The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 6, 2009):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
http://www.sciencemag.org/cgi/content/full/322/5900/434

Supporting Online Material can be found at:
http://www.sciencemag.org/cgi/content/full/1160930/DC1

A list of selected additional articles on the Science Web sites related to this article can be found at:
http://www.sciencemag.org/cgi/content/full/322/5900/434#related-content
This article cites 29 articles, 10 of which can be accessed for free:
http://www.sciencemag.org/cgi/content/full/322/5900/434#otherarticles

This article has been cited by 3 article(s) on the ISI Web of Science.

This article has been cited by 3 articles hosted by HighWire Press; see:
http://www.sciencemag.org/cgi/content/full/322/5900/434#otherarticles

This article appears in the following subject collections:
Genetics
http://www.sciencemag.org/cgi/collection/genetics

Information about obtaining reprints of this article or about obtaining permission to reproduce this article in whole or in part can be found at:
http://www.sciencemag.org/about/permissions.dtl
Reports


7. See supporting online material text.


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 NNG04GC86G 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 de 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 were 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), Ministerio de Ciencia e Tecnología (Brazili, and Secretaria de Ciencia y Tecnologia (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,1,2* Nuno L. Barbosa-Morais,1,2,6 Dominic Schmidt,1,2 Caitlin M. Conboy,3 Lesley Vannes,4 Victor L. J. Tybulewicz,4 Elizabeth M. C. Fisher,5 Simon Tavare,1,2,6 Duncan T. Odom1,2

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 (J). 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 (J–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 hepaticocyte 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-
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 (Fig. 1). 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 TcMmChr16 (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–encoded 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-
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-conformed 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).

Adding further evidence to the human origin of these regions. Although the use of 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 ≈ 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 = 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).
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
11. B. E. Bernstein et al., Cell 120, 149 (2005).
15. Materials and methods are available as supporting material on Science Online.

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†

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