nature genetics

Tissue-specific transcriptional regulation has diverged significantly between human and mouse

Duncan T Odom[1,5,6], Robin D Dowell[2,6], Elizabeth S Jacobsen[1], William Gordon[3], Timothy W Danford[2], Kenzie D MacIsaac[4], P Alexander Rolfe[2], Caitlin M Conboy[1,5], David K Gifford[1,2] & Ernest Fraenkel[2,3]

We demonstrate that the binding sites for highly conserved transcription factors vary extensively between human and mouse. We mapped the binding of four tissue-specific transcription factors (FOXA2, HNF1A, HNF4A and HNF6) to 4,000 orthologous gene pairs in hepatocytes purified from human and mouse livers. Despite the conserved function of these factors, from 41% to 89% of their binding events seem to be species specific. When the same protein binds the promoters of orthologous genes, approximately two-thirds of the binding sites do not align.

Elements of transcriptional regulation have central roles in evolution[1–3]. In many cases, conserved biological processes are controlled by evolutionarily conserved regulatory programs, and evolving phenotypes are associated with cross-species variation in transcription regulation[4]. However, in the absence of suitable genome-wide data, it is unclear what fraction of all protein-DNA interactions are under either positive or negative selective pressure[1]. A preliminary effort to compare genome-wide binding sites for two stem cell–specific transcription factors in human and mouse has suggested that large differences exist between mouse and human[5,6], but because the data were obtained using different methodologies, there remains the possibility that observed changes are the result of purely technical differences.

To compare systematically the binding of transcriptional regulators to promoter regions across species, we designed carefully matched chromatin immunoprecipitation (ChIP)-chip experiments[7] in human and mouse. We created custom DNA microarrays that array 10 kb of sequence surrounding the known transcription start sites of over 4,000 orthologous pairs of mouse and human genes. We selected these genes because we were able to unambiguously assign their orthology and design oligonucleotides to represent the putative regulatory regions at high density (**Fig. 1a** and **Supplementary Methods** online). We included 47 hand-curated, tissue-specific genes in the array design as controls.

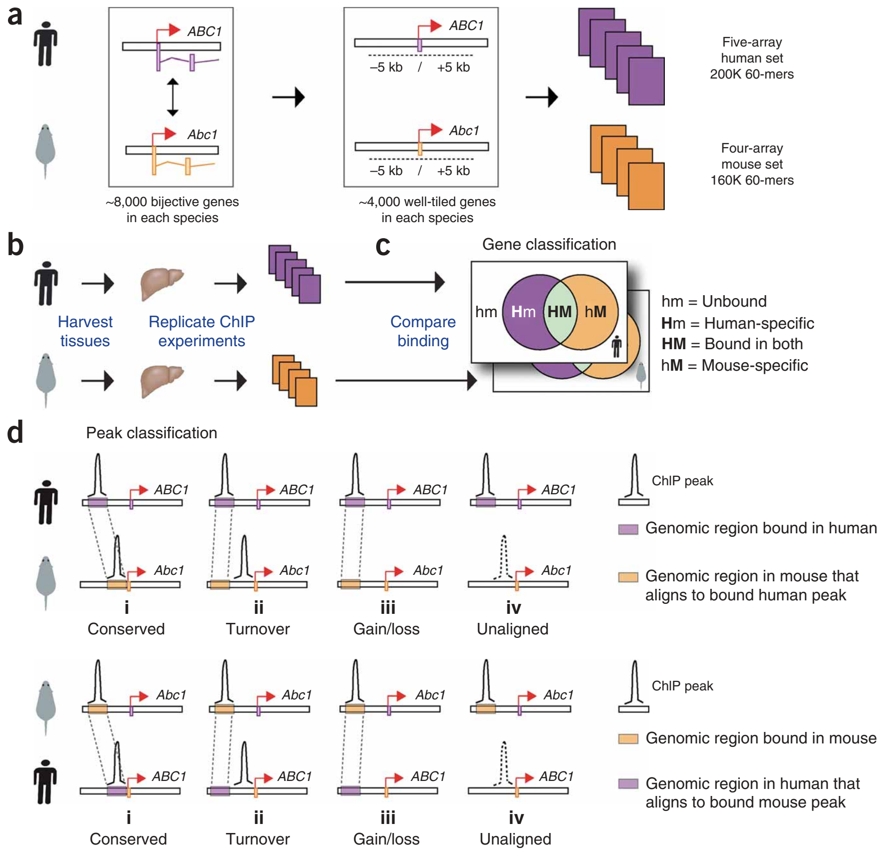
We performed ChIPs independently in primary hepatocytes directly isolated from mouse and human liver using antibodies against four tissue-specific transcription factors (FOXA2, HNF1A, HNF4A and HNF6) involved in liver development and regulation (**Fig. 1b** and **Supplementary Table 1** online)[7]. Hepatocytes were chosen as a representative tissue for these experiments because (i) they are functionally and structurally conserved among mammals[8], (ii) their gene expression programs are similar across species (**Supplementary Table 1**), (iii) their gene expression patterns are largely unperturbed by isolation procedures[9] and (iv) the transcription factors responsible for hepatocyte development and function are highly conserved[8]. We amplified and fluorescently labeled the DNA from these binding experiments, hybridized it to the microarrays and then scored binding events[10]. MIAME-compliant data have been deposited in ArrayExpress under accession code E-TABM-108. The author's supporting website (http://fraenkel.mit.edu/TxEvol) contains analysis files (binary binding call files for all genes using two error models with two binding thresholds for each method and motif discovery input files) and a downloadable PDF file with graphs of binding data for all curated tissue-specific genes.

Several possible outcomes can be distinguished when comparing a binding event in one species with data from a second species (**Fig. 1**). First, one can determine if a particular transcription factor binds anywhere within the arrayed region of the human and/or mouse ortholog (a 'gene-centric' approach) (**Fig. 1c**). Second, one can determine if the positions of individual binding events are maintained, to the resolution limits of the ChIP assay (a 'peak-centric' approach). As DNA sequences may have undergone rearrangements between human and mouse, we considered whether a binding event detected in one species occurred at the corresponding aligned region in the second species, resulting in four possible outcomes (**Fig. 1d**).

We were surprised to find that 41%–89% of the orthologous promoters bound by a protein in one species were not bound by the same protein in the second species, depending on the transcription factor (**Fig. 2a** and **Supplementary Fig. 1** and **Supplementary Table 2** online). In some of these cases, a transcription factor may continue to regulate both orthologs through binding sites that lie beyond the >10 kb of promoter sequence represented on our arrays. The sets of gene pairs with promoters that are bound in both species by each factor (category 'HM' in **Fig. 1c**) were significantly enriched for an independently determined set of liver-specific genes (**Supplementary Fig. 1**), consistent with known functional conservation of the transcription factors we profiled. The extent of species-specific binding was much greater than would be expected based on our

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Computer Science and Artificial Intelligence Laboratory, [3]Department of Biological Engineering and [4]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [5]Present address: Cancer Research UK–Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. [6]These authors contributed equally to this work. Correspondence should be addressed to E.F. (fraenkel-admin@mit.edu).

Received 17 November 2006; accepted 27 April 2007; published online 21 May 2007; doi:10.1038/ng2047



a
ABC1
Abc1
~8,000 bijective genes in each species
ABC1
–5 kb / +5 kb
Abc1
–5 kb / +5 kb
~4,000 well-tiled genes in each species
Five-array human set 200K 60-mers
Four-array mouse set 160K 60-mers
b
Harvest tissues
Replicate ChIP experiments
c
Compare binding
Gene classification
hm
Hm
HM
hM
hm = Unbound
Hm = Human-specific
HM = Bound in both
hM = Mouse-specific
d
Peak classification
ABC1
Abc1
i
Conserved
ii
Turnover
iii
Gain/loss
iv
Unaligned
ChIP peak
Genomic region bound in human
Genomic region in mouse that aligns to bound human peak
Abc1
ABC1
i
Conserved
ii
Turnover
iii
Gain/loss
iv
Unaligned
ChIP peak
Genomic region bound in mouse
Genomic region in human that aligns to bound mouse peak

Figure 1 Strategy to analyze interactions between transcription factors and DNA in mouse and human. (**a**) Left: approximately 8,000 high-confidence human (purple) and mouse (orange) gene orthologs were identified. Center: 60-mer oligonucleotides were designed against a region 5 kb upstream and 5 kb downstream of the complete set of transcription start sites in both species (colored boxes on genome track); orthologous genes with incomplete coverage, low oligonucleotide quality or substantial gaps in one or both species were removed from the final design (**Supplementary Methods**). Right: a human five-array set and a mouse four-array set capturing the transcription start sites for approximately 4,000 genes in each species were created using these oligonucleotides. (**b**) Mouse and human hepatocytes were isolated from liver samples and used for ChIPs, which were hybridized against the array sets. (**c**) Gene-centric analysis classifies orthologous gene pairs by whether they are bound in neither species (hm), bound uniquely in human (Hm), bound in both species (HM) or bound uniquely in mouse (hM). (**d**) Peak-centric analysis classifies peaks relative to whether corresponding aligned regions exist in the second species and whether these aligned regions are bound. The four possible outcomes are shown in both the human-to-mouse and the mouse-to-human panels: in the first three cases (i, ii, iii), the aligned locus is present in the arrayed region of the ortholog. In case (i) (conserved), the aligned regions are bound in both species; in case (ii) (turnover), the orthologous gene is bound, but not at the aligned locus; in case (iii) (gain/loss), no binding is detected in the arrayed region of the second species, including the aligned sequence, and in case (iv) (unaligned), the aligned sequence is not present within the arrayed region, so the binding event cannot be classified definitively, regardless of the presence or absence of a binding event in the other species.

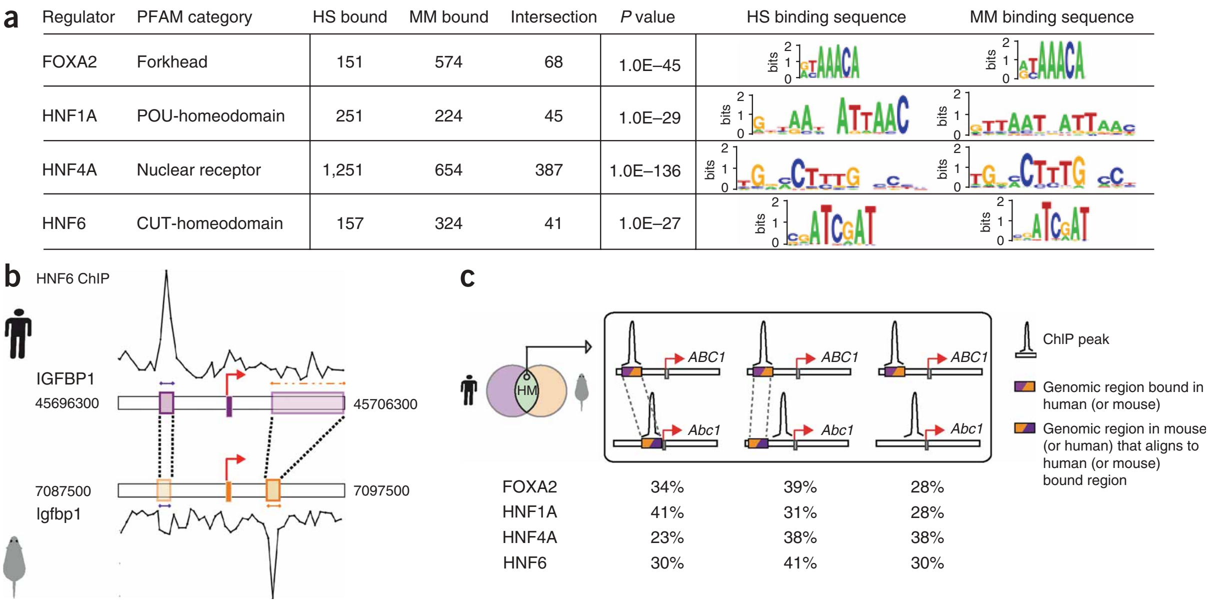
experimentally determined error rates and did not depend on the computational technique used to identify the bound region (**Supplementary Fig. 2** and **Supplementary Table 2** online).

To estimate the maximum variation in binding that could be attributed to environmental and intraspecies genetic sources (as opposed to interspecies evolutionary sources), we compared HNF6 genomic occupancy in primary human hepatocytes to corresponding HNF6 occupancy in the human cell line HepG2 (**Supplementary Table 2**). Despite the fact that HepG2 cells are an immortalized hepatocellular carcinoma that is severely aneuploid and has been propagated in culture for over two decades[11], we found that 66% of the genes bound in primary human liver were bound in HepG2. In contrast, only 26% of orthologous gene pairs bound by HNF6 in human hepatocytes are also bound in orthologous regions in mouse hepatocytes.

Using the THEME algorithm[12], we determined that the observed changes in binding patterns across species did not arise from changes in the DNA-binding specificity of the transcription factors and that transcription factor binding in each species was highly correlated with the presence of sequences matching the protein's motif (**Supplementary Table 3** and **Supplementary Fig. 3** online). To determine whether binding differences between orthologs arise from sequence differences at potential binding sites, we scanned previously reported mouse-human genome alignments[13] for conserved motif sequences. As expected, the frequency of conserved motif sequences near binding peaks was highest for conserved peaks (case (i) in **Fig. 1d**); the frequency of conserved motif sequences was lower near binding events that are unique to one species but was still above background (case (ii), turnover, and case (iii), gain/loss) (**Supplementary Table 3**). The conserved sequences that are not bound in our assay may be functional in both species under particular conditions, during alternative developmental stages or in tissues not analyzed in our study.

Most crucially, the location of binding events varies widely between species in ways that cannot be predicted from human-mouse sequence alignments alone. For instance, the binding site for HNF6 at IGFBP1 shifted over 4 kb from the promoter region in humans to the first intron in mice (**Fig. 2b**). More broadly, in the 41 orthologous pairs of promoters that were bound by HNF1A in both species, there were 47 binding events in humans and 51 binding events in mice. Of these, only 20 occurred in sequences that were aligned to each other. The fraction of aligned binding events was even lower for other factors (**Fig. 2c** and **Supplementary Table 3**).

Our findings have implications for the use of the mouse as a model organism. For example, HNF1A bound strongly to SEL1L in human liver, yet this binding was entirely absent from the corresponding mouse region (**Supplementary Fig. 1**). Polymorphisms around the *SEL1L* locus seem to influence the onset of disease in individuals with maturity-onset diabetes of the young type III, which is caused by haploinsufficiency of HNF1A[14]. The lack of HNF1A binding in the mouse suggests that this susceptibility may be species specific. In contrast to the variation in cross-species binding sites, the location of binding events within a species is robust to substantial environmental and genetic perturbations. Of genes bound in both human hepatocytes and in the human carcinoma cell line HepG2, over 95% had peaks within 100 nucleotides of each other (**Supplementary Table 2**).



Regulator	PFAM category	HS bound	MM bound	Intersection	*P* value	HS binding sequence	MM binding sequence
FOXA2	Forkhead	151	574	68	1.0E–45		
HNF1A	POU-homeodomain	251	224	45	1.0E–29		
HNF4A	Nuclear receptor	1,251	654	387	1.0E–136		
HNF6	CUT-homeodomain	157	324	41	1.0E–27		

FOXA2	34%	39%	28%
HNF1A	41%	31%	28%
HNF4A	23%	38%	38%
HNF6	30%	41%	30%

Figure 2 Comparison of binding in human and mouse. (**a**) Most binding events within 5 kb of a transcription start site are species specific (the gene-centric approach). Shown are the number of genes bound by liver master regulators in each species, the *P* value (using a hypergeometric distribution) that the cross-species overlap is due to random chance and the THEME-derived binding motifs in human and mouse. (**b**) The location of binding events varies between species. Here, ChIP enrichments are shown as traces. The 500-bp sequence underlying the ChIP peak in each species (purple, human; orange, mouse) is aligned with the corresponding sequence in the second species using dashed lines. For clarity, mouse ChIP enrichments are displayed as a negative *y*-axis, but orientation of the transcription start site is from left to right. IGFPB1 is bound by HNF6 in both species, but the binding events do not align. The human sequence aligned with the mouse HNF6 peak in IGFBP1 contains large insertions overlapping a substantial portion of the human first intron (outlined with a dashed orange box) and is not bound by HNF6. (**c**) Shared binding events are frequently found in non-aligned regions (the peak-centric approach). From left to right: aligned regions (colored boxes) that are bound in both species (**Fig. 1d**, case (i)); aligned regions present on both human and mouse arrays but bound only in one species (**Fig. 1d**, case (ii)); regions bound in both species, but lacking aligned sequences on the orthologous array (**Fig. 1d**, case (iv), with a binding peak present). Typically, only about one-third of the binding events detected in both species occur in sequences that align with each other (see also **Supplementary Table 3**).

The *in vivo* binding of four distinct tissue-specific transcription factors (FOXA2, HNF1A, HNF4A and HNF6) responsible for liver gene expression has diverged substantially between human and mouse. The most notable feature of this divergence is the high mobility of transcription factor binding sites. Analysis of genomic regions that are bound by the same factors in both species shows that approximately two-thirds of the binding events are not aligned between the mouse and human genomes. The cross-species variation cannot be explained by changes in the sequence specificity of the transcription factors, nor can it be predicted based solely on the conservation of binding sequences in the two species. Other effects, including the concentration of these transcription factors, other interacting proteins and chromatin modifications, are likely to contribute to the observed variations[15] (**Supplementary Note** online). Differences between human and mouse physiology and behavior may also contribute to the observed binding changes, and these physiological and behavioral differences will affect all studies that use the mouse as a model for human biology. The marked plasticity of transcription factor binding indicates that accurate mapping of functional genomic elements responsible for gene expression will require direct measurements of transcription factor occupancy in multiple species.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS
We are grateful to S. Strom and K. Dorko (University of Pittsburgh) for human liver samples (DK92310); T. DiCesare, S. Gupta, R. Kumar, J. Rodriguez and K. Walker for technical assistance and R. Young and G. Gerber for discussions. This work was supported by funding from US National Institutes of Health (grants DK68655 and DK70813 (D.T.O.) and DK076284 (R.D.D.)), Cancer Research UK (D.T.O.) and the Whitaker Foundation (E.F.).

AUTHOR CONTRIBUTIONS

D.T.O., R.D.D. and E.F. designed experiments; R.D.D. designed the arrays; D.T.O., E.S.J., W.G. and C.M.C. performed experiments; R.D.D., T.W.D, K.D.M., P.A.R. and E.F. analyzed the data and D.T.O, R.D.D., D.K.G. and E.F. created the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Bird, C.P., Stranger, B.E. & Dermitzakis, E.T. *Curr. Opin. Genet. Dev.* **16**, 559–564 (2006).
2. Moses, A.M. *et al. PLoS Comput. Biol.* **2**, e130 (2006).
3. Prabhakar, S., Noonan, J.P., Paabo, S. & Rubin, E.M. *Science* **314**, 786 (2006).
4. King, M.C. & Wilson, A.C. *Science* **188**, 107–116 (1975).
5. Boyer, L.A. *et al. Cell* **122**, 947–956 (2005).
6. Loh, Y.H. *et al. Nat. Genet.* **38**, 431–440 (2006).
7. Odom, D.T. *et al. Mol. Syst. Biol.* **2**, 2006.0017 (2006).
8. Zaret, K.S. *Nat. Rev. Genet.* **3**, 499–512 (2002).
9. Richert, L. *et al. Drug Metab. Dispos.* **34**, 870–879 (2006).
10. Qi, Y. *et al. Nat. Biotechnol.* **24**, 963–970 (2006).
11. Natarajan, A.T. & Darroudi, F. *Mutagenesis* **6**, 399–403 (1991).
12. Macisaac, K.D. *et al. Bioinformatics* **22**, 423–429 (2006).
13. Schwartz, S. *et al. Genome Res.* **13**, 103–107 (2003).
14. Kim, S.H. *et al. Diabetes* **53**, 1375–1384 (2004).
15. Guccione, E. *et al. Nat. Cell Biol.* **8**, 764–770 (2006).