Top 5 Cell Biology Papers

These papers provide a breadth of information about Cell Biology that is generally useful and interesting from a biological science perspective.

Contents

Published in final edited form as:
Science. 2010 June 4; 328(5983): 1281–1284. doi:10.1126/science.1188462.

FCHo Proteins are Nucleators of Clathrin-Mediated Endocytosis

William Mike Henne[1,2,*], **Emmanuel Boucrot**[1,*,†], **Michael Meinecke**[1], **Emma Evergren**[1], **Yvonne Vallis**[1], **Rohit Mittal**[1], and **Harvey T. McMahon**[1,†]

[1]MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 0QH, UK

Abstract

Clathrin-mediated endocytosis, the major pathway for ligand internalization into eukaryotic cells, is thought to be initiated by clustering of clathrin and adaptors around receptors destined for internalization. However, here we report that the membrane-sculpting F-BAR domain-containing FCHo1 and 2 (FCHo1/2) proteins were required for plasma membrane clathrin-coated vesicle (CCV) budding and marked sites of CCV formation. Changes in FCHo1/2 expression levels correlated directly with numbers of CCV budding events, ligand endocytosis, and synaptic vesicle marker recycling. FCHo1/2 proteins bound specifically to the plasma membrane and recruited the scaffold proteins, eps15 and intersectin, which in turn engaged the adaptor-complex AP2. The FCHo F-BAR membrane-bending activity was required, leading to the proposal that FCHo1/2 sculpt the initial bud site and recruit the clathrin machinery for CCV formation.

Clathrin-mediated endocytosis is the process by which cargo is internalized into vesicles with the aid of adaptors (like AP2) and the coat-protein clathrin (1, 2). Amphiphysins and sorting nexin 9 likely recruit the membrane scission protein dynamin to membranes of high curvature by their N-terminal BAR (Bin/Amphiphysin/Rvs) domains (3, 4). Here we investigate the possibility that membrane-sculpting proteins play an early role in invagination even prior to adaptor and clathrin recruitment. We studied the F-BAR-containing protein family FCHo1/2 (Fer/Cip4 homology domain only proteins 1 and 2) whose F-BAR homodimer module can recognize less extreme curvatures than BAR modules (5-7). FCHo1/2 are ubiquitously expressed (fig. S1A and B) and have a twisted shape distinct from the F-BAR dimers of FBP17 and CIP4 (5-7). The yeast homolog, Syp1, is recruited early to sites of actin-dependent endocytosis (8-10). We confirmed that FCHo1/2 are localized to clathrin-coated pits (CCPs) only on the plasma membrane (PM) (fig. S1C-F). Furthermore, an FCHo signal defined where a CCP forms, because it was detected before the visible appearance of clathrin or its PM specific adaptor, AP2 (Fig. 1A,C and fig. S2A-B). The FCHo1/2 signal decreased before the clathrin signal intensity reached its maximum but in some rare cases the FCHo protein did not leave, and defined sites where clathrin returned multiple times, thus marking endocytic ‘hot-spots’ (fig. S2C). FBP17, another F-BAR protein implicated in clathrin-mediated endocytosis (6) was, in contrast, recruited at later stages to some (3±1%) CCPs (fig. S2E). FCHo2 was detected by cryoimmuno-electron microscopy at early to late stages of CCPs consistent with live cell imaging kymographs

[†]To whom correspondence should be addressed: hmm@mrc-lmb.cam.ac.uk and eboucrot@mrc-lmb.cam.ac.uk.
[*]These authors contributed equally to this work.
[2]present address: Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853, USA.

Supporting Online Material
www.sciencemag.org
Materials and Methods
Supporting Text
Figs. S1-S12
Movies S1, S2

Europe PMC Funders Author Manuscripts

(Fig. 1B). A complete loss of CCPs was observed when FCHo1+2 levels were greatly reduced using double RNAi (Fig. 1C-D and fig. S3A-D) with a concomitant reduction in internalization of three known cargoes for clathrin-mediated endocytosis: transferrin (Tf), low-density lipoprotein (LDL), and epidermal growth factor (EGF) (Fig. 1E). In the absence of FCHo proteins, both AP2 and clathrin were cytosolic. These phenotypes were rescued by an RNAi-resistant form of FCHo2 (Fig. 1C-E and fig. S3E) (11). FCHo1/2 function was not limited to fibroblasts but was also associated with clathrin-mediated endocytosis in primary astrocytes and recycling of synaptic vesicle markers (synaptotagmin1 and synaptophysin) following stimulated exocytosis in 4 DIV (days in vitro) hippocampal neurons (fig. S4). Overexpression of FCHo1 or FCHo2 led to a dramatic increase in CCP density. This increase wasnot due to slowed or inhibited CCV budding (as with epsin1 overexpression (fig. S5B) or dynamin inhibition by dynasore(12)) as CCPs were dynamic (increased nucleation rate) and functional (increased Tf uptake) during FCHo1/2 overexpression (Fig. 1F-H, fig. S5, Movie S1). Because CCP numbers directly correlate with FCHo1/2 levels, FCHo proteins appear to act as CCP nucleators.

The presence of a membrane bending protein early in CCP formation caused us to seek an explanation for how FCHo1/2 recruitment connects to clathrin recruitment. We looked for FCHo2 C-terminal AP2-μ homology domain (μHD) interactors in brain and HeLa cell extracts, and found that the main interaction partners were known CCP proteins eps15, eps15R and intersectin 1 and 2 (Fig. 2A and fig. S6A-B). The interaction of eps15 with the μHD was direct (fig. S6C and (8)) as was that with intersectin1. The CCP localization of eps15 and intersectin was dependent on FCHo1/2 (Fig. 2B). Notably, these proteins interact directly with AP2 but not with clathrin, unlike many other CCV accessory proteins (2). Eps15 and intersectin appearance coincided with that of FCHo2 (Fig. 2C), suggesting that these proteins constitute an early module for nascent CCP assembly. FCHo1/2 are necessary for CCP formation and yet their fluorescent intensity diminished before vesicle budding (Fig. 1A), just as dynamin intensity increased (fig. S6D). Similarly, the yeast Syp1 intensity decreased as Abp1 increased (9, 10). In purified CCVs, FCHo1/2, eps15 and intersectin levels were reduced compared to total extracts (Fig. 2D and fig. S6E), consistent with their absence in previous mass spectrometry studies (13). Thus, FCHo1/2 initiate CCPs but are excluded from mature vesicles, with FCHo1/2 being primarily PM-associated, consistent with their localization on constricted CCP necks (Fig. 2E).

RNAi of AP2 leads to a marked reduction in CCP numbers (14). Thus, we tested the localization of FCHo2 in the absence of AP2. While AP2 puncta were largely missing, FCHo2 puncta at the PM remained and still colocalized with eps15 and intersectin (Fig. 2G and fig. S7A). The FCHo2-μHD ligand interaction was also AP2-independent in vitro (Fig. 2F and fig. S7B). In contrast, the localization of epsin, another membrane sculpting molecule that binds clathrin and AP2, was dependent on the presence of AP2 (fig. S7C). An alternative strategy to disrupt CCP formation is the overexpression of the C-terminus of AP180, which binds to clathrin with high affinity (15). Overexpression led to an accumulation of AP2 puncta which co-localized with FCHo2, eps15 and intersectin but had no clathrin, and were static (fig. S7D-F). Thus FCHo2, eps15 and intersectin do not require AP2 or clathrin to cluster. An eps15/eps15R/intersectin1/intersectin2 quadruple knockdown affected FCHo2 clustering into puncta but not its PM localization, while AP2 was cytosolic (Fig. 2H and fig. S8A-E). RNAi of Dab2, a μHD interaction partner that arrives early at CCPs (fig. S6A and S8F) and that was not enriched in CCVs (Fig. 2D), did not lead to a reduction in CCPs (fig. S8F-G). Thus eps15 and intersectin cluster FCHo1/2 at nascent sites of CCP nucleation. Mutation of K797 in FCHo2 μHD (fig. S9A, a conserved residue that is equivalent to where AP2-β interacts with AP2-μ in pdb:2VGL) (16), abolished interactions with eps15 and intersectins (Fig. 2J and fig. S9B). In the FCHo1+2 RNAi background, K797E did not rescue CCP formation and Tf uptake, and was diffusely located on the PM

(Fig. 2I). Thus FCHo membrane recruitment and clustering by eps15 and intersectins initiates CCP maturation with subsequent recruitment of AP2 and clathrin leading to coated vesicle formation.

Functionality of F-BAR domains is mediated by three distinct properties: membrane binding, dimerization, and membrane sculpting (5, 6, 17). To test the importance of each property in FCHo2 function, we designed: i) a chimera replacing the F-BAR domain with a PM targeting PH domain, which dimerizes because of an EGFP tag (18), ii) a structure-based mutant of FCHo2 (F38E+W73E) which should disrupt dimer formation, and iii) a fluorescently-tagged SGIP1, a close relative of FCHo1/2 and a CCP component which has a C-terminal μHD but no F-BAR domain (19). All three proteins localized to CCPs in wild-type cells (fig. S10A), but upon depletion of endogenous FCHo1+2 did not rescue CCP formation (Fig 3A). As expected, both the PH chimera and SGIP1 localized to the PM (sometimes in large sheets) whereas the dimer mutant remained cytosolic. Thus, the dimeric, membrane sculpting F-BAR module is necessary for CCP formation. The region following the FCHo1/2 F-BAR domain (residues 263-430) is rich in positively charged amino acids and has a high homology with the N-terminus of SGIP1 (9). An extended F-BAR module containing this homology region (F-BAR-x for e**x**tended), showed enhanced membrane binding and tubulation in vitro (fig. S10B,C). The F-BAR-x module co-sedimented preferentially with liposomes enriched with $Pi(4,5)P_2$, helping to explain why FCHo proteins are PM targeted (Fig. 3B). Acute decrease of cellular $Pi(4,5)P_2$ levels by the addition of 1-butanol (20) led to acute relocalization of FCHo2 to the cytosol (fig. S10D), supporting the role of $Pi(4,5)P_2$ in the targeting of FCHo1/2 to the PM. The F-BAR-x module caused extensive tubulation of $Pi(4,5)P_2$-liposomes to high curvatures (from 130 to 18nm tubules and many small vesicles) in a protein concentration-dependent manner (Fig. 3C). Protein density surrounding tubules sometimes exhibited striations where the angle correlated with the degree of membrane curvature (fig. S10E and Movie S2). Narrower tubules displayed more oblique angles and the narrowest ones were twisted (Fig 3C) (5, 17). This provides a mechanistic explanation for the generation of increasing curvature required for CCP budding and dynamin recruitment. To test the contribution of membrane sculpting in FCHo2 function, we mutated two conserved lysines (K146E+K165E) on the concave face of the F-BAR module as well as a conserved residue (I268N) associated with a macrophage-induced autoimmune disease (21) and its interacting residue (L136E) along the F-BAR 'wing' (fig. S11A,B). As expected, the K146E+K165E mutant displayed reduced membrane binding in vitro and relocalized to the cytosol, whereas I268N and L136E mutations did not abrogate membrane binding but failed to tubulate the PM (fig. S11C,D). When placed into full-length FCHo2, I268N and L136E induced enlarged and static aberrant CCPs and could not rescue FCHo1+2 RNAi-induced Tf uptake defect (Fig. 3E). Thus FCHo2-mediated membrane sculpting is essential for normal CCP nucleation.

We showed that FCHo1/2 proteins nucleate CCPs and that AP2 is a later component recruiting clathrin, cargo and accessory proteins (fig. S12). We also uncovered a role for membrane sculpting in the initiation of clathrin-mediated endocytosis. Thus curvature generation appears to be fundamental to plasma membrane CCV formation from neurons to fibroblasts and FCHo 1 and 2 represent key initial proteins ultimately controlling cellular nutrient uptake, receptor regulation and synaptic vesicle retrieval.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Europe PMC Funders Author Manuscripts

Acknowledgments

We thank S.-Y. Peak-Chew for mass spectrometry, M. Daly for cell sorting, G. Lingley for movie design, J.E. Tyrrell for summer assistance and Perkin Elmer for support with spinning-disk microscopy. Support was provided to H.M.M. (Medical Research Council, UK), W.M.H. (LMB PhD Scholarship), E.B. (Human Frontiers Science Program Organization and MRC), M.M. and E.E. (EMBO fellowships).

References and Notes

1. Traub LM. Nat Rev Mol Cell Biol. 2009; 10:583. [PubMed: 19696796]
2. Schmid EM, McMahon HT. Nature. 2007; 448:883. [PubMed: 17713526]
3. Peter BJ, et al. Science. 2004; 303:495. [PubMed: 14645856]
4. Lundmark R, Carlsson SR. J Biol Chem. 2003; 278:46772. [PubMed: 12952949]
5. Henne WM, et al. Structure. 2007; 15:839. [PubMed: 17540576]
6. Shimada A, et al. Cell. 2007; 129:761. [PubMed: 17512409]
7. http://www.endocytosis.org/F-BAR_proteins/BAR-Comparisons.html
8. Reider A, et al. Embo J. 2009; 28:3103. [PubMed: 19713939]
9. Stimpson HE, Toret CP, Cheng AT, Pauly BS, Drubin DG. Mol Biol Cell. 2009; 20:4640. [PubMed: 19776351]
10. Boettner DR, et al. Curr Biol. 2009; 19:1979. [PubMed: 19962315]
11. Materials and methods are available as supporting material on Science Online.
12. Macia E, et al. Dev Cell. 2006; 10:839. [PubMed: 16740485]
13. Blondeau F, et al. Proc Natl Acad Sci U S A. 2004; 101:3833. [PubMed: 15007177]
14. Motley A, Bright NA, Seaman MN, Robinson MS. J Cell Biol. 2003; 162:909. [PubMed: 12952941]
15. Ford MG, et al. Science. 2001; 291:1051. [PubMed: 11161218]
16. Collins BM, McCoy AJ, Kent HM, Evans PR, Owen DJ. Cell. 2002; 109:523. [PubMed: 12086608]
17. Frost A, et al. Cell. 2008; 132:807. [PubMed: 18329367]
18. Zacharias DA, Violin JD, Newton AC, Tsien RY. Science. 2002; 296:913. [PubMed: 11988576]
19. Uezu A, et al. J Biol Chem. 2007; 282:26481. [PubMed: 17626015]
20. Boucrot E, Saffarian S, Massol R, Kirchhausen T, Ehrlich M. Exp Cell Res. 2006; 312:4036. [PubMed: 17097636]
21. Grosse J, et al. Blood. 2006; 107:3350. [PubMed: 16397132]

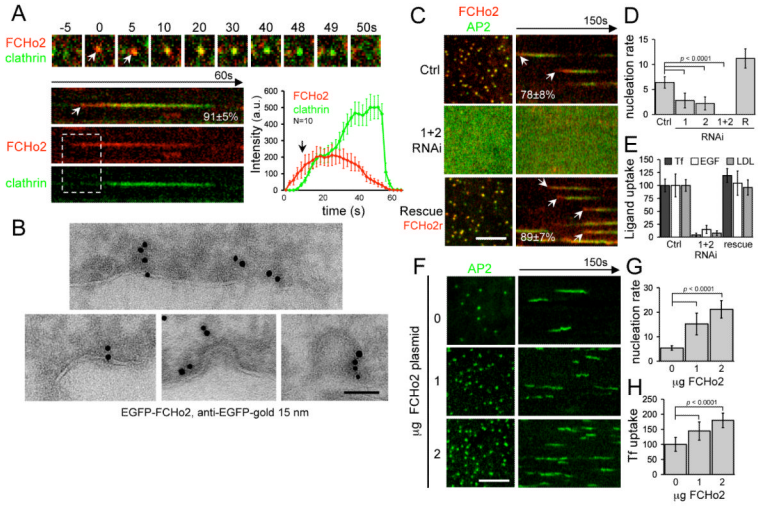


Fig. 1.

FCHo1/2 proteins are clathrin/AP2 nucleators. (**A**) Dynamic cell surface localization (top) and kymograph (bottom) of representative CCPs labelled with RFP-FCHo2 (FCHo2) and GFP-LCa (clathrin). FCHo2 was detected before clathrin (white arrows and graph). a.u.: arbitary units. (**B**) Cryoimmuno-electron microscopy localized GFP-FCHo2 at CCPs. Scale bar 100nm. (**C**) CCVs, labelled by σ2-GFP (AP2), did not form with double RNAi of FCHo1+2 (FCHo1+2 RNAi) where AP2 became cytosolic (contrast enhanced to show the diffuse signal at the plasma membrane). Inhibition was relieved by co-expression of RNAi-resistant RFP-FCHo2 (rescue). (**D**) Nucleation rates (number of new CCPs/10^4µm^2/second) in cells treated with scrambled RNAi (Ctrl), RNAi against FCHo1 (1), FCHo2 (2), FCHo1+2 (1+2) or rescue (R). (**E**) Clathrin ligands, transferrin (Tf) epidermal growth factor (EGF) and low-density lipoprotein (LDL) uptake in cells treated as in C. (**F**) Clathrin vesicles (AP2) in BSC1 cells transfected with 0, 1 or 2µg of untagged-FCHo2 for 2×10^5 cells. (**G**) Nucleation rate and (**H**) Tf uptake in cells treated as in F. Scale bars, 5µm (C,F) and 200nm (B). Displayed kymographs were representative (percentage, n=319 CCPs (11)).

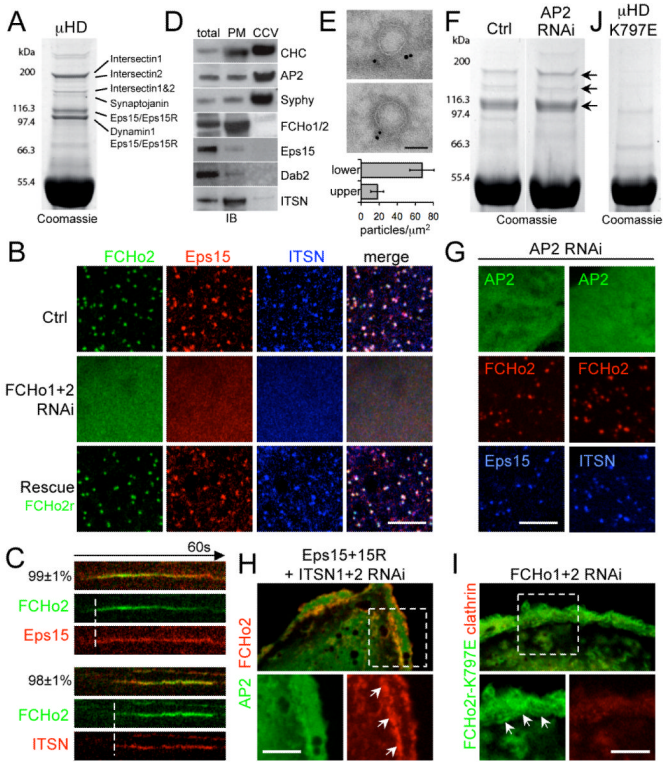
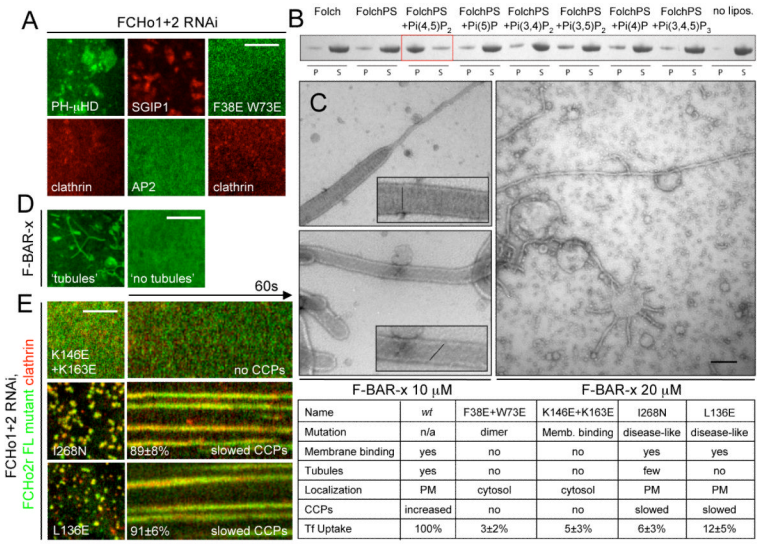


Fig. 2.

FCHo2 directly binds and recruits eps15 and intersectin to initiate CCP maturation. (**A**) Pull down with GST-FCHo2-μHD and rat brain lysate. Interacting proteins were identified by mass spectrometry. (**B**) Eps15 and intersectin (ITSN) formed puncta at the PM colocalizing with FCHo2. In double FCHo1+2 RNAi cells, Eps15 and ITSN were cytosolic (contrast enhanced to show the diffuse signal); co-expression of RNAi-resistant FCHo2 (FCHo2r) rescued PM-targeting (Rescue). (**C**) Kymograph of representative CCPs (percentage, (11)) labeled with FCHo2 and Eps15 or ITSN. (**D**) FCHo2, eps15, and ITSN were CCV de-enriched. Clathrin (CHC), AP2, and vesicle marker synaptophysin (Syphy) displayed enrichment in CCV fractions (CCV). FCHo2 and ITSN were PM enriched. IB: immunoblot. (**E**) Cryoimmuno-electron microscopy of GFP-FCHo2 localized it to the CCP neck. Bar graph shows gold particle density in the upper and lower half of constricted CCPs ($p<0.01$). (**F**) Pull downs with GST-μHD from scrambled (left) or AP2 RNAi (right) treated HeLa cells. Eps15 and ITSN bands were visible in both (arrows). (**G**) Upon AP2 depletion (μ2 RNAi) σ2-GFP (another AP2 subunit) was cytosolic (contrast enhanced to show the diffuse signal), but Eps15 and ITSN still co-localize with FCHo2 at the PM (arrows). (**H**) FCHo2 and AP2 (σ2-GFP) puncta disappeared under Eps15 + Eps15R + ITSN1 + ITSN2 quadruple RNAi. AP2 became cytosolic (diffuse signal) whereas FCHo2 remained at the PM (inset). (**I**) In FCHo1+2 RNAi cells, RNAi-resistant FCHo2-K797E (FCHo2r-K797E) bound to the PM (arrows) but did not cluster nor rescue CCP formation, reported by RFP-LCa (clathrin). In these cells FCHo1+2 RNAi inhibition of Tf uptake was also not rescued (7.2±3.5% of control uptake ($p<0.0001$). (**J**) GST-μHD K797E no longer pulled-down the protein bands visible in A. Green, red and blue panels indicate GFP-, RFP-, and BFP-tagged proteins respectively. Scale bars, 5μm. (B,H,G,I) and 100nm (E).

Europe PMC Funders Author Manuscripts



Name	wt	F38E+W73E	K146E+K163E	I268N	L136E
Mutation	n/a	dimer	Memb. binding	disease-like	disease-like
Membrane binding	yes	no	no	yes	yes
Tubules	yes	no	no	few	no
Localization	PM	cytosol	cytosol	PM	PM
CCPs	increased	no	no	slowed	slowed
Tf Uptake	100%	3±2%	5±3%	6±3%	12±5%

Fig. 3.

Lipid binding and membrane sculpting FCHo1/2 abilities are both essential for CCP formation. (**A**) Chimeric GFP-PLC-PH+FCHo2μHD (PH-μHD), a dimer interface mutant GFP-FCHo2(F38E+W73E), and RFP-SGIP1 all could not rescue CCP nucleation - monitored by following either clathrin or AP2 fluorescence - in double RNAi FCHo1+2 cells. (**B**) Lipid co-sedimentation assay of 15μM F-BAR-x in presence of 1mg/mL liposomes: Folch (Avanti brain lipid), FolchPS (80% Folch, 20% phosphatidylserine) or FolchPS +5% of indicated PiPs. Liposome-bound proteins were pelleted (P) by ultracentrifugation, unbound protein remained in the supernatant (S). (**C**) Folch+15%PS +5%Pi(4,5)P_2 liposomes incubated with either 10 or 20μM F-BAR-x and spotted onto EM grids gave mainly tubules of diameters of 50-80nm (10 μM) or 18nm (20μM, most were visibly twisted). Insets show enlargements with protein density striations. (**D**) F-BAR-x-induced in vivo tubulation: representative images of 'tubules' and 'no tubules'. (**E**) RNAi-resistant form of full-length FCHo2 (FCHo2r) with membrane-binding mutation (K146E +K163E) remained cyosolic and could not rescue FCHo1/2 RNAi-mediated absence of CCPs, whereas FCHo2r containing membrane-sculpting mutations (I268N and L136E) displayed slowed and aberrant CCPs. Displayed kymographs were representative (percentage, (11)). Table summarizing the mutants and their phenotypes. Scale bars, 5μm (A,D,E) and 100nm (C).

Published in final edited form as:
Nature. 2010 July 8; 466(7303): 243–247. doi:10.1038/nature09115.

Functionally defective germline variants of sialic acid acetylesterase in autoimmunity

Ira Surolia[1,*], **Stephan P. Pirnie**[1,*], **Vasant Chellappa**[1,*], **Kendra N. Taylor**[1,*], **Annaiah Cariappa**[1,*], **Jesse Moya**[1], **Haoyuan Liu**[1], **Daphne W. Bell**[1,†], **David Driscoll**[1], **Sven Diederichs**[1,¶], **Khaleda Haider**[1], **Ilka Netravali**[1], **Sheila Le**[1], **Roberto Elia**[1], **Ethan Dow**[1], **Annette Lee**[6], **Jan Freudenberg**[6], **Philip L. De Jager**[7], **Yves Chretien**[10], **Ajit Varki**[8], **Marcy E. MacDonald**[2], **Tammy Gillis**[2], **Timothy W. Behrens**[9], **Donald Bloch**[5], **Deborah Collier**[5], **Joshua Korzenik**[4], **Daniel K. Podolsky**[4,††], **David Hafler**[7], **Mandakolathur Murali**[3], **Bruce Sands**[4], **John H. Stone**[5], **Peter K. Gregersen**[6], and **Shiv Pillai**[1,§]

[1]Cancer Center, Massachusetts General Hospital, Harvard Medical School, Boston MA 02114

[2]Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, Boston MA 02114

[3]Clinical Immunology Laboratory, Massachusetts General Hospital, Harvard Medical School, Boston MA 02114

[4]Gastrointestinal Unit, Massachusetts General Hospital, Harvard Medical School, Boston MA 02114

[5]Division of Rheumatology, Massachusetts General Hospital, Harvard Medical School, Boston MA 02114

[6]Feinstein Institute for Medical Research, North Shore LIJ Health System, Manhasset New York, 11030

[7]Division of Molecular Immunology, Center for Neurologic Diseases, Brigham & Women's Hospital and Partners Center for Personalized Genetic Medicine, Boston, MA, 02115 and Program in Medical & Population Genetics, Broad Institute of Harvard University and Massachusetts Institute of Technology, Cambridge, MA 02142

[8]Departments of Medicine and Cellular & Molecular Medicine, University of California, San Diego, La Jolla, CA 92093

[9]Genentech Inc., South San Francisco, CA 94080

[§]To whom correspondence should be addressed. pillai@helix.mgh.harvard.edu.
[*]These authors contributed equally
[†]Present address, Cancer Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, Maryland
[¶]Present address, German Cancer Research Center and Institute of Pathology, University of Heidelberg, Germany
[††]Present address, University of Texas Southwestern Medical Center, Dallas, Texas

Author Contributions SP was responsible for overall study design and writing the manuscript. SPP, HL, JM, DRD, DWB, SL, TG, MEM, KNT, RE, AC, ED and SP contributed to sequencing and sequence analysis. Full length human SIAE was cloned by IS. IS, VC, SD, and IN performed mutagenesis, and IS, SPP, KH, VC, KNT, and AC performed functional analyses. Association studies, dominant negative analyses, and metabolic labeling studies were performed by VC. JF, AL, and PKG, performed the Principal Components Analysis, and MM, PKG, JHS, TWB, BS, DKP, JK, DH, PD, DC and DB provided annotated clinical material. AV provided advice on enzymology. Statistical analyses were performed by YC, IN and SP.

Author Information: Reprints and permission information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for reprints should be addressed to SP (pillai@helix.mgh.harvard.edu).

Supplementary information is linked to the online version of the paper at www.Nature.com/Nature

[10]Harvard University, Department of Statistics, Cambridge MA 02138

Abstract

Sialic acid acetylesterase (SIAE) is an enzyme that negatively regulates B lymphocyte antigen receptor signaling and is required for the maintenance of immunological tolerance in mice[1, 2]. Heterozygous loss-of-function germline rare variants and a homozygous defective polymorphic variant of *SIAE* were identified in 24/923 Caucasian subjects with relatively common autoimmune disorders and in 2/648 Caucasian controls. All heterozygous loss-of-function *SIAE* mutations tested were capable of functioning in a dominant negative manner. A homozygous secretion-defective polymorphic variant of *SIAE* was catalytically active, lacked the ability to function in a dominant negative manner, and was seen in 8 autoimmune subjects but in no control subjects. The Odds Ratio for inheriting defective *SIAE* alleles was 8.6 in all autoimmune subjects, 8.3 in subjects with rheumatoid arthritis, and 7.9 in subjects with type I diabetes. Functionally defective *SIAE* rare and polymorphic variants represent a strong genetic link to susceptibility in relatively common human autoimmune disorders.

Our previous studies revealed a defect in B cell tolerance as evidenced by the spontaneous development of autoantibodies in *Siae* mutant mice on a C57BL/6 background[1]. Given this phenotype we sought to ask if this enzyme was linked to autoimmunity in human subjects. Although genome wide association studies had not revealed altered frequencies of common variants of *SIAE* in patients with autoimmunity, the possibility that loss-of-function rare variants of this gene might be enriched in patients with autoimmune disorders was addressed by complete re-sequencing of all the exons of *SIAE* in patients with autoimmunity and in healthy controls.

In the first phase of our studies we completely re-sequenced the *SIAE* gene from 188 subjects with autoimmunity and 190 healthy controls as described below. Initially we analyzed 19 subjects from Massachusetts General Hospital (MGH) selected only on the basis of their having high ANA titers. 13 of these 19 subjects had defined autoimmune disorders and were included in our studies. In this initial set of 13 Caucasian subjects, unique non-synonymous changes were observed in one subject with Crohn's disease and in one subject with rheumatoid arthritis (RA). As a result of these preliminary observations, we next analyzed 76 Caucasian subjects with RA from the NARAC (North American Rheumatoid Arthritis Consortium) collection, and 89 subjects with inflammatory bowel disease (IBD) from MGH, making an initial total of 188 autoimmune subjects. The only criterion used in selection was ethnicity. The control DNAs in this initial phase were obtained from 190 healthy volunteers at MGH primarily of European ancestry. Re-sequencing of all 10 exons of *SIAE* revealed the existence of point substitutions in *SIAE* in both patients and controls. A number of known SNPs were identified as expected (Supplementary Table I). A total of 19 out of 923 autoimmune subjects presented with one of 14 previously unidentified non-synonymous SNPs in the *SIAE* gene, while 8 other autoimmune subjects had a homozygous polymorphism resulting in a valine replacing methionine at position 89 (Table I). Among control subjects, 17/648 presented with one of 8 non-synonymous SNPs in the *SIAE* gene. No controls presented with the homozygous 89V/89V polymorphic form of SIAE (Table I). Functional analyses were performed on each SIAE variant as described below.

Since the initial analyses revealed a marked enrichment of loss-of-function *SIAE* variants in autoimmune subjects as compared with controls, a larger number of autoimmune subjects and controls were analyzed. Power calculations revealed that a sample size of 550 cases and 550 controls would be required to obtain a power of at least 0.80 (see power table and calculations in Supplementary Notes). Autoimmune subjects in this second phase included

more subjects with RA from NARAC, subjects from the MADGC (Multiple Autoimmune Disease Genetics Consortium) collection with systemic lupus erythematosus (SLE) and juvenile idiopathic arthritis (JIA), and subjects from MGH with IBD and rheumatic disorders including SLE, RA, mixed connective tissue disorder (MCTD) and Sjogren's syndrome (SjS). We also included subjects with multiple sclerosis (MS) from a collection at the Brigham and Womens' Hospital, and subjects with Type 1 diabetes (T1D) from the EDIC (Epidemiology of Diabetes Intervention and Complication study) collection of the NIDDK (National Institute of Diabetes, Digestive and Kidney Diseases). Additional healthy control DNAs, primarily from subjects of European ancestry, were obtained from the MGH Cancer Center, the Feinstein Institute and the phenogenetic collection at Brigham and Womens' Hospital.

In order to determine whether variants were functional or defective, we recreated the changes corresponding to all the coding *SIAE* variants that we had discovered in patients and controls into a C-terminal FLAG-tagged human *SIAE* cDNA cloned from MDA-MB 231 cells. Each cDNA was transfected into 293T cells, and lysates and supernatants were each divided into two equal aliquots. SIAE was immunoprecipitated with anti-FLAG antibodies, and one aliquot was saved for a quantitative Western blot assay while the other was utilized in an esterase assay using a fluorogenic substrate, 4-methylumbelliferyl acetate. Quantitative Western blotting was performed using a near infrared-dye labeled second antibody and detected using the Li-Cor Odyssey system. Each cDNA was transfected three or more times and the entire assay performed on at least three occasions for each cDNA.

As described in Table I we have now identified 27/923 autoimmune subjects with either rare heterozygous non-synonymous substitutions in *SIAE* that do not represent known SNPs or a specific defective homozygous polymorphism. In 24 of these patients the *SIAE* variants were found to be functionally defective either because of a defect in catalytic activity to below 50% of wild type or because of a profound defect in secretion (in the absence of a catalytic defect). A group of missense variants that are severely catalytically defective include c.935C>T, c.587G>T, c.926A>C, c.634G>A, c.1435C>T, c.1178G>A, and c. 688C>T and encode T312M SIAE, C196F SIAE, Q309P SIAE, G212R SIAE, R479C SIAE, R393H SIAE, and R230W SIAE. The analysis of these severely catalytically defective variants by transfection, immunoprecipitation, enzyme assays and immunoblot assays are shown in Fig. 1 and Supplementary Fig. 3. These variants are also very poorly secreted presumably because they are grossly misfolded proteins that fail to egress the endoplasmic reticulum.

More modest, but reproducible catalytic defects were seen in the c.1046A>G variant that encodes Y349C SIAE, and this variant also exhibits reduced secretion (see Fig. 1, bottom panels). The c.1211T>C variant encodes F404S SIAE that also appears to exhibit a less severe catalytic defect (but nevertheless below the 50% cutoff set), similar to that seen in Y349C SIAE (Fig. 1). F404S SIAE was found in four autoimmune patients including two blood relatives, one with SLE and the other with juvenile idiopathic arthritis. The c.796T>G variant found in one subject with Type I diabetes encodes C266G SIAE that is also defective (Fig. 1).

In contrast to the catalytically defective *SIAE* variants seen in patients with autoimmune diseases, with the exception of two variants (R314H and T312M) observed once each in controls (Table 1), most of the new *SIAE* variants found in normal subjects did not exhibit reduced catalytic activity, as shown in Fig 2. Interestingly the protein encoded by the 89V polymorphic allele of *SIAE* is catalytically active but is not secreted (see Fig. 2, third set of panels). The 89V polymorphism is quite common in controls in the heterozygous state (9.7%, see Supplementary Table S1). In order to more precisely establish that M89V SIAE

is secretion defective, 293T cells transfected with wild type and *M89V SIAE* respectively were metabolically labeled with ^{35}[S] methionine and chased for 10 min, 1h, 2h and 4h. As seen in Fig. 3b, a striking defect in secretion of M89V SIAE was confirmed by this analysis.

Since SIAE exists as a dimer or higher order oligomer (Supplementary Fig. 2), we examined whether catalytically dead mutants from patients with autoimmunity on the one hand, and the catalytically active but secretion-defective M89V variant on the other, could function in a dominant interfering manner (Fig. 3a). The K400N allele (Fig. 1) was also tested as a representative catalytically normal SIAE allele. Since the ultimate test of dominant negative function would be to re-create a heterozygous animal with one mutant allele, mutations were recreated in a murine *Siae* cDNA for these studies. As seen in Fig. 3a and Supplementary Fig.1, the murine equivalents of the C196F, G212R, Q309P, T312M, Y349C, F404S and R479C variants are capable of dominantly inhibiting wild type SIAE while M89V SIAE and K400N SIAE are not. Based on this finding it was clear that only subjects with homozygous 89V/89V SIAE polymorphisms (as opposed to subjects with heterozygous *M89V* changes) should be considered to be of potential functional relevance for predisposition to autoimmunity.

Strikingly, eight autoimmune subjects (3 with RA, 1 with SLE, 1 with MS, and three with Type I diabetes) are homozygous for c.[265A>G]+[265A>G] polymorphic alleles, which encode 89V/89V SIAE variants, whereas these homozygous genotypes were not observed in a single control. Given the defect in secretion of this variant, we consider it likely that in subjects with homozygous 89V/89V SIAE this esterase is unlikely to be able to effectively access the post-Golgi compartment in which it would normally de-acetylate 9-*O*-acetylated sialoproteins that serve as CD22 ligands. Hardy-Weinberg equilibrium tests for the *M89V* polymorphism showed a deviation from equilibrium for the cases but not for the controls (Supplementary Tables S2a, 2b, and 2c). This deviation is statistically significant. Given the overall similarities in the *89V* and *89M* allele frequencies in cases and controls this clearly reflects an enrichment of *89V* homozygotes in autoimmune subjects, strongly supporting a role for this homozygous polymorphism in disease susceptibility.

A number of *SIAE* rare variants were found in patients with autoimmunity that are probably not involved in the genetic predisposition of these subjects to autoimmunity. For example, a c.98A>G variant encoding N33S SIAE was discovered in a patient with RA and was found to be functionally normal based on the criteria used (Fig. 1, top panels). One patient with Crohn's disease inherited a c.8C>G variant encoding an A3G change in the signal peptide encoding portion of *SIAE*. The coding region of *SIAE* would be predicted to be intact in this variant though it is theoretically possible that A3G SIAE might not be readily translocated into the ER. We consider it unlikely that A3G SIAE is translocation-defective given the accumulation of SIAE in culture supernatants when *A3G SIAE* is transfected into 293 T cells (Fig. 1, third set of panels). A c.1200G>T variant that encodes K400N SIAE was discovered in a patient with Crohn's disease initially examined as part of a small subset of patients with high ANA titers. This enzyme is active and is efficiently secreted but always appears in supernatants as a protein doublet (Fig. 1, second set of panels). Lysine 400 is immediately adjacent to a consensus N-glycosylation site, and it may be that a particular N-glycan is added inefficiently in this variant. This variant is however catalytically active and we classify it as a non-defective allele.

An absolute correlation was not found between conservation of amino acid residues of SIAE across species and a requirement for catalytic activity. Of the 11 heterozygous variants that were found to be defective in autoimmune patients only one (C266G) was not conserved between primates and rodents. One of 3 catalytically normal variants identified in autoimmune subjects (N33S SIAE) was also not conserved across species.

In the first phase of this study defective variants were identified in 7/188 autoimmune patients and 0/190 controls. The Odds Ratio could only be calculated as an estimate (Peto Odds Ratio), and this approach yielded an Odds Ratio of 7.71. In the second phase of the study 17/735 autoimmune patients and 2/458 controls inherited defective *SIAE* alleles, and the calculated Odds Ratio was 5.40. In summary, the total number of patients with autoimmune disorders analyzed was 923, with 24/923 inheriting defective *SIAE* alleles, and the total number of ethnically matched controls was 648, with 2/648 controls inheriting defective *SIAE* alleles. The calculated Odds Ratio for all autoimmune disorders was 8.62 with a two-sided p-value of 0.0002.

While care was taken to include only Caucasian non-Jewish subjects in this study, an objective determination of shared ethnicity, a principal components analysis 3‚4, was conducted on samples with defective *SIAE* alleles and on controls (see Supplementary Fig. 5). The novel *SIAE* variants that we have observed in subjects with autoimmunity cannot be ascribed to population stratification with respect to controls.

Seven of the 648 controls inherited a non-synonymous rare variant of *SIAE* but only 2/648 inherited defective alleles (Fig. 2 and Table I). One of the rare variants (c.935C>T encoding T312M SIAE) found in one of the controls (all 10 exons were sequenced in 648 controls) was identical to a defective variant originally found in a patient with RA and also in a patient with MS. Another variant, (c.941G>A encoding R314H SIAE), was found in a single control and was also found to be defective (Fig. 2). The remaining rare variants found in controls (c.1340A>G, c.481C>A, c.185G>A, c.1368G>A, and c.1385A>G encoding the H447R, Q161K, R62H, M456I and Q462R versions of SIAE respectively) were completely normal as determined by the two assays described (see Fig. 2 and Table I for a summary). A polymorphic variant c.190G>A, encoding G64S SIAE, exhibits normal catalytic activity and is readily secreted (Fig. 2 and Table I), was found in controls, but not found in autoimmune subjects (Table I and Supplementary Table S1). It might in theory convey protection against disease but a biochemical basis for such a possible protective role is unclear.

All the defective heterozygous *SIAE* alleles in patients that were tested function in a dominant interfering manner in the transfection assay employed. While high Odds Ratios were observed in a number of autoimmune disorders, the results from rheumatoid arthritis subjects (Odds Ratio, 8.3, 2-tailed p= 0.0056) and Type I diabetes patients (Odds Ratio, 7.9, 2-tailed p= 0.0075) were particularly significant. While a dominant negative effect may contribute to disease susceptibility for all tested defective heterozygous variants, we also consider haploinsufficiency to be a possible mechanism for some of the defective variants which might result in the reduction of levels of catalytically active enzyme in B cells below a threshold. This would imply that the W48X alteration found in a patient with type I diabetes (Table 1) may be clinically relevant. We plan to examine the effects of haploinsufficiency in mutant mice both on a C57B/6 background as well as in a lupus–prone background.

A number of susceptibility loci for human autoimmune disorders have been uncovered by genome wide association studies, and the relative risks for these associations are generally modest 5‚ 6. A number of recent reports have supported the hypothesis that rare genetic variants can contribute to disease susceptibility. A pioneering study on rare variants in genes that are relevant to lipoprotein synthesis in patients susceptible to cardiovascular disease utilized a predictive algorithm (Polyphen) to determine which variants were probably non-functional7. Loss-of-function variants in the *Trex* gene, which has a single coding exon, have also been described in patients with SLE8 and a recent re-sequencing study has revealed rare variants in the cytosolic helicase MDA5/IF1H1, which mediates innate immune responses to pathogen encoded RNAs 9. Our results provide important support for a

role for rare variation in the predisposition to autoimmune diseases and strikingly illustrate the importance of performing functional assays for the variants being studied. All of the variants identified through re-sequencing are listed in Supplementary Table S1. There is clear enrichment of defective coding variants in autoimmune patients as compared to controls (Table 2), and it is notable that these variants primarily involved residues that are highly conserved across evolution. While our initial studies strongly support a role for defective *SIAE* rare and polymorphic variants in RA and T1D, a role for these variants in other autoimmune disorders including SLE, MS and IBD appears likely (Supplementary Table S3). More extensive studies on these and other autoimmune disorders are called for.

The contribution of B cells to disease, with or without a role for autoantibodies, is recognized in a growing number of diseases including rheumatoid arthritis, multiple sclerosis, and Type1 Diabetes. Mutant *Siae* in rodents results in enhanced B cell activation and a break in B cell tolerance1, but it remains formally possible that SIAE may be required in cell types other than B cells in humans as well as in rodents. One type of inflammatory bowel disease, Crohn's disease, like multiple sclerosis, is generally considered to be etiologically linked to T_H1 or T_H17 cells10. Although B cell depletion can result in marked clinical improvement in patients with multiple sclerosis11, B cells are not generally considered to be of etiopathogenic significance in Crohn's disease; it remains formally possible that autoantigen specific B cells may function as critical antigen presenting cells that secrete cytokines driving helper T cell polarization in certain disease situations. Interestingly, analysis of B cells from a Crohn's disease patient harboring a catalytically defective heterozygous *SIAE* variant (G212R SIAE; see Table I) revealed a marked enhancement of cell surface 9-*O*-acetyl sialic acid following BCR activation compared to B cells from a control subject (Fig. 3c). Enhanced 9-*O*-acetylation of sialic acid on B cells was also noted in a patient with Sjogren's syndrome with a heterozygous C196F SIAE variant, and in a patient with lupus with a R393H SIAE variant (Table I; Supplementary Fig.6). These analyses suggests that a defect in SIAE results in enhanced BCR mediated expression of surface 9-*O*-acetylated sialic acid, while the presence of normal SIAE prevents this enhanced 9-*O*-acetylation event. This phenomenon is being further explored in a range of subjects with autoimmune disorders. Further analyses will be necessary to determine whether there is a role for B cells in disease pathogenesis in a subset of patients with IBD.

SIAE contributes to a signaling mechanism that helps set a threshold for B cell activation, presumably preventing weakly self-reactive B cells from moving towards the T cell zone and consequently being at risk for somatic mutation and for the potential generation of high affinity self-reactive B cells; alternatively SIAE may possibly help maintain tolerance in germinal centers1, 2. The strong association of defective *SIAE* alleles to rheumatoid arthritis and type I diabetes may well represent only the tip of the iceberg for a pathway that includes Lyn, SHP-1, a sialic acid acetyltransferase, SIAE, CD22 and likely other Siglecs expressed in B cells2. The possibility that SIAE may be of functional relevance in innate immune cells and thus influence disease pathogenesis also deserves exploration.

METHODS

Analysis of the sequence of *SIAE*

Each exon of *SIAE* was amplified from genomic DNA from individual subjects and subjected to automated sequencing. Residue numbering was based on ENST00000263593 (Ensembl), corresponding to the Genbank accession number NP_733746. Genomic DNA was extracted from clotted blood specimens from patients with autoimmune disease using a QIAamp DNA blood mini kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. All 10 exons of the human *SIAE* gene were amplified by PCR using intronic primers (Supplementary Table S4). Purification of the amplified products, bidirectional

automated sequencing and sequence analysis were performed as described earlier13. All sequence variants were confirmed by sequencing at least two independent PCR amplicons. DNAs from controls were obtained from EBV-immortalized lymphoblastoid cell-lines established from healthy blood donors14. All blood samples were collected with approval from the MGH/Partners Human Studies Institutional Review Board and from the Institutional Review Board at North Shore Long Island Jewish Health System.

Site directed mutagenesis of human *SIAE* and assays for SIAE catalytic activity and secretion

A human *SIAE* cDNA was cloned from MDA-MB 231 cells and a full length FLAG-tagged human *SIAE* expression construct (in pcDNA3.1) was generated. This clone was targeted for mutagenesis using PfuTurbo DNA polymerase (Stratagene, CA, USA). Each *SIAE* variant (other than the W48X truncation) was recreated by site directed mutagenesis as a C-terminal FLAG tagged human *SIAE* cDNA in an expression vector Site directed mutagenesis was used to create the S127A variant with a defect in the catalytic site as well as each of the variants listed in Table I. The PCR products were digested overnight with DpnI (10 Units; Stratagene) and transformed into TOP10 chemically competent cells (Invitrogen, Carlsbad, CA). Clones containing the mutants were verified by DNA sequencing. All mutant and wild type cDNAs were transfected into HEK 293T cells. Lysates and supernatants were immunoprecipitated with anti-FLAG antibodies and catalytic activity of the immunoprecipitated esterase was assayed by a fluorimetric method15. Equivalent amounts of each lysate and supernatant were immunoprecipitated for the catalytic assay as well as for quantitation of the FLAG-tagged protein by an immunoblot assay on the LI-COR Odyssey, using a mouse monoclonal anti-FLAG antibody (Sigma) and an IR Dye 800CW labeled Goat anti-mouse IgG (LI-COR) as a secondary antibody. Immunoprecipitation, metabolic labeling, and pulse chase studies were performed as described in ref.1.

Assays for determining the dominant negative function of specific SIAE variants

These assays were carried out by co-transfecting cDNAs encoding V5-tagged wild type murine Siae, together with FLAG-tagged murine versions of *SIAE* mutants discovered in subjects with rheumatoid arthritis into 293T cells. The V5-tagged wild type proteins in cell lysates were immunoprecipitated using mouse monoclonal anti-V5 antibody (Invitrogen), for quantitative immunoblot and esterase activity assays. Expression of FLAG-tagged mutants was also monitored by immunoprecipitation and Western blot assays.

Analysis of Cell surface 9-*O*-acetylation of sialic acid on human B lymphocytes

The method used was described in murine lymphocytes in reference1and is based on the method described originally by Krishna and Varki16. Briefly human B cells were stained with antibodies to CD19 and CD27 (BD Pharmingen) and incubated either with or without $F(ab')_2$, polyclonal rabbit anti-human IgM (Dako). Cells were also stained either with FITC-$F(ab')_2$, goat anti-human IgG, Fcγ specific (Jackson Immunoresearch) alone, or the CHE-FcD reagent (an Influenza C hemagglutin esterase-fused to the Fc portion of human IgG, chemically treated with diisopropyl fluorophosphate) complexed with FITC-$F(ab')_2$ fragment goat anti-human IgG, Fcγ specific. Cells were analyzed by flow cytometry.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Melissa Cohen for coordinating the recall of IBD patients, and Katie Pedrick and Emma Chung for their contributions. These studies were supported by grants from the Alliance for Lupus Research, the Center for the Study of Inflammatory Bowel Disease at MGH, and the NIH (AI 064930, AI 076505, and AR 058481) to SP and a grant (NS 32765) to MEM. PKG acknowledges NIH support for NARAC (AR 044422 and AR 022263) and MADGC (AI 068759). The NIDDK is acknowledged for making samples available from the EDIC collection of its DNA repository. DWB acknowledges the Intramural program of the National Human Genome Research Institute at NIH.

References

1. Cariappa A, et al. B cell antigen receptor signal strength and peripheral B cell development are regulated by a 9-O-acetyl sialic acid esterase. J Exp Med. 2009; 206:125–138. [PubMed: 19103880]
2. Pillai S, Cariappa A, Pirnie SP. Esterases and autoimmunity: the sialic acid acetylesterase pathway and the regulation of peripheral B cell tolerance. Trends Immunol. 2009; 30:488–493. [PubMed: 19766537]
3. Nassir R, et al. An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. BMC Genet. 2009; 10:39. [PubMed: 19630973]
4. Tian C, et al. European Population Genetic Substructure: Further Definition of Ancestry Informative Markers for Distinguishing Among Diverse European Ethnic Groups. Mol Med. 2009
5. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science. 2008; 322:881–888. [PubMed: 18988837]
6. Gregersen PK, Behrens TW. Genetics of autoimmune diseases--disorders of immune homeostasis. Nat Rev Genet. 2006; 7:917–928. [PubMed: 17139323]
7. Cohen JC, et al. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science. 2004; 305:869–872. [PubMed: 15297675]
8. Lee-Kirsch MA, et al. Mutations in the gene encoding the 3'–5' DNA exonuclease TREX1 are associated with systemic lupus erythematosus. Nat Genet. 2007; 39:1065–1067. [PubMed: 17660818]
9. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science. 2009; 324:387–389. [PubMed: 19264985]
10. Xavier RJ, Podolsky DK. Unravelling the pathogenesis of inflammatory bowel disease. Nature. 2007; 448:427–434. [PubMed: 17653185]
11. Hauser SL, et al. B-cell depletion with rituximab in relapsing-remitting multiple sclerosis. N Engl J Med. 2008; 358:676–688. [PubMed: 18272891]
12. Jessani N, et al. Class assignment of sequence-unrelated members of enzyme superfamilies by activity-based protein profiling. Angew Chem Int Ed Engl. 2005; 44:2400–2403. [PubMed: 15765498]
13. Kwak EL, et al. Epidermal growth factor receptor kinase domain mutations in esophageal and pancreatic adenocarcinomas. Clin Cancer Res. 2006; 12:4283–4287. [PubMed: 16857803]
14. Bell DW, et al. Common nonsense mutations in RAD52. Cancer Res. 1999; 59:3883–3888. [PubMed: 10463575]
15. Shukla AK, Schauer R. Fluorimetric determination of unsubstituted and 9(8)-O-acetylated sialic acids in erythrocyte membranes. Hoppe Seylers Z Physiol Chem. 1982; 363:255–262. [PubMed: 7076126]
16. Krishna M, Varki A. 9-O-Acetylation of sialomucins: a novel marker of murine CD4 T cells that is regulated during maturation and activation. J Exp Med. 1997; 185:1997–2013. [PubMed: 9166429]

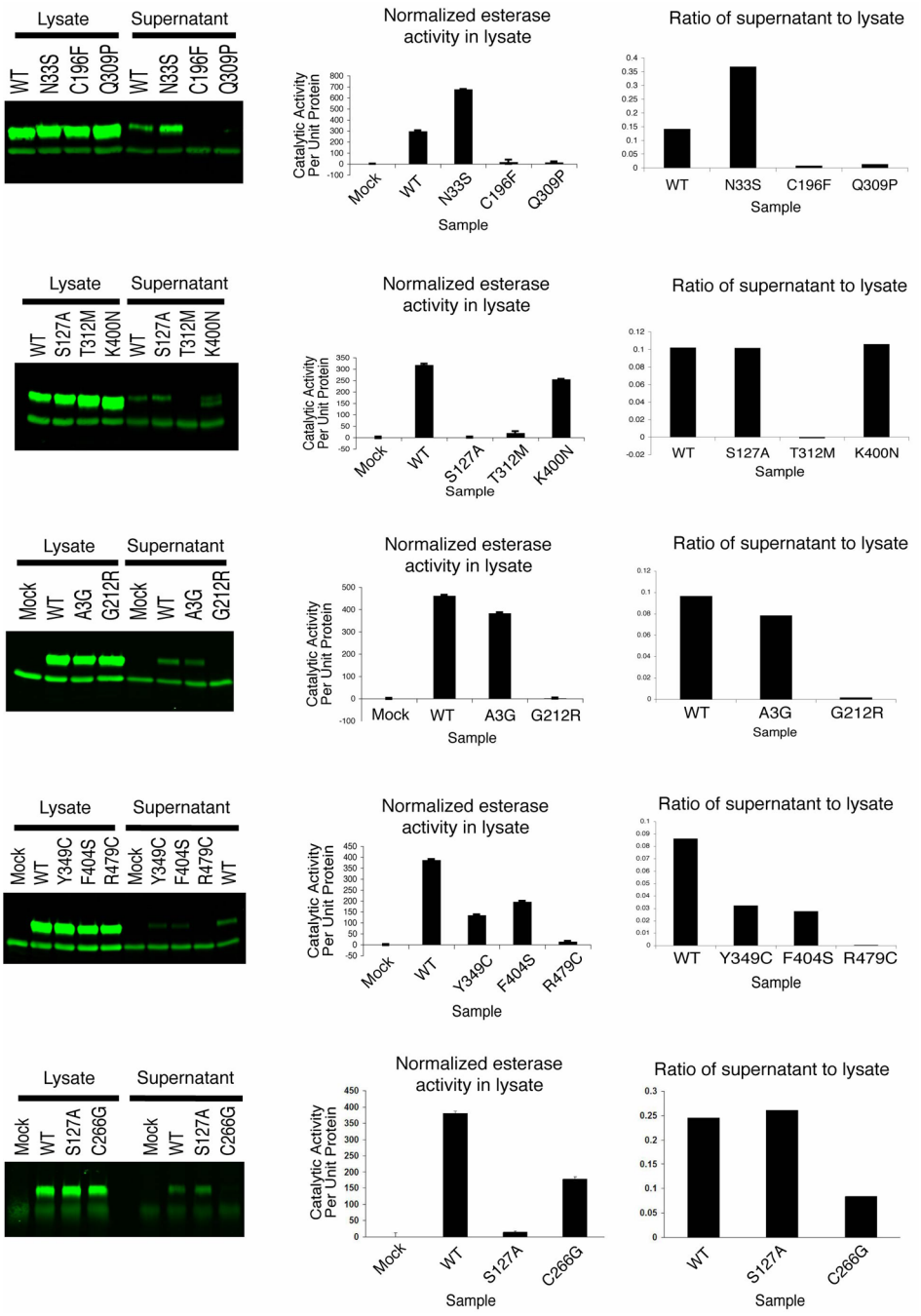


Figure 1. Analysis of *SIAE* variants from subjects with autoimmunity

Each *SIAE* variant found in subjects with autoimmunity was re-created by site-directed mutagenesis in a human *SIAE* cDNA, that was then sequenced along its entire length. Wild type (WT) *SIAE*, a known catalytic site mutant (*S127A SIAE*12), and each *SIAE* variant that was unique to autoimmune subjects were transfected into 293T cells. Assays were performed for *A3G SIAE, N33S SIAE, C196F SIAE, G2121R SIAE, C266G SIAE, Q309P SIAE, T312M SIAE, Y349C SIAE, K400N SIAE, F404S SIAE*, and *R479C SIAE*. Quantitative western blot analysis (using anti-FLAG antibodies) was performed on both the cell lysate and the culture supernatant, and a ratio of these two measurements is shown in the

right hand panels of the figure. “Mock” refers to cells which were not transfected but from which lysate and supernatant were analyzed.

Half of each lysate was immunoprecipitated with anti-FLAG antibodies and examined for esterase activity, presented following normalization for lysate SIAE protein content. Each row shows results from one representative transfection. Each variant was tested in this manner on at least three or more occasions to ensure reproducibility.

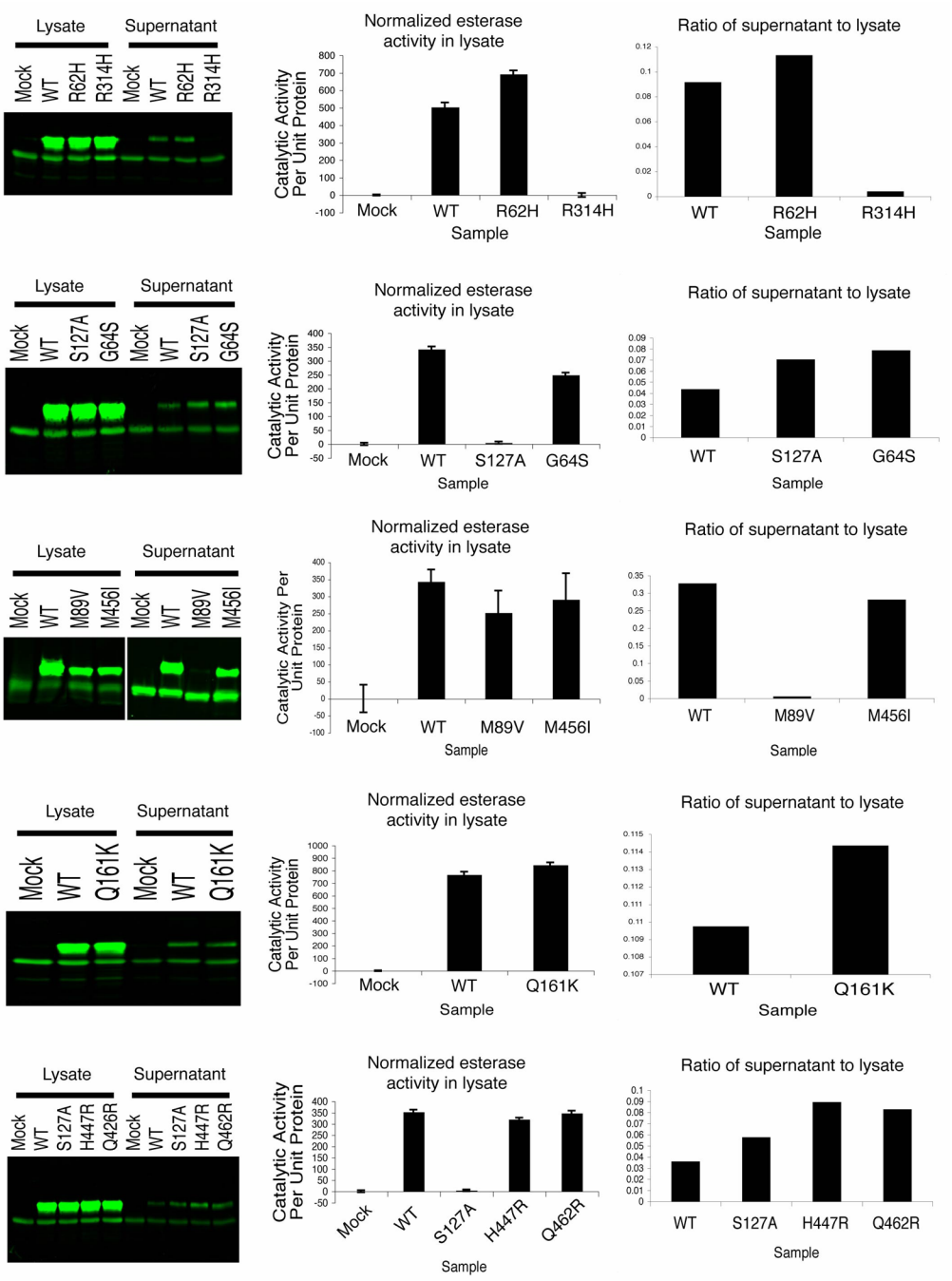


Figure 2. Analysis of *SIAE* variants from controls

Each variant identified in control subjects was recreated in an *SIAE* cDNA as described above for subjects with autoimmunity. Wild type (WT) *SIAE*, *S127A SIAE* and each *SIAE* variant that was unique to controls (*R62H SIAE, G64S SIAE, Q161K SIAE, R314H SIAE, H447R SIAE, M456I SIAE*, and *Q462R SIAE*) was transfected into 293T cells. Also shown are results from *M89V SIAE*, which was found in heterozygous form in both patients and controls and in homozygous form only in patients. *T312M SIAE* was observed in one control and in two patients. Results for this variant are included in Fig. 1. Analyses were performed as described in the legend for Fig. 1.

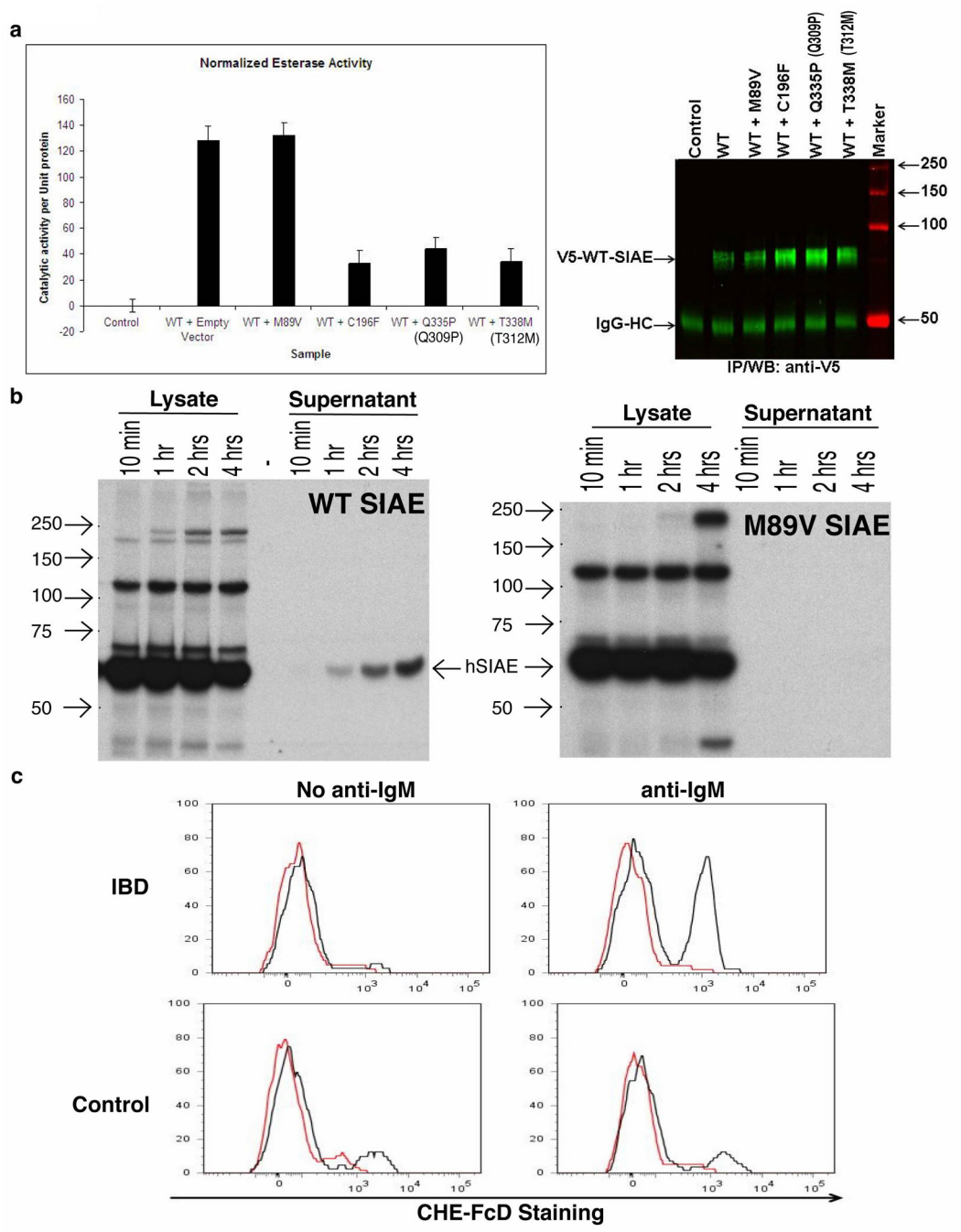


Figure 3. Analysis of SIAE mutants in terms of secretion, in vitro dominant interfering activity, and effect on induced cell surface 9-O-acetylation of sialic acid

a. Murine C196F Siae, and the murine equivalents of Q309P SIAE and T312M SIAE, (Q335P and T338M Siae), function in a dominant interfering fashion but M89V Siae does not. V5-tagged wild type Siae was transfected along with FLAG-tagged C196F Siae or FLAG-tagged M89V Siae and the enzyme activity of V5-tagged wild type Siae was assessed in transfectants as a function of its protein level. Expression of mutant Siae was monitored by an anti-FLAG Western blot of immunoprecipitated mutant proteins (see Supplementary Fig. 4).

b. Pulse-chase analysis comparing secretion of wild type SIAE and M89V SIAE. Transfected 293T cells were metabolically pulse-labeled with ^{35}[S] methionine and lysates and supernatants were immunoprecipitated with anti-FLAG antibodies after 10 minutes, 1 hour, 2 hours and 4 hours of chase. Proteins were separated by SDS-PAGE and revealed by autofluorography. The position of molecular weight markers is indicated on the left in kilodaltons.

c. Enhanced 9-*O*-acetylation of sialic acid following BCR ligation in B cells from a subject with a defective *SIAE* mutation. Naïve ($CD19^+CD27^-$) B cells from the peripheral blood of a subject with Crohn's disease (labeled IBD) with a heterozygous *SIAE* mutation (G212R) and from a control subject were analyzed for cell surface 9-*O*-acetylation with and without anti-IgM induced BCR ligation. Cell surface 9-*O*-acetylation was detected using CHE-FcD staining approach as described in Methods. The black tracing reflects CHE-FcD staining and the red represents staining with the second antibody alone.

Table I

SIAE variants identified in Caucasian autoimmune subjects and controls

SIAE Change	Esterase Activity	Secretion	Dom. Neg.	Disease	Source
Autoimmune Patients (n=923)					
T312M	Defective	Defective	Yes	RA	MGH
T312M	Defective	Defective	Yes	MS	BWH
Q309P	Defective	Defective	Yes	RA	NARAC
C196F	Defective	Defective	Yes	RA	NARAC
C196F	Defective	Defective	Yes	SjS	MGH
M89V/M89V	Normal	Defective	No	RA	NARAC
M89V/M89V	Normal	Defective	No	RA	NARAC
M89V/M89V	Normal	Defective	No	RA	NARAC
M89V/M89V	Normal	Defective	No	SLE	MADGC
M89V/M89V	Normal	Defective	No	MS	BWH
M89V/M89V	Normal	Defective	No	T1D	NIH
M89V/M89V	Normal	Defective	No	T1D	NIH
M89V/M89V	Normal	Defective	No	T1D	NIH
G212R	Defective	Defective	Yes	CD	MGH
F404S	Defective	Defective	Yes	JIA	MADGC
F404S	Defective	Defective	Yes	SLE	MADGC
F404S	Defective	Defective	Yes	UC	MGH
F404S	Defective	Defective	Yes	MS	BWH
Y349C	Defective	Reduced	Yes	SLE	MADGC
R479C	Defective	Defective	Yes	CD	MGH
W48X	Truncated/NT	Truncated/NT	NT	T1D	NIH
C266G	Defective	Defective	NT	T1D	NIH
R230W	Defective	Defective	NT	T1D	NIH
R393H	Defective	Defective	NT	SLE	MGH
K400N	Normal	Doublet	No	CD	MGH

SIAE Change	Esterase Activity	Secretion	Dom. Neg.	Disease	Source
A3G	Normal	Normal	NT	CD	MGH
N33S	Normal	Normal	NT	RA	NARAC
Ethnically Matched Controls (n=648)					
R314H	Defective	Defective	NT	Control	NS/LIJ
T312M	Defective	Defective	Yes	Control	NS/LIJ
Q161K	Normal	Normal	NT	Control	MGH
G64S	Normal	Normal	NT	Control	MGH
G64S	Normal	Normal	NT	Control	MGH
G64S	Normal	Normal	NT	Control	MGH
G64S	Normal	Normal	NT	Control	MGH
G64S	Normal	Normal	NT	Control	MGH
G64S	Normal	Normal	NT	Control	NS/LIJ
G64S	Normal	Normal	NT	Control	BWH
G64S	Normal	Normal	NT	Control	BWH
G64S	Normal	Normal	NT	Control	BWH
G64S	Normal	Normal	NT	Control	BWH
Q462R	Normal	Normal	NT	Control	MGH
H447R	Normal	Normal	NT	Control	MGH
R62H	Normal	Normal	NT	Control	NS/LIJ
M456I	Normal	Normal	NT	Control	NS/LIJ

Abbreviations: RA, rheumatoid arthritis; MS, multiple sclerosis; SLE, systemic lupus erythematosus; SjS, Sjogren's syndrome; JIA, juvenile idiopathic arthritis; T1D, Type1 diabetes; CD, Crohn's disease; UC, ulcerative colitis; MGH, Massachusetts General Hospital; BWH, Brigham and Womens' Hospital; NS/LIJ, North Shore Long Island Jewish; NIH, National Institutes of Health; NARAC, North American Rheumatoid Arthritis Consortium; MADGC, Multiple Autoimmune Disorders Genetics Consortium; NT, not tested.

Table II

Functionally defective SIAE coding variants in rheumatoid arthritis, Type I diabetes and all autoimmune diseases combined compared with controls *

Disease group	# of subjects	Odds Ratio (95% CI**)	Two-tailed p-value***
Rheumatoid Arthritis	234	8.31 (1.69 –40.87	0.0056
Type I diabetes	252	7.89 (1.58 – 39.30)	0.0075
All Autoimmune Disorders	923	8.62 (2.03 – 36.62)	0.0002

* Patients and controls (n=648) were of European ancestry; Jewish subjects were not included in these analyses.

** 95% CI= 95% Confidence Interval

*** 2 tailed p-value was determined using Fisher's exact test

Published in final edited form as:
Nat Cell Biol. 2010 June ; 12(6): 591–597. doi:10.1038/ncb2061.

Light-mediated activation reveals a key role for Rac in collective guidance of cell movement in vivo

Xiaobo Wang[1,*], **Li He**[1,*], **Yi I. Wu**[2], **Klaus M. Hahn**[2], and **Denise J. Montell**[1,*]

[1]Department of Biological Chemistry Center for Cell Dynamics Johns Hopkins School of Medicine 855 North Wolfe Street Baltimore, MD 21205 USA

[2]Department of Pharmacology Lineberger Comprehensive Cancer Center University of North Carolina, Chapel Hill Chapel Hill, North Carolina 27599 USA

The small GTPase Rac induces actin polymerization, membrane ruffling, and focal contact formation in cultured single cells1, but can either repress or stimulate motility in epithelial cells depending on the conidtions2, 3. Therefore the role of Rac in collective epithelial cell movements in vivo, which are important for both morphogenesis and metastasis4-7, is difficult to predict. Recently photoactivatable analogs of Rac (PA-Rac) have been developed, allowing rapid and reversible activation or inactivation of Rac using light8. In cultured single cells, light-activated Rac leads to focal membrane ruffling, protrusion, and migration. Here we show that focal activation of Rac is also sufficient to polarize an entire group of cells in vivo, specifically the border cells of the Drosophila ovary. Moreover activation, or inactivation, of Rac in one cell of the cluster caused a dramatic response in the other cells, suggesting that the cells sense direction as a group based on relative levels of Rac activity. Communication between cells of the cluster required Jun N-terminal kinase (JNK) but not guidance receptor signaling. These studies further show that photoactivatable proteins are effective tools in vivo.

Border cells are a group of 6-8 cells that arise from the monolayer of ~650 epithelial follicle cells that surround 15 nurse cells and one oocyte in a structure called an egg chamber (Figure 1a-c). Border cells migrate ~175 μm in between the nurse cells, as an interconnected group of two distinct cell types: 4-8 migratory cells surround two central polar cells (Figure 1d-i, k). Polar cells cannot migrate but secrete a cytokine that activates the JAK/STAT pathway rendering the outer cells motile9. The outer cells carry the polar cells and lose the ability to move in the absence of continuous JAK/STAT activation10. Thus each cell type requires the other. Border cells also require steroid hormone, receptor tyrosine kinase, Notch, and other signaling cascades11-16. Thus border cells experience a rich and complex signaling environment, as do most cells in vivo.

The requirement for Rac in border cell migration was one of the earliest demonstrations of its role in cell motility in vivo17. Expression of either dominant-negative or constitutively active Rac impedes migration13, 17, 18, suggesting that its activity must be spatially and/or temporally controlled. However the precise function of Rac remains unclear.

[3]To whom correspondence should be addressed dmontell@jhmi.edu .
[*]These authors contributed equally

Author contributions X. W. carried out the experiments documented in Figures 1-3 and S1-9. Li He carried out the experiments shown in Figures 4-5 and S10-12. X.W. made all of the transgenic flies and helped Li to collect FRET data in Figure 5. Y.W. developed and provided the PA-Rac constructs and advised X. W. on their use. K.M.H. and D.J.M. coordinated the study. D.J.M. prepared the final version of the manuscript, based on contributions from all authors.

To evaluate the effect of locally activating Rac in border cells we generated transgenic flies expressing the photoactivatable form of Rac (PA-RacQ61L) tagged with mCherry, under control of the Gal4/UAS system. When expressed in border cells using *slbo*-Gal4, the protein was distributed throughout the cells, in the cytoplasm, nuclei, and at cell surfaces (Figure 1g-i). In the absence of laser illumination, border cell migration was normal (Figure S1; Movies S1 and S2).

Upon exposure to repeated pulses of laser light, border cell migration could be redirected (Figure 1l-t; Movie S3). In this example, border cells were migrating along the path designated by the solid arrow and the leading cell extended a prominent forward-directed protrusion. The laser was applied to the cell next to the leading cell, which did not exhibit any detectable protrusion at the time. Following illumination the cluster retracted the original forward protrusion, changed direction and began moving to the side, a behavior never observed in wild-type19, 20. Light pulses were delivered once per minute due to the reversibility of PA-Rac8. The border cells reached the side of the egg chamber after ~60 minutes (Figure 1m,n and Movie S3). Although light pulses were continuously delivered, the cluster did not move further down the side of the egg chamber over the next 20 minutes (Figure 1o-q), suggesting there might be a barrier or repellent in this region. When we shifted the site of illumination toward the center of the egg chamber (Figure 1r), the cells responded by moving in that direction (Figure 1s,t; Movie S3). A single amino acid substitution in the LOV domain (C450M) renders the protein light-insensitive8 and this construct could not redirect border cell migration even in the presence of light (Figure 1u-w; Movie S4).

To determine if Rac activity was only required in the lead cell, we co-expressed dominant-negative Rac (RacT17N) together with PA-RacQ61L in all border cells and photoactivated Rac in one cell. RacT17N alone strongly inhibits border cell motility18 and photoactivation of Rac in the front cell failed to promote forward movement of the cluster in this background (Figure 1g-i). However, activating Rac in approximately half of the cells in the cluster caused them to move forward, albeit very slowly (Figure 1j-m). These results suggest that each cell requires some Rac activity for motility, and each cell contributes to the migration speed of the cluster, but the highest level of Rac activation determines the direction of movement.

We then tested whether PA-RacQ61L was sufficient to cause border cells to move in a direction opposite to their normal movement (Figure 2). Border cells expressing PA-RacQ61L were first driven in the normal direction, and prominent lamellipodia-like protrusion was evident at the site of illumination (Movie S5). Then, we illuminated the rear. Front protrusion ceased rapidly (Movie S5) but rearward movement was initially very slow. After a variable delay, clusters moved backwards (Figure 2d-i), sometimes reconnecting with a follicle cell within the epithelium (Movie S5). In contrast, the light-insensitive control protein did not reverse the migration direction (Movie S6). On average, the PA-Rac-induced forward migration speed exceeded the reverse migration speed by 4.5-fold (Figure 2p; PA-Rac Front vs. PA-Rac Back), suggesting an influence of endogenous directional signaling on the behavior induced by PA-Rac.

To explore the interaction between endogenous signals and PA-Rac, we compared the responses of wild-type cells to those of cells with reduced guidance receptor activity. PVR and EGFR are receptor tyrosine kinases that function redundantly to guide migrating border cells13-15. Border cells expressing dominant-negative forms of both guidance receptors, PVR^{DN} and $EGFR^{DN}$, extend protrusions in all directions and make little forward progress20. PA-Rac rescued both the morphological defect and directional movement in this genotype (Figure 2j-l), consistent with the idea that Rac normally functions downstream of

the receptors to determine the direction of movement. When clusters were illuminated at the front, the cells moved forward (Figure 2j-l). When the same clusters were illuminated at the back, rearward movement (Figure 2m-o) resulted. In contrast to the responses of wild-type clusters, average forward and reverse migration speeds were indistinguishable in border cells expressing PVR^{DN} and $EGFR^{DN}$ (Figure 2p), supporting the idea of competition between endogenous guidance receptor signaling and PA-Rac induced directionality.

After stimulating rearward protrusion, we stopped illuminating and observed the recovery (Figure S2). Both wild-type and PVR^{DN}- and $EGFR^{DN}$-expressing clusters rapidly protruded in response to rear illumination and retracted the rearward protrusion following cessation of the light. However, wild-type cells protruded less and retracted more (Figure S2q). Over longer time courses, wild-type cells typically stalled after cessation of rear illumination but eventually recovered movement in the normal forward direction (Figure S3a-m). In contrast, PVR^{DN} and $EGFR^{DN}$-expressing clusters failed to recover forward movement (Figure S3m-y). These results also suggest that endogenous PVR and EGFR signals compete with PA-RacQ61L-induced polarization.

The inability of PA-RacQ61L to cause border cells to move down the side of the egg chamber led us to probe the microenvironment further. Within the anterior ~1/3 of their normal travel path, focal Rac activation could steer border cells from the center path all the way to the follicle cells, or along the perimeter of the egg chamber (Figure 3a-c and m). However if we treated cells after they reached the center of the egg chamber, they could not be redirected to the follicle cell layer (Figure 3d-i and m), although PA-RacQ61L could still move them forwards or backwards. Within the posterior 1/3 of their normal path, the cells could again be directed off their normal course, in between the nurse cells (Figure 3j-l and m). A summary of the responses to PA-RacQ61L is shown in Figure 3m. Thus there are regions in the egg chamber that actively repel the border cells or lack important structural or chemical substrates for migration, suggesting that there is additional guidance information besides the ligands for PVR and EGFR.

PA-RacQ61L was also insufficient to cause border cells to migrate earlier than normal, possibly because high levels of JAK/STAT signaling, which are required for the border cells to initiate movement, are not achieved at earlier time points[21, 22]. Consistent with this, PA-RacQ61L did not cause protrusion or migration in border cells expressing a dominant-negative form of the receptor Domeless, which is required for STAT activation (Figure S4a-f). A key downstream target of STAT is the transcription factor Slow Border Cells (SLBO)[23], and *slbo* mutant border cells cannot extend protrusions. However PA-RacQ61L could not rescue protrusion or migration in *slbo* mutants (Figure S4g-l). PA-RacQ61L also failed to rescue guidance receptor deficiency after stage 10 (not shown). Together these findings demonstrate that PA-RacQ61L reveals temporal as well as spatial constraints on migrating cells.

To evaluate the effects of locally inhibiting Rac, we generated transgenic flies expressing PA dominant-negative Rac (UAS-PA-RacT17N). Illuminating the leading border cell arrested migration and, strikingly, led to protrusion at the cluster rear (Figure S5a-j; Movie S7). In contrast illumination of the rear of the cluster enhanced forward protrusion (Figure S5k-m) and migration (Figure 2p). The magnitude of the effect was smaller in PVR^{DN} and $EGFR^{DN}$-expressing cells (Figure S5n-p; Figure 2p).

The non-autonomous effects of PA-RacQ61L and PA-RacT17N were striking so we examined the morphological consequences at higher magnification. Specifically, activation of Rac in one cell of either a wild-type cluster (Figure 4a-c) or a cluster expressing PVR^{DN} and $EGFR^{DN}$ (Figure 4d-f) resulted in retraction of protrusions by the other cells and

movement of the cluster in the direction of the light. This was true whether the illumination was provided at the front of the cluster (not shown) or at the back. Strikingly PA-RacT17N had precisely the opposite effect in a polarized wild-type cluster (Figure 4j-l). Focal inhibition of Rac in the protruding lead cell caused a loss of polarization and random protrusion of all the cells in the cluster (Figure 4l).

To quantify these results, we developed an automated method to count the number of protruding cells (Figure S6) and calculated the directionality index, which measures the degree of polarization of the cell cluster20. PA-RacQ61L treatment rescued the PVR^{DN},$EGFR^{DN}$ polarization and the number of protruding cells nearly to wild-type (Figure 4m,n).

Inhibition of the JNK pathway also reduces the directionality index24 (Figure 4n). The JNK pathway helps to coordinate border cell movement by promoting cohesion between border cells. To test the hypothesis that cluster cohesion is important for the non-autonomous effects of Rac, we monitored the effect of PA-RacQ61L in cells with reduced JNK signaling. Photoactivation of Rac at the back of clusters with impaired JNK signaling did not cause retraction of protrusions that were extended in other directions and resulted in little net movement of the cluster (Figure 4g-i). The same effect was observed whether JNK signaling was reduced by expression of the JNK phosphatase Puckered (UAS-Puc2A) or by expressing a dominant-negative form of the kinase (not shown).

The inability of PA-RacQ61L to rescue the JNK knockdown phenotype could have been because JNK signaling is required autonomously downstream of Rac to generate lamellipodial protrusion. However PA-RacQ61L induced autonomous cell protrusion in the direction of illumination, even in cells over-expressing puc2A (Figure 4g-i). Therefore JNK signaling is not required downstream of Rac to promote protrusion, consistent with the published observation that reduction of JNK signaling does not lead to reduced protrusion24. Together these results suggest that JNK signaling is required for the non-autonomous propagation of directional information from the cell with highest Rac activity to the other cells of the cluster. This could be due to direct mechanical coupling of the cells or via signaling pathways downstream of adhesion receptors or both.

Our results suggested that Rac is normally active in all the cells of the cluster, that the leading cell has a higher level of Rac activity, and this asymmetry is lost in PVR^{DN}- and $EGFR^{DN}$-expressing cells. To test this we took advantage of a Rac fluorescence resonance energy transfer (FRET) biosensor25. When expressed in Drosophila S2 cells, biosensor activity increased in response to EGF stimulation, and the increase was blocked by co-expression of dominant-negative Rac (Figure S7a-k). We generated transgenic flies expressing the biosensor under the control of Gal4/UAS. When expressed with *slbo*-Gal4, we consistently observed a FRET signal in border cells (Figure S7l,m), and this was dramatically reduced upon co-expression of dominant-negative Rac (Figure S7n,o). Moreover the signal within the border cell cluster was asymmetric and appeared highest in elongating protrusions, which were most prominent in the leading cell (Figure 5a-e). This FRET signal was inhibited by co-expression of RacT17N (Figure 5l). To quantify the asymmetry we divided the border cell cluster into 30 sectors (where sector 0 represents the front of the cluster and −15 and +15 represent the rearmost sector), and measured the FRET efficiency in each sector for more than 30 clusters (Figure 5f-g). As predicted, the Rac activity was highest at the front (between sectors −5 and +5) and lowest at the back (Figure 5h, i, m). We then measured the Rac activity in more than 30 border cell clusters expressing PVR^{DN} and $EGFR^{DN}$ and found no difference between front and back (Figure j,k,m), consistent with the proposal that asymmetric Rac activation requires guidance receptor

input. In the absence of such asymmetry, non-directional signals activate Rac uniformly, stimulating random protrusion.

During normal morphogenesis and in tumor metastasis, many cells move in interconnected groups in a process termed collective cell migration4-7. Border cells represent one model for the study of such movements. We previously found that guidance receptor signaling not only promotes border cell protrusion at the front of the cluster but also polarizes the group so as to inhibit protrusion at the rear20. However it was unclear to what extent each cell sensed direction independently or whether they did so collectively and what intracellular signal(s) downstream of the receptors would be sufficient to polarize the group26. The results presented here demonstrate that a local increase in Rac activity is sufficient not only to stimulate protrusion autonomously in the treated cell but also to cause retraction of side and back cells, resulting in net cluster polarization and movement in the direction of highest Rac activity. Conversely inhibition of Rac in the lead cell caused the other cells to protrude in all directions as if guidance receptor activity were lost. These results suggest that elevated guidance activity at the front of the cluster activates Rac to a higher level in the front cell and this is sufficient to set the direction of migration for the whole group. Despite the fact that receptor tyrosine kinases activate myriad downstream signaling pathways, other pathways do not appear to be necessary, though they may play redundant or overlapping roles. Thus, asymmetric Rac activity is key for direction-sensing in vivo. We also show that JNK signaling is required to transmit the guidance signal between cells of the cluster. This work further suggests that photoactivatable proteins are likely to be a powerful new class of tools for the manipulation of protein activities with fine spatial and temporal control to address a variety of biological questions in animals.

Methods Summary

Drosophila strains

New transgenic fly lines were generated by Bestgene Inc. N-terminal-cherry tagged PA-RacQ61L, PA-RacT17N, the light insensitive control C450M-PA-RacQ61L8 and the Rac FRET probe were inserted into pUASt *Drosophila* expression vector using the Gateway recombination system (Invitrogen). *P[slbo-GAL4]*27 drives UAS transgene expression in outer, migratory border cells but not polar cells even though the endogenous *slbo* gene and protein product are expressed in both cell types23. *P[UAS-MCD8-GFP]*28, *P[UAS-moesin-GFP]*29, *P[UAS-DRacT17N]* and *P[UAS-DRacV12]*30 have been described previously. *P[UAS-PVR^{DN}]* and *P[UAS-EGFR^{DN}]* were obtained from P. Rørth13. *P[UAS-Puc2A]* and *P[UAS-DnBsk]* were obtained from E. Martin-Blanco24. All stocks were maintained at room temperature. Before dissection, flies were maintained at 29°C overnight to increase trangene expression levels. This incubation had no negative effect on border cell migration.

Imaging and photomanipulation

Drosophila egg chambers were dissected and mounted in Schneider's insect medium supplemented with 20% FBS and 0.10 mg/ml insulin as described19, 20. Photoactivation, time-lapse-imaging, and 3D morphological reconstruction were carried out using a Zeiss 510-Meta confocal microscope using a 63X, 1.4 numerical aperture lens with 2X zoom. To photoactivate, the 458 nm laser was set at 10% power for 0.1 ms per pixel in a 7μm spot and the photoactivation scan took approximately 25 seconds. After 30 seconds, border cells were imaged using 568nm. This series of steps was repeated for the duration of the timelapse experiment. Where indicated, 15-20 Z planes separated by 1.5μm were obtained before and after photoactivation (samples were illuminated every 80 seconds for one hour). 3D reconstructions were rendered using Imaris software.

S2 cells were transfected with the Rac FRET vector with or without the Rac^{DN} vector using the QIAGEN Effectine Kit. Cells were transferred to serum-free medium 48 hrs after transfection and cultured for another 6 hrs. Then the cells were transferred into 4-well Lab-Tek Chamber Slide for 1hr before imaging. A final concentration of 150ng/ml EGF was added to induce Rac activity. Rac FRET probe was kindly provided by Dr. Erez Raz. FRET experiment in S2 cells were carried out on Olympus IX81 microscope using 40X, 1.3 numerical aperture oil immersion objective. CFP and YFP signals were recorded using Chroma 86002BS dichroic mirror sets: CFP (excitation, 436/10nm; emission, 470/30nm), YFP(excitation, 436/10nm; emission, 535/30nm). A 25% neutral density filter was used to reduce bleaching.

FRET images of live cultured egg chambers were acquired with Zeiss LSM710 microscope. 458nm laser was used to excite the sample. CFP and YFP emission signals were collected through Channel I (470–510 nm) and Channel II (525-600 nm) respectively. To capture single, high-resolution, stationary images, 40X/1.1 water immersion objective was used. CFP and YFP images were acquired simultaneously for most of the experiments. Sequential acquisition of CFP and YFP channels with alternative orders were tested and gave the same result as simultaneous acquisition. CFP and YFP images were first processed by ImageJ software. A background ROI was subtracted from the original image. The YFP images were registered to CFP images by TurboReg pulgin. Gaussian smooth filter was then applied to both channels. The YFP image was thresholded and converted to binary mask with background set to zero. Final ratio image was generated by MATLAB program, during which only the unmasked pixel was calculated and all YFP/CFP ratios were adjusted to the initial FRET ratio to reduce the effect of bleaching. FRET images were analyzed using MATLAB. Border cell cluster was first isolated with its center calculated basing on its contour. Then the cluster was divided into 30 sectors, each of which occupies a 12-degree central angle. Because the center of the cluster contains the polar cells which do not express slbo-Gal4 and therefore were devoid of signal, only the signal within the distal 1/3 of each sector from the center was calculated. Average signal of each sector become a vector of length 31. The first and last element corresponding to the −15 and 15 sectors were the same, so the front of border cell was centered at zero. A heatmap was composed by 30 vectors from different egg chambers with the same genotype. All vectors for each genotype were further averaged and smoothed to generate a representative curve of the FRET distribution around the cluster.

Measurement of migration speed, protrusion number, directionality index and protrusion density.

The distance of the center of the border cell cluster between the first and last time points in a time lapse series was measured in Imaris software. This distance divided by the elapsed time gave the speed. Cell protrusions were counted as follows: a circle corresponding to the average cluster diameter was drawn and any extension more than 2 μm beyond that was considered a protrusion. The directionality index (DI) was calculated using the following equation:

$$DI = \left(\sum_{i=1}^{N} \vec{p}_i \cdot \vec{d} \right) / \sum_{i=1}^{N} \left\| \vec{p}_i \right\|$$

where N is the total number of major protrusions, $\vec{p}_i$ is the *i*th protrusion vector, and $\vec{d}$ is the unit vector of migration direction. Protrusion vector is calculated by fitting the major protrusion by a parabola whose peak together with the cluster center gives the vector's

direction and length. Protrusion density was generated by dividing the number of all the recognizable membrane protrusions by the estimated cell perimeter in micrometers. The morphology analysis and quantification were done in MATLAB.

Immunohistochemistry

Drosophila ovaries were dissected and fixed as described previously31 and incubated with 1.4 units Alexa 488-conjugated phalloidin (Molecular Probes) per ml and 1 μg/ml DAPI prior to imaging on a Zeiss 510-Meta confocal microscope and 3D reconstruction using Imaris software.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by GM046425 to D.J.M and GM 057464 to K.M.H. and by the Cell Migration Consortium.

References

1. Ridley AJ, Paterson HF, Johnston CL, Diekmann D, Hall A. The small GTP-binding protein Rac regulates growth factor-induced membrane ruffling. Cell. 1992; 70:401–410. [PubMed: 1643658]
2. Sander EE, Collard JG. Rho-like GTPases: their role in epithelial cell-cell adhesion and invasion. Eur J Cancer. 1999; 35:1302–8. [PubMed: 10658518]
3. Fukata M, Kaibuchi K. Rho-family GTPases in cadherin-mediated cell-cell adhesion. Nat Rev Mol Cell Biol. 2001; 2:887–97. [PubMed: 11733768]
4. Friedl P, Gilmour D. Collective cell migration in morphogenesis, regeneration and cancer. Nat Rev Mol Cell Biol. 2009; 10:445–57. [PubMed: 19546857]
5. Weijer CJ. Collective cell migration in development. J Cell Sci. 2009; 122:3215–23. [PubMed: 19726631]
6. Rorth P. Collective cell migration. Annu Rev Cell Dev Biol. 2009; 25:407–29. [PubMed: 19575657]
7. Bidard FC, Pierga JY, Vincent-Salomon A, Poupon MF. A “class action” against the microenvironment: do cancer cells cooperate in metastasis? Cancer Metastasis Rev. 2008; 27:5–10. [PubMed: 18066649]
8. Wu YI, et al. A genetically encoded photoactivatable Rac controls the motility of living cells. Nature. 2009; 461:104–8. [PubMed: 19693014]
9. Silver DL, Montell DJ. Paracrine signaling through the JAK/STAT pathway activates invasive behavior of ovarian epithelial cells in Drosophila. Cell. 2001; 107:831–41. [PubMed: 11779460]
10. Silver DL, Geisbrecht ER, Montell DJ. Requirement for JAK/STAT signaling throughout border cell migration in Drosophila. Development. 2005; 132:3483–92. [PubMed: 16000386]
11. Bai J, Uehara Y, Montell DJ. Regulation of Invasive Cell Behavior by Taiman, a Drosophila Protein Related to AIB1, a Steroid Receptor Coactivator Amplified in Breast Cancer. Cell. 2000; 103:1047–1058. [PubMed: 11163181]
12. Duchek P, Rorth P. Guidance of cell migration by EGF receptor signaling during Drosophila oogenesis. Science. 2001; 291:131–3. [PubMed: 11141565]
13. Duchek P, Somogyi K, Jekely G, Beccari S, Rorth P. Guidance of cell migration by the drosophila pdgf/vegf receptor. Cell. 2001; 107:17–26. [PubMed: 11595182]
14. McDonald JA, Pinheiro EM, Kadlec L, Schupbach T, Montell DJ. Multiple EGFR ligands participate in guiding migrating border cells. Dev Biol. 2006; 296:94–103. [PubMed: 16712835]
15. McDonald JA, Pinheiro EM, Montell DJ. PVF1, a PDGF/VEGF homolog, is sufficient to guide border cells and interacts genetically with Taiman. Development. 2003; 130:3469–78. [PubMed: 12810594]

16. Wang X, Adam JC, Montell D. Spatially localized Kuzbanian required for specific activation of Notch during border cell migration. Dev Biol. 2007; 301:532–40. [PubMed: 17010965]
17. Murphy AM, Montell DJ. Cell Type-specific Roles for Cdc42, Rac, and RhoL in *Drosophila* Oogenesis. JOURNAL OF CELL BIOLOGY. 1996; 133:617–630. [PubMed: 8636236]
18. Geisbrecht ER, Montell DJ. A role for Drosophila IAP1-mediated caspase inhibition in Rac-dependent cell migration. Cell. 2004; 118:111–25. [PubMed: 15242648]
19. Prasad M, Jang ACC, Montell D. A Protocol for culturing Drosophila melanogaster egg chambers for live imaging. Nature Protocols in preparation. 2007
20. Prasad M, Montell DJ. Cellular and Molecular Mechanisms of Border Cell Migration Analyzed Using Time-lapse Live-cell Imaging. Developmental Cell. 2007; 12:997–1005. [PubMed: 17543870]
21. Montell DJ. Border-cell migration: the race is on. Nat Rev Mol Cell Biol. 2003; 4:13–24. [PubMed: 12511865]
22. Jang AC, Chang YC, Bai J, Montell D. Border-cell migration requires integration of spatial and temporal signals by the BTB protein Abrupt. Nat Cell Biol. 2009; 11:569–79. [PubMed: 19350016]
23. Montell DJ, Rorth P, Spradling AC. slow border cells, a locus required for a developmentally regulated cell migration during oogenesis, encodes Drosophila C/EBP. Cell. 1992; 71:51–62. [PubMed: 1394432]
24. Llense F, Martin-Blanco E. JNK signaling controls border cell cluster integrity and collective cell migration. Curr Biol. 2008; 18:538–44. [PubMed: 18394890]
25. Kardash E, et al. A role for Rho GTPases and cell-cell adhesion in single-cell motility in vivo. Nat Cell Biol. 12:47–53. sup pp 1-11. [PubMed: 20010816]
26. Rorth P. Collective guidance of collective cell migration. Trends Cell Biol. 2007; 17:575–9. [PubMed: 17996447]
27. Rorth P, et al. Systematic gain-of-function genetics in Drosophila. Development. 1998; 125:1049–57. [PubMed: 9463351]
28. Lee T, Luo L. Mosaic analysis with a repressible cell marker for studies of gene function in neuronal morphogenesis. Neuron. 1999; 22:451–61. [PubMed: 10197526]
29. Edwards KA, Demsky M, Montague RA, Weymouth N, Kiehart DP. GFP-moesin illuminates actin cytoskeleton dynamics in living tissue and demonstrates cell shape changes during morphogenesis in Drosophila. Dev Biol. 1997; 191:103–17. [PubMed: 9356175]
30. Luo L, Liao YJ, Jan LY, Jan YN. Distinct morphogenetic functions of similar small GTPases: *Drosophila* Drac1 is involved in axonal outgrowth and myoblast fusion. Genes and Development. 1994; 8:1787–1802. [PubMed: 7958857]
31. McDonald JA, Montell DJ. Analysis of cell migration using Drosophila as a model system. Methods Mol Biol. 2005; 294:175–202. [PubMed: 15576913]

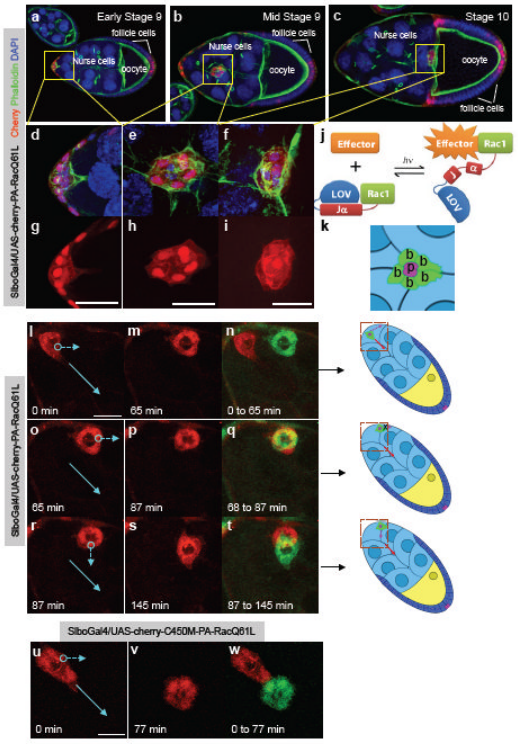


Figure 1.

Local activation of PA-Rac1 redirects an entire border cell group.

(**a-c**) Egg chambers labelled with DAPI (blue) to stain all nuclei, Alexa 488-phalloidin (green) to mark actin filaments, and mCherry (red) to show PA-RacQ61L. (**d-f**), Higher magnification views of border cells from each stage. (**g-i)** PA-RacQ61L expression only. (**j**) Schematic diagram from8 showing the mechanism of PA-Rac light-activation. (**k**) Schematic of border cell cluster composed of two non-migratory polar cells (purple, p) which do not express *slbo*-Gal4 and are therefore unlabeled in all subsequent images. Polar cells are surrounded by 4-6 migratory border cells (green, b). (**l-t**) Selected still images from a time-lapse film of the response of border cells to photoactivation of PA-RacQ61L. (**l-n**) Photoactivation diverts border cells to the edge of the egg chamber. (**o-q**) Continued photoactivation in same direction did not move them further along the edge. (**r-t**) Photoactivation of the same cluster in a different position drove movement towards the egg chamber center. In **n**, **q** and **t** the starting position of the cluster is shown in red and the final position in green. Schematics at right show the position of the treated cluster within the egg chamber. Red boxes indicate the regions shown in the micrographs. Red arrow indicates the normal direction of migration. Pink arrow shows the direction the cells move if they respond to the light. (**u-w**) Phototreatment of light insensitive control C450M-PA-Rac1Q61L. In **l**, **n**, **q** and **u**, solid arrows indicate the normal direction of migration; circles indicate where the laser light was applied. Dashed arrows indicate the direction the cells move if they respond to the light. Scale bars, 20 μm.

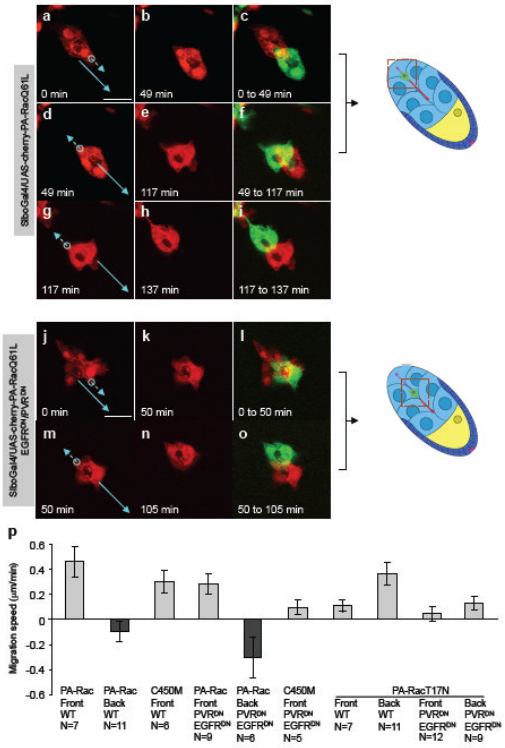


Figure 2.

Forward or backward movement in response to photoactivatable Rac.

(**a-i**) In an otherwise wild-type background, PA-RacQ61L can promote forward (**a-c**) or backward (**d-i**) movement. (**j-l**) Forward and (**m-o**) reverse migration of border cells expressing PVR^{DN}, $EGFR^{DN}$, and PA-RacQ61L. The schematics at the right show the position of the cluster within the egg chamber. Scale bars, 20μ. In panels with two colors, red represents the starting position and green shows the ending position over the indicated time period. **p**) Average migration speeds for clusters expressing the indicated proteins in response to illumination of the front or the back of the cluster. PA-Rac refers to PA-RacQ61L. C450M is the light-insensitive control. Values represent the average of the indicated number (N) of experiments and error bars show the standard deviation.

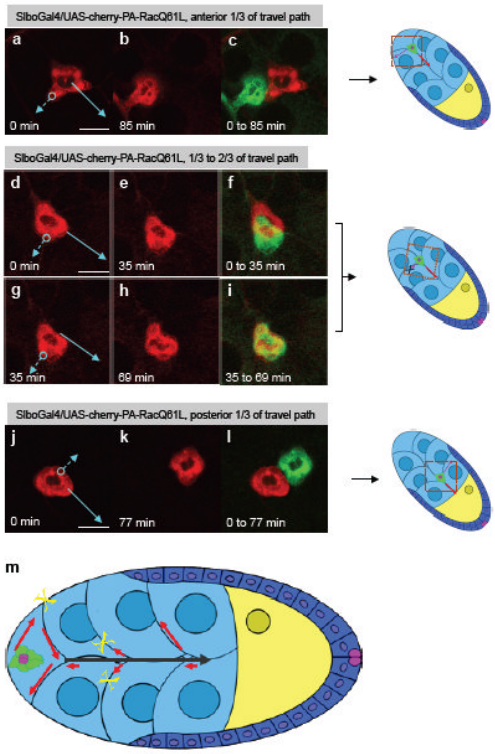


Figure 3.

Responsiveness of border cells to PA-RacQ61L depends on their location within the egg chamber.

(a-c) Within the anterior third of the egg chamber, photoactivation diverts border cells. (**d-i**) In the middle third, photoactivation has little effect. (**c**) In the posterior third, photoactivation again drives border cells toward the side. Scale bars represent 20 μm, elapsed time is shown in minutes. Schematics show border cell position within the egg chamber. In panels with two colors, red indicates the starting position and green shows the ending position. (**m**) Summary of experiments. The lengths of the arrows indicate the average distance migrated in the indicated direction in response to PA-RacQ61L, for border cells starting at the base of the arrow. The black arrow indicates the normal migration direction. Yellow Xs indicate positions beyond which border cells did not move. Each arrow summarizes at least five experiments.

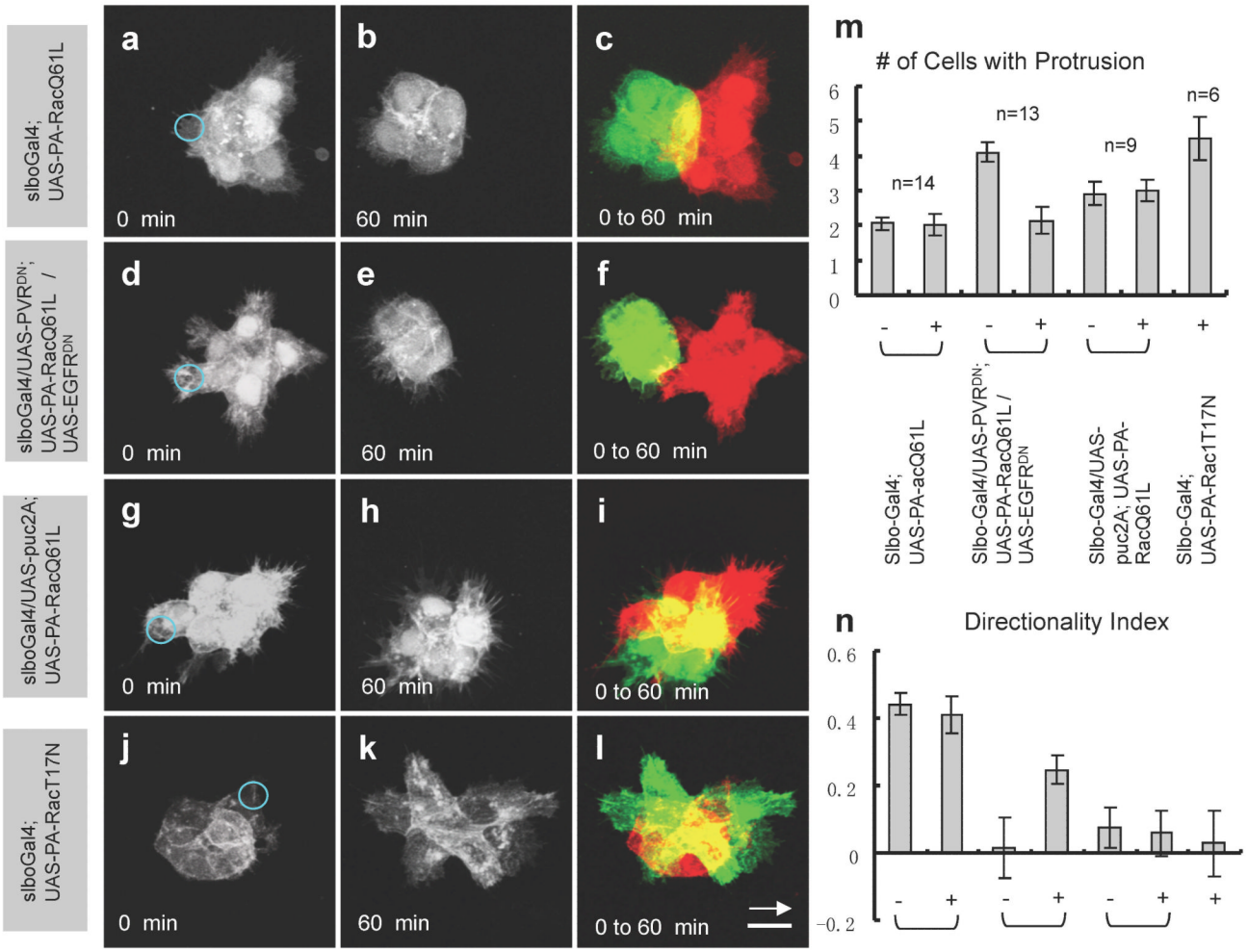


Figure 4.

Local photoactivation or photoinactivation of Rac in one cell affects the morphology and behavior of other cells in the group.

(**a-l**) Confocal images of border cell clusters before (0 min) and after (60 min) photoactivation. Circles indicate areas of laser treatment. The white arrow in panel **l** indicates the direction the border cells would normally migrate and applies to all panels. Scale bar is 10 μm. In **c**, **f**, **i** and **l**, red shows the starting position and green shows the ending position. **m**, The average number of cells sending protrusions simultaneously within one cluster was calculated from 3-D reconstructed images (see methods and Figure S6). "−" and "+" indicate before and after photoactivation. **n**, Directionality indices were calculated from the same samples (see methods).

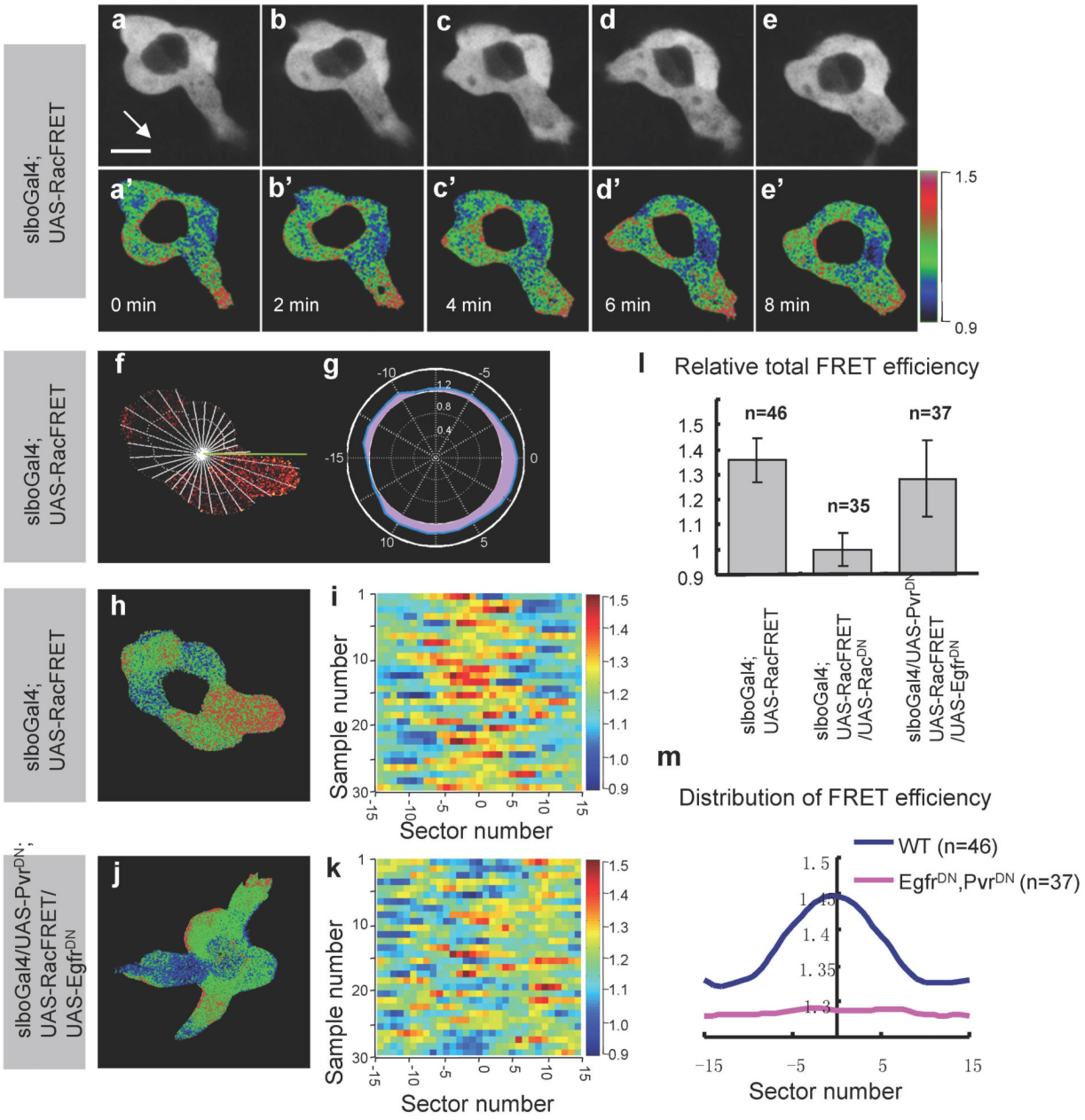


Figure 5.

Rac activity pattern in migrating border cell clusters.

(**a-e'**) A time lapse series of migrating border cells. (**a-e**) YFP channel only (**a'-e'**) processed FRET signal **f.** FRET image of wild-type border cells displayed with Red-Hot pseudocolor divided into 30 sectors. The yellow line shows the direction of migration. The white circle indcates the central region that was excluded from the analysis. (**g**) Average FRET efficiency from image **f** plotted in a radar map. Efficiency higher than 1.2 is highlighted in purple. **(h-j)** Representative FRET patterns in wild-type (h) and $EGFR^{DN}$- and PVR^{DN}-expressing (j) border cells. (**i**, **k)** Heatmaps from 30 examples of each genotype. Each row represents the FRET signal distribution of an individual border cell cluster.

Positions from −15 to 15 plotted on the x-axis correspond to the sectors, where 0 represents the front of the cluster. **l.** FRET efficiencies in border cells of the indicated genotypes. All results were normalized to the efficiency of Rac^{DN}. (**m)** Distributions of average FRET efficiencies in wild-type (blue) and $PVR^{DN}/EGFR^{DN}$ border cells, plotted as a function of sector number, where 0 represents the front.

Published in final edited form as:
Nature. 2010 May 27; 465(7297): 492–496. doi:10.1038/nature09075.

CD95/Fas promotes tumour growth

Lina Chen[1,*], **Sun-Mi Park**[1,*], **Alexei V. Tumanov**[2], **Annika Hau**[1], **Kenjiro Sawada**[3,†], **Christine Feig**[1,†], **Jerrold R. Turner**[2], **Yang-Xin Fu**[2], **Iris Romero**[3], **Ernst Lengyel**[3], and **Marcus E. Peter**[1]

[1]The Ben May Department for Cancer Research, The University of Chicago, 924 E 57th Street, Chicago, IL 60637

[2]Department of Pathology, The University of Chicago, 924 E 57th Street, Chicago, IL 60637

[3]Department of Obstetrics and Gynecology/Section of Gynecologic Oncology, The University of Chicago, 924 E 57th Street, Chicago, IL 60637

Abstract

CD95 (also called Fas and APO-1) is a prototypical death receptor that regulates tissue homeostasis mainly in the immune system through induction of apoptosis 1¯3. During cancer progression CD95 is frequently downregulated or cells are rendered apoptosis resistant 4,5 raising the possibility that loss of CD95 is part of a mechanism for tumour evasion. However, complete loss of CD95 is rarely seen in human cancers 4 and many cancer cells express large quantities of CD95 and are highly sensitive to CD95 mediated apoptosis *in vitro.* Furthermore, cancer patients frequently have elevated levels of the physiological ligand for CD95, CD95L 6. These data raise the intriguing possibility that CD95 could actually promote the growth of tumours through its nonapoptotic activities 7. Here we show that cancer cells in general, regardless of their CD95 apoptosis sensitivity, depend on constitutive activity of CD95, stimulated by cancer-produced CD95L, for optimal growth. Consistently, loss of CD95 in mouse models of ovarian cancer and liver cancer reduces cancer incidence as well as the size of the tumours. The tumorigenic activity of CD95 is mediated by a pathway involving JNK and c-Jun. These results demonstrate that CD95 plays a growth promoting role during tumorigenesis and suggest that efforts to inhibit its activity rather than to enhance its activation should be considered during cancer therapy.

To test the function of endogenous CD95 in tumour cells, expression of CD95 was reduced in various human cancer cell lines using CD95 specific shRNA lentiviruses (Supplementary Fig. 1). Infection of the CD95 high expressing ovarian cancer cell line HeyA8, which in vitro is sensitive to CD95 mediated apoptosis, with a lentiviral shRNA against CD95 (R#6) substantially reduced CD95 protein and surface expression resulting in loss of CD95 apoptosis sensitivity (Fig. 1a). Abrogation of CD95 expression also resulted in substantial

Correspondence and requests for materials should be addressed to M.E.P. (m-peter@northwestern.edu).
*these authors share first authorship
†present address: University of Osaka, Japan (K.S.); Cambridge Research Institute/Cancer Research UK, United Kingdom (C.F.); Northwestern University, Department of Medicine, Chicago, IL, USA (L.C, A.H. and M.E.P.).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Author Contributions L.C. and S.M.P performed the experiments; A.T. performed the PH; A.H., K.S., C.F. and I.R. performed some experiments; J.T. performed pathology analyses; Y.X.F. and E.L., supervised some experiments; M.E.P designed experiments and supervised the project.

Author Information Reprints and permission information is available at www.nature.com/reprints.

growth reduction. This was confirmed with another CD95 targeting shRNA (R#4) (Supplementary Fig. 2) and in another ovarian cancer cell line, SKOV3ip1, which expresses very low amounts of CD95 and is completely resistant to CD95L induced apoptosis (Fig. 1b and data not shown). Reconstitution of the SKOV3ip1 CD95 knock down cells with an siRNA resistant version of CD95 to endogenous levels restored the growth of SKOV3ip1 cells (Supplementary Fig. 3). A growth inhibiting effect of reducing CD95 expression was also found in cell lines derived from colon (HCT116), renal (CAKI-1), breast (MCF7), and liver (HepG2) cancer (Fig. 1c-f). Knocking down CD95 in MCF7 cells with the R#6 virus resulted, 6 days after infection, in growth arrest (data not shown). However, 3 weeks after infection cells expressing intermediate CD95 levels grew out, albeit with reduced growth as compared to vector control infected cells (Fig. 1e). Neither MCF7 nor CAKI-1 cells benefited from overexpressing CD95 (Supplementary Fig. 4) suggesting that cancer cells, regardless of the absolute level of CD95 expression, maintain expression of CD95 at a level sufficient to promote optimal growth. This is consistent with our previous report that in 22 tumour cell lines stimulation of CD95 did not result in increased proliferation 6. However, incubating cells with the neutralising CD95L monoclonal antibody (mAb), NOK-1, reduced cell growth (Supplementary Fig. 5a) suggesting that the small amount of CD95L produced by tumour cells (Supplementary Fig. 5b) contributes to their growth. To directly test this assumption we used lentivirus based shRNAs specific for CD95L which knocked down CD95L in a murine cell line expressing full length human CD95L (Supplementary Fig. 5c). We monitored growth of HepG2 cells infected with each of three independent CD95L specific shRNA viruses all of which caused reduction of CD95L expression (Fig. 1g). Paralleling the efficiency of the knock down the viruses caused different degrees of growth inhibition ranging from a reduction in growth (L#2) to a complete loss of growth (L#1) (Fig. 1h). Knocking down CD95L using an siRNA SmartPool (Supplementary Fig. 5c) also resulted in profound growth inhibition of HepG2 cells (Supplementary Fig. 5d). Finally, knocking down CD95L also reduced growth of all other cancer cell lines suggesting that CD95L is essential for the growth of many tumour cells (Fig. 1i). The knock down of CD95 in HepG2 only caused a moderate reduction in growth, possibly because this knock down may not have been efficient enough to cause a more pronounced effect. Given the small amount of CD95L expressed in tumour cells, reducing its expression is more likely to cause severe effects by falling below a threshold of minimal expression. This interpretation is consistent with a recent analysis that determined that the threshold for CD95 to signal apoptosis versus nonapoptotic signalling is 1000 times higher 8. Our data suggest that almost undetectable amounts of CD95 and CD95L are sufficient and in some cases required to promote growth of tumour cells.

CD95 is highly expressed in epithelial ovarian cancer but these cancer cells acquire resistance to CD95 mediated apoptosis 9. Large quantities of CD95L are found in patient ascites 10،11 and patient serum 12. As a first step toward understanding the role of CD95 in ovarian cancer *in vivo* we employed an intraperitoneal xenograft model of metastatic ovarian cancer 13. Nude mice injected with SKOV3ip1 cells infected with either the R#6 (Fig. 2a) or the R#4 (Fig. 2b) virus had about a 50% reduction in tumour mass, number of metastases and ascites formation. To test whether *in vivo* CD95 dependent tumour growth was driven by CD95L, SKOV3ip1 cells were grown for one week in nude mice and then treated with either a neutralising anti-CD95L mAb specific for mouse CD95L (MFL3) or for human CD95L (NOK-1). Tumour growth was attenuated only in mice injected with the human CD95L specific antibody (Fig. 2d) suggesting that it is predominantly the CD95L produced by the human cancer cells (see insert in Fig. 2d), rather than CD95L produced by the microenvironment in the mouse host, that drives CD95-mediated tumour growth.

To test whether the growth promoting role of CD95 could also be found in actual ovarian cancer we used a primary ovarian cancer cell line (MONTY-1 14) at an early passage

number. CD95 was knocked down in MONTY-1 cells (again resulting in reduced proliferation (Supplementary Fig. 6)) after which these cells were injected into nude mice. Similar to findings with SKOV3ip1 cells, MONTY-1 cells with reduced CD95 expression were completely resistant to CD95 mediated apoptosis and showed reduced intra-abdominal tumour load (Fig. 2e). Also, similar to findings with SKOV3ip1 cells (Fig. 2c) this result could, at least in part, be explained by a reduced proliferation rate (reduced Ki-67 staining) and reduced vascularization (reduced CD31 staining). The tumours showed an increase in TUNEL staining suggesting that the inhibition of the growth promoting CD95 receptor caused cancer cells to die (Fig. 2f).

Similar to other subtypes of epithelial ovarian cancers 9, the endometrioid subtype expresses CD95 (Fig. 3a) suggesting that in ovarian cancer CD95 could promote tumorigenesis. To test this we used a genetic model of epithelial endometrioid ovarian cancer based on the expression of activated mutant *K-ras* and deletion of *Pten* 15. For these studies we generated *K-ras/Pten* mutant mice carrying loxP sites in the *CD95* gene 16 (Supplementary Fig. 7). Cancer formation was initiated by injecting Adenovirus (AdV) Cre into the right ovarian bursa of the mice. Eight weeks after injection 9 out of 12 mice expressing wt CD95 had formed an ovarian cancer in the injected ovary with no cancer in the uninjected control ovary while only one of the 11 mice with knocked out CD95 had an early cancer (Fig. 3b-d). While wt mice started to die 8-9 weeks after AdV Cre injection in this model, only one of 4 mice lacking CD95 in the ovaries had an early cancer after 14 weeks. IHC confirmed that CD95 was highly expressed in the cancerous ovaries of wt mice and efficiently knocked out in the AdV-Cre injected ovaries of mice carrying the floxed CD95 alleles (Fig. 3d). PCR analysis to detect deleted *pten* confirmed that the Cre recombinase was equally active in cells of either genotype (Supplementary Fig. 7c). In summary, the data suggest that CD95 plays a role in the tumorigenesis of ovarian cancer.

Liver is the tissue with the highest constitutive CD95 expression. It was previously shown that mice expressing an apoptosis signalling deficient mutant of CD95 have a defect in liver regeneration after partial hepatectomy (PH) 17 raising the possibility that CD95 is involved in liver regeneration. To determine if it is CD95 expressed on hepatocytes that drives their proliferation, we generated hepatocyte-specific CD95 knock out mice by crossing CD95 loxP mice with mice expressing the Cre recombinase under control of the liver specific albumin promoter (*Alb-Cre* mice) (Supplementary Fig. 8). These mutant mice were then subjected to 2/3 PH. Proliferation of hepatocytes was severely impaired 48 hrs after PH in the absence of CD95 on hepatocytes (Fig. 4a,b) confirming that CD95 expression on hepatocytes is required for their proliferation. This raised a possibility that CD95 might also promote liver cancer.

Prolonged damage to the liver accompanied by compensatory hepatocyte proliferation can promote hepatocarcinogenesis 18 which can be recapitulated in a mouse model of diethylnitrosamine (DEN) induced hepatocellular carcinoma (HCC). As compared to wt mice, mice lacking expression of CD95 in the liver showed a clear reduction of HCC 8 months following a single injection of DEN (Fig. 4c). These data suggest that CD95 contributes to hepatocarcinogenesis. Ki-67 staining in the tumours of CD95 deficient livers was reduced (Fig. 4d,e), consistent with a contribution of CD95 to the growth of liver cancer. Because it was previously shown that expression of Cre recombinase can affect the apoptosis sensitivity of certain tissues 19, we confirmed our data in Cre expressing mice of a different genotype (Supplementary Fig. 9). DEN has been described to induce liver damage through a mechanism that involves ROS 18. Consistently, reduction of HCC in mice lacking CD95 expression in the liver was not due to a decrease in DEN induced damage because injection of DEN into 15 day old or adult wt or *CD95* k.o. mice did not induce significantly increased apoptosis and caused a similar activation of JNK (Supplementary Fig. 10).

However, the proliferative response (Ki-67 staining) of hepatocytes in DEN injected mice was strongly reduced in CD95 deficient livers (Supplementary Fig. 10a). In summary, our data support a role of CD95 in liver regeneration that translates into an increase in hepatocarcinogenesis in a chemically induced model of liver cancer.

JNK1 deficient mice have a reduced ability to recover from PH and are less susceptible to DEN induced liver cancer 20›21. Recently this link between JNK1 activation and HCC was confirmed in human HCC by two independent studies. HCC patients showed an increased activation of JNK1 in the cancer cells when compared to adjacent noncancerous tissue 20›22. Consistent with a function of JNK in proliferation of tumour cells we found that treatment of tumour cell lines with the JNK inhibitor SP600125 completely inhibited their growth (Supplementary Fig. 11a and Supplementary Results). Strikingly, hepatocytes from CD95 deficient mice showed reduced phosphorylation of JNK and the JNK substrate c-Jun (Fig. 4f) suggesting that loss of CD95 reduced the basal activity of JNK resulting in loss of proliferative capacity. To test whether stimulation of CD95 in the liver causes activation of JNK we injected wt mice with the agonistic anti-CD95 mAb, Jo2. Injection of Jo2 caused a massive increase in the phosphorylation of JNK and c-Jun both in mice that possessed an intact liver (these mice died when injected with Jo2) and in mice subjected to PH (known to protect mice from liver death 17, as evidenced by a lack of caspase-3 cleavage, TUNEL staining and increase in liver enzyme ALT) (Fig. 4g and Supplementary Fig. 12a). A clear increase in intracellular p-c-Jun staining of hepatocytes (Supplementary Fig. 12b) indicated that JNK was activated in hepatocytes in response to CD95 stimulation in mice protected from apoptosis by PH. The link between CD95 and the JNK pathway was also found in endometrioid tumours from *K-ras/Pten* mutant mice. Tumour tissue taken from cancerous ovaries expressing CD95 stained strongly for both CD95 and nuclear p-c-Jun, whereas tumour tissue from mice deficient for CD95 expression in the ovaries not only lacked expression of CD95, but was also devoid of p-c-Jun staining (Fig. 4h).

In aggregate the data suggest that cancer cells and the regenerating liver require a basal activity of CD95 to activate JNK resulting in phosphorylation of c-Jun and driving proliferation. To identify further downstream effectors in this pathway we isolated RNA from HeyA8 cells with knocked down CD95 using two independent shRNAs, and from HepG2 cells with knocked down CD95, and from CD95 deficient livers and compared gene expression profiles to the corresponding parental/wt cells. Interestingly, among all three systems, there were only two genes that were downregulated to a greater extent than CD95 itself—early growth 1 (Egr1) and the AP1 component c-Fos (Supplementary Fig. 13a-c). Both genes are downstream of JNK and are essential growth promoting transcription factors 23›24. Consistent with their function, both Egr1 and c-Fos were strongly upregulated in the liver of Jo2-injected, partial hepatectomised mice (Supplementary Fig. 13d) and their expression was strongly reduced in CD95 deficient ovarian cancer (Supplementary Fig. 13e).

Our data suggest that the CD95/CD95L system, rather than acting tumour suppressive, drives cancer growth joining the ranks of the TNFR1 and TNFα in stimulating tumour growth. In line with our results, studies suggest that CD95 activates neuronal stem cells 25 and acts as a tumour promoter for glioblastoma by activating src kinases 26. In addition it was recently shown that mice that only express soluble CD95L suffer from large histiocytic sarcomas in the liver 27 likely due to a lack of apoptosis induction and a tumorigenic activity of CD95L. Our data suggest that CD95 exerts a growth promoting activity mainly through a pathway involving JNK, c-Jun, Erg1 and c-Fos, and that tumour cells are stimulated by their own CD95L. The data provide an explanation for the long standing mystery of the role of CD95 expression in many tissues and in the majority of human

cancers without signs of apoptosis *in vivo*, and suggest that efforts to inhibit rather than to promote CD95 activity should be considered during cancer therapy.

Methods Summary

Endometrioid ovarian cancer induction

The method of ovarian cancer induction using *LSL-K-ras*$^{G12D/+}$*Pten*$^{loxP/loxP}$ and *LSL-K-ras*$^{G12D/+}$*Pten*$^{loxP/loxP}$*CD95*$^{loxP/loxP}$ female mice was described previously 28. Briefly, mice were sedated, the right ovary was exposed and the ovarian bursa was injected with AdCre (2.5×10^7 plaque-forming units) (University of Iowa Gene Transfer Vector Core). The left ovary was not injected and served as an internal control. Mice were evaluated weekly for palpable tumour and all were sacrificed 8 or 14 weeks after the injection of the virus. At the time of sacrifice the primary tumour was excised, weighed, measured and the number of metastatic nodules and volume of ascites were recorded. All tissue was fixed in 10% formalin, embedded in paraffin, and stained with haematoxylin and eosin.

Partial hepatectomy

2/3 partial hepatectomy was performed as previously described 29,30. In brief, mice were anaesthetised with ketamine (100 mg/kg) and xylazine (10 mg/kg) i.p., the liver was exposed through a midline incision, and the right and left lobes were sequentially ligated with a silk 4-0 suture, and resected. BrdU solution (50 mg/kg in PBS) was injected i.p. 2 hrs prior to analysis. DNA synthesis was measured by immunohistochemical staining of paraffin liver sections with anti-BrdU antibody. The percentage of BrdU positive hepatocyte nuclei among total hepatocyte nuclei was calculated using the Cellular Image Analysis System (ACIS, Clarient, San Juan Capistrano, CA).

Liver cancer induction

15 day old *CD95*$^{loxP/loxP}$ (wt) and *CD95*$^{loxP/loxP}$Albumin-Cre (*CD95* k.o.) male mice were injected with a single dose of DEN (25 μg/g body weight, i.p.). Eight months later, all mice were sacrificed and livers were excised. Parameters of total liver weight, number of liver surface nodules and maximum nodule diameters were recorded.

Full Methods and associated references are available in the online version of the paper at www.nature.com/nature.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to Drs. Alexander Chervonsky and Daniela Dinulescu for providing the *CD95* loxP/loxP mice and the *K-ras*$^{G12D/+}$*Pten*$^{fl/fl}$ mice, respectively, and to Dr. Syed Ahmed for help with one experiment. We are grateful to Terry Li for performing the IHC and to Dr. You-Jia Hua for analysing the gene array data. This work was funded by grants CA112240 from the NCI and CCFA 1661 from the Crohn's and Colitis Foundation of America.

References

1. Nagata S. Fas ligand-induced apoptosis. Annu Rev Genet. 1999; 33:29–55. [PubMed: 10690403]
2. Krammer PH. CD95's deadly mission in the immune system. Nature. 2000; 407:789–95. [PubMed: 11048730]
3. Strasser A, Jost PJ, Nagata S. The many roles of FAS receptor signaling in the immune system. Immunity. 2009; 30:180–92. [PubMed: 19239902]

4. Peter ME, Legembre P, Barnhart BC. Does CD95 have tumor promoting activities? Biochim Biophys Acta. 2005; 1755:25–36. [PubMed: 15907590]
5. Debatin KM, Krammer PH. Death receptors in chemotherapy and cancer. Oncogene. 2004; 23:2950–66. [PubMed: 15077156]
6. Barnhart BC, et al. CD95 ligand induces motility and invasiveness of apoptosis-resistant tumor cells. Embo J. 2004; 23:3175–85. [PubMed: 15272306]
7. Peter ME, et al. The CD95 receptor: apoptosis revisited. Cell. 2007; 129:447–50. [PubMed: 17482535]
8. Lavrik IN, et al. Analysis of CD95 threshold signaling: triggering of CD95 (FAS/APO-1) at low concentrations primarily results in survival signaling. J Biol Chem. 2007; 282:13664–71. [PubMed: 17347143]
9. Baldwin RL, Tran H, Karlan BY. Primary ovarian cancer cultures are resistant to Fas-mediated apoptosis. Gynecol Oncol. 1999; 74:265–71. [PubMed: 10419743]
10. Abrahams VM, et al. Epithelial ovarian cancer cells secrete functional Fas ligand. Cancer Res. 2003; 63:5573–81. [PubMed: 14500397]
11. Rabinowich H, et al. Lymphocyte apoptosis induced by Fas ligand- expressing ovarian carcinoma cells. Implications for altered expression of T cell receptor in tumor-associated lymphocytes. J Clin Invest. 1998; 101:2579–88. [PubMed: 9616229]
12. Taylor DD, Lyons KS, Gercel-Taylor C. Shed membrane fragment-associated markers for endometrial and ovarian cancers. Gynecol Oncol. 2002; 84:443–8. [PubMed: 11855885]
13. Sawada K, et al. C-Met overexpression is a prognostic factor in ovarian cancer and an effective target for inhibition of peritoneal dissemination and invasion. Cancer Res. 2007; 67:1670–1680. [PubMed: 17308108]
14. Kaur S, et al. {beta}3-integrin expression on tumor cells inhibits tumor progression, reduces metastasis, and is associated with a favorable prognosis in patients with ovarian cancer. Am J Pathol. 2009; 175:2184–96. [PubMed: 19808644]
15. Dinulescu DM, et al. Role of K-ras and Pten in the development of mouse models of endometriosis and endometrioid ovarian cancer. Nat Med. 2005; 11:63–70. [PubMed: 15619626]
16. Stranges PB, et al. Elimination of antigen-presenting cells and autoreactive T cells by Fas contributes to prevention of autoimmunity. Immunity. 2007; 26:629–41. [PubMed: 17509906]
17. Desbarats J, Newell MK. Fas engagement accelerates liver regeneration after partial hepatectomy. Nat Med. 2000; 6:920–3. [PubMed: 10932231]
18. Maeda S, Kamata H, Luo JL, Leffert H, Karin M. IKKbeta couples hepatocyte death to cytokine-driven compensatory proliferation that promotes chemical hepatocarcinogenesis. Cell. 2005; 121:977–90. [PubMed: 15989949]
19. Schmidt-Supprian M, Rajewsky K. Vagaries of conditional gene targeting. Nat Immunol. 2007; 8:665–8. [PubMed: 17579640]
20. Hui L, Zatloukal K, Scheuch H, Stepniak E, Wagner EF. Proliferation of human HCC cells and chemically induced mouse liver cancers requires JNK1-dependent p21 downregulation. J Clin Invest. 2008; 118:3943–53. [PubMed: 19033664]
21. Sakurai T, Maeda S, Chang L, Karin M. Loss of hepatic NF-kappa B activity enhances chemical hepatocarcinogenesis through sustained c-Jun N-terminal kinase 1 activation. Proc Natl Acad Sci U S A. 2006; 103:10544–51. [PubMed: 16807293]
22. Chang Q, et al. Sustained JNK1 activation is associated with altered histone H3 methylations in human liver cancer. J Hepatol. 2009; 50:323–33. [PubMed: 19041150]
23. Lim CP, Jain N, Cao X. Stress-induced immediate-early gene, egr-1, involves activation of p38/JNK1. Oncogene. 1998; 16:2915–26. [PubMed: 9671412]
24. Cavigelli M, Dolfi F, Claret FX, Karin M. Induction of c-fos expression through JNK-mediated TCF/Elk-1 phosphorylation. EMBO J. 1995; 14:5957–64. [PubMed: 8846788]
25. Corsini NS, et al. The death receptor CD95 activates adult neural stem cells for working memory formation and brain repair. Cell Stem Cell. 2009; 5:178–90. [PubMed: 19664992]
26. Kleber S, et al. Yes and PI3K Bind CD95 to Signal Invasion of Glioblastoma. Cancer Cell. 2008; 13:235–48. [PubMed: 18328427]

27. La OR, et al. Membrane-bound Fas ligand only is essential for Fas-induced apoptosis. Nature. 2009; 461:659–63. [PubMed: 19794494]

28. Romero IL, et al. Effects of oral contraceptives or a gonadotropin-releasing hormone agonist on ovarian carcinogenesis in genetically engineered mice. Cancer Prev Res (Phila Pa). 2009; 2:792–9. [PubMed: 19737983]

29. Greene AK, Puder M. Partial hepatectomy in the mouse: technique and perioperative management. J Invest Surg. 2003; 16:99–102. [PubMed: 12746193]

30. Mitchell C, Willenbring H. A reproducible and well-tolerated method for 2/3 partial hepatectomy in mice. Nat Protoc. 2008; 3:1167–70. [PubMed: 18600221]

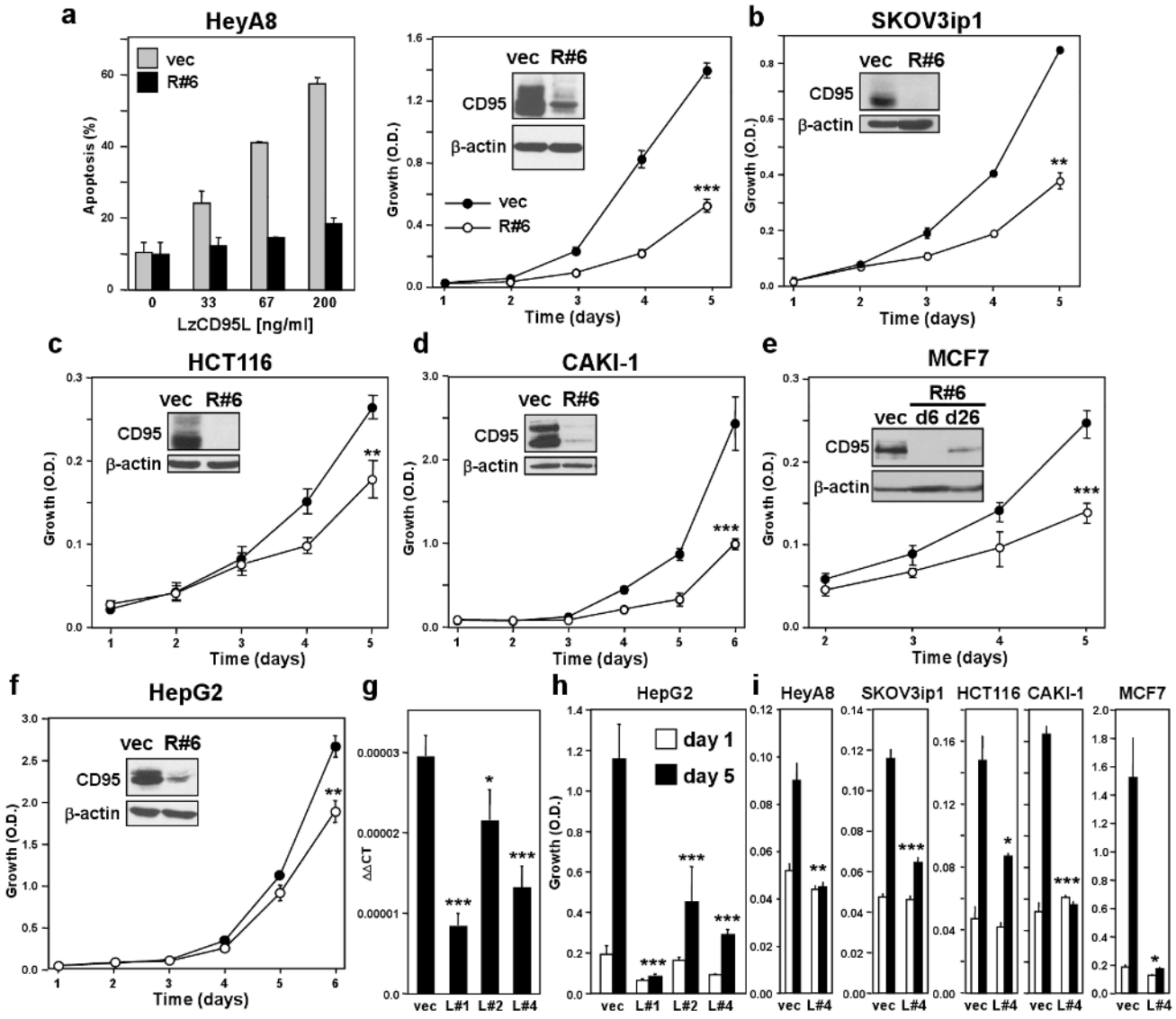


Figure 1. Reducing CD95 or CD95L expression inhibits cell proliferation of cancer cells

a-f, Growth of different cell lines infected with the CD95 specific shRNA lentivirus CD95shRNA#6 (R#6). Inserts show the total CD95 expression levels of cells expressing scrambled control (vec) or R#6 as determined by western blot analysis. A similar effect was also found with a CD95 expressing variant of the neuroblastoma cell line NB4 (not shown). Apoptosis sensitivity of HeyA8 vec and R#6 cells by LzCD95 ligand treatment was determined by quantifying DNA fragmentation (**a**). **g**, HepG2 cells were stably infected with three different CD95L specific shRNA lentiviruses and CD95L mRNA was quantified using real time PCR. **h**, Growth of cells in g over 5 days. **i**, Growth of different cell lines infected with the L#4 virus. Proliferation of cells was examined by SRB assay (a-f and h and i). * $p<0.05$, ** $p<0.01$, *** $p<0.001$. Values in graphs in **a** to **i** represent mean -/+ s.d. from three independent experiments.

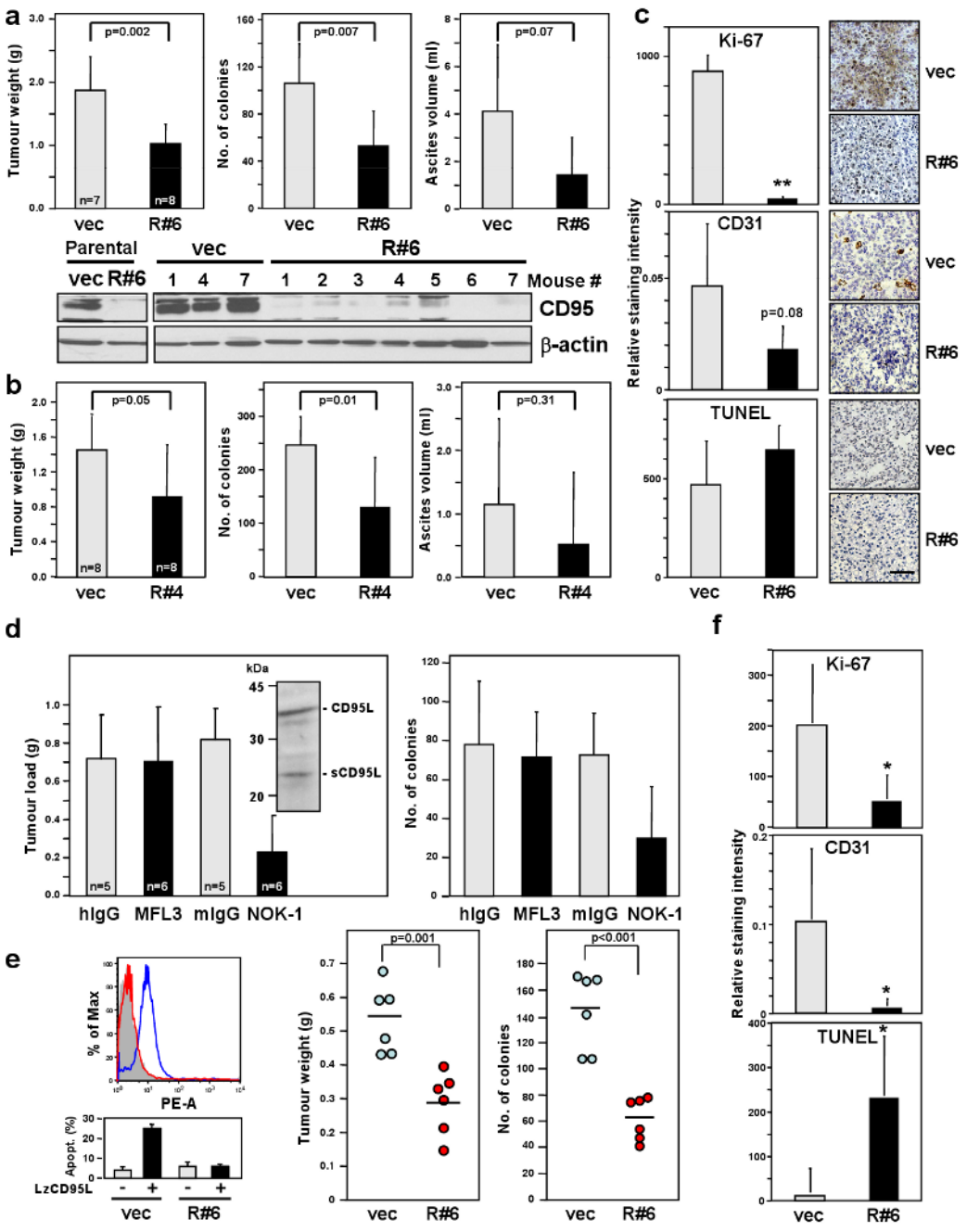


Figure 2. Loss of CD95 expression inhibits ovarian cancer in vivo

a Tumour weight, number of tumour colonies and ascites from mice injected with SKOV3ip1 vec or R#6 cells. Lysates of cells and tumour tissues were examined for CD95 level by western blot analysis. **b**, Same parameters as in (a) were measured from mice injected with SKOV3ip1 vec or R#4 cells. **c**, Histology and immunohistochemistry staining for Ki-67, TUNEL and CD31 of SKOV3ip1 vec and R#6 tumours. Scale bar = 100 μm. ** p-value <0.001. **d**, Tumour load and number of tumour colonies of mice treated with neutralising mAb for murine CD95L (MFL3), human CD95L (NOK-1) or corresponding isotype control mAbs were measured. Inset, Western blot analysis of SKOV3ip1 cell lysate

for CD95L. **e**, Surface CD95 staining (upper left) and apoptosis sensitivity by LzCD95L (100 ng/ml) treatment (lower left) of MONTY-1 vec and R#6 cells. Weight and number of colonies of tumours formed by MONTY-1 vec and R#6 cells are shown (right). **f**, The staining intensity for Ki-67, TUNEL and CD31 of tumours from MONTY-1 vec and R#6 cells were quantified. * p-value <0.05. Values in graphs in **a** to **e** and **f** represent mean -/+ s.d. from three independent experiments. The horizontal bars in right part part of **e** represent the mean of 6 animals.

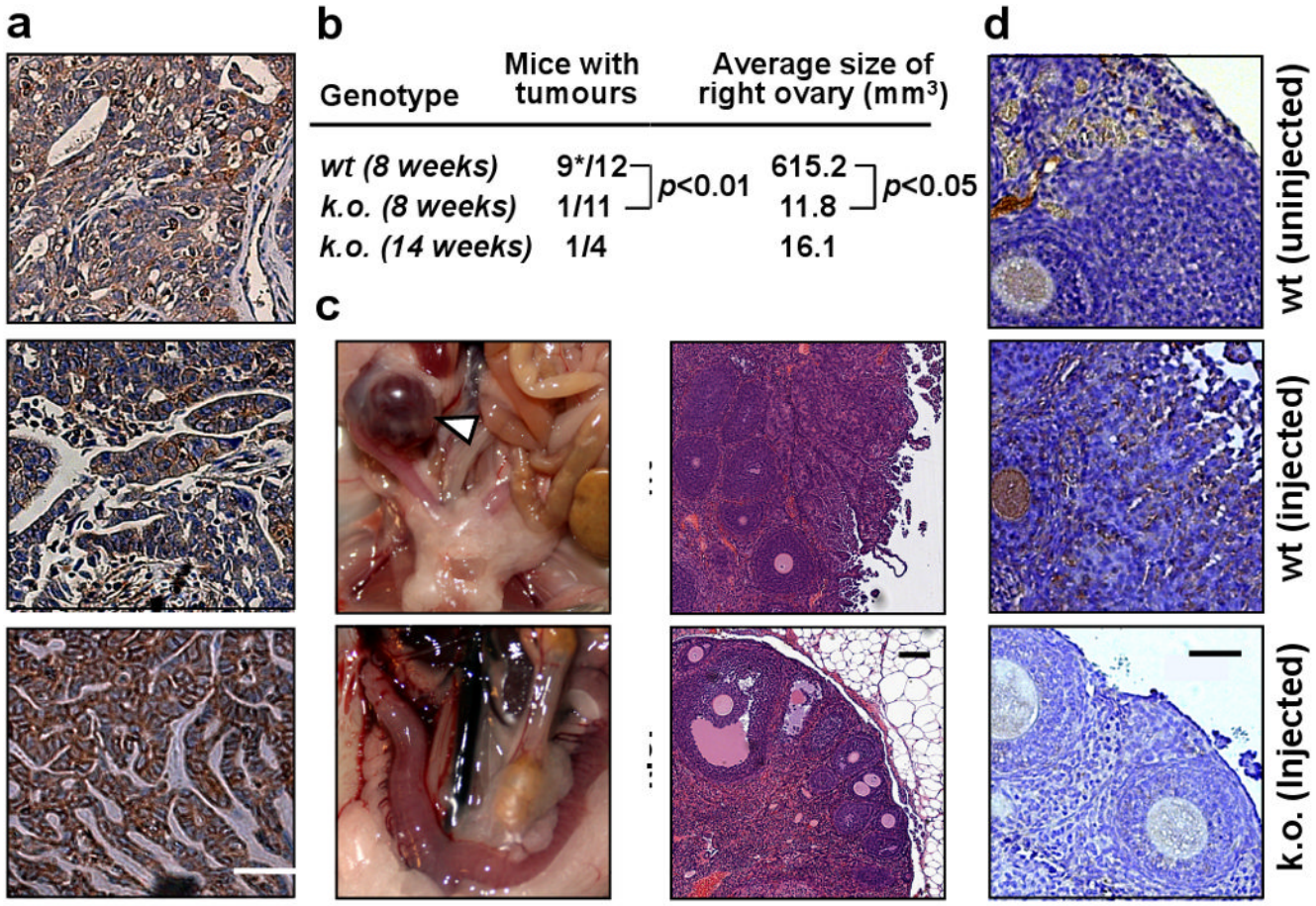


Figure 3. Deletion of CD95 leads to reduction in tumour formation in a spontaneous model of endometrioid ovarian cancer

a, Staining for CD95 in three primary endometrioid ovarian cancers. Scale bar = 50 μm. **b**, Number of mice that formed visible tumours either 8 or 14 weeks after injection of AdV Cre into the right ovarian bursa. * one mouse died from ovarian cancer 42 days after AdV Cre injection. **c**, Representative image and histology of right ovaries from wt and k.o. mice 8 weeks after injection of AdV Cre. Arrow head indicates ovarian tumour. Scale bar = 100 μm. **d**, CD95 staining of ovary from untreated wt, AdV Cre treated wt or AdV Cre treated k.o. mice. Scale bar = 50 μm. wt, *LSL-K-ras*$^{G12D/+}$*Pten*$^{loxP/loxP}$*CD95*$^{wt/wt}$; k.o., *K-ras*$^{G12D/+}$*Pten*$^{loxP/loxP}$*CD95*$^{loxP/loxP}$.

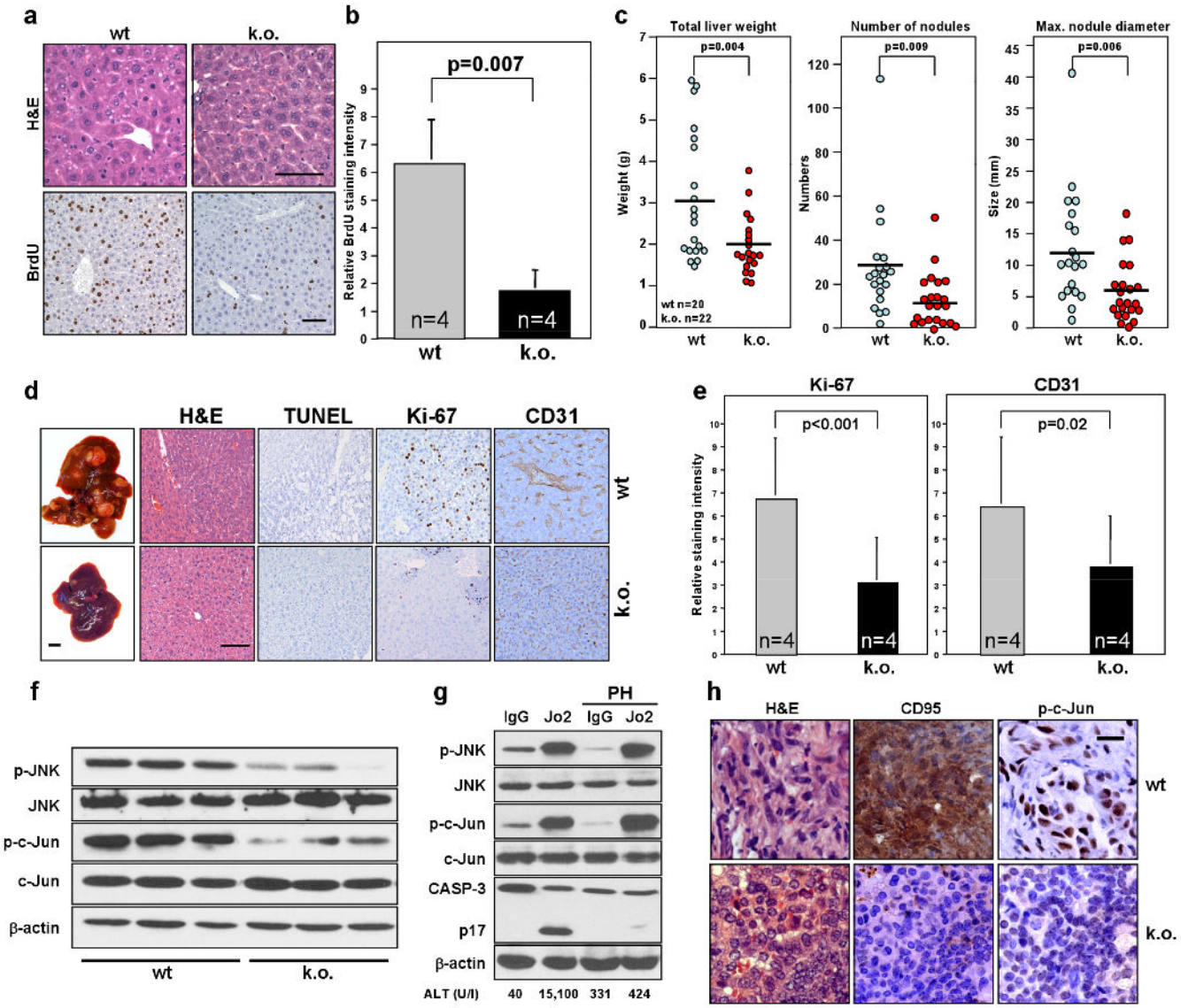


Figure 4. Deletion of CD95 in the liver leads to a decrease in tumour formation caused by reduced ability of hepatocytes to proliferate and to activate JNK

a, H&E and BrdU staining of livers from wild-type (wt) and liver specific *CD95* k.o. mice 48 hrs after partial hepatectomy. Scale bars = 100 μm. **b**, Quantification of relative BrdU staining intensity of the mice in (**a**). **c**, wt and liver specific *CD95* k.o. mice were injected with a single dose of DEN i.p. to induce liver tumour formation. 8 months later, all mice were sacrificed and parameters of total liver weight, number of liver surface nodules and maximum nodule diameter were recorded. **d**, Intact livers (scale bar = 1 cm), H&E staining and immunohistochemistry of TUNEL, Ki-67 and CD31 from mice in (**c**). Scale bar = 100 μm. **e**, Quantification of Ki-67 and CD31 staining for liver samples in (**d**). There was no detectable TUNEL staining. **f**, Western blot for phospho-JNK and phospho-c-Jun in three untreated wt and three *CD95* k.o. mouse livers. **g**, wt mice with or without PH were injected i.p with 10 μg of murine CD95-specific agonistic antibody, Jo2 or isotype matched control mAb. After 6 hours phospho-JNK, p-c-Jun levels and cleavage of caspase-3 in livers were measured by western blot analysis. Concentration of the liver enzyme ALT in the serum of injected mice is given. **h**, Immunohistochemistry staining of ovarian tumours from wt (*LSL-*

K-ras$^{G12D/+}$ *Pten*$^{loxP/loxP}$ *CD95*$^{wt/wt}$) or k.o. (*K-ras*$^{G12D/+}$ *Pten*$^{loxP/loxP}$ *CD95*$^{loxP/loxP}$) mice 8 weeks after injection of Adv Cre for CD95 and phospho-c-Jun. Scale bar = 20 μm. Values in graphs in **b** and **e** represent mean -/+ s.d. from three independent experiments. The horizontal bars in **c** represent the mean.

OPEN ACCESS Freely available online

Most "Dark Matter" Transcripts Are Associated With Known Genes

Harm van Bakel[1], Corey Nislow[1,2], Benjamin J. Blencowe[1,2], Timothy R. Hughes[1,2]*

1 Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada, **2** Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

Abstract

A series of reports over the last few years have indicated that a much larger portion of the mammalian genome is transcribed than can be accounted for by currently annotated genes, but the quantity and nature of these additional transcripts remains unclear. Here, we have used data from single- and paired-end RNA-Seq and tiling arrays to assess the quantity and composition of transcripts in PolyA+ RNA from human and mouse tissues. Relative to tiling arrays, RNA-Seq identifies many fewer transcribed regions ("seqfrags") outside known exons and ncRNAs. Most nonexonic seqfrags are in introns, raising the possibility that they are fragments of pre-mRNAs. The chromosomal locations of the majority of intergenic seqfrags in RNA-Seq data are near known genes, consistent with alternative cleavage and polyadenylation site usage, promoter- and terminator-associated transcripts, or new alternative exons; indeed, reads that bridge splice sites identified 4,544 new exons, affecting 3,554 genes. Most of the remaining seqfrags correspond to either single reads that display characteristics of random sampling from a low-level background or several thousand small transcripts (median length = 111 bp) present at higher levels, which also tend to display sequence conservation and originate from regions with open chromatin. We conclude that, while there are bona fide new intergenic transcripts, their number and abundance is generally low in comparison to known exons, and the genome is not as pervasively transcribed as previously reported.

Citation: van Bakel H, Nislow C, Blencowe BJ, Hughes TR (2010) Most "Dark Matter" Transcripts Are Associated With Known Genes. PLoS Biol 8(5): e1000371. doi:10.1371/journal.pbio.1000371

Academic Editor: Sean R. Eddy, HHMI Janelia Farm, United States of America

Received December 3, 2009; **Accepted** April 9, 2010; **Published** May 18, 2010

Funding: This work was supported by Genome Canada (http://www.genomecanada.ca) through the Ontario Genomics Institute, the Ontario Research Fund, and March of Dimes (http://www.marchofdimes.com). HvB was supported by the Netherlands Organization for Scientific Research (NWO; http://www.nwo.nl) (grant no. 825.06.033) and the Canadian Institutes of Health Research (CIHR; http://www.cihr-irsc.gc.ca/) (grant no. 193588). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: APA, alternative cleavage and polyadenylation; BW, bandwidth parameter; CAGE, capped analysis of gene expression; CNV, copy number variation; lincRNAs, large intervening noncoding RNAs; ncRNAs, noncoding RNAs; ORF, open reading frames; pasRNA, promoter-associated RNA; TSS, transcription start site; TTS, transcription termination site; TU, transcript unit; TUF, transcript of unknown function

* E-mail: t.hughes@utoronto.ca

Introduction

In recent years established views of transcription have been challenged by the observation that a much larger portion of the human and mouse genomes is transcribed than can be accounted for by currently annotated coding and noncoding genes. The bulk of these findings have come from experiments using "tiling" microarrays with probes that cover the non-repetitive genome at regular intervals [1–9], or from sequencing efforts of full-length cDNA libraries enriched for rare transcripts [10,11]. Additionally, capped analysis of gene expression (CAGE) in human and mouse show that a significant number of sequenced 5′ tags map to intergenic regions [12]. Estimates of the proportion of transcripts that map to locations separate from known exons range from 47% to 80% and are distributed approximately equally between introns and intergenic regions. Dubbed transcriptional "dark matter" [13], the "hidden" transcriptome [1], or transcripts of unknown function (TUFs) [4,14], the exact nature of much of this additional transcription is unclear, but it has been presumed to comprise a combination of novel protein coding transcripts, extensions of existing transcripts, noncoding RNAs (ncRNAs), antisense transcripts, and biological or experimental background. Determining the relative contributions of each of these potential sources is important for understanding the nature and possible biological function of transcriptional dark matter.

Homology searches for transcripts mapping outside known annotation boundaries [10], as well as cDNA sequencing efforts, indicate that it is still possible to find new exons of protein coding genes [10,15,16]. The genomic positions of TUFs are also biased towards known transcripts [8], suggesting that at least a portion may represent extensions of current gene annotations. Nevertheless, the majority of dark matter transcripts is thought to be noncoding [2,4,5,10]. Previous efforts to characterize dark matter transcripts have revealed the existence of thousands of ncRNAs with evidence for tissue-specific expression [17,18], as well as over a thousand large intervening noncoding RNAs (lincRNAs) originating from intergenic regions bearing chromatin marks associated with transcription [19]. Other studies have reported new classes of ncRNAs, such as those that cluster close to the transcription start sites (TSSs) of protein coding genes [20–24]. These promoter-associated RNAs (pasRNAs) typically initiate in the nucleosome free regions that mark a TSS, with transcription occurring in both directions. Finally, results from the ENCODE pilot project have suggested a highly interleaved structure of the

Author Summary

The human genome was sequenced a decade ago, but its exact gene composition remains a subject of debate. The number of protein-coding genes is much lower than initially expected, and the number of distinct transcripts is much larger than the number of protein-coding genes. Moreover, the proportion of the genome that is transcribed in any given cell type remains an open question: results from "tiling" microarray analyses suggest that transcription is pervasive and that most of the genome is transcribed, whereas new deep sequencing-based methods suggest that most transcripts originate from known genes. We have addressed this discrepancy by comparing samples from the same tissues using both technologies. Our analyses indicate that RNA sequencing appears more reliable for transcripts with low expression levels, that most transcripts correspond to known genes or are near known genes, and that many transcripts may represent new exons or aberrant products of the transcription process. We also identify several thousand small transcripts that map outside known genes; their sequences are often conserved and are often encoded in regions of open chromatin. We propose that most of these transcripts may be by-products of the activity of enhancers, which associate with promoters as part of their role as long-range gene regulatory sites. Overall, however, we find that most of the genome is not appreciably transcribed.

human transcriptome, with an estimate that as much as 93% of the human genome may give rise to primary transcripts [9]. Though this estimate was based on a combination of sources that included rapid amplification of cDNA ends coupled to detection on tiling arrays (RACE-tiling), manually curated GENCODE annotations, and paired-end sequencing of long cDNAs (GIS-PET), it was dominated by the results of RACE-tiling experiments that alone found 80% genome coverage, compared to 64.6% and 66.4% for GENCODE annotations and GIS-PET, respectively.

The fact that most TUFs do not appear to be under evolutionary selective pressure [25] has prompted suggestions that at least some of the transcriptional dark matter may constitute "leaky" background transcription [9,26]. Consistent with this notion, many of the intergenic and intronic transcripts are detected at low levels, close to the detection limit of qPCR or Northern blots [13]. Presumably as a consequence, validation rates for unannotated transcribed regions detected in tiling array experiments have varied between 25% and 70% [1,5,27], and a comparison [13] of human chromosome 22 data from three major tiling array studies done on different platforms [1,3,27] also revealed little overlap of expressed probes, with 89% of overlapping positive probes mapping to exons or introns of known transcripts. While this low overlap may be due to differences in the samples analyzed [4], there is also evidence that some dark matter transcripts may be due to experimental artifacts. For example, a reassessment of the analysis parameters used in the tiling array study by Kampa et al. [2] revealed a similar number of transcribed fragments in real and randomized microarray data [28]. These issues make it difficult to assess the level of false positives in tiling array experiments.

Transcriptome sequencing (RNA-Seq) has emerged as a new technology that does not suffer from many of the limitations of array platforms such as cross-hybridization [29]. The technique has a wide dynamic range spanning at least four to five orders of magnitude [30,31] and allows accurate quantitation of expression levels, as determined by experiments using externally spiked-in RNA controls and quantitative PCR [30]. These characteristics make RNA-Seq suitable to accurately assess the relative proportion of sequence from the known versus the dark matter transcriptome. Comparisons between studies of eukaryotic transcriptomes have shown that the estimated proportion of transcriptional dark matter reported in RNA-Seq studies is consistently lower than estimates from tiling arrays [32]. Although most RNA-Seq studies to date have focused on polyadenylated (PolyA+) RNA, which would be enriched for coding transcripts, this cannot fully account for the differences, as most tiling array studies show nearly the same degree of nonexonic transcription for PolyA+ as for total RNA sources [1–9]. Indeed, it was reported that even in the most mature form of PolyA+ RNA isolated from the cytosol, approximately half of the transcribed sequence does not correspond to known exons [5]. Moreover, RNA-Seq data from Arabidobsis rRNA-depleted total RNA samples contained a relatively small proportion (3.5%) of intergenic reads [33]. These results may not be characteristic of the larger and more complex human and mouse transcriptomes, but they do present an example in which the proportion of dark matter transcripts is relatively low in a more heterogeneous RNA pool. Other studies, in contrast, reported a higher proportion of nonexonic reads in yeast [34] and for total RNA in human [35], leaving unresolved the question of the quantity and character of dark matter transcripts.

To investigate the extent and nature of transcriptional dark matter, we have analyzed a diverse set of human and mouse tissues and cell lines using tiling microarrays and RNA-Seq. A meta-analysis of single- and paired-end read RNA-Seq data reveals that the proportion of transcripts originating from intergenic and intronic regions is much lower than identified by whole-genome tiling arrays, which appear to suffer from high false-positive rates for transcripts expressed at low levels. The majority of RNA-Seq reads that map to intergenic regions either display a high degree of correlation with neighboring genes or are associated with more than 10,000 potential novel exonic fragments we identified in human and mouse. A genome-wide analysis of "*de novo*" splice junctions in human samples further revealed 2,789 previously uncharacterized transcript fragments that have no overlap with exons of known gene annotations, 1,259 of which map to intergenic regions. We also find 4,544 additional exons for annotated transcripts, 723 of which extend transcripts at the 5′ end and include likely alternative promoters. The novel exons from spliced transcripts are supported by EST data, are generally more conserved, and derive from coding as well as noncoding transcripts. We conclude that analysis of data from tiling arrays leads to vast overestimates of the proportion of transcriptional dark matter. However, the mammalian transcriptome does contain thousands of unannotated transcripts, exons, promoters, and termination sites. Intriguingly, there is a strong overlap of short intergenic transcripts with DNase I hypersensitive sites, suggesting that they may be the equivalent of pasRNAs for distant enhancers.

Results

High False-Positive Rate from Tiling Arrays

We directly compared the accuracy of tiling arrays and RNA-Seq in identifying known transcribed regions from polyadenylated (PolyA+) RNA. To avoid potential genomic abnormalities of cell lines we mainly focused on transcriptome data from tissue sources. For microarray expression profiling, we used Affymetrix whole-genome tiling arrays at a 35 bp resolution for four human and four mouse tissues. In addition, we generated RNA-Seq data for cDNA fragments from human whole brain tissue (multiple donors) and a mixture of cell lines, which were sequenced at both ends on an

Illumina genome analyzer to an average depth of 23 M paired 50 nt reads per sample. To match coverage across a wider variety of tissues, we supplemented the paired-end RNA-Seq data with publicly available 32 nt single-end PolyA+ selected datasets, sequenced to an average depth of 22 M reads for 8 human tissues from single donors [16]. RNA-Seq data for mouse were obtained from Mortazavi et al. [36] and consisted of 25 nt single-end data for PolyA+ RNA from three tissues, sequenced to an average depth of 73 M reads. The resulting combined dataset contained tissue-matched RNA-Seq and tiling array data for 4 human and 3 mouse tissues. For our analyses, we only considered RNA-Seq reads that could be unequivocally mapped to unique positions in the genome. This avoided erroneous identification of transcribed regions and facilitated comparisons to data obtained from tiling arrays, which were designed for the non-repetitive part of the genome. Overall the total number of uniquely mapped reads numbered 185.6 M and 79.8 M for the human and mouse genomes, respectively (see Table S1 for a breakdown per tissue). Since the arrays contained only perfect-match probes, the raw intensity data were normalized against a genomic DNA reference to correct for any bias in probe sequence composition (Materials and Methods).

We compared the performance of tiling arrays and RNA-Seq for human total brain tissue, since it had the highest combined sequence coverage of any tissue used in this study (50.2 M uniquely mapped reads from three independent samples, corresponding to 2.1 Gb of sequencing data). Figures 1A and 1B show the relation between the fraction of detected transcript fragments on tiling arrays (transfrags) or in RNA-Seq data (seqfrags) that overlap known RefSeq exons (i.e., precision) and the total fraction of exons recovered (i.e., recall). Tiling array transfrags were identified by selecting consecutive probes that scored above a range of intensity thresholds, with additional limits on the minimum length of each transfrag (minrun) and the maximum gap between probes meeting the threshold (maxgap). The analysis was performed directly on the normalized intensity data, or after applying additional median smoothing across neighboring probes in the genome within a sliding window, to reduce intensity variability. Seqfrags were defined as consecutively transcribed regions in the uniquely mapped RNA-Seq data, and performance was evaluated over a range of thresholds set on the minimum number of reads per seqfrag. We find that RNA-Seq offers superior precision in identifying RefSeq exons compared to tiling arrays, while achieving a high level of recall (Figure 1A, 1B). This difference remains apparent even over a broad range of parameter settings typically used to identify transcribed regions in tiling array data. These observations do not directly demonstrate that tiling arrays have a higher false-positive rate, as a lower precision would also be expected if the majority of the genome were transcribed: the difference between platforms could also reflect a lack of sensitivity to detect unannotated transcripts expressed at lower levels in RNA-Seq data, due to insufficient sequencing depth. If this were the case, however, we would expect that the precision-recall curves for RNA-Seq data would look progressively more similar to those of the tiling arrays with increasing read counts. Instead, when we examined the effect of varying sequencing depth by sampling smaller subsets of reads from the combined human brain RNA-Seq datasets we found that increased sequencing improves recall without a loss in precision (Figure 1B). Thus, the discrepancy with tiling arrays increases rather than decreases with greater sequencing depths.

We also directly compared RNA-Seq read coverage with tiling array measurements at the same genomic location. Figure 1C shows a direct comparison between the number of reads and the normalized probe signal intensity. Consistent with the precision-recall curves that show that high precision in tiling array experiments is only achieved at the most stringent intensity thresholds (Figure 1A), we find that the agreement between sequencing data and array intensities data is poor for all but the most highly transcribed regions. Indeed, the normalized intensity distribution for tiling array probes overlapping transcribed regions in RNA-Seq data with single-read coverage is essentially random (Figure 1D), consistent with previous observations that the correlation between RNA-Seq data and tiling arrays is poor for transcripts expressed at low levels [29,36]. We do note, however, that the tiling arrays and RNA-Seq data generally agree on the location of the greatest transcript mass (Figure 1C, red line). The increased precision of RNA-Seq is presumably due to reduced ambiguity in detecting transcripts at lower expression levels, relative to microarrays, in which signal from cross-hybridization increasingly contributes to false-positive detection at low expression levels. It is thus conceivable that the proportion of dark matter transcripts based on tiling array experiments is considerably overestimated. Given the improved performance of RNA-Seq over tiling arrays, we therefore focused on RNA-Seq data to revisit the nature of dark matter transcripts.

Dark Matter Transcripts Make up a Small Fraction of the Total Sequenced Transcript Mass

To assess the proportion of unique sequence-mapping reads accounted for by dark matter transcripts in RNA-Seq data, we compared the mapped sequencing data to the combined set of known gene annotations from the three major genome databases (UCSC, NCBI, and ENSEMBL, together referred to here as "annotated" or "known" genes). When considering uniquely mapped reads in all human and mouse samples, the vast majority of reads (88%) originate from exonic regions of known genes (Figure 2A). These figures are consistent with previously reported fractions of exonic reads of between 75% and 96% for unique reads [16,33,36–38], including those of the original studies from which some of the RNA-Seq data in this study were derived. When including introns, as much as 92%–93% of all reads can be accounted for by annotated gene regions. A further 4%–5% of reads map to unannotated genomic regions that can be aligned to spliced ESTs and mRNAs from high-throughput cDNA sequencing efforts, and only 2.2%–2.5% of reads cannot be explained by any of the aforementioned categories. The proportions of mapped reads are consistent between tissues and cell lines and independent of read sequence length (Table S1). Altogether, dark matter transcripts only account for a small proportion of PolyA+ transcripts.

While annotated exons can explain the majority of reads, they make up a much smaller proportion of the total transcribed area of the genome: 22.3% in human and 50.6% in mouse (Figure 2B). Nevertheless, complete annotated gene structures in both organisms still account for ~75% of the total transcribed area. The apparent discrepancy in transcribed intronic versus exonic area in human versus mouse is directly related to the combined increased sequencing depth for the human samples (Table S2). This is illustrated in Figure 2C, which shows the relationship between the amount of sequence coverage in the combined PolyA+ RNA-Seq data from human brain samples and the transcribed area. While the exonic transcribed area levels off quickly at around 500 Mb of RNA-Seq coverage, intergenic and intronic areas keep increasing at roughly constant rates. When we extrapolate from the observed relationship between the amount of mapped sequence data and genomic area covered (Figure 2D), we find that given sufficient sequencing depth the whole genome may

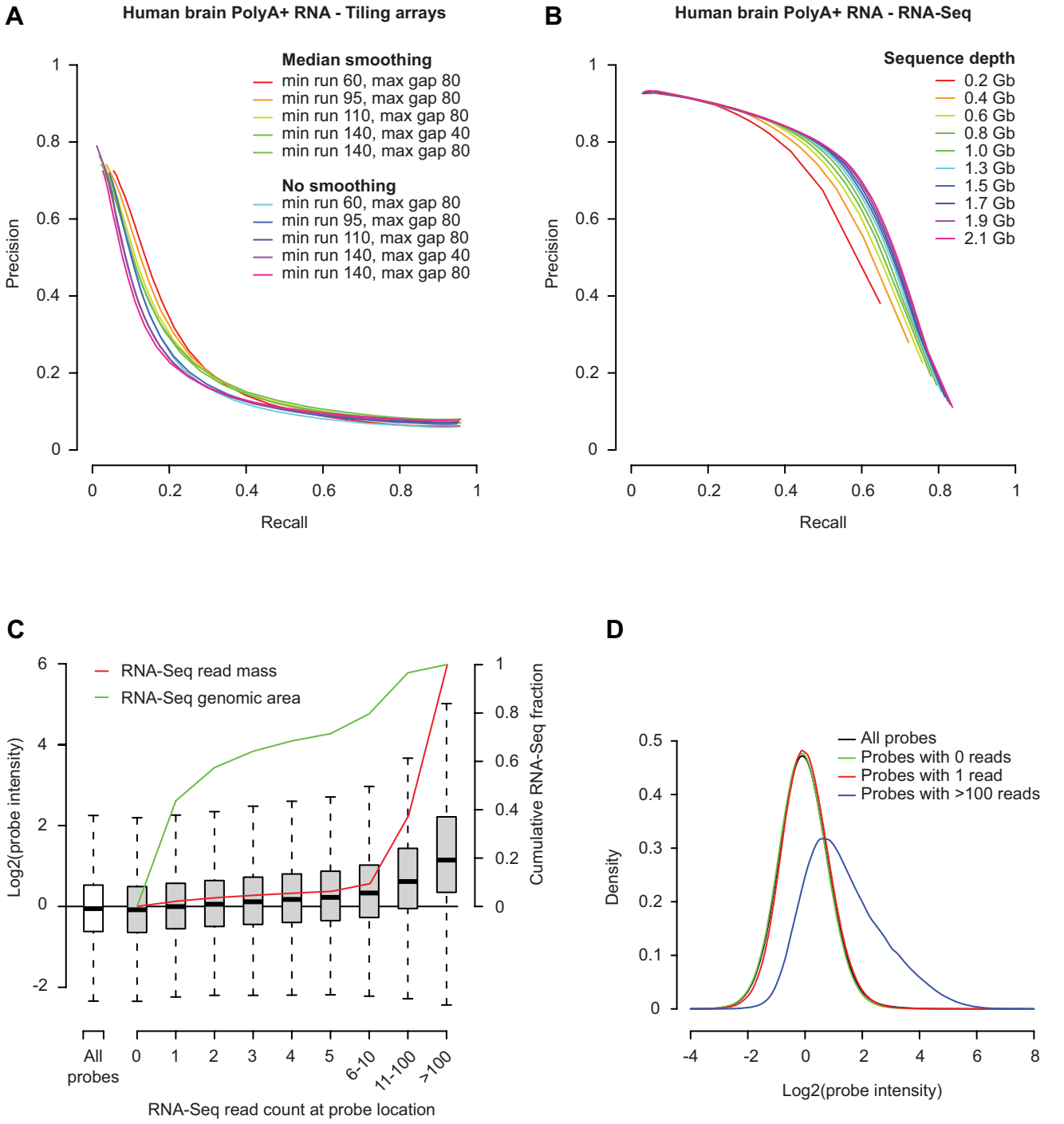


Figure 1. Low precision for tiling arrays compared to RNA-Seq data. (A) Precision-recall curves for detection of exons in human RefSeq gene annotations on tiling arrays. Transcribed genomic regions (transfrags) were selected based on a range of parameters that were applied before or after median smoothing with a bandwidth of 70 bp: max gap, the maximum distance between two positive probes; min run, the minimum size of a transcribed region. The log2 normalized intensity threshold used to select positive probes was varied between −1 and 2 to plot each line. (B) Precision-recall curves for the combined RNA-Seq data from three human brain samples, at different read depths (0.2 to 2.1 Gb). Transcribed regions (seqfrags) were identified on the basis of uniquely mapped reads, and the threshold for the minimal read count per seqfrag was varied between 1 and 100 to plot each line. (C) Comparison of RNA-Seq read counts and tiling array probe intensities for the pooled set of human brain RNA-Seq reads (three samples). The number of RNA-Seq reads overlapping each mapped probe coordinate was determined and used to draw a boxplot of the intensity distributions measured for probes overlapped by varying numbers of RNA-Seq reads, as indicated (gray boxes). The intensity distribution across all probes is shown in comparison (white box). Line graphs indicating the cumulative fraction of RNA-Seq read area (green) and read count (red) covered at each read coverage level are superimposed on the barplot, with the scale shown on the right. (D) Kernel-density plot of probe intensities for high- and low-coverage probe groups from (A), as indicated.
doi:10.1371/journal.pbio.1000371.g001

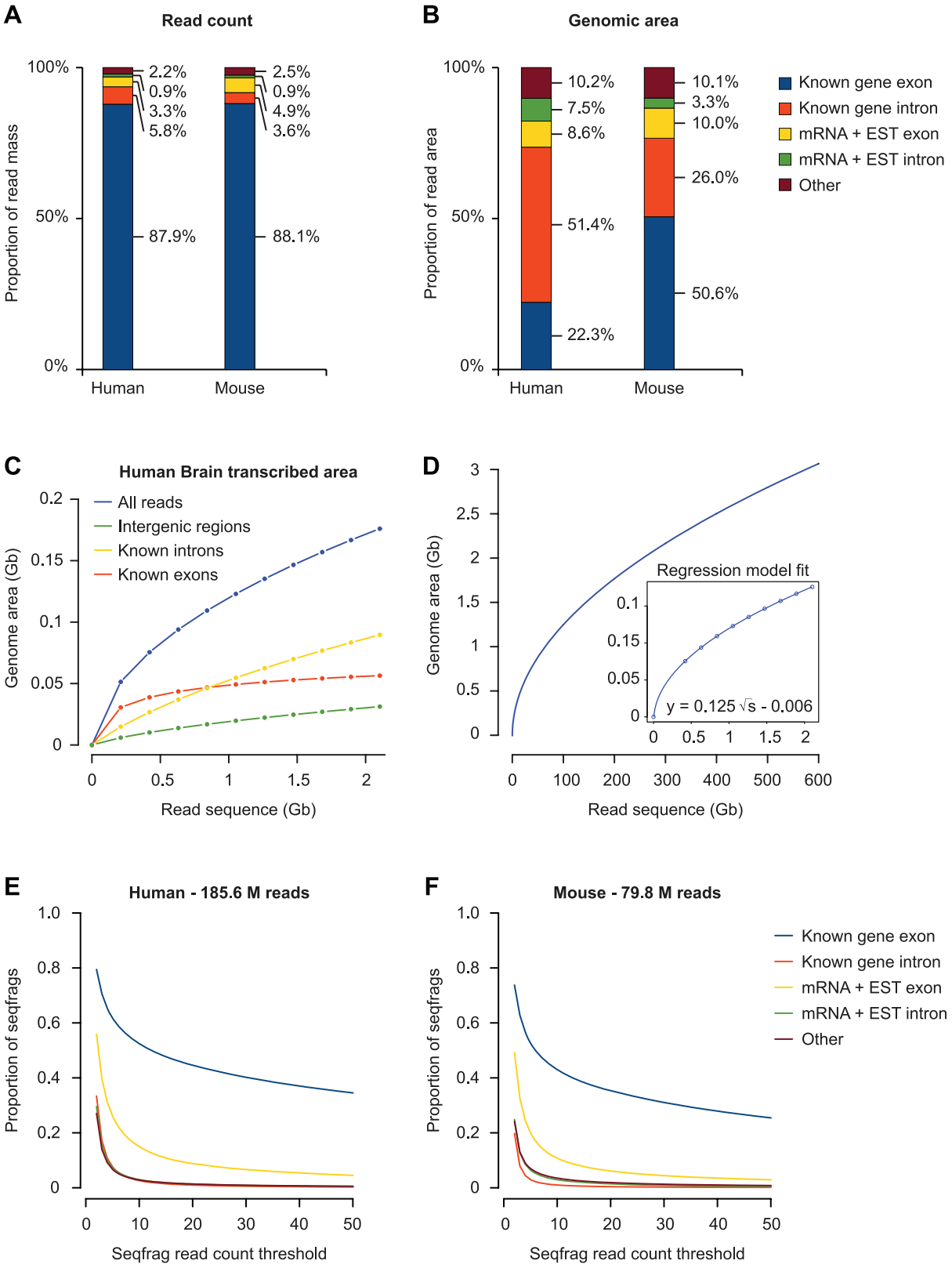


Figure 2. RNA-Seq read mapping overview. (A) Proportion of reads with a unique match in the genome mapping to known genes, mRNAs, and spliced ESTs. Reads were pooled across all human or mouse RNA-Seq samples and sequentially matched against a non-redundant set of known genes, mRNA, and spliced EST data. Any remaining reads were classified as "other." (B) Same as in (A) but considering the total amount of transcribed genomic area, rather than read count. (C) The relationship between the RNA-Seq read depth and the transcribed area in the genome for human brain RNA-Seq reads, based on 50.2 million reads pooled from the three independent samples that were assayed separately. The total transcribed area is indicated for all reads, as well as those that map to known exons, known introns, and intergenic regions. (D) Extrapolation of transcribed genomic

area at increasing read depths, based on the distribution of all reads in (C). The model fitted on the uniquely mapped reads is shown in the inset. (E, F) Cumulative fraction of seqfrags as a function of the number of reads mapped to each seqfrags in the combined set of human and mouse samples, respectively.
doi:10.1371/journal.pbio.1000371.g002

appear as transcripts. However, the fact that such pervasive transcription would only be detected at sequencing depths more than two orders of magnitude above current levels suggests that these transcripts may largely be attributed to biological and/or technical background. Indeed, the vast majority of intergenic and intronic seqfrags have very low sequence coverage (Figure 2E, 2F), exemplified by the fact that 70% (human) to 80% (mouse) of the transcribed area in these regions is detected by a single RNA-Seq read in only one sample, much of which is consistent with random placement (see below).

The low coverage and ubiquitous character of the intronic seqfrags suggests that they may represent random sampling from partially processed or unprocessed RNAs. We also note that 4.5% of all mapped (non-unique) human RNA-Seq reads correspond to rRNAs and sn(o)RNAs, suggesting that the PolyA+ selection did not fully exclude RNAs that are not polyadenylated. Alternatively, some of these transcripts may be polyadenylated under normal conditions, or they could correspond to degradation intermediates [39]. We note that, as the number of reads increases, the amount of transcribed area in intergenic regions increases at a much lower rate than in intronic regions (Figure 2C), even though intergenic regions make up a larger proportion of the human genome (1.7 Gb compared to 1.3 Gb for introns), further supporting the notion of random sampling of introns. In the complete set of uniquely mapped human brain RNA-Seq data, intergenic reads appear 3.8-fold less often than reads in intronic regions. In contrast, the cumulative read coverage is much higher for mRNA and EST exons than it is for either introns or intergenic regions (Figure 2E, 2F), indicating that many mRNAs and ESTs likely constitute valid transcripts that are not currently annotated in the three major genome databases. In summary, even though the genome may be randomly transcribed at very low levels, the vast majority of sequence reads in PolyA+ samples corresponds to known genes and transcripts, arguing against widespread transcription to the extent reported previously.

Most Intergenic Transcripts Are Adjacent to Known Genes

We next sought to gain further insight into the nature of dark matter seqfrags, focusing mainly on intergenic regions to avoid possible interference from unprocessed RNAs in introns. Potential sources of seqfrags in intergenic regions include 5′ and 3′ extensions of known genes, aberrant termination products, pasRNAs, and novel genes. We therefore began our characterization of intergenic seqfrags by examining their relationship to neighboring genes. In both human and mouse PolyA+ RNA-Seq data, we observed that the average read density in intergenic regions is dramatically higher near the starts and ends of annotated genes (Figure 3A) and can extend up to a distance of ~10 kb from both the transcription start and ends. We also observed bias towards genes in our tiling array analysis (unpublished data), as did a previous analysis using tiling arrays [8], but this study found the bias to be equal between 5′ and 3′ ends. In RNA-Seq data, the effect is stronger at the 3′ compared to the 5′ end of genes. Most transcripts at 3′ ends are consistent with alternative cleavage and polyadenylation (APA) site usage and unannotated UTR extensions of genes [16] or 3′ associated RNAs [12], rather than new exons, since in our splicing analysis (see below) we found very few instances of 3′ intergenic seqfrags linked to new 3′ exons (unpublished data). The increased number of transcripts at the 3′ end of genes is consistent with observations that RNA polymerase II can remain associated with DNA for up to 2 kb following the annotated ends of known mRNAs [40].

To determine the strand of origin of the positionally biased intergenic transcripts and to assess whether this bias was limited to PolyA+ RNA, we examined additional available sequencing-based transcriptome datasets. These included strand-specific RNA-Seq data from human rRNA-depleted whole brain and universal reference RNA [35], as well as from mouse brain PolyA+ [41] and rRNA-depleted total RNA (NCBI short read archive, SRX012528). We also incorporated data from CAGE-tag [12] and Paired-End diTag (GIS-PET) sequencing studies [42], which specifically targeted transcript ends. In all these datasets we find that most reads originate from known exons (Table S3), and among intergenic reads we find the same striking increase in read frequency in intergenic regions proximal to genes (Figure 3B, Figure S1, Table S3) as in PolyA+ samples. The enrichment of CAGE tags is consistent with peaks found at both the 5′ and 3′ ends of genes [12], and the majority of transcripts at the 3′ end of genes are in a sense orientation relative to the neighboring genes (Figure 3B). While CAGE tags are also enriched at 3′ ends of genes in the same orientation, the effect is less pronounced compared to RNA-Seq reads, suggesting that a significant number of transcripts in these regions result from alternative termination of protein-coding genes. Transcripts in intergenic regions flanking TSSs are approximately equally distributed between the sense and antisense strand (Figures 3B, S1A, and S1B), consistent with divergent transcription from promoter regions [12,20–24], as well as unannotated 5′ transcript ends.

To examine the relationship between genes and gene-associated transcripts in greater detail, we next determined whether the increased sequence coverage of seqfrags in intergenic regions flanking genes correlated with the coverage of genic trancripts across the 11 human PolyA+ RNA-Seq samples (the same analysis could not be done for the mouse data, as the number of available samples was too low to reliably estimate correlations). To this end, we first identified intergenic seqfrags by merging overlapping RNA-Seq reads from all human samples and then determined the sequence coverage for seqfrags and genes in each sample. Figure 3C shows that the correlation in coverage between intergenic seqfrags and neighboring genes is much higher than it is for randomly selected genes, indicating that expression in intergenic regions is positively associated with that of the flanking genes. This effect is strongest up to a distance of 10 kb from the gene but persists to a lesser degree over larger distances (Figure 3D). After setting a threshold of $p<0.05$, based on how often the correlation coefficient between a given seqfrag and neighboring gene was expected to occur at random (Materials and Methods), we find a significantly increased correlation with intergenic seqfrags for 2,970 annotated genes, 934 of which remain after multiple testing correction (Table 1). Consistent with the increased read frequency at 3′ ends of genes, the number of genes with correlated intergenic seqfrags at the 3′ end is 3-fold greater than at 5′ ends of genes (Table 1). Many of the correlated seqfrags at 3′ ends are directly adjacent to the annotated genes (see Figure 3E for a representative example), adding further support to our hypothesis that many of these transcripts are linked in their expression. Additionally, we found a small number of extensions at

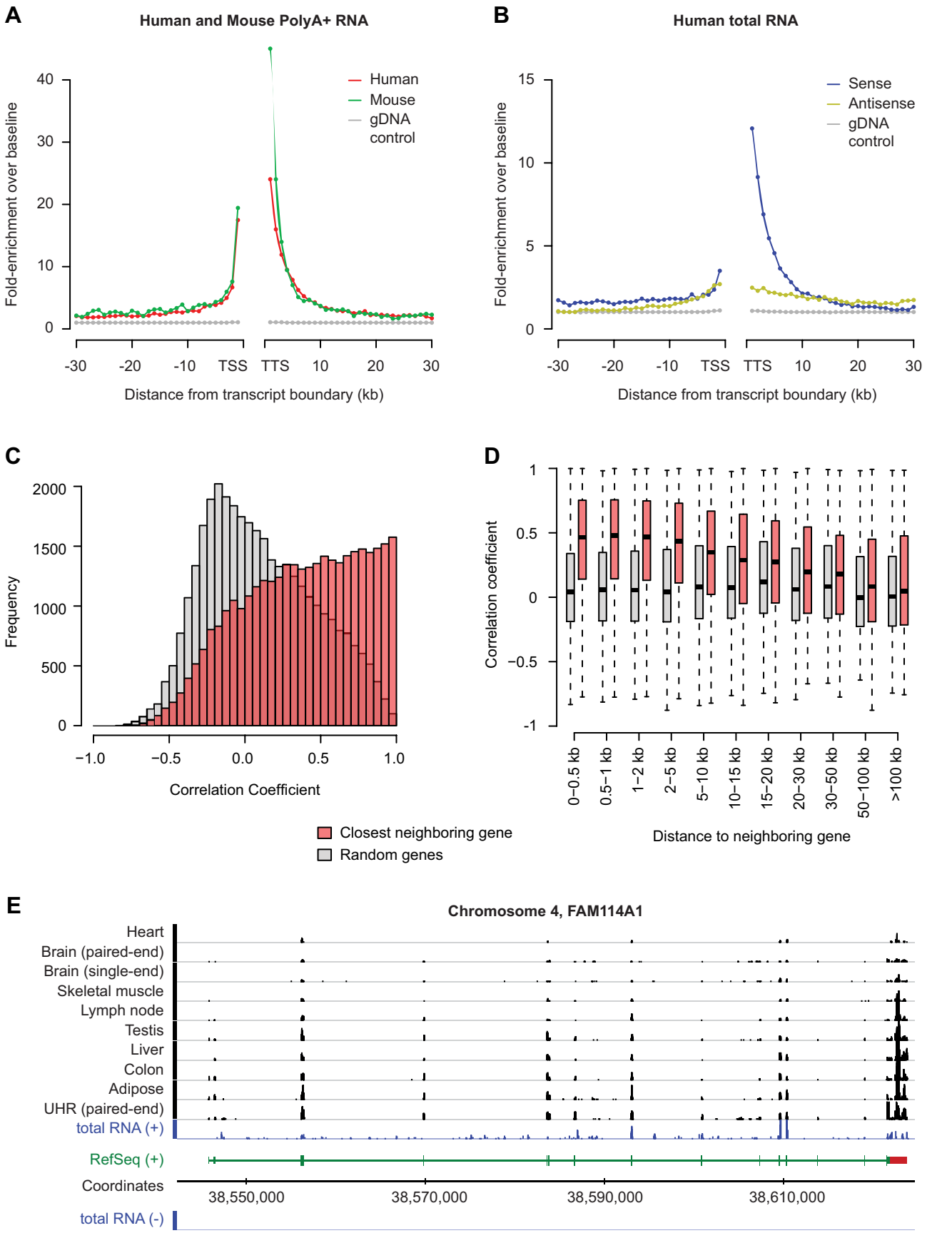

A
Human and Mouse PolyA+ RNA
Fold-enrichment over baseline
Human
Mouse
gDNA control
Distance from transcript boundary (kb)
B
Human total RNA
Sense
Antisense
gDNA control
C
Frequency
Correlation Coefficient
Closest neighboring gene
Random genes
D
Correlation coefficient
Distance to neighboring gene
E
Chromosome 4, FAM114A1
Heart
Brain (paired-end)
Brain (single-end)
Skeletal muscle
Lymph node
Testis
Liver
Colon
Adipose
UHR (paired-end)
total RNA (+)
RefSeq (+)
Coordinates
38,550,000
38,570,000
38,590,000
38,610,000
total RNA (-)

Figure 3. Intergenic expression is positionally biased towards known genes. (A) Relative enrichment of RNA-Seq read frequency in intergenic regions as a function of the distance to 5′ and 3′ ends of annotated genes in the human (red) and mouse genomes (green). The distribution in genomic DNA-Seq reads from HeLa cells [42] is shown as a control (gray). All intergenic regions in the human and mouse genomes were aligned relative to the annotated transcription start (TSS) or termination (TTS) sites of flanking genes. The robust average number of reads per 10 million uniquely mapped reads across all samples was then determined in 1 kb segments (RPKB) from the TSS or TTS, up to a distance of 30 kb, and the relative enrichment ratio in each segment was calculated by dividing by the median RPKB at distances more than 30 kb away from genes (baseline). Robust averages were calculated after removing the top 0.5% outliers, to avoid very highly expressed regions from having a disproportionate effect. (B) Same plots as in (A) for the combined reads from total RNA samples taken from human brain tissue and a universal human reference sample [35], uniquely mapped to the sense (blue) or antisense strand (yellow) relative to the neighboring gene region. (C) Histogram showing the distribution of correlation coefficients (red) between the read coverage in intergenic seqfrags and the nearest neighboring gene, across 11 human RNA-Seq samples. Read coverage was calculated as the number of reads per base per 10 million RNA-Seq reads across seqfrags and exonic regions of neighboring genes. Correlation coefficients were only calculated if the number of reads mapping to seqfrags and neighboring genes was greater than 10 in at least five out of eleven samples. The background distribution of correlation coefficients between seqfrags and randomly selected genes that met these thresholds is shown in comparison (gray). (D) Boxplot showing the correlation between the read coverage of intergenic transcripts and closest neighboring genes (red) or random genes (gray) across 11 human RNA-Seq tissue samples, as a function of their distance. (E) Representative example of intergenic transcription directly adjacent to the 3′ end of FAM114A1. The region with significant correlation is indicated by a red box. Mapped read coverage for the PolyA+ (black) and total RNA (blue) samples was standardized on a sequencing depth of 10 million reads and plotted in graphs scaled from 1- to 25-fold coverage.
doi:10.1371/journal.pbio.1000371.g003

larger distances, which are consistent with unannotated novel 3′ and 5′ exons (unpublished data and see below).

The total number of genes with correlated 5′ and 3′ intergenic seqfrags is likely underestimated in our analysis, as a minimum number of sequence reads in each sample are needed to calculate a correlation coefficient. Many transcribed intergenic regions detected at very low coverage had to be excluded from the correlation analysis, even though these low coverage regions are clearly enriched in regions flanking known genes (Figure S2). Consequently, some positional bias is still observed after removing the regions identified in this analysis (unpublished data), and correlated transcription in regions flanking genes is likely far more widespread. This is particularly relevant because while the 10 kb flanking regions make up only ~18% of the total intergenic area, they account for as much as 78% of the intergenic reads in human and mouse PolyA+ RNA. The same trend holds true for CAGE and GIS-PET datasets, as well as RNA-Seq datasets from rRNA-depleted human total RNA (Table S3). Although gene-flanking regions in rRNA-depleted mouse brain total RNA accounted for only 30.7% of intergenic reads, further inspection revealed that most of the reads outside these regions were linked to a small number of seqfrags (21) with excessive read counts (>10,000) confined to a small area (5 kb). This strongly suggests that there are a very small number of unannotated specific transcripts expressed at high levels, and after excluding these outliers, 71.1% of intergenic reads are found near genes (Table S3). The majority of intergenic dark matter transcripts are therefore linked to annotated protein-coding genes, either as extended transcripts or separate noncoding transcripts such as pasRNAs.

Table 1. Human transcripts with significantly correlated 3′ and 5′ seqfrags.

		Transcripts		Seqfrags	
Correlation Cutoff	**Category**	**Counts**	**Fraction**	**Counts**	**Fraction**
$p \leq 0.05$	All	2,970		6,109	
	3′ end	2,074	69.8%	4,612	75.5%
	5′ end	994	30.2%	1,497	24.5%
FDR ≤0.05	All	934		1,474	
	3′ end	698	74.7%	1,145	77.7%
	5′ end	251	26.9%	329	22.3%

doi:10.1371/journal.pbio.1000371.t001

Intergenic Regions Harbor a Limited Number of Novel Transcripts

Even when combining RNA-Seq data from all human or mouse tissues, read coverage in intergenic regions is very low (Figure 2B, 2C). To determine whether intergenic seqfrags are the result of low-level random background initiation, or whether they instead derive from a limited set of unannotated transcripts, we investigated the RNA-Seq read distribution in these regions. If the low-coverage intergenic seqfrags are indeed due to a uniform level of background initiation, reads should be spread evenly and the number of reads per kb of intergenic sequence should follow a random (Poisson) distribution. Given the observed transcriptional bias in regions flanking genes, we only considered intergenic regions that were at least 10 kb away from annotated genes (corresponding to ~82% of all intergenic sequence). These trimmed regions account for 0.8% of the total number of reads in the human PolyA+ RNA-Seq data (1.64% for mouse), with an average coverage that is 9.4-fold lower than in intronic regions (3.3-fold for mouse). We find a clear departure from a random distribution in the trimmed intergenic regions of both species (Figure 4A, 4B), including several thousand loci with greater than 20 reads, which should not occur under our null hypothesis. We also independently assessed seqfrags that are supported by only a single RNA-Seq read in one tissue (''singletons''), which account for ~70% of transcribed area in the trimmed intergenic regions in the human and mouse genomes. The distribution of singleton seqfrags is much closer to the random distribution (Figure 4D, 4E), although some deviation still persists for these low-coverage regions. To exclude that our observations are due to an inherent bias in cDNA library amplification or sequencing, e.g., due to GC content, we repeated the same analysis for an equal number of genomic DNA-Seq reads from HeLa cells [43] or a pool of human sperm DNA from four donors [44]. Both of these datasets were similarly generated on an Illumina genome analyzer and closely follow a random distribution (Figure S3). Taken together, these results indicate that while most reads >10 kb away from annotated genes are placed in a way that resembles random distribution across the genome, some have a non-random character, including several thousand regions with high read coverage that may be derived from unannotated novel transcripts.

To estimate the proportion of intergenic regions transcribed above background levels, we selected all 1 kb regions with a significantly higher read count compared to the random distribution ($p<0.05$) for all reads, or singleton reads only. At the lower thresholds based on singleton read frequencies, 3.0% (39.1 Mb) and 0.9% (11.4 Mb) of trimmed intergenic regions

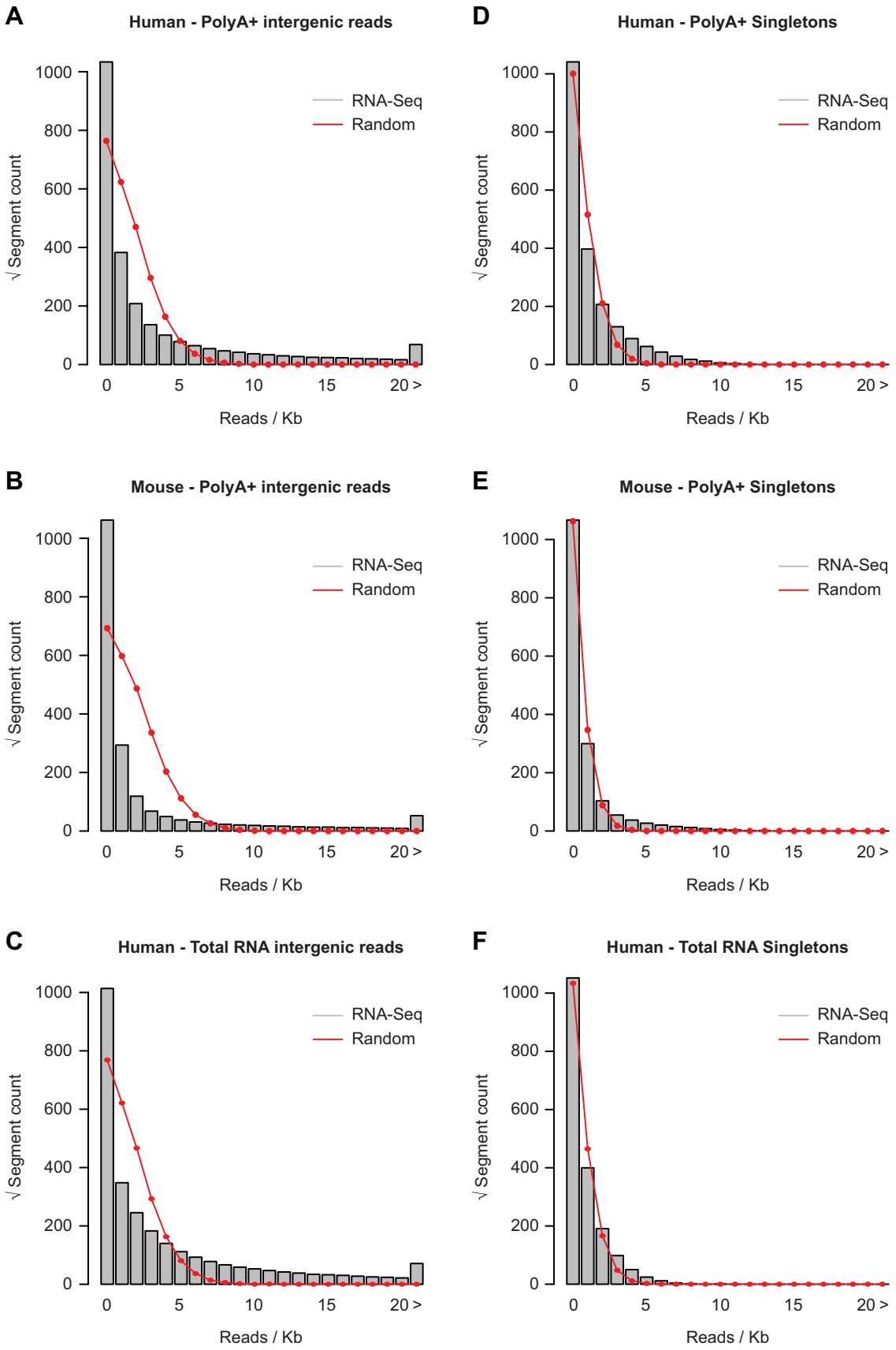


Figure 4. Evidence for specific expression in intergenic regions. Rootograms of the distribution of the number of the total number of RNA-Seq reads per kb of trimmed intergenic sequence for the combined (A) human PolyA+, (B) mouse PolyA+, and (C) human total RNA sequence data

(gray bars), in comparison to the expected random distribution for the same number of reads (red lines). Ten kb intergenic regions flanking known gene annotations were excluded from the analysis. (D, E, and F) Same as (A), (B), and (C), but considering only intergenic transcribed regions with single-read coverage (singletons). The derived random distribution was adjusted accordingly.
doi:10.1371/journal.pbio.1000371.g004

contain transcripts in the human and mouse genomes, respectively, decreasing to 1.2% (15.8 Mb) and 0.42% (5.25 Mb) at the more stringent thresholds. The increased area in the human compared to the mouse genome is consistent with the broader range of tissues assayed by RNA-Seq. The fraction of trimmed intergenic regions with significantly increased read counts is higher in human total RNA compared to PolyA+ RNA (Figure 4C, 4F): 4.1% (53.9 Mb) or 2.5% (32.8 Mb) at the lower and higher stringency levels, respectively. Considering that the total RNA sequence data was derived from a smaller sample set, this suggests that there are additional unprocessed and/or noncoding transcripts in intergenic regions not detected in PolyA+ RNA.

We also applied an additional threshold to identify putative novel exonic regions in the trimmed intergenic areas, selecting for seqfrags with a PolyA+ RNA-Seq read count greater than or equal to that of the top 5% of seqfrags detected in known introns (6 reads for human and 4 for mouse). At these thresholds we find 16,268 potentially "exonic" seqfrags in human (spanning 2.5 Mb) and 11,533 in mouse (spanning 0.66 Mb), which account for 56.9% and 87.4% of the reads in the trimmed intergenic regions in each organism, respectively. The area covered by the putative exonic seqfrags is 3.8% of the total area covered by seqfrags overlapping known exons in the human genome and 1.4% for the mouse genome. The putative exonic seqfrags tend to be well conserved at the sequence level compared to a random selection of intergenic sequences (Figure 5A, 5B), as judged by PhastCons conservation score based on multiple alignments among 18–22 mammalian genomes. This is significant, considering that the overall conservation for intergenic and intronic reads is close to random (Figure 5C, 5D). Taken together, our results show that a limited number of conserved novel exonic seqfrags can explain the majority of intergenic transcript mass detected in PolyA+ RNA, with a small proportion of low-level transcripts over a broad area that may be due to random initiation events.

Global Splice Junction Analysis Identifies New Transcript Structures

We next attempted to identify novel transcript structures by detecting splice junctions between transcribed regions in the genome using Tophat [45]. Tophat uses a two-stage approach that first aligns unspliced RNA-Seq reads to the genome to identify transcribed areas, which are then examined in the second stage to identify junction sequences spanning all possible 5′ and 3′ combinations of these regions, using the reads that could not be mapped in the first stage. The main advantage of this approach is that it does not require a predefined set of annotated exons and it can therefore identify splicing between unannotated regions of the genome. Moreover, as the analysis takes the canonical splice junction donor and acceptor sites (GT-AG) into account, it is possible to determine the strand of origin for each junction, despite the fact that the PolyA+ RNA-Seq data used in this study were not generated in a strand-specific manner. We restricted our analysis to human samples, since we found the reads in the mouse dataset to be too short to reliably detect junction sequences.

Overall, we found 160,516 unique splice junctions in the 11 PolyA+ human RNA-Seq samples, 151,708 (94.5%) of which can be classified as "known," meaning that they span any two exons within a single annotated transcript (Table S4). The remaining 8,808 novel junctions involved a single known exon or spanned two unannotated regions in the genome. In total, we could detect 57.8% of all exons in the combined set of gene annotations by at least one junction. Only 300 junctions bridged exons between transcripts, and almost all mapped to tandem-repeated regions in the genome (Table S5). Considering the high degree of sequence similarity between the repeated regions, some of these are presumably due to mapping inaccuracies. A significant proportion of bridging junctions (47%, 25% with confirmed deletions) also overlap regions with validated copy number variations (CNVs) that are common in the general population [46], suggesting that others may result from gene fusions following deletion events. These findings further argue against pervasive transcription to the extent reported in previous studies.

We assessed the false positive rate in the detected junctions by randomizing the sequences of potential splice junction reads and determined it to be 0.054% for paired-end reads and 2.7% for single-end reads (see Materials and Methods). The higher accuracy for paired-end reads demonstrates the considerable advantage of using longer reads to accurately assess splice junctions. Indeed, we found that the shorter 32 mer reads are particularly sensitive to false positive detections due to the presence of low-complexity regions and PolyA/T repeats, and we therefore applied additional filtering steps to exclude the affected junctions (see Materials and Methods for details). The longer read lengths of the paired-end compared to single-end RNA-Seq samples, combined with a 4-fold increase in sequencing depth, also resulted in a more than 3-fold higher splice junction detection rate.

The fact that short RNA-Seq reads typically cover only a single junction between exons makes it difficult to determine which combinations of alternative splice junctions correspond to transcripts observed in vivo. We therefore instead focused on identifying transcriptional units (TUs) that represent the aggregate assembly of all connected splice junctions. Thus, a completely reconstructed TU for an annotated gene will comprise the full complement of exonic regions, though these may be used in different configurations in alternatively spliced transcripts. Splice junctions were considered connected if they were directly adjacent to each other on the same strand, arranged in a head-to-tail configuration, and (i) the "facing" junction ends overlapped, or (ii) the complete region between facing splice junctions was transcribed, or (iii) facing junctions were within a distance of 200 bp (i.e., the approximate average exon size).

The vast majority of TUs we identified (91.2%) overlap with at least one exon of an annotated gene (Table 2), and 92.1% of exons in these TUs overlap known gene annotations (Table 3). We also detected 3,451 unannotated internal exons in 2,720 genes, as well as 723 and 370 unannotated 5′ and 3′ exons, affecting 544 and 290 genes, respectively. Among the TUs that are not connected to known gene annotations (i.e., independent TUs), 1,259 map to intergenic regions, the majority of which (82.6%) consist of a single junction. Only a minor fraction of independent TUs (4.8% of the total number of TUs) overlap genic regions on the sense or antisense strand. As it is possible that additional rare splice junctions are not detected in our analysis, some independent TUs overlapping genes in the sense direction may yet turn out to be connected to the gene they overlap. The majority of novel exons in the reconstructed TUs overlap with exons from the UCSC mRNA and spliced EST tracks (Table 3), providing further evidence that

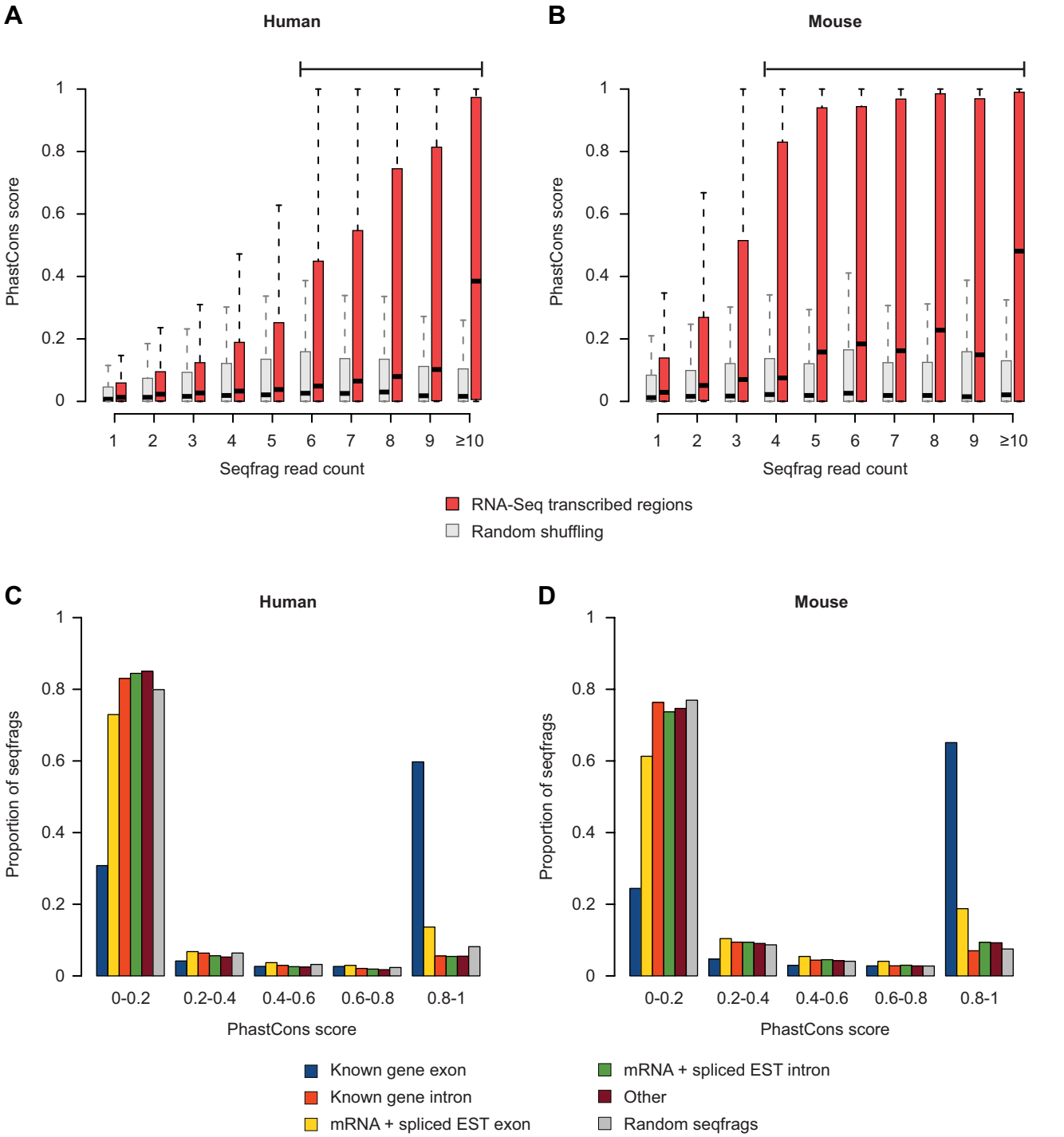


Figure 5. Seqfrags with read counts above background are conserved at the sequence level. Distribution of maximum PhastCons conservation score measured across seqfrags mapping to trimmed intergenic regions in the pooled (A) human and (B) mouse RNA-Seq samples as a function of read coverage (red). PhastCons scores were obtained from the UCSC genome browser and reflect the degree of conservation in multiple alignments of the human and mouse genomes with 18 and 20 other mammalian species, respectively. Conservation scores obtained from a random shuffling of seqfrag positions within trimmed intergenic regions are shown in gray for comparison. (C and D) Bar plots indicating the PhastCons score distribution for seqfrags mapping to different genomic regions. The bars are color-coded according to the class of seqfrags (legend), with the score distribution for randomly mapped seqfrags shown in gray.
doi:10.1371/journal.pbio.1000371.g005

they are derived from true splicing events. A small number (73) further overlap exons predicted by Wang et al. [16], which were derived from an analysis of splice junctions associated with computationally predicted exons. Taken together, our findings confirm that the vast majority of spliced transcripts in PolyA+ RNA are linked to known gene annotations and argue against widespread interleaved transcription of protein-coding genes in the human genome. The full set of TUs and junctions has been made

Table 2. Overview of transcript units identified in human RNA-Seq samples.

	Transcript Units		Breakdown by # of Exons		
Category	**Count**	**Fraction**	**Exons**	**Transcript Count**	**Fraction**
Exon overlap with known gene	29,029	91.2%	2	9,546	32.9%
			3	4,414	15.2%
			>3	15,069	51.9%
Non-exon gene overlap, sense strand	475	1.5%	2	422	88.8%
			3	29	6.1%
			>3	24	5.1%
Non-exon gene overlap, antisense strand	1,055	3.3%	2	927	87.9%
			3	94	8.9%
			>3	34	3.2%
Intergenic	1,259	4.1%	2	1,040	82.6%
			3	168	13.3%
			>3	51	4.1%

doi:10.1371/journal.pbio.1000371.t002

available on our supplementary website (http://hugheslab.ccbr.utoronto.ca/supplementary-data/hm_transcriptome/).

Characterization of Novel Exons and Multi-Exon Transcript Units

To further characterize the 4,544 novel exons connected to existing transcripts, as well as the 2,789 novel independent TUs (i.e., multi-exon transcripts), we assessed their expression levels, degree of conservation, and coding potential. As expected, novel exons detected as part of TUs that overlap annotated transcripts show evidence of increased conservation compared to randomly positioned exons (Figure 6A). Consistent with our analysis, a significant proportion of these exons overlap with Exoniphy predictions of evolutionary conserved protein-coding exons [47], most notably for novel 3′ (20.5%) and 5′ exons (18.9%) (Table S6A). The degree of overlap was significantly higher compared to random selections from intergenic regions ($p<0.0001$). In contrast, we observed little overlap with conserved RNA secondary structures as predicted by the Evofold [48] and RNAz algorithms [49] (Table S6A). We further examined whether the novel 5′ exons overlapped regions of open chromatin that typically mark regulatory regions [50–52] and which can be identified using digital DNase I hypersensitivity assays [53]. To this end, we used publicly available genome-wide data on DNase I hypersensitivity hotspots generated by the UW ENCODE group for 11 cell lines [54]. Consistent with their expected association with promoter regions, we found that the majority of novel 5′ exons overlapped the complete set of DNase I hypersensitivity zones identified by the HotSpot algorithm [53] in all 11 cell lines, as well as a more restricted set that only included hotspots found in both replicates for 8 cell lines ($p<0.0001$) (Table S6A).

Table 3. Exon overview for transcript units in human RNA-Seq samples.

	Exons		EST + mRNA Overlap	
Category	**Count**	**Fraction**	**Count**	**Fraction**
Known gene	174,693	94.2%		
New exon for known gene	4,544	2.5%	3,060	67.3%
Internal	3,451	1.9%	2,291	66.3%
External, 5′ end	723	0.4%	523	72.3%
External, 3′ end	370	0.2%	246	66.4%
Overlapping known gene	3,364	1.8%	2,223	66.0%
Sense strand	1,069	0.6%	609	57.0%
Antisense strand	2,295	1.2%	1,614	70.3%
Intergenic	2,821	1.5%	1,686	60.0%

doi:10.1371/journal.pbio.1000371.t003

Most of the novel exons are expressed at lower levels compared to the other exons of the gene they are linked to, which suggests that they derive from low-frequency alternative splicing events in the tissues we examined (Figure 6B). Indeed, we find direct evidence of alternative splicing for 2,526 (73%) of the novel internal exons and 2,370 of these (94%) are overlapped by junctions that bypass the novel exon. For novel exons at the 5′ and 3′ termini there is direct evidence for alternative splicing for 310 (43%) and 144 (39%), respectively. Among these are 145 cases of clear alternative promoter usage, where we find splice junctions between internal exons and the annotated promoter, as well as alternative junctions that link to a more distal promoter (Table S7). Figure 7A shows an example of one such alternative promoter for the SLC41A1 gene, encoding a solute carrier family protein.

In contrast to many of the transcribed fragments reported in tiling array studies, we find evidence for higher overall conservation for exons in independent TUs in intergenic regions, and those overlapping genes on the sense or antisense strand (Figure 6A). We assessed the coding potential of the independent TUs using a support vector machine classifier that incorporates quality measures of the available open reading frames (ORF) and blastx results [55]. Larger independent TUs with three or more exons show a general tendency to be coding: 60.8% in the case of intergenic TUs, and 70.8% and 41% for TUs overlapping genes on the sense and antisense strand, respectively. An example of a coding transcript with a translated ORF that has high sequence similarity to the elongation factor TU GTP binding domain is shown in Figure 7B. Some of the other translated TUs with clear similarities to existing proteins have stop codon mutations within the ORF, indicating that they could be pseudogenes.

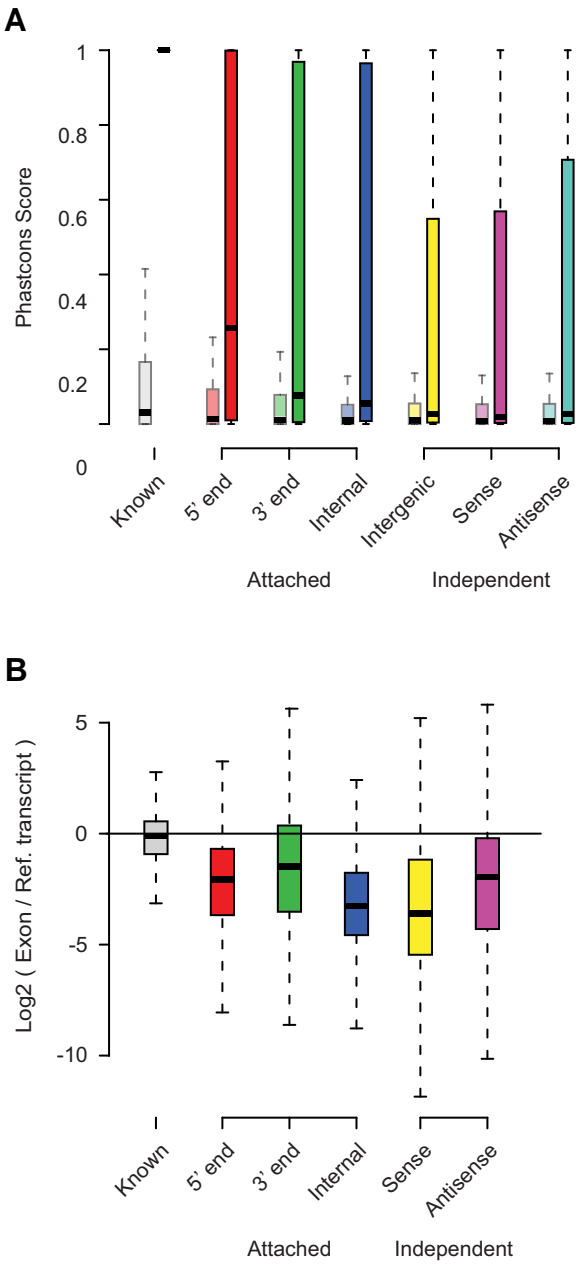


Figure 6. Conservation and usage of human TU exons. (A) The distribution of PhastCons scores for novel exons in each category as in (A) (darker bars), compared to the distribution of scores from the same set of exons after random reshuffling their positions in the genome (lighter bars). (B) Plot of the ratios between the read coverage of novel exons (calculated in RPB) and the genes they are associated with, either by overlap (sense or antisense) or as additions to known gene structures (5′ end, 3′ end, and internal). The ratios for predicted exons overlapping exons of known gene structures are shown in comparison.
doi:10.1371/journal.pbio.1000371.g006

None of the smaller intergenic TUs (containing only a single splice junction) were classified as coding. We note, however, that it is challenging to reliably detect the coding potential of small transcript fragments, and some of the TU fragments may in fact be part of larger coding transcripts. Indeed, when we extended the independent TUs by incorporating seqfrags overlapping the flanking junction sequences in the detected TUs, the proportion of potential coding transcripts increased to 8.3% for TUs overlapping gene regions on the antisense strand and to ~17% for TUs overlapping genic regions on the sense strand and intergenic TUs. Moreover, we find a significant overlap with Exoniphy predictions of coding exons, ranging between 10.5% for intergenic TUs and 21% for antisense TUs (Table S6B). Further investigation will be required to characterize these smaller TUs.

Even among the larger intergenic TUs with three or more exons, there is a subset of 116 transcripts that appear to be noncoding and are thus potential human lincRNAs, one example of which is shown in Figure 7C. The fact that we could not perform a comprehensive splice junction analysis in the mouse RNA-Seq data precludes us from making a detailed comparison with the previously identified mouse lincRNAs [19], however we do find a significant overlap between 95 of the mouse intergenic seqfrags with a read count above background and 30 of the lincRNA regions (Table S8A,B). The observation that there is little overlap (0%–1%) between reconstructed TUs and Evofold and RNAz predictions (Table S6B) suggests that most transcripts identified here do not fold into conserved RNA structures. In summary, our results reveal novel alternatively spliced exons and promoters in the human genome that are used at relatively low frequencies, as well as new lincRNA candidates.

Many Transcripts in Intergenic Regions Distal from Genes Are Short, Unspliced, and Associated with DNase I-Hypersensitive Regions

Only a small proportion (3.6%) of the 16,268 human intergenic seqfrags we identified with a read count above background were found to be part of TUs, which was surprising given that we could identify splice junctions for the majority of seqfrags in annotated exons. The lack of junctions connecting intergenic seqfrags cannot simply be explained by a reduced detection rate due to lower read counts compared to exonic seqfrags, as the proportion of intergenic seqfrags with detected junctions is consistently lower even at high coverage levels (Figure 8A). We therefore conclude that the majority of intergenic seqfrags are derived from unspliced single-exon transcripts. However, the remaining 15,646 human seqfrags that are not part of TUs are often spaced closely together, suggesting that they may be part of a single transcript, or are processed individually from larger precursor transcripts. Indeed, in many cases the intervening sequence between consecutive seqfrags is classified as transcribed when allowing reads mapping to multiple positions in the genome (see, for example, Figure 8B). When we group neighboring seqfrags with a maximum gap of 500 bp, 8,536 seqfrag clusters remain in human (7,976 of which show no evidence of splicing) and 5,506 in mouse.

We used the support vector machine classifier and Exoniphy predictions of coding exons, described above, to examine the coding potential of the unspliced intergenic seqfrags. Only 1.4% and 3.5% of human and mouse intergenic seqfrags with a read count above background overlap Exoniphy predictions, respectively (Table S8A,B). Moreover, out of the top 5% largest human intergenic seqfrags, ranging in size between 0.4 and 3.8 kb, only 12% were classified as coding. Taken together, these observations strongly suggest that the majority of the small intergenic seqfrags we identified are noncoding. As in the case of intergenic TUs, these transcripts also display little overlap with Evofold and RNAz regions.

The most striking property of the unspliced seqfrags is their strong association with open chromatin: 6,407 out of the 15,646 (40.9%) human intergenic seqfrags overlap with DNase I hypersensitivity hotspots identified in one of the 11 cell lines that were assayed, 3.4-fold more than would be expected by chance

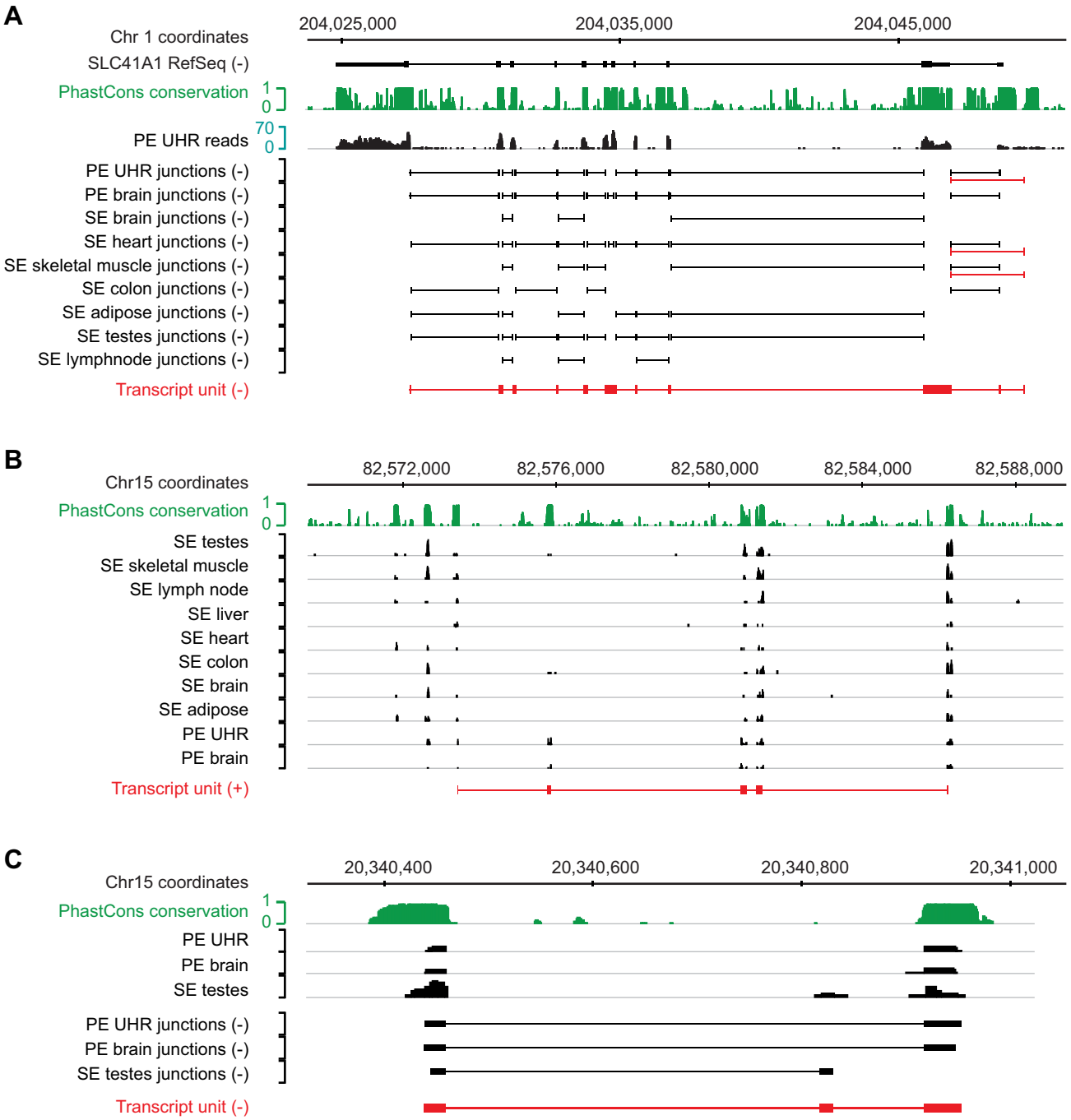


Figure 7. Examples of identified TUs. (A) Evidence for the presence of an alternative promoter at the human SLC41A1 gene. Splice junctions connecting to the alternative promoter region are indicated in red. Mapped RNA-Seq data for the UHR paired-end (PE) read sample is shown for reference (black). The PhastCons conservation track scores were based on multiple alignments of 28 vertebrates. (B) Protein-coding TU detected in an intergenic region on chromosome 17, with high similarity to the elongation factor Tu GTP binding domain. The two additional upstream transcribed regions may be part of the same transcript, though no junction sequences were detected. (C) Intergenic TU (red) detected on chromosome 15 based on junctions in the PE brain, PE UHR, and SE testes RNA-Seq samples.
doi:10.1371/journal.pbio.1000371.g007

(Table S8A). Figure 8C shows a clear enrichment in tags from hypersensitive sites for RA-differentiated SK-N-SH neuroblastoma cells across the full length of brain-expressed seqfrags. Moreover, the typical size of the unspliced seqfrags (median 111 bp) is smaller than that of the DNase I-hypersensitive regions (median 248 bp), and unlike coding transcripts and other ncRNAs, many of the seqfrags appear to be contained entirely within the DNase I-hypersensitive regions. We expect that the true number of seqfrags associated with DNase I hypersensitive regions may be larger, considering that the cell lines assayed only account for a small selection of the cell types represented in the tissues and cell types assayed by RNA-Seq. Thus, these analyses reveal the existence of thousands of small intergenic transcripts associated with open chromatin.

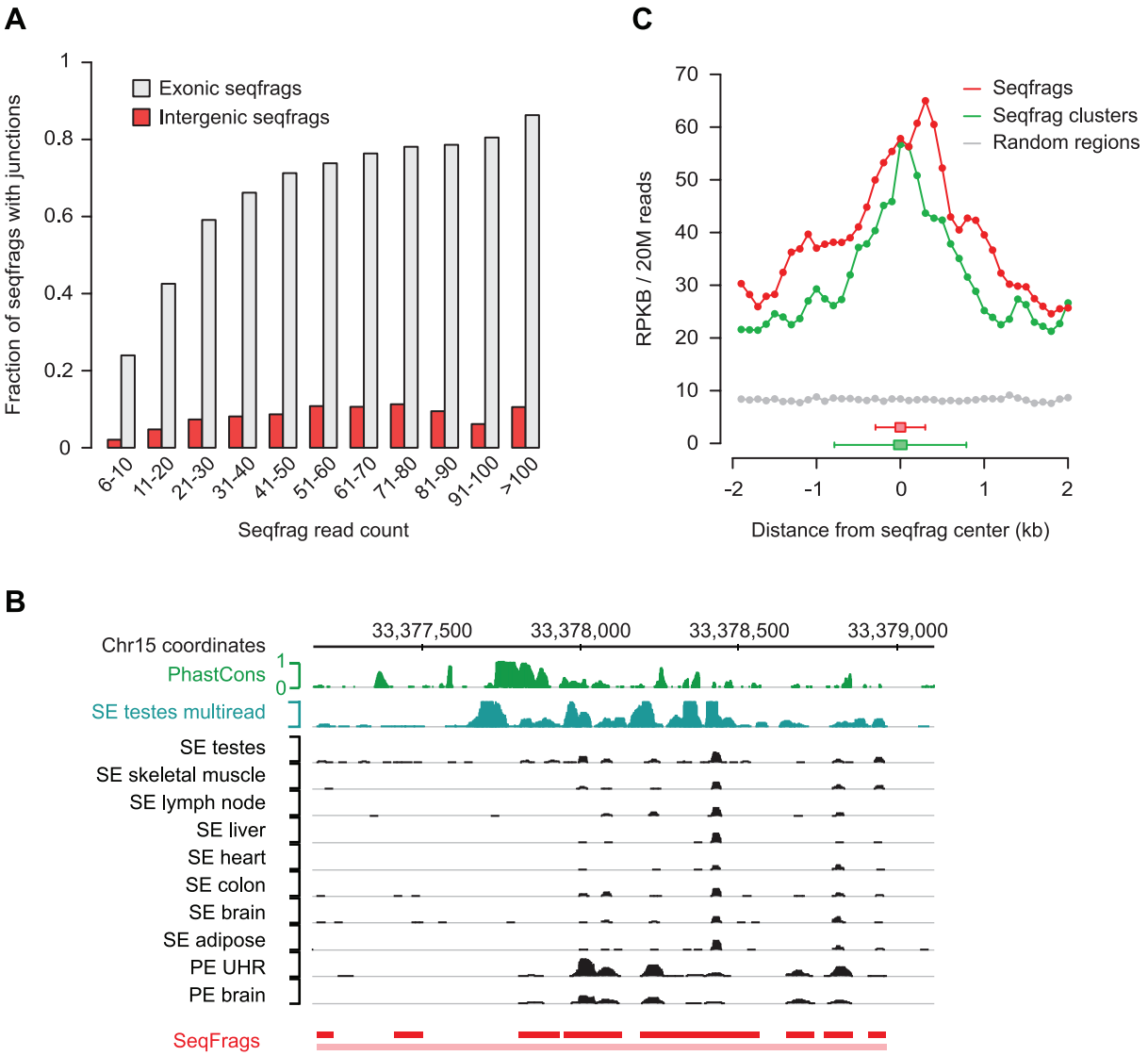


Figure 8. Most intergenic transcripts are unspliced and associated with open chromatin. (A) Relationship between read count and the fraction of seqfrags with at least one identified junction sequence for seqfrags in exonic (gray) or trimmed intergenic (red) regions. (B) Cluster of ubiquitously expressed seqfrags derived from uniquely mapped reads on chromosome 15. An additional track with multireads from SE testes RNA-Seq data (blue) shows that many of the uniquely mapped seqfrags are part of a larger, continuously transcribed region. (C) Digital DNase I hypersensitivity profiles in RA-differentiated SK-N-SH cells for 11,416 seqfrags (red) and 5,819 seqfrag clusters (green) expressed in human brain. Hypersensitivity is shown as the average density of in vivo cleavage fragment reads per kb (RPKB, normalized to 20 million reads) across all seqfrags or clusters, measured in 100 bp windows flanking the center position of each seqfrag or cluster up to a distance of 2 kb. The DNase I hypersensitivity at random positions in intergenic regions is shown as a control (gray). The box-and-whisker plots at the bottom of the graph indicate the median (box) and the 95th percentile (whiskers) of the seqfrag- (red) and seqfrag cluster size range (green).
doi:10.1371/journal.pbio.1000371.g008

Discussion

In contrast to earlier studies based on oligonucleotide tiling array analysis of RNA [1–9], GIS-PET [9], and RACE-tiling arrays [9], but consistent with other RNA-Seq studies [16,33,36–38], we find that the proportion of dark matter transcripts among polyadenylated RNA from a large variety of different tissue types is small. Our comparison between tiling arrays and RNA-Seq data from the same tissues indicates that tiling arrays are ill-suited to accurately detect transcripts expressed at low levels. The major fraction of nonexonic transcripts in RNA-Seq data is associated with known genes and includes thousands of new alternative exons and hundreds of alternative promoters. However, we do not find evidence for widespread interleaved transcripts as previously described [9]; virtually all exon-exon junctions detected correspond to junctions within the same gene. Aside from new exons, most of the transcripts that are within or proximal to known genes can be explained as pasRNAs or terminator-associated RNAs, pre-mRNA fragments, or by alternative cleavage and polyadenylation site usage. The relatively small fraction of seqfrags that are not associated with known genes corresponds strongly to DNase I-hypersensitive regions. Altogether, we propose that most of the dark matter transcriptome may result from the process of transcribing known genes. Pervasive transcription of intergenic regions as described in previous studies occurs at a significantly reduced level and is of a random character.

The intergenic regions that are transcribed above background consist of a mix of both coding and noncoding transcripts. In contrast to the extensive intergenic transcription reported in tiling array studies, we found relatively few transcripts in these regions (16,268 seqfrags expressed above background levels in human and 11,533 in mouse). These numbers may be smaller, as some adjacent seqfrags may be parts of a single transcript that contain regions with sequence mapping ambiguities, or they may be larger as more tissues and cell types are surveyed.

The fact that non-exonic transcripts do not overlap with Evofold or RNAz regions argues against widespread roles as structural RNA. The most compelling support that these transcripts may have an independent function comes from the fact that they overlap with DNase I hypersensitive regions and that, unlike the many transcripts found by tiling array studies and from deep sequencing of subtracted cDNA libraries [11], the transcripts found by RNA-Seq show a significantly higher degree of conservation between species. We note, however, that these same two properties are consistent with low-level transcription from enhancers. Indeed, in yeast, it is known that placement of a strong activating transcription factor binding site in random regions of the genome results in the formation of a promoter [56]. Thus, single-exon intergenic seqfrags may represent the analog of pasRNAs for enhancers.

Our findings are based primarily on analysis of PolyA+ enriched RNA; however, our conclusions are corroborated by CAGE tags, GIS-PET, and RNA-Seq analysis of rRNA-depleted total RNA. Similar conclusions to ours were also reached in an independent RNA-Seq analysis of rRNA-depleted human total RNA (G. Schroth, pers. communication). It does not appear as if additional sequencing would substantially alter our conclusions, since coverage bias towards known exons increases with the number of reads. Moreover, while RNA-Seq analysis of PolyA+ RNA biases against very long and very short RNAs, this would not be expected to affect our ability to detect the widespread and pervasive transcription reported previously. Nonetheless, analysis of further tissues and cell types would be expected to identify additional intergenic ncRNA seqfrags that are more abundant but expressed in rare or specialized cell types. It is also likely that total RNA harbors additional transcripts not seen in PolyA+ enriched RNA and that are not evident in current total RNA-Seq analyses due to limitations in read counts.

A major remaining question is the possible function of the novel intergenic transcripts, if any. Undoubtedly, there are many functional ncRNAs remaining to be characterized [57]. However, we and others have emphasized that expression, conservation, and even localization and physical interactions of these RNAs do not constitute direct evidence for function [32]. Promoters and terminators are known to produce transcripts that appear to be associated primarily with the mechanics of gene expression and do not have known independent functions. To be conservative, a null hypothesis should perhaps be that novel transcripts—particularly those that are small and low-abundance—are a by-product rather than an independent functional unit [58]. Searching for phenotypes caused by genetic perturbation may be the most useful approach to disproving the null hypothesis.

Materials and Methods

Sample Sources

Total and PolyA+ samples for tiling array hybridizations from pooled human and mouse heart, liver, testis, and whole brain tissues were obtained from Clontech (Table S9). All human RNA samples were derived from tissues of individuals that suffered sudden death. The human whole brain PolyA+ RNA used for paired-end sequencing came from a Microarray Quality Control (MAQC) sample (Ambion) that consisted of a mixture of RNA from 23 Caucasian males. The PolyA+ selected universal human reference sample (Stratagene) consisted of pooled RNA from 10 human cell lines (Adenocarcinoma, mammary gland; Hepatoblastoma, liver; Adenocarcinoma, cervix; Embryonal carcinoma, testis; Glioblastoma, brain; Melanoma; Liposarcoma; Histiocytic Lymphoma, hystocyte; Lymphoblastic leukemia, T lymphoblast; Plasmacytoma, B lymphocyte).

Microarray Hybridizations

All RNA samples were DNase treated with 10 units of DNase I (Fermentas) per 50 ug of RNA prior to cDNA synthesis and purified with RNeasy spin columns (Qiagen) using a modified protocol that retains small RNAs <200 nt. Double stranded cDNA synthesis was done as previously described in Kapranov et al. [5]. Briefly, 9 ug of total RNA was reverse transcribed in a reaction that contained 1,800 units of SuperScript II enzyme (Invitrogen) and 83.3 ng of random hexamers and Oligo(dT) primers per ug of RNA. The cDNA was then used for second strand synthesis, after which the double-stranded cDNA (ds-cDNA) was purified using PCR purification columns (Qiagen) in combination with the nucleotide cleanup kit protocols. Following fragmentation and biotin labeling, 7 ug of ds-dDNA was hybridized per array.

Mapping of Genomic Coordinates for Tiling Array Probes

The Affymetrix Human and mouse tiling arrays version 2.0R were originally designed for the NCBI genome assemblies v34 and v33, respectively, and were remapped to more recent genome builds (v36 for human and v37 for mouse) using BLAT [59], not allowing for any mismatches in the alignments. A small number of probes mapping to multiple locations in the genome were assigned a position that would conserve probe order relative to the original array design. In cases where this was not possible, the position on the same chromosome nearest to the original probe location was selected, or a match was randomly selected if none could be found on the same chromosome. In total, 99.5% of probe sequences could be remapped to the new mouse genome assembly, and for the human arrays this number was close to 100%. Updated bpmap files are available on request.

Microarray Data Analysis

Arrays were scanned using an Affymetrix GeneChip scanner 3000 and raw probe intensities were obtained using the Affymetrix GeneChip Operating Software. Each array was quantile normalized against a reference genomic DNA hybridization using the Affymetrix Tiling Array Software v1.1 to obtain intensities corrected for probe sequence bias (Figure S4). The probe intensity data were further smoothed by calculating the pseudomedian of genomic DNA-normalized intensity values of probes that lie within a genomic sliding window around each probe [5]. The size of the sliding window was determined by the bandwidth parameter (BW) as follows: (2× BW) +1. Transcribed regions (transfrags) in tiling array data were selected as previously described [5], by joining positive probes together using three parameters: (i) an intensity threshold to select positive probes, (ii) the maximal distance (MAXGAP) that two neighboring positive probes can be separated by, and (iii) the minimal transfrag length (MINRUN). A range of BW, MAXGAP, and MINRUN parameter combinations were applied and used to assess precision and recall of exons in known transcripts.

Illumina Sequencing Datasets

Libraries for paired-end sequencing were prepared according to the manufacturer protocols. After selecting for cDNA fragments with a size distribution around 200 bp, 50 bp on both ends were sequenced in an Illumina Genome analyzer II. Single-end RNA-Seq data with a read length of 32 nt for PolyA+ RNA for 8 human tissues from individual donors (Adipose, Brain (2×), Colon, Heart, Liver, Lymph Node, Skeletal Muscle and Testis) were obtained from a previous study by Wang et al. [16]. Twenty-five mer single-end read data for PolyA+ RNA from three mouse tissues (Brain, Liver, Skeletal muscle) were taken from Mortazavi et al. [36]. Both literature datasets were produced following similar protocols that included a fragmentation step followed by a size selection for fragments of ~200 bp and sequencing on an Illumina Genome analyzer. All paired-end RNA-Seq data are available on our supplementary website (http://hugheslab.ccbr.utoronto.ca/supplementary-data/hm_transcriptome/).

Mapping of Unspliced RNA-Seq Reads to Reference Genomes

Single-end read RNA-Seq data were mapped to the NCBI human and mouse genome assemblies v36 and v37, respectively, using Seqmap v1.0.10 [60]. Several parameter settings were tested, and the maximum number of uniquely mapped reads (best unique hit) was obtained by restricting the read length to the first 25 bases and allowing for only one mismatch (Figure S5). These settings were subsequently used for all single-end read mappings. Paired-end reads from human brain and UHR samples were split and independently mapped using bowtie [61], selecting only the unique best hits from alignments that had a maximum of two mismatches in the seed sequence (first 28 bases) and an overall sum of mismatch phred quality scores no greater than 70. Single-end reads or tags from strand-specific datasets (Table S3, Figure S1) were also mapped using bowtie [61] to maintain strand information.

For the overlap analysis with known gene annotations, we combined the following tracks from the University of California Santa Cruz (UCSC) genome browser: UCSC known genes, Refseq genes, ENSEMBL genes, RNA genes, miRNAs, and snoRNAs (February 2009). In addition, mRNA and spliced EST tracks were obtained from the same source (September 2009) for a secondary mapping of seqfrags or sequence reads that did not match known gene annotations. Non-redundant sets of genes, mRNAs, and spliced ESTs were prepared by merging overlapping features, where the resulting exonic regions were defined as the union of exons in the source annotations and introns as the intervening regions between merged exons. For the calculation of the proportion of reads accounted for by each annotation category, reads were considered exonic if they partially or fully overlapped a merged exon, and intronic or intergenic if they were fully contained in these respective regions. The proportion of transcribed area was calculated by intersecting the genomic coordinates of continuously transcribed genomic regions (seqfrags) and the various genome annotation categories.

Conservation of RNA-Seq Regions

PhastCons [62] conservation scores for the human and mouse genomes were obtained from the UCSC website and were based on multi-species alignments of 18 (hg18-phastCons18way) and 20 (mm9-phastCons20way) placental mammals, respectively. Conservation scores were assigned to each seqfrag by taking the maximum PhastCons score in the genomic region covered by the seqfrag. For comparison purposes, a background score was determined for each seqfrag in the same manner, after reassigning seqfrags to random positions in the genome or within intergenic regions.

Correlation Analysis for Positional Bias

To calculate Pearson correlation coefficients between the expression levels of transcribed intergenic regions and the closest neighboring genes, overlapping mapped reads from all 11 human RNA-Seq samples were first merged into seqfrags. For each seqfrag, the nearest neighboring known transcript with read coverage in at least five RNA-Seq samples was then selected from the full set of transcripts in the UCSC known gene, Refseq gene, ENSEMBL gene, RNA gene, miRNA, and snoRNA tracks. In case multiple transcripts were found at the same distance (e.g., alternatively spliced transcripts), the transcript was selected that maximized the number of available data points for correlation analysis. The transcript expression levels in each tissue were defined as the median read coverage per kb of exon sequence and further adjusted for the difference in sequence coverage between RNA-Seq samples. Read coverage for intergenic seqfrags was determined analogously. Pearson correlation coefficients between transcript and seqfrag expression levels were only calculated if the read coverage for both the seqfrag and transcript were above zero in at least five of eleven samples, and all other intergenic seqfrags were removed from the analysis. The significance of the correlations was determined by comparing seqfrag expression levels to those of 1,000 randomly selected genes that met the same cutoff criteria. Nominal p values were defined as the proportion of random permutations where the correlation coefficient exceeded the observed correlation with the closest neighboring gene. Nominal p values were further adjusted for multiple testing by applying a Benjamini-Hochberg FDR correction [63] using the multtest R package from Bioconductor [64].

Assessment of the Intergenic Read Distribution

To determine whether RNA-Seq reads that map outside genes follow a random (Poisson) distribution, intergenic regions were divided into 1 kb segments and the total number of reads and the number of singleton reads in each segment was counted. Regions flanking genes up to a distance of 10 kb were excluded, as reads in these regions are more frequent, and correlated with known genes. For comparison, a random distribution was derived by sampling an equal number of uniquely mapped random reads with the same size distribution as the mapped RNA-Seq reads. To avoid a potential bias from the paired-end reads, we only mapped one of the reads in a pair. Comparisons between random and observed distributions were visualized in rootograms, which plot the square root of the number of segments as a function of the number of reads in each segment, allowing for a better assessment of differences at the tail of the frequency distribution.

Splice Junction Discovery

Analysis of novel splice junctions was performed using Tophat [45], which uses a detection method outlined in Figure S6A. Briefly, Tophat searches for splice junctions by first mapping RNA-Seq reads to the genome to identify "islands" of expression, which are equivalent to seqfrags. In contrast to the mapping of unspliced RNA-Seq reads described above, Tophat allows multiple genomic matches for each read (up to a maximum of 40 copies) during this mapping step. Each expression island is then considered a potential exon and used to build a set of potential splice junctions, taking into account the canonical splice donor and acceptor sites (GT-AG) within each island and a small flanking region of 45 bp. Subsequently, each possible pairing of neighbor-

ing junction sequences up to a specified distance (determined by the maximum allowed intron size) is compared to the set of "missing" RNA-Seq reads that could not be matched to the genome in the first mapping step to identify sequences that span junctions. Islands with high coverage are also examined for internal junctions, to account for the possibility that the intervening intronic region between two highly expressed exons is fully transcribed at lower coverage. Paired-end reads were analyzed with Tophat version 1.0.10, which features improvements in splice junction detection specific to paired sequencing data by taking the distance between read pairs into account. In contrast, single-end read data were analyzed using Tophat version 0.8.3, as we found that this version offered greatly improved sensitivity for shorter unpaired reads.

Splice junctions in paired-end read data were mapped allowing for a maximum intron size of 500 kb, which is sufficient to encompass 99.99% of all introns and 99% of all transcripts in the complete set of annotated transcripts described above. The minimum required read match size at each junction end (i.e., anchor size) was set to 8 nt. Finally, the minimum isoform fraction was set to 0.15 to suppress junctions that were supported by too few alignments relative to the junction exons. The isoform fraction was calculated as S/D, where S is the number of reads supporting each junction and D is the average coverage of the junction exon with the highest coverage [45]. Splice junctions in single-end read data were mapped using the full 32 mer read length, rather than the shorter 25 mer reads used for the mapping of unspliced reads. The Tophat parameter settings for single-end reads were the same as for the paired-end reads, with the exception that the minimum anchor size was set to 11 and the intron size was set to 20 kb (sufficient to bridge the length of 93.98% of all annotated introns and 53% of all transcripts). The adjusted parameters for single-end read data increased precision due to the shorter read lengths (Figure S7), at the expense of a somewhat reduced ability to detect long-range splice junctions. Finally, junctions with identical sequence that mapped to more than one genomic location in both the single- or paired-end RNA-Seq data were dropped from the analysis. Alternative splicing events were defined as junctions that shared the same start position with another junction but ended at a different position, or vice versa.

Estimation of False Positive Rate in Novel Splice Junctions

In order to estimate the proportion of false positives in the splice junction prediction, we adjusted Tophat to use a modified set of reads in the splice junction detection step. The initial read mapping stage to identify islands of expression was unchanged, but the sequence for the set of "missing" reads used in the second stage to detect splice junctions between islands was reversed (Figure S6A), resulting in a scrambled set of potential junction sequences with very similar sequence properties, in particular for low-complexity and repetitive regions. In addition, the pairing of reads in the paired-end dataset was randomized. With the modified sets of "missing" reads, 62 junctions were detected in the brain and 60 in UHR sample, corresponding to an estimated false positive rate of 0.054% for paired-end read samples at the selected analysis thresholds.

At 7.3%, false positive rates for single-end reads were significantly higher, consistent with the shorter read lengths. Further examination of junction sequences revealed an over-representation of PolyA and PolyT repeats in junction sequences of single- compared to paired-end read samples (Figure S8). We believe that enrichment of these repeats is due to a bias in mapping short reads sequenced from PolyA tails, and additional filtering steps were therefore applied to exclude junctions with a PolyA/T repeat size larger than 5. Moreover, any junction found to contain more than 20% low-complexity regions as assessed by the DUST algorithm (http://compbio.dfci.harvard.edu/tgi/software/) and repeatmasker (http://www.repeatmasker.org) was discarded. After applying these filters to the real and randomized junction set, the false positive rate for detection of splice junctions in the single-end read set was reduced to 2.7%.

Assembly of Splice Junctions into Transcript Units

TUs were defined as described in the main text. Facing splice junctions arranged in a head-to-tail fashion were first assembled into tissue-specific TUs if (i) splice junction ends overlapped (ii) the complete region between facing splice junctions was transcribed or (iii) if facing splice junctions were within a distance of 200 bp (same range as the average exon size) (Figure S6B). TUs were then combined across tissues where TUs with at least one overlapping exon were merged to create a non-redundant set. Exons were detected either partially, with junctions on only one side (e.g., 5′ and 3′ terminal exons), or completely, with supporting junctions defining boundaries on both sides.

Assessment of Coding Potential of Novel TUs

The coding potential of novel transcript fragments was assessed using a support vector machine classifier [55] that assesses the protein-coding potential based on several sequence features that incorporates quality assessments of the predicted ORF as well as BLASTX comparisons with the NCBI non-redundant protein database.

Significance Testing for Overlaps between Transcripts and Genomic Feature Sets

Statistical significance for overlaps between genomic feature sets (i.e., Exoniphy predicted coding exons [47], RNAz [49], and EvoFold [48] conserved RNA structures, DNase I hypersensitivity sites generated by the UW ENCODE group [54], and enhancer sets [65,66]) and exons in transcript units or significant seqfrags in trimmed intergenic regions was calculated by permutation analysis. In each permutation round, seqfrags or TU exons were assigned random positions within intergenic regions (for novel 5′ and 3′ exons connected to annotated genes), trimmed intergenic regions (for seqfrags in intergenic regions at least 10 kb away from genes), or introns (for novel internal exons for annotated genes, as well as exons in independent sense and antisense TUs). p values were defined as the proportion of times that an overlap count greater than or equal to the number of observed overlaps was found in 10,000 permutations. Coordinates of genomic feature sets were obtained from the UCSC genome browser or the original publications and mapped to the hg18 genome build using the UCSC LiftOver tool when needed.

Accession Numbers

Affymetrix tiling array data are available at GEO (record GSE19289).

Supporting Information

Figure S1 Positional bias towards known genes in other genome-wide transcription datasets. Relative enrichment of intergenic read/tag frequency near annotated genes in a variety of datasets including (A) strand-specific RNA-Seq of mouse brain PolyA+ RNA [41], (B) strand-specific RNA-Seq of mouse brain rRNA-depleted total RNA (SRX012528, NCBI short read archive), (C) Cap analysis of gene expression (CAGE) tags from 41 different human libraries [12], (D) CAGE tags from 145

different mouse libraries [12], and (E) Gene Identification Signature paired-end tags (GIS-PET) from two human cancer cell lines (MCF7 and HCT116) [42]. RNA-Seq reads and CAGE tags were mapped using Bowtie as described in the Materials and Methods section. For the GIS-PET datasets, mapped ditag positions for the hg17 version of the human genome were obtained from the original publication [42] and converted to coordinates in the hg18 assembly using the UCSC LiftOver tool (http://genome.cse.ucsc.edu/). Relative enrichment ratios of reads and tags in gene-flanking regions were calculated as described for Figure 3A and 3B.
Found at: doi:10.1371/journal.pbio.1000371.s001 (0.14 MB PDF)

Figure S2 Low-coverage intergenic expression is positionally biased towards known genes. Relative enrichment of read frequency for low-coverage transcribed regions in the pooled RNA-Seq sets as a function of the distance to 5′ and 3′ ends of annotated genes in the human (red) and mouse (green) genome. The distribution for genomic DNA-Seq reads from HeLa cells is shown as a control (gray). Low coverage regions were defined as seqfrags that were detected by only a single read in the combined human and mouse RNA-Seq sets. Relative enrichment ratios of reads and tags in gene-flanking regions were calculated as described for Figure 3A and 3B.
Found at: doi:10.1371/journal.pbio.1000371.s002 (0.12 MB PDF)

Figure S3 Intergenic genomic DNA-Seq reads are approximately randomly distributed. A sample of intergenic reads was selected from public DNA-Seq datasets (gray bars) from human sperm genomic DNA and HeLa cells [43,44] and used to draw distribution plots analogous to Figure 5 in the main text. The number of selected DNA-Seq reads in the complete or singleton sets was equal to the number of intergenic reads in the pooled human RNA-Seq dataset. The expected random distribution is indicated by a red line.
Found at: doi:10.1371/journal.pbio.1000371.s003 (0.14 MB PDF)

Figure S4 Genomic DNA normalization reduces intensity bias due to probe GC content. (A) Affymetrix tiling array image of a mouse testis PolyA+ RNA hybridization, showing the probe signal intensity in the top half and a heatmap of the GC content of the same probes in the bottom half. Lighter shades of gray and orange correspond to higher probe intensities and GC content, respectively. (B) Running median average of probe signal intensities across mouse chromosome 18 for testes PolyA+ RNA (red) and genomic DNA (green), showing a similar baseline trend in both samples. After quantile normalization of the PolyA+ sample against genomic DNA, the non-specific baseline pattern is no longer present (blue).
Found at: doi:10.1371/journal.pbio.1000371.s004 (0.96 MB PDF)

Figure S5 Effect of alignment parameters on the number of uniquely mapped reads. Singleton 32 mer reads from 9 human tissues were mapped as either 25 mer or 32 mer, allowing for 0–2 mismatches. The number of uniquely mapped reads at each parameter combination is indicated.
Found at: doi:10.1371/journal.pbio.1000371.s005 (0.09 MB PDF)

Figure S6 Overview of splice junction detection and reconstruction of gene structures. (A) Splice junction detection by Tophat (modified from [45]). (B) Outline of the method used to merge splice junctions into gene structures. See Materials and Methods for a detailed description of this figure.
Found at: doi:10.1371/journal.pbio.1000371.s006 (0.11 MB PDF)

Figure S7 Precision-recall of known splice junctions in human brain single- (A, B) and paired-end (C, D) read data. Known junctions were defined as those that bridged any two exons of a single annotated reference transcript. The effects of three different parameters were tested: anchor size, junction read coverage, and the number of times the same junction sequence was found for different splice junctions. Numbering of points corresponding to different coverage thresholds is indicated in the top left panel and is analogous for all other lines drawn. The arrow indicates the precision-recall values for the parameter settings used in the Tophat analysis of single-end reads, before filtering junctions with low-complexity sequences.
Found at: doi:10.1371/journal.pbio.1000371.s007 (0.15 MB PDF)

Figure S8 PolyA/T repeat bias in junction sequences from single-end reads. Plots showing the percentage of junction sequences containing (A) PolyA/PolyT repeats or (B) PolyG/PolyC repeats, as a function of the repeat length. Lines represent different human RNA-Seq samples and are colored as indicated on the right.
Found at: doi:10.1371/journal.pbio.1000371.s008 (0.12 MB PDF)

Table S1 Read mass statistics for all RNA-Seq samples.
Found at: doi:10.1371/journal.pbio.1000371.s009 (0.05 MB PDF)

Table S2 Transcribed genomic area for all RNA-Seq samples.
Found at: doi:10.1371/journal.pbio.1000371.s010 (0.05 MB PDF)

Table S3 Proportion of intergenic reads in 10-kb regions flanking annotated genes.
Found at: doi:10.1371/journal.pbio.1000371.s011 (0.04 MB PDF)

Table S4 Human splice junction mapping statistics.
Found at: doi:10.1371/journal.pbio.1000371.s012 (0.04 MB PDF)

Table S5 Human splice junctions bridging exons between annotated genes.
Found at: doi:10.1371/journal.pbio.1000371.s013 (0.09 MB XLS)

Table S6 (A) Overlap between genomic features and novel exons in human TUs attached to known genes. (B) Overlap between genomic features and exons in human TUs independent from known genes.
Found at: doi:10.1371/journal.pbio.1000371.s014 (0.05 MB PDF)

Table S7 Alternative splice junctions connecting to unannotated upstream promoters in the human genome.
Found at: doi:10.1371/journal.pbio.1000371.s015 (0.07 MB XLS)

Table S8 (A) Overlap between significant seqfrags in trimmed intergenic regions and genomic features. (B) Overlap between seqfrag clusters in trimmed intergenic regions and genomic features.
Found at: doi:10.1371/journal.pbio.1000371.s016 (0.05 MB PDF)

Table S9 RNA sources for tiling array experiments.
Found at: doi:10.1371/journal.pbio.1000371.s017 (0.04 MB PDF)

Acknowledgments

We are grateful to Marinella Gebbia for assisting with the tiling array hybridizations, and Brendan J. Frey, Clement Chung, Jim Huang, and Sandy Pan for helpful discussions. Finally, we would like to thank Gary Shroth and colleagues at Illumina for sharing their paired-end sequencing data.

Author Contributions

The author(s) have made the following declarations about their contributions: Conceived and designed the experiments: HvB TRH. Performed the experiments: HvB. Analyzed the data: HvB. Contributed reagents/materials/analysis tools: CN BJB. Wrote the paper: HvB TRH.

References

1. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, et al. (2002) Large-scale transcriptional activity in chromosomes 21 and 22. Science 296: 916–919.
2. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, et al. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res 14: 331–342.
3. Schadt EE, Edwards SW, GuhaThakurta D, Holder D, Ying L, et al. (2004) A comprehensive transcript index of the human genome generated using microarrays and computational approaches. Genome Biol 5: R73.
4. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, et al. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308: 1149–1154.
5. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316: 1484–1488.
6. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, et al. (2004) Global identification of human transcribed sequences with genome tiling arrays. Science 306: 2242–2246.
7. Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, et al. (2004) A gene expression map for the euchromatic genome of Drosophila melanogaster. Science 306: 655–660.
8. Khaitovich P, Kelso J, Franz H, Visagie J, Giger T, et al. (2006) Functionality of intergenic transcription: an evolutionary comparison. PLoS Genet 2: e171. doi:10.1371/journal.pgen.0020171.
9. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816.
10. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559–1563.
11. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420: 563–573.
12. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38: 626–635.
13. Johnson JM, Edwards S, Shoemaker D, Schadt EE (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends Genet 21: 93–102.
14. Willingham AT, Gingeras TR (2006) TUF love for "junk" DNA. Cell 125: 1215–1220.
15. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, et al. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet 36: 40–45.
16. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470–476.
17. Numata K, Kanai A, Saito R, Kondo S, Adachi J, et al. (2003) Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. Genome Res 13: 1301–1306.
18. Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, et al. (2006) Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res 16: 11–19.
19. Guttman M, Amit I, Garber M, French C, Lin MF, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458: 223–227.
20. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322: 1845–1848.
21. Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, et al. (2008) RNA exosome depletion reveals transcription upstream of active human promoters. Science 322: 1851–1854.
22. Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon GJ, et al. (2009) Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. Nature 457: 1028–1032.
23. Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, et al. (2009) Tiny RNAs associated with transcription start sites in animals. Nat Genet 41: 572–578.
24. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, et al. (2008) Divergent transcription from active promoters. Science 322: 1849–1851.
25. Wang J, Zhang J, Zheng H, Li J, Liu D, et al. (2004) Mouse transcriptome: neutral evolution of "non-coding" complementary DNAs. Nature 431: 1 p following 757; discussion following 757.
26. Brosius J (2005) Waste not, want not–transcript excess in multicellular eukaryotes. Trends Genet 21: 287–288.
27. Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, et al. (2003) The transcriptional activity of human Chromosome 22. Genes Dev 17: 529–540.
28. Royce TE, Rozowsky JS, Bertone P, Samanta M, Stolc V, et al. (2005) Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. Trends Genet 21: 466–475.
29. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63.
30. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, et al. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5: 613–619.
31. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, et al. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453: 1239–1243.
32. van Bakel H, Hughes TR (2009) Establishing legitimacy and function in the new transcriptome. Brief Funct Genomic Proteomic 8: 424–436.
33. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, et al. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133: 523–536.
34. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344–1349.
35. Armour CD, Castle JC, Chen R, Babak T, Loerch P, et al. (2009) Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. Nat Methods 6: 647–649.
36. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621–628.
37. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, et al. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321: 956–960.
38. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW (2008) The antisense transcriptomes of human cells. Science 322: 1855–1857.
39. Slomovic S, Laufer D, Geiger D, Schuster G (2006) Polyadenylation of ribosomal RNA in human cells. Nucleic Acids Res 34: 2966–2975.
40. Lian Z, Karpikov A, Lian J, Mahajan MC, Hartman S, et al. (2008) A genomic analysis of RNA polymerase II modification and chromatin architecture related to 3′ end RNA polyadenylation. Genome Res 18: 1224–1237.
41. Parkhomchuk D, Amstislavskiy V, Soldatov A, Ogryzko V (2009) Use of high throughput sequencing to observe genome dynamics at a single cell level. Proc Natl Acad Sci U S A.
42. Chiu KP, Wong CH, Chen Q, Ariyaratne P, Ooi HS, et al. (2006) PET-Tool: a software suite for comprehensive processing and managing of Paired-End diTag (PET) sequence data. BMC Bioinformatics 7: 390.
43. Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, et al. (2009) Mapping accessible chromatin regions using Sono-Seq. Proc Natl Acad Sci U S A 106: 14926–14931.
44. Hammoud SS, Nix DA, Zhang H, Purwar J, Carrell DT, et al. (2009) Distinctive chromatin in human sperm packages genes for embryo development. Nature 460: 473–478.
45. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105–1111.
46. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, et al. (2010) Origins and functional impact of copy number variation in the human genome. Nature 464: 704–12.
47. Siepel A, Haussler D (2004) Computational identification of evolutionary conserved exons. In: Bourne PE, Gusfield D, eds. Proceedings of the eighth annual international conference on Research in computational molecular biology. San Diego: ACM. pp 177–186.
48. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, et al. (2006) Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol 2: e33. doi:10.1371/journal.pcbi.0020033.
49. Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat Biotechnol 23: 1383–1390.
50. Tuan D, Solomon W, Li Q, London IM (1985) The "beta-like-globin" gene domain in human erythroid cells. Proc Natl Acad Sci U S A 82: 6384–6388.
51. Groudine M, Kohwi-Shigematsu T, Gelinas R, Stamatoyannopoulos G, Papayannopoulou T (1983) Human fetal to adult hemoglobin switching: changes in chromatin structure of the beta-globin gene locus. Proc Natl Acad Sci U S A 80: 7551–7555.
52. Gross DS, Garrard WT (1988) Nuclease hypersensitive sites in chromatin. Annu Rev Biochem 57: 159–197.
53. Sabo PJ, Hawrylycz M, Wallace JC, Humbert R, Yu M, et al. (2004) Discovery of functional noncoding elements by digital analysis of chromatin structure. Proc Natl Acad Sci U S A 101: 16837–16842.
54. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, et al. (2009) Unlocking the secrets of the genome. Nature 459: 927–930.
55. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, et al. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35: W345–W349.
56. Dobi KC, Winston F (2007) Analysis of transcriptional activation at a distance in Saccharomyces cerevisiae. Mol Cell Biol 27: 5575–5586.
57. Mattick JS (2009) The genetic signatures of noncoding RNAs. PLoS Genet 5: e1000459. doi:10.1371/journal.pgen.1000459.
58. Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding RNAs: insights into functions. Nat Rev Genet 10: 155–159.
59. Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12: 656–664.

60. Jiang H, Wong WH (2008) SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics 24: 2395–2396.
61. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.
62. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034–1050.
63. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc B 57: 289–300.
64. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.
65. Narlikar L, Sakabe NJ, Blanski AA, Arimura FE, Westlund JM, et al. Genome-wide discovery of human heart enhancers. Genome Res 20: 381–392.
66. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, et al. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457: 854–858.