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
SOM Text
Figs. S1 to S12
Table S1
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).

<table>
<thead>
<tr>
<th>Species:</th>
<th>Human</th>
<th>Wild-type mouse</th>
<th>Tc1 mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Array:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HNF1α</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>HNF4α</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>HNF6</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

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 immuno precipitated 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 ONLINE MATERIAL**

**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.
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.
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.
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 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 CAGTGCCTGGACTTAGGAAA and 1-mm-rev GGCATTGCTCAAGACAGAAA; and human primers used: 1-hs-fw GGAATACGCCTGCTAGAT 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).

<table>
<thead>
<tr>
<th></th>
<th>Amino acid differences</th>
<th>Aligned amino acids</th>
<th>Percent amino acid identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>HNF1α</td>
<td>30</td>
<td>628</td>
<td>95</td>
</tr>
<tr>
<td>HNF4α</td>
<td>22</td>
<td>464</td>
<td>95</td>
</tr>
<tr>
<td>HNF6</td>
<td>5</td>
<td>503</td>
<td>99</td>
</tr>
</tbody>
</table>

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.
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.
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 (wHsChr21) 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 neuregulin-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.
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).

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 x 10^{-25}), whereas differences between the WtHsChr21/WtMmChr16 and TcHsChr21/TcMmChr16 comparisons are considerably closer to unity (p-value >>1 x 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.

**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).