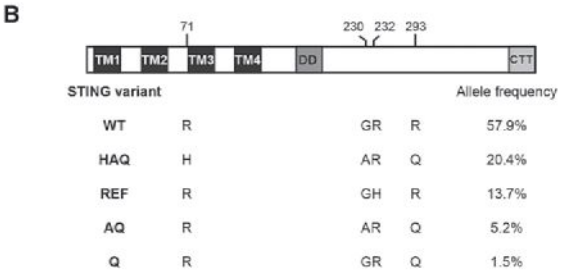
systemic circulation (Yan et al., 2008). We found that R_p, R_p (*R,R*) dithio-substituted diastereomer CDNs were both resistant to digestion with snake venom phosphodiesterase and induced higher expression of type I IFNs in human THP-1 cells compared to the R_p, S_p (*R,S*) dithio-substituted diastereomers or unmodified CDNs (data not shown).

To increase their affinity for STING, CDNs were also synthesized with a phosphate bridge configuration containing both 2′-5′ and 3′-5′ linkages, termed ''mixed linkage'' (ML), as found in endogenous human CDNs produced by cGAS (Ablasser et al., 2013; Diner et al., 2013; Gao et al., 2013b; Wolchok et al., 2013). The synthesis of dithio mixed-linkage CDNs, via modifications of literature procedures (Gaffney et al., 2010), resulted in both *R,R*- and *R,S* dithio diastereomers, which were purified and separated by a combination of silica gel and C18 reverse-phase prep-HPLC chromatography, affording CDNs with ≥95% purity as shown for ML RR-S2 CDA (Figure S3A, upper panel). The spectra for both ^{1}H NMR (data not shown) and the ^{31}P NMR (y axis of Figure S3A, upper panel) were consistent with ML RR-S2 CDA. Direct evidence for the regiochemistry of the phosphodiester linkages was obtained by ^{1}H-^{1}H COSY (correlation spectroscopy for assignment of ribose protons shown on x axis of Figure S3A, lower panel), in combination with a ^{1}H-^{31}P HMBC (heteronuclear multiple-bond correlation spectroscopy) 2D NMR (Figure S3A, lower panel). The 3D X-ray crystal structure of ML RR-S2 CDA confirms the presence of the 2′-5′, 3′-5′ mixed phosphodiester linkage and a dithio [R_P, R_P] diastereomer configuration (Figure S3B).

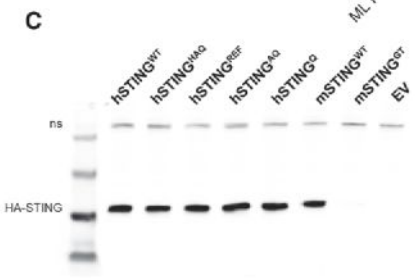
Synthetic CDNs Have Enhanced Binding Affinity to STING and Activate All Known hSTING Alleles

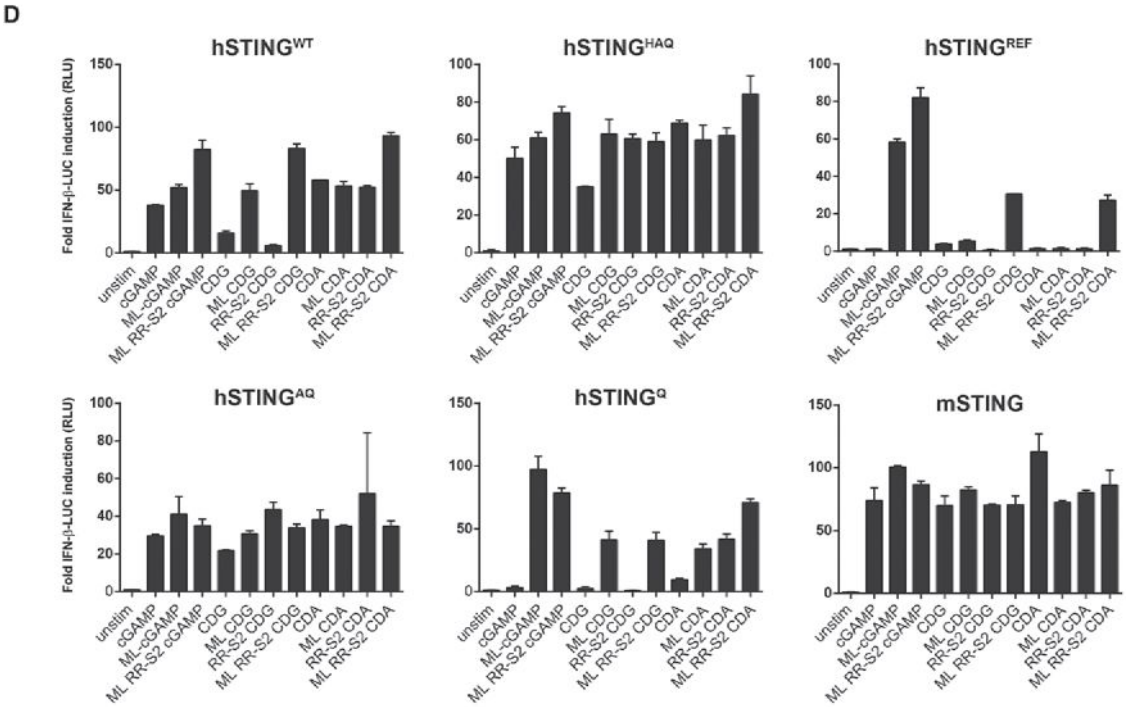
We purified recombinant hSTING and evaluated the relative binding affinity to various modified CDNs using differential scanning fluorimetry (DSF). DSF measures the stability of complex formation as a function of temperature as an indirect readout of protein-ligand association (Cavlar et al., 2013; Niesen et al., 2007). The increased shift in thermal stability of hSTING bound to ML RR-S2 CDA or ML RR-S2 cGAMP relative to unmodified

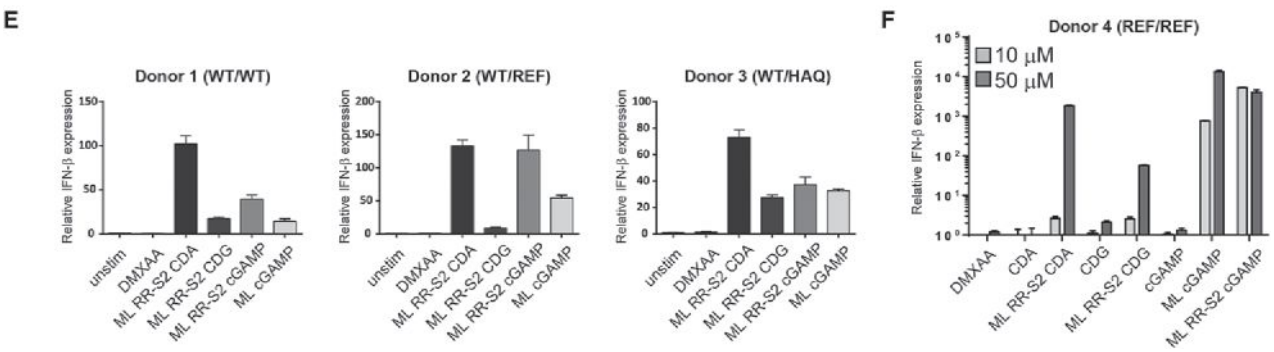




STING variant					Allele frequency
WT	R		GR	R	57.9%
HAQ	H		AR	Q	20.4%
REF	R		GH	R	13.7%
AQ	R		AR	Q	5.2%
Q	R		GR	Q	1.5%







(legend on next page)

CDNs indicates that R,R dithio and ML modifications enhance the binding affinity to STING (Figure 3A). Similar results were obtained using purified mSTING (Figure S3C). It has been shown recently that the bisphosphothionate analog of endogenous cGAMP (ML cGAMP) is resistant to hydrolysis by ENPP1 phosphodiesterase and thus is more potent at inducing IFN-β secretion in human THP1 cells (Li et al., 2014). Similarly, we found that R,R dithio-modified CDA compounds (ML RR-S2 CDA and RR-S2 CDA) showed enhanced type I IFN production over CDA in THP-1 human monocytes (Figure S3D).

SNPs in the hSTING gene have been shown to affect the responsiveness to bacterial-derived canonical CDNs (Diner et al., 2013; Gao et al., 2013b). Five haplotypes of hSTING have been identified (WT, REF, HAQ, AQ, and Q alleles), which vary at amino acid positions 71, 230, 232, and 293 (Figure 3B) (Jin et al., 2011; Yi et al., 2013). To test the responsiveness of the five hSTING variants to synthetic CDNs, we created stable HEK293T cell lines (defective in endogenous STING signaling) expressing each of the full-length hSTING variants. Similar levels of STING protein were expressed in each of the cell lines (Figure 3C). DMXAA potently activated mSTING and failed to activate any of the five hSTING alleles (Figure S3E), consistent with previous studies evaluating one of the hSTING alleles (Conlon et al., 2013). Cells expressing $hSTING^{REF}$ responded poorly to stimulation with the bacterial CDN compounds cGAMP, CDA, and CDG but were responsive to the endogenously produced cGAS product, ML cGAMP (Diner et al., 2013). Interestingly, the $hSTING^{Q}$ allele was also refractory to the bacterial CDNs. Cells expressing mSTING were responsive to all of the CDNs tested (Figure 3D). Cells transformed with either an empty vector or expressing a non-functional mutant (I199N) STING protein (*Goldenticket*; Sauer et al., 2011) were not responsive to any of the compounds (data not shown). In contrast, the dithio, mixed-linkage CDN derivatives (ML RR-CDA, ML RR-S2 CDG, and ML RR-S2 cGAMP) potently activated all five hSTING alleles, including the refractory $hSTING^{REF}$ and $hSTING^{Q}$ alleles (Figure 3D).

CDN Derivatives Potently Induce STING-Dependent Signaling in Murine and Human Immune Cells

To determine whether modified CDNs activated downstream STING signaling, we assessed murine bone marrow macrophages (BMMs) isolated from WT C57BL/6 and $STING^{-/-}$ (*Goldenticket*) mice (Sauer et al., 2011) for induction of IFN-β and other cytokines. Synthetic dithio mixed-linkage CDNs (ML RR-S2 CDA and ML RR-S2 CDG) induced the highest expression of IFN-β and the pro-inflammatory cytokines TNF-α, IL-6, and MCP-1 on a molar equivalent basis, as compared to endogenous ML cGAMP and the TLR3 agonist poly I:C (Figure S4A). The modified CDNs did not induce signaling in $STING^{-/-}$ BMMs, whereas, as expected, poly I:C agonists were still active. ML RR-S2 CDA was also found to induce aggregation of STING and induce phosphorylation of TBK1 and IRF3 in mouse BMM (Figures S4B and S4C). All of the modified CDNs tested also enhanced MHC class II and expression of co-stimulatory markers in a STING-dependent manner (Figure S4D).

To examine activation of STING signaling in primary human cells, we stimulated PBMCs from a panel of human donors harboring different STING alleles ($hSTING^{WT/WT}$, $hSTING^{WT/REF}$, and $hSTING^{WT/HAQ}$) and measured induction of IFN-β. In contrast to a lack of activation by DMXAA, dithio-modified ML CDNs induced IFN-β expression across these donors (Figure 3E). Importantly, dithio-modified ML CDNs, as well as the non-canonical cGAS product ML cGAMP, induced IFN-β expression in a donor homozygous for the $hSTING^{REF}$ allele, which was refractory to stimulation with canonical CDNs in HEK293T cells (Figures 3D and 3F). Looking at protein secretion by multiple donors that are homozygous for the $hSTING^{WT}$ allele, ML RR-S2 CDA induced significantly higher levels of IFN-α when compared to ML cGAMP (Figure S4E). Thus, ML RR-S2 CDNs are viable clinical candidates capable of activating STING pathway in human cells harboring different STING alleles and genotypes.

Intratumoral Delivery of Synthetic CDN Derivatives Results in Profound Anti-tumor Efficacy in Established B16 Melanoma

To evaluate whether modified dithio ML CDN compounds conferred increased anti-tumor activity, mice bearing established B16.F10 tumors were treated with three IT injections of CDN derivatives over a 1-week period. Whereas treatment with ML CDA and ML CDG had modestly reduced tumor growth, the R,R dithio derivatives profoundly inhibited tumor growth (Figure 4A). However, ML RR-S2 CDG was reactogenic, and these mice developed open wounds in the treated tumor that did not heal (data not shown). Lower dose levels of ML RR-S2 CDG were not efficacious (data not shown), indicating that this molecule had a narrow therapeutic index. In contrast, no injection site reactogenicity was observed with ML RR-S2 CDA and several mice developed vitiligo upon fur regrowth following complete eradication of the treated tumor (data not shown). Importantly,

Figure 3. Modified CDNs Potently Activate STING and Signal through All Human STING Allelic Variants

(A) Purified human STING binding to CDNs were analyzed by thermal shift assay. Temperature curves are the average from a representative experiment of three independent experiments performed in duplicate. Tm shift values are mean values ± SEM.

(B) Domain structure of hSTING is shown with the positions of the amino acid variations (bottom). The allelic frequencies of the hSTING isoforms shown on the left hand column were obtained from the 1000 Genome Project database as previously described (Yi et al., 2013).

(C) HEK293T cells were stably transfected with the indicated STING alleles. Whole-cell lysates from HEK293T cells stably expressing the indicated full-length STING-HA proteins were analyzed by western blot with anti-HA antibodies.

(D) HEK293T cells expressing the indicated STING alleles were transfected with an IFN-β-luciferase reporter construct. After 24 hr, cells were stimulated for 6 hr with the indicated CDN compound (10 μM) and assessed for IFN-β-reporter activity.

(E and F) Human PBMCs from donors with the indicated STING alleles were stimulated with 10 μM of the indicated CDN or 100 μg/ml DMXAA (E), or human PBMCs from a donor homozygous for the reference variant ($STING^{REF/REF}$) were stimulated with 10 μM and 50 μM of the indicated CDN or 100 μg/ml DMXAA (F). After a 6-hr stimulation, fold induction of IFN-β was measured by qRT-PCR and relative normalized expression was determined by comparison with untreated controls. Results are representative of at least two independent experiments.

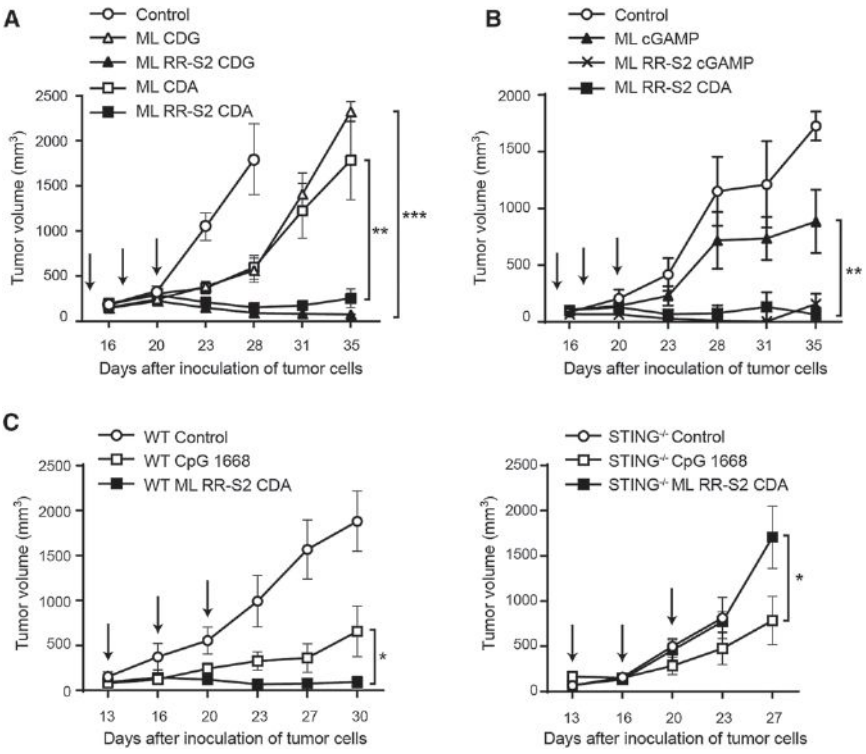


Figure 4. Synthetic CDN Modifications Significantly Improve Anti-tumor Efficacy in Established B16 Tumors

(A and B) WT C57BL/6 mice were inoculated with 5×10^4 B16.F10 cells in the left flank (n = 8). When tumor volumes were 100 mm^3, they received three 25 μg IT doses of ML CDA, ML CDG, ML RR-S2 CDG, ML RR-S2 CDA, or HBSS as control (A) or either three 50 μg doses of the endogenous cGAS product ML cGAMP, ML RR-S2 CDA, ML RR-S2 cGAMP, or HBSS as control (n = 8; B).

(C) WT C57BL/6 mice or STING$^{-/-}$ mice were treated with three IT doses of CDN ML RR-S2 CDA (50 μg), murine type B CpG ODN 1668 (50 μg), or HBSS vehicle control. Treatments were administered on the days indicated by the arrows, and tumor measurements were taken twice weekly. Data are representative of at least two independent experiments. Results are shown as mean tumor volume ± SEM. **p < 0.01; ***p < 0.001. ns, not significant.

these mice demonstrated significantly higher overall survival compared to mice treated with ML RR-S2 CDG (Figure S5A). ML RR-S2 CDA also showed higher anti-tumor control than the endogenous ML cGAMP (Figure 4B). Thus, ML RR-S2 CDA was selected for advancement to clinical development.

To determine whether the CDN-induced anti-tumor efficacy was STING-dependent, we compared activity in B16 tumor-bearing WT (C57BL/6) and STING$^{-/-}$ mice. CDN therapeutic efficacy was completely lost in STING$^{-/-}$ mice. In contrast, a CpG TLR9 agonist (Kawarada et al., 2001) did modestly reduce tumor growth in STING$^{-/-}$ mice, demonstrating that this mouse strain is capable of mounting an immune-mediated antitumor response (Figure 4C). A dose response of the ML RR-S2 CDA compound was performed in B16 tumor-bearing mice, which identified an optimal anti-tumor dose level that also elicited maximum tumor antigen-specific CD8$^+$ T cell responses (data not shown) and improved long-term survival to 50% (Figures S5B and S5C).

To evaluate which cell types are targeted by STING agonists in the TME, we analyzed the expression of IFN-β after ML RR-S2 CDA or DMXAA stimulation in BM-DC, bone marrow-derived macrophages, purified T cells from naive mice, B16 tumor cells, mouse embryonic fibroblasts (MEFs), and mouse primary dermal fibroblasts. Except for tumor cells, all cell types tested were found to express IFN-β. However, expression in WT BM-DCs was ten times higher than in the other cell types (Figure S5D). To determine whether this was also the case in the TME, we sorted four different cell populations from pre-established B16 tumors: DCs (CD45$^+$ CD11c$^+$ MHCII$^+$), macrophages (CD45$^+$ CD11b$^+$ F4/80$^+$ MHC-II$^+$), T cells (CD45$^+$ CD3$^+$), and endothelial cells (CD45$^-$ CD31$^+$). Sorted cells were then stimulated ex vivo with ML RR-S2 CDA or DMXAA. All subsets expressed IFN-β upon stimulation with the STING agonists. Expression in macrophages was highest, followed by DCs, which were both higher as compared with lymphocytes and endothelial cells (Figure S5E). These data demonstrate that the main source of type I IFN in the TME is likely APCs, although stromal cells and T cells might also contribute. Interestingly and in agreement with these data, we observed that the CD8α$^+$/CD103$^+$ DCs play a critical role in vivo, as the therapeutic effect of DMXAA was significantly diminished in *Batf3*$^{-/-}$ mice (Figure S5F). Altogether, these findings suggest that CD8α$^+$/CD103$^+$ DCs are critical for curative anti-tumor therapy by STING agonists.

ML RR-S2 CDA Induces Lasting Immune-Mediated Tumor Rejection in Multiple Tumor Types

To test anti-tumor efficacy in diverse tumor models, BALB/c mice bearing established 4T1 colon or CT26 mammary carcinomas were treated with ML RR-S2 CDA. All treated animals showed significant and durable tumor regression. Mice that were cured of their primary tumor were completely resistant to re-challenge when the same tumor cell line was used (Figures 5A and 5B). However, animals cured from CT26 tumors after ML RR-S2 CDA treatment showed no protection when they were re-challenged with the 4T1 tumor cells, demonstrating specificity (Figure 5B). Increased T cell responses were observed against the endogenous CT26 rejection antigen AH1 (Slansky et al., 2000) (Figure 5C). IT injection of ML RR-S2 CDA into one tumor in BALB/c mice bearing bilateral CT26 or 4T1 tumors also demonstrated significant regression of the contralateral, untreated tumor (Figures 5D and S5G). Using a different model to study the distal effect of ML RR-S2 CDA, we implanted B16.F10 melanoma in C57BL/6 mice and 7 days later intravenously infused B16.F10 melanoma cells to generate lung

metastases. The 2-week-old established flank tumors were treated with ML RR-S2 CDA, DMXAA, or HBSS control, and 2 weeks later, lung metastases were enumerated. Mice treated in the flank tumor with ML RR-S2 CDA showed substantial control of distant lung metastases (Figure 5E). Together, these results demonstrate that IT injection with ML RR-S2 CDA eradicates multiple tumor types and primes an effective systemic $CD8^+$ T cell immune response that significantly inhibits the growth of distal, untreated lesions.

DISCUSSION

Our results indicate that IT administration of STING agonists generates a potent anti-tumor immune T cell response and striking durable disease regression in multiple mouse tumor models. Although DMXAA has a potent therapeutic effect in mice, it lacks the ability to activate hSTING. The chemically modified CDN compounds described here have the capacity to activate all human polymorphic STING molecules while retaining the ability to engage mSTING and demonstrate a similarly impressive anti-tumor effect. The participation of the adaptive immune response for the observed anti-tumor responses is supported by the secondary rejection of non-injected tumors, the clearance of lung metastases, and long-term immunologic memory observed against autologous tumor re-challenge.

The mechanism of the therapeutic effect observed with the compounds tested in this study was absolutely dependent on host STING, and the majority of the anti-tumor effect was dependent upon T cells, specifically $CD8^+$ T cells, as described (Jassar et al., 2005; Wallace et al., 2007). However, a partial therapeutic effect was observed in $RAG^{-/-}$ and $TCR\alpha^{-/-}$ mice, indicating that innate immune cells are an important contributor to anti-tumor efficacy. Our in vitro data show that the various synthetic STING agonists tested induced IFN-β production by APCs via a mechanism that depended on the classical STING-TBK1-IRF3 signaling pathway (Ishikawa et al., 2009). In addition, STING agonists induced production of other cytokines, DC maturation, and also chemokine production in vitro. Previous work characterizing the vascular disrupting property of DMXAA demonstrated induction of TNF-α by stromal cells in vivo (Joseph et al., 1999) and attenuation of the therapeutic effect in $TNFR^{-/-}$ mice (Zhao et al., 2002). Thus, the T-cell-independent component of the therapeutic effect of STING agonists in vivo may be mediated through early TNF-α-mediated tumor vascular destruction. In addition, it seems likely that the induction of chemokines by STING agonists may also contribute to effective migration of activated T cells into the TME within the injected tumor site. This multi-faceted mechanism of action may explain the therapeutic potency of STING agonists against a range of cancers in vivo.

We developed synthetic CDN-derivative molecules based on rationally designed STING structure-function relationships. The lead molecule ML RR-S2 CDA has several features that improve both stability and lipophilicity, promoting significantly increased STING signaling as compared to endogenous and pathogen-derived CDNs. Whereas canonical CDNs have been evaluated as vaccine adjuvants and were recently shown to inhibit growth of 4T-1 tumors when given by intraperitoneal injection (Chandra et al., 2014), those investigations used canonical CDNs that may not be appropriate for clinical development, because there are hSTING alleles at significant frequencies in the population that are refractory to these structures. In the present study, we show the profound anti-tumor efficacy resulting from activation of the STING pathway with synthetic CDNs that not only activate all known hSTING alleles but have significantly higher potency than the natural STING ligands generated by cGAS. Although some human donors showed higher responsiveness to ML cGAMP, such as the donor bearing the refractory reference allele represented in Figure 3F, overall ML RR-S2 CDA activated all known hSTING allelic variants. This compound will therefore be attractive for clinical development.

A possible limitation of the treatment approach described herein is the necessity for IT injection to achieve maximal therapeutic effect. However, a practical advantage of this strategy is that it has the potential to generate T cell responses against tumor-specific antigens expressed by a patient's individual cancer. The attractiveness of IT injection approaches has been rekindled based on several recent clinical trial observations. IT injection of the oncolytic virus T-VEC has been shown in a randomized trial to provide improved clinical activity in melanoma patients compared with control (Goins et al., 2014). In addition, Levy and colleagues have shown that IT injection of the TLR9 agonist CpG along with local low-dose radiation therapy had clinical activity in patients with non-Hodgkin's lymphoma (Brody et al., 2010). Both of these studies demonstrated regression of non-injected lesions, consistent with the induction of tumor-specific T cells that could promote regression of tumors at distant sites. These observations, along with the impressive potency of STING agonists preclinically, support the development of clinical strategies for IT injection of STING agonists as a cancer therapeutic in patients.

EXPERIMENTAL PROCEDURES

Cells and Cell Isolations

The cells used for the in vivo experiments were: the C57BL/6-derived melanoma cell lines B16.F10 and B16.F10.SIY (henceforth referred to as B16.SIY), the breast cancer 4T1 cell line, and the colon cancer CT26 cell lines, all originally purchased from ATCC. All cells were maintained at 37°C with 5% CO_2 in DMEM supplemented with 10% heat-inactivated FCS, penicillin, streptomycin, L-arginine, L-glutamine, folic acid, and L-asparagine.

Immortalized WT and $STING^{-/-}$ macrophages were obtained as described in Roberson and Walker (1988). The WT macrophages were obtained from Dr. K. Fitzgerald (University of Massachusetts). Non-immortalized macrophages were derived from the bone marrow of WT (C57BL/6) or $STING^{-/-}$ mice and cultured in BMM media (RPMI media with 5% CSF, 5% FBS, 1× L-glutamine, and 1× pen/strep) for 7 days prior to use. Human PBMCs were isolated by density-gradient centrifugation using Ficoll-Paque Plus (GE Healthcare).

For stable overexpression of HA-STING in $STING^{-/-}$ macrophages, the full-length mSting-HA DNA sequence was generated by PCR. Sequence encoding full-length mSTING was amplified from pUNOI-mSTING plasmid (Invivogen) using a 5′ primer containing an EcoRI site, gcagacGAATTCATGCCATACTC CAACCTGCATCCAGCCATCCCACGGCCCAGAGGTCACCGCTCCAAATAT GTAGCCCTCATCTTTCTGGTGGCCAG, and a 3′ primer containing the HA-tag nucleotide sequence, followed by a TGA stop codon and a NotI site, tca catGCGGCCGCTCAGGCGTAGTCAGGCACGTCGTAAGGATAGATGAGGTC AGTGCGGAGTGGGAGAGGCTGATCC. The mSTING-HA PCR product was gel purified and double digested with EcoRI and NotI and then cloned into the multiple cloning site of pMXS-IRES-GFP with Quick ligation kit (New England Biolabs).

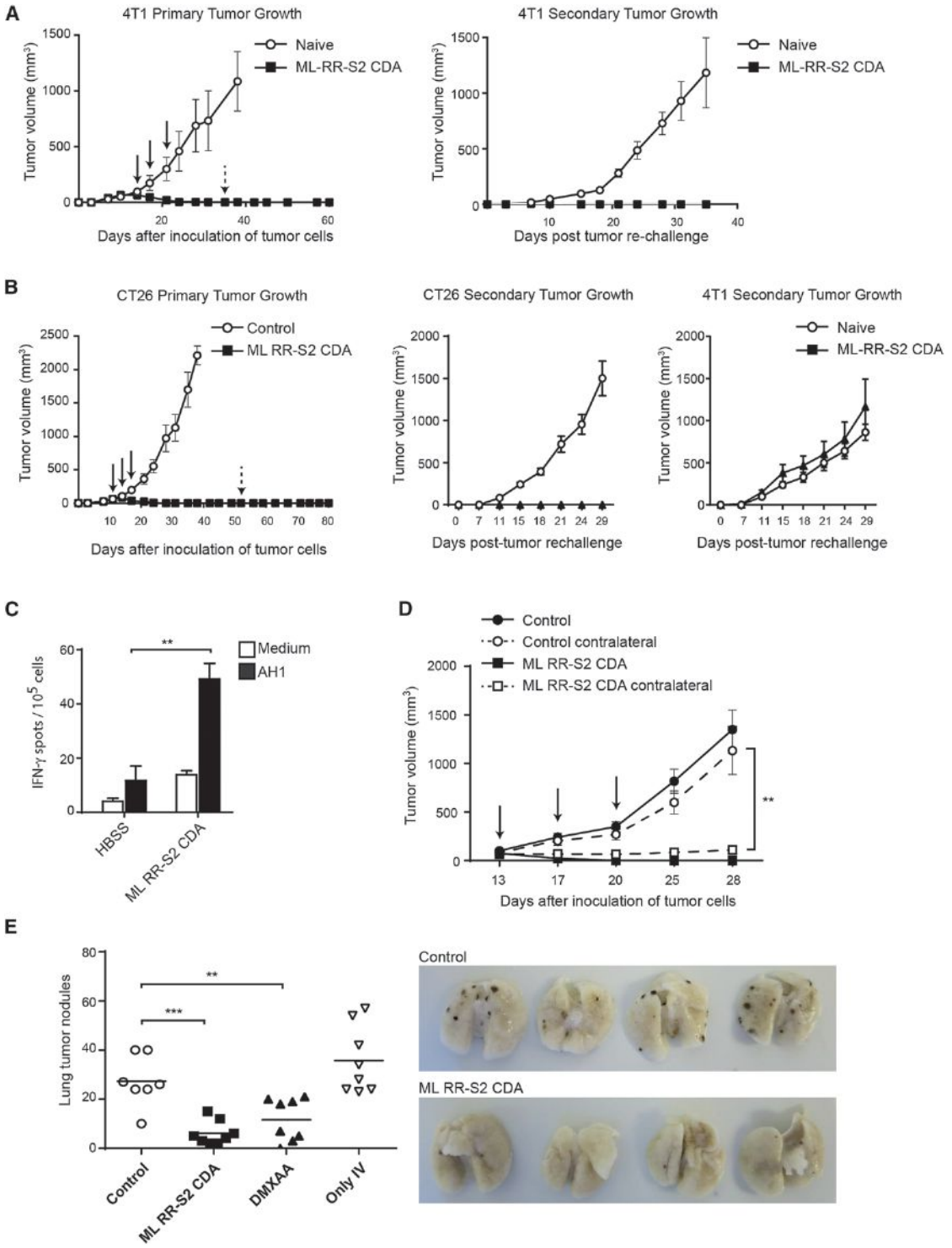


Figure 5. ML RR-S2 CDA Promotes Immune-Mediated Tumor Rejection

(A and B) WT BALB/c mice were inoculated with 10^5 4T-1 breast cancer (A) or CT26 (B) colon carcinoma cells in the left flank. When tumor volumes were 100 mm^3, they received three 50 μg doses IT of ML RR-S2 CDA or HBSS vehicle control (left graph). Mice were re-implanted with 10^5 4T1 (A and B) or CT26 (B) tumor cells on the opposite flank on day 55 post-initial tumor implantation. Naive mice were used as controls (right graph; n = 8).

(C) WT BALB/c mice were inoculated with 10^5 CT26 colon carcinoma cells in the left flank and treated on days 11, 14, and 18 with IT injections of ML RR-S2 CDA (25 μg each) or HBSS vehicle control (n = 4). 21 days post-implantation of CT26 tumors, PBMCs were stimulated with AH1 ($gp70_{423-431}$) and assessed by IFN-γ ELISPOT assay.

(legend continued on next page)

Stable HEK293T STING-expressing cell lines were generated with MSCV2.2 retroviral plasmids that contain STING cDNA cloned upstream of an IRES in frame with GFP. hSTING(REF)-HA, hSTING(WT)-HA, hSTING(HAQ)-HA, hSTING(Q)-HA, and mSTING(WT)-HA retroviral plasmids were obtained from the Vance Laboratory at UC Berkeley. hSTING(AQ)-HA was derived from hSTING(Q)-HA using a QuickChange Site-Directed Mutagenesis kit (Stratagene). Retroviral vectors were transfected into the amphotropic Phoenix packaging cell line using Lipofectamine (Invitrogen). After 2 days, viral supernatants were harvested and used for transduction of STING$^{-/-}$ macrophages or HEK293T cells. GFP^{+} cells were sorted in FACSAria (BD Bioscience) or MoFlow cell sorters.

Protein Expression and Purification

STING ligand-binding domains (human amino acids [aas] 140–379; mouse aas 139–378) were cloned via ligation-independent cloning into a custom pET-based vector containing a 6xHIS-SUMO-tobacco etch virus protease (TEV) site. All plasmids were confirmed by sequencing. Proteins were expressed in BL21 DE3 Rosetta 2 cells (EMD Millipore). Cells were grown in LB media at 37°C until an OD600 of about 0.6. Cells were then shifted to 18°C, induced with 0.25 mM isopropyl-beta-D-thiogalactopyranoside, and grown for 18–20 hr. Fusion proteins were purified on Ni-NTA agarose (QIAGEN). The 6xHIS-SUMO tag was removed by digestion with TEV protease (Sigma Aldrich) overnight during dialysis against buffer containing 20 mM Tris-HCl, 150 mM NaCl, 5 mM imidazole, 10% glycerol, and 0.5 mM Tris (2-carboxyethyl) phosphine (pH 7.5). After dialysis, the 6xHIS-SUMO tag and TEV protease were removed by Ni-NTA agarose. Proteins were concentrated to between 9 and 13 mg/ml. Aliquots were flash frozen in liquid nitrogen and stored at −80°C.

ImageStream Analysis of STING Aggregation in Cells

STING$^{-/-}$ macrophages overexpressing STING-HA tag were stimulated for 1 hr with 50 μg/ml of DMXAA resuspended in 7.5% of $NaHCO_3$, 50 μM of ML RR-S2 CDA resuspended in HBSS, or only the vehicles as control. After the incubation, cells were stained with anti-CD11b-APC (M1/70; BioLegend), rabbit anti-HA-tag (C29F4; Cell Signaling) and anti-rabbit IgG-PE (Invitrogen), and DAPI (Invitrogen). Single cell images were acquired in the ImageStreamx Mark II (Amnis), and data were analyzed using IDEAS software.

Western Blot Analysis

WT, STING$^{-/-}$ macrophages, and STING$^{-/-}$ macrophages overexpressing STING-HA or an empty vector were stimulated with 50 μg/ml DMXAA for 0, 15, 60, or 180 min; BM-DCs from WT or STING$^{-/-}$ mice were stimulated with 25 μg/ml DMXAA for the same time points. Proteins were extracted with Triton-X100 buffer (150 mM sodium chloride, 50 mM Tris, 1% Triton-X100 [pH 8.0]) with proteinase inhibitors (Thermo Scientific) and phosphatase inhibitors (Sigma). 30 μg of protein was electrophoresed in 10% SDS-PAGE gels and transferred onto Immobilon-FL membranes (Millipore). Blots were incubated with antibodies specific for phosphorylated TBK1 (Ser172), phosphorylated IRF3 (Ser396), total TBK1, STING, and GAPDH (Cell Signaling) or total IRF3 (Invitrogen). Proteins from HEK293T lines stably expressing STING were extracted with M-PER (Thermo Scientific). 6 μg of protein was loaded onto a 4–12% MES NuPAGE gel (Life Technologies), transferred to nitrocellulose, and probed with anti-HA antibody (Santa Cruz Biotechnology). Anti-rabbit IRDye 680RD label secondary antibody was used for visualization of bands with the Odyssey Imaging system (LI-COR Biosciences).

Differential Scanning Fluorimetry

Thermal shift assays were performed as in Cavlar et al. (2013). Assays were conducted with STING ligand binding domain at 1 mg/ml with or without various CDNs at 1 mM in 20 mM Tris-HCl, 150 mM NaCl (pH 7.5), and 1:500 dilution of SYPRO Orange Dye (Life Technologies). The fluorescence as a function of temperature was recorded in a CFX 96 real-time PCR machine (Bio-Rad) reading on the HEX channel EX 450–490 EM 560–580 nm. The temperature gradient was from 15°C to 80°C, ramping 0.5°C per 15 s. Curves were fit to a Boltzmann sigmoidal (Graph Pad Prism) to establish the midpoint of thermal unfolding (Tm).

Murine IFN-β ELISA

WT or STING$^{-/-}$ macrophages and BM-DCs from WT or STING$^{-/-}$ mice were stimulated with 50 μg/ml DMXAA. Conditioned media were collected after 4 hr. IFN-β concentration was assessed using VeriKine Mouse Interferon Beta ELISA Kit (PBL interferon source).

qRT-PCR Analysis of Cytokines

BM-DCs from WT or STING$^{-/-}$ mice were stimulated with 25 μg/ml DMXAA or 100 ng/ml LPS for 4 hr. Total RNA was isolated using the RNeasy kit (QIAGEN) and incubated with DNase I, Amplification Grade (Invitrogen). cDNA was synthesized using High Capacity cDNA Reverse Transcription Kit (Applied Biosystems), expression of cytokines was measured by real-time qRT-PCR using specific primers/probes for mouse INF-β, TNF-α, IL-6, and IL12p40, and pan-specific primers were to quantify expression of the IFN-α family. Primer sequences are listed in Table S1 in Supplemental Information. PCR reactions were performed in the 7300 Real Time PCR system (Applied Biosystems). The results are expressed as $2^{-\Delta Ct}$ using 18S as endogenous control.

WT BMM was stimulated with CDN at 5 μM in HBSS with the addition of Effectene (QIAGEN) transfection reagent (per kit protocol). Human PBMCs were stimulated as indicated. Stimulated cells were assessed by real-time qRT-PCR for gene expression of IFN-β1, MCP-1, TNF-α, and IL-6 using the PrimePCR RNA purification, and cDNA analysis system, and run on the CFX96 gene cycler (Bio-Rad). Relative normalized expression was determined by comparing induced target gene expression to unstimulated controls, using the reference genes Gapdh and Ywhaz (mouse) and GUSB and PGK1 (human), genes confirmed to have a coefficient variable (CV) below 0.5 and M value below 1, and thus did not vary with different treatment conditions.

Mice

All animals were used according to protocols approved by Institutional Animal Use Committee of the University of Chicago and Aduro Biotech and maintained in pathogen-free conditions in a barrier facility. C57BL/6, BALB/c, C3H/He, and TCRα$^{-/-}$ mice were obtained from Jackson and Charles River. RAG2$^{-/-}$ mice were obtained from Taconic. Tmem173$^{-/-}$ (STING-deficient) mice were provided by Dr. G. Barber (University of Miami), and STING$^{-/-}$ (*Goldenticket*; Sauer et al., 2011) and IFNAR$^{-/-}$ mice were purchased from Jackson Laboratories. Batf3$^{-/-}$ mice were provided by Dr. Kenneth M. Murphy (Washington University School of Medicine).

In Vivo Tumor Experiments

10^6 of B16-SIY tumor cells, 5×10^4 B16.F10 tumor cells, 10^5 4T1 and CT26, or 10^6 other tumor cells were injected s.c. in 100 μl DPBS or HBSS on the right flank of mice. Following tumor implantation, mice were randomized into treatment groups. When tumors were 100–200 mm^3 in volume (5–7 mm wide), either one single or three doses of DMXAA resuspended in 7.5% of $NaHCO_3$ or CDNs formulated in HBSS or vehicle control were injected IT. Measurements of tumors were performed twice per week using calipers, and the tumor volume was calculated with the formula $V = (\text{length} \times \text{width}^2)/2$. In some experiments, tumor-free survivors were re-challenged with tumor cells on the opposite flank several weeks after the injection of the primary tumor. Naive mice

(D) WT BALB/c mice were implanted with 10^5 of CT26 tumor cells on both flanks. On the days indicated, mice were treated in one flank only with ML RR-S2 CDA (50 μg) or HBSS vehicle control (n = 8).

(E) WT C57BL/6 were inoculated with 5×10^4 B16.F10 melanoma cells on the right flank at day 0 and implanted i.v. with 10^5 cells on day 7. Naive mice were implanted with cells i.v. only as a control. Flank tumors were treated on the days indicated with ML RR-S2 CDA (50 μg), DMXAA (150 μg), or HBSS control (n = 8). On day 28, lungs were harvested and lung tumor nodules counted. The histogram depicts total numbers of nodules in the ML RR-S2 CDA, DMXAA, or HBSS-control-treated mice, compared to the untreated i.v.-only tumor implanted mice. The images depict the ML RR-S2 CDA and HBSS-control-treated mice. Data are representative of at least two independent experiments. Results are shown as mean ± SEM. **$p < 0.01$; ***$p < 0.001$.

were used as controls. For the contralateral experiments, mice were implanted on both flanks and only one tumor was treated. For the B16 melanoma lung metastasis experiments, mice were implanted on the flank with 5×10^4 cells B16.F10 on day 0 and then injected intravenously with 1×10^5 cells on day 7. Lungs were harvested on day 28. Administration of compounds, measurements of tumors, and counting of lung tumors were performed in a blinded fashion.

IFN-γ ELISPOT and SIY-Pentamer Staining

Splenocytes were analyzed 5 days after the first IT injection of DMXAA or ML RR-S2 CDA. For the ELISPOTs, 10^6 splenocytes were plated per well and stimulated overnight with SIY peptide (160 nM) or AH1 (1 μM) peptide, with PMA (50 ng/ml) plus ionomycin (0.5 μM) as a positive control or medium as negative control. Spots were developed using the BD mouse IFN-γ kit according to the manufacturer's instructions, and the number of spots was measured using an Immunospot Series 3 Analyzer and analyzed using ImmunoSpot software (Cellular Technology). For SIY-pentamer staining, splenocytes were preincubated for 15 min with anti-CD16/32 monoclonal antibody (93) to block potential nonspecific binding and labeled with PE-MHC class I pentamer (Proimmune) consisting of murine H-2K^b complexed to SIYRYYGL (SIY) peptide, anti-TCRβ-AF700 (H57-597), anti-CD8-Pacific Blue (53-6.7), anti-CD4-Pacific Orange (RM4-5) (all antibodies from BioLegend), and the Fixable Viability Dye eFluor 450 (eBioscience). Stained cells were analyzed using LSR II cytometer with FACSDiva software (BD Biosciences). Data analysis was conducted with FlowJo software (Tree Star).

Luciferase Assay

10^4 HEK293T cells were seeded in 96-well plates and transiently transfected (Lipofectamine 2000) with human IFN-β firefly reporter plasmid (Fitzgerald et al., 2003) and TK-*Renilla* luciferase reporter for normalization. The following day, cells were stimulated with 10 μM of each CDN or 100 μg/ml DMXAA using digitonin permeabilization (50 mM HEPES, 100 mM KCL, 3 mM $MgCl_2$, 0.1 mM DTT, 85 mM sucrose, 0.2% BSA, 1 mM ATP, 0.1 mM GTP, and 10 μg/ml digitonin) to ensure uniform uptake. After 20 min, stimulation mixtures were removed and normal media was added. After a total of 6 hr, cell lysates were prepared and reporter gene activity measured using the Dual Luciferase Assay System (Promega) on a Spectramax M3 luminometer.

Statistical Analysis

Student's paired t test was used to calculate two-tailed p values to estimate statistical significance of differences between two treatment groups using Prism 6 software. The number of mice per group and the statistically significant p values are labeled in the figures and/or legends with asterisks.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2015.04.031.

AUTHOR CONTRIBUTIONS

L.C., L.H.G., T.W.D., and T.F.G. conceived the research and conducted the experiments with S.M.M., D.B.K., K.E.S., E.L., J.J.L., G.E.K., T.B., S.-R.W., and K.M. T.W.D., and T.F.G. conceived and supervised the entire project and wrote and revised the manuscript with L.C., L.H.G., K.E.S., and S.M.M.

ACKNOWLEDGMENTS

We thank Meredith Leong, Pete Lauer, and Russell Vance for discussions; Jake Bruml, Ryan Duggan, Michael Y.K. Leung, and Diana Ranoa for technical assistance; and Elie Diner for retroviral constructs. We thank Hector Nolla at the UC Berkeley Cancer Research Laboratory (CRL) for help with cell sorting. L.C. was supported by a post-doctoral fellowship from the Ramon Areces Foundation. This work was supported by P01 CA97296 and R01CA181160 from the National Cancer Institute. L.H.G., S.M.M., D.B.K., K.E.S., G.E.K., E.L, T.B., J.J.L., K.M., and T.W.D. are all paid employees of Aduro BioTech, hold stock in the company, and may be inventors on patent applications that apply to the CDN molecules described in the manuscript.

Received: November 26, 2014
Revised: February 24, 2015
Accepted: April 14, 2015
Published: May 7, 2015

REFERENCES

Ablasser, A., Goldeck, M., Cavlar, T., Deimling, T., Witte, G., Röhl, I., Hopfner, K.P., Ludwig, J., and Hornung, V. (2013). cGAS produces a 2′-5′-linked cyclic dinucleotide second messenger that activates STING. Nature *498*, 380–384.

Baguley, B.C., and Ching, L.-M. (1997). Immunomodulatory actions of xanthenone anticancer agents. BioDrugs *8*, 119–127.

Blank, C., Brown, I., Peterson, A.C., Spiotto, M., Iwai, Y., Honjo, T., and Gajewski, T.F. (2004). PD-L1/B7H-1 inhibits the effector phase of tumor rejection by T cell receptor (TCR) transgenic CD8+ T cells. Cancer Res. *64*, 1140–1145.

Brody, J.D., Ai, W.Z., Czerwinski, D.K., Torchia, J.A., Levy, M., Advani, R.H., Kim, Y.H., Hoppe, R.T., Knox, S.J., Shin, L.K., et al. (2010). In situ vaccination with a TLR9 agonist induces systemic lymphoma regression: a phase I/II study. J. Clin. Oncol. *28*, 4324–4332.

Burdette, D.L., and Vance, R.E. (2013). STING and the innate immune response to nucleic acids in the cytosol. Nat. Immunol. *14*, 19–26.

Burdette, D.L., Monroe, K.M., Sotelo-Troha, K., Iwig, J.S., Eckert, B., Hyodo, M., Hayakawa, Y., and Vance, R.E. (2011). STING is a direct innate immune sensor of cyclic di-GMP. Nature *478*, 515–518.

Cavlar, T., Deimling, T., Ablasser, A., Hopfner, K.P., and Hornung, V. (2013). Species-specific detection of the antiviral small-molecule compound CMA by STING. EMBO J. *32*, 1440–1450.

Chandra, D., Quispe-Tintaya, W., Jahangir, A., Asafu-Adjei, D., Ramos, I., Sintim, H.O., Zhou, J., Hayakawa, Y., Karaolis, D.K., and Gravekamp, C. (2014). STING ligand c-di-GMP improves cancer vaccination against metastatic breast cancer. Cancer Immunol Res *2*, 901–910.

Conlon, J., Burdette, D.L., Sharma, S., Bhat, N., Thompson, M., Jiang, Z., Rathinam, V.A.K., Monks, B., Jin, T., Xiao, T.S., et al. (2013). Mouse, but not human STING, binds and signals in response to the vascular disrupting agent 5,6-dimethylxanthenone-4-acetic acid. J. Immunol. *190*, 5216–5225.

Deng, L., Liang, H., Xu, M., Yang, X., Burnette, B., Arina, A., Li, X.-D., Mauceri, H., Beckett, M., Darga, T., et al. (2014). STING-dependent cytosolic DNA sensing promotes radiation-induced type I interferon-dependent antitumor immunity in immunogenic tumors. Immunity *41*, 843–852.

Diamond, M.S., Kinder, M., Matsushita, H., Mashayekhi, M., Dunn, G.P., Archambault, J.M., Lee, H., Arthur, C.D., White, J.M., Kalinke, U., et al. (2011). Type I interferon is selectively required by dendritic cells for immune rejection of tumors. J. Exp. Med. *208*, 1989–2003.

Diner, E.J., Burdette, D.L., Wilson, S.C., Monroe, K.M., Kellenberger, C.A., Hyodo, M., Hayakawa, Y., Hammond, M.C., and Vance, R.E. (2013). The innate immune DNA sensor cGAS produces a noncanonical cyclic dinucleotide that activates human STING. Cell Rep. *3*, 1355–1361.

Dubensky, T.W., Jr., Kanne, D.B., and Leong, M.L. (2013). Rationale, progress and development of vaccines utilizing STING-activating cyclic dinucleotide adjuvants. Ther Adv Vaccines *1*, 131–143.

Ebensen, T., Schulze, K., Riese, P., Link, C., Morr, M., and Guzmán, C.A. (2007a). The bacterial second messenger cyclic diGMP exhibits potent adjuvant properties. Vaccine *25*, 1464–1469.

Ebensen, T., Schulze, K., Riese, P., Morr, M., and Guzmán, C.A. (2007b). The bacterial second messenger cdiGMP exhibits promising activity as a mucosal adjuvant. Clin. Vaccine Immunol. *14*, 952–958.

Ebensen, T., Libanova, R., Schulze, K., Yevsa, T., Morr, M., and Guzmán, C.A. (2011). Bis-(3′,5′)-cyclic dimeric adenosine monophosphate: strong Th1/Th2/Th17 promoting mucosal adjuvant. Vaccine *29*, 5210–5220.

Fitzgerald, K.A., McWhirter, S.M., Faia, K.L., Rowe, D.C., Latz, E., Golenbock, D.T., Coyle, A.J., Liao, S.M., and Maniatis, T. (2003). IKKepsilon and TBK1 are essential components of the IRF3 signaling pathway. Nat. Immunol. *4*, 491–496.

Fuertes, M.B., Kacha, A.K., Kline, J., Woo, S.R., Kranz, D.M., Murphy, K.M., and Gajewski, T.F. (2011). Host type I IFN signals are required for antitumor CD8+ T cell responses through CD8alpha+ dendritic cells. J. Exp. Med. *208*, 2005–2016.

Gaffney, B.L., Veliath, E., Zhao, J., and Jones, R.A. (2010). One-flask syntheses of c-di-GMP and the [Rp,Rp] and [Rp,Sp] thiophosphate analogues. Org. Lett. *12*, 3269–3271.

Gajewski, T.F., Woo, S.R., Zha, Y., Spaapen, R., Zheng, Y., Corrales, L., and Spranger, S. (2013). Cancer immunotherapy strategies based on overcoming barriers within the tumor microenvironment. Curr. Opin. Immunol. *25*, 268–276.

Galon, J., Pagès, F., Marincola, F.M., Angell, H.K., Thurin, M., Lugli, A., Zlobec, I., Berger, A., Bifulco, C., Botti, G., et al. (2012). Cancer classification using the Immunoscore: a worldwide task force. J. Transl. Med. *10*, 205.

Gao, P., Ascano, M., Zillinger, T., Wang, W., Dai, P., Serganov, A.A., Gaffney, B.L., Shuman, S., Jones, R.A., Deng, L., et al. (2013a). Structure-function analysis of STING activation by c[G(2′,5′)pA(3′,5′)p] and targeting by antiviral DMXAA. Cell *154*, 748–762.

Gao, P., Ascano, M., Wu, Y., Barchet, W., Gaffney, B.L., Zillinger, T., Serganov, A.A., Liu, Y., Jones, R.A., Hartmann, G., et al. (2013b). Cyclic [G(2′,5′)pA(3′,5′)p] is the metazoan second messenger produced by DNA-activated cyclic GMP-AMP synthase. Cell *153*, 1094–1107.

Goins, W.F., Huang, S., Cohen, J.B., and Glorioso, J.C. (2014). Engineering HSV-1 vectors for gene therapy. Methods Mol. Biol. *1144*, 63–79.

Gray, P.M., Forrest, G., Wisniewski, T., Porter, G., Freed, D.C., DeMartino, J.A., Zaller, D.M., Guo, Z., Leone, J., Fu, T.M., and Vora, K.A. (2012). Evidence for cyclic diguanylate as a vaccine adjuvant with novel immunostimulatory activities. Cell. Immunol. *278*, 113–119.

Harlin, H., Meng, Y., Peterson, A.C., Zha, Y., Tretiakova, M., Slingluff, C., McKee, M., and Gajewski, T.F. (2009). Chemokine expression in melanoma metastases associated with CD8+ T-cell recruitment. Cancer Res. *69*, 3077–3085.

Hildner, K., Edelson, B.T., Purtha, W.E., Diamond, M., Matsushita, H., Kohyama, M., Calderon, B., Schraml, B.U., Unanue, E.R., Diamond, M.S., et al. (2008). Batf3 deficiency reveals a critical role for CD8alpha+ dendritic cells in cytotoxic T cell immunity. Science *322*, 1097–1100.

Ishikawa, H., and Barber, G.N. (2008). STING is an endoplasmic reticulum adaptor that facilitates innate immune signalling. Nature *455*, 674–678.

Ishikawa, H., Ma, Z., and Barber, G.N. (2009). STING regulates intracellular DNA-mediated, type I interferon-dependent innate immunity. Nature *461*, 788–792.

Jassar, A.S., Suzuki, E., Kapoor, V., Sun, J., Silverberg, M.B., Cheung, L., Burdick, M.D., Strieter, R.M., Ching, L.M., Kaiser, L.R., and Albelda, S.M. (2005). Activation of tumor-associated macrophages by the vascular disrupting agent 5,6-dimethylxanthenone-4-acetic acid induces an effective CD8+ T-cell-mediated antitumor immune response in murine models of lung cancer and mesothelioma. Cancer Res. *65*, 11752–11761.

Jin, L., Xu, L.G., Yang, I.V., Davidson, E.J., Schwartz, D.A., Wurfel, M.M., and Cambier, J.C. (2011). Identification and characterization of a loss-of-function human MPYS variant. Genes Immun. *12*, 263–269.

Joseph, W.R., Cao, Z., Mountjoy, K.G., Marshall, E.S., Baguley, B.C., and Ching, L.M. (1999). Stimulation of tumors to synthesize tumor necrosis factor-alpha in situ using 5,6-dimethylxanthenone-4-acetic acid: a novel approach to cancer therapy. Cancer Res. *59*, 633–638.

Kawarada, Y., Ganss, R., Garbi, N., Sacher, T., Arnold, B., and Hämmerling, G.J. (2001). NK- and CD8(+) T cell-mediated eradication of established tumors by peritumoral injection of CpG-containing oligodeoxynucleotides. J. Immunol. *167*, 5247–5253.

Kim, S., Li, L., Maliga, Z., Yin, Q., Wu, H., and Mitchison, T.J. (2013). Anticancer flavonoids are mouse-selective STING agonists. ACS Chem. Biol. *8*, 1396–1401.

Konno, H., Konno, K., and Barber, G.N. (2013). Cyclic dinucleotides trigger ULK1 (ATG1) phosphorylation of STING to prevent sustained innate immune signaling. Cell *155*, 688–698.

Lara, P.N., Jr., Douillard, J.Y., Nakagawa, K., von Pawel, J., McKeage, M.J., Albert, I., Losonczy, G., Reck, M., Heo, D.S., Fan, X., et al. (2011). Randomized phase III placebo-controlled trial of carboplatin and paclitaxel with or without the vascular disrupting agent vadimezan (ASA404) in advanced non-small-cell lung cancer. J. Clin. Oncol. *29*, 2965–2971.

Li, L., Yin, Q., Kuss, P., Maliga, Z., Millán, J.L., Wu, H., and Mitchison, T.J. (2014). Hydrolysis of 2′3′-cGAMP by ENPP1 and design of nonhydrolyzable analogs. Nat. Chem. Biol. *10*, 1043–1048.

McWhirter, S.M., Barbalat, R., Monroe, K.M., Fontana, M.F., Hyodo, M., Joncker, N.T., Ishii, K.J., Akira, S., Colonna, M., Chen, Z.J., et al. (2009). A host type I interferon response is induced by cytosolic sensing of the bacterial second messenger cyclic-di-GMP. J. Exp. Med. *206*, 1899–1911.

Niesen, F.H., Berglund, H., and Vedadi, M. (2007). The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nat. Protoc. *2*, 2212–2221.

Postow, M.A., Harding, J., and Wolchok, J.D. (2012). Targeting immune checkpoints: releasing the restraints on anti-tumor immunity for patients with melanoma. Cancer J. *18*, 153–159.

Prantner, D., Perkins, D.J., Lai, W., Williams, M.S., Sharma, S., Fitzgerald, K.A., and Vogel, S.N. (2012). 5,6-Dimethylxanthenone-4-acetic acid (DMXAA) activates stimulator of interferon gene (STING)-dependent innate immune pathways and is regulated by mitochondrial membrane potential. J. Biol. Chem. *287*, 39776–39788.

Roberson, S.M., and Walker, W.S. (1988). Immortalization of cloned mouse splenic macrophages with a retrovirus containing the v-raf/mil and v-myc oncogenes. Cell. Immunol. *116*, 341–351.

Römling, U., Galperin, M.Y., and Gomelsky, M. (2013). Cyclic di-GMP: the first 25 years of a universal bacterial second messenger. Microbiol. Mol. Biol. Rev. *77*, 1–52.

Sauer, J.D., Sotelo-Troha, K., von Moltke, J., Monroe, K.M., Rae, C.S., Brubaker, S.W., Hyodo, M., Hayakawa, Y., Woodward, J.J., Portnoy, D.A., and Vance, R.E. (2011). The N-ethyl-N-nitrosourea-induced Goldenticket mouse mutant reveals an essential function of Sting in the in vivo interferon response to Listeria monocytogenes and cyclic dinucleotides. Infect. Immun. *79*, 688–694.

Slansky, J.E., Rattis, F.M., Boyd, L.F., Fahmy, T., Jaffee, E.M., Schneck, J.P., Margulies, D.H., and Pardoll, D.M. (2000). Enhanced antigen-specific antitumor immunity with altered peptide ligands that stabilize the MHC-peptide-TCR complex. Immunity *13*, 529–538.

Sun, L., Wu, J., Du, F., Chen, X., and Chen, Z.J. (2013). Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. Science *339*, 786–791.

Wallace, A., LaRosa, D.F., Kapoor, V., Sun, J., Cheng, G., Jassar, A., Blouin, A., Ching, L.M., and Albelda, S.M. (2007). The vascular disrupting agent, DMXAA, directly activates dendritic cells through a MyD88-independent mechanism and generates antitumor cytotoxic T lymphocytes. Cancer Res. *67*, 7011–7019.

Wolchok, J.D., Kluger, H., Callahan, M.K., Postow, M.A., Rizvi, N.A., Lesokhin, A.M., Segal, N.H., Ariyan, C.E., Gordon, R.A., Reed, K., et al. (2013). Nivolumab plus ipilimumab in advanced melanoma. N. Engl. J. Med. *369*, 122–133.

Woo, S.R., Fuertes, M.B., Corrales, L., Spranger, S., Furdyna, M.J., Leung, M.Y., Duggan, R., Wang, Y., Barber, G.N., Fitzgerald, K.A., et al. (2014). STING-dependent cytosolic DNA sensing mediates innate immune recognition of immunogenic tumors. Immunity *41*, 830–842.

Woodward, J.J., Iavarone, A.T., and Portnoy, D.A. (2010). c-di-AMP secreted by intracellular Listeria monocytogenes activates a host type I interferon response. Science *328*, 1703–1705.

Yan, H., Wang, X., KuoLee, R., and Chen, W. (2008). Synthesis and immunostimulatory properties of the phosphorothioate analogues of cdiGMP. Bioorg. Med. Chem. Lett. *18*, 5631–5634.

Yi, G., Brendel, V.P., Shu, C., Li, P., Palanathan, S., and Cheng Kao, C. (2013). Single nucleotide polymorphisms of human STING can affect innate immune response to cyclic dinucleotides. PLoS ONE *8*, e77846.

Zhang, X., Shi, H., Wu, J., Zhang, X., Sun, L., Chen, C., and Chen, Z.J. (2013). Cyclic GMP-AMP containing mixed phosphodiester linkages is an endogenous high-affinity ligand for STING. Mol. Cell *51*, 226–235.

Zhao, L., Ching, L.M., Kestell, P., and Baguley, B.C. (2002). The antitumour activity of 5,6-dimethylxanthenone-4-acetic acid (DMXAA) in TNF receptor-1 knockout mice. Br. J. Cancer *87*, 465–470.

Cell Reports
Article

Manipulation of the Quorum Sensing Signal AI-2 Affects the Antibiotic-Treated Gut Microbiota

Jessica Ann Thompson,[1,5] Rita Almeida Oliveira,[1,5] Ana Djukovic,[2] Carles Ubeda,[2,3] and Karina Bivar Xavier[1,4,*]
[1]Instituto Gulbenkian de Ciência, 2780-156 Oeiras, Portugal
[2]Departamento de Genómica y Salud, Centro Superior de Investigación en Salud Pública, FISABIO, Valencia 46020, Spain
[3]Centers of Biomedical Research Network (CIBER) in Epidemiology and Public Health, Madrid 28029, Spain
[4]Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, 2780-157 Oeiras, Portugal
[5]Co-first author
*Correspondence: kxavier@igc.gulbenkian.pt
http://dx.doi.org/10.1016/j.celrep.2015.02.049

SUMMARY

The mammalian gut microbiota harbors a diverse ecosystem where hundreds of bacterial species interact with each other and their host. Given that bacteria use signals to communicate and regulate group behaviors (quorum sensing), we asked whether such communication between different commensal species can influence the interactions occurring in this environment. We engineered the enteric bacterium, *Escherichia coli,* to manipulate the levels of the interspecies quorum sensing signal, autoinducer-2 (AI-2), in the mouse intestine and investigated the effect upon antibiotic-induced gut microbiota dysbiosis. *E. coli* that increased intestinal AI-2 levels altered the composition of the antibiotic-treated gut microbiota, favoring the expansion of the Firmicutes phylum. This significantly increased the Firmicutes/Bacteroidetes ratio, to oppose the strong effect of the antibiotic, which had almost cleared the Firmicutes. This demonstrates that AI-2 levels influence the abundance of the major phyla of the gut microbiota, the balance of which is known to influence human health.

INTRODUCTION

The mammalian intestinal tract is home to approximately 10^{14} bacteria, encoding over 100-fold more genes than are within the human genome (Savage, 1977). This complement to the host's coding capacity provides a repertoire of additional metabolic functions including the digestion of complex polysaccharides, production of fatty acids, and vitamin biosynthesis (Arumugam et al., 2011; Louis et al., 2014). Interactions between commensal species and the host fulfill immensely important physiological roles, promoting the development of the intestinal tract, maturation of the immune system, and immunological tolerance to antigens (Berg, 1996; Hooper et al., 2012; Rakoff-Nahoum et al., 2004). These bacteria also provide a major protective barrier against pathogens, through a phenomenon known as colonization resistance (Bohnhoff and Miller, 1962; Lawley and Walker, 2013).

Though the gut microbiota is clearly beneficial in many ways, imbalances in this community (dysbiosis), including those induced by diet or antimicrobial usage, can pose a threat to host health. Antibiotics such as clindamycin or ampicillin that alter the commensal bacterial community also increase host susceptibility to opportunists such as *Clostridium difficile* and vancomycin-resistant enterococci (Buffie et al., 2012; Ubeda et al., 2010). Furthermore, shifts in the composition of the microbiota, particularly those involving the two predominant phyla in the mammalian gut, the Bacteroidetes and the Firmicutes, are associated with the pathogenesis of obesity, diabetes, chronic inflammatory bowel diseases, and gastrointestinal cancer, as well as autism and stress (Finegold et al., 2010; Frank et al., 2007; Galley et al., 2014; Ley et al., 2006; Qin et al., 2012; Turnbaugh et al., 2006; Wang et al., 2012). Consequently, the ability to drive this community from disease-associated to healthy states, by manipulating the native signals and interactions that occur between its members, to restore colonization resistance, for example, offers great potential for therapeutic benefit.

The mechanisms through which the resident microbial community inhibits the growth of invading microbes remain largely unknown, but there is increasing evidence that direct microbe-microbe interactions play a critical role in this process (Buffie et al., 2015; Hsiao et al., 2014; Kamada et al., 2012, 2013; Ng et al., 2013; Reeves et al., 2012). Cross-feeding and metabolic interactions clearly influence the composition of the microbiota (Buffie et al., 2015; Fabich et al., 2008; Flint et al., 2007; Kamada et al., 2012; Leatham et al., 2009; Ng et al., 2013). Bacteria also harbor vast arrays of mechanisms to sense and respond to features of their environment, including the presence of other bacteria. Small diffusible molecules known as autoinducers are synthesized and released in accordance with cell number; their subsequent detection enables bacteria to synchronously regulate behaviors at the population level in a process known as quorum sensing (Rutherford and Bassler, 2012). Attachment, biofilm formation, motility, and virulence are among the many phenotypes controlled in this manner. As quorum sensing and the behaviors it regulates are important in many bacteria-bacteria interactions in the host context, both symbiotic and pathogenic (Ruby, 2008; Rutherford and Bassler, 2012), this

phenomenon is likely to also contribute to the interactions between bacteria inhabiting the mammalian gut.

Many quorum sensing signals are species specific; however, production of and responses to one molecule, autoinducer-2 (AI-2), are observed throughout the bacterial kingdom (Chen et al., 2002; Miller et al., 2004; Pereira et al., 2013). As AI-2 produced by one species can influence gene expression in another, this signal can foster interspecies communication and enable bacteria to modify behaviors such as virulence, luminescence, and biofilm formation across different species (Armbruster et al., 2010; Cuadra-Saenz et al., 2012; Pereira et al., 2008; Xavier and Bassler, 2005a). This feature makes AI-2 an excellent candidate for mediating cell-cell interactions in the mammalian gut, where hundreds of bacterial species co-exist and interact. Multiple gut-associated bacteria that encode the AI-2 synthase, LuxS, or produce AI-2 have already been identified (Antunes et al., 2005; Hsiao et al., 2014; Lukás et al., 2008; Schauder et al., 2001). Furthermore, the human commensal bacterium *Ruminococcus obeum* was recently reported to inhibit colonization of the mouse gut by *V. cholerae*, partially through AI-2 signaling, highlighting how AI-2 produced by commensal bacteria can affect invading pathogens (Hsiao et al., 2014). We hypothesized that signaling through AI-2 might also occur between commensal members of the microbiota and asked whether AI-2 can shape the species composition of this community under conditions of dysbiosis.

To investigate this hypothesis, we made use of the natural signal production and depletion capabilities of the enteric bacterium, *Escherichia coli*. This organism secretes large amounts of AI-2 into the environment; it also harbors a highly efficient mechanism for signal uptake and degradation, known as the Lsr transport system (Pereira et al., 2012; Xavier and Bassler, 2005b). By internalizing and processing AI-2 produced by itself as well as from other species, Lsr system-expressing bacteria can disrupt the ability of neighboring species to correctly determine population density and regulate AI-2-dependent behavior appropriately, as shown in vitro in mixed cultures of *E. coli* and *Vibrio* spp. (Xavier and Bassler, 2005a). As a result, this system has been explored as a potential means for AI-2 interspecies quorum quenching (Roy et al., 2010). Given that *E. coli* can also stably colonize the mouse gut following streptomycin treatment (Conway et al., 2004), we used this bacterium as a tool to manipulate AI-2 signaling in vivo and demonstrated the accumulation and depletion of AI-2 in the intestinal tract of gnotobiotic mice. Streptomycin induces gut dysbiosis and has been used extensively to study *Salmonella* pathogenesis and *E. coli* physiology following the disruption of colonization resistance (Barthel et al., 2003; Spees et al., 2013). We characterized the changes induced by streptomycin upon the composition of the microbiota and then determined the effect of *E. coli*-mediated AI-2 manipulation upon this antibiotic-treated community.

RESULTS

Engineered *E. coli* Mutants Manipulate AI-2 Availability in the Mouse Gut

To manipulate AI-2 levels in the mouse gut, we constructed a combination of *E. coli* mutants affected in the Lsr system that accumulate different levels of AI-2 in vitro. To obtain high levels of AI-2, we constructed a mutant in *lsrK,* which encodes the signal kinase required for phosphorylation and intracellular retention of the signal. In the absence of LsrK, *E. coli* cannot sequester nor degrade AI-2 intracellularly, so the molecule accumulates extracellularly. To deplete environmental AI-2, we deleted *lsrR*, as it encodes a repressor of the Lsr transport system. In this mutant, the Lsr transporter is constitutively expressed at high levels and this strain is highly efficient at internalizing and scavenging environmental AI-2. The signal synthase, LuxS, was also deleted to produce mutants that do not produce AI-2 and thus do not contribute to the extracellular pool of AI-2. These mutations were introduced into a YFP-expressing strain of *E. coli* to ease identification of bacteria recovered from mice. In vitro validation of these strains confirmed that the Δ*lsrK* mutant strain accumulated and maintained high extracellular AI-2 levels (Figure S1A). Bacteria affected in *lsrR* rapidly internalized the signal, whereas no AI-2 could be detected in cultures containing mutants in the AI-2 synthase gene, *luxS*. We also confirmed that the Δ*lsrR*Δ*luxS* mutant strain internalized exogenously supplied AI-2 and could be used as a tool to deplete environmental signal (Figure S1B). No uptake occurred in cultures of Δ*lsrK*Δ*luxS* mutant bacteria provided with the signal, confirming that this strain does not deplete AI-2 in vitro (Figure S1B).

Germ-free C57BL/6J mice were then mono-colonized with either wild-type (WT), Δ*lsrK*, or *ΔlsrRΔluxS* mutant *E. coli* to determine the effect of each strain on AI-2 accumulation, without interference from other bacteria resident in the intestinal tract that might also produce or import AI-2. An additional group of germ-free mice was gavaged with PBS to provide a negative control. All *E. coli* strains colonized the mice to the same level, with approximately 10^9 colony-forming units (CFU)/g feces recovered 5 days after inoculation (Figure S1C). To determine the levels of AI-2 in the intestinal tract at this time point, extracts of the cecal contents were analyzed using a *Vibrio harveyi* biosensor that produces light in response to AI-2. Extracts from WT- and Δ*lsrK*-gavaged mice induced almost 25-fold more light production than cecal extracts from the control PBS-gavaged germ-free mice (Figure 1), demonstrating that these strains produce AI-2 in vivo, which accumulates in the gastrointestinal tract. The similar levels of AI-2 observed in these extracts, from WT- and Δ*lsrK*-colonized mice, suggest that Lsr transporter expression in the WT may be repressed in the cecum. This could be the result of inhibition by metabolites present in the gut, as glucose and other compounds such as glycerol are known to negatively regulate expression of this system (Pereira et al., 2012; Xavier and Bassler, 2005b). No signal was detected in extracts from mice colonized with *ΔlsrRΔluxS*, a non-AI-2-producing mutant *E. coli* (Figure 1).

To determine whether *ΔlsrRΔluxS* mutant *E. coli* could scavenge AI-2 present in the gut, germ-free mice were gavaged with a 1:1 mix of *ΔlsrK E. coli* (to provide AI-2) in combination with either the *ΔlsrRΔluxS* mutant (which removes but does not produce signal in vitro) or the control strain, *ΔlsrKΔluxS* (which neither produces nor internalizes signal). As predicted from our in vitro results, no AI-2 was detected in cecal contents from mice colonized with the mixture of Δ*lsrK* and *ΔlsrRΔluxS* mutants, whereas those from mice co-colonized with the Δ*lsrK*

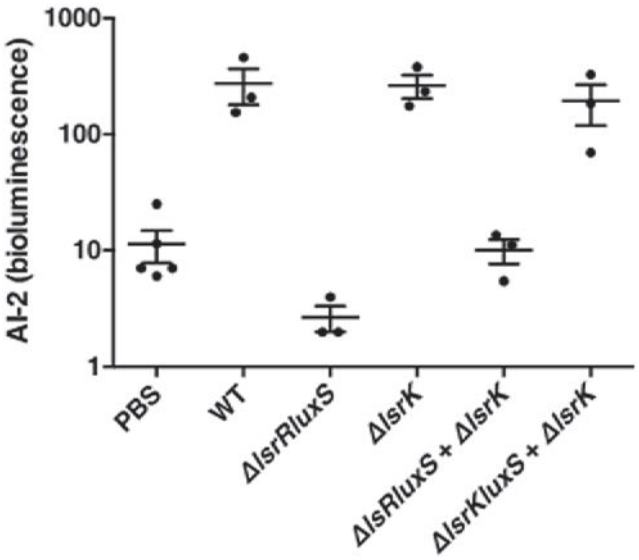


Figure 1. *E. coli* Accumulate and Deplete AI-2 in the Gut of Monocolonized Mice

AI-2 activity in cecal extracts harvested from germ-free mice 5 days after gavage with PBS, WT, *ΔlsrK* or *ΔlsrRΔluxS*; or a 1:1 mix of *ΔlsrK* and *ΔlsrRΔluxS*; or *ΔlsrK* and *ΔlsrKΔluxS E. coli,* as measured by *Vibrio harveyi* bioluminescence. Data shown are the mean, and the error bars correspond to SD; n = 3. See also Figure S1.

and *ΔlsrKΔluxS* mixture clearly contained signal (Figure 1). Total bacterial loads were similar in both groups of mice (Figure S1C), and the ratio between mutant bacterial strains remained close to one (Figure S1D). These data showed that there were no differences in the numbers of AI-2-producing Δ*lsrK* bacteria between the two groups, which could have otherwise explained the different levels of AI-2 detected. This demonstrates that the *ΔlsrRΔluxS* strain can efficiently internalize AI-2 in the gut and thus that deletion of *lsrR* relieved repression of the Lsr transporter. In summary, these results show that Δ*lsrK* and *ΔlsrRΔluxS* mutant *E. coli* can be used to manipulate AI-2 signaling, either accumulating or scavenging AI-2 in the mouse gut, respectively, and validate the use of the *ΔlsrKΔluxS* strain as a control that does not influence signal levels.

Streptomycin Induces Major Changes in the Fecal Microbiota

Streptomycin disrupts the microbiota, inducing a breakdown in colonization resistance, which enables *E. coli* to colonize the gut to high levels (Conway et al., 2004). We reasoned that this would provide a good system to determine the effect of AI-2 manipulation mediated by *E. coli* on the composition of the microbiota. In the absence of a detailed metagenomic description of the dysbiosis caused by prolonged exposure to streptomycin, we characterized the effect of this antibiotic upon the gut bacterial community during 28 days of treatment without manipulation of AI-2 levels. Antibiotic treatment decreased bacterial load: 4 days after the initial administration of antibiotic, density had dropped almost 20-fold from 1.61×10^{10} bacterial 16S rRNA gene copies/g feces in untreated mice to 8.36×10^{8} copies/g feces (Figure 2A). Despite continued exposure to streptomycin, the total bacterial load gradually increased after day 4 and stabilized 5- to 10-fold lower than that observed in untreated mice.

Streptomycin also induced major changes to the composition of the microbiota, as shown by high-throughput sequencing of the 16S rRNA genes amplified from DNA recovered from feces and analysis of the most abundant taxa prior to and during treatment (Figure 2B). Before treatment, the microbiota was composed of many different phylotypes, the majority of which belonged to the two phyla that commonly predominate among the gut bacterial community, the Firmicutes and the Bacteroidetes. These constituted 43% and 48%, respectively, of the bacteria detected (Figure 2C), with other phyla such as the Proteobacteria, Actinobacteria, and Deferribacteres present at low abundances, as previously observed in the mammalian gut (Stecher et al., 2007; Turnbaugh et al., 2010). Many of these taxa could no longer be detected 28 days into streptomycin treatment (Figures 2B and 2C), indicating an antibiotic-induced decrease in diversity of the gut bacterial community. Concurrently, only a few populations expanded. This trend of multiple losses and few increases was also visible at the lower taxonomic level when the relative frequencies of the 100 most-abundant operational taxonomic units (OTUs) (defined with 97% sequence similarity) were analyzed over time (Figure S2A). Many OTUs had already decreased in abundance below the level of detection only 2 days after the onset of treatment and did not subsequently recover (white boxes, Figure S2A). By day 28, just three OTUs, belonging to the Bacteroidetes phylum (OTU1, 2, and 4), constituted more than 60% of the bacteria detected in each mouse analyzed (Figure S3). This apparent decrease in community diversity was confirmed upon quantification of both the Chao richness index and the Shannon diversity index. These indexes both decreased by day 2 of treatment and remained low in streptomycin-treated microbiota when compared to those calculated from the untreated samples (Figures S2B and S2C). Thus, neither richness nor diversity of the microbiota recovered in the presence of antibiotic.

At the higher phylogenetic level, these changes drove a major shift in the balance between the Bacteroidetes and Firmicutes. From day 2, the Bacteroidetes had reached a relative abundance of approximately 90% and remained at this level during the treatment (Figure 2D). In contrast, the Firmicutes decreased hugely and only represented 0.7% of the bacterial community by the end of the experiment (Figures 2C and 2D). Despite fluctuations in the relative abundance of some phylotypes of the Proteobacteria (Figure S3, yellow), no significant difference was seen in the prevalence of this phylum.

Though streptomycin induced consistent changes in the ratio of Firmicutes to Bacteroidetes in all mice, its effect varied greatly at the individual OTU level during the early stages of antibiotic treatment. This is shown by the distinct changes in presence and abundance of specific OTUs in the different animals (Figures S2A and S3), which appeared to stabilize by day 28. The Jaccard distance, which provides a measure of dissimilarity taking into account the presence and absence of OTUs in the microbiota of each mouse, confirmed the apparent increase in inter-individual variation between the communities on days 4, 7, and 12 of

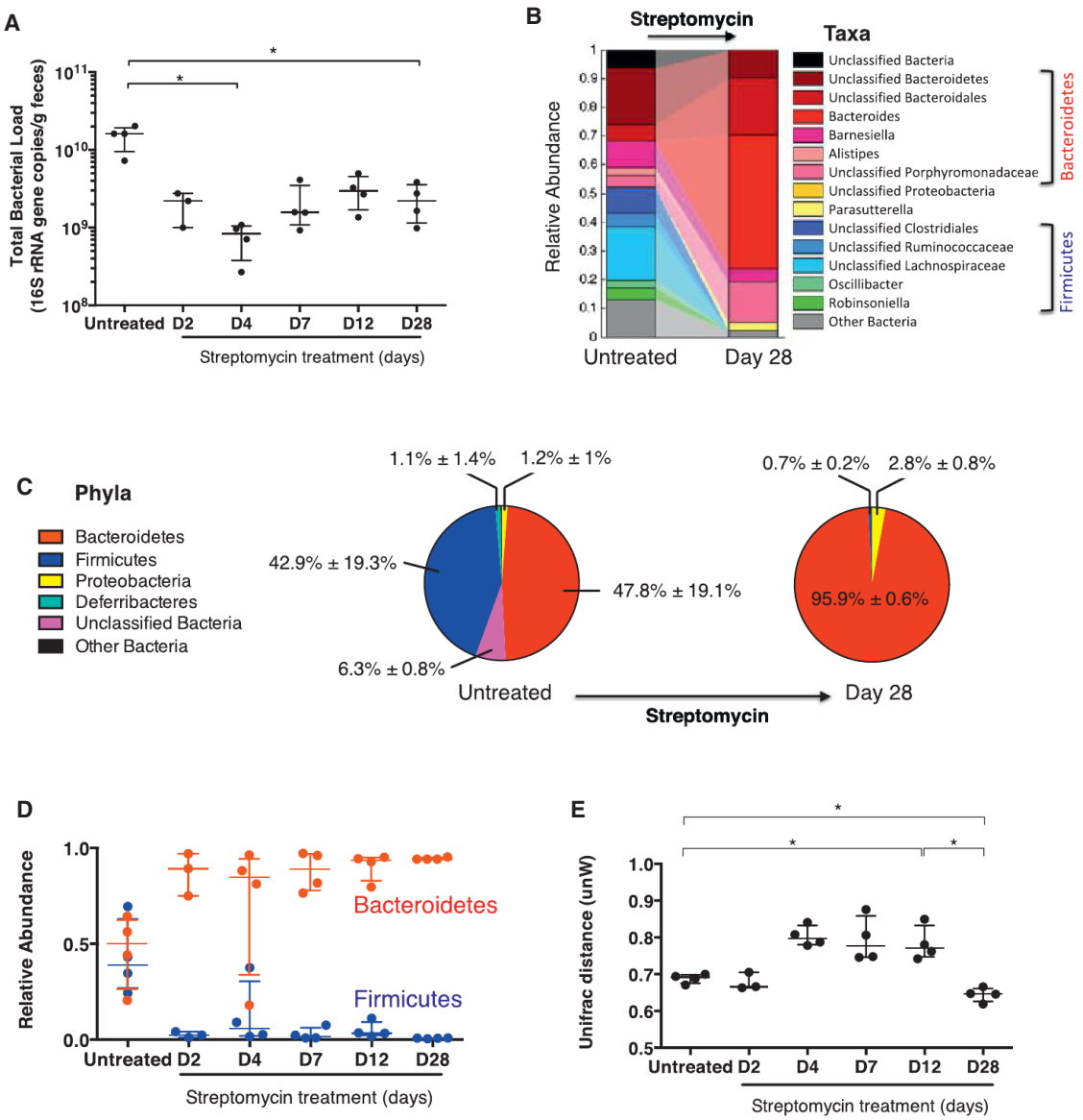


Figure 2. Streptomycin Changes Intestinal Microbiota Load and Composition

Streptomycin was given to four non-littermate mice in drinking water for 28 days. Animals were housed in separate cages, and fecal samples were collected for microbiota analysis prior to and 2, 4, 7, 12, and 28 days into streptomycin treatment.

(A) Total microbiota load measured from DNA by qPCR of the 16S rRNA gene copy number/g feces.

(B) Intestinal microbiota composition at the time points indicated. Each stacked bar represents the mean of the most abundant bacterial taxa in all the mice. The colored segments represent the relative fraction of each bacterial taxon.

(C) Relative abundance of the major phyla found in the gut microbiota before and after 28 days of treatment. Data shown are the mean ± the SD.

(D) Relative abundance of the Bacteroidetes (red circles) and Firmicutes (blue circles).

(E) Phylogenetic dissimilarities between microbial communities on each day determined by the mean unweighted Unifrac distance of the bacterial communities of each mouse versus all other mice.

Data shown in (A), (D), and (E) are the median, and the error bars show the interquartile range. Data were analyzed with the paired Student's t test (*$p < 0.05$). n = 4 except for day 2, where n = 3. See also Figures S2 and S3.

antibiotic treatment, followed by a decrease on day 28 (Figure S2D). This variability also affected the phylogenetic distance of the communities, as the unweighted UniFrac distance similarly increased after exposure to streptomycin and then lowered in the bacterial communities present on day 28 (Figure 2E). Distance measurements were in fact smaller for the microbiota after 28 days of streptomycin treatment than when untreated, showing a reduction in mouse-mouse variability upon prolonged antibiotic treatment (Figures 2E and S2D). These results demonstrate a highly variable effect of streptomycin upon

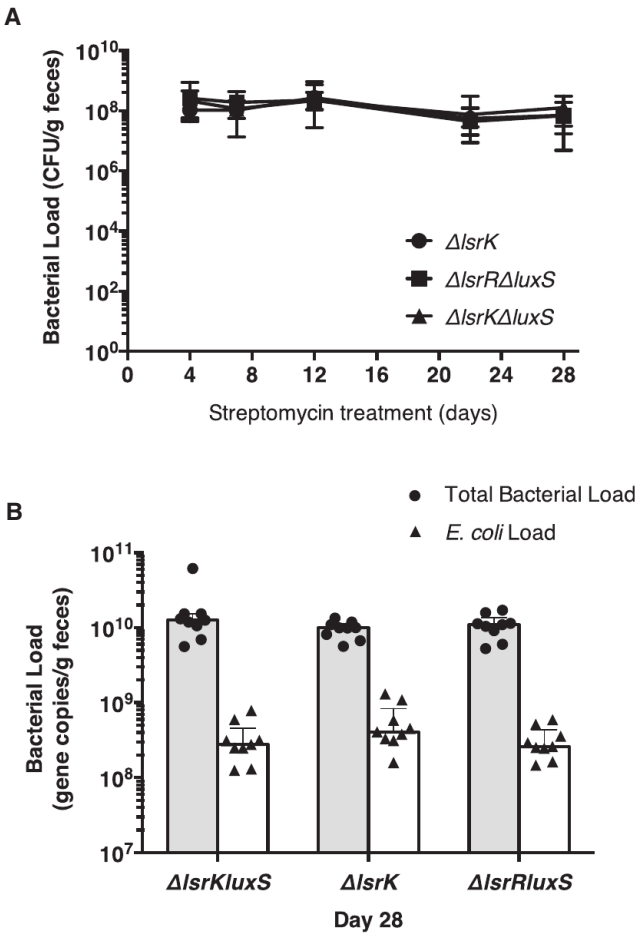


Figure 3. *E. coli* Colonization Levels and Total Microbiota Load in Streptomycin-Treated Mice

Individually housed mice were given streptomycin in drinking water and 2 days into treatment were colonized with the different *E. coli* mutants as specified. Fecal samples were collected at time points indicated.

(A) Loads of *E. coli* mutant strains as CFU/g feces from mice gavaged with the different *E. coli* mutants.

(B) Total microbiota (gray bars) and *E. coli* mutants (white bars) loads from day 28, determined by qPCR of 16S rRNA or *yfp* gene copy number/g feces from DNA extracted from fecal samples from mice from each of the different groups. Data shown are the median, and the error bars show the interquartile range; n = 9 per group.

the gut-associated bacteria within different mice during the early phases of antibiotic treatment that became more consistent across individuals by day 28. As the overall effect of the antibiotic was reproducible at this stage, this time point was selected for subsequent analysis of the influence of AI-2 on the bacterial community that emerges as a consequence of streptomycin-induced dysbiosis.

AI-2 Produced by *E. coli* Increases the Ratio of Firmicutes to Bacteroidetes during Streptomycin-Induced Dysbiosis

To determine whether AI-2 influences the species composition of the gut microbial community established during long-term streptomycin treatment, individually housed mice were gavaged 2 days after the start of antibiotic treatment with either *ΔlsrK*, *ΔlsrRΔluxS*, or *ΔlsrKΔluxS* YFP-expressing streptomycin-resistant *E. coli*. All three strains colonized the mice to similar levels, with approximately 10^8 CFU/g feces recovered throughout the experiment (Figure 3A). qPCR confirmed that *E. coli* colonization levels, and also the total bacterial load of the microbiota, were the same in all the three groups of mice 28 days after the start of antibiotic treatment (which corresponded to 26 days of *E. coli* colonization; Figure 3B).

Metagenomic analysis of the microbiota in fecal samples from day 28 of treatment using PCoA of the Jaccard index (Figure 4A) revealed that the structure of the microbiota in mice colonized with *ΔlsrK* mutant bacteria was significantly different from those in mice colonized with either of the other two groups (AMOVA test; p = 0.007, *ΔlsrK* versus *ΔlsrRΔluxS*; p = 0.006, *ΔlsrK* versus *ΔlsrKΔluxS*). No significant difference was observed between the communities of *ΔlsrRΔluxS*- and *ΔlsrKΔluxS*-colonized mice (p = 0.572). PCoA of the unweighted Unifrac showed the same trend, as the structure of the microbiota was also significantly different in terms of phylogeny between mice colonized by *ΔlsrK* and *ΔlsrRΔluxS* or *ΔlsrKΔluxS* mutant *E. coli* (AMOVA; p = 0.021 and p = 0.019, respectively; Figure 4B). This was due to the changes in relative abundance of multiple OTUs in *ΔlsrK* mutant-colonized mice compared to those in the mice containing either of the other two *E. coli* strains. The abundances of six OTUs were significantly different when compared between mice colonized by *ΔlsrK* and those containing *ΔlsrKΔluxS* or the *ΔlsrRΔluxS* bacteria (shown in Figures 4C–4H); other OTUs showed a similar trend for both comparisons but significant differences for only one of these cases (Figure S4). Some of the observed changes were considerable: OTU2, a member of the Bacteroidales order present at very high abundance on day 28, more than halved in frequency from 20.5% to 7.3% in the *ΔlsrKΔluxS*- and *ΔlsrK*-colonized mice, respectively (Figure S4A). No significant differences were observed when the abundances of the OTUs detected in mice colonized with either *ΔlsrKΔluxS* or *ΔlsrRΔluxS* were compared.

Most OTUs that changed in abundance were less prevalent in the presence of *ΔlsrK* mutant *E. coli* than when in the presence of either of the other two mutants (red colors, Figure 4I). These OTUs were all members of the Bacteroidetes phylum and included two from the Bacteroidales order and a member of the Porphyromonadaceae family. In contrast, one OTU classified as a Lachnospiraceae (a family within the Clostridiales order of Firmicutes) was increased in the presence of *ΔlsrK* mutant bacteria compared to both of the other groups (Figure 4C). Another member of the Lachnospiraceae family and an OTU correlating to the *Parasutterella* genus (classified within Burkholderiales order of Proteobacteria) were also found at significantly higher frequencies in the fecal microbiota of *ΔlsrK*-colonized mice compared to those harboring *ΔlsrKΔluxS* mutant bacteria (Figures S4B and S4D). These changes in abundance of multiple OTUs when in the presence of AI-2-producing (*ΔlsrK*) bacteria combined to give a significant effect at the phylum level (Figure 5). The Bacteroidetes, which again dominated the microbiota, were significantly lower in abundance in the *ΔlsrK* mutant-containing mice than in those

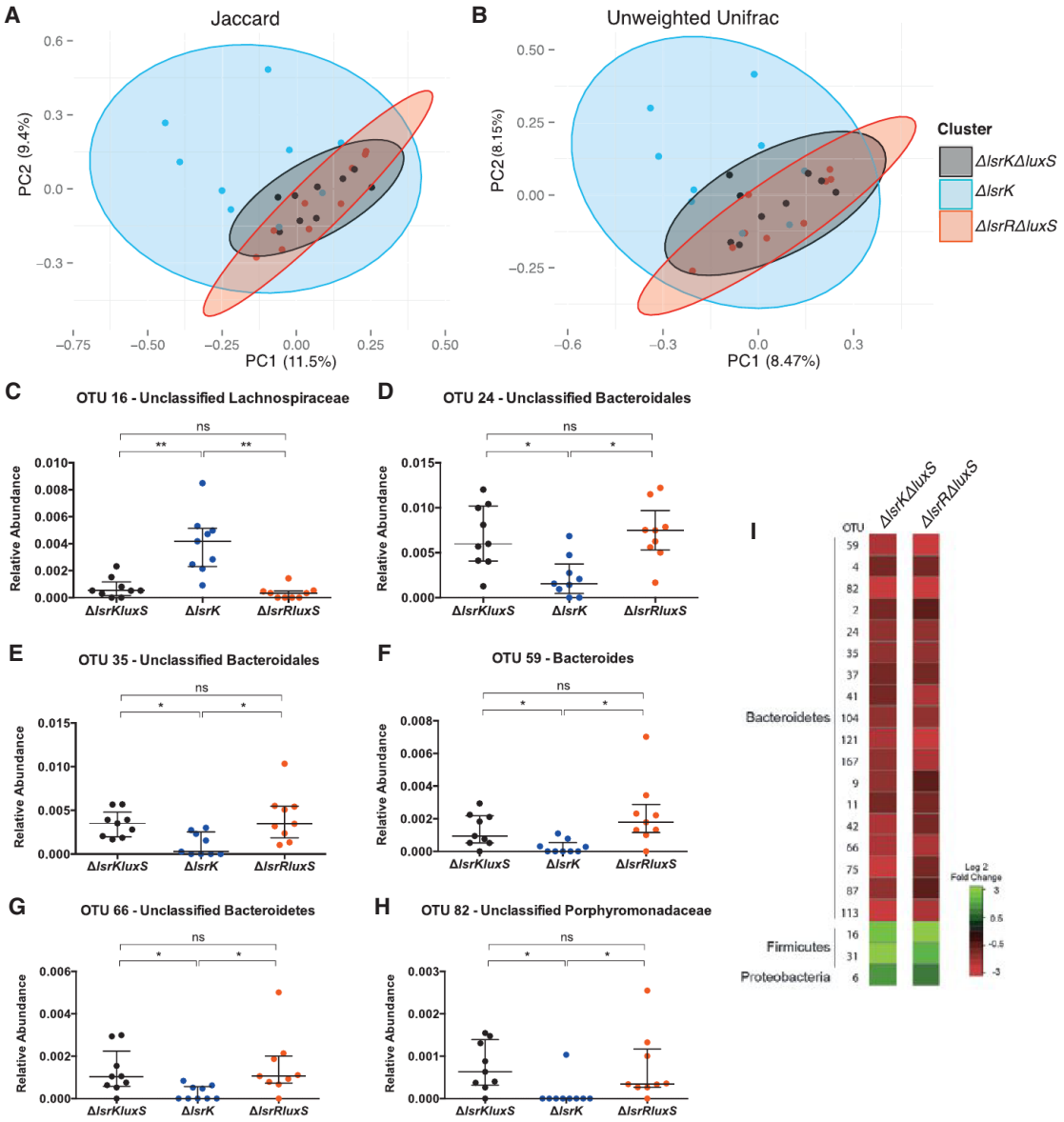


Figure 4. Microbiota Composition of Streptomycin-Treated Mice Differs in the Presence of *ΔlsrK* Mutant *E. coli*

Intestinal microbiota composition was analyzed in samples collected from the same mice as in Figure 3 after 28 days of antibiotic exposure (corresponding to 26 days of *E. coli* colonization; n = 9 per group).

(A and B) Analysis of the overall microbiota of the different *E. coli* groups by PCoA plots of Jaccard and unweighted Unifrac distances, respectively. The first two coordinates are shown. Each group is labeled with a different color, as indicated. Ellipses centered on the categorical averages of the metric distances with a 95% confidence interval for the first two coordinates of each group were drawn on the associated PCoA.

(C–H) Relative abundance of individual OTUs that exhibited a significant difference between the group colonized with Δ*lsrK* and both the other two groups are shown. Data shown are the median, and error bars show the interquartile range. Data were analyzed with the Wilcoxon test using the Benjamini-Hochberg correction (*q < 0.1; **q < 0.05; ns, not significant).

(I) Heatmap showing the fold change in relative abundance of OTUs in mice colonized with Δ*lsrK* mutant bacteria divided by the mean of the same OTU in mice colonized with either the Δ*lsrR*Δ*luxS* or the control group, Δ*lsrK*Δ*luxS*, *E. coli*. All the OTUs that exhibited a significant difference between the different groups are shown.

See also Figure S4.

colonized with either *ΔlsrRΔluxS* or *ΔlsrKΔluxS* bacteria (Figure 5A). In contrast, the Firmicutes were positively affected by the presence of AI-2-producing *ΔlsrK* mutant bacteria, undergoing 3- to 6-fold increase to 0.8% from a relative abundance of 0.24% and 0.13% in mice colonized with *ΔlsrKΔluxS* or *ΔlsrRΔluxS* mutant bacteria, respectively (Figure 5B). This

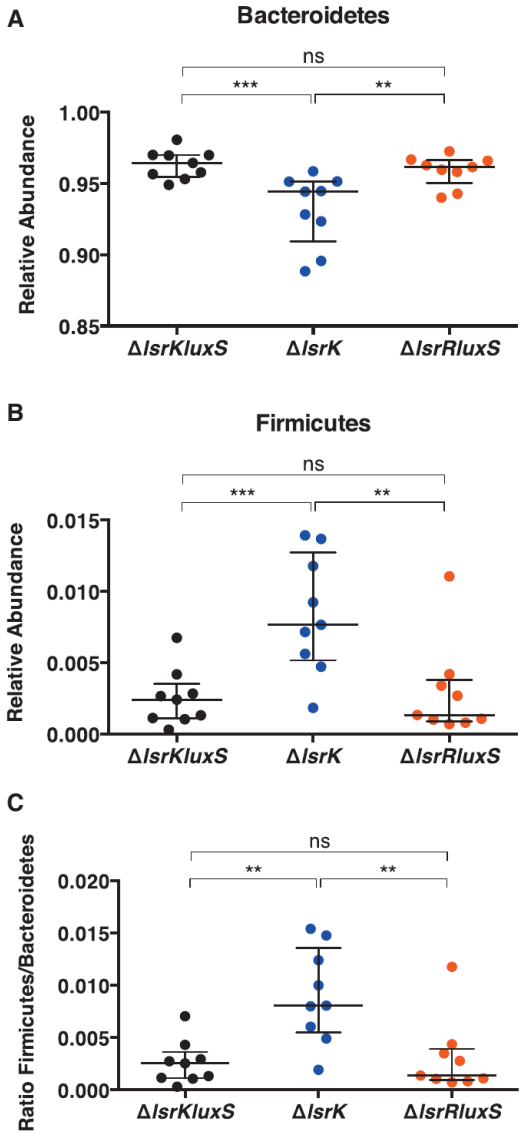


Figure 5. Colonization with *ΔlsrK* Mutant Bacteria Changes the Relative Abundance of the Major Phyla

(A and B) Relative abundances of the (A) Bacteroidetes and (B) Firmicutes phyla in streptomycin-treated mice colonized with either *ΔlsrKΔluxS*, *ΔlsrK*, or *ΔlsrRΔluxS* mutant *E. coli* (samples were collected 28 days into streptomycin treatment).

(C) The corresponding ratio of the relative abundances of Firmicutes to Bacteroidetes from the data described above.

Data shown are the median, and the error bars show the interquartile range of n = 9 per group. Data were analyzed with the Wilcoxon test using the Benjamini-Hochberg correction (**q < 0.05; ***q < 0.01; ns, not significant).

resulted in an increase in the ratio of Firmicutes to Bacteroidetes. Though this increase was significant (Figure 5C), the ratio in these *ΔlsrK* mutant-colonized mice remained much lower than that observed in the microbiota of untreated animals (0.897). Interestingly, the Bacteroidetes and Firmicutes are predicted to have very different AI-2 production capabilities: 17% and 83% of currently sequenced genomes (KEGG database) corresponding to these two phyla, respectively, encode homologs to the AI-2 synthase (Figure 6).

In conclusion, the presence of *ΔlsrK* mutant bacteria, which accumulate high levels of AI-2, changed the abundance of the Bacteroidetes and Firmicutes. The latter group were favored despite the continued presence of streptomycin, providing evidence that the bacterial communication molecule, AI-2, can modulate the composition of gut microbiota.

DISCUSSION

There is increasing evidence that microbe-microbe interactions influence the composition of the gut microbiota and affect the balance of this community. Our data support this notion by revealing that the administration of AI-2-producing bacteria into antibiotic-treated gut microbiota affected the abundance of multiple phylotypes and changed the overall structure of the emerging microbial community. Streptomycin caused a drop in bacterial load consistent with those seen in previous studies that analyzed single-dose or short-term treatments (Garner et al., 2009; Sekirov et al., 2008; Stecher et al., 2007). We also characterized the effect of streptomycin treatment using metagenomics and observed major changes to the bacterial community. Richness and diversity decreased due to a decrease in abundance, often to below the level of detection, of many OTUs: 80% of those present in untreated mice were undetectable by day 28 of streptomycin treatment. Such losses are likely to free niches for colonization and expansion by other members of this community that are better adapted to growth in the presence of antibiotics that can take advantage of the lower abundance of competitors.

Streptomycin is generally thought to create an environment favorable for growth of the Proteobacteria (Stecher et al., 2007). Though dose-dependent increases in prevalence of a bacterial group that included the Bacteroidetes were also previously reported (Sekirov et al., 2008), we had expected to see the expansion of the resident members of Proteobacteria. However, we observed only transient expansions of this phylum whereas the Bacteroidetes dominated, perhaps as a result of our longer antibiotic treatment (Stecher et al., 2007). Streptomycin resistance is easily acquired by spontaneous mutations: it is possible that the prevalence of Bacteroidetes was due to a major competitive advantage conferred upon bacteria that were resistant to streptomycin. Whether pre-existing or acquired during the course of treatment, this resistance could therefore be a major cause for the expansion of the OTUs most abundant at the end of treatment. These OTUs, all members of the Bacteroidetes phylum, were consistently present across all mice tested (OTUs 1, 2, and 4; Figure S3). As acquisition of resistance by independent spontaneous mutation events in the same OTUs in each of the mice during this treatment seems unlikely, this suggests that streptomycin-resistant members were already present in untreated mice and were subsequently selected for on exposure to antibiotic. Additionally, our data suggest that either the prevalence of streptomycin resistance or the propensity to gain it is higher among the Bacteroidetes than the Firmicutes.

Though selection of resistant strains is likely to be a considerable force in shaping the community that emerges during

	Putative LuxS gene (%)
Bacteroidetes	16.83
Firmicutes	83.33
- Bacilli	96.80
- Clostridia	48.68
- Other	27.27

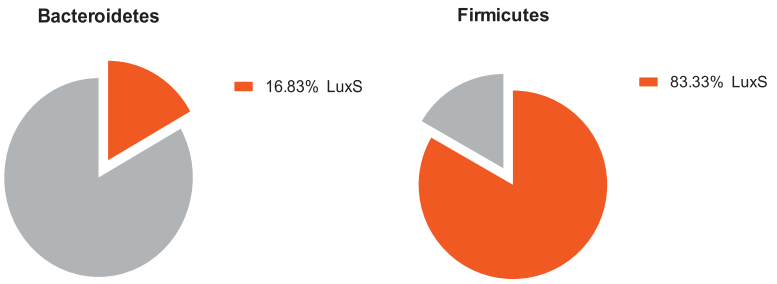


Figure 6. Higher Prevalence of LuxS Orthologs in the Complete Genomes of Bacteria Belonging to the Firmicutes
Percentage of full genome sequences corresponding to members of the Bacteroidetes and Firmicutes containing protein orthologs of LuxS.

antibiotic treatment, marked differences in relative abundance of the different phylotypes present at the end of our treatment were observed. This finding suggests that individual species have different abilities to exploit the space freed upon antibiotic treatment and that other antagonistic or synergistic interactions such as competition between species with similar metabolic demands and functions must also influence the growth of these microbes (Stecher et al., 2010). Given the high bacterial densities found within the multispecies gut environment, we asked whether the interspecies quorum sensing signal AI-2 could influence the presence and prevalence of species interacting within the microbiota during the emergence of the streptomycin-treated community.

Using *E. coli* strains engineered to either accumulate or deplete AI-2 in vivo, we demonstrated that manipulation of signal levels had no effect upon the numbers of *E. coli* itself (validating the use of this bacterium as a tool) nor upon the overall density of the microbiota upon colonization of streptomycin-treated mice. However, analysis of the species composition revealed differences in the microbiota of mice colonized with the Δ*lsrK* mutant *E. coli*, compared with those of the mice colonized with the other two mutants. As this strain accumulates AI-2 extracellularly, our data suggest that increased AI-2 availability leads to changes in the gut-associated bacteria that, at the phylum level, increase the ratio of Firmicutes to Bacteroidetes. This opposes the effect of streptomycin, which massively favors the Bacteroidetes. Such changes in the abundance of the major phyla have the potential to further influence AI-2 levels among the microbiota: signal production capabilities vary greatly between the major phyla, with a greater proportion of Firmicutes than Bacteroidetes encoding LuxS orthologs. By clearing most Firmicutes, streptomycin treatment is likely to create an environment containing relatively little AI-2, due to a paucity of AI-2 producers among the resulting gut community. With this in mind, it is perhaps unsurprising that the *E. coli* mutant strain that scavenges AI-2 had no detectable effect upon the streptomycin-treated microbiota. However, increased AI-2 availability, in favoring the Firmicutes, promotes the group of bacteria with a higher frequency of AI-2 producers. This positive feedback might be necessary to provide a context to achieve quorum and to potentiate further AI-2-dependent responses among the microbiota.

Members of both the Firmicutes and Bacteroidetes fulfill many functions important to host physiology, including the fermentation of diverse dietary polysaccharides too complex to be digested by the host. This process produces short-chain fatty acids (SCFAs) such as propionate, butyrate, and acetate (Louis et al., 2014). These molecules, particularly butyrate, provide approximately 10% of the host's calorie intake (McNeil, 1984) and also influence host gene expression, proinflammatory cytokine secretion, and T_{reg} induction (Arpaia et al., 2013; Chang et al., 2014; Furusawa et al., 2013; Smith et al., 2013). As a result, signaling through butyrate and other SCFAs modulates inflammation within the intestinal tract (Maslowski et al., 2009). Individual species of the Bacteroidetes and Firmicutes are thought to make distinct contributions to the pools of each SCFA: Firmicutes are proposed to include the major butyrate producers whereas increased prevalence of Bacteroidetes has been correlated with increased proportions of propionate in the total SCFA pool (Salonen et al., 2014). Thus, changes in abundance of these bacteria (such as those observed here in response to streptomycin or AI-2 availability) could alter the concentration of these metabolites within the gut, with consequent downstream effects upon host physiology, and potentially explain the altered concentrations of SCFAs and increase in inflammatory tone previously observed in the ceca of streptomycin-treated mice (Garner et al., 2009; Spees et al., 2013). Moreover, the increase in abundance of Firmicutes observed in the presence of the AI-2-producing *E. coli* mutant offers the exciting possibility that AI-2 signaling might have ameliorated the effect of streptomycin on the microbiota-derived functions: it favored the group of bacteria most detrimentally affected by antibiotic treatment and may have influenced their functions via signaling responses. Heightening the impact of signal manipulation among the microbiota in this way can be used to aid the identification of AI-2-regulated functions in this important bacterial community and explore the possibility of using AI-2 signaling to restore the protective functions of the gut microbiota or influence microbiota-induced host responses. The results presented here will facilitate the identification of candidate bacteria that are more likely to be sensitive to this signal molecule (those that favor signal producers, for example), knowledge which can be used to design models that further potentiate AI-2-dependent effects. This work highlights the potential gain from understanding and manipulating the bacterial chemical repertoire operating within the bacterial community inhabiting the gut, towards the aim of tailoring the composition of the microbiota to our benefit.

EXPERIMENTAL PROCEDURES

Bacterial Strains and Culture Conditions

All *E. coli* strains and primers used in this study are listed in supplemental tables. These are all *E. coli* K-12 MG1655 derivatives with a streptomycin-resistant mutation in *rpsL-1*(*K43N*) that express CFP or YFP constitutively. For details of culture conditions, genetic manipulation, and strain construction, see Supplemental Experimental Procedures.

Animal Studies

All experiments involving mice were approved by the Institutional Ethics Committee at the Instituto Gulbenkian de Ciência and the Portuguese National Entity (Direção Geral de Alimentação e Veterinária; ref. no. 008957, approval date 19/03/2013) following the Portuguese legislation (PORT 1005/92), which complies with the European Directive 86/609/EEC of the European Council.

To demonstrate AI-2 production in the gut, germ-free mice were gavaged with sterile PBS or 10^5 CFUs of wild-type, *ΔlsrK*, or *ΔlsrRΔluxS E. coli* strains. To demonstrate removal of AI-2 from the gut by the *E. coli ΔlsrRΔluxS* mutant strain, germ-free mice were gavaged with a 1:1 mix of 10^5 CFUs of *ΔlsrK* and *ΔlsrKΔluxS* or *ΔlsrK* and *ΔlsrRΔluxS* mutant bacteria labeled with either CFP or YFP. Fecal samples were collected 5 days after colonization and plated to determine bacterial load; cecal contents were harvested for analysis of AI-2 levels.

6- to 8-week-old male C57BL/6J mice conventionally raised under specific pathogen-free (SPF) conditions were used to analyze the effects of streptomycin and colonization by *E. coli* mutant strains upon the intestinal microbiota. To determine the effects of streptomycin treatment on the gut microbiota composition, four non-littermate mice were housed individually and maintained under 5 g/l streptomycin ad libitum in the drinking water (Conway et al. 2004). Fresh fecal samples were collected prior to antibiotic administration (day 0) and 2, 4, 7, 12, and 28 days during treatment for subsequent DNA extraction. To assess the effects of different *E. coli* mutants upon the gastrointestinal flora, five groups of mice (n = 6) were treated and maintained under streptomycin, as described above. On day 2 of treatment, two mice per group were gavaged with 100 μl PBS containing 10^8 colony-forming units (CFU) of either ARO071 (*ΔlsrK*), ARO093 (*ΔlsrKΔluxS*), or ARO081 (*ΔlsrRΔluxS*) and individually caged (so that n = 10 per treatment). Fecal samples were collected, part was homogenized in 1 ml sterile PBS and plated to determine colonization levels (CFU/g feces), and the remainder of each sample was used for subsequent DNA extraction.

Detection of AI-2 Activity

AI-2 activity was measured as previously described (Taga and Xavier, 2011) using the *V. harveyi* AI-2 reporter strain TL26 (*ΔluxN ΔluxS ΔcqsS*; Long et al., 2009). To determine AI-2 activity in mouse cecal extracts, the cecal contents were homogenized at a 10% weight/volume concentration in 0.1 M MOPS (pH 7). Samples were centrifuged and filtered, and then an equal volume of methanol was added to precipitate further debris. Supernatants were vacuum-dried, resuspended at 50% weight/volume in sterile water, and analyzed by bioassay as above. Enumeration of *V. harveyi* CFUs demonstrated no significant differences in growth of the reporter strain across the samples.

Sample Collection and DNA Extraction

DNA was extracted from fecal samples using the QIAamp DNA Stool Mini Kit (QIAGEN) following the manufacturer's instructions plus an additional membrane disruption step using 0.1-mm glass beads and high-speed shaking. Samples were stored at −20°C. Total DNA obtained was quantified with Qubit dsDNA BR Assay Kit (Invitrogen); samples with more than 10 ng/μl were sequenced and further analyzed (n = 9 per treatment).

Quantification of Bacterial Load by qPCR

Quantitative PCR (qPCR) was performed on DNA extracted from fecal samples using 16S rRNA universal primers and YFP (CFP)-specific primers to determine total bacterial and YFP-expressing *E. coli* loads/g feces, respectively.

16S rRNA Gene Amplification, Pyrosequencing, and Analysis

For each sample, the V1–V3 region of the 16S rRNA gene was amplified by PCR and sequenced using a 454 GS FLX Titanium platform following Roche recommendations. Sequences were processed using mothur (Schloss et al., 2009) as previously described (Ubeda et al., 2013), with some modifications. See Supplemental Experimental Procedures for further details of all methods and statistical analyses.

ACCESSION NUMBERS

The NCBI Sequence Read Archive accession number for the 16s rRNA sequences reported in this paper is SRP051373.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2015.02.049.

AUTHOR CONTRIBUTIONS

J.A.T., R.A.O., C.U., and K.B.X. designed the experiments. J.A.T. and R.A.O. performed all animal and in vitro experiments and most analyses. R.A.O., A.D., and C.U. performed sample preparation for 16S rRNA DNA sequencing and analysis. J.A.T. and K.B.X. wrote the manuscript with input from co-authors.

ACKNOWLEDGMENTS

We thank Jocelyne Demengeot, Isabel Gordo, Miguel P. Soares, Sandrine Isaac, and Joao B. Xavier for helpful discussion and critically reading the manuscript. We also acknowledge Joana Amaro for technical assistance. This work was funded by grants from the Howard Hughes Medical Institute (International Early Career Scientist; HHMI 55007436) and the Fundação para a Ciência e Tecnologia (PTDC/BIA-EVF/118075/2010 and RECI/IMI-IMU/0038/2012). C.U. was supported by grants from Ministerio de Ciencia e Innovacion (MICINN; SAF2011-29458) and a Marie-Curie Career Integration Grant (PCIG09-GA-2011-293894).

Received: December 18, 2014
Revised: February 18, 2015
Accepted: February 20, 2015
Published: March 19, 2015

REFERENCES

Antunes, L.C., Ferreira, L.Q., Ferreira, E.O., Miranda, K.R., Avelar, K.E., Domingues, R.M., and Ferreira, M.C. (2005). Bacteroides species produce Vibrio harveyi autoinducer 2-related molecules. Anaerobe *11*, 295–301.

Armbruster, C.E., Hong, W., Pang, B., Weimer, K.E.D., Juneau, R.A., Turner, J., and Swords, W.E. (2010). Indirect pathogenicity of *Haemophilus influenzae* and *Moraxella catarrhalis* in polymicrobial otitis media occurs via interspecies quorum signaling. mBio. *1*, e00102-10.

Arpaia, N., Campbell, C., Fan, X., Dikiy, S., van der Veeken, J., deRoos, P., Liu, H., Cross, J.R., Pfeffer, K., Coffer, P.J., and Rudensky, A.Y. (2013). Metabolites produced by commensal bacteria promote peripheral regulatory T-cell generation. Nature *504*, 451–455.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.-M., et al.; MetaHIT Consortium (2011). Enterotypes of the human gut microbiome. Nature *473*, 174–180.

Barthel, M., Hapfelmeier, S., Quintanilla-Martínez, L., Kremer, M., Rohde, M., Hogardt, M., Pfeffer, K., Rüssmann, H., and Hardt, W.-D. (2003). Pretreatment of mice with streptomycin provides a Salmonella enterica serovar Typhimurium colitis model that allows analysis of both pathogen and host. Infect. Immun. *71*, 2839–2858.

Berg, R.D. (1996). The indigenous gastrointestinal microflora. Trends Microbiol. *4*, 430–435.

Bohnhoff, M., and Miller, C.P. (1962). Enhanced susceptibility to Salmonella infection in streptomycin-treated mice. J. Infect. Dis. *111*, 117–127.

Buffie, C.G., Jarchum, I., Equinda, M., Lipuma, L., Gobourne, A., Viale, A., Ubeda, C., Xavier, J., and Pamer, E.G. (2012). Profound alterations of intestinal microbiota following a single dose of clindamycin results in sustained susceptibility to Clostridium difficile-induced colitis. Infect. Immun. *80*, 62–73.

Buffie, C.G., Bucci, V., Stein, R.R., McKenney, P.T., Ling, L., Gobourne, A., No, D., Liu, H., Kinnebrew, M., Viale, A., et al. (2015). Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. Nature *517*, 205–208.

Chang, P.V., Hao, L., Offermanns, S., and Medzhitov, R. (2014). The microbial metabolite butyrate regulates intestinal macrophage function via histone deacetylase inhibition. Proc. Natl. Acad. Sci. USA *111*, 2247–2252.

Chen, X., Schauder, S., Potier, N., Van Dorsselaer, A., Pelczer, I., Bassler, B.L., and Hughson, F.M. (2002). Structural identification of a bacterial quorum-sensing signal containing boron. Nature *415*, 545–549.

Conway, T., Krogfelt, K., and Cohen, P. (2004). The life of commensal *Escherichia coli* in the mammalian intestine. EcoSal Plus *2004* http://dx.doi.org/10.1128/ecosalplus.8.3.1.2.

Cuadra-Saenz, G., Rao, D.L., Underwood, A.J., Belapure, S.A., Campagna, S.R., Sun, Z., Tammariello, S., and Rickard, A.H. (2012). Autoinducer-2 influences interactions amongst pioneer colonizing streptococci in oral biofilms. Microbiology *158*, 1783–1795.

Fabich, A.J., Jones, S.A., Chowdhury, F.Z., Cernosek, A., Anderson, A., Smalley, D., McHargue, J.W., Hightower, G.A., Smith, J.T., Autieri, S.M., et al. (2008). Comparison of carbon nutrition for pathogenic and commensal Escherichia coli strains in the mouse intestine. Infect. Immun. *76*, 1143–1152.

Finegold, S.M., Dowd, S.E., Gontcharova, V., Liu, C., Henley, K.E., Wolcott, R.D., Youn, E., Summanen, P.H., Granpeesheh, D., Dixon, D., et al. (2010). Pyrosequencing study of fecal microflora of autistic and control children. Anaerobe *16*, 444–453.

Flint, H.J., Duncan, S.H., Scott, K.P., and Louis, P. (2007). Interactions and competition within the microbial community of the human colon: links between diet and health. Environ. Microbiol. *9*, 1101–1111.

Frank, D.N., St Amand, A.L., Feldman, R.A., Boedeker, E.C., Harpaz, N., and Pace, N.R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc. Natl. Acad. Sci. USA *104*, 13780–13785.

Furusawa, Y., Obata, Y., Fukuda, S., Endo, T.A., Nakato, G., Takahashi, D., Nakanishi, Y., Uetake, C., Kato, K., Kato, T., et al. (2013). Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. Nature *504*, 446–450.

Galley, J.D., Nelson, M.C., Yu, Z., Dowd, S.E., Walter, J., Kumar, P.S., Lyte, M., and Bailey, M.T. (2014). Exposure to a social stressor disrupts the community structure of the colonic mucosa-associated microbiota. BMC Microbiol. *14*, 189.

Garner, C.D., Antonopoulos, D.A., Wagner, B., Duhamel, G.E., Keresztes, I., Ross, D.A., Young, V.B., and Altier, C. (2009). Perturbation of the small intestine microbial ecology by streptomycin alters pathology in a Salmonella enterica serovar typhimurium murine model of infection. Infect. Immun. *77*, 2691–2702.

Hooper, L.V., Littman, D.R., and Macpherson, A.J. (2012). Interactions between the microbiota and the immune system. Science *336*, 1268–1273.

Hsiao, A., Ahmed, A.M.S., Subramanian, S., Griffin, N.W., Drewry, L.L., Petri, W.A., Jr., Haque, R., Ahmed, T., and Gordon, J.I. (2014). Members of the human gut microbiota involved in recovery from Vibrio cholerae infection. Nature *515*, 423–426.

Kamada, N., Kim, Y.-G., Sham, H.P., Vallance, B.A., Puente, J.L., Martens, E.C., and Núñez, G. (2012). Regulated virulence controls the ability of a pathogen to compete with the gut microbiota. Science *336*, 1325–1329.

Kamada, N., Seo, S.-U., Chen, G.Y., and Núñez, G. (2013). Role of the gut microbiota in immunity and inflammatory disease. Nat. Rev. Immunol. *13*, 321–335.

Lawley, T.D., and Walker, A.W. (2013). Intestinal colonization resistance. Immunology *138*, 1–11.

Leatham, M.P., Banerjee, S., Autieri, S.M., Mercado-Lubo, R., Conway, T., and Cohen, P.S. (2009). Precolonized human commensal Escherichia coli strains serve as a barrier to E. coli O157:H7 growth in the streptomycin-treated mouse intestine. Infect. Immun. *77*, 2876–2886.

Ley, R.E., Turnbaugh, P.J., Klein, S., and Gordon, J.I. (2006). Microbial ecology: human gut microbes associated with obesity. Nature *444*, 1022–1023.

Long, T., Tu, K.C., Wang, Y., Mehta, P., Ong, N.P., Bassler, B.L., and Wingreen, N.S. (2009). Quantifying the integration of quorum-sensing signals with single-cell resolution. PLoS Biol. *7*, e68.

Louis, P., Hold, G.L., and Flint, H.J. (2014). The gut microbiota, bacterial metabolites and colorectal cancer. Nat. Rev. Microbiol. *12*, 661–672.

Lukás, F., Gorenc, G., and Kopecný, J. (2008). Detection of possible AI-2-mediated quorum sensing system in commensal intestinal bacteria. Folia Microbiol. (Praha) *53*, 221–224.

Maslowski, K.M., Vieira, A.T., Ng, A., Kranich, J., Sierro, F., Yu, D., Schilter, H.C., Rolph, M.S., Mackay, F., Artis, D., et al. (2009). Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. Nature *461*, 1282–1286.

McNeil, N.I. (1984). The contribution of the large intestine to energy supplies in man. Am. J. Clin. Nutr. *39*, 338–342.

Miller, S.T., Xavier, K.B., Campagna, S.R., Taga, M.E., Semmelhack, M.F., Bassler, B.L., and Hughson, F.M. (2004). Salmonella typhimurium recognizes a chemically distinct form of the bacterial quorum-sensing signal AI-2. Mol. Cell *15*, 677–687.

Ng, K.M., Ferreyra, J.A., Higginbottom, S.K., Lynch, J.B., Kashyap, P.C., Gopinath, S., Naidu, N., Choudhury, B., Weimer, B.C., Monack, D.M., and Sonnenburg, J.L. (2013). Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens. Nature *502*, 96–99.

Pereira, C.S., McAuley, J.R., Taga, M.E., Xavier, K.B., and Miller, S.T. (2008). Sinorhizobium meliloti, a bacterium lacking the autoinducer-2 (AI-2) synthase, responds to AI-2 supplied by other bacteria. Mol. Microbiol. *70*, 1223–1235.

Pereira, C.S., Santos, A.J., Bejerano-Sagie, M., Correia, P.B., Marques, J.C., and Xavier, K.B. (2012). Phosphoenolpyruvate phosphotransferase system regulates detection and processing of the quorum sensing signal autoinducer-2. Mol. Microbiol. *84*, 93–104.

Pereira, C.S., Thompson, J.A., and Xavier, K.B. (2013). AI-2-mediated signalling in bacteria. FEMS Microbiol. Rev. *37*, 156–181.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55–60.

Rakoff-Nahoum, S., Paglino, J., Eslami-Varzaneh, F., Edberg, S., and Medzhitov, R. (2004). Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. Cell *118*, 229–241.

Reeves, A.E., Koenigsknecht, M.J., Bergin, I.L., and Young, V.B. (2012). Suppression of Clostridium difficile in the gastrointestinal tracts of germfree mice inoculated with a murine isolate from the family Lachnospiraceae. Infect. Immun. *80*, 3786–3794.

Roy, V., Fernandes, R., Tsao, C.-Y., and Bentley, W.E. (2010). Cross species quorum quenching using a native AI-2 processing enzyme. ACS Chem. Biol. *5*, 223–232.

Ruby, E.G. (2008). Symbiotic conversations are revealed under genetic interrogation. Nat. Rev. Microbiol. *6*, 752–762.

Rutherford, S.T., and Bassler, B.L. (2012). Bacterial quorum sensing: its role in virulence and possibilities for its control. Cold Spring Harb. Perspect. Med. *2*, pii: a012427.

Salonen, A., Lahti, L., Salojärvi, J., Holtrop, G., Korpela, K., Duncan, S.H., Date, P., Farquharson, F., Johnstone, A.M., Lobley, G.E., et al. (2014). Impact of diet and individual variation on intestinal microbiota composition and fermentation products in obese men. ISME J. *8*, 2218–2230.

Savage, D.C. (1977). Microbial ecology of the gastrointestinal tract. Annu. Rev. Microbiol. *31*, 107–133.

Schauder, S., Shokat, K., Surette, M.G., and Bassler, B.L. (2001). The LuxS family of bacterial autoinducers: biosynthesis of a novel quorum-sensing signal molecule. Mol. Microbiol. *41*, 463–476.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. *75*, 7537–7541.

Sekirov, I., Tam, N.M., Jogova, M., Robertson, M.L., Li, Y., Lupp, C., and Finlay, B.B. (2008). Antibiotic-induced perturbations of the intestinal microbiota alter host susceptibility to enteric infection. Infect. Immun. *76*, 4726–4736.

Smith, P.M., Howitt, M.R., Panikov, N., Michaud, M., Gallini, C.A., Bohlooly-Y, M., Glickman, J.N., and Garrett, W.S. (2013). The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. Science *341*, 569–573.

Spees, A.M., Wangdi, T., Lopez, C.A., Kingsbury, D.D., Xavier, M.N., Winter, S.E., Tsolis, R.M., and Bäumler, A.J. (2013). Streptomycin-induced inflammation enhances *Escherichia coli* gut colonization through nitrate respiration. mBio. *4*, e00430-13.

Stecher, B., Robbiani, R., Walker, A.W., Westendorf, A.M., Barthel, M., Kremer, M., Chaffron, S., Macpherson, A.J., Buer, J., Parkhill, J., et al. (2007). Salmonella enterica serovar typhimurium exploits inflammation to compete with the intestinal microbiota. PLoS Biol. *5*, 2177–2189.

Stecher, B., Chaffron, S., Käppeli, R., Hapfelmeier, S., Freedrich, S., Weber, T.C., Kirundi, J., Suar, M., McCoy, K.D., von Mering, C., et al. (2010). Like will to like: abundances of closely related species can predict susceptibility to intestinal colonization by pathogenic and commensal bacteria. PLoS Pathog. *6*, e1000711.

Taga, M.E., and Xavier, K.B. (2011). Methods for analysis of bacterial autoinducer-2 production. Curr. Protoc. Microbiol. *Chapter 1*, Unit 1C.1.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., and Gordon, J.I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. Nature *444*, 1027–1031.

Turnbaugh, P.J., Quince, C., Faith, J.J., McHardy, A.C., Yatsunenko, T., Niazi, F., Affourtit, J., Egholm, M., Henrissat, B., Knight, R., and Gordon, J.I. (2010). Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. Proc. Natl. Acad. Sci. USA *107*, 7503–7508.

Ubeda, C., Taur, Y., Jenq, R.R., Equinda, M.J., Son, T., Samstein, M., Viale, A., Socci, N.D., van den Brink, M.R.M., Kamboj, M., and Pamer, E.G. (2010). Vancomycin-resistant Enterococcus domination of intestinal microbiota is enabled by antibiotic treatment in mice and precedes bloodstream invasion in humans. J. Clin. Invest. *120*, 4332–4341.

Ubeda, C., Bucci, V., Caballero, S., Djukovic, A., Toussaint, N.C., Equinda, M., Lipuma, L., Ling, L., Gobourne, A., No, D., et al. (2013). Intestinal microbiota containing Barnesiella species cures vancomycin-resistant Enterococcus faecium colonization. Infect. Immun. *81*, 965–973.

Wang, T., Cai, G., Qiu, Y., Fei, N., Zhang, M., Pang, X., Jia, W., Cai, S., and Zhao, L. (2012). Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. ISME J. *6*, 320–329.

Xavier, K.B., and Bassler, B.L. (2005a). Interference with AI-2-mediated bacterial cell-cell communication. Nature *437*, 750–753.

Xavier, K.B., and Bassler, B.L. (2005b). Regulation of uptake and processing of the quorum-sensing autoinducer AI-2 in Escherichia coli. J. Bacteriol. *187*, 238–248.

OPEN ACCESS
CellPress

Cell Reports
Resource

Mapping Social Behavior-Induced Brain Activation at Cellular Resolution in the Mouse

Yongsoo Kim,[1] Kannan Umadevi Venkataraju,[1] Kith Pradhan,[1] Carolin Mende,[1] Julian Taranda,[1] Srinivas C. Turaga,[2] Ignacio Arganda-Carreras,[2] Lydia Ng,[3] Michael J. Hawrylycz,[3] Kathleen S. Rockland,[1,4] H. Sebastian Seung,[2] and Pavel Osten[1,*]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
[2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Boston, MA 02139, USA
[3]Allen Institute for Brain Science, Seattle, WA 98103, USA
[4]Boston University School of Medicine, Boston, MA 02118, USA
*Correspondence: osten@cshl.edu
http://dx.doi.org/10.1016/j.celrep.2014.12.014

SUMMARY

Understanding how brain activation mediates behaviors is a central goal of systems neuroscience. Here, we apply an automated method for mapping brain activation in the mouse in order to probe how sex-specific social behaviors are represented in the male brain. Our method uses the immediate-early-gene *c-fos*, a marker of neuronal activation, visualized by serial two-photon tomography: the c-*fos*-GFP+ neurons are computationally detected, their distribution is registered to a reference brain and a brain atlas, and their numbers are analyzed by statistical tests. Our results reveal distinct and shared female and male interaction-evoked patterns of male brain activation representing sex discrimination and social recognition. We also identify brain regions whose degree of activity correlates to specific features of social behaviors and estimate the total numbers and the densities of activated neurons per brain areas. Our study opens the door to automated screening of behavior-evoked brain activation in the mouse.

INTRODUCTION

Central to the understanding of brain functions is insight into the distribution of neuronal activity that drives behavior. Local measurements of brain activity in behaving mice can be made with electrodes and fluorescent calcium indicators (Buzsáki, 2004; Grewe and Helmchen, 2009), but such approaches provide information regarding only a very small fraction of the ~70 million neurons that comprise the mouse brain. The detection of elevated levels of the immediate-early genes (IEGs) linked to recent neuronal activity (Clayton, 2000; Guzowski et al., 2005) is a more spatially comprehensive technique. While it lacks the time resolution of electrophysiological recordings or calcium imaging, it does have the potential of providing a complete view of recent whole-brain activity. Once determined, the whole-brain IEG-based map can be used to generate structure-function hypotheses to be probed by high-resolution recordings as well as optogenetic and chemogenetic methods (Fenno et al., 2011; Lee et al., 2014).

Here, we use a pipeline of computational methods that permits automated unbiased mapping of c-fos induction in mouse brains at single-cell resolution, in a similar way as recently described for mapping the induction of the IEG Arc (Vousden et al., 2014). Specifically, we use serial two-photon (STP) tomography (Ragan et al., 2012) to image the expression of c-*fos*-GFP, a transgenic c-*fos* green fluorescent protein reporter (Reijmers et al., 2007), across the entire mouse brain. The activated c-*fos*-GFP+ cells are computationally detected, their location is mapped at stereotaxic coordinates within a reference brain, and their numbers and densities per anatomical brain areas are determined within the Allen Mouse Brain Atlas. Finally, region of interest (ROI)-based and voxel-based statistical tests are applied to identify brain areas with behaviorally evoked c-*fos*-GFP activation.

To demonstrate the application of the computational pipeline to the mapping of behavior-evoked brain activation, we focus on mouse social behavior and generate activation maps representing sex-specific social behaviors in the male brain. Rodent social behavior is an area of intense research, and c-fos mapping, lesion studies, and other functional approaches have been used to identify brain regions that are activated and contribute to male and female sexual behaviors as well as male-male aggressive behaviors (Anderson, 2012; Biały and Kaczmarek, 1996; Brennan and Zufall, 2006; Coolen et al., 1996; Pfaus and Heeb, 1997; Veening et al., 2005; Yang and Shah, 2014). Much less is known, on the other hand, about the brain areas activated during the initial period of sex discrimination and social recognition before the manifestation of the correct behavioral response.

Here, we explore the question of sex discrimination and social recognition by limiting the male-female and male-male interactions to a brief 90 s period, during which the behavioral repertoire comprises only social exploratory activity, such as anogenital sniffing, close following, and nose-to-nose sniffing, without mating or aggression. A side-by-side comparison of the female and male interaction-evoked whole-brain activation revealed (1) a broad activation of areas downstream of both the main and accessory olfactory bulb (MOB and AOB) in the male-female

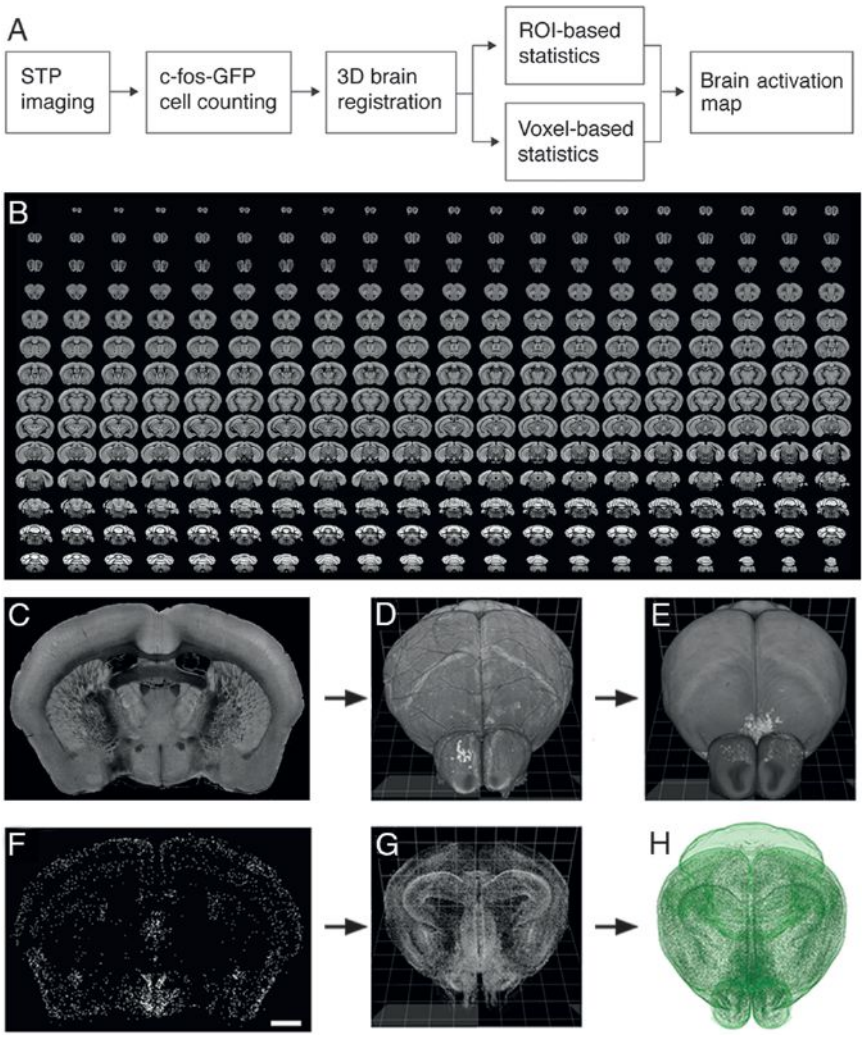


Figure 1. STP Tomography and Computational Detection of c-*fos*-GFP+ Cells

(A) Imaging and data processing pipeline for mapping whole-brain activation in c-*fos*-GFP mice.

(B) A sample 280-serial section data set of a c-*fos*-GFP mouse brain imaged by STP tomography.

(C–H) Registration of CN-detected c-*fos*-GFP+ cells in the RSTP brain. (C) A coronal section shows the autofluorescence signal, which is used for registering the 3D reconstructed sample brain (D) onto the RSTP brain (E). (F) A total of 2,177 c-*fos*-GFP+ cells were detected in the same coronal section; scale bar, 1 mm. (G) A total of 360,183 c-*fos*-GFP+ cells were detected in the whole brain, reconstructed in 3D, and (H) registered onto the RSTP brain using the image registration parameters established in the (D) and (E) step.

interaction and a bias toward structures downstream of the MOB in the male-male interaction; (2) activation of structures related to behavioral motivation during the male-female, but not male-male, interaction; and (3) sex-specific as well as shared hypothalamic activation. Taking advantage of the cellular resolution of the whole-brain data, we then identified brain regions whose level of activation was correlated to specific features of the social behaviors, including regions linked to anogenital sniffing that lie downstream of the pheromone-activated AOB and regions linked to close following that belong to the striatopallidothalamocortical circuitry. Finally, we calculated the total numbers and the densities of c-*fos*-GFP+ cells per activated brain region of the female- and male-specific brain data sets, providing a quantitative estimate of whole-brain activation evoked by social behaviors.

RESULTS

Whole-Brain Detection of c-*fos*-GFP+ Cells in STP Tomography Data Sets

We have established an automated and quantitative whole-brain method for mapping behaviorally evoked c-fos induction in transgenic reporter mice expressing c-fos-GFP from a recombinant *c-fos* promoter (Reijmers et al., 2007) (Experimental Procedures). This necessitated the development and optimization of (1) computational detection of c-*fos*-GFP+ cells in the mouse brain imaged by STP tomography (Ragan et al., 2012), (2) 3D registration of the STP data sets to a reference mouse brain, and (3) statistical analyses of the whole-brain distribution of the c-*fos*-GFP+ cells (Figure 1A).

The mouse brains were imaged by STP tomography as data sets of 280 serial coronal sections, with *x-y* resolution 1.0 μm and *z*-spacing 50 μm, which required an imaging time of ~21 hr per brain (Figure 1B) (Ragan et al., 2012). To achieve a reliable computational detection of the c-*fos*-GFP+ cells throughout the whole brain, we used convolutional networks (CNs) that can learn to recognize image features in complex data sets (V. Jain et al., 2007, IEEE, conference; Turaga et al., 2010) (Figure S1; Experimental Procedures). Since nearby c-*fos*-GFP+ cells were sometimes merged in the CN output, a postprocessing step was devised that could separate such "touching" cells (Figure S1). The CN performance was then quantified on a new set of marked-up fields of view from a second c-*fos*-GFP brain using the F-score measure, which represents the harmonic mean of the precision and recall (i.e., the false positive and false negative error rate), where F score 1 is the best and 0 the worst. The CN performance reached F-score 0.88 (precision 0.86, recall 0.90), which was comparable to human interuser variability represented by F-score 0.90 (precision 0.90, recall 0.90) (Figure S1; Experimental Procedures). We conclude that the trained CN provides an automated and highly accurate method for detection of c-*fos*-GFP+ cells in whole mouse brains imaged by STP tomography.

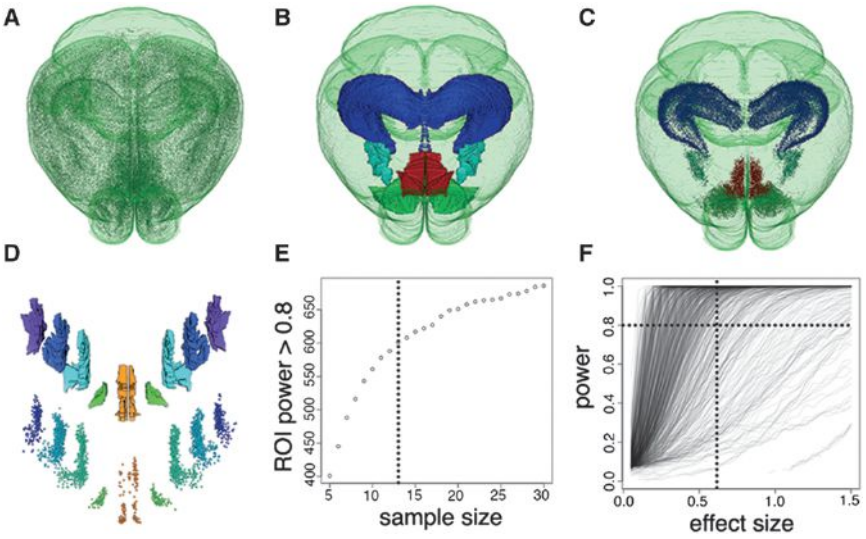


Figure 2. ROI-Based Segmentation and Sample Size Calculation

(A–D) ROI-based segmentation of the whole-brain c-*fos*-GFP+ cell count. (A) Whole-brain view of 360,183 c-*fos*-GFP+ cells (same brain as in Figure 1H). (B and C) Examples of ABA ROI segmentation (B) and the corresponding c-*fos*-GFP+ cell counts (C): hippocampus: dark blue; 33,508 cells; medial amygdalar nucleus: light blue; 3,035 cells; nucleus accumbens: green; 13,627 cells; and infralimbic cortical area: red; 4,665 cells. (D) Further segmentation of the infralimbic region by cortical layers; top shows the layer ROIs, from layer 1 (orange) to layer 6 (purple), and bottom shows the c-*fos*-GFP+ cell counts (ILA1 = 223, ILA2 = 243, ILA2/3 = 1,572, ILA5 = 1,731, ILA6 = 896 c-*fos*-GFP+ cells). The spacing between the layers was enlarged for better visualization.

(E and F) Estimation of the sample size based on power analysis of c-*fos*-GFP+ cell counts. (E) The simulation of the relationship between the number of sufficiently powered ROIs and the sample size shows a steep increase until ~N = 10, which then begins to plateau. For the current study, we chose a sample size of N = 13 (dashed line). (F) The plot of the relationship between the statistical power of each ROI and the effect size for N = 13 group. Of the total 763 ROIs analyzed, 601 (78.8%) showed sufficient statistical power at the effect size 0.6 and 699 (91.6%) at the effect size 1.0.

Anatomical Registration of the Whole-Brain c-*fos*-GFP Data

Results from the CN-based cell counting produce a number of c-*fos*-GFP+ cells per the individual 280-section data sets, with each cell having an *xyz* location. To be able to compare patterns of c-fos activation between experimental groups in one common brain volume, we created a Reference STP (RSTP) brain coregistered to the digital Allen Brain Atlas (ABA) for 8-week-old C57BL/6 mice (Sunkin et al., 2013) (Figure S2A; Experimental Procedures; Movie S1). The image registrations were done by a 3D affine transformation, followed by a 3D B-spline transformation with Mattes Mutual information as the similarity measure (Mattes et al., 2003). The 3D registration accuracy was calculated to be 65.0 ± 39.9 μm (mean ± SD; (Figures S2B and S2C; Experimental Procedures), which is also the accuracy for the registration of all STP experimental data sets to the RSTP brain for data analysis. The alignment of the RSTP and ABA Nissl brains was further improved by 2D affine and B-spline transformations using STP tomography-imaged CAG-Keima brain, which has a Nissl-like fluorescent labeling from the broadly expressing CAG (cytomegalovirus-IE/chicken β-actin) promoter (Figure S3A; Experimental Procedures). Finally, the alignment of many ABA anatomical labels was validated, and in some cases manually corrected, based on a comparison to brain structures delineated by tissue autofluorescence or fluorescent protein expression in parvalbumin-, glutamic acid decarboxylase-, and somatostatin-specific transgenic reporters (Taniguchi et al., 2011) (Figures S3B–S3D).

Calculation of the Sample Size for c-*fos*-GFP-Based Mapping of Mouse Brain Activation

The RSTP brain allows us to calculate the number of c-*fos*-GFP+ cells per anatomical ABA regions in the 280-section data sets. To estimate the required sample size for statistical comparisons, we used power analysis on data from a baseline group of mice (Experimental Procedures). The brains of 7 c-*fos*-GFP mice (no experimental manipulation) were imaged by STP tomography, warped to the RSTP brain, and the c-*fos*-GFP+ cells were counted per each anatomical ROI. To determine the optimal sample size, Monte Carlo methods were applied to this data to simulate ROI counts for two groups at various effect sizes. As shown in Figure 2, the number (N) of sufficiently powered ROIs ($\alpha < 0.05$, power > 0.80) increased at an approximately constant rate until N = 10, where it started to plateau. We chose N = 12–13 as sample size per group, which assures high statistical power for most ROIs.

The Selection of the Social Behavioral Protocols and Characterization of c-*fos*-GFP Induction

Interactions between a male and a female mouse, and between a male and a male mouse, include initial common social behaviors, such as anogenital sniffing and close following, and consequent sex-specific behaviors, such mounting and fighting. In the current study, we wished to focus on the comparison of brain activation evoked during the initial social exploration-based phase of the male-female and male-male interactions, during which the male is expected to recognize the social stimulus and to discriminate the sex of the interacting partner.

The social comparison was based on two experimental groups. In the male-female interaction group, an ovariectomized (OVX) conspecific female was introduced for 90 s in the home cage of a naive c-*fos*-GFP+ male, while in the male-male interaction group, a conspecific male was used as the 90 s stimulus (Movie S2). As described before in studies of social recognition (Ferguson et al., 2000, 2001), the brief interaction period included exploratory behavioral activities of anogenital sniffing, close following, and nose-to-nose sniffing, but no sexual behavior or aggression (Figure S4). The OVX female, which was recognized by the male as a social stimulus comparable

to an intact female (Figure S4), was chosen to limit experimental variability due to the estrous cycle (Ferguson et al., 2000; Winslow, 2003).

For control, we included four groups. Baseline group included mice that were not handled or otherwise manipulated. The handling group included mice that were transferred to the experimental area for 90 s, the object group included mice that received a novel object for 90 s, the olfactory group included mice that were exposed for 90 s to a novel object enriched with banana-like odor (isoamyl acetate [ISO]; note that ISO is a monomolecular odor and as such it is likely to induce simpler activation patterns compared to complex volatile odors.).

In order to characterize the time course of c-*fos*-GFP induction, we used the 90 s ISO stimulation and tested c-*fos*-GFP increase in the main olfactory bulb at 0.5, 1.5, 3, and 5 hr poststimulus. This protocol revealed a peak induction at 3 hr after the stimulation, which returned to the baseline level at 5 hr (Figures S5A–S5C). The time of 3 hr poststimulus was selected for analysis of all behavioral experiments.

In order to compare the c-*fos*-GFP signal to native c-*fos* signal, we analyzed female interaction-driven induction in eight selected brain regions by anti-c-*fos* immunohistochemistry in wild-type C57BL/6 mice and by STP tomography in c-*fos*-GFP mice (note that the c-*fos* signal was analyzed at 1 hr poststimulus because of the short half-life of the native c-*fos* protein). Overall, the c-*fos*-GFP signal represented 59% ± 6% (mean ± SEM) of anti-c-*fos* immunosignal, indicating that the direct c-*fos*-GFP fluorescence detects approximately 60% of all c-*fos* induced cells (Figures S5D–S5I). Importantly, the female interaction-driven increase was also highly comparable between the wild-type and c-*fos*-GFP mice (Figures S5D–S5I).

ROI- and Voxel-Based Statistical Analyses

The distribution of the c-*fos*-GFP+ cells among the different behavioral groups was compared using ROI- and voxel-based statistical tests corrected for multiple comparisons by false discovery rate (FDR) (Experimental Procedures). The 694 ROIs analyzed represent the segmentation of the RSTP Brain volume by the ABA anatomical regions and the c-*fos*-GFP cell counts are compared ROI-to-ROI between the experimental groups (Experimental Procedures). The RSTP brain voxelization (done by overlapping sphere voxels of 100 μm diameter) generates discrete digitization unbiased of anatomical regions and the c-*fos*-GFP cell counts are compared voxel-to-voxel (Experimental Procedures). The voxel-based statistics can reveal ''hot spot'' areas of activation and subregional differences within the anatomical ROIs (Figure S6).

In the first ROI analysis, the comparison of the male-female, male-male, olfactory, and handling groups to the baseline group revealed broad patterns of brain activation, with ~69%, 76%, 79%, and 35% of ROIs activated by the respective manipulations (Table S1). Since all ROIs activated in the handling group were also activated in the other three groups, the handling-induced brain activation represents nonspecific shared stimuli, such as moving the cage to the experimental area. In order to determine the stimulus-specific brain activations, we next compared the male-female, male-male, and olfactory groups to the handling group by both ROI and voxel-based analysis.

Female and Male Interaction-Evoked Brain Activation

It has been proposed that the detection of volatile pheromones by the main olfactory epithelium (MOE) and MOB is necessary for sex discrimination (Baum and Kelliher, 2009) (see Discussion). However, the mechanism of such detection at the level of downstream brain structures is not known. A comparison of the female and male interaction-evoked activation by ROI statistics revealed largely overlapping c-*fos*-GFP induction among MOB-connected brain regions, including the anterior olfactory nucleus (AON), piriform cortex (PIR), nucleus of the lateral olfactory tract (NLOT), anterior amygdala area (AAA), piriform-amygdala area (PAA), anterior and posterior lateral cortical amygdala (COAa, COApl), and entorhinal cortex lateral (ENTl), in addition to a female specific activation of taenia tecta (TT) and postpiriform transition area (TR) (Figure 3; Table S2; note that the heat map data in Figures 3, 4, and 5 show statistical significance, while the magnitude of c-*fos* upregulation is provided in Figure S7). A further analysis by voxel-based statistics revealed a mainly dorsal MOB activation by both stimuli and a clear dorsal-ventral separation between the two stimuli in the PIR and ENT (Figures 3C–3F; it should be noted that overlapping voxel activation between the male and female data sets, seen as yellow areas in Figures 3C–3F, represents activation of the same area, but not necessarily the same neurons). These data suggest that spatial organization of the dorsal MOB outputs leads to activation of distinct neuronal populations in the PIR and ENT, which may contribute to sex discrimination in the male brain.

The sensing of nonvolatile pheromones by the vomeronasal organ (VNO) and AOB has been proposed to play a critical role in mate recognition and behavioral motivation (Baum and Kelliher, 2009). Our analysis of brain regions downstream of the AOB revealed a strong bias toward the female interaction-evoked brain activation, including the AOB granular cell layer (AOBgr), the posterior medial cortical amygdala (COApm), the entire medial amygdala (MEA), bed nucleus of the accessory olfactory tract (BA), and bed nuclei of the stria terminalis (BST) (Figure 4; Table S2; Movie S3). In contrast, male-male interaction induced activation in fewer AOB-linked areas, including the BA and MEA anterior dorsal (ad), anterior ventral (av), and posterior dorsal (pd) (Figure 4; Table S2). Voxel analysis revealed focal activation in the AOB in the male-male interaction (Figure 4C) and a largely overlapping activation in the MEAad, av, and pd in the male-female and male-male data sets (Figures 4D and 4E; Table S2; Movie S3).

The male-female interaction also showed strongly evoked activation of brain areas linked to behavioral motivation, including the olfactory tubercle (OT) and nucleus accumbens shell (ACBsh) of the ventral striatum, prelimbic, infralimbic, and orbital medial (PL, ILA, and ORBm) prefrontal cortical areas, agranular insular cortex (AI), substantia innominata (SI; also known as ventral pallidum), medial dorsal thalamus (MDm), hippocampal ventral subiculum (SUBv), and serotonergic dorsal raphe (DR) (Figure 5A; Table S2). In contrast, the male-male interaction had a comparable induction only in the AI; much weaker activation in the prefrontal cortices, SI, OT, and SUBv; and no significant activation in the MDm and DR (Figure 5B; Table S2). Voxel-based analysis revealed that activation in the medial

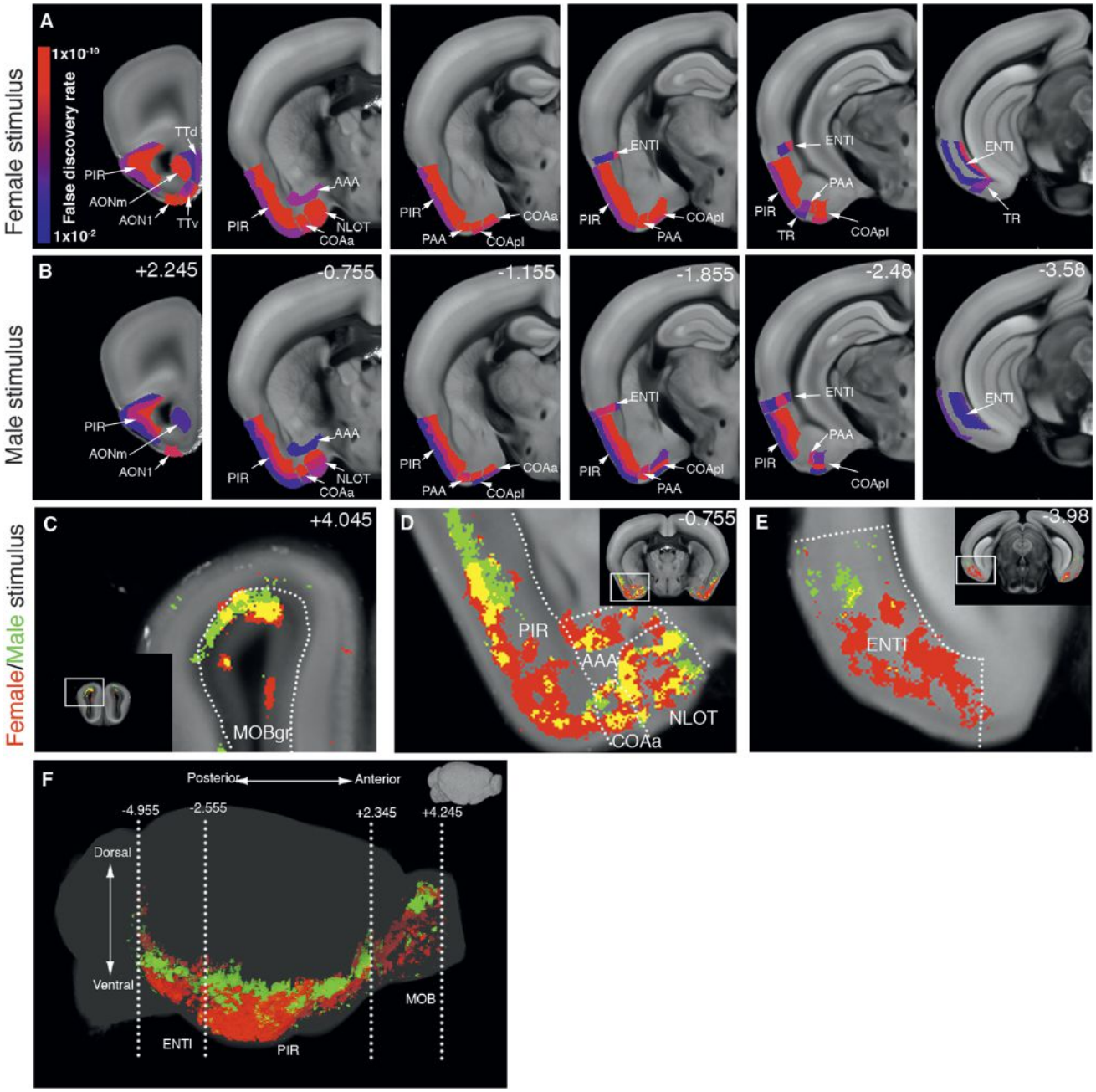


Figure 3. Social Behavior-Activated Areas: The MOB and Its Direct Downstream Circuitry

(A and B) ROI analysis:.The male-female (A) and male-male groups (B) are compared to the handling group and significantly activated ROIs downstream of the MOB are displayed. Most of the regions were activated by both stimuli. Heatmap in (A) represents FDR corrected statistical significance. Numbers in (B) represent bregma A/P coordinates. See Table S2 for ROI full names.

(C–F) Voxel-based analysis revealed activation pattern selective for the female stimulus (red), the male stimulus (green), and shared by both stimuli (yellow). (C) Both male and female stimuli induced dorsal activation in the MOB. (D–F) Dorsoventral separation was detected between the male- and female-evoked activation in the PIR (D and F) and ENT (E and F).

See also Movie S3 for the full data set.

prefrontal cortices in the male-male interaction was limited to superficial cortical layers (Figure 5C; Movie S3). We also observed a focal activation in the DR at a specific A/P bregma location in the male-female data set (Figure 5E; Movie S3).

The activation of the septal and hypothalamic nuclei is known to mediate both sexual and defensive/aggressive behaviors (Anderson, 2012; Swanson, 2000). We therefore asked whether the brief interaction used in our experiments was

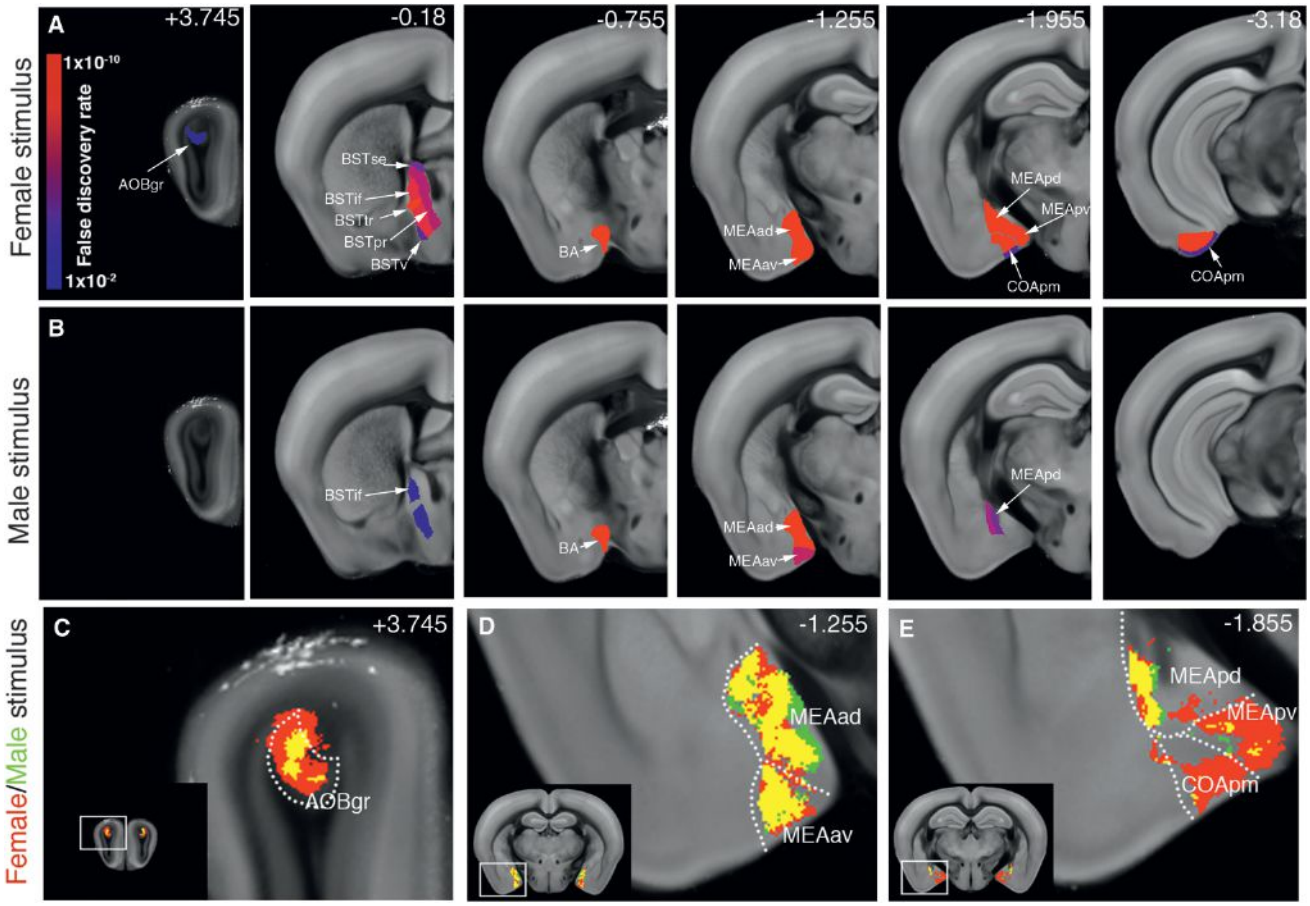


Figure 4. Social Behavior-Activated Areas: the AOB and Its Direct Downstream Circuitry

(A and B) ROI analysis. The male-female (A) and male-male groups (B) are compared to the handling group, and significantly activated ROIs downstream of the AOB are displayed. The female stimulus activated all AOB downstream regions, while the male stimulus induced only a partial activation of these areas. Heatmap in (A) represents FDR-corrected statistical significance. Numbers in (A) represent bregma A/P coordinates. See Table S2 for ROI full names.

(C–E) Voxel-based analysis revealed a largely overlapping activation pattern (yellow) in the coactivated AOB (C), MEAad and MEAav (D), and MEApd (E), and selective female-evoked activation in the MEApv and COApm (E).

See also Movie S3 for the full data set.

sufficient to activate these regions even though it lacked overt mating and fighting. The ROI analysis revealed that the female stimulus induced activation of the rostral lateral septum (LSr) and neuroendocrine nuclei, including the medial preoptic nucleus (MPN), medial preoptic area (MPO), ventral premammillary nucleus (PMv), ventrolateral part of the ventromedial nucleus (VMHvl), paraventricular hypothalamic nucleus (PVH), dorsomedial hypothalamus (DMH), anteroventral periventricular nucleus (AVPV), posterior periventricular hypothalamic nucleus (PVp), and tuberal nucleus (TU) (Figure 6A; Table S2). The male stimulus activated the VMHvl, DMH, PVH, PVp, and TU from the structures of the male-female data set, in addition to a male-specific activation of the dorsomedial part of the ventromedial nucleus (VMHdm), the anterior, preoptic, and intermediate periventricular nuclei (PVa, PVpo, PVi), retrochiasmatic area (RCH), subparaventricular zone (SBPV), supraoptic nucleus (SO), and arcuate nucleus (ARH) (Figure 6B; Figure S7; Table S2). Voxel-based analysis revealed very distinct and focal LSr activation at A/P coordinates between +0.345 and −0.145 (Figure 6C; Movie S3). The activation in the VMHvl, which was previously shown to play a role in both sexual and aggressive behaviors (Lin et al., 2011), was highly overlapping between the male-female and male-male data sets (Figure 6D; Movie S3). In addition, only a medial part of the PMv was activated in the male-male data set, suggesting a functional subdivision within this structure (Figure 6E; Movie S3).

Finally, among additional brain areas, the claustrum (CLA), basomedial amygdala (BMA), and intercalated amygdala (IA) were activated by both the female and male interactions; the capsular central amygdala (CEAc), basolateral amygdala (BLA), and thalamic parataenial nucleus (PT) were activated only in response to the female stimulus; and the temporal associational, perirhinal, and ectorhinal (TEa, PERI, and ECT) cortical areas were activated only in response to the male stimulus (Table S2). The activation of the hippocampal CA2 region linked to

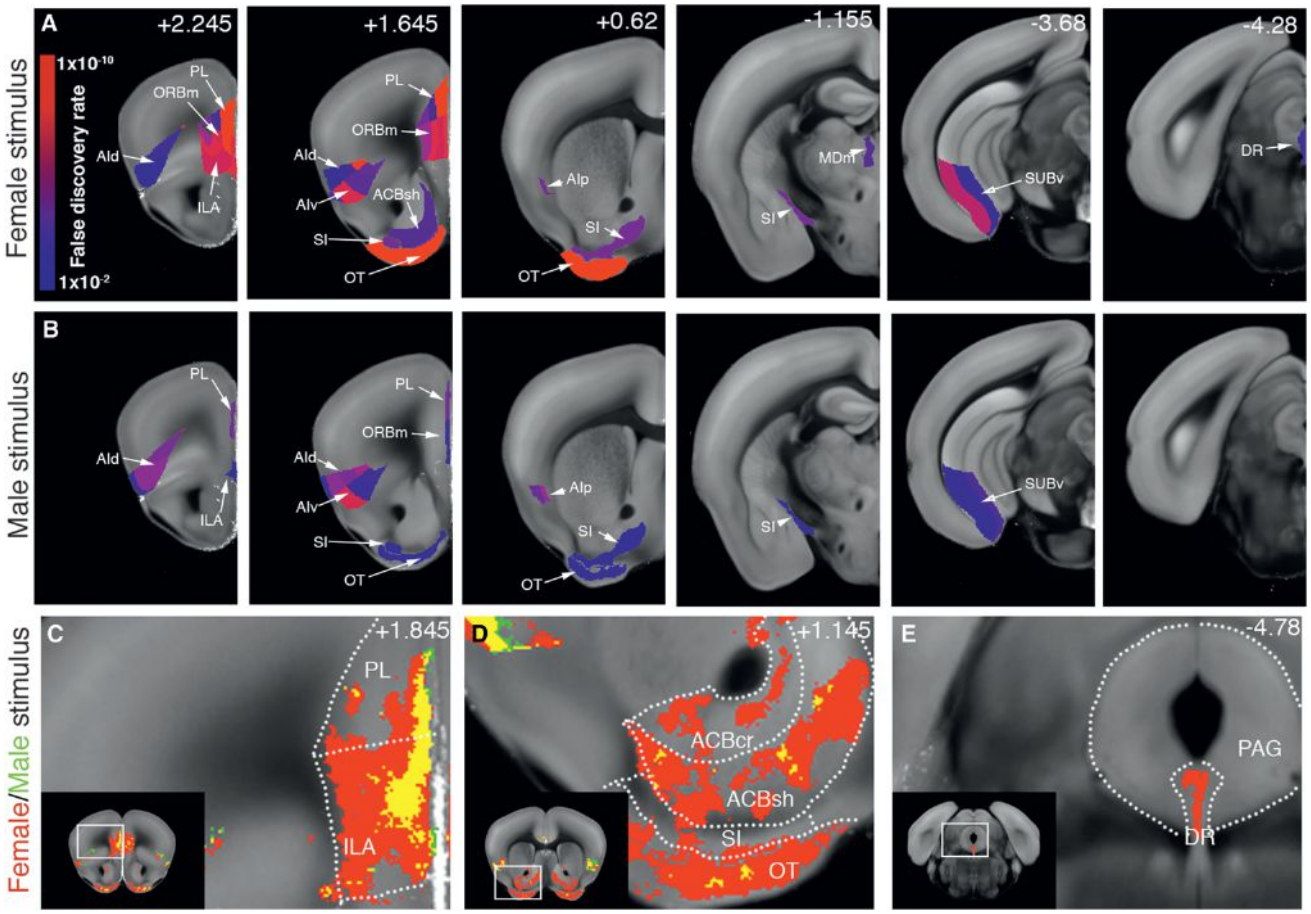


Figure 5. Social Behavior-Activated Areas: Motivational Circuitry

(A and B) ROI analysis. The male-female (A) and male-male groups (B) are compared to the handling group, and significantly activated ROIs previously implicated in behavioral motivation are displayed. The female stimulus activated frontal cortical areas (PL, ORBm, ILA, AI), ventral striatum (OT, ACB, SI), midline thalamus (MDm), ventral hippocampus (SUBv), and seretonergic DR, while the male stimulus activated AI; only superficially layer of PL, ORBm, and ILA; and weakly SI, OT, and SUBv. Heatmap in (A) represents FDR corrected statistical significance. Numbers in (A) represent bregma A/P coordinates. See Table S2 for ROI full names. (C–E) Voxel analysis showed that (C) the entire ventral part of PL and dorsal half of ILA was activated by the female stimulation, while only the upper layers of the same regions were activated by male stimulation. (D) Ventral striatum (ACB, SI, OT) showed a patch-shaped, strong activation pattern by female stimulus, but not by male stimulus. (E) Voxel analysis pinpointed the maximal activation in the DR by the female stimulus at A/P coordinate −4.78.

social memory (Hitti and Siegelbaum, 2014) was also detected in both the male-female and male-male data sets (Table S2).

Social Behavior-Specific Brain Activation

In addition to the male versus female comparison described above, we also asked which of the activated brain regions are specific to social behavior, i.e., are shared between the male-female and male-male data sets and are not activated in response to a nonsocial stimulus represented by a novel object enriched with a volatile odor (banana-like ISO).

First, we compared the ISO data set to the handling control. This analysis revealed the expected activation of the PIR and other areas downstream of the MOB, which was similar to the social behavior-evoked activation (Table S2). The activation throughout the rest of the brain, however, was highly divergent from the pattern evoked by the social stimuli, as it included many cortical areas, the entire hippocampus, and the hypothalamic subfornical organ (SFO) regulating autonomic functions (Smith and Ferguson, 2010); the suprachiasmatic nucleus (SCH) regulating sleep, waking, and locomotor activity (Saper et al., 2005); and the arcuate nucleus (ARH) linked to feeding (Sternson, 2013) (Table S2).

Second, we compared the shared male-female and male-male brain activation to the ISO data set. This analysis revealed the subset of areas specific to social behavior, which included the amygdalar regions BA, COApl, MEAav, MEApd, BMAp, BLAv, and PA, the hypothalamic VMHvl and PVH, and the SI (Figure 7; Table S3).

Correlation of c-*fos* Activation to Time Spent in Social Behaviors

The time spent in a specific behavioral activity may be expected to correlate to the number of c-*fos*-GFP+ cells in brain regions driving this activity. We next tested whether this correlation

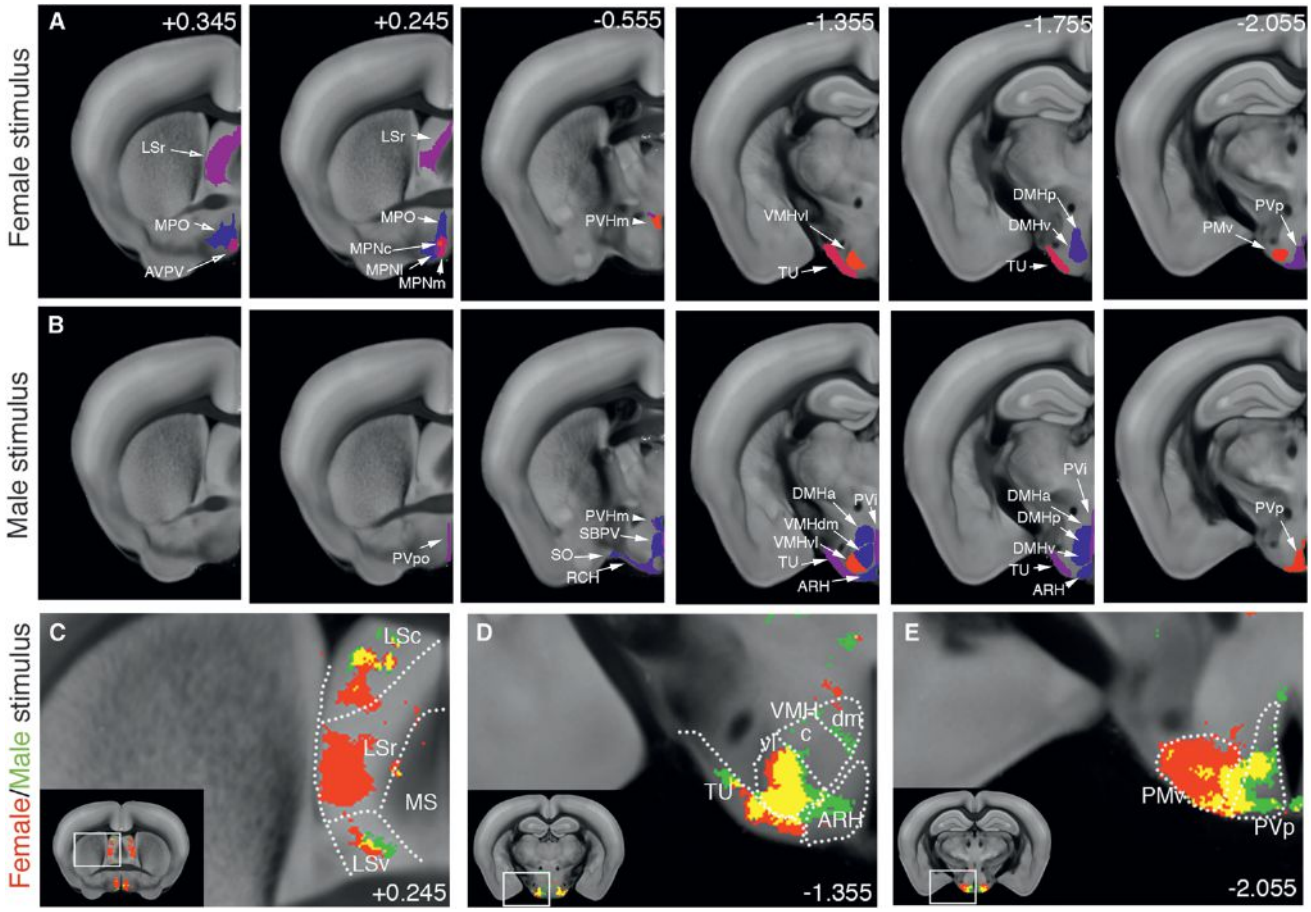


Figure 6. Social Behavior-Activated Areas: Septal and Hypothalamic Activation

(A and B) ROI analysis. The male-female (A) and male-male groups (B) are compared to the handling group, and significantly activated ROIs of the septum and hypothalamus are displayed. The female stimulus activated the rostral lateral septum (LSr), AVPV, medial preoptic area (MPO, MPN), PVH, TU, VMHvl, posterior and ventral DMH, PMv, and PVp. The male stimulus also activated the PVH, VMHvl, DMH (anterior part), TU, and PVp, in addition to a selective activation of the periventricular hypothalamic nuclei (PVpo, PVi), SBPV, RCH, SO, ARH, and VMHdm. Heatmap in (A) represents FDR-corrected statistical significance. Numbers in (A) represent bregma A/P coordinates. See Table S2 for ROI full names.

(C and D) Voxel analysis. (C) A distinct voxel activation was observed in the LSr only by female stimulation. (D) VMHvl showed largely overlapping activation by both stimuli, while VMHdm and ARH showed activation only by the male stimulus.

(E) PMv is highly activated by the female stimulus, while the medial part of PMv was also activated by the male stimulus.

See also Movie S3 for the full data set.

may be used to functionally link the activated brain areas in the male-female and male-male data sets to specific features of the social behavior.

The correlation to the time spent in anogenital sniffing identified mainly areas connected to volatile and nonvolatile olfactory signaling, such as the COAa, COApl, COApm, MEA, and BST, and hypothalamic neuroendocrine areas including the MPN, PMv, and VMHvl (Table 1). Correlation to the time spent in close following identified some of the same areas, such as the MEA and BST, but also areas linked to behavioral motivation, including the ACB, OT, SI, ILA, PL, ORBm, MDm, and DR (Table 1). Finally, the correlation to the time spent in nose-to-nose sniffing did not identify any positive association, suggesting that this behavioral feature is not quantitatively linked to any brain regions in our data sets. These data suggest that distinct aspects of the social behavior engage distinct sets of brain areas and that whole-brain cellular c-*fos*-GFP analysis is able to reveal this structure-function relationship.

Calculation of the Density of c-*fos*-GFP+ Cells per ROIs

While the above analyses identified the activated brain areas, the cellular resolution of our data allowed us to also estimate the total numbers and the densities of c-*fos*-GFP+ cells per anatomical ROIs. Since the *z* planes in the 280-section data sets are spaced 50 μm apart, we transformed the serial 2D data into 3D whole-brain estimates with a stereological method (Williams and Rakic, 1988) applied to a high-resolution 5,600-section data set with *z* spacing of 2.5 μm (Experimental Procedures). The obtained 2D-to-3D conversion factor of 2.5 was then used to multiply the 2D ROI counts in order to

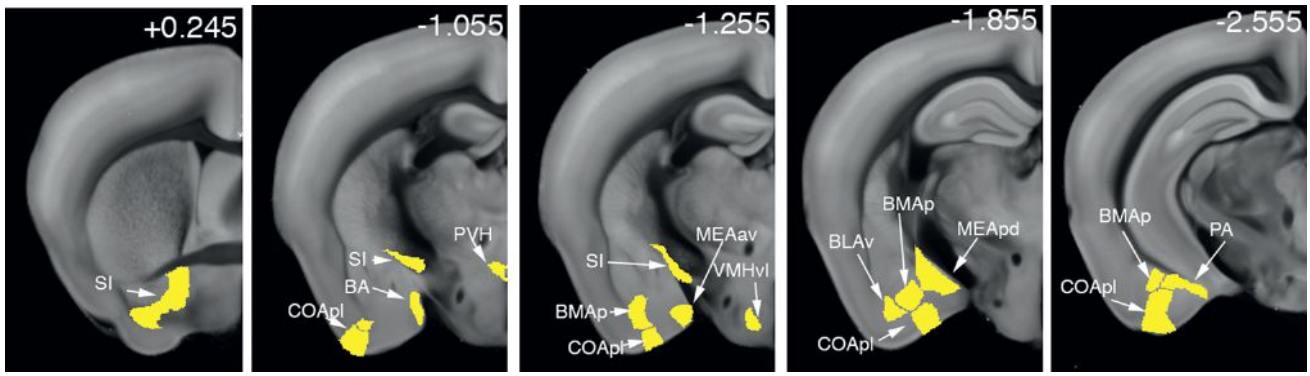


Figure 7. Social Behavior-Specific Brain Areas

Brain areas activated by both female and male stimulus, but not by ISO stimulation, are displayed as ROIs. Unique social behavior-activated areas included the amygdalar BA, COApl, MEAav, MEApd, BLAv, BMAp, PA, hypothalamic PVH, VMHvl, and ventral pallidum (SI).

estimate the total numbers of c-*fos*-GFP+ cells, and the total counts were divided by the ROI volumes in order to estimate the densities of c-*fos*-GFP+ cells per activated ROIs (Figure S7).

The average cell density in the structures significantly activated in the female and male data sets were, respectively, 4,993 ± 400 and 4,519 ± 283 per cubic mm (mean ± SEM), whereas the average density in these structures in the handling control was 3,127 ± 201 per cubic mm (Figure S7). Therefore, the social interactions evoked on average ~1,500 to 2,000 c-*fos*-GFP+ cells per cubic mm compared to the handling control, suggesting a sparse activation of a few percent of neurons per brain areas (see Discussion).

DISCUSSION

While the general organization of the brain structures regulating sexual and aggressive behavior is beginning to be understood (Anderson, 2012; Sokolowski and Corbin, 2012), much remains unknown about how information is processed from the sensory periphery (the olfactory system in rodents) to give rise to sex-specific behavioral responses. Here, utilizing a pipeline of computational methods, including ROI-based whole-brain mapping of c-*fos* activation, voxel-based mapping of subregional differences in c-*fos* activation, and correlation analysis linking ROI activation to behavior, we compared brief female interaction-evoked activation in the brain of a male mouse to the activation evoked by brief interaction with a male. Some more salient findings from our analyses are discussed below following the method discussion, while the complete ROI- and voxel-based results are provided as a resource in Tables S1, S2, and S3 and Movie S3.

The Method Pipeline for c-*fos*-GFP-Based Mouse Brain Screening

The entire method pipeline is automated, highly standardized and operates at a reasonably high-throughput: the imaging time per one brain is ~21 hr, while the imaged processing and computational analyses take ~24 hr that occur in parallel with the STP imaging (Ragan et al., 2012; Vousden et al., 2014).

The first key part of the computational pipeline is the detection of c-*fos*-GFP+ cells in the STP data sets. We chose to use CNs, because these algorithms rely on the learning procedure to account for signal to noise ratio variability and improved performance is achieved by simply increasing the training data set (V. Jain et al., 2007, IEEE, conference; Turaga et al., 2010). The trained CN performance (F-score = 0.88) was in fact close to human expert performance (F-score = 0.9), demonstrating the power of this approach for analysis of fluorescent labeling in STP tomography-imaged mouse brains. We have also tested two other cell detection methods—cell counting in the Volocity Image analysis software (Perkin Elmer) and cell counting based on watershed algorithm (Kopec et al., 2011)—but these were considerable less reliable (F score < 0.5) compared to the CN-based detection.

The second critical step of the method is the registration of the data sets to the RSTP Brain and the Allen Mouse Brain Atlas. Fixation-induced tissue autofluorescence provides rich image content for the registration by the warping algorithm Elastix (Mattes et al., 2003). As a result, we were able to achieve a high level of precision (~60 μm jitter) for the registration of the experimental data sets to the RSTP brain (Figure S2). The alignment of the ABA Nissl-stained sections to the RSTP brain was further helped by the use of the transgenic CAG-Keima brain with a cellular fluorescent protein labeling that matched in most brain regions the cellular Nissl signal and by several interneuron-specific reporter mice (Taniguchi et al., 2011) that helped to validate and improve the matching of the labels to specific brain nuclei (Figure S3). Consequently, the precision ABA labels became closely aligned to the RSTP brain, as judged based on brain landmarks, such as the corpus callosum, hippocampal pyramidal layers, and many structural borders visible in the autofluorescence signal (Figure S3).

The last part of the method pipeline includes statistical analyses of the brain-wide c-*fos*-GFP+ cell counts. Since it was first established in rat models of seizure, the inducibility of c-*fos* has been utilized to map neuronal activation in many behavioral and pharmacological experiments, demonstrating that c-*fos* can be used as an activity reporter in most if not all areas of the brain (Dragunow and Robertson, 1987; Morgan et al., 1987). The

Table 1. c-fos-GFP Count to the Social Behavioral Correlation

ROIs	Anogentical Sniffing	Close Following
Isocortex		
ILA		+
PL	+	++
ORBm	+	++
TT		++
Olfactory area		
DP		++
AOBgr	++	
COAa	+	
COApl	+++	+
COApm	+++	+
PAA	+++	+
TR	+	
Hippocampal formation		
ENTmv	++	
CA3		+
Cortical subplate		
EP	+	+
BLAa	++	+
BLAp	+	
BLAv	++	
BMAa	++	
BMAp	+++	
PA	+++	
Cerebral nuclei		
ACBsh		+
OT	+	++
AAA	+	
LSr		+
CEAc	+	+
IA	++	
MEAad	+++	
MEAav	+++	
MEApd	+++	+
MEApv	+++	+
SI		+
BSTmg	+	++
BSTv	+	++
BSTp	++	
Thalamus		
MDm		+
PT		+
Hypothalamus		
AVPV	++	+
MPN	++	+
PMv	+++	+
VMHvl	+	
TU	++	
Midbrain		
DR		+

Pearson correlation between the time spent in anogenital sniffing and close following and c-fos-GFP cell counts in the regions activated in the male-female and male-male data sets. Significance is based on FDR q value adjusted for multiple comparisons: (+) = 0.01 $\leq$ FDR q < 0.05; (++) = 0.001 $\leq$ FDR q < 0.01; and (+++) = FDR q < 0.001.

ROI- and voxel-based statistical analyses established here transform the traditional laborious immunostaining or in situ hybridization based c-*fos* mapping into an automated whole-brain assay.

In addition to the current application, these methods can also be used to detect and quantify other fluorescent protein-expressing transgenic mouse brains by simply training new CN on a different ground-truth data. This makes our pipeline easily adaptable to many other applications in quantitative whole-brain mapping, such as the generation of whole-brain cell counts in cell type-specific GFP reporter mice (Taniguchi et al., 2011).

Female- and Male-Evoked Maps of Whole-Brain Activation in the Male Brain

By focusing on the initial period of social exploratory behaviors between a naive male and a novel conspecific female or male mouse, we set out to determine the brain activation patterns that underlie social recognition and sex discrimination in the male brain. Our results revealed that while the brief interactions led to an activation of the expected sex-specific response at the hypothalamic level (indicating that the behaviors were sufficient for correct sex discrimination), the upstream patterns of brain activation strongly diverged between the two stimuli.

At the level of the AOB and MOB signaling, the female stimulus evoked activation of all downstream connected brain structures, while the male stimulus showed activation of all MOB-linked structures but only a subset of the AOB-linked structures. The strong MOB-driven brain activation in both behaviors agrees with the role of volatile signaling in sex discrimination proposed by studies using chemical lesion of the MOE (Keller et al., 2006) or genetic disruption of cellular signaling in the MOE (Mandiyan et al., 2005). The finding that the male and female stimuli activate different parts of the PIR and ENT areas suggests that topologically distinct MOB cortical outputs may discriminate the sex-specific stimuli. This dorsoventral separation is an example of a novel spatial organization in the piriform cortex, which until now has been considered to lack gross sensory input-based topology (Ghosh et al., 2011; Sosulski et al., 2011).

The role of the VNO and AOB-driven activation in social behaviors appears to be less clear than that of the MOE/MOB signaling. Lesioning of the VNO failed to affect sex discrimination in male mice (Pankevich et al., 2004), even though it did impair vocalization after nasal contact with female urine (Bean, 1982), while genetic disruption of VNO signaling caused male-male mounting instead of aggressive behavior without affecting male-female behavior (Stowers et al., 2002). Our data point to a more prominent role of the AOB-connected brain structures

in the male-female interaction, as MEA, BST, BA, and COApm were all activated in the male-female data set, but only MEA and BA were activated in the male-male data set. The selective BST activation in the male-female data set included the posterior division nuclei (principal, interfascicular, and transverse) proposed to function in reproductive behaviors (Dong and Swanson, 2004), and the magnocellular nucleus of the anterior division proposed to control neuroendocrine functions and pelvic functions, including penile erection (Dong and Swanson, 2006).

The female, but not the male, stimulus also evoked activation of brain areas of the striatopallidothalamocortical circuit known to positively regulate behavioral motivation (Ikemoto, 2007; Sesack and Grace, 2010), including the ventral striatum (OT, ACB), ventral pallidum (SI), thalamus (MDm), and prefrontal cortex (ILA, PL, ORB). While we did not detect activation of the dopaminergic neurons of the ventral tegmental area (VTA), which are known to reinforce ACB functions within this circuit during sexual behavior (Ikemoto, 2007; Sesack and Grace, 2010), we did detect activation of the serotonergic DR, which was recently shown to be necessary for ACB functions in social reward (Dölen et al., 2013). The switch between the DR and VTA modulation of ventral striatum may contribute to a transition between exploratory and consummatory male-female behavior.

The analysis of the hypothalamic brain areas revealed activation of structures regulating sexual and aggressive behaviors (Anderson, 2012; Swanson, 2000): the MPN and PMv regulating male reproductive behavior (Simerly, 2002; Yang and Shah, 2014) were selectively activated in the male-female data set, the VMHvl regulating both male sexual behavior and aggression (Anderson, 2012; Lin et al., 2011; Yang et al., 2013) was activated in response to both female and male stimuli, and the VMHdm regulating male defensive behaviors (Lin et al., 2011; Sokolowski and Corbin, 2012) was activated only in the male-male data set. Since the brief social interactions did not comprise mating or aggression, these data show that the activation of the hypothalamic nuclei can precede the manifestation of these behaviors as part of the male-female and male-male social exploration-based behaviors.

The Quantification of the Whole-Brain Activation Maps at Cellular Level

The cellular resolution of our data also allowed us to search for correlations between behavioral activity and brain activation and to estimate the density of activated cells per brain area.

The correlation between behavior and c-*fos* activation can be expected to identify the most behaviorally relevant brain regions in which the number of c-*fos* activated cells reflects the behavioral performance in individual animals. In agreement with this hypothesis, regions correlated to the time spent in anogenital sniffing included mainly amygdalar and hypothalamic areas of the vomeronasal sensory-motor system transforming the chemosensory information into sexual or aggressive behavior (Swanson, 2000), while the brain areas correlated to the time spent in following included the structures linked to behavioral motivation and described above as part of the striatopallidothalamocortical circuit.

The correlation analysis can be used to add functional significance to activated regions that were not previously known to be involved in social behaviors. For example, the activation of the amygdalar IA and CEAc nuclei was correlated to the anogenital sniffing time, while the PT thalamus activation was correlated to following. Since both IA and CEAc can inhibit the medial central amygdala (CEAm), which is the output fear pathway (Pitkänen et al., 1997), these data suggest that the IA and CEAc are activated by chemosensory cues and may act to modulate fear behaviors during social exploration. The PT, a part of the dorsal group of thalamic nuclei, projects to the ACB (Kelley and Stinus, 1984) and may play a role in motivational modulation of the male-female social behavior.

Finally, the quantification of the numbers of c-*fos*-GFP+ cells per brain area can provide information about the approximate percentage of neurons behaviorally recruited in the identified brain areas. For example, we observed on average 2-fold increase (~3,600 c-*fos*-GFP+ cells per cubic mm) in the prefrontal cortical areas in the male-female data sets, compared to the handling control (Figure S7). Since neuronal density in the mouse cortex is estimated at ~80,000–100,000 per cubic mm (Herculano-Houzel et al., 2006; Keller and Carlson, 1999; Meyer et al., 2010), these data suggest less than 5% of neurons is recruited in response to the female social stimulus. Further, as most brain areas showed similar c-*fos*-GFP+ densities, the behavioral recruitment of a few percent of neurons is likely a general feature of c-*fos* activation. This may represent c-*fos* induction occurring only in the most strongly activated cells, and such sparse c-*fos* induction may be relevant for the proposed sparse coding of sensory inputs (Olshausen and Field, 2004).

Caveats of the Current Study

There are several caveats associated with our study. First, while the behavior is limited to 90 s, our assay cannot determine whether the observed c-*fos*-GFP induction occurred entirely during this brief time period or whether some downstream activation occurred during a longer time interval. Second, the behavioral paradigm includes both the introduction and removal of the stimulus animal from the home cage of the c-*fos*-GFP male, and some of the observed activation pattern thus may reflect stress induced by these manipulations. Third, the use of the OVX females in our study restricts the interpretation of the male-female activation data to social exploration that lacks the effects of estrous hormones. Thus, male-female interaction with, for example, estradiol-induced OVX mice may be expected to induce brain activation partially distinct from the one described in the current study. Fourth, since the c-*fos*-GFP reporter labels ~60% of all c-*fos*+ cells detected by immunostaining, it is possible that some areas with native c-*fos* activation were missed in our assay. Finally, fifth, c-*fos* is a member of a family of IEGs regulated by neuronal activity, and the detection of other IEGs (such as Arc, homer-1A, or zif-268) can be expected to reveal partially overlapping activation maps compared to the c-*fos*-GFP map identified in our paper. Because neuronal activation in some brain areas may induce other IEGs but fail to induce c-*fos*, the c-*fos*-GFP-based network of brain areas described here should not be interpreted as a complete brain activation map evoked by social behavior.

CONCLUSIONS

Our method of c-*fos*-GFP-based screening generates cellular-resolution maps of behaviorally evoked whole-brain activation in the mouse. The patterns of female and male interaction-evoked brain activation revealed clear separation between the two stimuli, including at the level of brain structures downstream of both volatile and nonvolatile chemosensory signaling. These activation patterns were also markedly different from the activation pattern evoked during nonsocial olfactory-enhanced exploratory behavior. These findings demonstrate that our method can be used for screening behavior-evoked whole-brain activation, and we envision that future experiments will yield brain-map-like descriptions for other innate behaviors, such as aggression and defensive behaviors, or cognitive behaviors, such as attention and decision making. Further, the same method can be applied to genetic mouse models of neurodevelopmental disorders with the aim of identifying circuit deficits underlying changes in social, cognitive, and other higher-order brain functions.

EXPERIMENTAL PROCEDURES

Animals

Animal procedures were approved by the Cold Spring Harbor Laboratory Animal Care and Use Committee. The c-*fos*-GFP mice, Tg(Fos-tTA,Fos-EGFP) line, were obtained from The Jackson Laboratory. In our study, we used the direct c-*fos*-GFP signal, whereas several other studies used the tTA protein to drive other reporter molecules (Garner et al., 2012; Liu et al., 2012; Matsuo et al., 2008; Reijmers et al., 2007).

Behavioral Tests and c-fos-GFP Induction Time Course

Heterozygous c-*fos*-GFP male mice (8–11 weeks old) were individually housed for 1 week before the test. The behavioral stimuli were transfer of the animal to the experimental arena (handling control) or plus introduction of an OVX conspecific female (male-female group), conspecific male (male-male group), 50 ml falcon tube (object group), and 50 ml falcon tube with a side opening in which was cotton ball with isoamyl acetate (1:100 in mineral oil, 40 μl per experiment, freshly made each day). The stimulus was placed in the home cage for 90 s and then removed. The behavior was video-recorded and manually scored. After the behavioral stimulus was removed, the mice remained in the home cage for additional 3 hr and then killed by transcardial perfusion. For the time course of c-*fos*-GFP induction, isoamylacetate was introduced into the mouse home cage for a brief period of 90 s. The mice were killed at selected time points of 0.5, 1.5, 3, and 5 hr poststimulation.

Brain Preparation, STP Tomography Imaging, and Data Processing

The brains were prepared as described in our previous study (Ragan et al., 2012). Briefly, the brains were embedded in oxidized 4% agarose, crosslinked, and imaged as 280 serial sections. The raw image tiles were corrected for illumination and stitched in 2D in MATLAB and aligned in 3D in Fiji (Ragan et al., 2012). The CNs for detection of c-*fos*-GFP+ cells was trained based on ground truth data marked up by an expert biologist. The CN performance was scored based on the F-score (the harmonic mean of the precision and recall). Stereological procedure was used to calculate how CNs 2D based counting can be converted into 3D counting to calculate the densities of c-*fos*-GFP+ cells per activated ROIs. 3D registration methods with Elastix were the same as described previously (Ragan et al., 2012), but with modified parameters. See the Supplemental Experimental Procedures for more details.

c-*fos* Immunohistochemistry and Comparison to c-*fos*-GFP+ Cell Counting

Wild-type C57BL/6 mice (8–10 weeks old) underwent the same behaviors as the c-*fos*-GFP mice of the male-to-female and handling groups. The mice were killed and perfused 1 hr later and the brains were fixed overnight in 4% paraformaldehyde, then cut as 50 μm coronal sections. For immunohistochemistry, sections were exposed to rabbit anti-c-*fos* antibody (1:10,000, Santa Cruz SC052) and labeled by DAB solution. FIJI (ImageJ) and Volocity (Perkin-Elmer) were used for cell counting.

Statistics

We ran statistical comparisons between different behavioral groups based on either ROIs or evenly spaced voxels. Voxels were overlapping 3D spheres with 100 μm diameter each and spaced 20 μm apart from each other. The cell count of each voxel was calculated as the number of nuclei found within 100 μm from the center of the voxel in all 3D. To account for multiple comparisons across all voxel/ROI locations, we thresholded the p values and reported FDRs. For correlation between c-*fos*-GFP cell counts and social behavior, Pearson correlation R values were calculated between c-*fos*-GFP cell counts and time spent in social behaviors. See the Supplemental Experimental Procedures for more details.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, three tables, and three movies and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.12.014.

ACKNOWLEDGMENTS

We thank G. Fitzgerald, R. Palaniswamy, and J. Zambratto for expert technical assistance, J. Kuhl for comments and graphics work, W. Denk and K. Baldwin for comments on the manuscript, and members of the P.O. lab for helpful discussions. P.O. is supported by the National Institute of Mental Health grant 1R01MH096946-01, Simons Foundation for Autism Research grants 204719 and 253447, and CSHL; Y.K. is supported by NARSAD Young Investigator Grant, M.H. is supported by the Allen Institute for Brain Science, and H.S.S. is supported by Howard Hughes Medical Institute and Gatsby Charitable Foundation.

Received: August 24, 2014
Revised: October 21, 2014
Accepted: December 5, 2014
Published: December 31, 2014

REFERENCES

Anderson, D.J. (2012). Optogenetics, sex, and violence in the brain: implications for psychiatry. Biol. Psychiatry *71*, 1081–1089.

Baum, M.J., and Kelliher, K.R. (2009). Complementary roles of the main and accessory olfactory systems in mammalian mate recognition. Annu. Rev. Physiol. *71*, 141–160.

Bean, N.J. (1982). Olfactory and vomeronasal mediation of ultrasonic vocalizations in male mice. Physiol. Behav. *28*, 31–37.

Biały, M., and Kaczmarek, L. (1996). c-Fos expression as a tool to search for the neurobiological base of the sexual behaviour of males. Acta Neurobiol. Exp. (Warsz.) *56*, 567–577.

Brennan, P.A., and Zufall, F. (2006). Pheromonal communication in vertebrates. Nature *444*, 308–315.

Buzsáki, G. (2004). Large-scale recording of neuronal ensembles. Nat. Neurosci. *7*, 446–451.

Clayton, D.F. (2000). The genomic action potential. Neurobiol. Learn. Mem. *74*, 185–216.

Coolen, L.M., Peters, H.J., and Veening, J.G. (1996). Fos immunoreactivity in the rat brain following consummatory elements of sexual behavior: a sex comparison. Brain Res. *738*, 67–82.

Dölen, G., Darvishzadeh, A., Huang, K.W., and Malenka, R.C. (2013). Social reward requires coordinated activity of nucleus accumbens oxytocin and serotonin. Nature *501*, 179–184.

Dong, H.W., and Swanson, L.W. (2004). Projections from bed nuclei of the stria terminalis, posterior division: implications for cerebral hemisphere regulation of defensive and reproductive behaviors. J. Comp. Neurol. *471*, 396–433.

Dong, H.W., and Swanson, L.W. (2006). Projections from bed nuclei of the stria terminalis, magnocellular nucleus: implications for cerebral hemisphere regulation of micturition, defecation, and penile erection. J. Comp. Neurol. *494*, 108–141.

Dragunow, M., and Robertson, H.A. (1987). Kindling stimulation induces c-fos protein(s) in granule cells of the rat dentate gyrus. Nature *329*, 441–442.

Fenno, L., Yizhar, O., and Deisseroth, K. (2011). The development and application of optogenetics. Annu. Rev. Neurosci. *34*, 389–412.

Ferguson, J.N., Young, L.J., Hearn, E.F., Matzuk, M.M., Insel, T.R., and Winslow, J.T. (2000). Social amnesia in mice lacking the oxytocin gene. Nat. Genet. *25*, 284–288.

Ferguson, J.N., Aldag, J.M., Insel, T.R., and Young, L.J. (2001). Oxytocin in the medial amygdala is essential for social recognition in the mouse. J. Neurosci. *21*, 8278–8285.

Garner, A.R., Rowland, D.C., Hwang, S.Y., Baumgaertel, K., Roth, B.L., Kentros, C., and Mayford, M. (2012). Generation of a synthetic memory trace. Science *335*, 1513–1516.

Ghosh, S., Larson, S.D., Hefzi, H., Marnoy, Z., Cutforth, T., Dokka, K., and Baldwin, K.K. (2011). Sensory maps in the olfactory cortex defined by long-range viral tracing of single neurons. Nature *472*, 217–220.

Grewe, B.F., and Helmchen, F. (2009). Optical probing of neuronal ensemble activity. Curr. Opin. Neurobiol. *19*, 520–529.

Guzowski, J.F., Timlin, J.A., Roysam, B., McNaughton, B.L., Worley, P.F., and Barnes, C.A. (2005). Mapping behaviorally relevant neural circuits with immediate-early gene expression. Curr. Opin. Neurobiol. *15*, 599–606.

Herculano-Houzel, S., Mota, B., and Lent, R. (2006). Cellular scaling rules for rodent brains. Proc. Natl. Acad. Sci. USA *103*, 12138–12143.

Hitti, F.L., and Siegelbaum, S.A. (2014). The hippocampal CA2 region is essential for social memory. Nature *508*, 88–92.

Ikemoto, S. (2007). Dopamine reward circuitry: two projection systems from the ventral midbrain to the nucleus accumbens-olfactory tubercle complex. Brain Res. Brain Res. Rev. *56*, 27–78.

Keller, A., and Carlson, G.C. (1999). Neonatal whisker clipping alters intracortical, but not thalamocortical projections, in rat barrel cortex. J. Comp. Neurol. *412*, 83–94.

Keller, M., Douhard, Q., Baum, M.J., and Bakker, J. (2006). Sexual experience does not compensate for the disruptive effects of zinc sulfate—lesioning of the main olfactory epithelium on sexual behavior in male mice. Chem. Senses *31*, 753–762.

Kelley, A.E., and Stinus, L. (1984). The distribution of the projection from the parataenial nucleus of the thalamus to the nucleus accumbens in the rat: an autoradiographic study. Exp. Brain Res. *54*, 499–512.

Kopec, C.D., Bowers, A.C., Pai, S., and Brody, C.D. (2011). Semi-automated atlas-based analysis of brain histological sections. J. Neurosci. Methods *196*, 12–19.

Lee, H.M., Giguere, P.M., and Roth, B.L. (2014). DREADDs: novel tools for drug discovery and development. Drug Discov. Today *19*, 469–473.

Lin, D., Boyle, M.P., Dollar, P., Lee, H., Lein, E.S., Perona, P., and Anderson, D.J. (2011). Functional identification of an aggression locus in the mouse hypothalamus. Nature *470*, 221–226.

Liu, X., Ramirez, S., Pang, P.T., Puryear, C.B., Govindarajan, A., Deisseroth, K., and Tonegawa, S. (2012). Optogenetic stimulation of a hippocampal engram activates fear memory recall. Nature *484*, 381–385.

Mandiyan, V.S., Coats, J.K., and Shah, N.M. (2005). Deficits in sexual and aggressive behaviors in Cnga2 mutant mice. Nat. Neurosci. *8*, 1660–1662.

Matsuo, N., Reijmers, L., and Mayford, M. (2008). Spine-type-specific recruitment of newly synthesized AMPA receptors with learning. Science *319*, 1104–1107.

Mattes, D., Haynor, D.R., Vesselle, H., Lewellen, T.K., and Eubank, W. (2003). PET-CT image registration in the chest using free-form deformations. IEEE Trans. Med. Imaging *22*, 120–128.

Meyer, H.S., Wimmer, V.C., Oberlaender, M., de Kock, C.P., Sakmann, B., and Helmstaedter, M. (2010). Number and laminar distribution of neurons in a thalamocortical projection column of rat vibrissal cortex. Cereb. Cortex *20*, 2277–2286.

Morgan, J.I., Cohen, D.R., Hempstead, J.L., and Curran, T. (1987). Mapping patterns of c-fos expression in the central nervous system after seizure. Science *237*, 192–197.

Olshausen, B.A., and Field, D.J. (2004). Sparse coding of sensory inputs. Curr. Opin. Neurobiol. *14*, 481–487.

Pankevich, D.E., Baum, M.J., and Cherry, J.A. (2004). Olfactory sex discrimination persists, whereas the preference for urinary odorants from estrous females disappears in male mice after vomeronasal organ removal. J. Neurosci. *24*, 9451–9457.

Pfaus, J.G., and Heeb, M.M. (1997). Implications of immediate-early gene induction in the brain following sexual stimulation of female and male rodents. Brain Res. Bull. *44*, 397–407.

Pitkänen, A., Savander, V., and LeDoux, J.E. (1997). Organization of intra-amygdaloid circuitries in the rat: an emerging framework for understanding functions of the amygdala. Trends Neurosci. *20*, 517–523.

Ragan, T., Kadiri, L.R., Venkataraju, K.U., Bahlmann, K., Sutin, J., Taranda, J., Arganda-Carreras, I., Kim, Y., Seung, H.S., and Osten, P. (2012). Serial two-photon tomography for automated ex vivo mouse brain imaging. Nat. Methods *9*, 255–258.

Reijmers, L.G., Perkins, B.L., Matsuo, N., and Mayford, M. (2007). Localization of a stable neural correlate of associative memory. Science *317*, 1230–1233.

Saper, C.B., Lu, J., Chou, T.C., and Gooley, J. (2005). The hypothalamic integrator for circadian rhythms. Trends Neurosci. *28*, 152–157.

Sesack, S.R., and Grace, A.A. (2010). Cortico-Basal Ganglia reward network: microcircuitry. Neuropsychopharmacology *35*, 27–47.

Simerly, R.B. (2002). Wired for reproduction: organization and development of sexually dimorphic circuits in the mammalian forebrain. Annu. Rev. Neurosci. *25*, 507–536.

Smith, P.M., and Ferguson, A.V. (2010). Circulating signals as critical regulators of autonomic state—central roles for the subfornical organ. Am. J. Physiol. Regul. Integr. Comp. Physiol. *299*, R405–R415.

Sokolowski, K., and Corbin, J.G. (2012). Wired for behaviors: from development to function of innate limbic system circuitry. Front. Mol. Neurosci. *5*, 55.

Sosulski, D.L., Bloom, M.L., Cutforth, T., Axel, R., and Datta, S.R. (2011). Distinct representations of olfactory information in different cortical centres. Nature *472*, 213–216.

Sternson, S.M. (2013). Hypothalamic survival circuits: blueprints for purposive behaviors. Neuron *77*, 810–824.

Stowers, L., Holy, T.E., Meister, M., Dulac, C., and Koentges, G. (2002). Loss of sex discrimination and male-male aggression in mice deficient for TRP2. Science *295*, 1493–1500.

Sunkin, S.M., Ng, L., Lau, C., Dolbeare, T., Gilbert, T.L., Thompson, C.L., Hawrylycz, M., and Dang, C. (2013). Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Res. *41*, D996–D1008.

Swanson, L.W. (2000). Cerebral hemisphere regulation of motivated behavior. Brain Res. *886*, 113–164.

Taniguchi, H., He, M., Wu, P., Kim, S., Paik, R., Sugino, K., Kvitsiani, D., Fu, Y., Lu, J., Lin, Y., et al. (2011). A resource of Cre driver lines for genetic targeting of GABAergic neurons in cerebral cortex. Neuron *71*, 995–1013.

Turaga, S.C., Murray, J.F., Jain, V., Roth, F., Helmstaedter, M., Briggman, K., Denk, W., and Seung, H.S. (2010). Convolutional networks can learn to generate affinity graphs for image segmentation. Neural Comput. *22*, 511–538.

Veening, J.G., Coolen, L.M., de Jong, T.R., Joosten, H.W., de Boer, S.F., Koolhaas, J.M., and Olivier, B. (2005). Do similar neural systems subserve aggressive and sexual behaviour in male rats? Insights from c-Fos and pharmacological studies. Eur. J. Pharmacol. *526*, 226–239.

Vousden, D.A., Epp, J., Okuno, H., Nieman, B.J., van Eede, M., Dazai, J., Ragan, T., Bito, H., Frankland, P.W., Lerch, J.P., and Henkelman, R.M. (2014). Whole-brain mapping of behaviourally induced neural activation in mice. Brain Struct. Funct.

Williams, R.W., and Rakic, P. (1988). Three-dimensional counting: an accurate and direct method to estimate numbers of cells in sectioned material. J. Comp. Neurol. *278*, 344–352.

Winslow, J.T. (2003). Mouse social recognition and preference. Curr. Protoc. Neurosci. *Chapter 8*, 16.

Yang, C.F., Chiang, M.C., Gray, D.C., Prabhakaran, M., Alvarado, M., Juntti, S.A., Unger, E.K., Wells, J.A., and Shah, N.M. (2013). Sexually dimorphic neurons in the ventromedial hypothalamus govern mating in both sexes and aggression in males. Cell *153*, 896–909.

Yang, C.F., and Shah, N.M. (2014). Representing sex in the brain, one module at a time. Neuron *82*, 261–278.

OPEN ACCESS
CellPress

Cell Reports
Resource

Insights into the Evolution of Longevity from the Bowhead Whale Genome

Michael Keane,[1,18] Jeremy Semeiks,[2,18] Andrew E. Webb,[3,18] Yang I. Li,[4,18,19] Víctor Quesada,[5,18] Thomas Craig,[1] Lone Bruhn Madsen,[6] Sipko van Dam,[1] David Brawand,[4] Patrícia I. Marques,[5] Pawel Michalak,[7] Lin Kang,[7] Jong Bhak,[8] Hyung-Soon Yim,[9] Nick V. Grishin,[2] Nynne Hjort Nielsen,[10] Mads Peter Heide-Jørgensen,[10] Elias M. Oziolor,[11] Cole W. Matson,[11] George M. Church,[12] Gary W. Stuart,[13] John C. Patton,[14] J. Craig George,[15] Robert Suydam,[15] Knud Larsen,[6] Carlos López-Otín,[5] Mary J. O'Connell,[3] John W. Bickham,[16,17] Bo Thomsen,[6] and João Pedro de Magalhães[1,*]

[1]Integrative Genomics of Ageing Group, Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK
[2]Howard Hughes Medical Institute and Departments of Biophysics and Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75390-9050, USA
[3]Bioinformatics and Molecular Evolution Group, School of Biotechnology, Dublin City University, Glasnevin, Dublin 9, Ireland
[4]MRC Functional Genomics Unit, University of Oxford, Oxford OX1 3QX, UK
[5]Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología (IUOPA), Universidad de Oviedo, 33006 Oviedo, Spain
[6]Department of Molecular Biology and Genetics, Aarhus University, 8830 Tjele, Denmark
[7]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA 24061, USA
[8]Personal Genomics Institute, Genome Research Foundation, Suwon 443-270, Republic of Korea
[9]KIOST, Korea Institute of Ocean Science and Technology, Ansan 426–744, Republic of Korea
[10]Greenland Institute of Natural Resources, 3900 Nuuk, Greenland
[11]Department of Environmental Science, Center for Reservoir and Aquatic Systems Research (CRASR) and Institute for Biomedical Studies, Baylor University, Waco, TX 76798, USA
[12]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[13]The Center for Genomic Advocacy (TCGA) and Department of Biology, Indiana State University, Terre Haute, IN 47809, USA
[14]Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN 47907, USA
[15]North Slope Borough, Department of Wildlife Management, Barrow, AK 99723, USA
[16]Battelle Memorial Institute, Houston, TX 77079, USA
[17]Department of Wildlife and Fisheries Sciences, Texas A&M University, College Station, TX 77843, USA
[18]Co-first author
[19]Present address: Department of Genetics, Stanford University, Stanford, CA 94305, USA
*Correspondence: jp@senescence.info
http://dx.doi.org/10.1016/j.celrep.2014.12.008

SUMMARY

The bowhead whale (*Balaena mysticetus*) is estimated to live over 200 years and is possibly the longest-living mammal. These animals should possess protective molecular adaptations relevant to age-related diseases, particularly cancer. Here, we report the sequencing and comparative analysis of the bowhead whale genome and two transcriptomes from different populations. Our analysis identifies genes under positive selection and bowhead-specific mutations in genes linked to cancer and aging. In addition, we identify gene gain and loss involving genes associated with DNA repair, cell-cycle regulation, cancer, and aging. Our results expand our understanding of the evolution of mammalian longevity and suggest possible players involved in adaptive genetic changes conferring cancer resistance. We also found potentially relevant changes in genes related to additional processes, including thermoregulation, sensory perception, dietary adaptations, and immune response. Our data are made available online (http://www.bowhead-whale.org) to facilitate research in this long-lived species.

INTRODUCTION

The lifespan of some animals, including quahogs, tortoises, and certain whale species, is far greater than that of humans (Austad, 2010; Finch, 1990). It is remarkable that a warm-blooded species such as the bowhead whale (*Balaena mysticetus*) has not only been estimated to live over 200 years (estimated age of one specimen 211 SE 35 years), suggesting it is the longest-lived mammal, but also exhibits very low disease incidence until an advanced age compared to humans (George et al., 1999; Philo et al., 1993). As in humans, the evolution of longevity in whales was accompanied by low fecundity and longer developmental time (Tacutu et al., 2013), as predicted by evolutionary theory. The cellular, molecular, and genetic mechanisms underlying longevity and resistance to age-related diseases in bowhead

Table 1. Statistics of the Bowhead Whale Genome Sequencing

Sequence Data Generated

Libraries	Total Data (Gb)	Sequence Coverage (for 2.91 Gb)
200 bp paired-end	149.1	51.2×
500 bp paired-end	141.7	48.7×
3 kb mate-paired	57.3	19.7×
5 kb mate-paired	72.5	24.9×
10 kb mate-paired	28.5	9.8×
Total	449.1	154.3×

Genome Assembly Statistics

Assembly	N50 (kb)	Number	Total Size (Gb)
Contigs	34.8	113,673	2.1
Scaffolds	877	7,227	2.3

See also Figures S1 and S2.

whales are unknown, but it is clear that, in order to live so long, these animals must possess preventative mechanisms against cancer, immunosenescence, and neurodegenerative, cardiovascular, and metabolic diseases. In the context of cancer, whales, and bowhead whales, in particular, must possess effective antitumor mechanisms. Indeed, given their large size (in extreme cases adult bowhead whales can weigh up to 100 tons and are therefore among the largest whales) and exceptional longevity, bowhead whale cells must have a significantly lower probability of neoplastic transformation relative to humans (Caulin and Maley, 2011; de Magãlhaes, 2013). Therefore, studying species such as bowhead whales that have greater natural longevity and resistance to age-related diseases than humans may lead to insights on the fundamental mechanisms of aging. Here, we report the sequencing and analysis of the genome of the bowhead whale, a species of the right whale family *Balaenidae* that lives in Arctic and sub-Arctic waters. This work provides clues regarding mechanisms underlying mammalian longevity and will be a valuable resource for researchers studying the evolution of longevity, disease resistance, and basic bowhead whale biology.

RESULTS

Sequencing and Annotation of the Bowhead Whale Genome

We sequenced the nuclear genome of a female bowhead whale (*Balaena mysticetus*) using the Illumina HiSeq platform at ~150× coverage. We followed established standards in the field in terms of sequencing paired-end libraries at high coverage plus mate-paired libraries of varying (3, 5, and 10 kb) insert sizes (Table 1). Contigs and scaffolds were assembled with ALLPATHS-LG (Gnerre et al., 2011). In line with other genomes sequenced with second-generation sequencing platforms, the contig N50 was 34.8 kb and scaffold N50 was 877 kb (Table 1); the longest scaffold in our assembly was 5,861 kb. In total, our assembly is ~2.3 Gb long. Genome size was estimated experimentally to be 2.91 Gb in another female and 2.87 Gb averaged with one male (see Supplemental Results and Figure S1), but this discrepancy likely reflects highly repetitive regions, as observed for the genomes of other species with similar reported sizes such as the minke whale (Yim et al., 2014).

The full and partial completeness of the bowhead whale draft genome assembly was evaluated as 93.15% and 97.18%, respectively, by the CEGMA pipeline (Parra et al., 2007), which is comparable to the minke whale genome assembly (Yim et al., 2014). We also generated RNA sequencing (RNA-seq) data from seven adult bowhead whale tissues (cerebellum, kidney, muscle, heart, retina, liver, and testis) from specimens from Greenland and Alaska, resulting in two transcriptome assemblies (see Experimental Procedures) and annotated the genome using MAKER2, which combines ab initio methods, homology-based methods, and transcriptome data to derive gene models (Holt and Yandell, 2011). Our annotation contains 22,672 predicted protein-coding genes with an average length of 417 (median 307) amino acid residues. In addition, based on transcriptome data from two Alaskan individuals (Table S1), we estimated 0.5–0.6 SNPs per kilobase of RNA (Table S2). To begin annotation of the bowhead genome, we identified orthologs based on similarity with cow, human, and mouse genes/proteins (see Experimental Procedures), which allowed us to assign predicted gene symbols to 15,831 bowhead genes.

Moreover, to annotate microRNAs in the bowhead genome, we sequenced small RNA libraries prepared from kidney and skeletal muscle. The miRDeep algorithm (Friedländer et al., 2008, 2012) was used to integrate the sequencing data into a model of microRNA biogenesis by Dicer processing of predicted precursor hairpin structures in the genome, thus identifying 546 candidate microRNA genes. Of the 546 candidate miRNAs identified in the bowhead, 395 had seed sequences previously identified in miRNAs from human, cow, or mouse, whereas 151 did not. All of our data are available online from our Bowhead Whale Genome Resource portal (http://www.bowhead-whale.org).

Analysis of the Draft Bowhead Whale Genome

Repeat sequences make up 41% of the bowhead genome assembly, most of which (78%) belong to the group of transposable elements (TEs). Although long interspersed nuclear elements (LINEs), such as L1, and short interspersed nuclear elements (SINEs) are widespread TEs in most mammalian lineages, the bowhead genome, similar to other cetacean genomes—minke, orca, and common bottlenose dolphin—is virtually devoid of SINEs (Supplemental Folder 1). LINE-1 (L1) is the most abundant TE, particularly in orca (90%) and minke whale (89%) (Figure S2). In comparison, TE diversity (measured with Shannon's index) in the bowhead genome (0.947) is higher than in orca (0.469) and minke whale (0.515) but lower than in dolphin (1.389) and cow (Bovine Genome Sequencing and Analysis Consortium et al., 2009) (1.534).

As a first assessment of coding genes that could be responsible for bowhead whale adaptations, we used bowhead coding sequences to calculate pairwise dN/dS ratios for 9,682, 12,685, and 11,158 orthologous coding sequences from minke whale (*Balaenoptera acutorostrata*), cow (*Bos taurus*), and dolphin (*Tursiops truncatus*), respectively. It is interesting to note that there are high levels of sequence conservation in the protein coding regions between bowhead and these species: 96% (minke),

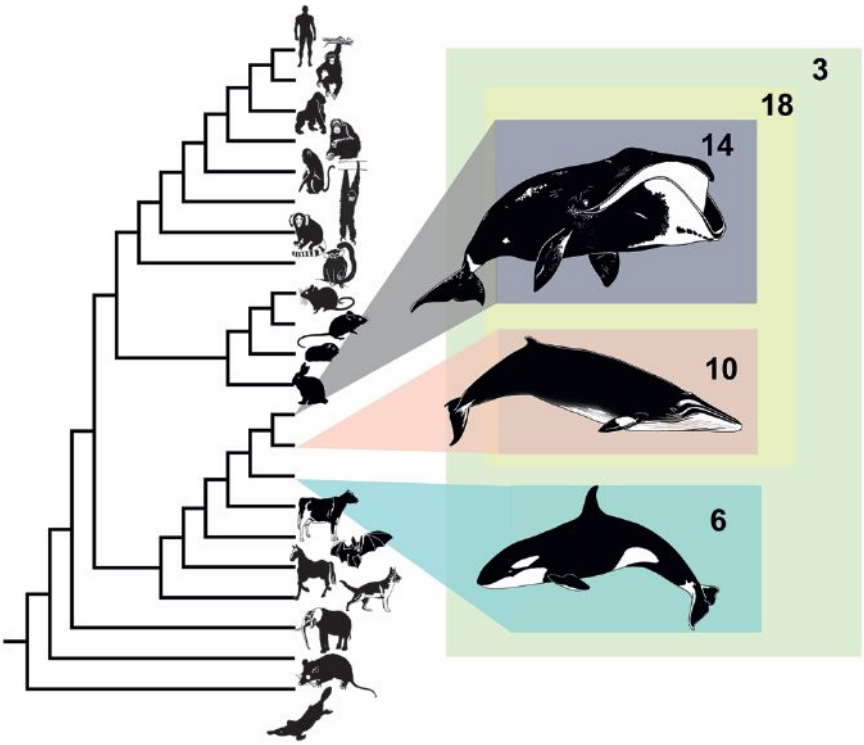


Figure 1. Phylogeny of Mammals Used in Codon-Based Maximum Likelihood Comparison of Selective Pressure Variation
The number of candidate genes under positive selection on each lineage is indicated.

92% (dolphin), and 91% (cow). This is not surprising, however, given the long generation time of cetaceans and of the bowhead whale, in particular, with animals only reaching sexual maturity at >20 years (Tacutu et al., 2013).

Because the minke whale is the closest relative to the bowhead (divergence time 25–30 million years ago [Gatesy et al., 2013]) with a sequenced genome and is smaller (<10 tons) and probably much shorter lived (maximum lifespan ~50 years) (Tacutu et al., 2013), comparisons between the bowhead and minke whale genomes may provide insights on the evolution of bowhead traits and of longevity, in particular. A number of aging- and cancer-associated genes were observed among the 420 predicted bowhead-minke orthologs with dN/dS exceeding 1, including *suppressor of cytokine signaling 2* (*SOCS2*), *aprataxin* (*APTX*), *noggin* (*NOG*), and *leptin* (*LEP*). In addition, the top 5% genes with high dN/dS values for bowhead-minke relative to the values for minke-cow and minke-dolphin orthologs included *forkhead box O3* (*FOXO3*), *excision repair cross-complementing rodent repair deficiency, complementation group 3* (*ERCC3*), and *fibroblast growth factor receptor 1* (*FGFR1*). The data on dN/dS ratios are also available on our portal to allow researchers to do their own analysis and quickly retrieve gene(s) of interest.

In a complementary and more detailed analysis of selective pressure variation, we used codon-based models of evolution (Yang, 2007) to identify candidate genes with evidence of lineage-specific positive selection (see Experimental Procedures). Using bowhead, minke, and orca protein-coding data along with a variety of available high-quality completed genomes from Laurasiatheria, Euarchontoglires, marsupial, and monotreme species, we identified a total of 866 single-gene ortholog families (SGOs) (i.e., these gene families have no more than one copy in each species). We tested each of the extant whale lineages, the ancestral baleen whale, and the most recent common ancestor (MRCA) of bowhead, minke, and orca, a total of five lineages (Figure 1), for evidence of lineage-specific positive selection.

Of the two extant whales analyzed, the number of SGOs exhibiting signatures of lineage-specific positive selection were as follows: bowhead (15 gene families) and minke (ten gene families). The small number of candidates under positive selection likely reflects the high level of protein conservation between bowhead and other cetaceans as well as the stringent filtering of candidates due to data-quality concerns; all results and alignments are provided in Supplemental Folder 1. A few genes associated with disease were identified, including *BMP and activin membrane-bound inhibitor* (*BAMBI*), which has been associated with various pathologies, including cancer, and also poorly studied genes of potential interest like *GRB2-binding adaptor protein, transmembrane* (*GAPT*).

In addition to the codon-based models of evolution, we wished to identify bowhead whale specific amino acid replacement substitutions. To this end, we aligned orthologous sequences between the bowhead whale and nine other mammals—a total of 4,358 alignments (see Experimental Procedures). Lineage-specific residues identified in this way have previously been shown to be indicative of significant changes in protein function (Tian et al., 2013). Our analysis revealed several proteins associated with aging and cancer among the top 5% of unique bowhead residues by concentration (i.e., normalized by protein length), including ERCC1 (excision repair cross-complementing rodent repair deficiency, complementation group 1), HDAC1 (histone deacetylase 1), and HDAC2 (Figure 2A). ERCC1 is a member of the nucleotide excision repair pathway (Gillet and Schärer, 2006), and disruption results in greatly reduced lifespan in mice and accelerated aging (Weeda et al., 1997). Histone deacetylases play an important role in the regulation of chromatin structure and transcription (Lee et al., 1993) and have been associated with longevity in *Drosophila* (Rogina et al., 2002). As such, these represent candidates involved in adaptive genetic changes conferring disease resistance in the bowhead whale. The full results are available in Supplemental Folder 1.

In addition to genes related to longevity, several interesting candidate genes emerged from our analysis of lineage-specific residues of potential relevance to other bowhead traits. Of

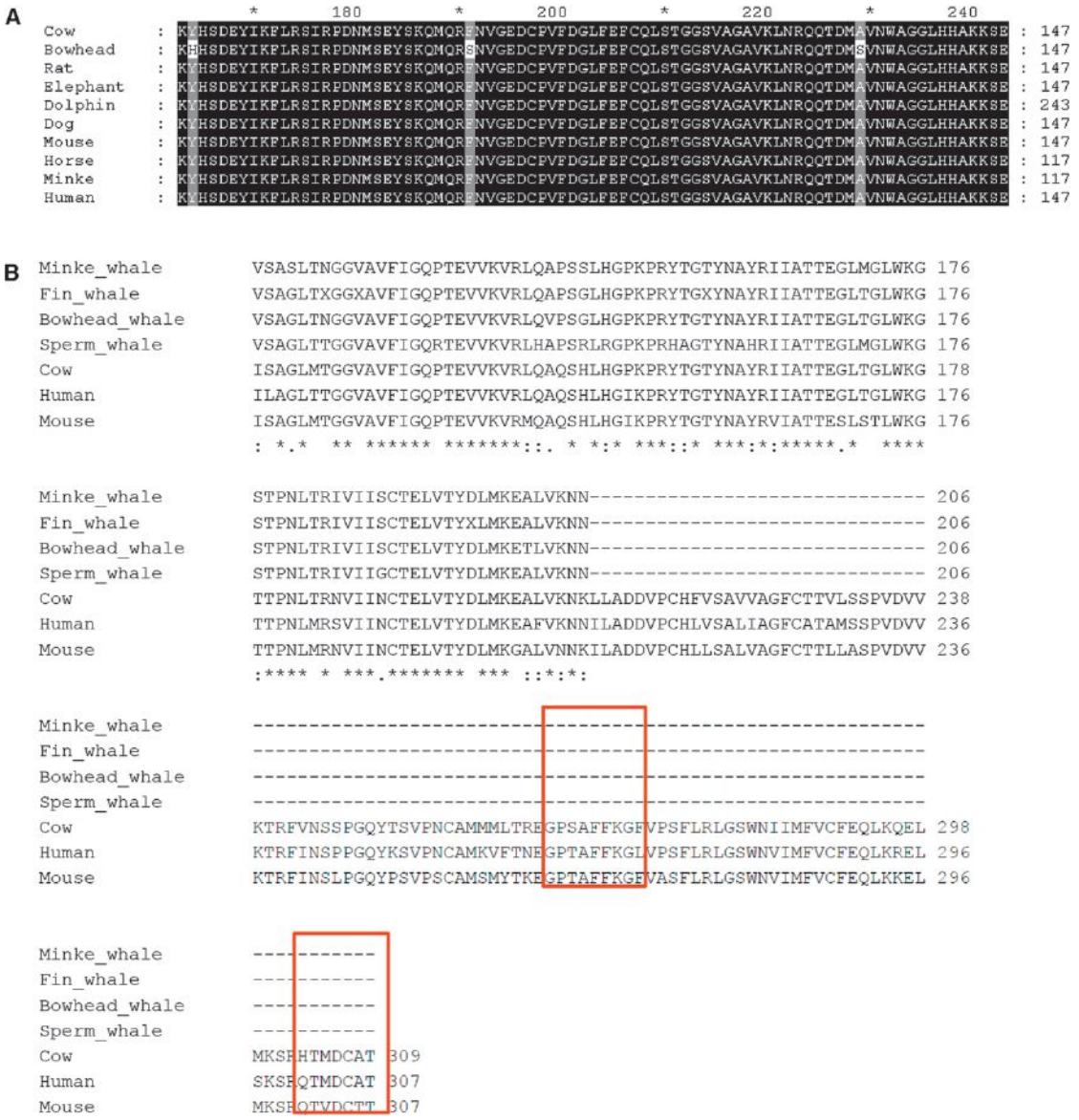


Figure 2. Multiple Protein Sequence Alignments of HDAC2 and UCP1

(A) Partial alignment of bowhead HDAC2 with mammalian orthologs. Unique bowhead residues are highlighted at human positions 68, 95, and 133.

(B) Partial alignment of whale UCP1 with mammalian orthologs. Conserved regions involved in UCP1 are marked in red.

note, a number of proteins related to sensory perception of sound were also identified with bowhead-specific mutations, including otoraplin (OTOR) and cholinergic receptor, nicotinic, alpha 10 (CHRNA10), which could be relevant in the context of the bowhead's ability to produce high- and low-frequency tones simultaneously (Tervo et al., 2011). In addition, many proteins must play roles in the large differences in size and development between the bowhead and related species and our results reveal possible candidates for further functional studies; for example, in the top ten proteins, SNX3 (sorting nexin 3) has been associated in one patient with eye formation defects and microcephaly (Vervoort et al., 2002), and WDR5 (WD repeat-containing protein 5) has been associated with osteoblast differentiation and bone development (Gori et al., 2006).

In the naked mole rat, a poikilotherm with a low metabolic rate and body temperature when compared to other mammals, unique changes in uncoupling protein 1 (UCP1), which is used to generate heat, have been previously found (Kim et al., 2011). Because the specific metabolic power output of cells in vivo for large whales must be much less than for smaller mammals (West et al., 2002), it is interesting to note that UCP1 of whales has a premature stop codon in C-terminal region, which is functionally important and conserved in other mammals (Figure 2B). It is tempting to speculate that these changes are related

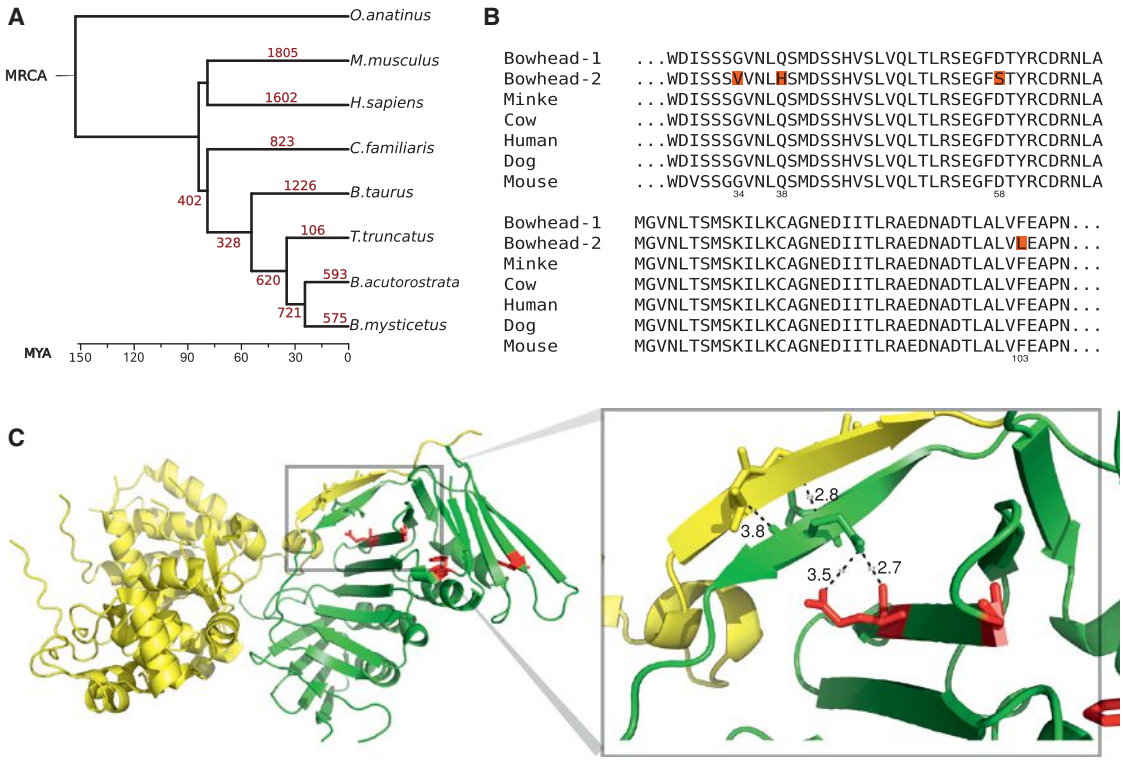


Figure 3. Gene Family Expansion and PCNA

(A) Gene family expansion. Numbers in red correspond to the predicted number of gene expansion events during mammalian evolution. Mean divergence time estimates were used from TimeTree (Hedges et al., 2006) for scaling.

(B) Multiple sequence alignment of PCNA residues 28–107, showing bowhead whale-specific duplication (gene IDs: bmy 16007 and bmy 21945). Lineage-specic amino acids in the duplicated PCNA of bowhead whales are highlighted in red.

(C) Crystal structure of the PCNA (green) and FEN-1 (yellow) complex. Lineage-specific residues on the PCNA structure are colored in red. A zoom in on the structures reveals a putative interaction between two β sheets, one within PCNA and another within FEN-1. This interaction may be altered through a second interaction between the PCNA β sheet and a lineage-specic change from glutamine to histidine within PCNA. Distance measurements between pairs of atoms are marked in black. PDB accession number: 1UL1.

See also Table S3 and Figure S3.

to differences in thermoregulation between whales and smaller mammals.

Potential Gene Duplications and Gene Losses

Gene duplication is a major mechanism through which phenotypic innovations can evolve (Holland et al., 1994; Kaessmann, 2010). Examples of mammalian phenotypic innovations associated to gene duplication include duplication of *RNASE1*, a pancreatic ribonuclease gene, in leaf-eating monkeys that contributed to adaptative changes in diet and digestive physiology (Zhang et al., 2002), a duplication of *GLUD1* in hominoids that subsequently acquired brain-specific functions (Burki and Kaessmann, 2004), and domestication of two syncytin gene copies that contributed to the emergence of placental development in mammals (Dupressoir et al., 2009). We surveyed the bowhead whale genome for expanded gene families that may reflect lineage-specific phenotypic adaptations and traits.

In the bowhead whale lineage, 575 gene families were predicted to have expanded (Figure 3). However, because gene expansion predictions are susceptible to false-positives owing to pseudogenes and annotation artifacts among other biases, we applied a stringent filter based on percentage of identity (Experimental Procedures) that reduced the number of candidate expansions to 41 (see Supplemental Folder 1 for the complete list). A functional enrichment analysis of these gene families, using default parameters in DAVID (Huang et al., 2009), only revealed a statistically significant enrichment (after correction for multiple hypothesis testing; Bonferroni <0.001) for genes associated with translation/ribosome. Given the association between translation and aging, for instance, in the context of loss of proteostasis (López-Otín et al., 2013), it is possible that these results reflect relevant adaptations in the bowhead whale.

Upon manual inspection of the gene expansion results, we found several duplicates of note. For instance, *proliferating cell nuclear antigen* (*PCNA*) is duplicated in bowhead whales with one copy harboring four lineage-specific residue changes (Figure 3B). Based on our RNA-seq data mapped to the genome (see Experimental Procedures and full results in Supplemental Folder 1), both *PCNA* copies are expressed in bowhead whale muscle, kidney, retina, and testis. By mapping the lineage-specific residues onto the structure of PCNA in complex with

FEN-1, we uncovered one amino acid substitution (Q38H), which may affect the interaction between PCNA and FEN-1 (Figure 3C). A subsequent branch-site test for selective pressure variation (see Experimental Procedures and Table S3) revealed that one substitution, D58S, may have undergone positive selection in the bowhead-whale lineage (with a posterior probability score of 0.983). The duplication of *PCNA* during bowhead-whale evolution is of particular interest due to its involvement in DNA damage repair (Hoege et al., 2002) and association with aging in that its levels in aged rat liver seem to relate to the decrease in the rate of cell proliferation (Tanno et al., 1996).

Another notable duplicated gene is *late endosomal/lysosomal adaptor, MAPK and MTOR activator 1 (LAMTOR1)*, in which six bowhead-specific amino acid changes were identified (Figure S3). LAMTOR1 is involved in amino acid sensing and activation of mTORC1, a gene strongly associated with aging and cancer (Cornu et al., 2013). The original *LAMTOR1* copy was expressed in all bowhead whale adult tissues for which we have data, with the duplicate having much lower (but detectable) expression in heart and retina. Also of note, putative duplications of *26S proteasome non-ATPase regulatory subunit 4 (PSMD4)* and *ubiquitin carboxyl-terminal esterase L3* (*UCHL3*) were identified with evidence of expression, which is intriguing considering the known involvement of the proteasome-ubiquitin system in aging (López-Otín et al., 2013) and given previous evidence that this system is under selection specific to lineages where longevity increased (Li and de Magalhães, 2013); *UCHL3* has also been involved in neurodegeneration (Kurihara et al., 2001). Other gene duplications of potential interest for their role in mitosis, cancer, and stress response include *cAMP-regulated phosphoprotein 19* (*ARPP19*), which has three copies even though we only detected expression of two copies, *stomatin-like 2* (*STOML2*), *heat shock factor binding protein 1* (*HSBP1*) with four copies of which two appear to be expressed, *spermine synthase (SMS)* and *suppression of tumorigenicity 13* (*ST13*).

Similar to previous genome characterizations, we chose the complete set of known protease genes for a detailed supervised analysis of gene loss (Quesada et al., 2009). This procedure highlighted multiple gene loss events potentially related to the evolution of several cetacean traits, including adaptations affecting the immune system, blood homeostasis, digestive system, and dentition (Figure S4). Thus, the cysteine protease CASP12, a modulator of the activity of inflammatory caspases, has at least one conserved premature stop codon in bowhead and minke whales. Interestingly, whereas this protease is conserved and functional in almost all of the terrestrial mammals, most human populations display different deleterious variants (Fischer et al., 2002), presumably with the same functional consequences as the premature stop codons in whales. Likewise, two paralogues of carboxypeptidase A (CPA2 and CPA3) have been pseudogenized in bowhead and minke whales. Notably, CPA variants have been associated with increased risk for prostate cancer in humans (Ross et al., 2009), which could be of interest in the context of reduced cancer susceptibility in whales compared with humans (de Magalhães, 2013).

Additionally, we found that multiple coagulation factors have been lost in bowhead and minke whales. The finding of bowhead whale-specific changes is also noteworthy because it could be related to the special characteristics of this mammal. For example, OTUD6A, a cysteine protease with a putative role in the innate immune system (Kayagaki et al., 2007), is specifically lacking in the assembled genome and expressed sequences of the bowhead whale. In addition, whereas the enamel metalloprotease MMP20 has been lost in bowhead and minke whales (Yim et al., 2014), our analysis suggests that these genomic events happened independently (see alignments in Supplemental Folder 1). Finally, as aforementioned, the cysteine protease UCHL3 seems to have been duplicated through a retrotranscription-mediated event in a common ancestor to bowhead and minke whales, although only the genome of the bowhead whale shows a complete, putatively functional open reading frame for this extra copy of the gene. UCHL3 may play a role in adipogenesis (van Beekum et al., 2012), which indicates that this duplication might be related to the adaptation of the bowhead whale to the challenging arctic environment. These results suggest specific scenarios for the role of proteolysis in the evolution of *Mysticetes*. Specifically, given the relationship between immunity and aging (López-Otín et al., 2013), some of these findings might open new approaches for the study of this outstanding cetacean.

DISCUSSION

The genetic and molecular mechanisms by which longevity evolves remain largely unexplained. Given the declining costs of DNA sequencing, de novo genome sequencing is rapidly becoming affordable. The sequencing of genomes of long-lived species allows comparative genomics to be employed to study the evolution of longevity and has already provided candidate genes for further functional studies (de Magalhães and Keane, 2013). Nonetheless, deciphering the genetic basis of species differences in longevity has major intrinsic challenges (de Magalhães and Keane, 2013), and much work remains to uncover the underlying mechanisms by which some species live much longer than others. In this context, studying a species so long lived and with such an extraordinary resistance to age-related diseases as the bowhead whale will help elucidate mechanisms and genes conferring longevity and disease resistance in mammals. Remarkably, large whales with over 1,000 times more cells than humans do not exhibit an increased cancer risk (Caulin and Maley, 2011), suggesting the existence of natural mechanisms that can suppress cancer more effectively in these animals. Having the genome sequence of the bowhead whale will allow researchers to study basic molecular processes and identify maintenance mechanisms that help preserve life, avoid entropy, and repair molecular damage. When compared to transcriptome data (Seim et al., 2014), the genome's greater completeness and quality permits additional (e.g., gene loss and duplication) and more thorough analyses. Besides, whereas the genomes of many commercially important agricultural species have been reported, the bowhead genome sequence is the first for a species key to a subsistence diet of indigenous communities. One of the outputs of this project will be to facilitate and drive research in this long-lived species. Data and results from this project are thus made freely available to the scientific community on an online portal (http://www.bowhead-whale.org/). We provide

this key resource for studying the bowhead whale and its various traits, including its exceptional longevity and resistance to diseases.

EXPERIMENTAL PROCEDURES

DNA and RNA Sampling in Greenland

Bowhead (*Balaena mysticetus*) DNA used for genome sequencing was isolated from muscle tissue sampled from a 51-year-old female (ID no. 325) caught in the Disko Bay, West Greenland in 2009 (Heide-Jørgensen et al., 2012). Tissue samples were stored at –20°C immediately after collection. Age estimation was performed using the aspartic acid racemization technique (Garde et al., 2007). CITES no. 12GL1003387 was used for transfer of biological material. Bowhead RNA used for RNA-seq and small RNA analysis was isolated from two different individuals: kidney samples were from a 44-year-old female (ID no. 500) and muscle samples were isolated from a 44-year-old male (ID no. 322). For more details of the individual whales, see Heide-Jørgensen et al. (2012).

Genome Sequencing

DNA was extracted following standard protocols, quantified using Qubit and run on an agarose gel to ensure no degradation had occurred. We then generated ~150× coverage of the genome using the Illumina HiSeq 2000 platform with 100 bp reads, sequencing paired-end libraries, and mate-paired libraries with insert sizes of 3, 5, and 10 kb (Table 1). Sequencing was performed at the Liverpool Centre for Genomic Research (CGR; http://www.liv.ac.uk/genomic-research/).

Genome Assembly

Libraries were preprocessed in-house by the CGR to remove adaptor sequences. The raw fastq files were trimmed for the presence of the Illumina adaptor sequence using Cutadapt and then subjected to window-based quality trimming using Sickle with a minimum window quality score of 20. A minimum read-length filter of 10 bp was also applied. Libraries were then assembled with ALLPATHS-LG (Gnerre et al., 2011), which performed all assembly steps including read error correction, initial read alignment, and scaffolding. ALLPATHS-LG build 43762 was used with the default input parameters, including K = 96. Several build parameters were automatically determined by the software at run time per its standard algorithm. Of 2.88×10^9 paired fragment reads and 1.87×10^9 paired jumping reads, 0.015% were removed as poly(A) and 1.5% were removed due to low-frequency kmers; 54% of jumping read pairs were error-corrected, and overall 33% of jumping pairs were redundant. In total, we used 216 Gbp for the 2.3 Gb assembly, meaning that coverage retained for the assembly was ~95×. Full assembly and read usage data are shown in Supplemental Folder 2. Assembly completeness was assayed with CEGMA by searching for 248 core eukaryotic genes (Parra et al., 2007).

Genome Size Determination

To determine the genome size for bowhead whale, spleen tissues were acquired from one male (10B17) and one female (10B18). Both whales were harvested in 2010 as part of the native subsistence hunt in Barrow, Alaska. Sample processing and staining followed the methods of Vindeløv and Christensen (1994). Instrument description and additional methodological details are provided in Oziolor et al. (2014). Briefly, flow cytometric genome size determination is based on propidium iodide fluorescent staining of nuclear DNA. Mean fluorescence is calculated for cells in the G0 and G1 phases of the cell cycle. This method requires direct comparison to known standards to convert measured fluorescence to pg of DNA. The primary standard used in this study was the domestic chicken (*Gallus gallus domesticus*). Chicken red blood cells are widely used as a genome size standard, with an accepted genome size of C = 1.25 pg. Chicken whole blood was purchased from Innovative Research. Mouse (*Mus musculus*) and rat (*Rattus norvegicus*) were included as internal checks, with estimates for both falling within 3% of previously published genome size estimates (Vinogradov, 1998). Spleen tissues from three male 129/SvEvTac laboratory mice and a single male Harlan SD Sprague-Dawley laboratory rat were used.

Transcriptome Sequencing and Assembly: Greenland Samples

Total RNA was extracted from the kidney and muscle employing the mirVanaTM RNA extraction kit (Ambion). RNA integrity of the individual RNA samples was assessed on a 1% agarose gel using an Agilent 2100 Bioanalyzer (Agilent Technologies). Library preparation was performed using the ScriptSeqTM mRNA-seq library preparation kit from Epicenter according to the manufacturer's protocol (Epicenter) and sequenced (100 bp paired end) as multiplexed samples using the Illumina HiSeq 2000 analyzer. Fastq generation and demultiplexing were performed using the CASAVA 1.8.2 package (Illumina). The fastq files were filtered for adapters, quality, and length using Trimmomatic (v.0.27), with a window size of 4, a base quality cutoff of 20, and a minimum length of 60 (Lohse et al., 2012). De novo transcriptome assembly was performed using the short read assembler software Trinity (release 2013-02-25), which is based on the de Bruijn graph method for assembly, with default settings (Grabherr et al., 2011).

Transcriptome Sequencing and Assembly: Alaskan Samples

Tissue biopsies were obtained from two male bowhead whales harvested by Inupiat hunters at Barrow, Alaska during the Fall hunt of 2010; heart, cerebellum, liver, and testes were biopsied from male bowhead number 10B16, and retina from male bowhead 10B20. Samples were immediately placed in liquid nitrogen and transported in a dry shipper to Purdue University. RNA was extracted using TRIZOL reagent (Invitrogen) following the manufacturer's protocol. RNA was purified using an Invitrogen PureLink Micro-to-Midi columns from the Total RNA Purification System using the standard protocol. RNA quantity and quality was estimated with a spectrophotometer (Nanodrop) and by gel electrophoresis using an Agilent model 2100 Bioanalyzer. cDNA libraries were constructed by random priming of chemically sheared poly A captured RNA. Randomly primed DNA products were blunt ended. Products from 450–650 bp were then isolated using a PippenPrep. After the addition of an adenine to the fragments, a Y primer amplification was used to produce properly tailed products. Paired-end sequences of 100 bp per end were generated using the Illumina HiScan platform. Sequences with primer concatamers, weak signal, and/or poly A/T tails were culled. The Trinity software package for de novo assembly (Grabherr et al., 2011) was used for transcript reconstruction (Table S1).

Small RNA Sequencing and Annotation

To annotate microRNA genes in the bowhead genome, we conducted deep sequencing of two small RNA libraries prepared from muscle and kidney tissues (Greenland samples). Total RNA was isolated using mirVana miRNA Isolation Kit (Ambion). Small RNA in the 15-40 nucleotides range was gel purified and small RNA libraries were prepared for next-generation sequencing using the ScriptMiner Small RNA-Seq Library Preparation Kit (Epicenter). The two libraries were sequenced on an Illumina Hi-Seq 2000 instrument to generate single end sequences of 50 nucleotides. Primary data analysis was done using the Illumina CASAVA Pipeline software v.1.8.2, and the sequence reads were further processed by trimming for adapters and filtering for low quality using Trimmomatic (Lohse et al., 2012). Identification of conserved and novel candidate microRNA genes in the bowhead genome was accomplished by applying the miRDeep2 algorithm (Friedländer et al., 2008, 2012).

Evaluation of Repeat Elements

To evaluate the percentage of repeat elements, RepeatMasker (v.4.0.3; http://www.repeatmasker.org/) was used to identify repeat elements, with parameters set as "-s -species mammal." RMBlast was used as a sequence search engine to list out all types of repeats. Percentage of repeat elements was calculated as the total number of repeat region divided by the total length of the genome, excluding the N-region. Genomes of minke whale (*Balaenoptera acutorostrata*), orca (*Orcinus orca*), common bottlenose dolphin (*Tursiops truncates*), and cow (*Bos taurus*) were downloaded from NCBI and run in parallel for comparison with the bowhead genome.

Genome Annotation

Putative genes were located in the assembly by structural annotation with MAKER2 (Holt and Yandell, 2011), which combined both bowhead

transcriptomes with comparative and de novo prediction methods including BLASTX, Exonerate, SNAP, Genemark, and Augustus. In addition to the RNA-seq data, the entire SwissProt database and the draft proteome of dolphin were used as input to the comparative methods. Repetitive elements were found with RepeatMasker (http://www.repeatmasker.org/). The complete set of MAKER input parameters, including training sets used for the de novo prediction methods, are listed in Supplemental Folder 2. In total, 22,672 protein-coding genes were predicted with an average length of 417 (median 307) amino acid residues.

The RNA-seq data from seven adult bowhead tissues described above were then mapped to the genome: FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used for quality control to make sure that data of all seven samples was of acceptable quality. STAR (Dobin et al., 2013) was used to generate genome files from the bowhead assembly and to map the reads to the bowhead genome with 70.3% of reads mapping, which is in line with other results including those in the minke whale (Yim et al., 2014). To count the reads overlapping genes, we used ReadCounter (van Dam et al., 2015). The results obtained from all seven samples were combined into a single file describing the number of nonambiguously mapping reads for each gene (full results in Supplemental Folder 1). Of the 22,672 predicted protein-coding genes, 89.5% had at least ten reads mapping and 97.5% of predicted genes had at least one read mapping to them, which is again comparable to other genomes like the minke whale genome (Yim et al., 2014).

To allow the identification of orthologous relationships with bowhead proteins, all cow protein sequences were downloaded from Ensembl (Flicek et al., 2013). Cow was initially used because it is the closest relative to the bowhead with a high-quality annotated genome available. First, BLASTP (10^{-5}) was used to find the best hit in the cow proteome for every predicted bowhead protein, and then the reciprocal best hit for each cow protein was defined as an ortholog. In addition, human and mouse orthologs from the OPTIC pipeline (see below) were used to assign predicted gene symbols to genes and proteins. A total of 15,831 bowhead genes have a putative gene symbol based on these predictions. Homologs in minke whale and dolphin were also derived and are available on our bowhead genome portal.

Genome Portal

To facilitate further studies of these animals, we constructed an online genome portal: The Bowhead Whale Genome Resource (http://www.bowhead-whale.org/). Its database structure, interface, and functionality were adapted from our existing Naked Mole Rat Genome Resource (Keane et al., 2014). Our data and results are available from the portal, and supplemental methods and data files are also available on GitHub (https://github.com/maglab/bowhead-whale-supplementary).

Pairwise dN/dS Analysis

The CodeML program from the PAML package was used to calculate pairwise dN/dS ratios (Yang, 2007). This is done using the ratio of nonsynonymous substitutions per nonsynonymous site (dN) to synonymous substitutions per synonymous site (dS), dN/dS, or ω (Yang, 2007). Specifically, these pairwise dN/dS ratios were calculated for bowhead coding sequences and orthologous sequences from minke, cow, and dolphin, excluding coding sequences that were less than 50% of the length of the orthologous sequence. The results were then ranked by decreasing dN/dS and are available on our bowhead genome portal. In addition, the ratio of the bowhead-minke dN/dS value to the higher of the dN/dS values for minke-cow and minke-dolphin was calculated to identify genes that evolved more rapidly on the bowhead lineage.

Assessment of Selective Pressure Variation across Single-Gene Orthologous Families Using Codon-Based Models of Evolution

To accurately assess variation in selective pressure on the bowhead, minke, and orca lineages in comparison to extant terrestrial mammals, we created a protein-coding database spanning the placental mammals. Along with the orca (http://www.ncbi.nlm.nih.gov/bioproject/189949), minke (Yim et al., 2014), and bowhead data described above, we extracted protein coding sequences from Ensembl Biomart v.73 (Flicek et al., 2013) for the following 18 genomes: chimpanzee, cow, dog, elephant, gibbon (5.6× coverage), gorilla, guinea pig, horse, human, macaque, marmoset, microbat, mouse, opossum, orangutan, platypus, rabbit, and rat. These genomes were all high coverage (mostly >6× coverage) with the exception of gibbon (Supplemental Folder 2). Sequence similarity searches were performed using mpi-BLAST (v 1.6.0) (Altschul et al., 1990) (http://www.mpiblast.org/) on all proteins using a threshold of 10^{-7}. Gene families were identified using in-house software that clusters genes based on reciprocal BLAST hits (Altschul et al., 1990). We identified a total of 6,630 gene families from which we extracted the single-gene orthologous families (SGOs). Families were considered SGOs if we identified a single-gene representative in each species (one-to-one orthologs), and to account for lower coverage genomes and missing data we also considered cases where a specific gene was not present in a species, i.e., one-to-zero orthology. SGOs were only considered for subsequent analysis if they contained more than seven species in total and if they contained no internal stop codons (indicative of sequencing errors). In total, we retained 866 SGOs for further analysis. Multiple sequence alignments (MSAs) were generated using default parameters in PRANK (v.100802) (Löytynoja and Goldman, 2008). To minimize potential false-positives due to poor sequence quality, the MSAs of the 866 SGOs underwent strict data-quality filtering. The first filter prohibited the presence of gaps in the MSA if created by unique insertions (>12 bp) in either bowhead or minke sequences. The second filter required unaligned bowhead or minke sequences to be at least half the length of their respective MSA. These two filters refined the number of testable SGOs to 319. The gene phylogeny of each SGO was inferred from the species phylogeny (Morgan et al., 2013). CodeML from the PAML software package (v.4.4e) (Yang, 2007) was employed for our selective pressure variation analyses. We analyzed each of the 319 refined SGOs using the nested codon-based models of evolution under a maximum likelihood framework. We employed the likelihood ratio test (LRT) using nested models of sequence evolution to evaluate a variety of models of codon sequence evolution (Yang, 2007). In general, these codon models allow for variable dN/dS ratios (referred to as ω throughout) among sites in the alignment, along different lineages on our phylogenetic tree, or a combination of both variations across lineages and sites. To assess the significance of fit of each model to the data, we used the recommended LRTs in CodeML (Yang, 2007) for comparing nested models (see Supplemental Folder 2). The LRT test statistic approximates the chi-square ($\chi 2$) distribution critical value with degrees of freedom equal to the number of additional free parameters in the alternative model. The goal of the codon-based modeling is to determine the selective pressures at work in a lineage and site-specific manner.

The models applied follow the standard nomenclature (i.e., model M1, M2, A, and A null) (Yang, 2007). Model M1 assumes that there are two classes of sites—those with an ω value of zero and those with an ω value of 1. Model M2 allows for three classes of sites—one with an ω value of zero, one with an ω value of one and one with an ω value that is not fixed to any value. Given the relationship between M1 and M2, they can be tested for the significance of the difference of the fit of these two models using an LRT with df = 2. Finally, we used model A that allows the ω value to vary across sites and across different lineages in combination. With model A, we can estimate the proportion of sites and the dN/dS ratio in the foreground lineage of interest in comparison to the background lineages and the estimated dN/dS ratio is free to vary above 1 (i.e., positive selection). Model A can be compared with its site-specific counterpart (model M1) using the LRT with df = 2. In addition, the lineage and site-specific model model A null was applied as a second LRT with model A. In model A null, the additional site category is fixed at neutral rather than being estimated from the data, and this LRT provides an additional test for model A (Zhang et al., 2005). In this way, we performed independent tests on each of the extant cetacean lineages (orca, minke, and bowhead), as well as testing each ancestral cetacean branch (the MRCA of the two baleen whales and the MRCA of all three cetaceans), to determine if there were signatures of positive selection that are unique to each lineage (Yang and dos Reis, 2011). Using empirical Bayesian estimations, we identified the specific residues that are positively selected in each lineage tested. Positive selection was inferred if all of the following criteria were met: (1) if the LRT was significant, (2) if the parameters estimated under that model were concurrent with positive selection, and (3) if the alignment in that region was of high quality (as judged by alignment

completeness and quality in that region). The posterior probability (PP) of a positively selected site is estimated using two calculations: Naive Empirical Bayes (NEB) or Bayes Empirical Bayes (BEB) (Yang, 2007). If both NEB and BEB are predicted, we reported the BEB results as they have been shown to be more robust under certain conditions (Yang et al., 2005). For all models used in the analysis where ω is estimated from the data, a variety of starting ω values was used for the calculation of likelihood estimates. This ensures that the global minimum is reached.

Identification of Proteins with Bowhead-Unique Residues

An in-house Perl pipeline was used to align each bowhead protein with orthologs from nine other mammals: human (*Homo sapiens*), dog (*Canis familiaris*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), minke whale (*Balaenoptera acutorostrata*), cow (*Bos taurus*), dolphin (*Tursiops truncatus*), horse (*Equus caballus*), and elephant (*Loxodonta africana*) and then identify the unique bowhead amino acid residues. Gaps were excluded from the analysis, and a maximum of one unknown residue was allowed in species other than the bowhead. The results were ranked by the number of unique residues normalized by the protein length (full results in Supplemental Folder 1).

Gene Expansion Analysis, Filtering, and Expression

Human, mouse, dog, cow, dolphin, and platypus genomes and gene annotations were obtained from Ensembl (Flicek et al., 2013), the genome and gene annotation of minke whale were obtained from Yim et al. (2014). In total, 21,069, 22,275, 19,292, 19,988, 15,769, 17,936, 20,496, and 22,733 human, mouse, dog, cow, dolphin, platypus, minke whale, and bowhead whale genes, respectively, were used to construct orthology mappings using OPTIC (Heger and Ponting, 2007). Briefly, OPTIC builds phylogenetic trees for gene families by first assigning orthology relationships based on pairwise orthologs computed using PhyOP (Goodstadt and Ponting, 2006). Then, a tree-based method, PhyOP, is used to cluster genes into orthologous groups, and, last, gene members are aligned and phylogenetic trees built with TreeBeST (Vilella et al., 2009). Further details are available in the OPTIC paper (Heger and Ponting, 2007). Predicted orthology groups can be accessed at http://genserv.anat.ox.ac.uk/clades/vertebrates_bowhead.

To identify gene families that underwent expansion, gene trees were reconciled with the consensus species tree, and duplicated nodes were identified. The tree used, derived from TimeTree (Hedges et al., 2006), was: (mm_oanatinus5, ((mm_cfamiliaris3, (mm_btaurus, (mm_ttruncatus, (mm_balaenoptera, mm_bmysticetus))))), (mm_hsapiens10, mm_mmusculus5))). The following algorithm was used to reconcile gene and species trees.

A stringent filter was applied to the data so that gene duplicates in bowhead whales were required to differ by at most 10% in protein sequence from a cognate copy but were also required to differ by at least 1% to avoid assembly artifacts and to remove recently duplicated copies with no function. Further manual inspection of the alignments was performed. Gene expression inferred from our RNA-seq data was used to check the expression of duplicates.

An in-house peptide-sensitive approach was used to align the PCNA cDNA into codons, and CodeML/PAML was used to test M0, a one-rate model that assumes the same rate of evolution in all branches against M2^a, a branch site test with one rate for the background and one rate for the bowhead whale branch (Yang, 2007).

ACCESSION NUMBERS

Our data and results can be downloaded from the Bowhead Whale Genome Resource (http://www.bowhead-whale.org/downloads/). In addition, data are available at the NCBI BioProject PRJNA194091 with raw sequencing reads in the Sequence Read Archive (SRP050351).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Results, four figures, three tables, and two supplemental data files and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.12.008.

AUTHOR CONTRIBUTIONS

G.M.C., J.C.G., R.S., J.W.B., B.T., and J.P.M. conceived and designed the study; L.B.M., E.M.O., and C.W.M. performed the experiments; M.K., J.S., A.E.W., Y.I.L., V.Q., L.B.M., S.v.D., D.B., P.I.M., P.M., L.K., J.B., H.-S.Y., G.W.S., J.C.P., C.L.-O., M.J.O., J.W.C., and J.P.M. analyzed the data; T.C., N.V.G., N.H.N., M.P.H.-J., R.S., and K.L. contributed reagents/materials/analysis tools; and M.K. and J.P.M. wrote the paper.

ACKNOWLEDGMENTS

This project was supported by grants from the Life Extension Foundation and the Methuselah Foundation to J.P.M. We thank the Inuit whaling captains of the Alaska Eskimo Whaling Commission and the Barrow Whaling Captains' Association for allowing us to sample their whales and their willingness to support the transcriptome study. This study was also partly funded by the Augustinus Foundation to K.L. T.C. was supported by a Wellcome Trust grant (WT094386MA) to J.P.M. and M.K. was supported by a studentship from the University of Liverpool's Faculty of Health and Life Sciences. J.S. was supported by grants from the NIH (GM094575 to N.V.G.) and the Welch Foundation (I-1505 to N.V.G.). Y.I.L. was supported by a Nuffield Department of Medicine Prize studentship from the University of Oxford. C.L.-O. is an Investigator of the Botin Foundation also supported by grants from Ministerio de Economía y Competitividad-Spain and Instituto de Salud Carlos III (RTICC)-Spain. M.J.O. and A.E.W. are funded by Science Foundation Ireland Research Frontiers Programme Grant (EOB2763) to M.J.O. M.J.O. would also like to acknowledge the Fulbright Commission for the Fulbright Scholar Award 2012-2013. M.J.O. and A.E.W. thank the SFI/HEA Irish Centre for High-End Computing (ICHEC) for processor time and technical support. This work was also supported by the Korea Institute of Ocean Science and Technology (KIOST) in-house program (PE99212). Further thanks to Prof. Chris Ponting and Dr. Andreas Heger for hosting gene orthology predictions from OPTIC, the University of Liverpool High Performance Computing facilities for processor time, Eric de Sousa for help with RNA-seq data analysis, and to Louise Crompton for assistance in compiling and formatting the bibliography. Last, we are grateful to the staff at the Liverpool Centre for Genomic Research for advice during this project.

Received: September 7, 2014
Revised: November 21, 2014
Accepted: December 3, 2014
Published: December 24, 2014

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Austad, S.N. (2010). Methusaleh's Zoo: how nature provides us with clues for extending human health span. J. Comp. Pathol. *142* (*Suppl 1*), S10–S21.

Burki, F., and Kaessmann, H. (2004). Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. Nat. Genet. *36*, 1061–1063.

Caulin, A.F., and Maley, C.C. (2011). Peto's Paradox: evolution's prescription for cancer prevention. Trends Ecol. Evol. *26*, 175–182.

Bovine Genome Sequencing and Analysis Consortium, Elsik, C.G., Tellam, R.L., Worley, K.C., Gibbs, R.A., Muzny, D.M., Weinstock, G.M., Adelson, D.L., Eichler, E.E., Elnitski, L., et al. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science *324*, 522–528.

Cornu, M., Albert, V., and Hall, M.N. (2013). mTOR in aging, metabolism, and cancer. Curr. Opin. Genet. Dev. *23*, 53–62.

de Magalhães, J.P. (2013). How ageing processes influence cancer. Nat. Rev. Cancer *13*, 357–365.

de Magalhães, J.P., and Keane, M. (2013). Endless paces of degeneration—applying comparative genomics to study evolution's moulding of longevity. EMBO Rep. *14*, 661–662.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dupressoir, A., Vernochet, C., Bawa, O., Harper, F., Pierron, G., Opolon, P., and Heidmann, T. (2009). Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. Proc. Natl. Acad. Sci. USA *106*, 12127–12132.

Finch, C. (1990). Longevity, Senescence, and the Genome (Chicago: University of Chicago Press).

Fischer, H., Koenig, U., Eckhart, L., and Tschachler, E. (2002). Human caspase 12 has acquired deleterious mutations. Biochem. Biophys. Res. Commun. *293*, 722–726.

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., et al. (2013). Ensembl 2013. Nucleic Acids Res. *41*, D48–D55.

Friedländer, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. Nat. Biotechnol. *26*, 407–415.

Friedländer, M.R., Mackowiak, S.D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. *40*, 37–52.

Garde, E., Heide-Jorgensen, M.P., Hansen, S.H., Nachman, G., and Forchhammer, M.C. (2007). Age-specific growth and remarkable longevity in narwhals (Monodon monoceros) from West Greenland as estimated by aspartic acid racemization. J. Mammal. *88*, 49–58.

Gatesy, J., Geisler, J.H., Chang, J., Buell, C., Berta, A., Meredith, R.W., Springer, M.S., and McGowen, M.R. (2013). A phylogenetic blueprint for a modern whale. Mol. Phylogenet. Evol. *66*, 479–506.

George, J.C., Bada, J., Zeh, J., Scott, L., Brown, S.E., O'Hara, T., and Suydam, R. (1999). Age and growth estimates of bowhead whales (Balaena mysticetus) via aspartic acid racemization. Can. J. Zool. *77*, 571–580.

Gillet, L.C.J., and Schärer, O.D. (2006). Molecular mechanisms of mammalian global genome nucleotide excision repair. Chem. Rev. *106*, 253–276.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA *108*, 1513–1518.

Goodstadt, L., and Ponting, C.P. (2006). Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Comput. Biol. *2*, e133.

Gori, F., Friedman, L.G., and Demay, M.B. (2006). Wdr5, a WD-40 protein, regulates osteoblast differentiation during embryonic bone development. Dev. Biol. *295*, 498–506.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Hedges, S.B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics *22*, 2971–2972.

Heger, A., and Ponting, C.P. (2007). Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. Genome Res. *17*, 1837–1849.

Heide-Jørgensen, M.P., Garde, E., Nielsen, N.H., Andersen, O.N., and Hansen, S.H. (2012). A note on biological data from the hunt of bowhead whales in West Greenland 2009-2011. J. Cetacean Res. Manag. *12*, 329–333.

Hoege, C., Pfander, B., Moldovan, G.L., Pyrowolakis, G., and Jentsch, S. (2002). RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO. Nature *419*, 135–141.

Holland, P.W., Garcia-Fernàndez, J., Williams, N.A., and Sidow, A. (1994). Gene duplications and the origins of vertebrate development. Dev. Suppl., 125–133.

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics *12*, 491.

Huang, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Res. *20*, 1313–1326.

Kayagaki, N., Phung, Q., Chan, S., Chaudhari, R., Quan, C., O'Rourke, K.M., Eby, M., Pietras, E., Cheng, G., Bazan, J.F., et al. (2007). DUBA: a deubiquitinase that regulates type I interferon production. Science *318*, 1628–1632.

Keane, M., Craig, T., Alföldi, J., Berlin, A.M., Johnson, J., Seluanov, A., Gorbunova, V., Di Palma, F., Lindblad-Toh, K., Church, G.M., and de Magalhães, J.P. (2014). The Naked Mole Rat Genome Resource: facilitating analyses of cancer and longevity-related adaptations. Bioinformatics *30*, 3558–3560.

Kim, E.B., Fang, X., Fushan, A.A., Huang, Z., Lobanov, A.V., Han, L., Marino, S.M., Sun, X., Turanov, A.A., Yang, P., et al. (2011). Genome sequencing reveals insights into physiology and longevity of the naked mole rat. Nature *479*, 223–227.

Kurihara, L.J., Kikuchi, T., Wada, K., and Tilghman, S.M. (2001). Loss of Uch-L1 and Uch-L3 leads to neurodegeneration, posterior paralysis and dysphagia. Hum. Mol. Genet. *10*, 1963–1970.

Lee, D.Y., Hayes, J.J., Pruss, D., and Wolffe, A.P. (1993). A positive role for histone acetylation in transcription factor access to nucleosomal DNA. Cell *72*, 73–84.

Li, Y., and de Magalhães, J.P. (2013). Accelerated protein evolution analysis reveals genes and pathways associated with the evolution of mammalian longevity. Age (Dordr.) *35*, 301–314.

Lohse, M., Bolger, A.M., Nagel, A., Fernie, A.R., Lunn, J.E., Stitt, M., and Usadel, B. (2012). RobiNA: a user-friendly, integrated software solution for RNA-seq-based transcriptomics. Nucleic Acids Res. *40*, W622–W627.

López-Otín, C., Blasco, M.A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. Cell *153*, 1194–1217.

Löytynoja, A., and Goldman, N. (2008). A model of evolution and structure for multiple sequence alignment. Philos. Trans. R. Soc. Lond. B Biol. Sci. *363*, 3913–3919.

Morgan, C.C., Foster, P.G., Webb, A.E., Pisani, D., McInerney, J.O., and O'Connell, M.J. (2013). Heterogeneous models place the root of the placental mammal phylogeny. Mol. Biol. Evol. *30*, 2145–2156.

Oziolor, E.M., Bigorgne, E., Aguilar, L., Usenko, S., and Matson, C.W. (2014). Evolved resistance to PCB- and PAH-induced cardiac teratogenesis, and reduced CYP1A activity in Gulf killifish (Fundulus grandis) populations from the Houston Ship Channel, Texas. Aquat. Toxicol. *150*, 210–219.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics *23*, 1061–1067.

Philo, L.M., Shotts, E.B., and George, J.C. (1993). Morbidity and mortality. In The Bowhead Whale, J.J. Burns, J.J. Montague, and C.J. Cowles, eds. (Lawrence, Kansas: Allen Press), pp. 275–312.

Quesada, V., Ordóñez, G.R., Sánchez, L.M., Puente, X.S., and López-Otín, C. (2009). The Degradome database: mammalian proteases and diseases of proteolysis. Nucleic Acids Res. *37*, D239–D243.

Rogina, B., Helfand, S.L., and Frankel, S. (2002). Longevity regulation by Drosophila Rpd3 deacetylase and caloric restriction. Science *298*, 1745.

Ross, P.L., Cheng, I., Liu, X., Cicek, M.S., Carroll, P.R., Casey, G., and Witte, J.S. (2009). Carboxypeptidase 4 gene variants and early-onset intermediate-to-high risk prostate cancer. BMC Cancer *9*, 69.

Seim, I., Ma, S., Zhou, X., Gerashchenko, M.V., Lee, S.G., Suydam, R., George, J.C., Bickham, J.W., and Gladyshev, V.N. (2014). The transcriptome of the bowhead whale Balaena mysticetus reveals adaptations of the longest-lived mammal. Aging (Albany, N.Y. Online) *6*, 879–899.

Tacutu, R., Craig, T., Budovsky, A., Wuttke, D., Lehmann, G., Taranukha, D., Costa, J., Fraifeld, V.E., and de Magalhães, J.P. (2013). Human Ageing

Genomic Resources: integrated databases and tools for the biology and genetics of ageing. Nucleic Acids Res. *41*, D1027–D1033.

Tanno, M., Ogihara, M., and Taguchi, T. (1996). Age-related changes in proliferating cell nuclear antigen levels. Mech. Ageing Dev. *92*, 53–66.

Tervo, O.M., Christoffersen, M.F., Parks, S.E., Kristensen, R.M., and Madsen, P.T. (2011). Evidence for simultaneous sound production in the bowhead whale (Balaena mysticetus). J. Acoust. Soc. Am. *130*, 2257–2262.

Tian, X., Azpurua, J., Hine, C., Vaidya, A., Myakishev-Rempel, M., Ablaeva, J., Mao, Z., Nevo, E., Gorbunova, V., and Seluanov, A. (2013). High-molecular-mass hyaluronan mediates the cancer resistance of the naked mole rat. Nature *499*, 346–349.

van Beekum, O., Gao, Y., Berger, R., Koppen, A., and Kalkhoven, E. (2012). A novel RNAi lethality rescue screen to identify regulators of adipogenesis. PLoS ONE *7*, e37680.

van Dam, S., Craig, T., and de Magalhães, J.P. (2015). GeneFriends: a human RNA-seq-based gene and transcript co-expression database. Nucleic Acids Res. http://dx.doi.org/10.1093/nar/gku1042

Vervoort, V.S., Viljoen, D., Smart, R., Suthers, G., DuPont, B.R., Abbott, A., and Schwartz, C.E. (2002). Sorting nexin 3 (SNX3) is disrupted in a patient with a translocation t(6;13)(q21;q12) and microcephaly, microphthalmia, ectrodactyly, prognathism (MMEP) phenotype. J. Med. Genet. *39*, 893–899.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. *19*, 327–335.

Vindeløv, L.L., and Christensen, I.J. (1994). Detergent and proteolytic enzyme-based techniques for nuclear isolation and DNA content analysis. Methods Cell Biol. *41*, 219–229.

Vinogradov, A.E. (1998). Genome size and GC-percent in vertebrates as determined by flow cytometry: the triangular relationship. Cytometry *31*, 100–109.

Weeda, G., Donker, I., de Wit, J., Morreau, H., Janssens, R., Vissers, C.J., Nigg, A., van Steeg, H., Bootsma, D., and Hoeijmakers, J.H.J. (1997). Disruption of mouse ERCC1 results in a novel repair syndrome with growth failure, nuclear abnormalities and senescence. Curr. Biol. *7*, 427–439.

West, G.B., Woodruff, W.H., and Brown, J.H. (2002). Allometric scaling of metabolic rate from molecules and mitochondria to cells and mammals. Proc. Natl. Acad. Sci. USA *99* (*Suppl 1*), 2473–2478.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Yang, Z., and dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. Mol. Biol. Evol. *28*, 1217–1228.

Yang, Z., Wong, W.S.W., and Nielsen, R. (2005). Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. *22*, 1107–1118.

Yim, H.S., Cho, Y.S., Guang, X., Kang, S.G., Jeong, J.Y., Cha, S.S., Oh, H.M., Lee, J.H., Yang, E.C., Kwon, K.K., et al. (2014). Minke whale genome and aquatic adaptation in cetaceans. Nat. Genet. *46*, 88–92.

Zhang, J., Zhang, Y.P., and Rosenberg, H.F. (2002). Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat. Genet. *30*, 411–415.

Zhang, J., Nielsen, R., and Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. *22*, 2472–2479.

With Cell Press Reviews

Looking for authoritative reviews on the forefront of science? Turn to Cell Press. Our editors stay abreast of the latest developments so they can commission expert, cutting-edge reviews.

To propel your research forward, faster, turn to Cell Press Reviews. Our insightful, authoritative reviews—published across the life sciences in our primary research and *Trends* journals—go beyond synthesis and offer a point of view.

Find your way with Cell Press Reviews

www.cell.com/reviews

CellPress
Your work is our life

The 2016 Keystone Symposia Meeting Series

Systems Immunology: From Molecular Networks to Human Biology (A1)
Jan 10–14, 2016 | Big Sky, Montana | USA

Cytokine JAK-STAT Signaling in Immunity & Disease (A2)
Jan 10–14, 2016 | Steamboat Springs, Colorado | USA

Molecular & Cellular Basis of Growth & Regeneration (A3)
Jan 10–14, 2016 | Breckenridge, Colorado | USA

Nuclear Receptors: Full Throttle (J1)
joint with **Metabolism, Transcription & Disease (J2)**
Jan 10–14, 2016 | Snowbird, Utah | USA

Biology of Down Syndrome: Impacts Across the Biomedical Spectrum (A4)
Jan 24–27, 2016 | Santa Fe, New Mexico | USA

Traumatic Brain Injury: Clinical, Pathological & Translational Mechanisms (J3)
joint with **Axons: From Cell Biology to Pathology (J4)**
Jan 24–27, 2016 | Santa Fe, New Mexico | USA

Drug Discovery for Parasitic Diseases (A5)
Jan 24–28, 2016 | Tahoe City, California | USA

Small RNA Silencing: Little Guides, Big Biology (A6)
Jan 24–28, 2016 | Keystone, Colorado | USA

Purinergic Signaling (J5) *joint with* **Cancer Immunotherapy: Immunity & Immunosuppression Meet Targeted Therapies (J6)**
Jan 24–28, 2016 | Vancouver, British Columbia | Canada

Neurological Disorders of Intracellular Trafficking (A7)
Jan 31–Feb 4, 2016 | Keystone, Colorado | USA

Cell Biology & Immunology of Persistent Infection (A8)
Jan 31–Feb 4, 2016 | Banff, Alberta | Canada

The Cancer Genome (Q1) *joint with* **Genomics & Personalized Medicine (Q2)**
Feb 7–11, 2016 | Banff, Alberta | Canada

Fibrosis: From Basic Mechanisms to Targeted Therapies (Q3)
joint with **Stromal Cells in Immunity (Q4)**
Feb 7–11, 2016 | Keystone, Colorado | USA

Plant Epigenetics: From Genotype to Phenotype (B1)
Feb 15–19, 2016 | Taos, New Mexico | USA

Obesity & Adipose Tissue Biology (B2)
Feb 15–19, 2016 | Banff, Alberta | Canada

Noncoding RNAs in Health & Disease (Q5)
joint with **Enhancer Malfunction in Cancer (Q6)**
Feb 21–24, 2016 | Santa Fe, New Mexico | USA

G Protein-Coupled Receptors: Structure, Signaling & Drug Discovery (B3)
Feb 21–25, 2016 | Keystone, Colorado | USA

New Frontiers in Understanding Tumor Metabolism (Q7)
joint with **Immunometabolism in Immune Function & Inflammatory Disease (Q8)**
Feb 21–25, 2016 | Banff, Alberta | Canada

T Follicular Helper Cells & Germinal Centers (B4)
Feb 26–Mar 1, 2016 | Monterey, California | USA

Immunity in Skin Development, Homeostasis & Disease (B5)
Feb 28–Mar 2, 2016 | Tahoe City, California | USA

Tuberculosis Co-Morbidities & Immunopathogenesis (B6)
Feb 28–Mar 3, 2016 | Keystone, Colorado | USA

Stem Cells & Cancer (C1)
Mar 6–10, 2016 | Breckenridge, Colorado | USA

Cancer Vaccines: Targeting Cancer Genes for Immunotherapy (X1)
joint with **Antibodies as Drugs (X2)**
Mar 6–10, 2016 | Whistler, British Columbia | Canada

Ubiquitin Signaling (X3) *joint with* **NF-κB & MAP Kinase Signaling in Inflammation (X4)**
Mar 13–17, 2016 | Whistler, British Columbia | Canada

Islet Biology: From Cell Birth to Death (X5)
joint with **Stem Cells & Regeneration in the Digestive Organs (X6)**
Mar 13–17, 2016 | Keystone, Colorado | USA

Chromatin & Epigenetics (C2)
Mar 20–24, 2016 | Whistler, British Columbia | Canada

HIV Persistence: Pathogenesis & Eradication (X7) *joint with* **HIV Vaccines (X8)**
Mar 20–24, 2016 | Olympic Valley, California | USA

Cancer Pathophysiology: Integrating the Host & Tumor Environments (C3)
Mar 28–Apr 1, 2016 | Breckenridge, Colorado | USA

Modern Phenotypic Drug Discovery: Defining the Path Forward (D1)
Apr 2–6, 2016 | Big Sky, Montana | USA

Mitochondrial Dynamics (D2)
Apr 3–7, 2016 | Steamboat Springs, Colorado | USA

Heart Failure: Genetics, Genomics & Epigenetics (Z1)
joint with **Cardiac Development, Regeneration & Repair (Z2)**
Apr 3–7, 2016 | Snowbird, Utah | USA

Myeloid Cells (D3)
Apr 10–14, 2016 | Killarney, County Kerry | Ireland

New Therapeutics for Diabetes & Obesity (G1)
Apr 17–20, 2016 | La Jolla, California | USA

Gut Microbiota, Metabolic Disorders & Beyond (D4)
Apr 17–21, 2016 | Newport, Rhode Island | USA

Epigenetic & Metabolic Regulation of Aging & Aging-Related Diseases (E1)
May 1–5, 2016 | Santa Fe, New Mexico | USA

Positive-Strand RNA Viruses (N1)
May 1–5, 2016 | Austin, Texas | USA

Nucleic Acid Sensing Pathways: Innate Immunity, Immunobiology & Therapeutics (E2)
May 8–12, 2016 | Dresden | Germany

State of the Brain (R1)
May 22–26, 2016 | Alpbach | Austria

New Approaches to Vaccines for Human & Veterinary Tropical Diseases (M1)
May 22–26, 2016 | Cape Town | South Africa

B Cells at the Intersection of Innate & Adaptive Immunity (E3)
May 29–Jun 2, 2016 | Stockholm | Sweden

Understanding the Function of Human Genome Variation (K1)
May 31–Jun 4, 2016 | Uppsala | Sweden

Autophagy: Molecular & Physiological Mechanisms (V1)
Jun 5–9, 2016 | Whistler, British Columbia | Canada

Common Mechanisms of Neurodegeneration (Z3)
joint with **Microglia in the Brain (Z4)**
Jun 12–16, 2016 | Keystone, Colorado | USA

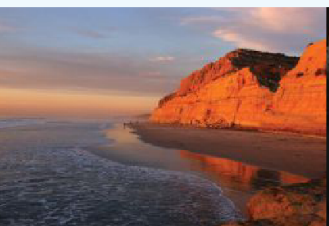
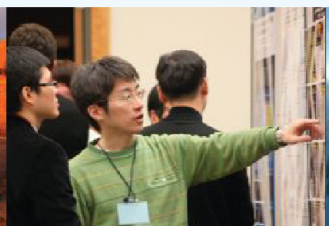
Exosomes/Microvesicles: Novel Mechanisms of Cell-Cell Communication (E4)
Jun 19–22, 2016 | Keystone, Colorado | USA

Translational Vaccinology for Global Health (S1)
Oct 26–30, 2016 | London | United Kingdom

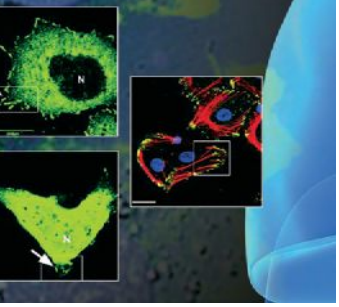
Phytobiomes: From Microbes to Plant Ecosystems (S2)
Nov 8–12, 2016 | Guarujá, São Paulo | Brazil

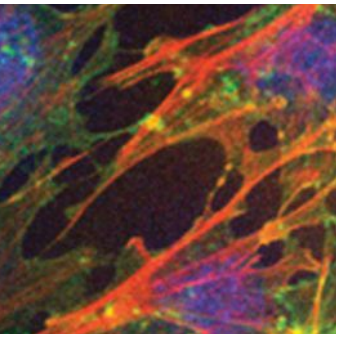
Hemorrhagic Fever Viruses (S3)
Dec 4–8, 2016 | Santa Fe, New Mexico | USA

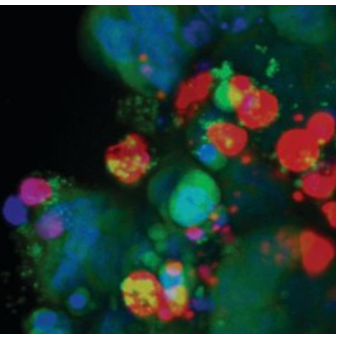
Cellular Stress Responses and Infectious Agents (S3)
Dec 4–8, 2016 | Santa Fe, New Mexico | USA

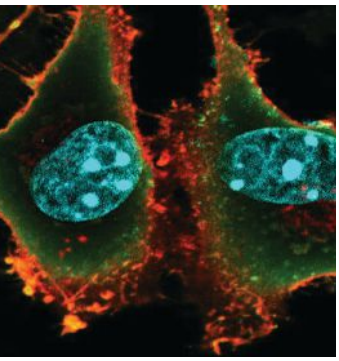



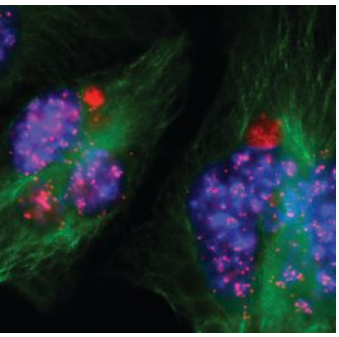

KEYSTONE SYMPOSIA™
on Molecular and Cellular Biology
Accelerating Life Science Discovery
www.keystonesymposia.org/meetings | 1.800.253.0685 | 1.970.262.1230











2016 SCIENTIFIC CONFERENCES

Presenting the most significant research on cancer etiology, prevention, diagnosis, and treatment

Fourth AACR-IASLC International Joint Conference: Lung Cancer Translational Science From the Bench to the Clinic
Conference Chairpersons: Karen L. Kelly and Alice T. Shaw
Co-Chairpersons: Stephen B. Baylin, Jeffrey A. Engelman, Roy S. Herbst, and Pierre P. Massion
January 4-7, 2016 • San Diego, CA

The Function of Tumor Microenvironment in Cancer Progression
Conference Co-Chairpersons: Raghu Kalluri, Robert A. Weinberg, Douglas Hanahan, and Morag Park
January 7-10, 2016 • San Diego, CA

Patient-Derived Cancer Models: Present and Future Applications from Basic Science to the Clinic
Conference Co-Chairpersons: Manuel Hidalgo, Hans Clevers, S. Gail Eckhardt, and Joan Seoane
February 11-14, 2016 • New Orleans, LA

10th AACR-JCA Joint Conference on Breakthroughs in Cancer Research: From Biology to Therapeutics
Co-Chairpersons: Frank McCormick and Tetsuo Noda
February 16-20, 2016 • Maui, HI

AACR Precision Medicine Series: Cancer Cell Cycle-Tumor Progression and Therapeutic Response
Conference Chairperson: Julien Sage
Co-Chairpersons: Karen E. Knudsen and J. Alan Diehl
February 28-March 2, 2016 • Orlando, FL

AACR Annual Meeting 2016
Program Committee Chairperson: Scott A. Armstrong
April 16-20, 2016 • New Orleans, LA

Pancreatic Cancer
Conference Co-Chairpersons: Manuel Hidalgo, Christine Iacobuzio-Donahue, and Robert H. Vonderheide
May 12-15, 2016 • Orlando, FL

AACR Precision Medicine Series: Targeting the Vulnerabilities of Cancer
Conference Co-Chairpersons: Stephen W. Fesik, Jeffrey E. Settleman, and Paul Workman
May 16-19, 2016 • Miami, FL

Engineering and Physical Sciences in Oncology
Conference Co-Chairpersons: Rakesh Jain, Robert Langer, and Joan Brugge
June 25-28, 2016 • Boston, MA

Translational Control of Cancer: A New Frontier in Cancer Biology and Therapy
Conference Co-Chairpersons: Jennifer A. Doudna, Frank McCormick, Davide Ruggero, and Nahum Sonenberg
October 27-30, 2016 • San Francisco, CA

DNA Repair: Tumor Development and Therapeutic Response
Conference Co-Chairpersons: Robert G. Bristow, Theodore S. Lawrence, and Maria Jasin
November 2-5, 2016 • Montreal, Quebec, Canada

EORTC-NCI-AACR Molecular Targets and Cancer Therapeutics Symposium
November 29-December 2, 2016 • Munich, Germany

Learn more and register at
www.AACR.org/Calendar

AACR American Association for Cancer Research
FINDING CURES TOGETHER℠

We make human cells.

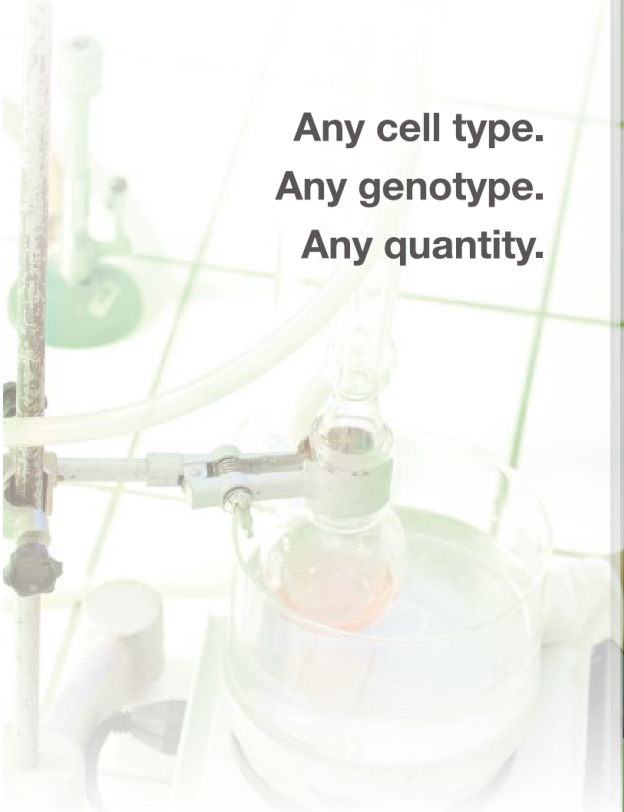
We can't state it more simply: We manufacture human cells from induced pluripotent stem cells. Our highly pure iCell® and MyCell® products are consistent and reproducible from lot-to-lot, enabling you to focus on the biology . . . and what drives your research.

True Human Biology. Why use an animal or immortalized cell model as a proxy when a true human model, consistent across lots, is available? If you have a phenotype of interest, we can make iPSCs and terminal cells from your donors

Highly Characterized. Highly Predictive. Our cells have the appropriate gene and protein expression and the relevant physiological response of the target cell type, with numerous peer-reviewed publications detailing their characterization and predictivity of human response

MORE Cells. CDI provides more cells than other providers: 96-well capacity per vial. GUARANTEED

Any cell type.
Any genotype.
Any quantity.



Cellular Dynamics international
a FUJIFILM company

You make:

safer medicines
discoveries
therapies
breakthroughs

Check out the iCell community of scientists who discuss the difference iCell products have made in their research:

www.cellulardynamics.com/community