Comparison of the transcriptional landscapes between human and mouse tissues

Shin Lin[a,b,1], Yiing Lin[c,1], Joseph R. Nery[d], Mark A. Urich[d], Alessandra Breschi[e,f], Carrie A. Davis[g], Alexander Dobin[g], Christopher Zaleski[g], Michael A. Beer[h], William C. Chapman[c], Thomas R. Gingeras[g,i], Joseph R. Ecker[d,j,2], and Michael P. Snyder[a,2]

[a]Department of Genetics, Stanford University, Stanford, CA 94305; [b]Division of Cardiovascular Medicine, Stanford University, Stanford, CA 94305; [c]Department of Surgery, Washington University School of Medicine, St. Louis, MO 63110; [d]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037; [e]Centre for Genomic Regulation and UPF, Catalonia, 08003 Barcelona, Spain; [f]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Spain; [g]Functional Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11742; [h]McKusick-Nathans Institute of Genetic Medicine and the Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205; [i]Affymetrix, Inc., Santa Clara, CA 95051; and [j]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037

Contributed by Joseph R. Ecker, July 23, 2014 (sent for review May 23, 2014)

Although the similarities between humans and mice are typically highlighted, morphologically and genetically, there are many differences. To better understand these two species on a molecular level, we performed a comparison of the expression profiles of 15 tissues by deep RNA sequencing and examined the similarities and differences in the transcriptome for both protein-coding and -noncoding transcripts. Although commonalities are evident in the expression of tissue-specific genes between the two species, the expression for many sets of genes was found to be more similar in different tissues within the same species than between species. These findings were further corroborated by associated epigenetic histone mark analyses. We also find that many noncoding transcripts are expressed at a low level and are not detectable at appreciable levels across individuals. Moreover, the majority lack obvious sequence homologs between species, even when we restrict our attention to those which are most highly reproducible across biological replicates. Overall, our results indicate that there is considerable RNA expression diversity between humans and mice, well beyond what was described previously, likely reflecting the fundamental physiological differences between these two organisms.

transcriptome | epigenome | species comparison | noncoding transcripts

The mouse has served as a valuable model organism for human biology and disease. It is widely assumed that biochemical, cellular, and developmental pathways in the mouse are highly conserved with humans and that many processes are clearly preserved at a molecular and genetic level. Moreover, recent detailed studies have examined gene expression in a limited number of tissues in humans and mice. These studies have indicated that gene expression is often conserved and is more similar between the comparable tissues of different organisms rather than within tissues of the same organism. In contrast, the transcript isoform repertoire was found to be markedly different between species (1, 2).

Gene Expression Is More Similar Among Tissues Within a Species Than Between Corresponding Tissues of the Two Species

To examine the similarities between humans and mice in much greater detail, we produced RNA-seq data from 13 human tissues [as part of the Encyclopedia Of DNA Elements (ENCODE)], another 11 human tissues [as part of the Roadmap Epigenomics Mapping Consortium (REMC) (3)], and 13 mouse tissues (for mouse ENCODE). We also included in our analysis other data from mouse ENCODE and the Illumina Human BodyMap 2.0 (HBM) (*SI Materials and Methods*). Sequencing was performed to a depth of 11,313,824–166,188,101 mappable reads (median of 68,399,538 with and an interquartile range of 31,557,381–81,836,199). In total, our analysis used 93 datasets encompassing the most tissue-diverse RNA-seq dataset to date spanning several major projects. Thirteen of the mouse and human orthologous datasets were produced by the same laboratory. For our analysis regarding noncoding transcripts, we incorporated an additional 294 RNA-seq datasets from the Genotype-Tissue Expression (GTEx) project (4).

We first explored gene expression similarities and differences by analyzing the expression of ∼15,106 protein-coding orthologs; this list was generated by the modENCODE and mouse ENCODE consortia and represents the most recent mouse–human ortholog list to date (biorxiv.org/content/biorxiv/early/2014/05/31/005736.full.pdf). Fragments per kilobase of transcript per million (FPKM) values were obtained from each dataset, and principal component analysis (PCA) was used to compare gene expression (*Materials and Methods*). In contrast to what was reported previously (1, 2, 5), surprisingly, we found that the mouse and human samples cluster by species when the data are projected onto the first three principal components (Fig. 1*A*). Because the same tissues of the same species produced by different laboratories did not cluster together, the possibility of methodologic differences among laboratories confounding our results was considered. To address this issue, analysis of only the 13 paired samples processed under one experimental protocol yielded the same species-specific clustering (Fig. 1*C*). The same species-specific clustering was observed when other combinations of 10 or more tissues were examined, indicating that the clustering is not due to the particular 13–15 tissues selected. Finally,

Significance

To date, various studies have found similarities between humans and mice on a molecular level, and indeed, the murine model serves as an important experimental system for biomedical science. In this study of a broad number of tissues between humans and mice, high-throughput sequencing assays on the transcriptome and epigenome reveal that, in general, differences dominate similarities between the two species. These findings provide the basis for understanding the differences in phenotypes and responses to conditions in humans and mice.

Author contributions: S.L., Y.L., W.C.C., T.R.G., J.R.E., and M.P.S. designed research; S.L., Y.L., J.R.N., M.A.U., and A.D. performed research; S.L., Y.L., and W.C.C. contributed new reagents/analytic tools; S.L., Y.L., A.B., C.A.D., A.D., C.Z., and M.A.B. analyzed data; and S.L., Y.L., J.R.E., and M.P.S. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE36025). The Roadmap Epigenomics Mapping Consortium RNA-seq data can be accessed under GSE16256.

[1]S.L. and Y.L. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: ecker@salk.edu or mpsnyder@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1413624111/-/DCSupplemental.

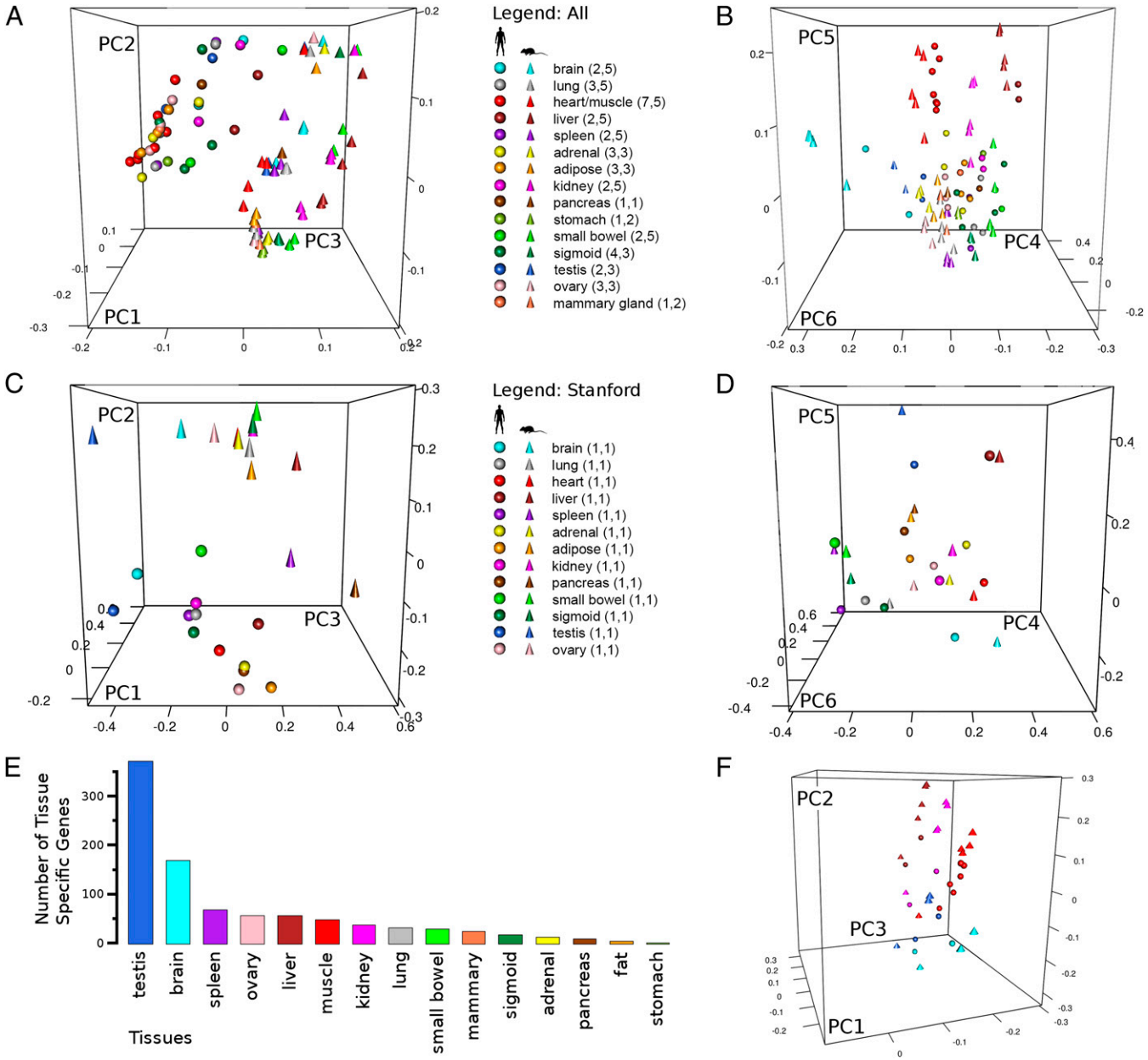

Fig. 1. Loading plots from PCA on human and mouse gene expression data. (*A*) PCA is performed on the combined Stanford (human, mouse), Salk (human), HBM (human), LICR (mouse), and CSHL (mouse) expression datasets using 15 tissue types, 15,106 orthologs (biorxiv.org/content/biorxiv/early/2014/05/31/005736.full.pdf), and Pearson's correlation as the distance measure. The loadings on principal components 1–3 are plotted. (*B*) Same as in *A* except loadings on principal components 4–6 are plotted. (*C*) The loadings on principal components 1–3 are plotted from a PCA performed as in *A* except only 13 human and mouse tissue sets processed at Stanford. (*D*) The loadings on principal components 4–6 for the analysis in *C* are used. (*E*) Barplot of number of tissue specific-genes per tissue. (*F*) PCA is performed as in *A* except the tissue set is restricted to testis, brain, heart, liver, and kidney, which have higher numbers of tissue-specific genes. The loadings on principal components 1–3 are plotted.

different normalization methods (e.g., quantile normalization) applied to the data produced similar groupings.

To understand the differences between our results and those of others (1, 2, 5), we performed extensive additional analyses. We first varied the ortholog list and similarity measure, but these changes did not significantly alter our results (Fig. S1*A*). We next applied our analytic process to human and mouse data produced from other studies that reported tissue dominated clustering (1, 5), and we were able to reproduce their findings (Fig. S1 *B* and *C*). Moreover, in our own dataset, we observed a more tissue-dominated clustering for principal components 4–6 (Fig. 1 *B* and *D*). Thus, gene expression profiles of different organism tissues do exhibit similarities in gene expression but of lower strength relative to organismal signals.

To determine whether we could further reconcile these various observations, we identified the groups of genes that are tissue specific and those present in all tissues (i.e., housekeeping) using Shannon entropy (H) (6). H is a parameter commonly used to assess tissue specificity, with lower values signifying expression in a smaller fraction of the total set. We calculated H for each gene using our expression data and considered genes with values below two to be tissue specific. We found that testes, brain, liver, muscle (cardiac and/or skeletal), and kidney were among the tissues that expressed the most tissue-specific genes (Fig. 1*E*),

and interestingly, tissue-specific genes are generally more highly expressed than housekeeping ones (Fig. S2). Other studies restricted their analyses to the same tissues expressing high numbers of tissue-specific genes, thereby likely explaining their finding that gene expression clustering is dominated by tissues. Indeed, when we limit our analysis to the same tissues (testes, brain, liver, muscle, and kidney), we obtain tissue-specific clustering in the first three principal components (Fig. 1*F*). Overall, our results indicate that for the human–mouse comparison, tissues appear more similar to one another within the same species than to the comparable organs of other species when examining a more complete set of tissue types. When other clustering algorithms are used, species-specific clustering is still observed with the exception of one to three tissues.

A Subset of Housekeeping Genes Drives Species-Specific Expression

To better understand the genes driving the species-dominated clustering, we identified those most different in the two species by using the nonparametric Mann–Whitney test on expression levels between human and mouse across all tissues for our set of orthologous genes. A total of 4,767 genes are statistically significant at a false discovery rate (FDR) of 0.0005; 2,569 of these are expressed at higher levels in the human relative to the mouse, and 2,198 are expressed at lower levels. As expected, removal of genes that are differentially expressed between human and mouse changes the clustering in the PCA from a pattern that is species oriented to one that is more dominated by tissues (Fig. S1*D*).

To determine the types of genes that drive the species-specific expression across different tissues, genes were examined for enrichments in Gene Ontology (GO) biological processes (7). Among those associated with the most proteins are cellular nitrogen compound metabolic process, biosynthetic process, signal transduction, anatomical structure development, and transport (Tables S1 and S2). Thus, the genes differentially expressed between human and mouse generally participate in basic cellular functions, and indeed, the H of these genes is weakly higher than background (median H of differentially expressed genes is 3.40 vs. background of 3.20, $P < 2.2 \times 10^{-16}$ by the Mann–Whitney test), signifying they are composed of more housekeeping genes.

Histone Mark Differences for Differentially Expressed Genes Between Humans and Mice

To further explore and extend the observation that gene expression is more similar between tissues in the same organism than in the same tissues across organisms, we used an independent assay in which chromatin marks were analyzed. Recent studies with human cell lines studied by ENCODE (8) have reported detailed relationships between histone modification and expression levels (9). For a limited set of tissues (heart, lung, small bowel, spleen, liver, and brain), data on the same histone modifications exist from the human REMC and mouse ENCODE projects. For the 4,767 differentially expressed genes between humans and mice, we examined the histone marks H3K4me3, H3K4me1, H3K27me3, H3K9me3, H3K9ac, and H3K27ac modification levels using available ChIP-seq data in the 1-kb flanking regions of

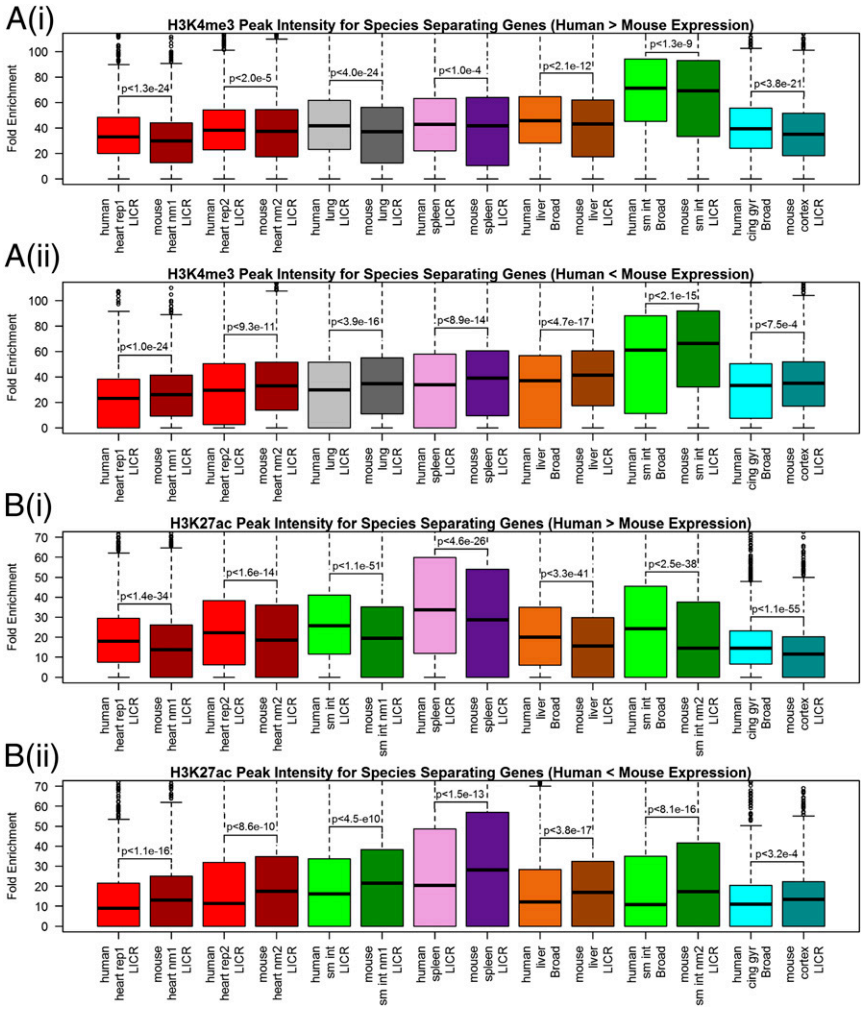

Fig. 2. Histone peak intensities for species-separating genes. Fold enrichment over control of (*A*) H3K4me3 and (*B*) H3K27ac present at promoters of 4,767 of the most differentially expressed genes between humans and mice are visualized with boxplots for various tissues. Histone intensities are quantile normalized in human and mouse pairs for each tissue on all orthologs beforehand. The rep1 and rep2 designation represents biological replicates. The nm1 and nm2 (i.e., normalization 1 and 2) boxplots originate from the same sample but are scaled differently, because they are quantile normalized to different samples. Sm int and cing gyr are abbreviations for small intestines and cingulate gyrus, respectively. (*i*) Boxplots show intensities for the 2,569 genes in which human expression is higher than in mouse. (*ii*) Boxplots show intensities for the 2,198 genes in which human expression is lower than in mouse. *P* values are generated by the nonparametric paired Wilcoxon test between the human and mouse ChIP intensity values.

transcription start sites (Fig. 2). We found that the signals for active promoter marks (10) H3K4me3 and H3K27ac correspond to gene expression levels: i.e., the 2,569 genes in which human genes are more highly expressed than mouse have higher H3K4me3 and H3K27ac mark levels in human tissues compared with mouse and vice versa for the 2,198 remaining genes. These patterns were consistent across all other REMC available tissue data which can be compared with corresponding mouse ENCODE data regardless of laboratory of origin [either Ludwig Institute for Cancer Research (LICR) or the Broad (11)]. Because histone ChIP-seq represents a technically orthogonal assay to RNA-seq, the H3K4me3 and H3K27ac patterns support the observation that differential expression of a large number of genes between human and mouse is biological and not an experimental artifact. Moreover, because the gene expression findings were produced across five laboratories [Stanford, Salk, Illumina, LICR, and Cold Spring Harbor Laboratory (CSHL)] and are concordant with features of the H3K4me3 and H3K27ac data

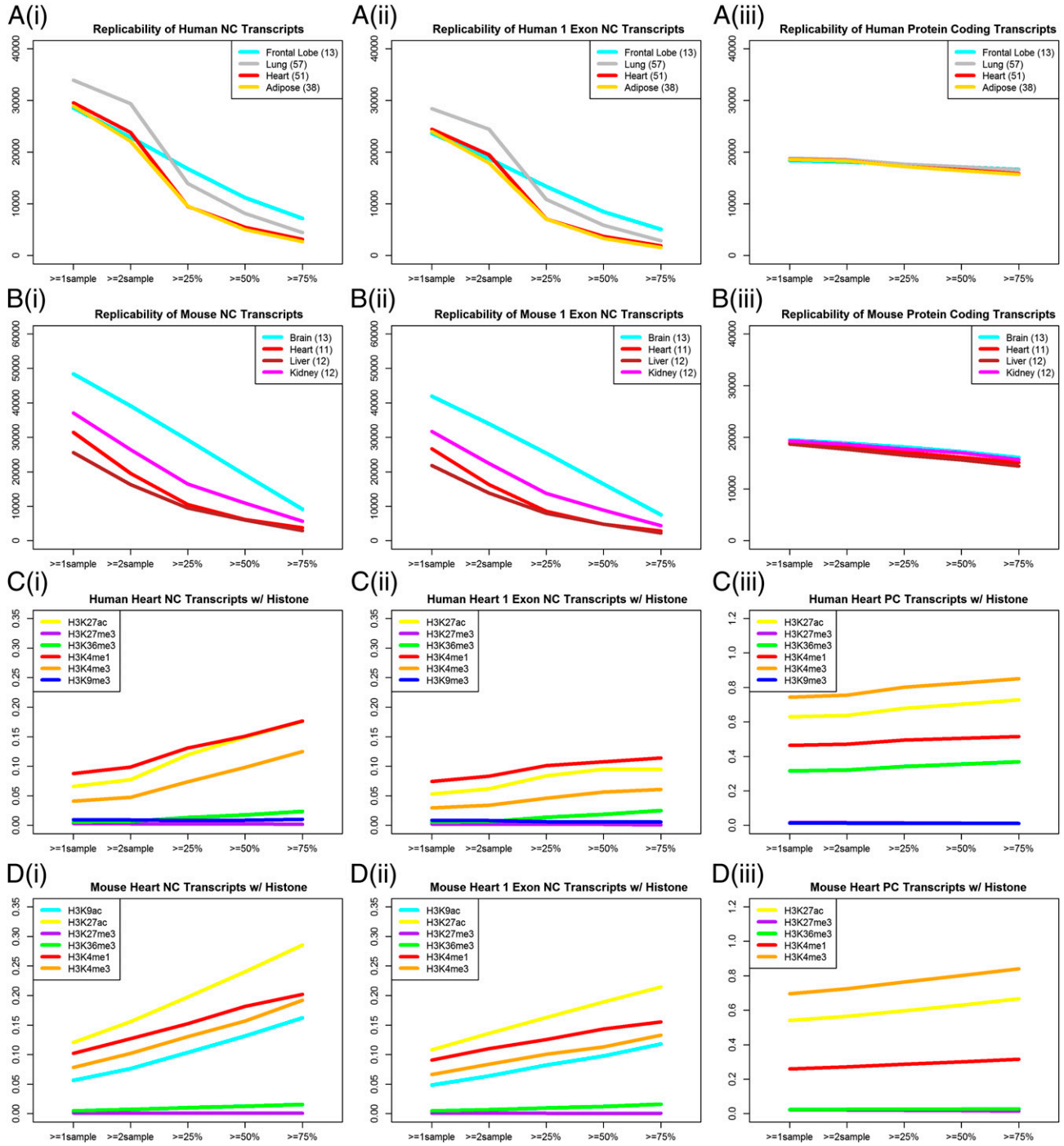

Fig. 3. Features of intergenic noncoding transcripts. The number of noncoding (NC) (*i*), noncoding single exon (NC 1 exon) (*ii*), and protein-coding (PC) (*iii*) transcripts expressed in various tissues is shown for (*A*) humans and (*B*) mice at varying levels of replicability. The number of biological replicates for each tissue is listed in parentheses in the legend. FPKM greater than zero is considered expressed. For the same transcript types, the fraction of transcripts with various histone marks present around the transcription start sites for (*C*) humans and (*D*) mice is plotted against different levels of replicability.

generated by two laboratories (LICR and Broad), our findings are reproducible by analyzing data generated by several groups.

In addition to the examination of protein-coding sequences, we also examined noncoding RNA conservation and expression in different tissues across multiple biological replicates of humans and mice. Because the HBM and GTEx RNA-seq data do not contain strand information, we focused our analysis on intergenic noncoding transcripts, which includes long intergenic noncoding RNAs (lincRNAs), as well as single exon noncoding transcripts. eRNAs are not expected to be represented in our collection because they are not captured in poly(A)+ fractions (12). In total, using approaches described in *Materials and Methods*, we arrived on a set of 70,390 and 116,526 noncoding transcripts for humans and mice, respectively. For humans, our number of noncoding transcripts is much higher than the 8,195 compiled by Cabili et al. (13), mainly because we include single exon transcripts. Because we analyzed a large number of biological replicates for several tissues in both humans and mice, we could assess the ability to detect reproducibly each transcript in more than one sample (defined here as replicability), which has not been performed previously. We found that as the threshold of replicability increases, we observe a marked drop-off in the number of noncoding transcripts detected (Fig. 3 *A* and *B*), a pattern that holds for the Cabili et al. (13), Gencode 14 (14), and Washietl et al. (15) sets of lincRNAs (Fig. S3). For example, of the 29,530 noncoding transcripts expressed in at least one sample of human heart, only 3,116 can be found in 75% or more of the 51 biological replicates analyzed. This trend is not observed when the same analysis is performed on protein-coding genes, which are generally quite reproducible across multiple experiments (Fig. 3 *A, iii* and *B, iii*). Of note, many thousands of single exon noncoding transcripts are reproducible over scores of biological replicates. Overall, the replicability of noncoding transcripts is positively correlated with higher expression, but only for the extreme end of transcripts that are detected by many experiments (Fig. S4); between 50% and 80% reproducibility, there is minimal relationship between replicability and median expression.

Next, we examined replicability in the context of promoter histone marks generated from a single heart sample studied in REMC. We observed that as the replicability threshold is increased, the proportion of genes with activating histone marks increases (Fig. 3 *C* and 3 *D* for heart; see Figs. S5 and S6 for other tissues). The presence of activating histone marks (e.g., H3K27ac and H3K4me1) at noncoding transcript promoters is thus also associated with higher replicability. The histone associations revealed in our work extend the H3K4me3 and H3K36me3 results described in Guttman et al. (16) in that other activating marks (e.g., H3K27ac and H3K4me1) are also positively associated with noncoding transcripts. We also observe that a majority of highly replicable, noncoding transcripts is not associated with histones. The failure to find histone marks at transcript boundaries may be due to degradation of RNA samples, leading to imputation of 5′ truncated noncoding transcripts. The same observations were also found for the lincRNAs called by Cabili et al. (13) and Gencode 14 (17) (Fig. S7). Comparison with an analogous analysis on protein-coding genes yields several other findings of note (Fig. 3 *C* and *D*; see Fig. S6 for other tissues). First, the number of heart noncoding transcripts with H3K4me1 is greater than those with H3K4me3; for protein-coding genes, the opposite is true. Second, unlike for noncoding transcripts, the proportion of protein-coding genes with activating histone marks is comparatively higher and does not change as appreciably with replicability threshold. For example, the proportion of protein-coding genes with H3K27ac in the human heart increases by a factor of only 1.16, from 62.9% to 72.9%, across different thresholds of replicability, compared with a factor of 2.7, from 6.6% to 17.6%, for noncoding transcripts.

We further extended the concept of replicability as a gauge for noncoding transcript validity by examining the tissue-specific/housekeeping composition of noncoding transcripts. Heretofore, the vast majority of noncoding transcripts were thought to be tissue-specific, with a recent estimate as high as 78% in humans (13). However, if putative noncoding transcripts are actually biological noise or very lowly expressed, they are more likely to be deemed tissue specific rather than to be replicated across multiple tissues. We prepared boxplots of Shannon entropy *H* across increasing thresholds of expression replicability for noncoding transcripts and found that the proportion of tissue-specific noncoding transcripts decreases in both humans and mice (Fig. 4 *A* and *C* for heart; see Fig. S8 for other tissues). This trend is not observed for protein-coding RNAs (Fig. 4 *B* and *D*). Determining the precise number of noncoding transcripts that are tissue-specific will require multiple biological replicates of RNA-seq runs across a wide range of tissues beyond what has been analyzed thus far.

In the aforementioned aspects of noncoding transcripts, we see characteristics change markedly as the replicability threshold is increased. One feature with a less pronounced change is the number of human and mouse noncoding transcript orthologs, which had been previously found to be low (17, 18). We used the University of California, Santa Cruz (UCSC) liftOver tool (19) to find conserved noncoding regions in the human and mouse genomes. Table 1 shows the number of orthologous noncoding transcripts over a range of replicability thresholds for human and mouse heart. At increasing thresholds of expression replicability, the proportion of orthologs increases but remains low. Even at the threshold of transcription in greater than or equal to 75% of biological replicates, of the approximate 3K noncoding transcripts in human and mouse, the proportion of orthologs is no more than 20%. Thus, noncoding transcripts represent another aspect of gene expression that is markedly different between humans and mice.

Discussion

Overall, we demonstrate that, by examining a wide range of tissues, differences in gene expression between humans and mice predominate, a feature that is corroborated in the epigenome by examining H3K4me3 and H3K27ac patterns. Thus, differences in gene transcription level between species are evident from the split of the human and mouse lineages, a time point much later than

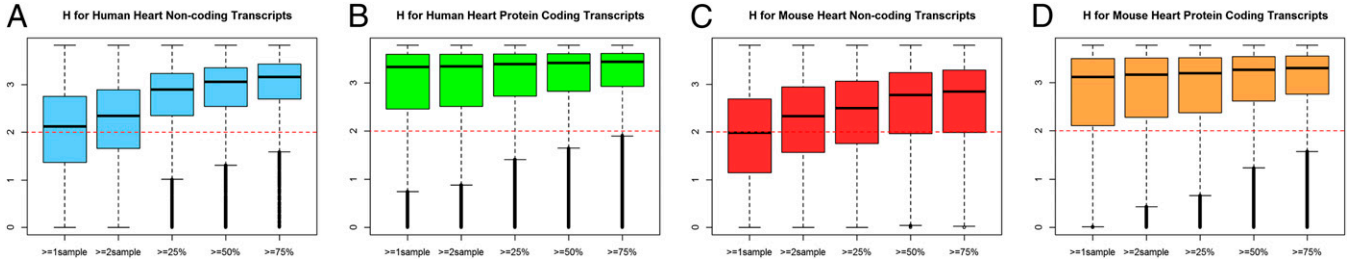

Fig. 4. Boxplots of tissue specificity of noncoding transcripts in heart. Tissues specificity, as measured by the Shannon entropy *H*, is shown in boxplots at increasing thresholds of expression replicability for (*A*) noncoding transcripts in humans, (*B*) protein coding transcripts in human, (*C*) noncoding transcripts in mice, and (*D*) protein coding transcripts in mice. The number of genes represented in each boxplot in *A–D* correspond to those plotted in Fig. 3 *A, i* and *iii* and *B, i* and *iii*, respectively.

Table 1. Number of heart noncoding transcript orthologs

Replicability	Noncoding transcript	Orthologs (%)
Human to mouse		
≥1 of 51	29,530	3,511 (11.9)
≥2 of 51	23,844	3,000 (12.6)
≥13 of 51	9,513	1,478 (15.5)
≥26 of 51	5,455	949 (17.4)
≥38 of 51	3,116	603 (19.4)
Mouse to human		
≥1 of 11	31,450	2,952 (9.4)
≥2 of 11	19,554	2,047 (10.5)
≥3 of 11	13,934	1,513 (10.9)
≥6 of 11	6,189	794 (12.8)
≥8 of 11	3,810	527 (13.8)

was described by previous studies that examined chickens and other distantly related organisms (1, 2, 20). We presume that this extensive difference in expression likely reflects the underlying biology of the two species. For example, mice and humans have very different metabolism and physiology—differences that are overall greater than tissue-specific differences even though tissue-specific genes have higher gene expression signals than broadly expressed genes. Indeed, differences in the transcriptional program between humans and mice have also been observed on a limited basis by other groups, for example, in the context of the inflammatory response (21).

In this study, we also examined replicability using a large number of datasets from biological replicates. More recently, work by Washietl et al. (15) suggested that noncoding transcripts are as reproducible as protein-coding ones. We did not find this to be the case, and our use of a much higher number of replicates for certain tissues may offer advantages over the limited number of samples that they examined. Moreover, the transcripts they considered were somewhat different from ours (a quarter of their set was antisense), and our analysis methods were simpler. In any case, we find that most noncoding transcripts cannot be found at an appreciable level in multiple individuals, in contrast to protein-coding transcripts. Such expression might be the consequence of biological noise or that many noncoding transcripts are expressed at a very low level or in a limited number of cells and not detected. Regardless, we found that even those noncoding transcripts that are highly replicable are not conserved between mice and humans. When the biological function of lincRNAs can be found, they appear to be involved in regulation; our results are consistent with the idea that regulatory information in general, such as transcription factor binding, is highly diverged (22, 23). Overall, our study demonstrates the extensive divergence in the expression of both noncoding genes as well as conserved, protein-coding genes that likely mediates the extensive differences between humans and mice.

Materials and Methods

Detailed experimental procedures are provided in *SI Materials and Methods*. RNA-seq data from all sources were processed in an identical fashion. Gene expression values were determined using Tophat and Cufflinks (24). For noncoding transcripts, we first assembled RNA-seq datasets from ENCODE, mouse ENCODE, HBM, and REMC with mouse and human data from Brawand et al. (5), Barbosa-Morais et al. (1), and Merkin et al. (2), as well as human brain, heart, lung, and adipose samples from the GTEx project (4). We performed ab initio transcript construction by Cufflinks on each human (348 samples) and mouse (103 samples) tissue sample, and then merged all results with Cuffmerge (24), and imputed their noncoding status with the Coding-Potential Assessment Tool (CPAT) (25).

ACKNOWLEDGMENTS. We thank Bing Ren and Bradley Bernstein for kindly providing us Roadmap Epigenomics Mapping Consortia histone data generated from their laboratories. We also thank Eurie Hong, Rama Balakrishnan, and Venkat Malladi for guidance in our Gene Ontology analysis and the support of Mid-America Transplant Services in this research effort. Stanford datasets were generated by the Stanford Center for Genomics and Personalized Medicine. This research was supported by Grant U54HG006996 (to M.P.S.). This work was also supported by National Institutes of Health (NIH) Fellowship Grants F32HL110473 and K99HL119617 (to S.L.). J.R.E. is an investigator with the Howard Hughes Medical Institute. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the NIH (commonfund.nih.gov/GTEx). Additional funds were provided by the National Cancer Institute (NCI), National Human Genome Research Institute, National Heart, Lung, and Blood Institute, National Institute on Drug Abuse, National Institute of Mental Health, and National Institute of Neurological Disorders and Stroke. Donors were enrolled at Biospecimen Source Sites funded by the NCI\Science Applications International Corporation (SAIC)-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care (X10S172). The Laboratory, Data Analysis, and Coordinating Center was funded through Contract HHSN268201000029C to the The Broad Institute. Biorepository operations were funded through an SAIC-F subcontract to the Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by a supplement to University of Miami Grant DA006227. Statistical methods development grants were made to the University of Geneva (Grant MH090941), the University of Chicago (Grants MH090951 and MH090937), the University of North Carolina–Chapel Hill (Grant MH090936), and to Harvard University (Grant MH090948). The datasets used for the analyses described in this article were obtained from dbGaP through dbGaP accession no. phs000424.v3.p1.c1 on February 19, 2013.

1. Barbosa-Morais NL, et al. (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science* 338(6114):1587–1593.
2. Merkin J, Russell C, Chen P, Burge CB (2012) Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* 338(6114):1593–1599.
3. Bernstein BE, et al. (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* 28(10):1045–1048.
4. Lonsdale J, et al.; GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 45(6):580–585.
5. Brawand D, et al. (2011) The evolution of gene expression levels in mammalian organs. *Nature* 478(7369):343–348.
6. Schug J, et al. (2005) Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol* 6(4):R33.
7. Ashburner M, et al.; The Gene Ontology Consortium (2000) Gene ontology: Tool for the unification of biology. *Nat Genet* 25(1):25–29.
8. Dunham I, et al.; ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.
9. Dong X, et al. (2012) Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol* 13(9):R53.
10. Wang Z, et al. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* 40(7):897–903.
11. Zhu J, et al. (2013) Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* 152(3):642–654.
12. Kim TK, et al. (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465(7295):182–187.
13. Cabili MN, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 25(18):1915–1927.
14. Harrow J, et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* 22(9):1760–1774.
15. Washietl S, Kellis M, Garber M (2014) Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res* 24(4):616–628.
16. Guttman M, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458(7235):223–227.
17. Derrien T, et al. (2012) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* 22(9):1775–1789.
18. Church DM, et al.; Mouse Genome Sequencing Consortium (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol* 7(5):e1000112.
19. Hinrichs AS, et al. (2006) The UCSC Genome Browser Database: Update 2006. *Nucleic Acids Res* 34(Database issue):D590–D598.
20. Papasaikas P, Valcárcel J (2012) Evolution. Splicing in 4D. *Science* 338(6114): 1547–1548.
21. Seok J, et al.; Inflammation and Host Response to Injury, Large Scale Collaborative Research Program (2013) Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc Natl Acad Sci USA* 110(9):3507–3512.
22. Borneman AR, et al. (2007) Divergence of transcription factor binding sites across related yeast species. *Science* 317(5839):815–819.
23. Cheng Y, et al. (2014) Principles of regulatory information conservation revealed by comparing mouse and human transcription factor binding profiles. *Nature*, in press.
24. Trapnell C, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7(3):562–578.
25. Wang L, et al. (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* 41(6):e74.