GENETICS

It's the Sequence, Stupid!

Differences in regulatory DNA sequences drive species-specific gene expression.

Hilary A. Coller[1] and Leonid Kruglyak[2]

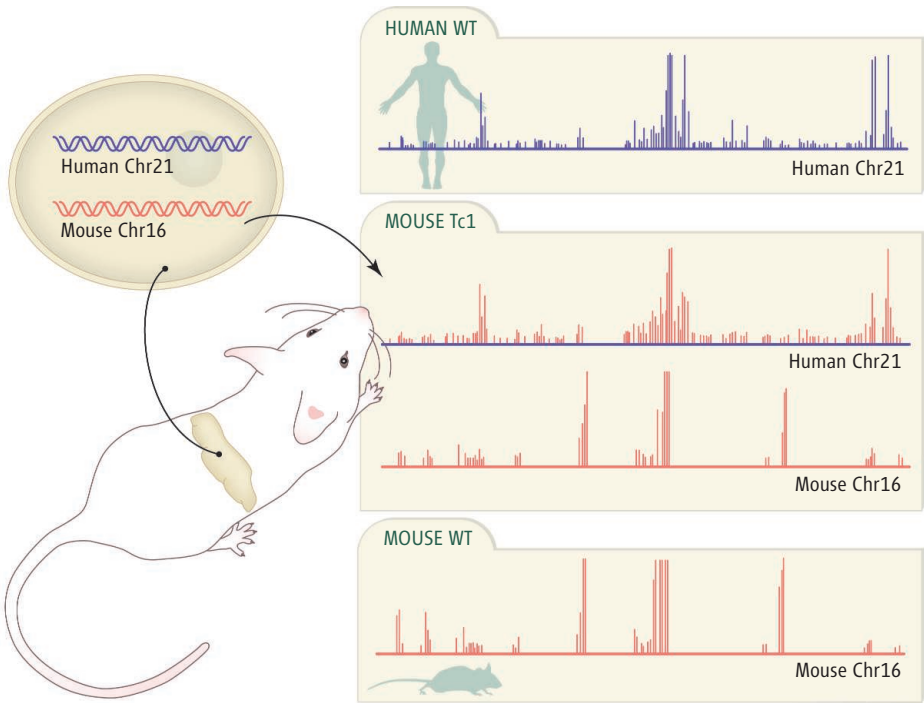
One of the surprises revealed by comparative genome sequencing is that closely related species share remarkably similar complements of genes. For example, a recent evaluation of the human gene catalog found at most 168 genes without close homologs in mouse or dog, with perhaps as few as 12 representing newly evolved protein-coding regions (*1*). Moreover, the corresponding genes tend not to differ much in their coding sequences: Nearly 80% of amino acids are identical between orthologous human and mouse proteins (*2*). Although this leaves many potentially functional coding changes, these observations lend further credence to the proposal, first made more than 30 years ago, that many of the observed differences between species likely stem from when and where the products of the genes are made (*3*). But what governs these changes in gene expression? There is no shortage of possible explanations—differences in external cues and cellular milieus, in how genomes are packaged, in the proteins that control transcription, and in the regulatory sequences to which they bind. Strikingly, on page 434 of this issue (*4*), Wilson *et al*. show that in human and mouse liver cells, the differences in regulatory sequences dominate all other factors.

Wilson *et al*. took advantage of an ideal system: a mouse model of Down syndrome in which mouse cells contain a copy of human chromosome 21 in addition to the complete mouse genome (*5*). In these cells, the human DNA sequence is placed in an otherwise murine context, including all external and cellular cues as well as regulatory proteins. This system allowed the authors to ask an otherwise impossible question: Is regulation of the genes on human chromosome 21 in these mouse cells (Tc1 hepatocytes) determined by the human DNA sequence, or by the mouse cellular environment and transcriptional machinery?

The authors compared the regulation of human genes in Tc1 cells to those of their mouse orthologs in these same cells. They then compared the observed patterns to those in mouse hepatocytes from littermates that did not inherit the extra human chromosome, as well as to those in normal human hepatocytes. The authors first confirmed that the protein binding and expression patterns of mouse genes in Tc1 hepatocytes match those in normal mouse hepatocytes, and that both differ from patterns for orthologous genes in human cells. What about the human genes in the Tc1 hepatocytes? If regulation is driven largely by sequence, then these genes should be regulated just as they are in normal human hepatocytes, whereas if species-specific developmental context, epigenetic factors, or differences in transcription factors themselves play a defining role, then the genes should most closely mimic their mouse orthologs.

The authors compared regulation at three levels: binding of transcription factors to DNA, modification of histones [proteins that bind chromosomal DNA and determine its packing and accessibility for binding (*6*)], and gene expression. The results were clear. The binding patterns of transcription factors HNF1α, HNF4α, and HNF6 on human chromosome 21 in mouse cells matched those seen in human cells, not those observed in mouse cells, with only a few exceptions (see the figure). Similarly, although histone H3K4me3 modifications at canonical transcription start sites were largely shared between the human and mouse chromosomes, these same modifications at other sites (thought to represent unannotated promoters) showed human-specific patterns on human chromosome 21 in Tc1 cells. Finally, gene expression (the amount of messenger RNA transcribed) from human chromosome 21 genes in Tc1 hepatocytes was more closely correlated to the expression of human chromosome 21 genes in human hepatocytes than to the expression of their mouse orthologs in the Tc1 cells. The authors thus concluded that it is the regulatory DNA sequence, rather than any other species-specific factor, that is the single most important determinant of gene expression.

This result raises many interesting questions about the transcriptional regulatory code. Wilson *et al*. show that the information required for species-specific regulation is encoded in cis-regulatory DNA sequence. Yet because essentially all of human chromosome



CREDIT: P. HUEY/SCIENCE

Reading the regulatory code. Transcription factor proteins in mouse Tc1 cells carrying a human chromosome (**middle**) bind to the human DNA in a human-specific pattern (**top**) and to the corresponding mouse DNA in a mouse-specific pattern (**bottom**).

[1]Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA. [2]Howard Hughes Medical Institute, Lewis-Sigler Institute for Integrative Genomics, and Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA. E-mail: hcoller@princeton.edu; leonid@genomics.princeton.edu

Published by AAAS

21 is present, the data do not address whether the code is local. The results are consistent with either a few regulatory sites for each gene close to the corresponding start of transcription, or many interacting regulatory sites scattered across large swaths of human chromosome 21. Experiments that replaced a mouse gene by its human ortholog along with varying lengths of upstream and downstream regions might address how much human sequence is needed to recapitulate a human pattern of binding and expression. Would a few kilobases upstream be sufficient, as would be expected if only proximal transcription factor binding sites matter, or would much larger segments upstream and downstream be required, indicating that multiple types of poorly understood sequence elements are acting in concert?

The paper's findings also call into question one of the basic tenets of comparative genomics: that evolutionary conservation can serve as the primary tool for finding functional sequences (*2*, *7*, *8*). Clearly, nonconserved sequences are responsible for the observed functional differences in binding and expression of human and mouse genes in the same cells. Thus, although many conserved noncoding sequences are functional, and interspecies comparisons can help us to identify these motifs, narrowing our attention only to these sequences must result in an incomplete understanding of the regulatory code (*9*). Indeed, this approach guarantees missing the species-specific regulatory instructions that make us different from mice.

Finally, the transcriptional machinery of a mouse cell is able to read out human-specific gene expression instructions based solely on the sequence of the human chromosome, but today's bioinformatic methods cannot. Substantial progress has been made in predicting expression from sequence in yeast (*10*, *11*), whereas in mammals, known regulatory sequences are too short and degenerate, and extend too far from the start of transcription, for us to accurately predict gene expression from sequence information alone. So what would it take for us to predict how mouse cells would read out the regulatory code of, say, an armadillo chromosome, without doing the experiments? That is, how can we move toward reading the regulatory code as easily as we read the genetic code, which allows us to seamlessly go from the DNA sequence to the protein complement of any species? We anticipate that deciphering the regulatory code will require a combination of computational and experimental approaches, in concert with improved physical models of protein-DNA interaction (*12*). It will also require an understanding of the cellular context to an extent not necessary for the genetic code, because the complement of regulatory proteins operating to control transcription varies with the species, the cell type, and the environment. The ENCODE project provides one model of experimentally monitoring all accessible regulatory readouts, such as transcription itself, binding of regulatory proteins, histone modification states, and nucleosome positioning on a global scale (*13*). Our hope is that innovative approaches to the analysis of ENCODE-like data will ultimately allow us to crack the regulatory code.

References

1. M. Clamp *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19428 (2007).
2. R. H. Waterston *et al.*, *Nature* **420**, 520 (2002).
3. M. C. King, A. C. Wilson, *Science* **188**, 107 (1975).
4. M. D. Wilson *et al.*, *Science* **322**, 434 (2008); published online 11 September 2008 (10.1126/science.1160930).
5. A. O'Doherty *et al.*, *Science* **309**, 2033 (2005).
6. T. Jenuwein, C. D. Allis, *Science* **293**, 1074 (2001).
7. Q. F. Wang *et al.*, *Genome Biol.* **8**, R1 (2007).
8. X. Xie *et al.*, *Nature* **434**, 338 (2005).
9. X. Y. Li *et al.*, *PLoS Biol.* **6**, e27 (2008).
10. M. A. Beer, S. Tavazoie, *Cell* **117**, 185 (2004).
11. C. T. Harbison *et al.*, *Nature* **431**, 99 (2004).
12. A. M. Moses *et al.*, *PLoS Comput. Biol.* **2**, e130 (2006).
13. E. Birney *et al.*, *Nature* **447**, 799 (2007).

10.1126/science.1165664

MATERIALS SCIENCE

In Praise of Pores

Advanced processing methods are used to tailor the properties of porous ceramics.

Paolo Colombo

Highly porous structures are found extensively in the natural world (*1*), because their design enables the efficient optimization of characteristics such as the strength-to-density and stiffness-to-density ratios. Moreover, synthetic porous ceramics have advantages over metallic or polymeric components, especially when resistance to high-temperature or corrosive environments or compatibility with biological materials is required. Recent progress in fabrication procedures has considerably widened the range of morphologies and properties achievable for porous ceramics, resulting in their use in an ever-expanding range of applications, including catalyst supports and chemical reactors (*2*), biomedical tracking and delivery platforms (*3*), electrodes, insulators, and heat exchangers (*4*). The introduction of nanopores in ceramics has opened possibilities for the development of smart devices such as photoactivated sensors and switches and drug delivery capsules.

All these applications require the porous component to have a specific range of values for different properties (see the figure) (*5*), which can only be achieved through judicious choice of starting materials and well-controlled processing. The manufacturing method strongly influences the amount of porosity (from a few percent to more than 95% in volume), the pore size (from nanometers to millimeters), the distribution of the pores within the solid, the shape and interconnectivity of the pores and other characteristics such as the flaw population (amount, size, and morphology of defects), and the size and cost of the component. The introduction of porosity is therefore an extremely versatile and powerful tool for greatly extending the range of properties offered by a ceramic component. No other single strategy enables the value of a given property to be varied to such an extent, often by orders of magnitude, in a single material.

Macroporous cellular (foam-like) ceramic structures are conventionally produced by dip-coating a polymeric foam into a ceramic slurry, followed by burnout of the preform and sintering. This approach leads to high-porosity, low-cost parts of limited strength, suitable for molten metal filters or kiln furniture. Similar porosity can also be created through the elimination of "sacrificial filler" materials by burnout or dissolution, providing tighter control on the average pore size distribution, with a wider range of cell size and amount of porosity. More recent methods use direct foaming of a ceramic slurry (for example, through mechanical frothing, gas injection, or in situ gas generation by decomposing an

Dipartimento di Ingegneria Meccanica–Settore Materiali, Università di Padova, 35131 Padova, Italy, and Department of Materials Science and Engineering, Pennsylvania State University, University Park, PA 16802, USA. E-mail: paolo.colombo@unipd.it

Species-Specific Transcription in Mice Carrying Human Chromosome 21

Michael D. Wilson, *et al.*

Science **322**, 434 (2008);
DOI: 10.1126/science.1160930

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

1.1×10^{18} and 1.5×10^{18} kg. With the photometrically derived nominal size of $r = 54$ km for each component (assumed albedo of 0.16), the density of 2001 QW_{322} (Fig. 2B) is probably 0.8 to 1.2 g cm^{-3}. This is a little higher than that of comparably sized outer solar system bodies [figure 5 of (*13*); 0.6 to 0.8 g cm^{-3}]. Our nominal albedo of 0.16 is approximately double that estimated from optical and thermal infrared photometry for similar-size KBOs (*14*, *15*) but about a factor of 2 below that of (58534) Logos/Zoe ($p = 0.37 \pm 0.04$) (*2*), which is of comparable size. Estimated density from eqs. S2 and S3 is proportional to the assumed albedo to the power of $^3/_2$. Halving our albedo would increase our radius estimates by $\sqrt{2}$ and decrease the estimated density by a factor of $2^{3/2} = 2.8$, below the range of published densities (*13*) for such small bodies.

The nominal densities shown in Fig. 2 are at the boundary between the density of a low-porosity, pure-water ice body and that of a mixture of water ice and silicate rocks (*13*). A thermal detection, mutual eclipse, or stellar occultation by the binary (all unlikely) would be necessary to further constrain the size, albedo, density, and hence the bulk composition of 2001 QW_{322}.

Given the very large separation (Fig. 3), such a binary is difficult to create and maintain. Of all the proposed KBO binary-formation scenarios (*16*–*19*), only the collision of two bodies close to a third one (*16*) can simply explain the primordial formation of such a system (*7*).

A study of the long-term stability of the large-separation KB binaries (*8*) led to the conclusion that the major destabilizing factor is unbinding due to direct collisions of impactors on the secondary. Applying their method to the newly determined orbital and physical parameters for 2001 QW_{322} and our nominal albedo, we find that the lifetime of this binary is 0.3 to 1 billion years, which is two to three times shorter than the previous estimate. This finding implies one of two things: (i) Either 2001 QW_{322} was created with its current mutual orbit early in the history of the solar system, in which case it is one of the few survivors of a population at least 50 to 100 times larger, or (ii) this is a transitory object, evolving because of perturbation from interactions with smaller KBOs, from a population of more tightly bound binaries. Asserting this latter hypothesis would require better orbital statistics for moderately large KB binaries (separation of 1 to 2″).

For the likely mutual-orbit parameters, the average orbital speed is $\langle v \rangle \simeq 0.85$ m/s or a mere 3 km $hour^{-1}$, a slow human walking pace. An observer standing on one of the components (a very precarious situation, as the gravity is only 0.02 m/s^2 or nearly 600 times smaller than on Earth) would see the other component subtend an angle of only 3 arc min, which corresponds to a pinhead seen at arm's length. The existence of the other component would not be in doubt, however, because when viewed at full phase it would be as luminous as Saturn seen from Earth, and it would move perceptibly from week to week.

References and Notes

1. W. J. Merline *et al.*, in *Asteroids III*, W.F. Bottke Jr., A. Cellino, P. Paolicchi, R.P. Binzel, Eds. (Univ. of Arizona Press, Tucson, AZ, 2002), pp. 289–312.
2. K. S. Noll, W. M. Grundy, E. I. Chiang, J. L. Margot, S. D. Kern, in *The Solar System Beyond Neptune*, A. Barucci, H. Boehnhardt, D. Cruikshank, A. Morbidelli, Eds. (Univ. of Arizona Press, Tucson, AZ, 2008), pp. 345–363.
3. J. L. Margot, M. E. Brown, C. A. Trujillo, R. Sari, J. A. Stansberry, *Bull. Am. Astron. Soc.* **37**, 737 (2005).
4. J. J. Kavelaars, J.-M. Petit, G. Gladman, M. Holman, *IAU Circ.* **7749**, 1 (2001).
5. W. J. Merline *et al.*, *Bull. Am. Astron. Soc.* **32**, 1017 (2000).
6. B. Gladman, B. G. Marsden, C. Van Laerhoven, in *The Solar System Beyond Neptune*, A. Barucci, H. Boehnhardt, D. Cruikshank, A. Morbidelli, Eds. (Univ. of Arizona Press, Tucson, AZ, 2008), pp. 43–57.
7. See supporting online material text.
8. J.-M. Petit, O. Mousis, *Icarus* **168**, 409 (2004).
9. J. Burns, V. Carruba, B. Gladman, B.G. Marsden, *Minor Planet Electron. Circ.* **L30,** 1 (2002).
10. O. R. Hainaut, A. C. Delsanti, *Astron. Astrophys.* **389**, 641 (2002).
11. A. A. S. Gulbis, J. L. Elliot, J. F. Kane, *Icarus* **183**, 168 (2006).
12. D. Nesvorný, J. L. A. Alvarellos, L. Dones, H. F. Levison, *Astron. J.* **126**, 398 (2003).
13. W. M. Grundy *et al.*, *Icarus* **191**, 286 (2007).
14. J. A. Stansberry *et al.*, *Astrophys. J.* **643**, 556 (2006).
15. J. R. Spencer, J. A. Stansberry, W. M. Grundy, K. S. Noll, *Bull. Am. Astron. Soc.* **38**, 546 (2006).
16. S. J. Weidenschilling, *Icarus* **160**, 212 (2002).
17. P. Goldreich, Y. Lithwick, R. Sari, *Nature* **420**, 643 (2002).
18. Y. Funato, J. Makino, P. Hut, E. Kokubo, D. Kinoshita, *Nature* **427**, 518 (2004).
19. S. A. Astakhov, E. A. Lee, D. Farrelly, *Mon. Not. R. Astron. Soc.* **360**, 401 (2005).

20. This work was partially supported by NASA/Planetary Astronomy Program grant NNG04GI29G. A.C.B. also acknowledges support from Ministerio de Educacion y Ciencia (Spain), National project n. AYA2005-07808-C03-03. J.L.M. was partially supported by grant NNX07AK68G from the NASA Planetary Astronomy program. This research used the facilities of the Canadian Astronomy Data Centre operated by the National Research Council of Canada with the support of the Canadian Space Agency. The Canada-France-Hawaii Telescope is operated by the National Research Council of Canada, the Institut National des Sciences de l'Univers of the Centre National de la Recherche Scientifique of France, and the University of Hawaii. Observations at Palomar Observatory are carried out under a collaborative agreement between Cornell University and the California Institute of Technology. Observations made with European Southern Observatory Telescopes at the La Silla or Paranal Observatories under program IDs 069.C-0460, 071.C-0497, 072.C-0542, 074.C-0379, 075.C-0251, and 380.C-0791. The Gemini Observatory is operated by the Association of Universities for Research in Astronomy, under a cooperative agreement with NSF on behalf of the Gemini partnership: NSF (US), the Science and Technology Facilities Council (UK), the National Research Council (Canada), Comisión Nacional de Investigación Científica y Tecnológica (Chile), the Australian Research Council (Australia), Ministério da Ciência e Tecnologia (Brazil), and Secretaría de Ciencia y Tecnología (Argentina). Observations were obtained at the WIYN Observatory, a joint facility of the University of Wisconsin–Madison, Indiana University, Yale University, and the National Optical Astronomy Observatories; the William Herschel Telescope, at Roque de los Muchachos Observatory (La Palma, Canary Islands, Spain), operated by the Instituto de Astrofisica de Canarias; and the MMT Observatory, a joint facility of the Smithsonian Institution and the University of Arizona.

Supporting Online Material

www.sciencemag.org/cgi/content/full/322/5900/432/DC1
SOM Text
Figs. S1 and S2
Tables S1 to S4
References

11 July 2008; accepted 12 September 2008
10.1126/science.1163148

Species-Specific Transcription in Mice Carrying Human Chromosome 21

Michael D. Wilson,[1]* Nuno L. Barbosa-Morais,[1,2]* Dominic Schmidt,[1,2] Caitlin M. Conboy,[3] Lesley Vanes,[4] Victor L. J. Tybulewicz,[4] Elizabeth M. C. Fisher,[5] Simon Tavaré,[1,2,6] Duncan T. Odom[1,2]†

Homologous sets of transcription factors direct conserved tissue-specific gene expression, yet transcription factor–binding events diverge rapidly between closely related species. We used hepatocytes from an aneuploid mouse strain carrying human chromosome 21 to determine, on a chromosomal scale, whether interspecies differences in transcriptional regulation are primarily directed by human genetic sequence or mouse nuclear environment. Virtually all transcription factor–binding locations, landmarks of transcription initiation, and the resulting gene expression observed in human hepatocytes were recapitulated across the entire human chromosome 21 in the mouse hepatocyte nucleus. Thus, in homologous tissues, genetic sequence is largely responsible for directing transcriptional programs; interspecies differences in epigenetic machinery, cellular environment, and transcription factors themselves play secondary roles.

Higher eukaryotes are organized collections of different cell types, each of which is created from differential transcription of a common genome (*1*). Evolutionarily conserved sets of tissue-specific transcription factors establish each cell's transcription during development and maintain it during adulthood by binding to DNA in a sequence-specific manner (*1*–*3*). These proteins typically recognize short consensus motifs, often between 6 and 16 nucleotides, found at high frequency throughout a genome. How transcription factors discriminate among nearly identical motifs is poorly understood, although chromatin state, cellular environ-

ment, and surrounding regulatory sequences have all been suggested to direct transcription factors to specific cognate sites (*4*, *5*). Sequence comparisons alone can identify only a fraction of regulatory regions (*6*), because the protein–DNA binding events linking transcription factors with genetic control sequences, and thus gene expression, change on a rapid evolutionary time scale (*7*–*10*). For instance, the targeted genes and precise binding locations of conserved, tissue-specific transcription factors for mouse and human differ significantly (*7*). Even when transcription factors bind near orthologous genes in two species, the precise locations of the large majority of the binding events do not align (*7*, *9*). In numerous cases, transcription factors frequently bind one highly conserved motif near a gene in one species and a different conserved motif near the orthologous gene in a second species (*7*, *9*). This divergence of transcription factor–binding locations among related species is a widely occurring phenomenon, and similar observations have been made in yeast, *Drosophila*, and mammals (*7*–*10*). Thus, the mechanisms that determine tissue-specific transcriptional regulation must be more complex than simple gain and loss of the immediately bound, local sequence motifs.

The role that DNA sequence plays in directing histone modifications is also not well understood. It has been previously shown on human chromosomes 21 and 22 that, at the sequence level, sites of methylation at lysine 4 of histone H3 (H3K4) are no more conserved relative to mouse genome than background sequence (*11*). Genomic locations where H3K4 methylation occurred in both species did not show high levels of overall sequence conservation (*11*). One interpretation of this observation is that sequence comparisons alone have a limited capability for identifying epigenetic landmarks.

Ultimately, transcription factor binding and epigenetic state contribute to tissue-specific gene expression (*4*, *5*). A complete understanding of the mechanisms underlying divergence of transcriptional regulation and transcription itself is central to the debate surrounding the relative roles that cis-regulatory mutations and protein-coding mutations play during evolution (*12*, *13*).

Here, we isolate the role that genetic sequence plays in transcription by using a mouse model of Down syndrome that stably transmits human chromosome 21 (*14*, *15*). In this mouse, we compared transcriptional regulation of orthologous human and mouse sequences in the same nuclei and, thereby, eliminated most environmental and experimental variables otherwise inherent to interspecies comparisons.

Tc1 mice are partially mosaic, and ~60% of their hepatic cells contain human chromosome 21, which we confirmed by quantitative genotyping (fig. S1). Historically, human chromosome 21 has been extensively studied to explore transcription and transcriptional regulation on a chromosomewide basis (*11*, *16*, *17*), and the corresponding orthologous mouse regions are located primarily in chromosome 16, with additional regions in chromosomes 10 and 17 (*14*).

We chose liver as a representative tissue for these experiments because most liver cells are hepatocytes that are easy to isolate and highly conserved in structure and function. A set of conserved, well-characterized transcription factors (including HNF1α, HNF4α, and HNF6) are responsible for hepatocyte development and function (*2*, *18*), and orthologous liver-specific mouse and human transcription factors recognize the same consensus sequences (*7*). Despite almost perfect conservation in their DNA binding domains, the mouse orthologs of HNF1α, HNF4α, and HNF6 can vary in amino acid composition by up to 5% from their human orthologs in regions that could mediate protein-protein interactions (table S1) (*19*, *20*). No liver-specific transcription factor genes we profiled reside on human chromosome 21 (HsChr21); therefore, binding events identified are due to mouse transcription factors.

Because approximately three-quarters of the conserved synteny between human chromosome 21 and the mouse genome resides on mouse chromosome 16, we used tiling microarrays to obtain genomic information in four chromosome-nuclear combinations: human chromosome 21 located in human hepatocytes (indicated as WtHsChr21), human chromosome 21 located in Tc1 mouse hepatocytes (TcHsChr21), mouse chromosome 16 located in Tc1 mouse hepatocytes (TcMmChr16), and mouse chromosome 16 located in wild-type mouse hepatocytes (WtMmChr16).

For every experiment, we subtracted all potentially mouse-human degenerate probes computationally, as well as experimentally, by cross-hybridizing each platform with nucleic acids from the heterologous species [details in (*15*)]. Taken together, our genomic microarrays, in principle, could interrogate more than 28 Mb of human and mouse DNA sequence shared in both HsChr21 and MmChr16, which would capture information on ~145 genes embedded in their native chromosomal context. After subtraction of regions deleted from TcHsChr21, ~20 Mb and 105 genes are interrogated herein.

Three aspects of this system are of particular note: (i) the primary Tc1 hepatocytes used in these experiments are indistinguishable in liver function, tissue architecture, and mouse genome–based gene expression and transcription factor binding from that profiled from wild-type littermates (see below); (ii) TcHsChr21 and TcMmChr16 are in an identical dietary, developmental, nuclear, organismal, and metabolic environment in Tc1 hepatocytes; and (iii) as all profiled transcription factors arise from the mouse genome, species-specific effects are eliminated for antisera used in chromatin immunoprecipitation (ChIP) experiments.

We first confirmed the substantial divergence in transcription factor binding between wild-type mouse and human hepatocytes by performing ChIP assays against HNF1α, HNF4α, and HNF6, which are members of three different protein families (Fig. 1). As expected, most transcription factor–binding events were species-specific (*7*) and were located distal to transcriptional start sites (TSSs) (*10*, *21*). We define human-specific (or human-unique) as ChIP enrichment on the human genome that does not have detectable signal in the orthologous region of the mouse genome (and vice versa) (Fig. 1A, and fig. S2).

To determine the role that human DNA sequence can play in directing mouse transcription factor binding, we performed ChIP experi-

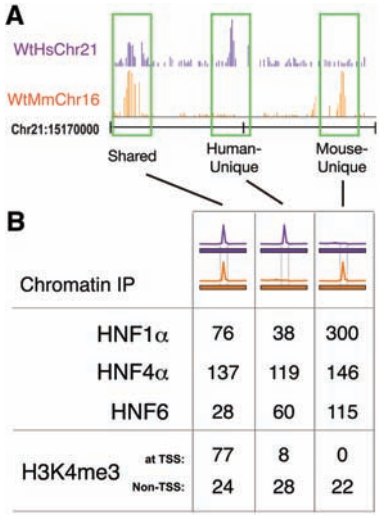


Fig. 1. Transcriptional regulation of human hepatocytes varies from mouse hepatocytes across a complete chromosome. (**A**) Genome track showing ChIP enrichment of HNF1α binding in wild-type mouse and human hepatocytes across 30 kb of genomic sequence. The species of bound DNA sequences and ChIP signal are indicated by color: Purple represents human; orange represents mouse. Highlighted in green are HNF1α-bound regions that are shared by both species, human-unique, or mouse-unique. (**B**) The total number of genomic regions occupied by three transcription factors (HNF1α, HNF4α, and HNF6) and H3K4me3 that are shared between the species, human-unique, or mouse-unique. ChIP data were obtained in wild-type mouse and human hepatocytes across the homologous regions of human chromosome 21 and mouse chromosome 16.

[1]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. [2]Department of Oncology, Hutchison/MRC (Medical Research Council) Research Centre, Hills Road, Cambridge CB2 0XZ, UK. [3]Medical Scientist Training Program, University of Minnesota Medical School, Minneapolis, MN 55455, USA. [4]Division of Immune Cell Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK. [5]Institute of Neurology, University College London, Queen Square, London WC1N 3BG, UK. [6]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, UK.

*These authors contributed equally to this work
†To whom correspondence should be addressed. E-mail: duncan.odom@cancer.org.uk

ments against HNF1α, HNF4α, and HNF6 in hepatocytes from the Tc1 mouse (Fig. 2). For each transcription factor, we simultaneously hybridized DNA from replicate ChIP enrichment experiments to microarrays representing human chromosome 21 and mouse chromosome 16 (*15*). We found that transcription factor binding on TcMmChr16 and WtMmChr16 is largely identical; thus, the presence of an extra human chromosome does not perturb transcription factor binding to the mouse genome (fig. S3).

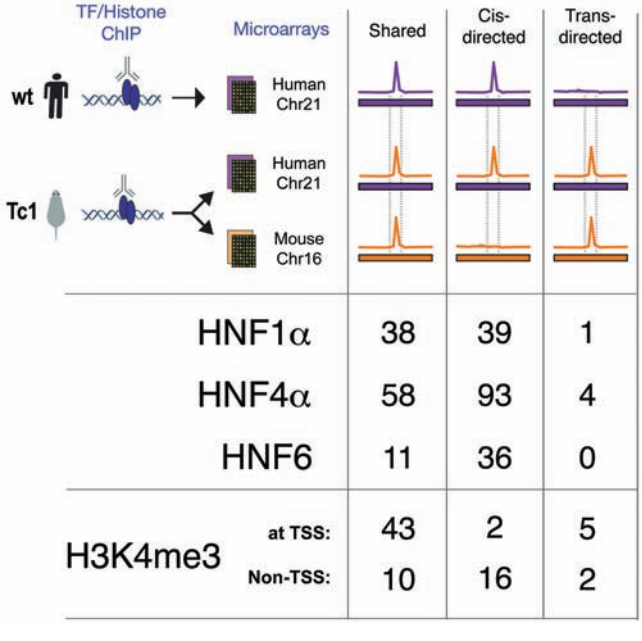
We then asked whether transcription factor binding to transchromic TcHsChr21 aligned with the positions found on (human) WtHsChr21 or (mouse) TcMmChr16. Although binding events could also be present uniquely on TcHsChr21 that do not align to either WtHsChr21 or TcMmChr16, this was rarely observed. If the transcription factor–binding positions on TcHsChr21 align with positions found on WtHsChr21, then that would indicate that this binding is largely determined by cis-acting DNA sequences, as the transcription factors are present in both mouse and human hepatocytes and regulate key liver functions. If more than a small number of binding events on TcHsChr21 were found at locations that align elsewhere in the genome (for instance, with binding events on TcMmChr16), then other mechanistic influences besides genome sequence, such as chromatin structure, interspecies differences in developmental remodeling, diet, and/or environment must contribute substantially toward directing the location of transcription factor binding.

Remarkably, almost all of the transcription factor–binding events on HsChr21 are found in both human and Tc1 mouse hepatocytes (85 to 92%) (Fig. 2A and fig. S4). The few peaks that appear to be unique to WtHsChr21 or TcHsChr21 are generally of lower intensity and difficult to evaluate reliably by using standard peak-calling algorithms (fig. S5). Indeed, as can be seen in Fig. 3, the pattern of conservation and divergence in transcription factor binding found in both WtHsChr21 (located in human liver) and WtMmChr16 (located in mouse liver) is recapitulated in TcHsChr21 and TcMmCh16 (both located in mouse liver) (see also figs. S6 and S7). Because transcription factors often bind to regions that do not contain their canonical binding sequences (*7*, *9*, *21*), this result is further notable.

Despite the evolutionary divergence of primate and rodent lineages, mouse genome–encoded transcription factors can bind to human sequences in a manner identical to the human genome–coded transcription factors in a homologous tissue. These data eliminate the possibility that protein concentration differences or small coding variations in the mouse versions of transcription factors (or within larger transcriptional complexes) could redirect transcription factor binding to locations different from those found in human. Taken together, underlying genetic sequences appear to be the dominant influence on where transcription factors bind in homologous mammalian tissues.

We then explored how the mouse chromatin remodeling machinery interacts with TcHsChr21 (Fig. 1) (*22*). Using ChIPs, we isolated nucleosomes containing the trimethylated lysine 4 of histone H3 (H3K4me3) to identify the genomic anchor points for basal transcriptional machinery (*11*, *22*–*25*). Although most H3K4me3 enrichment occurs at TSSs and correlates with gene expression, it recently has been shown that most TSSs are H3K4me3-enriched, regardless of whether they are being actively elongated (*11*, *22*–*25*). Depending on the cell type, approximately a quarter of genes can show differential H3K4 methylation, and many of these genes have been shown to be cell type–specific (*22*).

We first identified how well trimethylation of the H3K4 position is shared in both the wild-



		Shared	Cis-directed	Trans-directed
HNF1α		38	39	1
HNF4α		58	93	4
HNF6		11	36	0
H3K4me3	at TSS:	43	2	5
	Non-TSS:	10	16	2

Fig. 2. Comparison of the binding of the liver-specific transcription factors HNF1α, HNF4α, and HNF6, and enrichment of H3K4me3 on TcHsChr21 with the corresponding data obtained in mouse TcMmChr16 and human WtHsChr21 regions. The color scheme is the same as in Fig. 1; notably, the primary difference from Fig. 1 is the addition of the human chromosome in a mouse environment, which is indicated as a purple bar (representing the human chromosomal sequences) with an orange peak (from mouse transcription factor binding). The binding events on TcHsChr21 are sorted into categories on the basis of whether they align with similar peaks in mouse and human (shared), align only with peaks in human (cis-directed), or align only with peaks in mice (trans-directed).

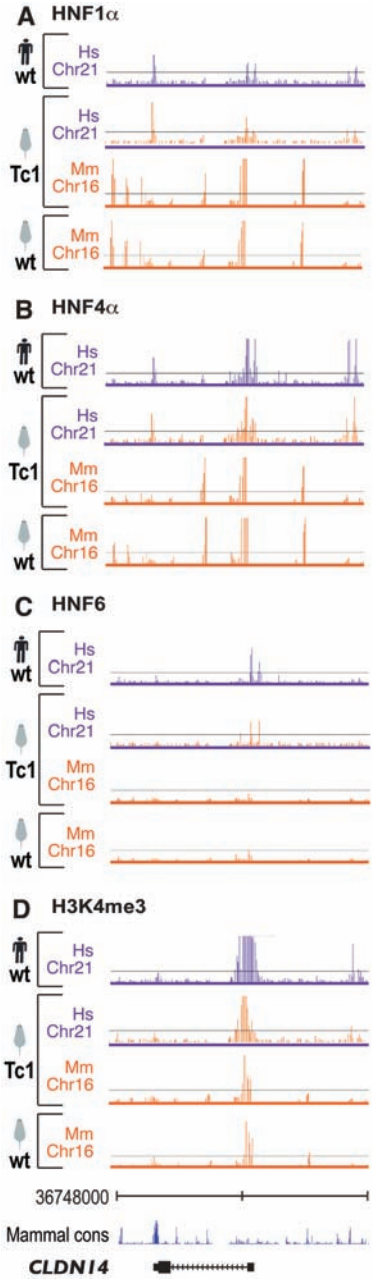


Fig. 3. Patterns of transcription factor binding and transcription initiation are determined by genetic sequence. ChIP enrichment for (**A**) HNF1α, (**B**) HNF4α, (**C**) HNF6, and (**D**) H3K4me3 are shown across a 50-kb region surrounding the liver-expressed gene *CLDN14*. The human chromosome 21 coordinates and the vertebrate sequence conservation track (Seq Cons; genome.ucsc.edu) are shown flanking *CLDN14*. Each panel shows the species of genetic sequence as a bar colored by species (human, purple; mouse, orange) below a track showing ChIP enrichment, similarly colored by species.

type mouse and human hepatocytes. We found that 77% of the regions of H3K4me3 enrichment were shared in both WtHsChr21 and WtMmChr16. These regions are similar in a number of features, including proximity to TSSs (77 out of 101) and presence of CpG islands (80 out of 101). Consistent with H3K4me3 serving as an anchor for the basal transcriptional machinery, for almost every shared region enriched for H3K4me3 in human hepatocytes (97 out of 101), RNA transcripts were found in the liver-derived cell line HepG2 (*16*).

Regions enriched in trimethylation of H3K4 located distal to known TSSs are thought to represent unannotated promoter regions (*11*, *25*). The vast majority of the species-specific regions enriched in H3K4me3 in human hepatocytes (28 out of 36) and mouse hepatocytes (22 out of 22) were distal to TSSs (Fig. 1 and fig. S8). These species-specific sites of H3K4me3 enrichment were less likely to have CpG islands (3 out of 36 and 2 out of 22, respectively) and showed somewhat lower enrichment than the conserved regions (fig. S8). Consistent with their association with unannotated TSSs, human-specific regions enriched for trimethylation of H3K4 also showed evidence of transcription in HepG2 (26 out of 36 and 12 out of 22, respectively). In sum, H3K4me3 enrichment was found to be shared in both wild-type mouse and human hepatocytes at the majority of TSSs, yet largely divergent elsewhere.

On the basis of the presence of the trimethylated form of H3K4 in both mouse and human we observed at TSSs, we expected that a human chromosome subject to mouse developmental remodeling would have enrichment of H3K4me3 at similar positions near TSSs. It was unclear, however, whether the mouse transcriptional machinery would successfully recreate the human-specific histone modifications at uncharacterized promoters distal to known TSSs. Observing H3K4me3 enrichment on TcHsChr21 at either the human-unique sites on WtHsChr21 or the mouse-unique sites on WtMmChr16 could suggest what mechanisms direct the location of transcriptional initiation.

We found that virtually all of the TSSs and about three-quarters of non-TSS H3K4me3-enriched regions on WtHsChr21 were found at the same location on TcHsChr21 (Fig. 2 and fig. S4). We found a minority of cases (7 out of 78) where H3K4me3 enrichment occurred at sites on the TcHsChr21 that aligned with H3K4me3-enriched sites on TcMmChr16, without significant signal in WtHsChr21 (Fig. 2). Although these could be examples where human sequence in a mouse environment is handled in a mouse-specific manner, most are marginally enriched for H3K4me3 (see supporting online text 1). Taken as a whole, close inspection of the patterns of enrichment of H3K4me3 on TcHsChr21 reveals that 85% of H3K4me3-enriched regions found on WtHsChr21 were reproduced on TcHsChr21 (fig. S4); the remarkable extent of this similarity is shown for the liver-expressed gene *CLDN14* as a typical example (Fig. 3). Independent ChIP sequencing (ChIP-seq) experiments confirmed 93% (77 out of 82) of the sites of H3K4me3 enrichment on TcHsChr21 and 73% of sites on TcMmChr16 (70 out of 95); the majority of non-confirmed sites on TcMmChr16 (20 out of 25) were mouse-unique, half of which (13 out of 25) were found in the *Tiam1* gene (see supporting online text 1 and fig. S9).

In addition to expanding the examples of functionally conserved H3K4me3 sites, our results demonstrate that the regions of differential H3K4 methylation between divergent species are primarily dictated by cis-acting genetic sequence. Neither the cellular environment nor differences among the mouse and human chromatin–remodeling complexes substantially influence the placement of key chromatin landmarks associated with transcriptionally active regions.

Having shown that transcription factor binding and transcription initiation occurred in positions largely determined by underlying genetic sequences, we finally examined how the Tc1 mouse environment affects gene expression originating from the human chromosome. Using human gene expression microarrays that had been computationally and experimentally confirmed to be unaffected by the presence of mouse transcripts, we identified a distinct set of human genes that was expressed reproducibly in Tc1 mouse hepatocytes (Fig. 4A). Genes located in regions known to be deleted from TcHsChr21 were not detected as expressed (fig. S10) (*14*). Unsupervised clustering and principal component analysis of transcriptional data from the human gene expression microarrays clearly separated Tc1 and wild-type littermates by the presence of TcHsChr21 (fig. S10). Conversely, we asked whether the presence of the human chromosome perturbs mouse genome–based gene expression. No differential expression of mouse hepatocyte mRNA between Tc1 mice and wild-type littermates was detected by mouse-specific Illumina BeadArrays [note vertical scale in (Fig. 4B)]. Unsupervised clustering of the normalized mouse array data accurately grouped mice by litter and strain, independently of the absence or presence of the human chromosome (fig. S10).

We asked how well the transcripts originating from TcHsChr21 correlated with the transcripts originating from WtHsChr21 in human hepatocytes (Fig. 4C and fig. S11). Gene expression in Tc1 mouse hepatocytes originating from the human chromosome was determined by using the probes representing the 121 genes present on TcHsChr21 and then compared with matching gene expression data for the same 121 genes obtained from human hepatocytes. We found a strong correlation between the expression levels of the human genes located in Tc1 mouse hepatocytes and their counterparts located in wild-type human hepatocytes (Fig. 4C and fig. S11). This correlation ($R \approx 0.90$) was slightly lower than that found between replicate individual human livers (fig. S12), yet appears to be higher than similar correlations previously reported between human and other primates (*26*, *27*). The expression of orthologous genes within Tc1 hepatocytes (i.e., TcHsChr21 versus TcMmChr16) is substantially more divergent, with $R \approx 0.28$ (Fig. 4D). It is possible that the correlation between mouse and human orthologs could be influenced by the experimental differences between platforms, as well as by microarray design peculiarities. To address this concern, we determined the relative rank-order of expression among the genes on WtHsChr21, TcHsChr21, and TcMmChr16 and then compared the ranked results. We found correlation trends similar to the above (fig. S11) (*15*).

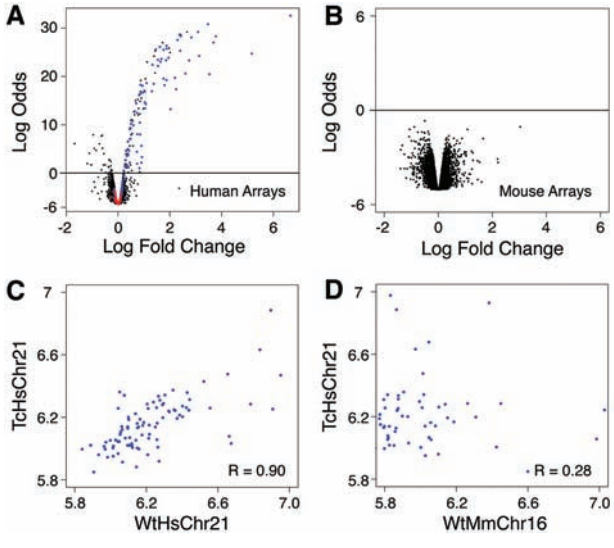


Fig. 4. Gene expression in the Tc1 mouse originating from the mouse and human chromosomes is largely indistinguishable from comparable wild-type nuclear environments. Volcano plots (empirical Bayes log odds of differential expression versus average log fold change) make several points. (**A**) Tc1 hepatocytes have high transcription occurring from the transplanted human chromosome 21, when we used human genomic arrays and wild-type littermate mRNA as a reference (black probes map to human genes; blue probes map to genes located on HsChr21; red probes map to regions absent from TcHsChr21); however, (**B**) wild-type and Tc1 mouse gene expression on mouse genomic arrays have indistinguishable patterns of transcription (black probes map to mouse genes). (**C**) Plot of the log expression of TcHsChr21 (*y* axis) transcripts versus WtHsChr21 (*x* axis) transcripts ($R \approx 0.90$). (**D**) Plot of the log expression of TcHsChr21 (*y* axis) transcripts versus WtMmChr16 (*x* axis) orthologous transcripts ($R \approx 0.28$).

Our results test the hypothesis that variation in gene expression is dictated by regulatory regions, extending recent studies of expression by quantitative trait-loci mapping and comparative expression studies that have been confined to closely related species (*26*–*30*). The apparent absence of overt trans influences could be explained by the modest amount of human DNA provided by a single copy of human chromosome 21 when compared with the complete mouse genome, as well as the absence of liver-specific transcriptional regulators on chromosome 21. The extent to which protein coding and cis-regulatory mutations contribute to changes in morphology, physiology, and behavior is actively debated in evolutionary biology (*3*, *12*, *13*). Myriad points of control influence gene expression; however, it has also been an unresolved question as to which of these mechanisms has the most influence globally. Here, we show that each layer of transcriptional regulation within the adult hepatocyte, from the binding of liver master regulators and chromatin remodeling complexes to the output of the transcriptional machinery, is directed primarily by DNA sequence. Although conservation of motifs alone cannot predict transcription factor binding, we show that within the genetic sequence there must be embedded adequate instructions to direct species-specific transcription.

References and Notes

1. E. H. Davidson, D. H. Erwin, *Science* **311**, 796 (2006).
2. K. S. Zaret, *Mech. Dev.* **92**, 83 (2000).
3. G. A. Wray, *Nat. Rev. Genet.* **8**, 206 (2007).
4. B. Li, M. Carey, J. L. Workman, *Cell* **128**, 707 (2007).
5. E. Guccione *et al.*, *Nat. Cell Biol.* **8**, 764 (2006).
6. L. Elnitski, V. X. Jin, P. J. Farnham, S. J. Jones, *Genome Res.* **16**, 1455 (2006).
7. D. T. Odom *et al.*, *Nat. Genet.* **39**, 730 (2007).
8. A. M. Moses *et al.*, *PLOS Comput. Biol.* **2**, e130 (2006).
9. A. R. Borneman *et al.*, *Science* **317**, 815 (2007).
10. E. Birney *et al.*, *Nature* **447**, 799 (2007).
11. B. E. Bernstein *et al.*, *Cell* **120**, 169 (2005).
12. H. E. Hoekstra, J. A. Coyne, *Evolution* **61**, 995 (2007).
13. S. B. Carroll, *Cell* **134**, 25 (2008).
14. A. O'Doherty *et al.*, *Science* **309**, 2033 (2005).
15. Materials and methods are available as supporting material on *Science* Online.
16. D. Kampa *et al.*, *Genome Res.* **14**, 331 (2004).
17. J. S. Carroll *et al.*, *Cell* **122**, 33 (2005).
18. S. Cereghini, *FASEB J.* **10**, 267 (1996).
19. J. Eeckhoute, B. Oxombre, P. Formstecher, P. Lefebvre, B. Laine, *Nucleic Acids Res.* **31**, 6640 (2003).
20. F. M. Sladek, M. D. Ruse Jr., L. Nepomuceno, S. M. Huang, M. R. Stallcup, *Mol. Cell. Biol.* **19**, 6509 (1999).
21. A. Rada-Iglesias *et al.*, *Hum. Mol. Genet.* **14**, 3435 (2005).
22. M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch, R. A. Young, *Cell* **130**, 77 (2007).
23. M. Vermeulen *et al.*, *Cell* **131**, 58 (2007).
24. R. J. Sims 3rd *et al.*, *Mol. Cell* **28**, 665 (2007).
25. A. Barski *et al.*, *Cell* **129**, 823 (2007).
26. Y. Gilad, A. Oshlack, G. K. Smyth, T. P. Speed, K. P. White, *Nature* **440**, 242 (2006).
27. P. Khaitovich *et al.*, *Science* **309**, 1850 (2005).
28. P. J. Wittkopp, B. K. Haerum, A. G. Clark, *Nat. Genet.* **40**, 346 (2008).
29. C. C. Park *et al.*, *Nat. Genet.* **40**, 421 (2008).
30. Y. Gilad, S. A. Rifkin, J. K. Pritchard, *Trends Genet.* **24**, 408 (2008).

31. We are grateful to E. Jacobsen, R. Stark, I. Spiteri, B. Liu, J. Marioni, A. Lynch, J. Hadfield, N. Matthews, the Cambridge Research Institute (CRI) Genomics Core, CRI Bioinformatics Core, and Camgrid for technical assistance, and B. Gottgens and J. Ferrer for insightful advice. Supported by the European Research Council (D.T.O.); Royal Society Wolfson Research Merit Award (S.T.); Hutchison Whampoa (D.T.O., ST); Medical Research Council (E.F., VT); Wellcome Trust (E.F., V.T.); University of Cambridge (D.T.O., D.S., N.B.M., S.T.); and Cancer Research U.K. (D.T.O., M.D.W., N.B.M., S.T., D.S.). Data deposited under ArrayExpress accession numbers E-TABM-473 and E-TABM-474. M.D.W., N.B.M., D.S., D.T.O., and C.M.C. designed and performed experiments; N.B.M., M.D.W., and D.S. analyzed the data; L.V., V.T., M.D.W., and E.F. created, prepared, and provided Tc1 mouse tissues; and M.D.W., N.B.M., D.T.O., and S.T. wrote the manuscript. D.T.O. oversaw the work. The authors declare no competing interests.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1160930/DC1
Materials and Methods
SOM Text
Figs. S1 to S12
Table S1

27 May 2008; accepted 3 September 2008
Published online 11 September 2008;
10.1126/science.1160930
Include this information when citing this paper.

Surface Sites for Engineering Allosteric Control in Proteins

Jeeyeon Lee,[1]* Madhusudan Natarajan,[2]* Vishal C. Nashine,[1] Michael Socolich,[2] Tina Vo,[2] William P. Russ,[2] Stephen J. Benkovic,[1] Rama Ranganathan[2]†

Statistical analyses of protein families reveal networks of coevolving amino acids that functionally link distantly positioned functional surfaces. Such linkages suggest a concept for engineering allosteric control into proteins: The intramolecular networks of two proteins could be joined across their surface sites such that the activity of one protein might control the activity of the other. We tested this idea by creating PAS-DHFR, a designed chimeric protein that connects a light-sensing signaling domain from a plant member of the Per/Arnt/Sim (PAS) family of proteins with *Escherichia coli* dihydrofolate reductase (DHFR). With no optimization, PAS-DHFR exhibited light-dependent catalytic activity that depended on the site of connection and on known signaling mechanisms in both proteins. PAS-DHFR serves as a proof of concept for engineering regulatory activities into proteins through interface design at conserved allosteric sites.

Proteins typically adopt well-packed three-dimensional structures in which amino acids are engaged in a dense network of contacts (*1*, *2*). This emphasizes the energetic importance of local interactions, but protein function also depends on nonlocal, long-range communication between amino acids. For example, information transmission between distant functional surfaces on signaling proteins (*3*), the distributed dynamics of amino acids involved in enzyme catalysis (*4*–*6*), and allosteric regulation in various proteins (*7*) all represent manifestations of nonlocal interactions between residues. To the extent that these features contribute to defining biological properties of protein lineages, we expect that the underlying mechanisms represent conserved rather than idiosyncratic features in protein families.

On the basis of this conjecture, methods such as statistical coupling analysis (SCA) quantitatively examine the long-term correlated evolution of amino acids in a protein family—the statistical signature of functional constraints arising from conserved communication between positions (*8*, *9*). This approach has identified sparse but physically connected networks of coevolving amino acids in the core of proteins (*8*–*12*). The connectivity of these networks is remarkable, given that a small fraction of total residues are involved and that no tertiary structural information is used in their identification. Empirical observation in several protein families shows that these networks connect the main functional site with distantly positioned secondary sites, enabling predictions of allosteric surfaces at which binding of regulatory molecules (or covalent modifications) might control protein function. Both literature studies and forward experimentation in specific model systems confirm these predictions (*8*–*12*). Thus, techniques such as SCA may provide a general tool for computational prediction of conserved allosteric surfaces.

The finding that certain surface sites might be statistical "hotspots" for functional interaction with active sites suggests an idea for engineering new regulatory mechanisms into proteins. What if two proteins were joined at surface sites such that their statistically correlated networks were juxtaposed and could form functional interactions (Fig. 1A)? If the connection sites are functionally linked to their respective active sites

[1]Department of Chemistry, Pennsylvania State University, University Park, PA 16802, USA. [2]Green Center for Systems Biology and Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: rama.ranganathan@utsouthwestern.edu

www.sciencemag.org/cgi/content/full/1160930/DC1

Supporting Online Material for

Species-Specific Transcription in Mice Carrying Human Chromosome 21

Michael D. Wilson, Nuno L. Barbosa-Morais, Dominic Schmidt, Caitlin M. Conboy, Lesley Vanes, Victor L. J. Tybulewicz, Elizabeth M. C. Fisher, Simon Tavaré, Duncan T. Odom*

*To whom correspondence should be addressed. E-mail: duncan.odom@cancer.org.uk

Published 11 September 2008 on *Science* Express
DOI: 10.1126/science.1160930

This PDF file includes

MATERIALS AND METHODS

SUPPORTING TEXT 1: Detailed analysis of H3K4me3 enrichment between WtHsChr21, TcHsChr21 and WtTcMmChr16.

SUPPLEMENTAL FIGURES

Figure S1. Genotyping of hepatocytes from nine Tc1 mice shows that one copy of TcHsChr21 is present on average in 61% of cells.

Table S1. Percent identity of transcription factors in this study.

Figure S2. The distributions of ratios in probes that are unbound, shadow, or bound.

Figure S3. Transcription factor binding and transcription initiation events on TcMmChr16 in the Tc1 mouse are not perturbed by the presence of the transplanted TcHsChr21.

Figure S4. Most transcription factor binding and H3K4me3 enriched regions on TcHsChr21 were consistent with those found in human hepatocytes.

Figure S5. Human transcription initiation and transcription factor binding events that are recapitulated in Tc1 hepatocytes show stronger enrichment signal than events which are not.

Figure S6. Comparison of transcription factor binding and H3K4me3 enrichment between TcHsChr21 and TcMmChr16.

Figure S7. p-value calculations obtained by chi-squared tests of associations.

Figure S8. Human transcription initiation events at TSS are significantly more enriched than events which are distal to TSS.

Figure S9. Independent validation of HNF4a, HNF6 and H3K4me3 microarray data using ChIP-seq.

Figure S10. Gene expression comparison of hepatic transcription in wild-type human, wild-type mouse, and Tc1 mouse.

Figure S11. Correlation in gene expression originating from HsChr21 and MmChr16 in wild-type human, wild-type mouse, and Tc1 mice in hepatocytes.

Figure S12. Gene expression correlations among replicates.

DATA ACCESSION NUMBERS - ARRAYEXPRESS

Gene expression: E-TABM-473
Chromatin immunoprecipitation microarrays: E-TABM-474

MATERIALS AND METHODS

Molecular Biology and Genomics

Mouse material. The Tc1 mouse line was generated as previously described (O'Doherty et al 2005). Tc1 mice used in this study were bred by crossing female Tc1 mice to a male (129S8 x C57BL/6J)F1 mouse. Liver material was prepared for chromatin immunoprecipitation (ChIP) and mRNA expression analysis as previously described (Odom et al 2007). For each mouse ChIP, and each mRNA expression experiment, biological replicates consisted of hepatocytes from a single animal.

Human material. Crosslinked, healthy human hepatocytes were obtained from the Liver Tissue Distribution Program (NIDDK Contract #N01-DK-9-2310) at the University of Pittsburgh (K. Dorko, S. Strom). After receipt, these cells were resuspended into HBSS, portioned into aliquots of 2.5 x 10^7 hepatocytes, and stored frozen at -80°C until used in experiments. Human ChIPs were performed with either individual or pooled mixtures of hepatocytes from donors of mixed gender and ages. Expression analysis was performed on total RNA extracted from two individual flash frozen adult liver samples as well as a commercial mixed donor total RNA sample from Ambion (AM7960).

Species:	Human	Wild-type mouse		Tc1 mouse	
Array :	HsChr21	HsChr21	MmChr16	HsChr21	MmChr16
HNF1α	3	2	2	2	2
HNF4α	3	3	2	3	4
HNF6	2	3	2	2	2

Number of biological replicates used for ChIP-chip experiments reported in this study. Mouse wild-type refers to Tc1 littermates that do not carry human chromosome 21 except for HNF6 where biological replicates from previous experiments were used ((C57BL/6 x A)F1/J; Odom *et al.* 2007).

Chromatin immunoprecipitations (ChIP). ChIP experiments with human and mouse hepatocytes cells were performed in replicate as previously described (Odom et al. 2007). Antibodies used were: HNF1α (sc-6547); HNF4α (sc-8987); HNF6 (sc-13050) and H3K4me3 (ab8580).

Microarrays. ChIP-chip experiments were hybridized to commercially available Agilent Technologies microarrays designed against human chromosome 21 (AMADID 014841) and mouse chromosome 16 (AMADID 015340) as recommended by the manufacturer's "Agilent Mammalian ChIP-on-chip" protocol version 9.1.

Briefly, the immunoprecipitated material was labelled with Cyanine 5-dUTP and the input control was labelled with Cyanine 3-dUTP (Enzo life sciences) using BioPrime Array CGH Genomic Labeling System kit following the manufacturer's protocol. Unincorporated dyes were removed using QIAquick PCR clean-up kit. Equal amounts of Cy5 and Cy3 labelled DNA was combined and hybridized at 65 deg C to microarrays

using 2X Hi-RPM Hybridization Buffer Gene expression and manufacturer's protocols. After 40 hours hybridization arrays were washed with Agilent Array CGH wash buffers 1 and 2 following the manufacturer's protocol and scanned using the Agilent scanner. Raw data was extracted using the Agilent Feature Extraction Software and processed as mentioned below.

Gene expression experiments. Flash frozen mouse and human liver material was homogenized in QIAzol reagent using a Precellys bead grinder homogenizer. Samples were extracted with chloroform and total RNA was isolated with Qiagen miRNeasy kit using the manufacturer's protocol. Total RNA samples were labelled with Illumina-Totalprep RNA Amplification kit (Ambion) following manufacturer's instructions. Briefly: 1. 250ng of input total RNA was used for First strand cDNA synthesis (2 hours at 42 deg C) using oligo(dT) primer and ArrayScript enzyme; 2. Second strand cDNA synthesis (2 hours at 16 deg C) using DNA polymerase and RNAse H; 3. cDNA purification using purifying columns; 4.cRNA in vitro transcription using Biotin-NTP and cRNA purification using purifying columns; and 5. Quality and quantitative QC's were done separately. Hybridization was done using the IntelliHyb Seal method according to the manufacturer's instructions. Analysis was performed using Illumina Sentrix Human-6 version 2 and Mouse-6 version 1.1 Expression BeadChip microarrays. Default scanner settings for DirectHyb gene expression protocol were used in this experiment.

ChIP-sequencing. Solexa libraries were prepared following the instructions of Illumina (Sample preparation for genomic DNA — version 2.2) with the following modifications. The ChIP-enriched DNA and input DNA were not further fragmented. After end-repair and addition of an 'A' base to the 3' ends, the adapters were ligated to the ends of the DNA Fragments using 2 µl of 'Adapter oligo mix' in a total reaction volume of 25 µl. Between these steps, the DNA was purified using the DNA Clean&Concentrator-5 kit (Zymo Research). Subsequently, the DNA was amplified by 18 cycles of PCR, purified with QIAquick PCR purification Kit, and eluted with 33.5 µl of 10 mM tris buffer at pH7.0. The PCR-product was sized fractionated on 2% agarose gel and a gel slice containing the 200-300 bp fragments was excised. The flowcells were prepared and processed according to the manufacturer's protocols, with single-end sequencing for 36 cycles.

Computational Biology and Data Analysis

ChIP-chip. Raw ChIP-chip data were read into the statistical software environment *R*. Quality assessment, within-array median normalization and enrichment analysis (computation of average ratios and associated statistics) were performed using tools included in the *limma* package available through the Bioconductor project. Integration of genome mapped enrichment ratios and associated B-statistics drove the preliminary automated detection of putative binding sites. Ratios and genomic locations were also used to provide an estimate for enrichment intensities. Those ChIP-derived binding sites and associated classification were manually curated by visualizing the corresponding tracks on the UCSC Genome Browser. Curation included automated and visual analysis of ChIP-chip data from cross-hybridizations of each platform with DNA from the heterologous species. Binding sites potentially resulting from heterologous cross-

hybridization were removed. Mouse (mm8) to Human (hg18) genomic cross-mapping relied on the Golden Path chained blastz alignments downloaded from UCSC.

Validation of array data with ultra highthroughput sequencing. Raw Solexa data for H3K4me3, HNF6 and HNF4a were aligned to a 'Tc1 genome' that included the mouse genome (mm9) as well as human chromosome 21 (hg18). The sequencing of the input Tc1 genome DNA identified known duplicated and deleted regions (data not shown). The Genome analyser pipeline 0.3.0 using default parameters (Illumina) was used to align the 36-mer reads to the hybrid Tc1 genome sequence. Significantly enriched peaks were called using Model-based Analysis of ChIP-seq data (MACS; http://liulab.dfci.harvard.edu/MACS/) algorithm and inspected manually in order to validate the human and mouse array peaks.

Gene expression. Illumina bead level data were summarized, pre-processed and analyzed in *R* using the *beadarray* package available through Bioconductor. Summarized data were quantile normalized and log2 transformed. Analysis of differential expression relied on the B-statistic. In-house probe annotation and BioMart were used in the selection and assignation of human chromosome 21 genes and their orthologs in mouse for correlation studies.

Transcription start site analysis (TSS). Throughout this study, we define a transcription start site (TSS) as any region of the genome that overlaps with a transcription start site as annotated by any RefSeq or UCSC gene model in the human and mouse genomes; this definition includes all known alternative transcriptional start sites that may be utilized, even rarely or transiently.

SUPPORTING TEXT 1: *Detailed analysis of H3K4me3 enrichment between WtHsChr21, TcHsChr21 and WtTcMmChr16*

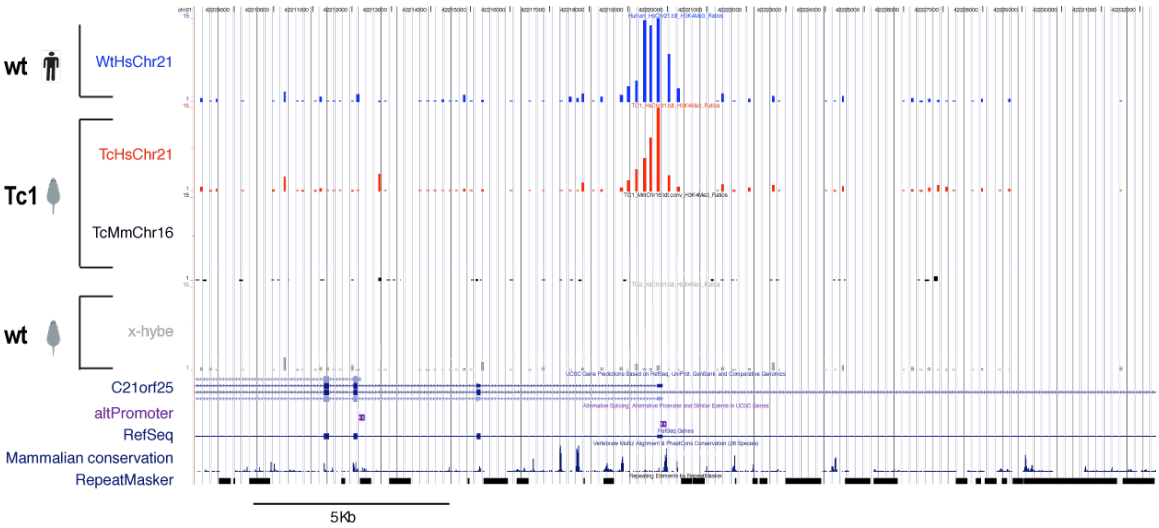
Wild-type mouse and human conserved H3K4me3 events [the regions missing from TcHsChr21 are not included in this discussion]:

67 percent (53/79) of the human H3K4me3 enriched regions were present in the orthologous position in the mouse genome. 81 percent of these (43/53) shared H3K4me3 enriched positions occurred at predicted transcription starts sites (TSS) as determined by overlap with the RefSeq or UCSC gene models in the human and mouse genomes. Ninety-three percent (40/43) of the conserved TSS H3K4me3 occupied regions contained CpG islands, all of which fell in regions transcribed in HepG2, a human liver cancer cell line, as determined by an in-depth, 5 bp resolution analysis of small and large RNA expression in the human liver cell line HepG2 (Kampa et al. 2004). Seven of the ten H3K4me3 conserved regions designated as putative non-TSS lacked a CpG island; remarkably, five of these were found within the hypothetical C21orf34 gene, which is known to give rise to several microRNAs. Of the remaining 5 events not associated with C21orf34, four are intronic, and one occurs on a conserved CpG island with no gene annotation. Most of the mouse TcMmChr16 (48/53) and human TcHsChr21 (49/53) conserved events were confirmed by Solexa DNA sequencing of an independent ChIP of H3K4me3 in a Tc1 mouse, where the human chromosome recapitulates the histone modification pattern found on WtHsChr21 (Figure S9).

Human and TcHsChr21 shared H3K4me3 events:

Only 2/18 H3K4me3 events shared by WtHsChr21 and TcHsChr21 (yet absent from TcMmChr16) were located at a TSS. In contrast to the conserved mouse and human H3K4me3 events, only 16 percent (3/18) of these human only events possessed a CpG island. However, 17/18 of these human-chromosome specific events fell within regions transcribed in HepG2 cells suggesting that they may have functional roles. Solexa sequencing confirmed 94 percent (17/18) of these events on TcHsChr21, and unambiguously confirmed that H3K4me3 is not present on the orthologous regions of TcMmChr16.

SUPPORTING TEXT 1 (cont'd):

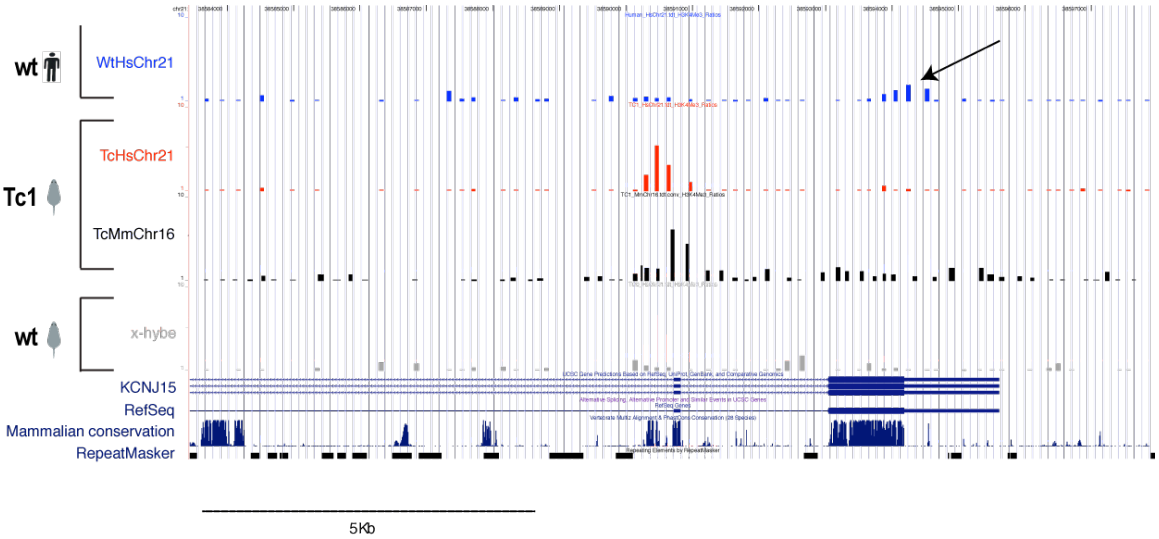


Example 1.1: *C21orf25*. Two of the 18 human-unique H3K4me3 enriched regions recapitulated on the TcHsChr21 are located at TSS. *C21orf25* is a clear example of a human specific H3K4me3 event occurring at an alternative promoter. The absence of signal in the mouse genome was confirmed by ChIP-seq. Cross hybridization (x-hybe) of wild-type mouse on human chromosome 21 microarrays is shown in grey.

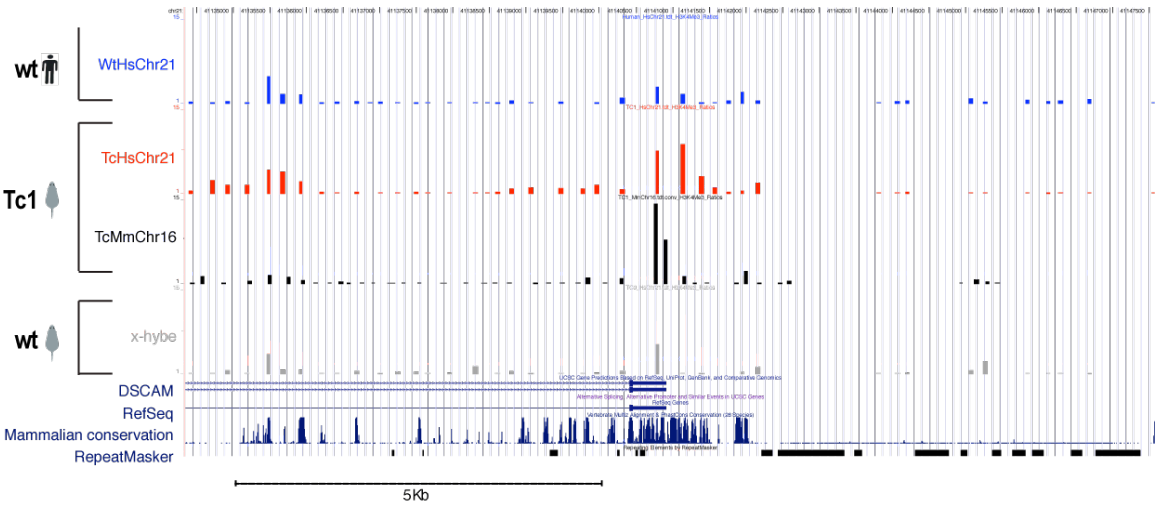
H3K4me3 enrichment events shared between TcHsChr21 and TcMmChr16:

Seven examples where H3K4me3 occurred at TcHsChr21 and TcMmChr16 orthologous sites without significant signal in WtHsChr21 were identified. 5/7 of these events occurred at TSS locations, one of which possessed a CpG island. 6/7 showed evidence of HepG2 expression. Solexa sequencing supported all of these events on TcHsChr21 and 5/7 of the orthologous regions of the mouse genome. These serve as examples where the human sequence in a mouse environment can be handled in a mouse specific manner. It is important to point out that some of these examples are marginal.

SUPPORTING TEXT 1 (cont'd):

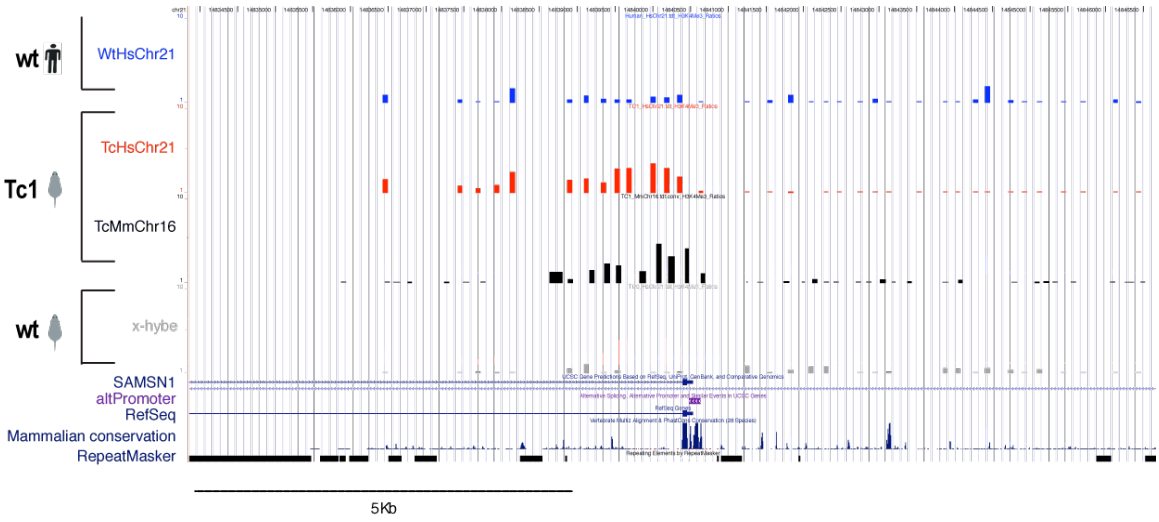


Example 2.1: *KCNJ15*. This is an example of a mouse dominant peak (black) that is also present in the TcHsChr21 (red) but to a much less extent in the human (blue). The weak human peak (blue) above the TSS (black arrow) suggests that this may be an example where transcription initiation has been redirected on TcHsChr21 to a wild-type mouse location. Cross hybridization (x-hybe) of wild-type mouse on human chromosome 21 microarrays is shown in grey.



Example 2.2: *DSCAM*. This is an example of a mouse peak (black) that is also present in the TcHsChr21 (red) but to a much lesser extent in the human (blue). This occurs at the TSS of *DSCAM*. Cross hybridization (x-hybe) of wild-type mouse on human chromosome 21 microarrays is shown in grey.

SUPPORTING TEXT 1 (cont'd):



Example 2.3: *SAMSN1*. This example occurs at an alternative TSS where TcMmChr16 and TcHsChr21 share an H3K4me3 enrichment while the WtHsChr21 has only weak signal (blue). Cross hybridization (x-hybe) of wild-type mouse on human chromosome 21 microarrays is shown in grey.

TcMmChr16 only H3K4me3:

Of the 33 mouse only events, 15 fell within one gene, *Tiam1*. 45 percent (15/33) of the TcMmChr16 only events showed evidence of HepG2 transcription in orthologous location on human Chromosome 21. Only 3 of the 33 events occurred at a TSS and one possessed CpG island. Solexa sequencing validated only 2/15 of H3K4me3 events on the *Tiam1* locus were whereas 13/18 of the remaining mouse-only events were validated.

HsChr21 only but no Tc1 H3K4me3:

Less than 10 percent (6/79) of the HsChr21 H3K4me3 events were not recapitulated in the mouse nuclear environment (neither TcHsChr21 or TcMmChr16). Only 2/6 of these showed evidence of HepG2 transcription one of which was located at a TSS.

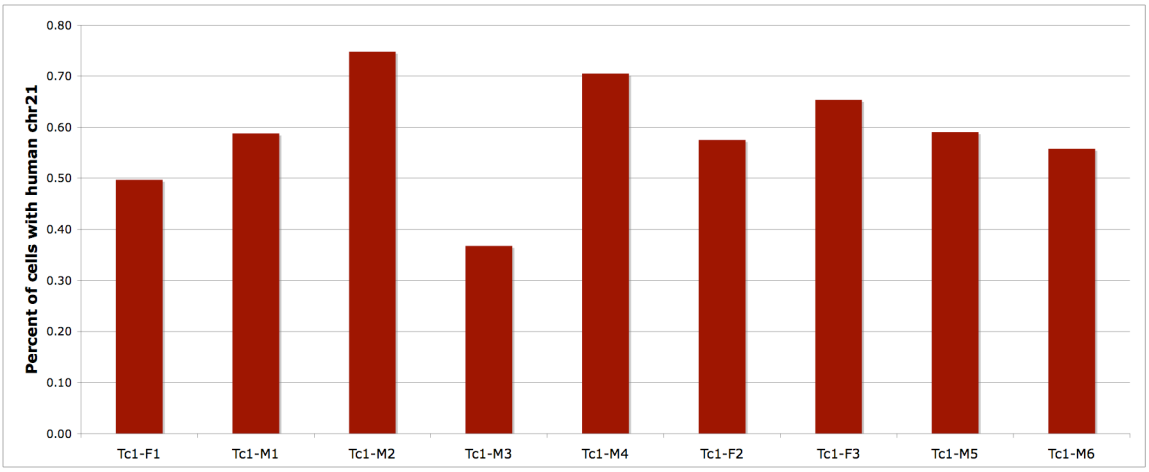


Figure S1. Genotyping of hepatocytes from nine Tc1 mice shows that one copy of TcHsChr21 is present on average in 61% of cells (n=9; SD =0.08). Genotyping was performed in triplicate using primers designed against regions of MmChr16 and HsChr21 respectively (mouse: 1-mm-fw CAGTGCGTGGACTTAGGAAA and 1-mm-rev GGCATTGCTCAAGACAGAAA; and human primers used: 1-hs-fw GGAAATCACGCCTGGTAGAT and 1-hs-rev GGTATCTGCAGCCCTCTCTC). Real Time PCR analysis was performed using the ABI7900HT and the Power SYBR Green kit (Applied BioSystems) according to the manufacturer's protocols. Both primer sets were determined to amplify species-specific products with similar efficiencies (data not shown).

	Amino acid differences	Aligned amino acids	Percent amino acid identity
HNF1α	30	628	95
HNF4α	22	464	95
HNF6	5	503	99

Table S1. Percent identity of transcription factors in this study. Reference protein sequences were aligned using CLUSTALW and gaps were removed before calculating amino acid differences.

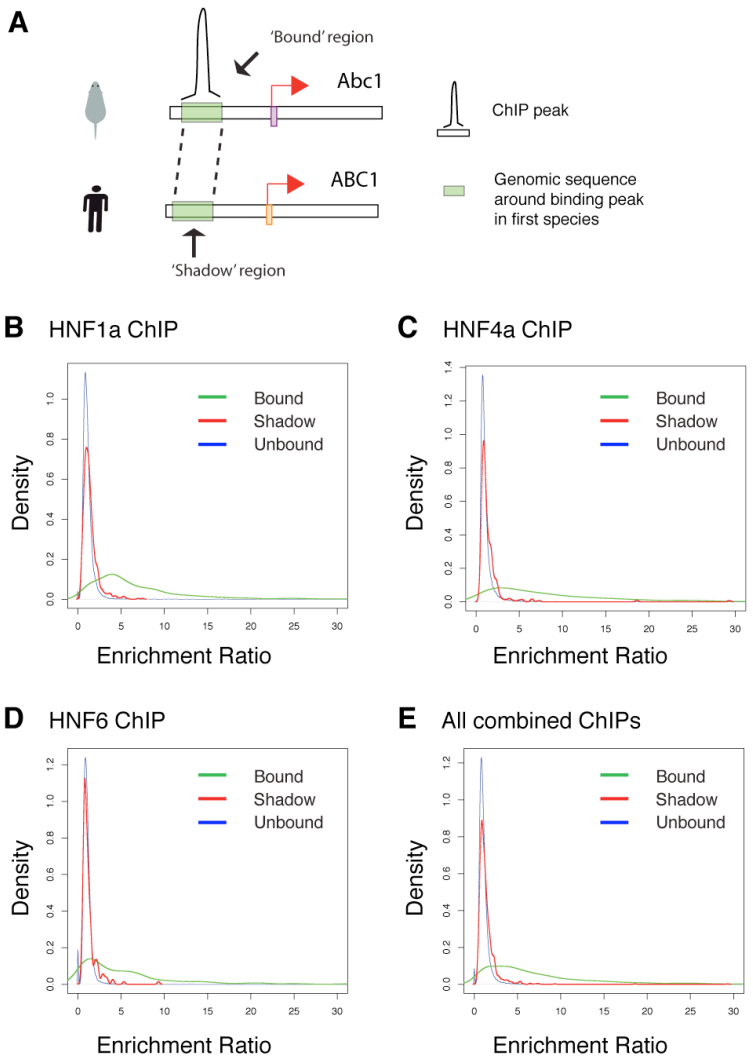
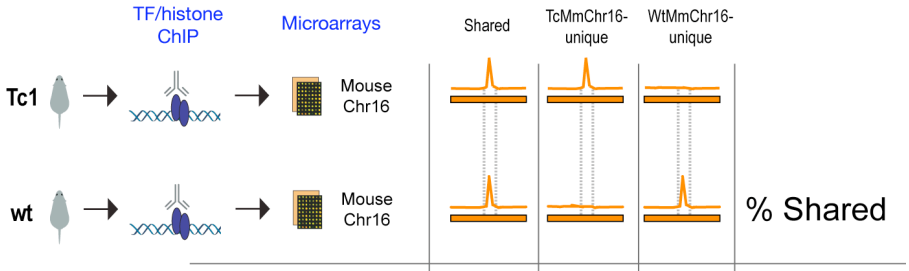
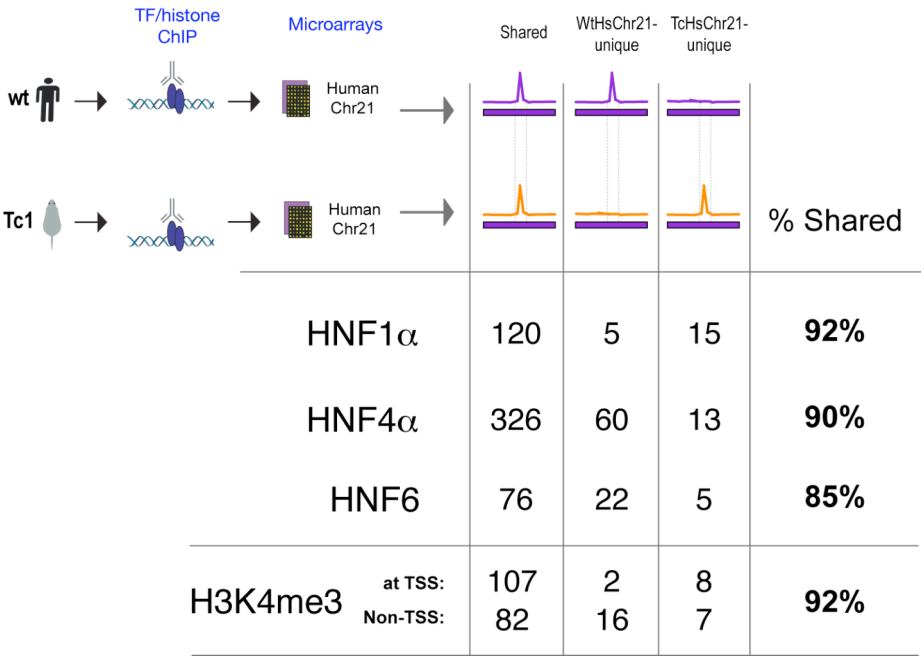


Figure S2. The distributions of ratios in probes that are unbound, shadow, or bound. Panel A shows how probes were assigned. In short, probes called as bound were categorized into a set. Probes in the second species in homologous regions not ChIP enriched for the same factor were then placed into a category called 'shadow', followed by all other probes ('unbound'). Panels B, C, and D show the distribution of ratios among these categories for three transcription factors in this study. Panel E represents all ChIP data combined into one plot. Shadow regions have a slight enrichment shift possibly due to the inclusion of some false negatives, but largely are indistinguishable from unbound probes. In contrast, bound probes in green typically have much more enrichment.



HNF1α		636	0	0	**100%**
HNF4α		419	2	0	**99%**
HNF6		295	8	26	**95%**
H3K4me3	**at TSS:**	90	0	0	**100%**
	Non-TSS:	105	0	0	

Figure S3. Transcription factor binding and transcription initiation events on TcMmChr16 in the Tc1 mouse are not perturbed by the presence of the transplanted TcHsChr21. All enriched mouse chromosome 16 events (including those that are not alignable to human chromosome 21) were determined using ChIP of three transcription factors as well as H3K4me3 followed by hybridization to Agilent 244K chromosome 16 microarrays. Enriched regions were compared between Tc1 and wild-type mice (TcMmChr16 vs WtMmChr16). Mouse wild-type refers to Tc1 littermates that do not carry human chromosome 21 except for HNF6 where biological replicates from previous experiments were used ((C57BL/6 x A)F1/J; Odom *et al.* 2007). Percent shared was determined by adding the complete number of binding events together, and dividing the shared number by the total.



HNF1α		120	5	15	**92%**
HNF4α		326	60	13	**90%**
HNF6		76	22	5	**85%**
H3K4me3	at TSS:	107	2	8	**92%**
	Non-TSS:	82	16	7	

Figure S4. Most transcription factor binding and H3K4me3 enriched regions on TcHsChr21 were consistent with those found in human hepatocytes. All enriched human chromosome 21 events (including those that are not alignable to the mouse genome) were determined using ChIP of three transcription factors as well as H3K4me3 followed by hybridization to Agilent 244K human chromosome 21 microarrays. Notably, all of the transcription factors profiled in the Tc1 mouse are derived from the mouse genome.
A few genes harboured multiple wt-human unique (WtHsChr21) or Tc1-mouse unique (TcHsChr21) enriched events in more than one ChIP experiment. The most prevalent example of a human gene that was differentially regulated in the mouse nuclear environment comes from *C21orf34*. *C21orf34* encodes a short hypothetical protein and several non-coding RNAs (mir-99a, let-7c and mir-125b-2) and harbours a significant number of wt-human unique (wtHsChr21) events comprising: 10/18 H3K4me3 events, 3/60 HNF4α events, 2/5 HNF1α events, and 6/22 HNF6 events. Similarly, at least one wt-human unique event for experiments with HNF1α, HNF4α, HNF6 and H3K4me3 are observed for the solute carrier family 37 member 1 gene *SLC37A1*. Several Tc1-mouse (TcHsChr21) unique events for *RUNX1* can be observed for H3K4me3, HNF4α and HNF1α and these aberrant binding sites may be in part explained by their proximity to a deleted region of TcHsChr21. Disco interacting protein 2 homolog (*DIP2A*) contains TcHsChr21-unique for both H3K4me3 and HNF4a. The hormonally upregulated neu-associated kinase (*HUNK*) harbours a TcHsChr21 unique event for both HNF1α and HNF4α and a wt-human unique event for H3K4me3. Finally the following genes have at least one species unique event for two different factors:
DSCAM, HLCS, CLDN14, BC039377, APP and DOPEY2. While these species-specific events are the most likely candidates for being susceptible to trans-influences, overall these events are statistically of lower intensity (see Fig S5). Furthermore, within all of the above genes, several strong examples of shared events from the above ChIP experiments can also be found.

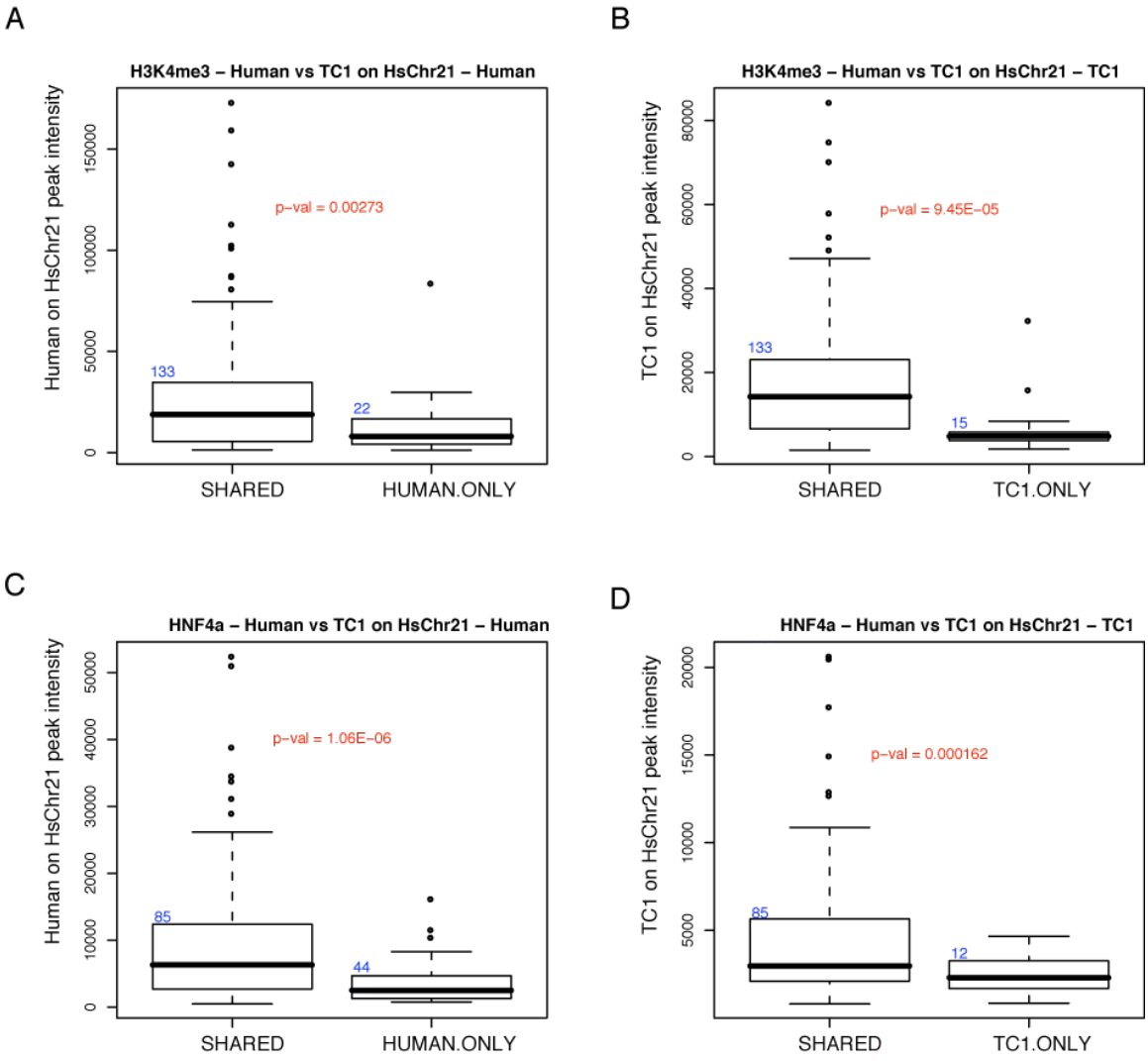
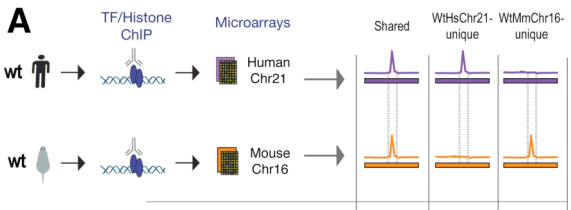
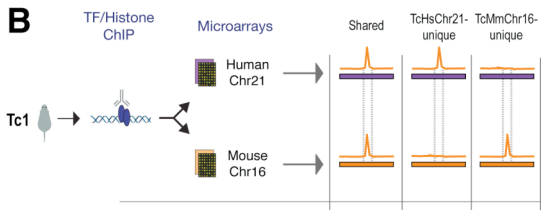


Figure S5. Human transcription initiation and transcription factor binding events that are recapitulated in Tc1 hepatocytes show stronger enrichment signal than events which are not. Panels A and B shows that recapitulated H3K4me3 events are more enriched than those which are WtHsChr21 only and TcHsChr21 only respectively. Panels C and D show the same trend for HNF4α transcription factor binding.



		Shared	WtHsChr21-unique	WtMmChr16-unique
HNF1α		76	38	300
HNF4α		137	119	146
HNF6		28	60	115
H3K4me3	at TSS:	77	8	0
	Non-TSS:	24	28	22



		Shared	TcHsChr21-unique	TcMmChr16-unique
HNF1α		39	45	116
HNF4α		62	99	140
HNF6		11	38	83
H3K4me3	at TSS:	48	2	3
	Non-TSS:	12	20	32

Figure S6. Comparison of transcription factor binding and H3K4me3 between TcHsChr21 and TcMmChr16. Panel (A) reproduces the data in Figure 1, and shows the divergence of transcription factor binding and histone modifications on orthologous regions of WtHsChr21 and WtMmChr16 between wild-type mouse and human hepatocytes. Panel (B) is the comparable data to (A) from TcHsChr21 and TcMmChr16 obtained from Tc1 mouse hepatocytes. Some numbers in panel B are lower due to deletions from TcHsChr21 caused by creation of the Tc1 mouse (O'Doherty, *et al.* 2005).

	WtHsChr21:WtMmChr16 ***versus:***		
TF binding	WtMmChr16:TcMmChr16	WtHsChr21:TcHsChr21	TcHsChr21:TcMmChr16
HNF1a	9.8×10^{-168}	2.0×10^{-46}	0.82
HNF4a	9.6×10^{-89}	5.7×10^{-42}	1.2×10^{-4}
HNF6	3.6×10^{-67}	4.5×10^{-25}	0.18

Figure S7. p-value calculations obtained by chi-squared tests of associations, comparing the proportions of shared and unshared binding events of WtHsChr21 and WtMmChr16 to the proportions found between each relevant pair combination of other chromosomes. Chi-squared tests indicate that, relative to the proportions found for WtHsChr21 and WtMmChr16, the proportions found between WtHsChr21 and TcHsChr21 as well as between WtMmCh16 and TcMmChr16 are significantly different (each p-value $<< 1 \times 10^{-25}$), whereas differences between the WtHsChr21/WtMmChr16 and TcHsChr21/ TcMmChr16 comparisons are considerably closer to unity (p-value $>> 1 \times 10^{-4}$). Together, these data thus indicate a high degree of similarity between the reference patterns (WtHsChr21 v WtMmChr16) and the test patterns (TcHsChr21 v TcMmChr16) (as in Figure S6).

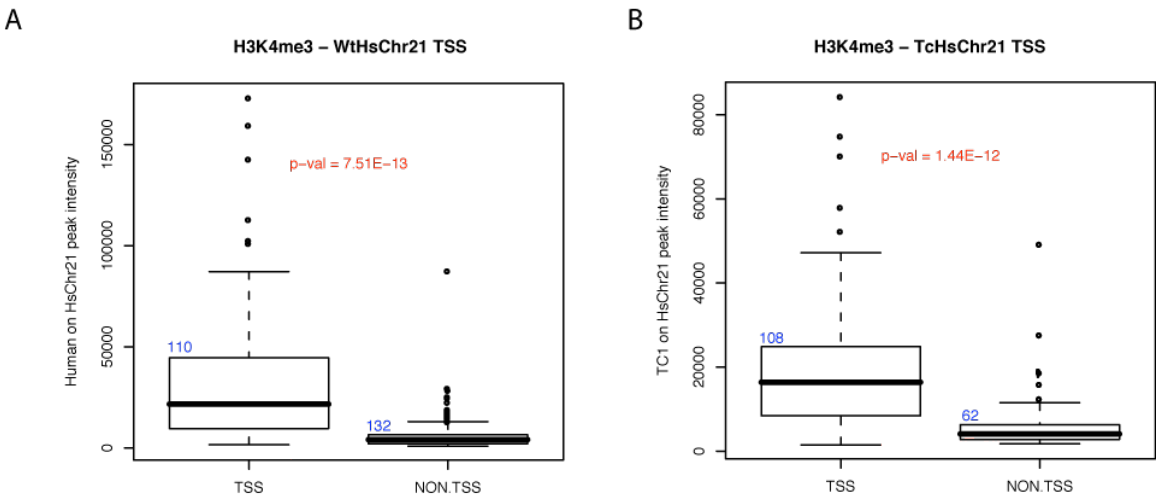
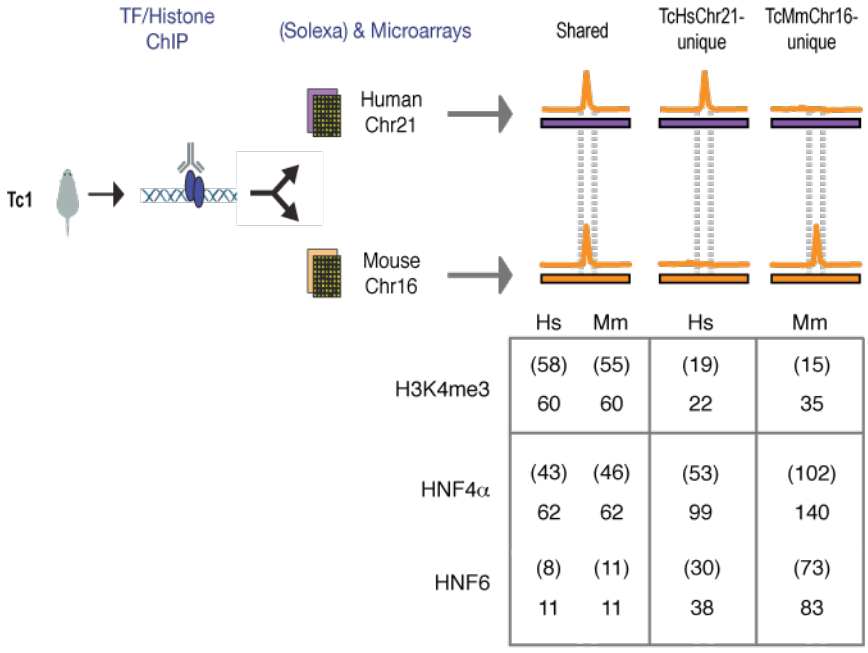


Figure S8. Human transcription initiation events at TSS are significantly more enriched than events distal to TSS. H3K4me3 events at TSS and non-TSS were compared within (A) WtHsChr21 and (B) TcHsChr21. p-values for each comparison are shown in red.



	Hs	Mm	Hs	Mm
H3K4me3	(58) 60	(55) 60	(19) 22	(15) 35
HNF4α	(43) 62	(46) 62	(53) 99	(102) 140
HNF6	(8) 11	(11) 11	(30) 38	(73) 83

Figure S9. Independent validation of H3K4me3, HNF4α, HNF6 and microarray data using Solexa sequencing. The number of validated events is shown in brackets above the total number of peaks called using microarrays. Most (13/20) TcMmChr16-unique H3K4me3 events that were not validated fell within a single gene (*Tiam1*; see **Supporting text 1** for a detailed explanation).

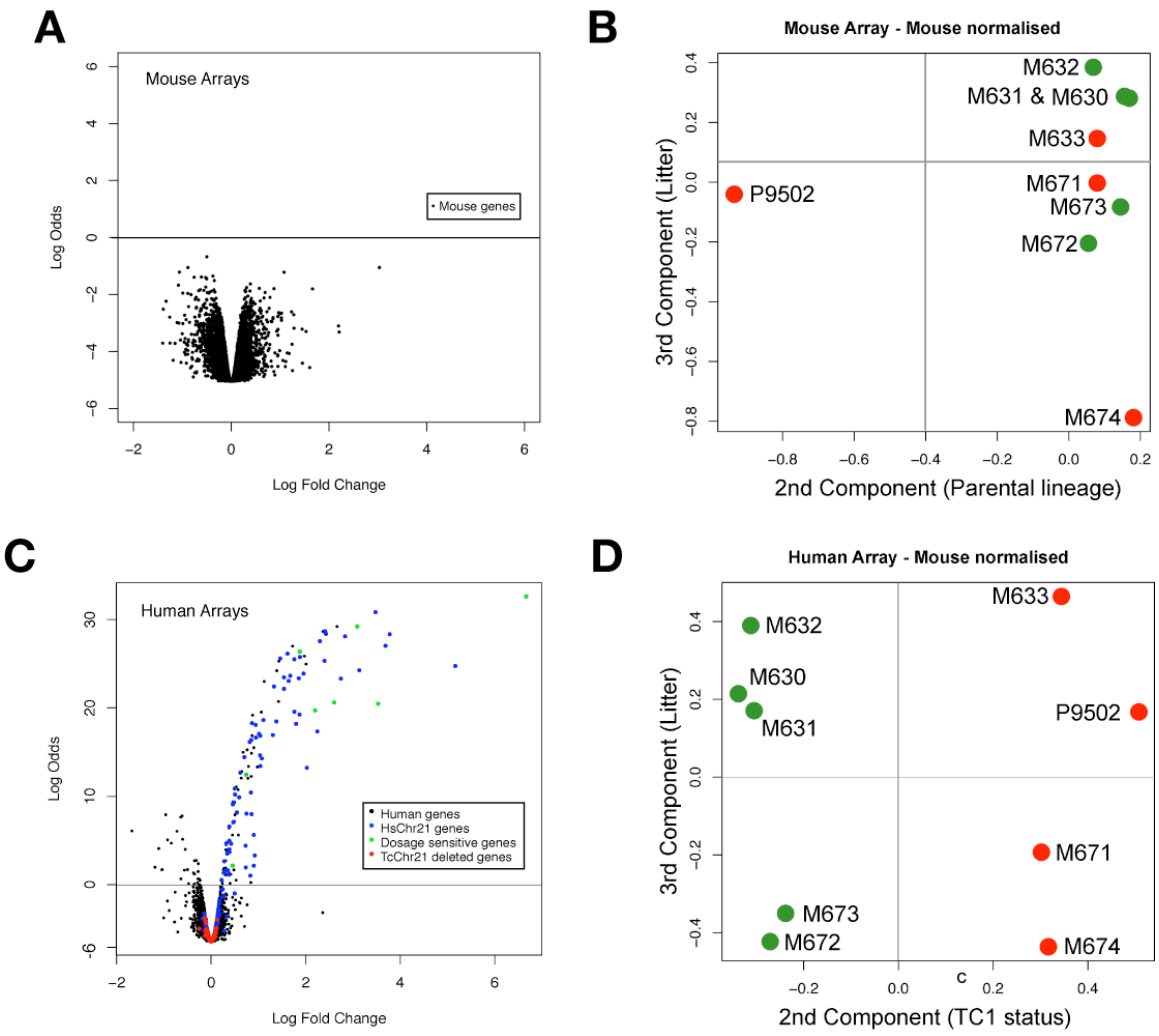


Figure S10. Gene expression comparison of hepatic transcription in wild-type human, wild-type mouse, and Tc1 mouse. (A) Volcano plot of wild-type (green) vs Tc1 (red) mouse-genome-driven gene expression (as in Fig 4). (B) Principal component analysis (PCA) of data from panel (A) clusters mice based on litter and background, but shows no substantial effects from the presence or absence of HsChr21. Mouse designations starting with either M63-, M73-, M67-, or P95- are age-matched siblings. (C) Volcano plot of transcripts in Tc1 mouse hepatocytes versus control mRNA obtained from wild-type littermates on human microarrays. Note that genes deleted from TcHsChr21 are colored red, and none of these show significant signal. Blue indicates the gene is located on TcHsChr21. In addition, a number of known dosage-dependent genes, indicated in green, are strongly expressed in Tc1 mice. (D) Principal component analysis of data from (C) clusters mice based on whether they carry the Tc1 chromosome or not, and secondarily by litter.

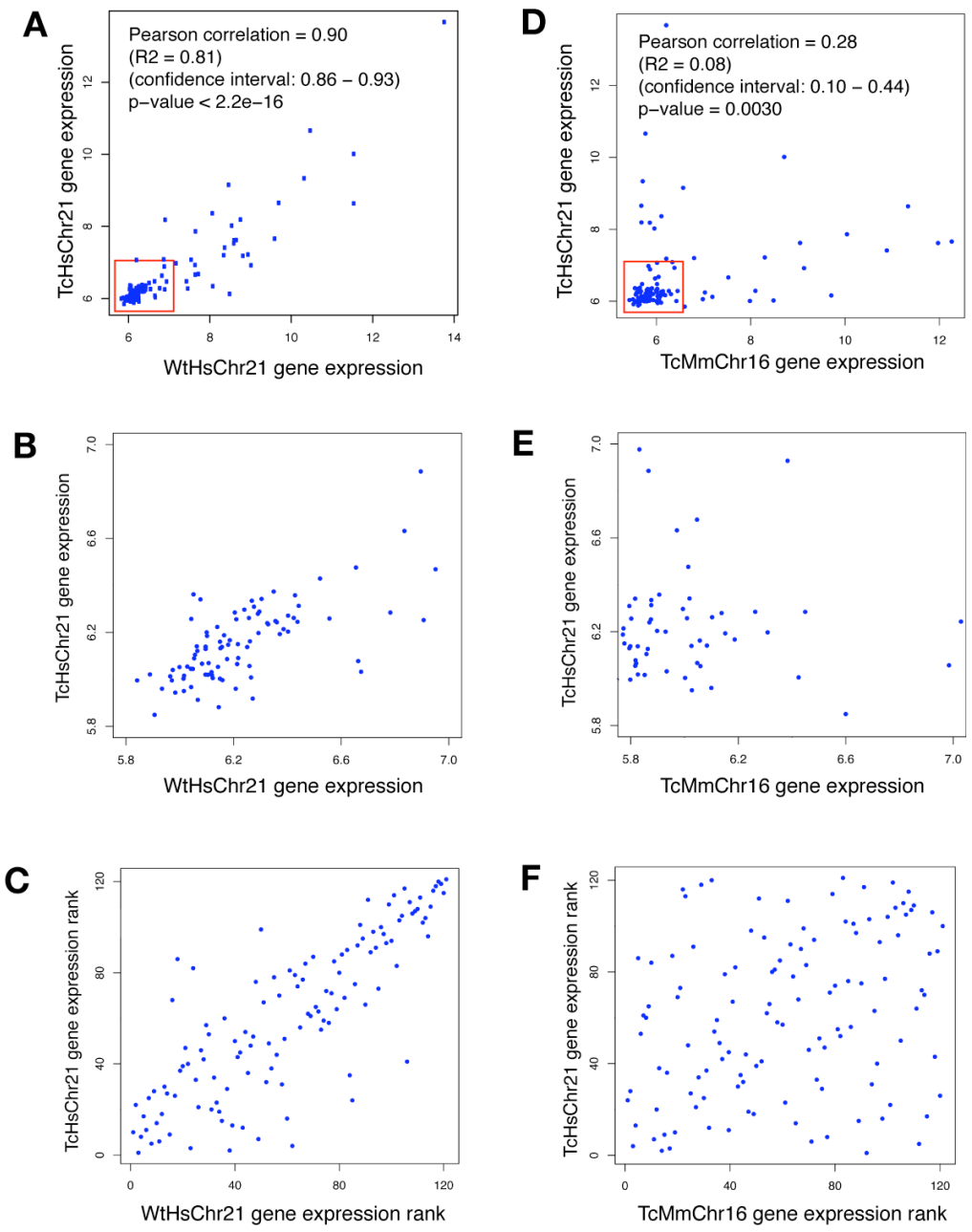


Figure S11. Correlation in gene expression originating from WtHsChr21 and WtMmChr16 in wild-type human, wild-type mouse, and Tc1 mice in hepatocytes. (A-C) Gene expression correlations were made between genes expressed on WtHsChr21 and TcHsChr21 and (D-F) TcMmChr16 and TcHsChr21. (A) Gene expression at all genes on WtHsChr21 and TcHsChr21, with panel (B) demonstrating the high correlation even of low-intensity genes (outlined in red in (A) and zoomed in both (B) and Fig 4C) where noise is historically a larger problem. (C) Rank ordering of the absolute expression of genes found on HsChr21 for both wild-type human and Tc1 mouse hepatocytes. (D-F) similar analysis comparing TcMmChr16 and TcHsChr21 show almost complete loss of correlation and rank order.

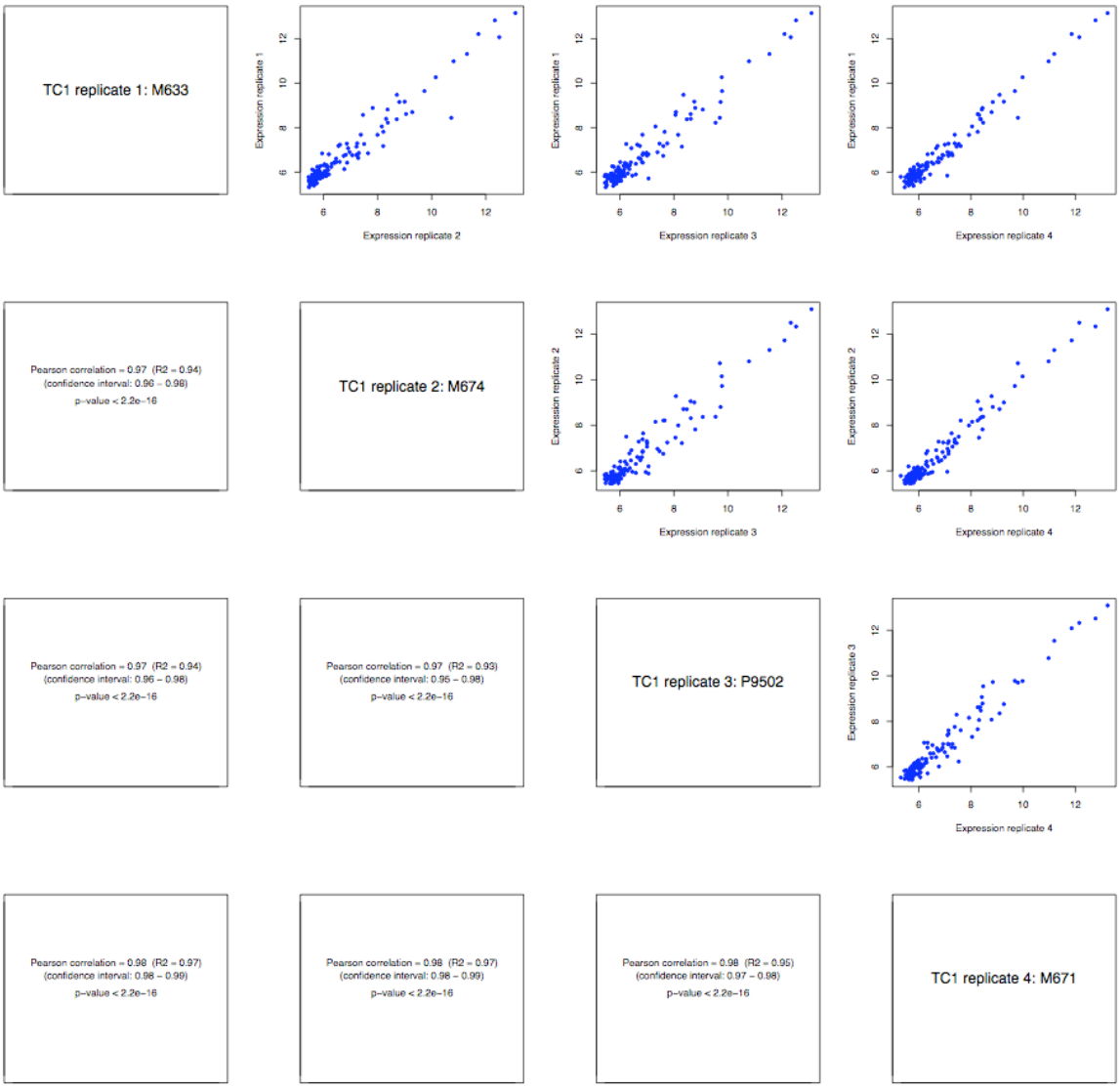


Figure S12. Gene expression correlations among replicates. *Panel A*: TcMmChr16 gene expression in the Tc1 mouse liver using mouse bead arrays. The gene expression replicates (diagonal) cross-plotted (upper right panels) and the correlation associated p-values for each pair (lower left panels).

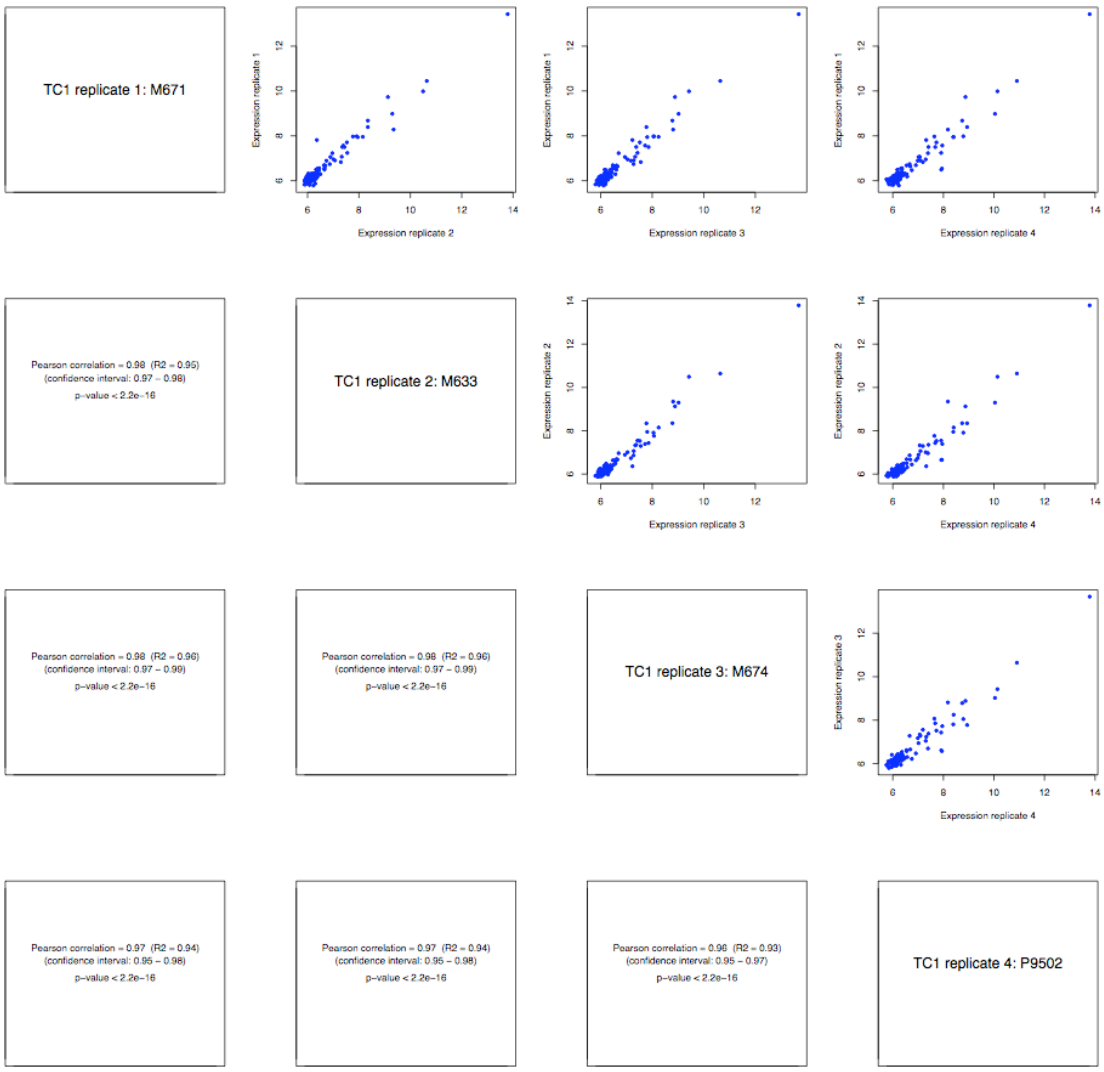


Figure S12. Gene expression correlations among replicates. *Panel B*: TcHsChr21 gene expression in Tc1 mouse liver on human bead arrays. The gene expression replicates (diagonal) cross-plotted (upper right panels) and the correlation associated p-values for each pair (lower left panels).

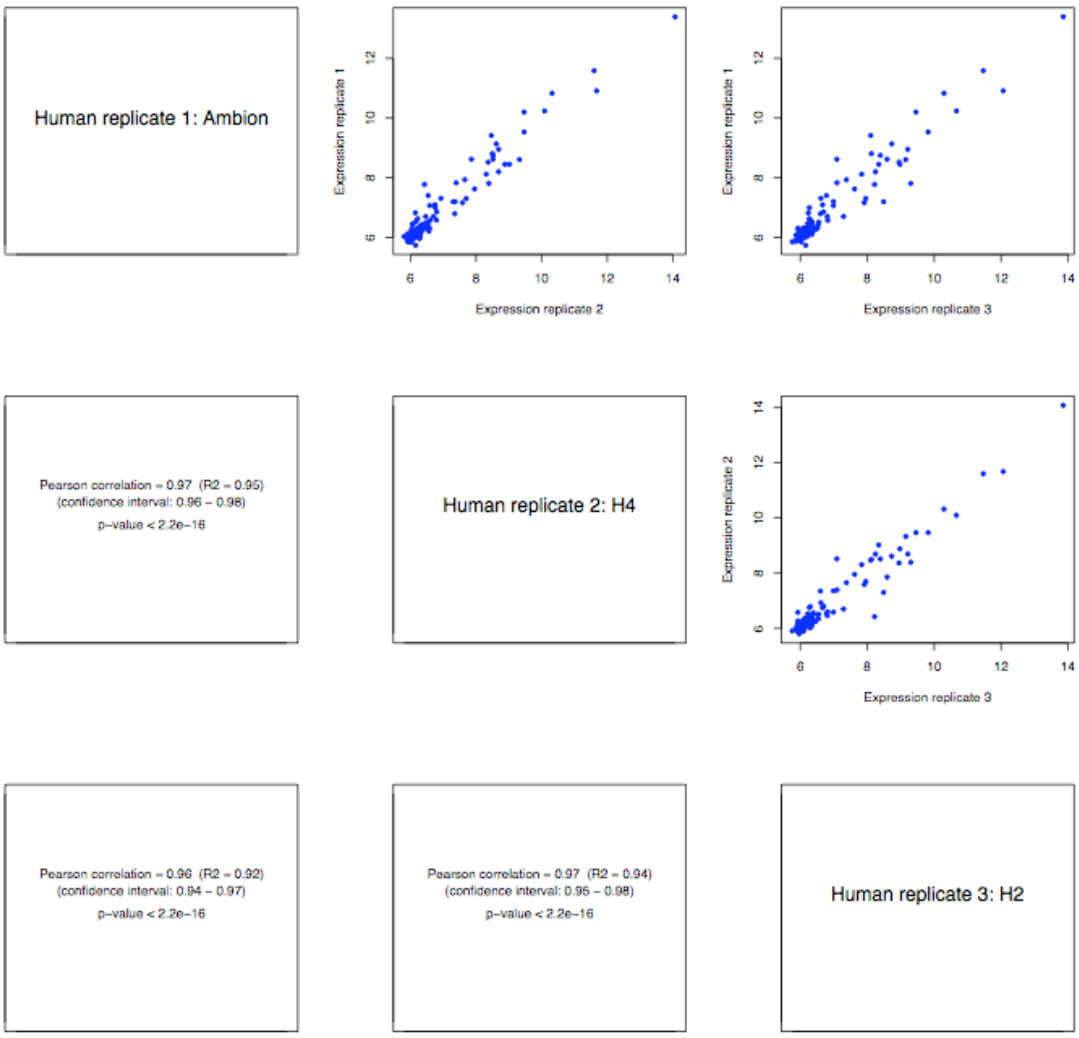


Figure S12. Gene expression correlations among replicates. *Panel C*: WtHsChr21 expression in human liver on human bead arrays. The gene expression replicates (diagonal) cross-plotted (upper right panels) and the correlation associated p-values for each pair (lower left panels).