Predicting Tissue-Specific Enhancers in the Human Genome

Len A. Pennacchio[1,2], Gabriela G. Loots[3], Marcelo A. Nobrega[4] and Ivan Ovcharenko[2,5,*]

*[1]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA; [2]U.S. Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA; [3]Genome Biology Division, [4]Department of Human Genetics, University of Chicago, Chicago, IL 60637, and [5]Computational Directorate, Lawrence Livermore National Laboratory, 7000 East Avenue L-441, Livermore, CA. 94550, USA; *to whom correspondence should be addressed, tel. 925.422.5035; fax 925.422.2099; email, ovcharenko1@llnl.gov*

ABSTRACT

Determining how transcriptional regulatory signals are encoded in vertebrate genomes is essential for understanding the origins of multi-cellular complexity; yet the genetic code of vertebrate gene regulation remains poorly understood. In an attempt to elucidate this code, we synergistically combined genome-wide gene expression profiling, vertebrate genome comparisons, and transcription factor binding site analysis to define sequence signatures characteristic of candidate tissue-specific enhancers in the human genome. We applied this strategy to microarray-based gene expression profiles from 79 human tissues and identified 7,187 candidate enhancers that defined their flanking gene expression, the majority of which were located outside of known promoters. We cross-validated this method for its ability to *de novo* predict tissue-specific gene expression and confirmed its reliability in 57 of the 79 available human tissues, with an average precision in enhancer recognition ranging from 32% to 63%, and a sensitivity of 47%. We used the sequence signatures identified by this approach to assign tissue-specific predictions to ~328,000 human-mouse conserved noncoding elements in the human genome. By overlapping these genome-wide predictions with a large *in vivo* dataset of enhancers validated in transgenic mice, we confirmed our results with a 28% sensitivity and 50% precision. These results indicate the power of combining complementary genomic datasets as an initial computational foray into the global view of tissue-specific gene regulation in vertebrates.

INTRODUCTION

Increasing lines of evidence support the notion that the majority of functional elements in the human genome do not code for proteins [1, 2], yet our ability to systematically categorize and predict their function remains limited. For instance, most progress in elucidating transcriptional regulatory mechanisms has stemmed from computational and experimental analyses of transcription factors (TFs) acting within promoter regions of functionally related cohorts of genes. While informative [3-6], these studies did not assess distant-acting regulatory elements and thereby only sampled a limited portion of the vertebrate gene regulatory network [7, 8]. Several recent studies have provided conclusive evidence that the complex transcriptional expression pattern of human genes is mediated through multiple discrete sequences, often located hundreds of kilobases (kb) away from their core promoters [9, 10]. In these studies, evolutionary sequence conservation has served as a reliable indicator of biological activity, with an increasing number of distant noncoding evolutionary conserved regions (ECRs) validated as tissue-specific enhancers during development [9-15]. Although genome comparisons have provided a powerful approach for identifying noncoding ECRs that are under selective pressure, we have yet to develop reliable high-throughput computational methods for the discovery of distant regulatory elements with predetermined functional specificity. Here we describe a strategy for translating noncoding sequence data into transcriptional regulatory information that serves two vital purposes: to define the genetic vocabulary of tissue-specific gene regulation and to use this information to predict tissue-specific enhancers in the entire human genome, *de novo*. This approach combines genome-wide tissue-specific gene expression profiling data [16], vertebrate genome comparisons, and pattern analysis of transcription factor binding sites (TFBS), thus providing an initial foundation for deciphering vertebrate gene regulation from a purely computational strategy.

RESULTS

Predicting candidate regulatory elements for tissue-specific genes

As a first step towards directly relating gene expression to comparative sequence data, we clustered overlapping gene transcripts in the human genome and identified 18,504 unique protein-coding loci (the boundaries of each locus were defined by the neighbouring genes, independent of the absolute size of the locus; *see* Materials and Methods). We next assigned transcriptional information obtained from the GNF Atlas2 gene expression database (*gnfAtlas2)* [16] to these genomic loci. This included 79 human tissues with the majority of human loci (85%) successfully linked to their corresponding gene expression pattern. For each represented tissue, we defined two sets of genes: high expressors and low expressors. The high expressor group included the top 300 most highly expressed genes, while the low expressors included the bottom 5,000 least expressed genes. Our goal was to compare the genomic loci containing these two contrasting gene sets (across available tissues) to search for possible shared DNA sequence features in the vicinity of genes highly expressed in a given tissue.

We initially observed a strong correlation between the tissue specificity of a gene and the size of the locus, such that loci of highly expressed genes in the central nervous system (CNS) were on average significantly larger than the global median locus length. In contrast, loci corresponding to highly expressed genes in the immune system or various tumour tissues were significantly shorter (Figure 1; Figure S1). For example, the median locus length of a human gene highly expressed in fetal brain was 245kb, while genes

highly expressed in testis were on average 3.6 times shorter (68kb) (Figure S1). We also found that 10% of the brain and CNS loci coincided with vast noncoding regions termed gene deserts [9] in the human genome (a 2-fold increase over the expected value; *p*-value < 1e-7), consistent with the observation that most enhancers identified within gene deserts, to date, are biased towards brain and/or CNS expression during vertebrate development [9, 11, 17]. Finally, we observed a linear correlation between locus length and the number of human/mouse noncoding ECRs regardless of the tissue under investigation (Figure S1E).

Recent studies suggest that the most highly conserved noncoding ECRs within a locus commonly possess gene regulatory function [9, 18, 19]. Therefore, we selected the three most conserved human/mouse noncoding ECRs for each of the 18,504 human genes in our study, as well as noncoding ECRs overlapping with the gene's promoter region (defined as the 1.5kb region upstream of the transcription start site). These selection criteria generated a dataset of 60 thousand (k) candidate regulatory elements in the human genome, comprising ~1% of the entire genomic sequence (on average 4.2 candidate regulatory elements per locus). For comparison, functional noncoding elements have been previously estimated to span ~2-3% of the human genome [20], while the dataset defined in this study corresponds to the most highly conserved portion of this functional dataset. Classification of these elements based on their genomic location annotated 32% of these candidate regulatory elements as intergenic, 28% as promoter, 20% as intronic, 13% as 3'UTR, and 8% as 5'UTR. Approximately 24k of these elements flanked 6,059 genes with the highest gene expression in at least one of the 79 tissues and ~55k of these elements flanked 15,632 genes with the lowest gene expression (serving as a negative control dataset).

To explore the sequence motifs of these noncoding ECRs linked to genes displaying high versus low expression in the same tissue, we used a previously described motif identification strategy [21] and identified 1.8 million (M) evolutionarily conserved putative TFBS within this dataset (*see* Materials and Methods). We found that several individual motifs were significantly enriched in 43 human tissues (Table S1). For example, we observed a strong association among NRF1, OCT, MEF2 and CREB, transcription factors known to play key roles in brain and neuronal development [22-26] in candidate regulatory elements from loci highly expressed in human fetal brain (Table S1A). However, as described in further detail below, no single TF by itself was sufficient to predict where a candidate enhancer will drive gene expression.

Determining sequence signatures of candidate tissue-specific enhancers

Based on the presumed combinatorial nature of multiple TFs to mediate a given enhancer's activity, we employed an analysis strategy that simultaneously scored the impact of multiple TFBS motifs in an attempt to classify noncoding ECR candidate enhancers based on sequence signatures that define gene expression in a particular tissue. This was accomplished by assigning a weight to each TF that quantifies its association with a given tissue. By summing these TFBS motif weights, we were thus able to generate a regulatory potential tissue-specificity score for each of the 24k candidate enhancers of highly expressed genes as well as 55k background elements of the low expressed genes. This scoring scheme provided the means to optimize TF weights in an

effort to enrich for positively scoring candidate enhancers in tissue-specific loci of highly expressed genes while simultaneously minimizing their presence in loci of genes with low expression (independently performed for each tissue; *see* Materials and Methods). We named this approach *Enhancer Identification* (or *EI*) and its application allowed us to select candidate tissue-specific enhancers from the pool of conserved noncoding elements in loci of genes highly expressed in a given tissue (Figure 2). We performed *EI* analysis independently on both human and mouse gene expression data and while we primarily utilized human statistics in our discussion, mouse data analysis is provided in the Supplementary Materials (Figures S1 and S2; Tables S2 and S7).

The *EI* scoring optimization allowed us to maximize our resolving power to the point where 60% (+/-5%) of genes highly expressed in a tissue group contain signatures that are present in less than 15% of the low expressed genes, for any given tissue (Figure 3B). For example, *EI* identified at least one fetal lung candidate enhancer for 65% of genes with high fetal lung expression, while no such candidates were identified in the non-intergenic regions (promoter, UTR, or intronic) of greater than 86% of genes with low fetal lung expression (intergenic regions were excluded from the negative control group to prevent potential associations with neighboring genes' regulation; *see* Materials and Methods). Of the original 24k candidate regulatory elements linked to genes highly expressed in one or more of the 79 available tissues, *EI* optimization identified 7,187 candidate enhancers with signatures that defined tissue-specific expression. Through this consolidation of the dataset, we found that 47% of human noncoding ECRs defined as candidate enhancers were predictive of expression in more than one tissue, consistent with our finding that 66% of the human genes in this study are highly expressed in multiple tissues. Since these

candidate enhancers were assigned to different tissues that are functionally related (for example, CD4 and CD8 T-cells) (Table S6), it is possible that the transcriptional regulation of genes expressed in similar tissues could be achieved through shared gene regulatory mechanisms. These findings are consistent with *in vivo* expression data derived from enhancer scans in transgenic mice indicating that one-third of embryonic enhancers active during a single time-point in development drive expression in more than one tissue type [9, 11, 12, 27]. Finally, we also found that 20% of highly expressed genes within our dataset harbor more than one distinct candidate enhancer predicted to be active in the same tissue, supporting the hypothesis that certain genes contain multiple discrete regulatory elements that overlap in their enhancer activity [11, 28].

Since the *EI* method is based on the weighting of multiple TFs for their association with tissue-specific expression, we sought to further explore the nature of this combinatorial TF scoring scheme. We found that in no case was a single TF sufficient to predict tissue-specific gene expression, supporting the notion that tissue-specific gene regulation is a direct result of interplay among multiple TFs. To quantify the impact of an individual *i*-th TF on predicting gene expression in a particular tissue *t*, we calculated the *TF importance* parameter (I_i^t) defined as the product of the TF occurrence (percentage of tissue-specific candidate enhancers with a particular conserved TFBS) and its weight, in a tissue-specific group of candidate enhancers (Table S2). Since TF importance compounds the effects of TF occurrence and weight, it presents an integrative measure of the TF's role in generating high positive scores of tissue-specific candidate regulatory elements. At the same time, it minimizes the impact of TFs that are rare or have small weights and thus do not contribute significantly to establishing either a positive or a negative tissue-specificity score. This

quantification allowed for the identification of cohorts of TFs in candidate enhancers potentially involved in tissue-specific regulatory networks, i.e. those TFs both with high weights and high occurrences. As an example of a high TF impact on tissue-specific regulation, the photoreceptor-specific CRX TF has the highest importance parameter value in eye development (Table S2) consistent with the known function of this regulatory protein in Cone-Rod Dystrophy (CRD), an inherited progressive disease that causes deterioration of the cone and rod photoreceptor cells and leads to blindness [29].

To illustrate this method's ability to predict functional enhancers, we examined two well characterized enhancers, one for skeletal muscle and one for liver, flanking the human cardiac/slow skeletal muscle troponin C (*TNNC1*) and the apolipoprotein B (*APOB*) genes, respectively (Figure 4). An *EI* scan of the *TNNC1* locus first identified 4 noncoding ECRs (out of 12 total) as candidate regulatory elements (two intergenic, one intronic, and one promoter element). Subsequent *EI* optimization then correctly predicted the noncoding ECR in intron 1 as a skeletal muscle enhancer in precise agreement with the previously defined *TNNC1* skeletal muscle enhancer [30, 31]. In a second example, *EI* correctly identified the *APOB* promoter element as a fetal liver (and adult liver) enhancer and predicted transcription factors HNF4 and C/EBP to be activating *APOB* expression, in concordance with previous experimental studies [32].

To explore the possibility of synergistic TF linkage that may be biologically required for directing tissue-specific gene expression, we extracted the top 10 scoring TFs for each tissue based on their importance in predicting tissue-specific expression. As an example, we focused on the TF characteristics of two similar tissue types: heart and skeletal muscle

(Figure 5A) (a complete list of the top TF for each tissue is provided in Table S2). We observed that 5 of the top 10 TF predictions for both these muscle types are shared, four of which (MEF2, SRF, Myogenin, and ERR1) are strongly linked to transcriptional regulation in muscle tissue and associated to various human cardiac myopathies [33-36]. As a second example, the top 10 TFs predictors of liver expression included Hepatocyte Nuclear Factor 1 (HNF1), HNF4, PPAR, SREBP1, HNF4-DR1, NR2F2, and FRX-IR1 (Figure 5A), all of which are known regulatory proteins important in liver function [37-39]. These two examples highlight the biological plausibility of the *EI*'s method of tissue-specific gene expression prediction.

To globally address the power of the predicted TF-tissue associations in addition to the support gained from the above selected examples, we mapped TFs to the human genome and determined the tissue *gnfAtlas2* expression profile for each TF gene. Our rationale was that if tissue-specific gene expression predictiveness is based on TFBS density in candidate enhancer sequences, then the TF required for this function should be expressed in the tissue of activity. Thus, we attempted to correlate positive TF importance with the level of TF gene expression in the available 79 human tissues. This was accomplished by adjusting the minimal TF importance threshold increasingly from -0.25 to +0.25 (thus gradually increasing the ratio of TFs with positive importance values in the group) to determine if TF expression and enhancer predictiveness were positively correlated (Figure 5B). Indeed, we observed that ~60% of predicted positive TF-tissue associations corresponding to TF importance thresholds of $\geq$0.1 were supported by an increased level of gene expression in the associated tissue (Figure 5B). One possible explanation for the lack of total concordance between the predicted TF-tissue associations and tissue-

specificity is the ubiquitous nature of TF gene expression that often leads to ambiguous definitions of tissue-specificity with increased and decreased level of gene expression in *gnfAtlas2*. Manual curation of these interactions, revealed that 90% (142/158) of predicted TF-tissue associations with ≥0.25 TF importance threshold are supported by published literature or alternative sources of experimental evidence (*see* Supplementary Materials; Table S3).

Since any parametric optimization approach could potentially introduce "over-fitting" - the identification of random profiles that separate genes with high versus low expression purely by chance - we attempted to cross-validate our results. This was accomplished by characterizing the ability of the *EI* method to annotate tissue-specific enhancers in loci of highly expressed genes without any *a priori* knowledge of tissue-specificity of gene expression (i.e. these genes were excluded from the training set; *see* Materials and Methods). This approach allowed us to quantify both the method's precision (defined as the proportion of predicted elements that act as tissue-specific enhancers) and sensitivity for each tissue (Figure 3). Through this analysis, we observed a high variability in *EI* precision across the 79 sampled human tissues, and hence were classified into three quality groups (Figure 3A): (1) poor (lower-bound precision, $P^{\downarrow}$, less than 20%), (2) good (lower-bound precision, $P^{\downarrow}$ between 20-40%), and (3) excellent (lower-bound precision, $P^{\downarrow}$ greater than 40%) (Table 1). Next, we grouped lower- and upper-bound precision values to use their average as an estimate for the true precision and found that 72% (57/79) of human tissues have an average precision of 40%. These data allowed us to conclude that over-fitting did not account for the majority of signals obtained from the *EI* predictive method. In contrast, *EI* was suboptimal for the remaining 22 human tissues

which fell into the poor category where the average precision was below 30%, indicating that over-fitting likely explained a significant fraction of the signature derived for these tissues. Tissues comprised within this category mainly consisted of multiple gland and germ tissues as well as structures such as the appendix and olfactory bulb. Based on these observations, we placed low significance values on the predictions derived for these tissues and their enhancer predictions should be treated cautiously as they are likely to represent false-positives. In contrast, the average precision of the excellent group was above 50% for 12 tissues including heart, liver, tongue, blood and several immune tissues (Table 1). Thus, these tissue-types bear the highest confidence of *EI* predictions.

Assigning tissue-specific predictions to conserved noncoding sequences in the human genome

Since the *EI* method can generate tissue-specific predictions for any conserved element, we used this approach to score 364k previously reported candidate enhancers in the human genome [19] (*see* Materials and Methods). In total, *EI* was able to assign a tissue-specificity to 90% (328k) of these elements covering 4.0% of the human genome. This large dataset comprises tissue-specificity predictions for the majority (86%) of genes in the human genome and represents a useful resource for detailed experimental follow-up studies by gene-centric investigators. We anticipated this latter approach to feature a relatively high-rate of false-positive predictions since tissue-specificity of the predictions could not always be supported by high expression of flanking genes. However, this genome dataset could represent an important resource for prioritizing tissue-specific enhancers in loci of genes with known functions when one is interested in sifting through multiple

evolutionary conserved elements and prioritizing only those that correspond to candidate enhancers with matching tissue-specificity. We should note that we observed an overlap in tissue-specificity predictions as a result of several related tissues having similar *EI* recognition profiles (Tables S6 and S7). For example, 24% of CD4+ T-cell predicted elements were also classified as CD8+ T-cell, while 14% of liver predicted elements were simultaneously classified as fetal liver. In contrast, only 0.8% of CD4+ T-cell predicted elements were simultaneously classified as fetal liver predictions. This suggests that the direct *EI* tissue-specificity annotation of conserved elements may fail to distinguish between closely related tissues, but can possibly distinguish between major tissue categories or different organs.

Experimental validation of tissue-specific enhancer predictions

Based on our genome-wide predictions of tissue-specific activities for all noncoding ECRs, we sought to determine their performance against existing enhancer data of gene expression derived from transgenic mouse studies. As a test bed, we examined the *EI* tissue-specific predictions for 5 previously characterized enhancers expressed in the brain and nervous system in the 1Mb region upstream of the *DACH1* gene [9]. We found that 3 of these elements were predicted to have enhancer activity limited to brain tissues while the 2 remaining elements were not assigned to any tissue (Table S4). While these initial correlations were based on a small sample set, the statistical significance of this match is supported by a *p*-value of 0.004 (*see* Supplementary Materials).

To expand these data beyond the limited published *in vivo* data for distant-acting enhancer elements, we next performed a large-scale analysis of our whole genome predictions

against a publicly available dataset of 106 elements that have been shown to act as tissue-specific enhancers in the mouse at embryonic day 11.5 of development (E11.5) (data available at http://enhancer.lbl.gov) [27] (Table S5). In this dataset we found 71 (67%) enhancers to dictate expression in forebrain, midbrain, hindbrain, and/or neural tube. We thus assessed whether whole genome *EI* tissue-specific predictions overlap with these *in vivo* characterized enhancers. Indeed, 28% (20/71) of these elements were selectively predicted as enhancers active in the brain and/or the nervous system. In addition, another 7% (5/71) of these validated CNS-specific enhancers had *EI* predictions with a mixed annotation of brain/CNS and another organ/tissue, suggesting these elements are possibly multi-functional. We also observed 21% (15/71) of predictions provided tissue annotations inconsistent with the experimental data, while the remaining 44% (31/71) elements had no tissue-specific prediction(s) (Table S5). This corresponded to 28% sensitivity and 50% precision in recognition of brain and CNS-specific enhancers using the *EI* method, *de novo*. By calculating the distribution of brain/CNS predictions in a large random dataset (*see* Supplementary Materials) we found the overlap of this analysis to be 2.5-fold greater than what would be expected by chance, corresponding to a *p*-value of 0.0001.

To further explore the relationship between the 20 concordant *EI* whole genome predictions and the existing *in vivo* nervous system dataset described above, we examined the distribution of the predictions within the 18 different brain tissues present in *gnfAtlas2* database. While we found 4 or less of these *in vivo* defined CNS-enhancers were predicted to be expressed in each of the 17 adult brain tissues present in the expression annotation, 11 of them were annotated solely to the fetal brain category in the *gnfAtlas2*

[the probability of this observation being random is less than 1e-7 (*see* Supplementary Materials)]. This high ratio of fetal brain predictions is consistent with the entire *in vivo* expression dataset that corresponds to a single time point of enhancer analysis during embryonic development at E11.5. This suggests that the fetal brain enhancer recognition profile of *EI* is a specific signature of *in vivo* embryonic brain enhancers, in contrast to enhancers active in specialized compartments of the adult brain. Additional *in vivo* datasets based on non-embryonic time points will further aid in assessing the ability of this approach to predict enhancer elements active in adult tissues.

DISCUSSION

Deciphering the genetic code of gene regulation in vertebrate genomes remains a significant challenge that has been partially aided by the availability of the human and other vertebrate genome sequences. However, while techniques such as comparative genomics can enrich for putative enhancer sequences based on evolutionary conservation, predicting their tissue-specificity has been difficult. Nevertheless, several proof of principle studies have demonstrated that there is a vaguely defined but computationally recognizable genetic code of gene regulatory elements corresponding to selected biological functions [40-42]. Additional studies have also revealed the power of microarray expression data to correlate the distribution of evolutionary conserved putative TFBS in the promoters of co-expressed human (and mouse) genes with the level and dynamics of gene expression [43, 44]. These early focused studies suggest that computationally predicting enhancer function is a solvable problem. We therefore developed a multi-faceted approach coupling TF binding specificities, comparative genomics, and microarray expression data in an attempt to recognize sequence signatures within putative enhancer

elements in the human genome. Through these efforts, we show that it is possible to identify tissue-specific enhancers for 72% of human sampled tissues by constructing sequence recognition profiles based on the distribution of TFBS in noncoding ECRs linked to genes expressed in similar tissues.

One of the inferences we can formulate based on the results of the *EI* method introduced here is the proportion of enhancer activity assigned to promoters versus more distant-acting sequences. This measurement was possible since the *EI* approach utilizes the 3 most highly conserved human-mouse elements neighboring the gene under investigation and thus goes beyond promoter only exploration of *cis*-regulatory features; the dominant method currently employed in regulatory genomics. Through the comparison of the *EI* signal strength in promoter versus non-promoter conserved elements, we found that only 23% of *EI* candidate enhancers map to promoter regions of corresponding genes. While a caveat to this analysis is the incomplete status of precisely defined promoter boundaries, this result is consistent with ChIP-chip and *in vivo* enhancer studies which also suggest that more than half of human genes potentially rely on distant mechanisms of gene regulation [7-10].

Since this method can be applied to the analysis of any set of co-expressed genes, this provides a rapid and efficient approach for translating gene expression data into function-specific gene regulatory principles. Thereby, it should be straightforward to extend this method to other tissues, developmental time-points, or functional gene categories (such as Gene Ontology and KEGG datasets [45, 46], for example). In addition, the elements identified in this study represent a dataset of tissue-specific candidate enhancers that could

be used to guide the ongoing large-scale experimental efforts aimed at exploring transcriptional regulatory function in human, mouse and other vertebrate genomes. Since the backbone of the *EI* optimization method is the association of TFs with tissue-specificities, we were able to predict over 7k such associations and retrieve experimental evidence for 90% of them in a selected group of 158 TF-tissue associations (at a TF importance threshold of 0.25). Furthermore, characterization of TF spacing in predicted tissue-specific enhancers allowed us to extract approximately one thousand TF pairs significantly enriched as putative synergistic activators in a given tissue (*see* Supplementary Materials). While we were able to bring forth published evidence for several predicted TF co-occurrences, the vast majority of TF-tissue linkages and their potential interactions represent novel regulatory associations that could be used in facilitating future studies of the complexities of gene regulatory pathways.

It is likely that general approaches that assign tissue-specificity to enhancer function will greatly improve over time. Current challenges include the varying quality of the human and mouse microarray expression data and their primary adult material source that serves to define gene expression tissue-specificities, the lack of *in vivo* spatial and temporal enhancer data to further serve as training sets, and our incomplete knowledge of TFs as well as their precise sequence-based binding properties currently available in the TRANSFAC database [47]. In addition, the comparative analysis exploited here was limited to human-mouse genome alignments under one alignment and conservation scoring method. Nevertheless, despite these limitations, our finding of *EI*'s ability to identify tissue-specific enhancers with the available datasets is encouraging, and represents a platform for further efforts in this area.

In summary, the data presented here lend further support to the notion that sequence-based features in vertebrate *cis*-regulatory elements are computationally recognizable, similar to previous successes in the inference of coding, intron-exon, core promoter, and repetitive DNA sequence signatures. Even though our study is limited by the availability and reliability of position weight matrices (PWM) of known TFs, the methods introduced here present a universal framework for the *de novo* prediction of regulatory elements with shared biological function, as well as for defining novel interactions among transcription factors that can explain tissue-specific function of enhancer elements. Future computational efforts linked to topics such as human disease and vertebrate phenotypic diversity are likely to provide insights into gene regulatory mechanisms of unexplained biological phenomena.

MATERIALS AND METHODS

Gene annotation and expression data. The UCSC Genome Browser [48] database was used to extract gene positional information. Human and mouse "*knownGene*" transcripts[49] were mapped to the NCBI Build 35 of the human (hg17) and mouse (mm7) genomes and grouped into 18,504 and 17,636 non-overlapping loci, respectively. GNF Novartis Atlas2 tissue-specific gene expression [16] was extracted from the *gnfAtlas2* table and mapped to genes using the *knownToGnfAtlas2* table (both tables are available in the UCSC Genome Browser database). At least one tissue-expression profile was available for each of the 15,690 human and 14,303 mouse genes.

Profiling putative TFBS in the human genome. Human-mouse ECR Brower genome alignments [50] were processed by rVista 2.0 [21] to identify evolutionary conserved putative TFBS in the human and mouse genomes. We utilized previously described optimized PWM thresholds [51, 52] to limit the appearance of predictions to 5 TFBS per 10kb of random sequence. In total, 13.4M conserved putative TFBS were identified using 554 TRANSFAC 9.4 PWMs [47] and 3 manually-curated PWMs for TBX5, NKX2.5, and GLI TFs. These putative TFBS were then grouped into 364 separate TF families (as several TFs have multiple overlapping definitions in the TRANSFAC database) – we refer to those TF families simply as TFs in the text – and superimposed with the 60k candidate enhancers to construct a dataset of 1.8M putative TFBS in candidate enhancers for the study.

Tissue-specific enrichment of individual putative TFBS. We calculated the ratio of highly-to-lowly expressed gene loci containing putative TFBS in candidate enhancers for different TFs and utilized the hypergeometric distribution with Bonferroni correction for

multiple testing to identify significantly enriched putative TFBS in different tissues (Table S1).

Assigning tissue specificity scores to candidate enhancers. We utilized the distribution of putative TFBS inside a candidate enhancer (or noncoding ECR) to assign a tissue-specificity score to that element. First, we assigned a tissue specificity weight w_i^t to each *i*-th TF as a measure of its association with the tissue *t*. Next, the distribution of putative TFBS in the *k*-th candidate enhancer was scored to define candidate enhancer tissue-specificity:

$$S_k^t = \sum_{i=1..n_{TF}} w_i^t N_k^i ,$$

where N_k^i is the number of *i*-th TF putative TFBS located in the *k*-th candidate enhancer and the summation is performed over all n_{TF} TFs. TF weights were allowed to vary from -1 to 10. Large positive weights w_i^t indicate a strong correlation between the *i*-th TF and the *t*-th tissue-specificity, while large negative weights indicate the unlikely presence of the *i*-th TF in a candidate enhancer that is active in the tissue *t*.

***EI* optimization to define TF tissue specificity weights.** To identify tissue-specific candidate enhancers, we performed the Brent's method optimization of TF weights w_i^t that maximizes the number of positively scoring candidate enhancers in loci of highly expressed genes ($L+$) and simultaneously minimizes the number of positively scoring candidate enhancers in loci of lowly expressed genes ($L-$). Optimization was performed independently for different tissues. To ensure a reliable and specific identification of noncoding features in loci of highly expressed tissue-specific genes we included a large

background dataset comprising 5,000 loci of lowly expressed genes assigned to each tissue. A scoring function F^t,

$$F^t = \sum_{k \subset L+}^{S_k^t > 0} \log(1 + S_k^t) - \lambda \cdot \frac{N_{E+}}{N_{E-}} \cdot \sum_{l \subset L-}^{S_l^t > 0} \log(1 + S_l^t)$$

containing summations over all positively scoring candidate enhancer associated with $L+$ ($k \subset L+$) and $L-$ ($l \subset L-$) was maximized to perform the optimization of weights (the distribution of positively scoring candidate enhancers in $L+$ and $L-$ was free to change following the change in TF weights). The ratio of the total number of candidate enhancers in $L+$ (N_{E+}) to the total number of candidate enhancers in $L-$ (N_{E-}) was introduced to the scoring function to account for differences in the number of highly and lowly expressed genes and the number of corresponding candidate enhancers. λ, or the signal enrichment coefficient served to increase the negative impact of positively scoring noncoding ECRs in $L-$. λ was selected as 1 during the initial optimization step and then gradually increased to 10,000 to achieve the greatest separation between loci of highly and lowly expressed genes. Optimization was initialized with TF weights estimated using the density of putative TFBS in $L+$ and $L-$ as

$$w_i^t = \frac{\sum_{k \subset L+} N_i^k \Big/ N_{E+}}{\sum_{k \subset L-} N_i^k \Big/ N_{E-}} - 1.$$

Initial TF weights were upper-bounded by 1 and the optimization was performed contiguously and recursively for each *i*-th TF. It was interrupted after achieving an increase less than 0.1 in the scoring function during a cycle of TF weights optimizations across all TFs. An important property of this optimization is the dynamic selection of the

positively scoring subset of tissue-specific candidate enhancers from the original set of candidate enhancers.

Cross-validation.

Cross-validations was performed by expanding the dataset of highly expressed genes to 400 and further subdividing this set into 2 groups consisting of 300 genes for *EI* optimization and 100 genes for testing the signal recognition (test genes were not included in the optimization). Cross-validation was repeated four times to estimate the statistical error in precision and sensitivity. The four cross-validation replicas of 100 test genes did not overlap with each other to ensure that four independent quantifications are carried out. Similarly, each time a different group of 500 genes was removed from the background dataset for each cycle of *EI* optimization. Using this approach we performed four independent rounds of *EI* optimization with 300 signal (highly expressed) and 4,500 background (lowly expressed) genes and subsequently applied the generated TF profiles to independently calculate the percentage of tissue-specific candidate enhancers from the 100 test (R^+) and 500 control (R^-_{int}) datasets. Optimization and testing of control genes was restricted to non-intergenic regions to avoid potential cross-talk with tissue-specific enhancers controlling the expression of neighboring genes. Therefore, *EI* precision in recognizing tissue-specific enhancers (which measures the ratio of true positive tissue-specific enhancers in the full dataset of predicted elements)

$$P^{\uparrow} = \frac{R^+ - R^-_{\text{int}}}{R^+}$$

represents the upper-bound estimate of the precision. By excluding the non-intergenic component of loci of test highly expressed genes from the quantification, after that, one decreases the percentage of recognized test gense to R^+_{int} and the corresponding precision

$$P^{\downarrow} = \frac{R^{+}_{\text{int}} - R^{-}_{\text{int}}}{R^{-}_{\text{int}}}$$

then represents the lower-bound of the precision of the method. By averaging these two values we were able to estimate the real precision of the method ($\overline{P}$). Also, R^{+} served as an estimate for the lower-bound sensitivity of the method ($Sn^{\downarrow}$) in the *de novo* recognition of tissue-specific enhancers (which measures the probability of a tissue-specific enhancer to be detected by *EI*) in cases where the corresponding gene does not belong to a specific group of highly expressed genes.

Mapping TFs to known transcripts. We used information on TF gene names corresponding to PWM used by the TRANSFAC database for automated (and manually-curated after that) GenBank queries to identify the name and chromosomal location of the human gene best matching each TF. For example, we were able to map the AML1 TF matrix to the human runt-related transcription factor 1 (RUNX1) residing at chr21 (q22.12). In total, 314 of 364 utilized TRANSFAC TFs were successfully mapped to human genes. In several instances, a TF mapped to more than one gene locus (in the case of E2F1DP2 hetero-dimer, the TF complex mapped to E2F1 and TFDP2 genes; similarly the SREBP TF mapped to both SREBP1 and SREBP2 genes); in such cases the expression profiles were averaged across all genes corresponding to the TF or TF-complex.

Permutation analysis to identify significant tissue-specific inter-TF interactions. We analyzed the distribution of positively scoring TFBS in tissue-specific candidate enhancers independently of each tissue. Only TFs with individual TF occurrence >= 5% or TF importance >= 0.05 were sub-selected for the analysis. The number of TF-TF pairs with the minimal and maximal inter-TF distances of 5 and 100, respectively, was calculated for each pair of TFs. 10,000 permutations randomizing the distribution of TF name labels

among different TFBS were performed. The total number of TFBS for each TF as well as positions of individual TFBS was kept intact during the randomization. We then extracted a subset of TF pairs that occur less frequently in 95% of permutation tests than in the original distribution (corresponding to a p-value < 0.05 to observe the original distribution by chance) and that corresponded to at least a 2-fold increase in their density in the original distribution as compared to an average pair density in permutation tests.

Assigning tissue-specific enhancer predictions to a whole genome dataset of human-mouse noncoding ECRs. We profiled TFBS distributions for 364k previously catalogued human/mouse conserved noncoding sequences [19], and a comprehensive 1.4M noncoding ECRs set for the entire human genome, to identify 328k and 588k elements, respectively, that have a positive tissue-specificity score according to *EI* tissue-specificity profiles. We used a p-value 0.05 cut-off for the 364K set that corresponds to an estimate of 0.05 false positive enhancer predictions per 10kb of random sequence [19]. In cases of multiple tissue associations assigned to an elements we selected up to three top scoring associations with the score of at least 50% of the most top scoring tissue-association (data available at http://www.dcode.org/EI). The same score selection procedure was applied for the analysis of organ specificities.

ACKNOWLEDGEMENTS

We would like to thank Shyam Prabhakar and Alex Poliakov for providing Gumby enhancer predictions in the human genome. G.G.L. and I.O were supported by LLNL LDRD-04-ERD-052 grant; and I.O. was in part supported by LLNL LDRD-06-ERD-004 grant. The work was performed under the auspices of the United States Department of Energy by the University of California, Lawrence Livermore National Laboratory

Contract W-7405-Eng-48. L.A.P. was supported by the Grant HL066681, Berkeley-PGA, under the Programs for Genomic Application, funded by National Heart, Lung, & Blood Institute, and HG003988 funded by National Human Genome Research Institute and performed under Department of Energy Contract DE-AC02-05CH11231, University of California, E.O. Lawrence Berkeley National Laboratory.

COMPETING INTEREST STATEMENT

The authors declare they have no competing financial interests.

FIGURE LEGENDS

Figure 1. Locus length of human genes highly expressed in several tissues. Bound horizontal lines represent the inter-quartile range (the distance between the 25th and 75th percentiles) of the tissue-specific distributions, solid colored rectangles measure the standard error in median calculations (median locus length is depicted by a white line inside colored rectangles. Statistically significant (<5%) distributions that deviate from the global median (represented by a solid vertical black line) are marked by an asterisk on the left side bar. Tissues with a median value two-fold smaller or larger than the median are marked by a vertical line on the left side bar.

Figure 2. Schematic of the general *EI* strategy for defining signatures of tissue-specific enhancers.

Figure 3. Precision (A) and sensitivity (B) of the *EI* method in recognizing human tissue-specific enhancers. Lower- and upper- bound estimates of precision along with their average are given in red, blue, and black on precision plots (A), respectively. Standard deviation is also depicted for each lower and upper bound estimates. Tissues are split into poor, good, and excellent groups based on the lower-bound estimate of the precision. See Figure S2 for corresponding mouse data. Navy and red curves on sensitivity plots (B) measure the percentage of high and low expressed gene loci with tissue-specific enhancers, respectively; while the purple curve estimates *EI* sensitivity for *de novo* enhancer recognition.

Figure 4. *EI* annotation of *TNNC1* skeletal muscle (A) and *APOB* (B) liver enhancer. Zoomed-in view of Mulan [51] human/mouse evolutionary conservation profiles for these loci depicts candidate enhancer elements followed by profile of conserved TFBS present within.

Figure 5. Importance and occurrence of individual TFs in candidate enhancers corresponding to mouse liver, B-cells, heart, and skeletal muscle (A). Binning of 25k predicted TF-tissue associations by a minimal TF importance threshold (B). The number of TF-tissue associations almost does not change in the area of negative TF importance thresholds and rapidly decreases in the area of positive TF importance thresholds (dark red graph; right y-axis) indicative of a small number of TFs with large positive importance values and an even smaller number of TFs with large negative importance values. The percentage of TF-tissue associations that are confirmed by an increase in TF gene expression (orange bars) increases with the increase of minimal TF importance (followed by the corresponding decrease in the number of non-confirmed associations – blue bars). As ~60% of predicted TF-tissue associations with a minimal TF importance of 0.1 are supported by an increased level of TF gene expression in the corresponding tissue this threshold could serve as a cut-off of reliability in TF-tissue association predictions.

Figure 1.

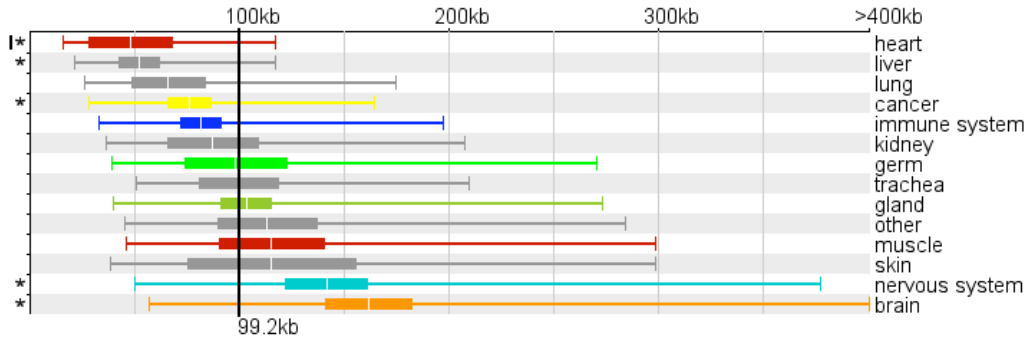

100kb
200kb
300kb
>400kb
heart
liver
lung
cancer
immune system
kidney
germ
trachea
gland
other
muscle
skin
nervous system
brain
99.2kb

Figure 2.

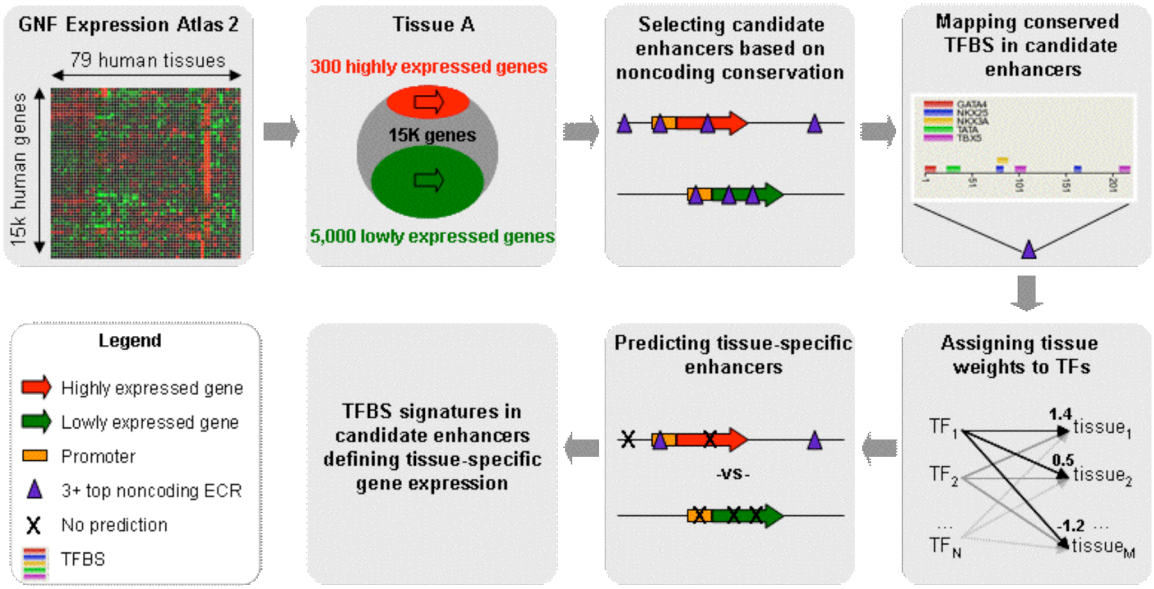

GNF Expression Atlas 2
79 human tissues
15k human genes
Tissue A
300 highly expressed genes
15K genes
5,000 lowly expressed genes
Selecting candidate enhancers based on noncoding conservation
Mapping conserved TFBS in candidate enhancers
Assigning tissue weights to TFs
TF1
TF2
TFN
1.4
tissue1
0.5
tissue2
-1.2
tissueM
Predicting tissue-specific enhancers
-vs-
TFBS signatures in candidate enhancers defining tissue-specific gene expression
Legend
Highly expressed gene
Lowly expressed gene
Promoter
3+ top noncoding ECR
No prediction
TFBS

Figure 3.

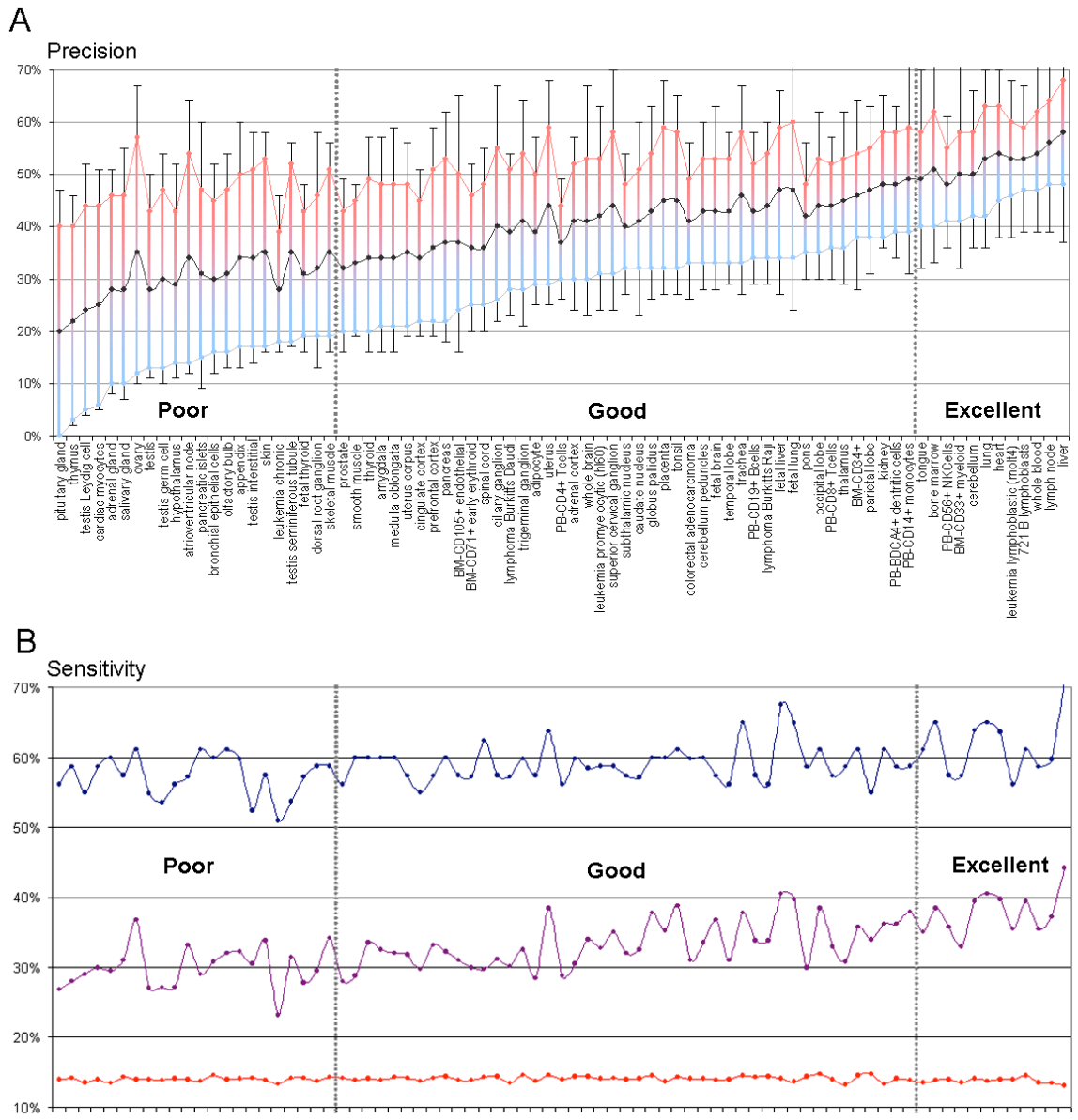

A
Precision
70%
60%
50%
40%
30%
20%
10%
0%
Poor
Good
Excellent
pituitary gland
thymus
testis Leydig cell
cardiac myocytes
adrenal gland
salivary gland
ovary
testis
testis germ cell
hypothalamus
atrioventricular node
pancreatic islets
bronchial epithelial cells
olfactory bulb
appendix
testis interstitial
skin
leukemia chronic
testis seminiferous tubule
fetal thyroid
dorsal root ganglion
skeletal muscle
prostate
smooth muscle
thyroid
amygdala
medulla oblongata
uterus corpus
cingulate cortex
prefrontal cortex
pancreas
BM-CD105+ endothelial
BM-CD71+ early erythroid
spinal cord
ciliary ganglion
lymphoma Burkitts Daudi
trigeminal ganglion
adipocyte
uterus
PB-CD4+ Tcells
adrenal cortex
whole brain
leukemia promyelocytic (HL60)
superior cervical ganglion
subthalamic nucleus
caudate nucleus
globus pallidus
placenta
tonsil
colorectal adenocarcinoma
cerebellum peduncles
fetal brain
temporal lobe
trachea
PB-CD19+ Bcells
lymphoma Burkitts Raji
fetal liver
fetal lung
pons
occipital lobe
PB-CD8+ Tcells
thalamus
BM-CD34+
parietal lobe
kidney
PB-BDCA4+ dentritic cells
PB-CD14+ monocytes
tongue
bone marrow
PB-CD56+ NKCells
BM-CD33+ myeloid
cerebellum
lung
heart
leukemia lymphoblastic (molt4)
721 B lymphoblastis
whole blood
lymph node
liver
B
Sensitivity
70%
60%
50%
40%
30%
20%
10%
Poor
Good
Excellent

Figure 4.

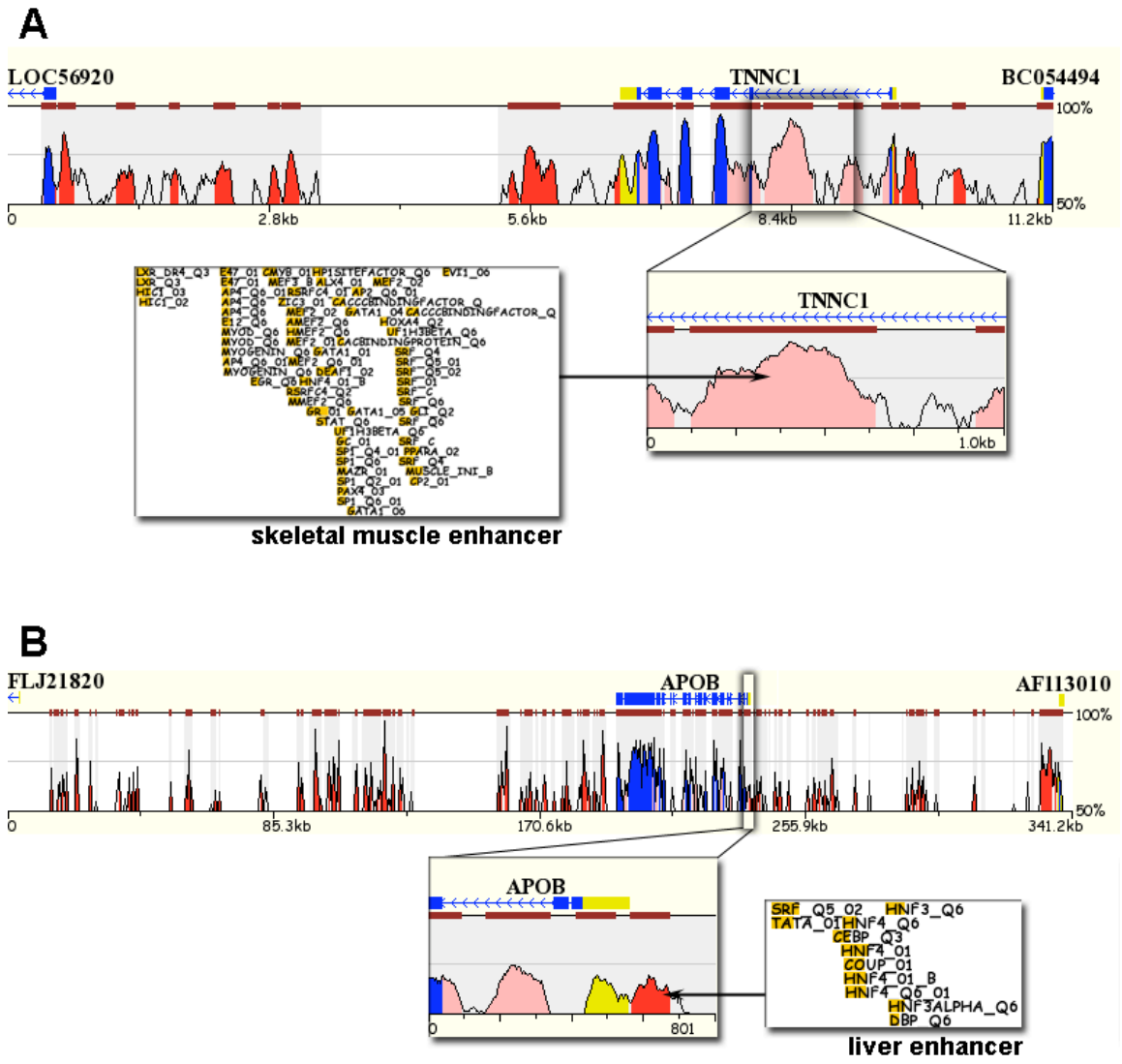
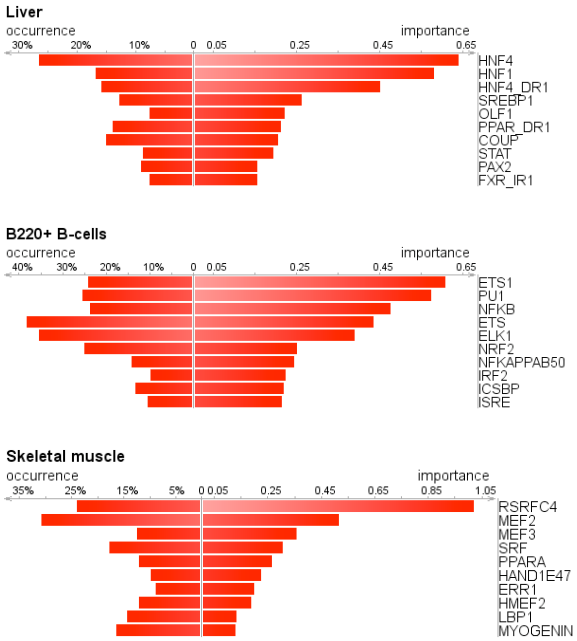

A
LOC56920
TNNC1
BC054494
100%
50%
0
2.8kb
5.6kb
8.4kb
11.2kb
TNNC1
1.0kb
skeletal muscle enhancer
B
FLJ21820
APOB
AF113010
100%
50%
0
85.3kb
170.6kb
255.9kb
341.2kb
APOB
801
liver enhancer

Figure 5.

A)


Liver
occurrence
importance
30% 20% 10% 0 0.05 0.25 0.45 0.65
HNF4
HNF1
HNF4_DR1
SREBP1
OLF1
PPAR_DR1
COUP
STAT
PAX2
FXR_IR1
B220+ B-cells
occurrence
importance
40% 30% 20% 10% 0 0.05 0.25 0.45 0.65
ETS1
PU1
NFKB
ETS
ELK1
NRF2
NFKAPPAB50
IRF2
ICSBP
ISRE
Skeletal muscle
occurrence
importance
35% 25% 15% 5% 0 0.05 0.25 0.45 0.65 0.85 1.05
RSRFC4
MEF2
MEF3
SRF
PPARA
HAND1E47
ERR1
HMEF2
LBP1
MYOGENIN

B)

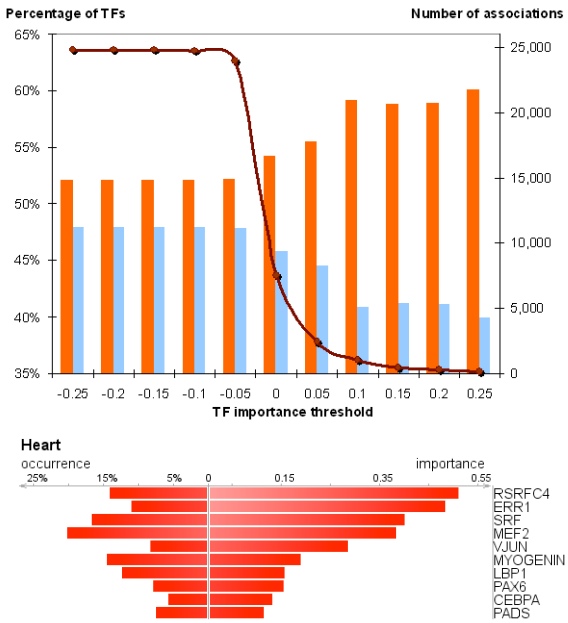

Percentage of TFs
Number of associations
65%
60%
55%
50%
45%
40%
35%
25,000
20,000
15,000
10,000
5,000
0
-0.25 -0.2 -0.15 -0.1 -0.05 0 0.05 0.1 0.15 0.2 0.25
TF importance threshold
Heart
occurrence
importance
25% 15% 5% 0 0.15 0.35 0.55
RSRFC4
ERR1
SRF
MEF2
VJUN
MYOGENIN
LBP1
PAX6
CEBPA
PADS

Table 1. Separation of human tissues into three precision groups. Tissues in a group list are sorted by the average *EI* precision so that tissues with the highest precision are listed lastly.

Precision group	Tissues
poor	pituitary gland, thymus, testis Leydig cell, cardiac myocytes, adrenal gland, salivary gland, testis, leukemia chronic myelogenous (k562), hypothalamus, testis germ cell, bronchial epithelial cells, pancreatic islets, olfactory bulb, fetal thyroid, dorsal root ganglion, atrioventricular node, appendix, testis interstitial, thyroid, ovary, skin, testis seminiferous tubule, skeletal muscle
good	prostate, smooth muscle, thyroid, amygdala, medulla oblongata, cingulate cortex, uterus corpus, prefrontal cortex, BM-CD71+ early erythroid, spinal cord, pancreas, BM-CD105+ endothelial, PB-CD4+ Tcells, lymphoma Burkitts Daudi, adipocyte, ciliary ganglion, subthalamic nucleus, trigeminal ganglion, adrenal cortex, whole brain, caudate nucleus, colorectal adenocarcinoma, leukemia promyelocytic (hl60), pons, globus pallidus, cerebellum peduncles, fetal brain, temporal lobe, PB-CD19+ Bcells, uterus, superior cervical ganglion, lymphoma Burkitts Raji, occipital lobe, PB-CD8+ Tcells, placenta, tonsil, thalamus, trachea, BM-CD34+, fetal liver, fetal lung, parietal lobe, kidney, PB-BDCA4+ dentritic cells, PB-CD14+ monocytes
excellent	PB-CD56+ NKCells, tongue, BM-CD33+ myeloid, cerebellum, bone marrow, lung, leukemia lymphoblastic (molt4), 721 B lymphoblasts, heart, whole blood, lymph node, liver

References

1. Altshuler, D. et al. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).
2. Lindblad-Toh, K. et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* (2005).
3. Sharan, R., Ben-Hur, A., Loots, G.G. & Ovcharenko, I. CREME: Cis-Regulatory Module Explorer for the human genome. *Nucleic Acids Res* **32**, W253-256 (2004).
4. Kim, T.H. et al. A high-resolution map of active promoters in the human genome. *Nature* **436**, 876-880 (2005).
5. Bajic, V.B., Tan, S.L., Suzuki, Y. & Sugano, S. Promoter prediction analysis on the whole human genome. *Nat Biotechnol* **22**, 1467-1473 (2004).
6. Xie, X. et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338-345 (2005).
7. Cawley, S. et al. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499-509 (2004).
8. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147-151 (2003).
9. Nobrega, M.A., Ovcharenko, I., Afzal, V. & Rubin, E.M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).
10. Lettice, L.A. et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**, 1725-1735 (2003).
11. de la Calle-Mustienes, E. et al. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res* **15**, 1061-1072 (2005).
12. Woolfe, A. et al. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**, e7 (2005).
13. Loots, G.G. et al. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136-140 (2000).
14. Dermitzakis, E.T., Reymond, A. & Antonarakis, S.E. Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat Rev Genet* **6**, 151-157 (2005).
15. Emison, E.S. et al. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature* **434**, 857-863 (2005).
16. Su, A.I. et al. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* **99**, 4465-4470 (2002).
17. Uchikawa, M., Takemoto, T., Kamachi, Y. & Kondoh, H. Efficient identification of regulatory sequences in the chicken genome by a powerful combination of embryo electroporation and genome comparison. *Mech Dev* **121**, 1145-1158 (2004).
18. Ovcharenko, I., Stubbs, L. & Loots, G.G. Interpreting mammalian evolution using Fugu genome comparisons. *Genomics* **84**, 890-895 (2004).
19. Prabhakar, S. et al. Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res* (2006).

20. Waterston, R.H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).
21. Loots, G.G. & Ovcharenko, I. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res* **32**, W217-221 (2004).
22. Riccio, A. et al. A nitric oxide signaling pathway controls CREB-mediated gene expression in neurons. *Mol Cell* **21**, 283-294 (2006).
23. Shalizi, A.K. & Bonni, A. Brawn for Brains: The Role of MEF2 Proteins in the Developing Nervous System. *Curr Top Dev Biol* **69**, 239-266 (2005).
24. Chang, W.T., Chen, H.I., Chiou, R.J., Chen, C.Y. & Huang, A.M. A novel function of transcription factor alpha-Pal/NRF-1: increasing neurite outgrowth. *Biochem Biophys Res Commun* **334**, 199-206 (2005).
25. Ilia, M., Sugiyama, Y. & Price, J. Gender and age related expression of Oct-6--a POU III domain transcription factor, in the adult mouse brain. *Neurosci Lett* **344**, 138-140 (2003).
26. Okuda, T. et al. Oct-3/4 repression accelerates differentiation of neural progenitor cells in vitro and in vivo. *Brain Res Mol Brain Res* **132**, 18-30 (2004).
27. Pennacchio, L.A. et al. *Nature* (submitted).
28. Frazer, K.A. et al. Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res* **14**, 367-372 (2004).
29. Itabashi, T. et al. Novel 615delC mutation in the CRX gene in a Japanese family with cone-rod dystrophy. *Am J Ophthalmol* **138**, 876-877 (2004).
30. Christensen, T.H., Prentice, H., Gahlmann, R. & Kedes, L. Regulation of the human cardiac/slow-twitch troponin C gene by multiple, cooperative, cell-type-specific, and MyoD-responsive elements. *Mol Cell Biol* **13**, 6752-6765 (1993).
31. Parmacek, M.S. et al. A novel myogenic regulatory circuit controls slow/cardiac troponin C gene transcription in skeletal muscle. *Mol Cell Biol* **14**, 1870-1885 (1994).
32. Novak, E.M., Dantas, K.C., Charbel, C.E. & Bydlowski, S.P. Association of hepatic nuclear factor-4 in the apolipoprotein B promoter: a preliminary report. *Braz J Med Biol Res* **31**, 1405-1408 (1998).
33. Parlakian, A. et al. Temporally controlled onset of dilated cardiomyopathy through disruption of the SRF gene in adult heart. *Circulation* **112**, 2930-2939 (2005).
34. Sakuma, K. et al. Serum response factor plays an important role in the mechanically overloaded plantaris muscle of rats. *Histochem Cell Biol* **119**, 149-160 (2003).
35. Kadi, F., Johansson, F., Johansson, R., Sjostrom, M. & Henriksson, J. Effects of one bout of endurance exercise on the expression of myogenin in human quadriceps muscle. *Histochem Cell Biol* **121**, 329-334 (2004).
36. Huss, J.M., Torra, I.P., Staels, B., Giguere, V. & Kelly, D.P. Estrogen-related receptor alpha directs peroxisome proliferator-activated receptor alpha signaling in the transcriptional control of energy metabolism in cardiac and skeletal muscle. *Mol Cell Biol* **24**, 9079-9091 (2004).
37. Shih, D.Q. et al. Hepatocyte nuclear factor-1alpha is an essential regulator of bile acid and plasma cholesterol metabolism. *Nat Genet* **27**, 375-382 (2001).
38. Zannis, V.I., Kan, H.Y., Kritis, A., Zanni, E. & Kardassis, D. Transcriptional regulation of the human apolipoprotein genes. *Front Biosci* **6**, D456-504 (2001).

39. Cheng, W. et al. HNF factors form a network to regulate liver-enriched genes in zebrafish. *Dev Biol* (2006).
40. Hallikas, O. et al. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47-59 (2006).
41. Sun, Q. et al. Defining the mammalian CArGome. *Genome Res* **16**, 197-207 (2006).
42. Thompson, W., Palumbo, M.J., Wasserman, W.W., Liu, J.S. & Lawrence, C.E. Decoding human regulatory circuits. *Genome Res* **14**, 1967-1974 (2004).
43. Sharan, R., Ovcharenko, I., Ben-Hur, A. & Karp, R.M. CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics* **19 Suppl 1**, i283-291 (2003).
44. Das, D., Nahle, Z. & Zhang, M.Q. Adaptively inferring human transcriptional subnetworks. *Mol Syst Biol* **2**, 2006 0029 (2006).
45. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).
46. Harris, M.A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**, D258-261 (2004).
47. Wingender, E. et al. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* **28**, 316-319 (2000).
48. Kent, W.J. et al. The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).
49. Karolchik, D. et al. The UCSC Genome Browser Database. *Nucleic Acids Res* **31**, 51-54 (2003).
50. Ovcharenko, I., Nobrega, M.A., Loots, G.G. & Stubbs, L. ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res* **32**, W280-286 (2004).
51. Ovcharenko, I. et al. Mulan: multiple-sequence local alignment and visualization for studying function and evolution. *Genome Res* **15**, 184-194 (2005).
52. Cartharius, K. et al. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics* **21**, 2933-2942 (2005).