LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-216861

Genome-wide de Novo Prediction of Proximal and Distal Tissue-Specific Enhancers

G. G. Loots, I. V. Ovcharenko

November 4, 2005

Genome Research

Genome-wide *de Novo* Prediction of Proximal and Distal Tissue-Specific Enhancers

Summary sentence: We introduce a novel method for deciphering the transcriptional regulatory code in eukaryotic genomes and describe its utility in predicting tissue specific enhancers, in humans.

Gabriela G. Loots[1] and Ivan Ovcharenko[2,3]

[1]Genome Biology and [2]EEBI Divisions, Lawrence Livermore National Laboratory, 7000 East Avenue L-441, Livermore, CA. 94550

[3]To whom correspondence should be addressed, tel. 925.422.5035; fax 925.422.2099; email, ovcharenko1@llnl.gov

Determining how transcriptional regulatory networks are encoded in the human genome is essential for understanding how cellular processes are directed. Here, we present a novel approach for systematically predicting tissue specific regulatory elements (REs) that blends genome-wide expression profiling, vertebrate genome comparisons, and pattern analysis of transcription factor binding sites. This analysis yields 4,670 candidate REs in the human genome with distinct tissue specificities, the majority of which reside far away from transcription start sites. We identify key transcription factors (TFs) for 34 distinct tissues and demonstrate that tissue-specific gene expression relies on multiple regulatory pathways employing similar, but different cohorts of interacting TFs. The methods and results we describe provide a global view of tissue specific gene regulation in humans, and propose a strategy for deciphering the transcriptional regulatory code in eukaryotes.

Increasing lines of evidence point to the observation that the majority of functional elements in the human genome are noncoding in nature, yet our ability to systematically predict them remains limited. Most progress in elucidating transcriptional regulatory mechanisms has stemmed from computational and experimental analyses of transcription factors acting at promoter regions of functionally related cohorts of genes. Whereas informative (*1-3*), studies restricted to promoter-specific regulation sample a small portion of the regulatory network in a vertebrate genome and overlook contributions by distant regulatory elements (RE) (*4, 5*). Several recent studies have provided conclusive evidence that the complex transcriptional expression patterns of human genes is mediated

through multiple unique sequences located hundred of kilobases (kb) away from transcription start sites (*6, 7*). In these studies, evolutionary sequence conservation has served as a reliable indicator of biological activity, and the majority of distant noncoding evolutionary conserved regions (ncECRs) examined experimentally in vertebrates, function as tissue-specific enhancers (*6, 8-10*), some of which are linked to human disorders (*11, 12*). In addition, ultraconserved and core ECRs have been shown to control the basal gene regulatory activity during key aspects of vertebrate development (*9, 13, 14*). Although genome comparisons have provided a powerful approach for identifying evolutionary conserved elements that are under selective pressure, we have yet to develop reliable high-throughput computational methods for the discovery of distant REs with required functional specificity. Here we introduce a new strategy for converting noncoding sequence data into transcriptional regulatory information which serves two vital purposes: (1) to define combinatorial arrays of regulatory motifs associated with tissue-specific gene expression, and (2) to predict tissue-specific distant enhancers in the human genome, *de novo*. This approach combines genome-wide tissue-specific gene expression profiling (*15*), vertebrate genome comparisons, and pattern analysis of transcription factor binding sites (TFBS).

For this analysis, the human transcriptome was first parsed into expression groups based on the site and level of gene expression. Thirty four principal tissue specific groups were constructed, each including 60 to 250 genes, where each gene was expressed at a minimum of 5-fold above the average expression level in that tissue (*16*); we refer to these genes as *overexpressors*. Next, we defined the genomic boundaries of each

overexpressor by associating each intergenic element with the nearest transcript. For several tissues we observed that the length of the genomic interval and the number of associated tissue-specific ncECRs are highly dependent on the tissue-specificity of gene expression (Figure 1). The median intergenic size was determined to be 63kb in length, with a median number of 39 ncECRs/locus. Loci harboring heart and liver specific genes were on average twice as short, and contained a significantly smaller number of ncECRs. In contrast, loci of genes highly expressed in the fetal brain, uterus, spinal cord, and kidney were significantly longer than the median and/or were enriched in ncECRs. Several other expression groups had either elevated (fetal liver, skin, smooth muscle, adipocyte, thyroid, and kidney) or reduced (lung and whole blood) densities of ncECRs or loci lengths. A bias towards large loci abundant in ncECRs was observed in embryonic and development-related tissues supporting the hypothesis that developmental genes benefit from complex transcriptional regulation established through an interplay of multiple distant gene regulatory elements (*17*).

Based on the rationale that highly conserved ncECRs participate in controlling gene expression (*6, 18*), for each transcript in the human genome, we selected the top 5 most highly conserved human-mouse ncECRs as candidate transcriptional regulatory elements (cREs) to be subjected to sequence pattern analysis (*19*). This set of cREs was augmented with an additional, less conserved group of ncECRs overlapping promoter regions (*16, 20*), to construct a collection of the most probable proximal and distal cREs. Over 103k human ncECRs were selected by this approach with an average density of 5.7 ncECRs per gene locus. TFBS mapping and conservation-based filtering (*21*) across

these elements identified 2.3 million (M) evolutionarily conserved TFBS or cTFBS (*22*). For 15 different tissues, TFBS analysis revealed a significant enrichment in cTFBS of transcription factors (TF) known to drive tissue-specific expression (Table S2), thus computationally confirming that regulatory proteins known to be active in a particular tissue leave a footprint of overrepresented binding sites in genomic loci of highly expressed gene targets. For example, we find MEF2 and SRF cTFBS to be enriched in cREs associated with overexpressors in skeletal muscle (*23, 24*), while HEN1 TFBS (neuronal stem cell leukemia TF) were overabundant in loci of genes expressed in the nervous system (*25, 26*), consistent with experimental data. These results indirectly confirm that the enrichment of noncoding functional elements matches the tissue specificity of gene expression and thus support the computational methods of cRE selection. We find that footprints of individual TF are unreliable markers of tissue-specificity due to low signal to noise ratio. For example, while 30% of overexpressors in the skeletal muscle expression group contain at least one MEF2 cTFBS, the same is observed in 13% of loci that are neutral in this tissue (Table S2).

Gene regulation in vertebrates is believed to be established through an interplay of multiple TF that bind to RE in combinatorial fashion (*27*), and do not simply represent an array of repeated TFBS, commonly found in *Drosophila* (*28*). To dissect the function of long-range vertebrate REs we performed combinatorial cTFBS cluster analysis adapted from methods used in vertebrate promoter studies (*1, 29*). Statistical analysis of TF pairs of cTFBS enriched in expression groups did not detect a noticeable improvement over individual TF enrichment analysis. For example, only two TF pairs, ‘MAZR+SRF’ and

'SP1+SRF' had enriched densities in skeletal muscle loci. While the latter pair has been shown to synergistically activate skeletal muscle genes (*30, 31*), these two pairs of TFs combined are overrepresented in less than 13% of ECRs linked to skeletal muscle overexpressor genes.

To overcome this problem, we developed a new combinatorial analysis strategy that simultaneously scores the impact of multiple TFs on gene regulation in a particular tissue. This is accomplished by assigning a weight to each TF to measure its regulatory potential to contribute to tissue-specific transcription. The sum of individual TF weights assigned to the cTFBS profile of a ncECR generates a score for each ncECR and determines its regulatory potential as a tissue-specific enhancer [see (*16*) for details]. TF weights were optimized independently in each different expression group to maximize the enrichment of positively scoring ncECRs in loci of overexpressed genes and to minimize their presence in neutral loci. Several types of regulatory information were generated using this method of regulatory element decoding optimization or *RED optimization*. We were able to (1) catalogue putative tissue-specific REs, (2) identify candidate proximal and distant REs for most overexpressor gene loci in the human genome, (3) determine the functional impact of each individual TF on tissue-specificity, (4) identify sets of TFs that cooperatively function in different tissues, and (5) construct several different gene regulatory pathways describing a single tissue-specific gene regulatory mechanism. The *RED optimization* method allowed us to distinguish the tissue specificity of more than 70% overexpressor gene loci according to their noncoding sequence, while detecting less than15% neutral loci (Figure 2 and Table S3). In summary, the genomic analysis of 4,188

human genes overexpressed in adult tissues identified 4,670 tissue-specific cREs corresponding to 3,559 unique loci. On average, 33% of these cREs overlapped promoter regions, 10% - first introns, 2% - 5'UTRs, 7% - 3'UTRs, 23% - other introns, while the remaining 25% were intergenic in nature (Table S4); estimating that ~55% of the human transcriptome relies on REs distal to the transcriptional start site - consistent with previous estimates (*4, 5*). We found ~20% of genes to harbor more than one cRE associated with a particular tissue specificity, suggesting that certain tissues potentially require multiple redundant and/or functionally additive regulatory elements; as previously proposed in evolutionary studies for the SIM2 locus (*32*).

Thirty percent of the studied loci (1,253) harbor genes highly expressed in more than one tissue. By performing *RED optimization* for these loci independently in each tissue we were able to quantify the ratio of cREs with multiple tissue-specificities by dissecting cases of multi-tissue activation that depend either on a single or on multiple cREs. In general, we observed that individual ncECR assigned to multiple different tissues were mainly detected in cell-types that are either functionally related or spatially congruent (Fig. 3). Sixteen percent of ncECRs (1,986 of 12,147) distributed across 1,105 of these loci were classified as tissue-specific cREs, 57% (1,146) of which were assigned by the *RED optimization* method to more than one tissue category (p-value$<10^{-8}$), through independent analysis of each expression group. Our results suggest that transcriptional regulation in tissues that are functionally or spatially interconnected is often achieved through shared REs responsible for a specific yet all-encompassing pattern of expression, rather than through multiple tissue-specific restrictive REs. Our finding are in accord

with similar broad expression patterns determined in *in vivo* experiments of candidate enhancer elements identified by REs scans of intergenic regions of evolutionary deeply conserved genes (*6, 8, 9*).

Currently, no extensive and/or centralized tissue-specific database of characterized vertebrate enhancers exists that would allow one to directly assess the performance of new computational methods. While a handful of distant (or long-range) cREs have been tested *in vivo* in several model organisms *(9, 25-34)*, the majority of available validated enhancers have been selected from a subset of highly conserved gene loci, known as *trans-dev* genes. In the adult, *trans-dev* genes commonly have a broad and sometimes ubiquitous expression pattern, as revealed by microarray expression data (*15*) and thus were not classified as overexpressors in a particular tissue in the current study. From a list of five *trans-dev* genes examined *in vivo* for the presence of tissues-specific enhancers in vertebrate embryos (*9*), only PAX6 and HLXB9 were highly expressed in an adult tissue, the pancreatic islets. *In vivo* analysis of these genes did not address their activity in pancreatic islets, thus preventing us from comparing their enhancers to predictions derived from *RED optimization* analysis. Nonetheless several published examples stand out. Human cardiac/slow skeletal muscle troponin C (TNNC1) is associated with cardiomyopathies (*33, 34*), and was found to be one of the 10 most highly expressed genes in skeletal muscles, in this study. Its regulation in skeletal muscle has been examined in great detail, and a critical RE has been identified in the first intron (*35, 36*). *RED optimization* identified three cREs in the TNNC1 locus (Fig. 4), one of which corresponds with the previously characterized skeletal muscle RE of this gene (*35, 36*)

and contains a pair of MEF2 sites. GNF Expression profiling of TNNC1 indicates that it is also highly expressed in two other muscle types – heart and tongue. Our method predicted a heart-specific cRE in the second intron and a tongue-specific one in TNNC1's promoter region (Fig. 4). These predictions do not overlap with the skeletal muscle cRE, despite the similar nature of these muscle types.

Similarly to TNNC1, an element in the promoter region of atrial natriuretic factor (ANF), a gene associated with the Holt-Oram syndrome, has been shown to activate expression during cardiac development (*37, 38*). Also, a 3'UTR element has been recently shown to play a regulatory role in modeling the ANF heart expression through NRSF-dependent repression in ventricular myocytes (*39, 40*). GNF Expression profiles depicted high ANF expression in the adult heart tissue, thus permitting us to include this gene in the analysis. *RED optimization* correctly identified the 3'UTR RE of the ANF but not the promoter element, possibly due to differences in transcriptional regulation during embryonic patterning and the adult heart. Likewise, *RED optimization* correctly identified the ApoB promoter element as a fetal liver cRE and predicted HNF4 and C/EBP binding to activate ApoB expression, in concordance with previous findings (*41*). *RED optimization* is therefore an efficient approach for deciphering the location of tissue-specific cREs in these known cases and can easily be applied on a global scale to identify and prioritize candidate tissue-specific enhancers.

Analysis of the distribution of positively scoring TFs across predicted tissue-specific cREs revealed that a single TF has a limited impact on the tissue-specificity of gene

expression (Table S5), supporting the hypothesis that tissue-specific gene regulation is a direct result of an elaborate interplay among multiple TFs. In our analysis, OCT1 TF solely stood out as the exception. It exhibited 30% occupancy in the fetal brain expression group, while the majority of all other tissues displayed a strong dependency on cohorts of multiple TFs (Fig. 5; Table S5). To quantify the impact of an individual *i*-th TF on regulating gene expression in a particular tissue *t* we introduced the parameter I_i^t defined as the product of TF occurrences and its weight, in a tissue-specific group of cREs (*16*). Interestingly, OCT1 groups with nine other TFs determined to have a high impact across multiple tissues, such as spinal cord, skeletal muscle, and T cells, to name a few. This suggests a key functional and basal position for OCT1 in vertebrate genome regulation. Indeed, OCT1 is a member of the POU family of transcription factors known to regulate many different processes in mammalian development, activating multiple transcriptional regulatory networks and thus residing upstream of many regulatory pathways. Non-intuitively OCT1 knockout mice are viable (*42*), suggesting functional redundancy among different members of the POU family. General OCT cTFBS occurrences ranked second to OCT1 in several tissues, including fetal brain and spinal cord, further strengthening this hypothesis.

In fetal liver, a different TF stands out in terms of its potential regulatory impact, the hepatic nuclear factor (HNF1). HNF1 is known to be strongly associated with multiple hepatic abnormalities including type 2 diabetes, dwarfism, renal Fanconi syndrome, hepatic dysfunction and hypercholesterolemia (*43*). In liver, it has been shown to cooperate with another member of its family, HNF4, and several other hepatic related

TFs including peroxisome proliferator activated receptor (PPAR) and nuclear transcription factor-Y (NFY) to stimulate transcription. Consistent with our expectation, these TFs showed a high impact on adult liver regulation (Fig. 5). Also, in the kidney we found strong support for HNFs partnering with GATAs and COUPs to co-activate transcription in this organ. Analysis of cREs of genes overexpressed in skeletal muscle immediately pinpointed to the well characterized MEF2 TF as a key regulator of gene expression in skeletal muscle, as well as highlighted several other skeletal muscle related TFs such as serum response factor (SRF), MyoD, and muscle-specific TATA-box (MTATA). Regulatory profile analysis in T- and B-cells generated similar results, where we determined that TF containing ETS domains such as ELK-1 play an important regulatory role (*44, 45*).

One of the most critical component in these methods that allowed us to achieve a high degree of separation between tissue-specific and neutral loci (Fig. 2) has been the binning of tissue-specific regulation into separate gene regulatory pathways (achieved by attributing different TF weights to each pathway). From 2 to 7 pathways were assigned to each tissue specificity (*16*). For example, in the liver, we observed five distinct gene regulatory pathways with a different interplay of TFs cohorts constituting each pathway (Fig. 6). As previously mentioned, both HNF1 and HNF4 are actively involved in regulating liver-specific genes and this analysis demonstrates some of the details of their versatility. While HNF4 is the only TF shared by all five pathways, HNF1 segregates uniquely to the pathway that encompasses the largest number of liver overexpressors (46%), and therefore has the highest impact on this pathway. Five other TFs (or TF

dimers, such as HNF4:DR1 complex) contribute to more than 2 pathways and 18 others are specific to only one of the pathways suggesting that liver specific transcriptional regulation can not be described universally by a single cohort of interacting TFs, but rather employs several distinguishable regulatory networks.

Cross-species conservation analysis of predicted cREs detects rapid evolution of REs across vertebrates consistent with previous studies (*46*). Ten percent of cREs are conserved in chicken, 5.5% in frog, and 3% in fish (Fugu and/or zebrafish). However, we detected distinct differences in the rates of evolutionary change in several tissues (*16*) (Fig. 7). Only 2% of trachea specific cREs were conserved in chicken, and none in frog or fish, while cREs associated with genes overexpressed in the uterus ranked more than 50% above the average in the number of cRE for each species. Conservation of smooth and skeletal muscle cREs displayed an interesting evolutionary profile. While these elements were relatively similarly conserved in chicken, far fewer smooth muscle cREs were conserved in frog and/or fish compared to those in skeletal muscle. Phenotypic adaptations of amphibians and fish may provide an explanation for these computational observations. At a different extreme, in the pancreas and pancreatic islets we observed a 3-fold decrease in the percentage of cREs conserved in chicken, while intact values were recorded for frog and fish (Fig 7).

Our study provides an initial systematic whole-genome analysis of the noncoding segment of the human genome to computationally predict the location and tissue-specificity of proximal and distal REs. While similar analyses have been carried out for

promoter sequences, this is the first attempt to comprehensively include potential regulatory sequences that act at a distance. The approach we developed combines tissue-specific gene expression profiling (*15*), genome comparisons and combinatorial TFBS pattern analysis to predict and catalogue putative tissue-specific distant enhancers in the human genome. Most importantly, our analysis was able to evaluate the potential functional contribution of each individual TF to tissue-specific gene activation, recapitulating known interactions, and describing novel relationships between TFs. We also present a new method of quantifying combinatorial interactions that allows us to distinguish between different cooperative interactions of the same TF, in different tissues. This method can be applied to the analysis of any set of co-expressed genes; thus providing a rapid and efficient approach for translating microarray expression data into process- and/or tissue-specific gene regulatory principles. The results presented here contribute to the ongoing efforts of identifying and cataloguing all functional elements in the human genome to create a foundation for computationally characterizing gene-regulatory sequences and elucidating gene-regulatory networks.

This work was performed under the auspices of the U. S. Department of Energy by University of California, Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

References and notes.

1. R. Sharan, A. Ben-Hur, G. G. Loots, I. Ovcharenko, *Nucleic Acids Res* **32**, W253-6 (Jul 1, 2004).
2. T. H. Kim *et al.*, *Nature* **436**, 876-80 (Aug 11, 2005).
3. V. B. Bajic, S. L. Tan, Y. Suzuki, S. Sugano, *Nat Biotechnol* **22**, 1467-73 (Nov, 2004).
4. S. Cawley *et al.*, *Cell* **116**, 499-509 (Feb 20, 2004).
5. M. Levine, R. Tjian, *Nature* **424**, 147-51 (Jul 10, 2003).
6. M. A. Nobrega, I. Ovcharenko, V. Afzal, E. M. Rubin, *Science* **302**, 413 (Oct 17, 2003).
7. L. A. Lettice *et al.*, *Hum Mol Genet* **12**, 1725-35 (Jul 15, 2003).
8. E. de la Calle-Mustienes *et al.*, *Genome Res* **15**, 1061-72 (Aug, 2005).
9. A. Woolfe *et al.*, *PLoS Biol* **3**, e7 (Jan, 2005).
10. G. G. Loots *et al.*, *Science* **288**, 136-40 (Apr 7, 2000).
11. E. T. Dermitzakis, A. Reymond, S. E. Antonarakis, *Nat Rev Genet* **6**, 151-7 (Feb, 2005).
12. E. S. Emison *et al.*, *Nature* **434**, 857-63 (Apr 14, 2005).
13. G. Bejerano *et al.*, *Science* **304**, 1321-5 (May 28, 2004).
14. A. Sandelin *et al.*, *BMC Genomics* **5**, 99 (Dec 21, 2004).
15. A. I. Su *et al.*, *Proc Natl Acad Sci U S A* **99**, 4465-70 (Apr 2, 2002).
16. *Supplementary Materials*.
17. C. E. Nelson, B. M. Hersh, S. B. Carroll, *Genome Biol* **5**, R25 (2004).
18. I. Ovcharenko, L. Stubbs, G. G. Loots, *Genomics* **84**, 890-5 (Nov, 2004).
19. *The top 5 noncoding ECRs with the largest number of identical nucleotides.*
20. *Promoters defined as 1.5kb upstream of each transcriptional start site.*
21. G. G. Loots, I. Ovcharenko, *Nucleic Acids Res* **32**, W217-21 (Jul 1, 2004).
22. *Human-mouse alignments (hg17-mm5) retrieved from the ECR Browser were processed by rVista 2.0 to identify cTFBS in the human genome. We utilized previously described optimized position weight matrix (PWM) thresholds that limit predictions to 3 TFBS/10kb of random sequence. 11.9M human cTFBS were identified using 486 TRANSFAC PWMs and manually curated PWMs for TBX5, NKX2.5, and GLI TFs; 92% (11M) overlapping with ECRs (≥70%/≥ 100bps).*
23. F. J. Naya, E. Olson, *Curr Opin Cell Biol* **11**, 683-8 (Dec, 1999).
24. S. Li *et al.*, *Proc Natl Acad Sci U S A* **102**, 1082-7 (Jan 25, 2005).
25. J. Bao, D. A. Talmage, L. W. Role, J. Gautier, *Development* **127**, 425-35 (Jan, 2000).
26. M. Kruger, T. Braun, *Mol Cell Biol* **22**, 792-800 (Feb, 2002).
27. W. W. Wasserman, A. Sandelin, *Nat Rev Genet* **5**, 276-87 (Apr, 2004).
28. B. P. Berman *et al.*, *Proc Natl Acad Sci U S A* **99**, 757-62 (Jan 22, 2002).
29. K. Cartharius *et al.*, *Bioinformatics* **21**, 2933-42 (Jul 1, 2005).
30. E. Biesiada, Y. Hamamori, L. Kedes, V. Sartorelli, *Mol Cell Biol* **19**, 2577-84 (Apr, 1999).
31. I. Irrcher, D. A. Hood, *J Appl Physiol* **97**, 2207-13 (Dec, 2004).

32. K. A. Frazer *et al.*, *Genome Res* **14**, 367-72 (Mar, 2004).
33. J. Erdmann *et al.*, *Clin Genet* **64**, 339-49 (Oct, 2003).
34. J. Mogensen *et al.*, *J Am Coll Cardiol* **44**, 2033-40 (Nov 16, 2004).
35. T. H. Christensen, H. Prentice, R. Gahlmann, L. Kedes, *Mol Cell Biol* **13**, 6752-65 (Nov, 1993).
36. M. S. Parmacek *et al.*, *Mol Cell Biol* **14**, 1870-85 (Mar, 1994).
37. B. G. Bruneau *et al.*, *Cell* **106**, 709-21 (Sep 21, 2001).
38. E. M. Small, P. A. Krieg, *Dev Biol* **261**, 116-31 (Sep 1, 2003).
39. K. Kuwahara *et al.*, *Mol Cell Biol* **21**, 2085-97 (Mar, 2001).
40. A. C. Houweling, M. M. van Borren, A. F. Moorman, V. M. Christoffels, *Cardiovasc Res* **67**, 583-93 (Sep 1, 2005).
41. E. M. Novak, K. C. Dantas, C. E. Charbel, S. P. Bydlowski, *Braz J Med Biol Res* **31**, 1405-8 (Nov, 1998).
42. J. W. Jonker, E. Wagenaar, S. Van Eijl, A. H. Schinkel, *Mol Cell Biol* **23**, 7902-8 (Nov, 2003).
43. D. Q. Shih *et al.*, *Nat Genet* **27**, 375-82 (Apr, 2001).
44. B. Knebel, H. Avci, C. Bullmann, J. Kotzka, D. Muller-Wieland, *Exp Clin Endocrinol Diabetes* **113**, 94-101 (Feb, 2005).
45. L. Y. Bourguignon, E. Gilad, K. Rothman, K. Peyrollier, *J Biol Chem* **280**, 11961-72 (Mar 25, 2005).
46. L. W. Hillier *et al.*, *Nature* **432**, 695-716 (Dec 9, 2004).

Figure legends.

Figure 1. Number of ncECRs (A) and locus length (B) of genes overexpressed in different tissues. Bounded horizontal lines represent the interquartile range (the distance between the 25th and 75th percentiles) of the tissue-specific distributions, solid colored rectangles measure the standard error in median calculations (white lined inside colored rectangles). Statistically significant (<5%) distributions that deviate from the median (represented by a solid vertical black line) are marked by an asterisk on the left side bar. Tissues where the median value is 2-fold smaller or larger than the median are marked by a vertical line on the left side bar. Tissues are color coded as follows: orange – brain; blue – nervous system; red – muscles; green – immune system; light blue – glands; yellow – testis; gray – other.

Figure 2. Percentage of loci identified by *RED optimization* after the TF weights optimization. Tissues sorted by the ratio of overexpressed to neutral percentage, which varied from 3.5 to 40.2 for trachea and fetal brain tissues, correspondingly. Dotted orange lines demarcate threshold of 15% and 70% utilized for the selection of the optimal number of gene regulatory pathways (*16*).

Figure 3. Venn diagrams representing overlaps in cREs from different tissues for 4 selected overlapping groups (A). Percentage of unique cREs and the tissue with the largest percentage of overlapping cREs (if larger than 5%) (B).

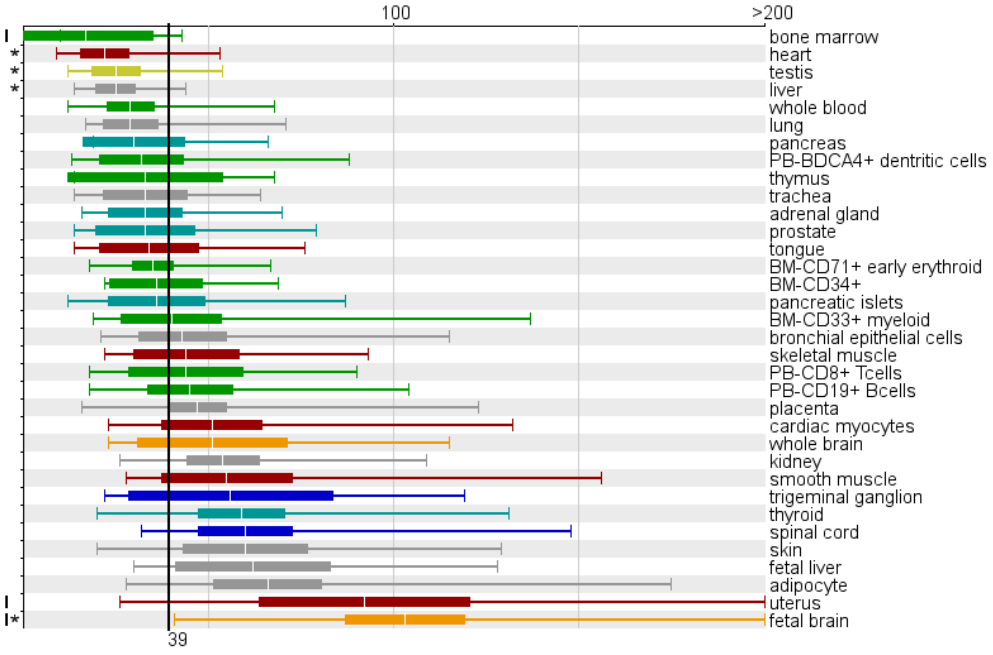
Figure 4. ECR Browser evolutionary conservation of the TNNC1 locus. ncECRs depicted as color coded regions (UTRs in yellow, introns in light red, intergenic in red) with gradient color bars above. Three tissue-specific cREs were detected in this locus, as marked.

Figure 5. Importance and occupancy of individual TFs in ten selected tissues (fetal brain, fetal liver, spinal cord, liver, thyroid, kidney, skeletal muscle, smooth muscle, T- and B-cells).

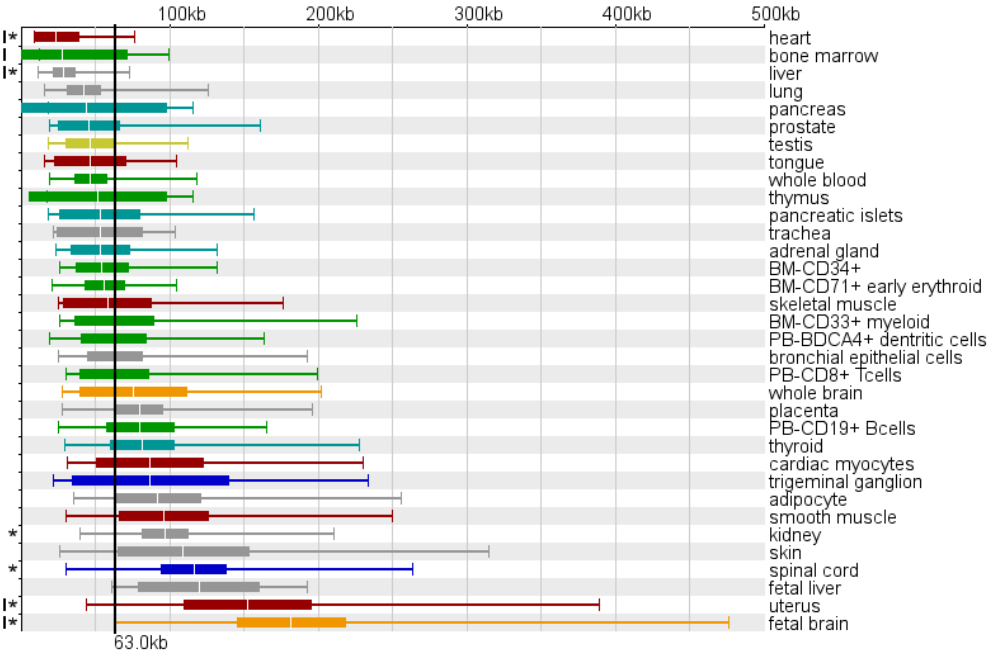
Figure 6. *RED optimization* identified five different regulatory pathways that describe liver specific gene regulation of overexpressor genes. We extracted the statistics of TF contributions to different pathways limiting to ten TFs with the largest occupancy per pathway and requiring a TF to be present at least twice and to contribute to at least 5% of the genes in a pathway. Gray boxes represent individual pathways with the count of genes given at the bottom of each box and a list of TFs that are unique to that pathway listed inside the box. TFs contributing to more than one pathway are listed outside of gray boxes with the lines connecting them to the pathways they contribute to.

Figure 7. Conservation of cREs in chicken, frog, and fish binned by tissue specificity and sorted by conservation in chicken.

Figure 1.


100
>200
bone marrow
heart
testis
liver
whole blood
lung
pancreas
PB-BDCA4+ dentritic cells
thymus
trachea
adrenal gland
prostate
tongue
BM-CD71+ early erythroid
BM-CD34+
pancreatic islets
BM-CD33+ myeloid
bronchial epithelial cells
skeletal muscle
PB-CD8+ Tcells
PB-CD19+ Bcells
placenta
cardiac myocytes
whole brain
kidney
smooth muscle
trigeminal ganglion
thyroid
spinal cord
skin
fetal liver
adipocyte
uterus
fetal brain
39

A)


100kb
200kb
300kb
400kb
500kb
heart
bone marrow
liver
lung
pancreas
prostate
testis
tongue
whole blood
thymus
pancreatic islets
trachea
adrenal gland
BM-CD34+
BM-CD71+ early erythroid
skeletal muscle
BM-CD33+ myeloid
PB-BDCA4+ dentritic cells
bronchial epithelial cells
PB-CD8+ Tcells
whole brain
placenta
PB-CD19+ Bcells
thyroid
cardiac myocytes
trigeminal ganglion
adipocyte
smooth muscle
kidney
skin
spinal cord
fetal liver
uterus
fetal brain
63.0kb

B)

Figure 2.

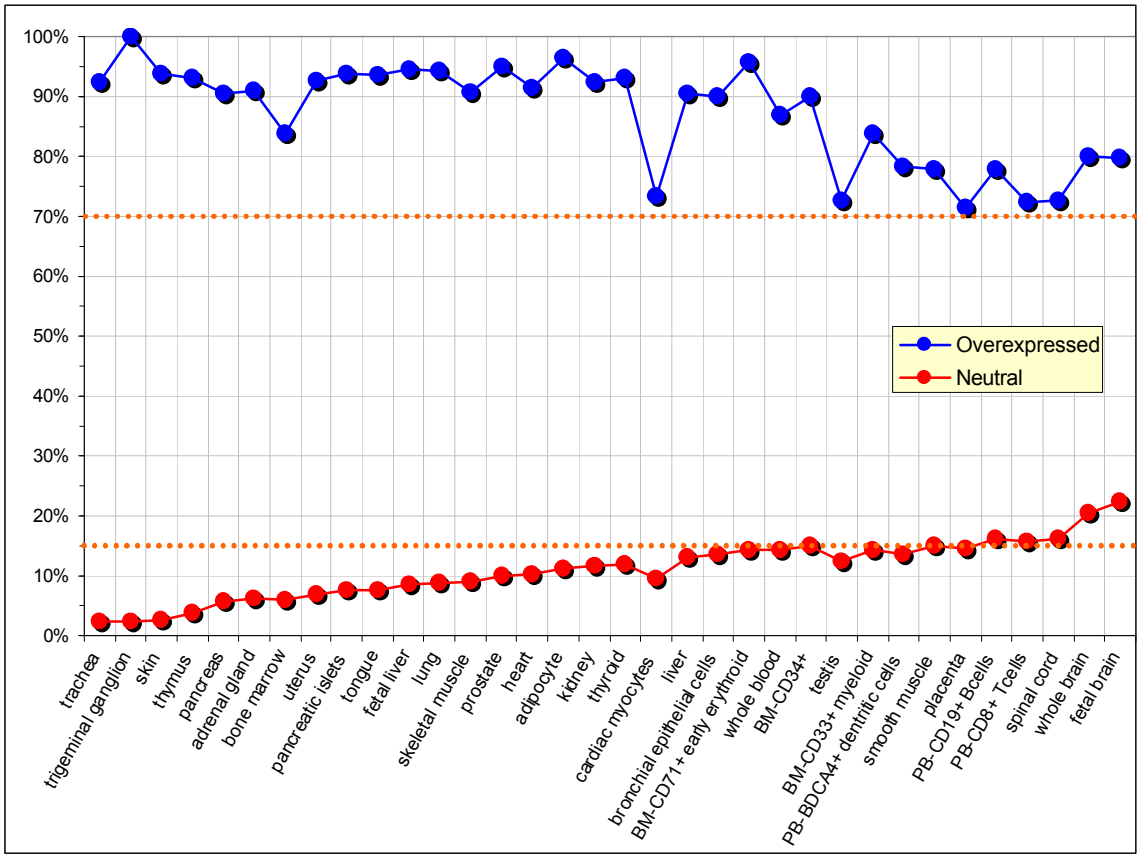
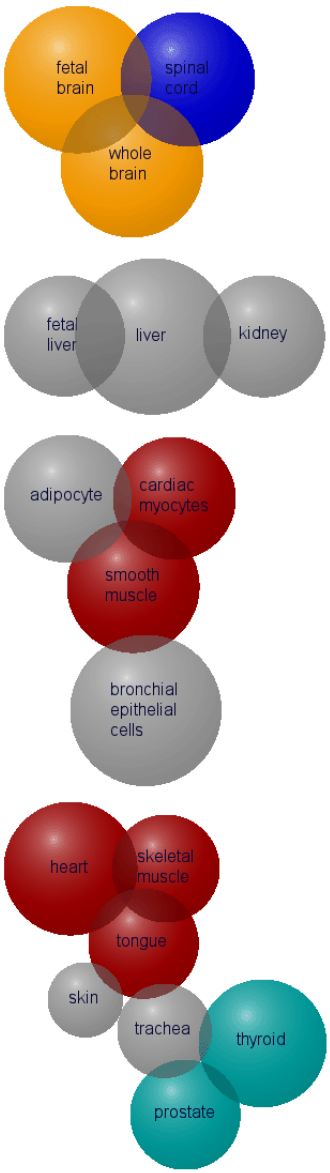

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Overexpressed
Neutral
trachea
trigeminal ganglion
skin
thymus
pancreas
adrenal gland
bone marrow
uterus
pancreatic islets
tongue
fetal liver
lung
skeletal muscle
prostate
heart
adipocyte
kidney
thyroid
cardiac myocytes
liver
bronchial epithelial cells
BM-CD71+ early erythroid
whole blood
BM-CD34+
testis
BM-CD33+ myeloid
PB-BDCA4+ dentritic cells
smooth muscle
placenta
PB-CD19+ Bcells
PB-CD8+ Tcells
spinal cord
whole brain
fetal brain

Figure 3.

A)



B)

Tissue	% of unique cREs	Tissue with the largest overlap in cREs; % of overlapping elements	
BM-CD33+ myeloid	46.2	whole blood	25.6
BM-CD34+	59.9	BM-CD71+ early erythroid	13.6
BM-CD71+ early erythroid	72.8	BM-CD34+	13.1
PB-BDCA4+ dentritic cells	45.2	BM-CD33+ myeloid	19.5
PB-CD19+ Bcells	54.7	PB-BDCA4+ dentritic cells	17.3
PB-CD8+ Tcells	59.1	whole blood	17.2
adipocyte	71.6	smooth muscle	7.4
adrenal gland	80.9	-	
bone marrow	51.8	whole blood	18.1
bronchial epithelial cells	76.0	-	
cardiac myocytes	66.0	smooth muscle	14.2
fetal brain	70.5	whole brain	17.9
fetal liver	53.2	liver	32.3
heart	77.1	skeletal muscle	8.9
kidney	75.0	liver	13.1
liver	60.1	fetal liver	19.4
lung	77.4	-	
pancreas	73.1	pancreatic islets	21.5
pancreatic islets	63.8	pancreas	15.7
placenta	79.5	-	
prostate	73.3	trachea	6.1
skeletal muscle	69.6	tongue	17.6
skin	80.3	tongue	8.2
smooth muscle	69.8	cardiac myocytes	12.2
spinal cord	64.9	whole brain	16.5
testis	89.8	-	
thymus	60.6	whole blood	15.5
thyroid	78.4	-	
tongue	61.8	skeletal muscle	16.8
trachea	66.7	prostate	8.3
trigeminal ganglion	90.1	-	
uterus	75.0	-	
whole blood	45.9	BM-CD33+ myeloid	24.8
whole brain	63.6	fetal brain	19.1

Figure 4.

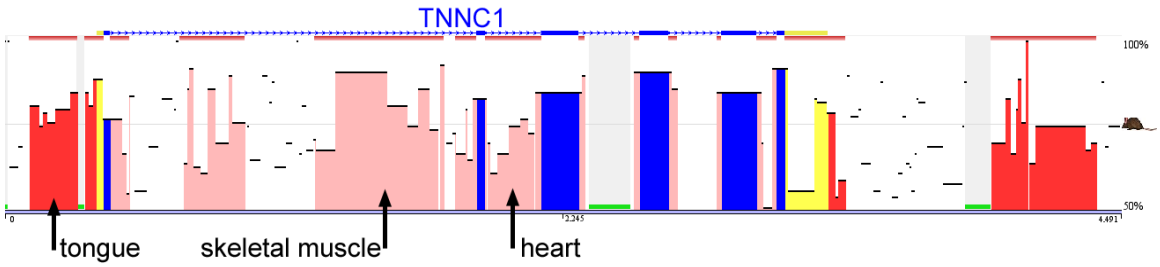


Figure 5.

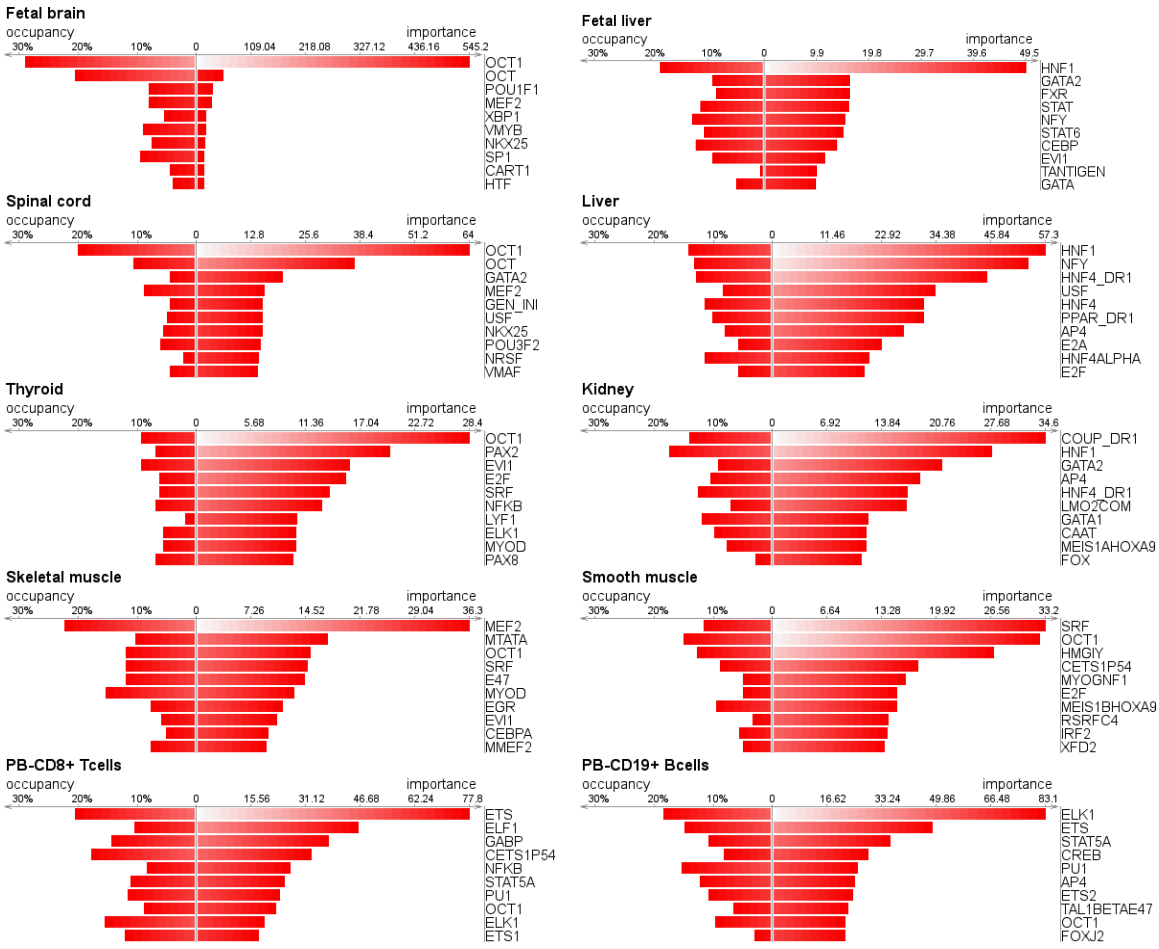

Fetal brain
occupancy
importance
30% 20% 10% 0 109.04 218.08 327.12 436.16 545.2
OCT1
OCT
POU1F1
MEF2
XBP1
VMYB
NKX25
SP1
CART1
HTF
Fetal liver
occupancy
importance
30% 20% 10% 0 9.9 19.8 29.7 39.6 49.5
HNF1
GATA2
FXR
STAT
NFY
STAT6
CEBP
EVI1
ITANTIGEN
GATA
Spinal cord
occupancy
importance
30% 20% 10% 0 12.8 25.6 38.4 51.2 64
OCT1
OCT
GATA2
MEF2
GEN_INI
USF
NKX25
POU3F2
NRSF
VMAF
Liver
occupancy
importance
30% 20% 10% 0 11.46 22.92 34.38 45.84 57.3
HNF1
NFY
HNF4_DR1
USF
HNF4
PPAR_DR1
AP4
E2A
HNF4ALPHA
E2F
Thyroid
occupancy
importance
30% 20% 10% 0 5.68 11.36 17.04 22.72 28.4
OCT1
PAX2
EVI1
E2F
SRF
NFKB
LYF1
ELK1
MYOD
PAX8
Kidney
occupancy
importance
30% 20% 10% 0 6.92 13.84 20.76 27.68 34.6
COUP_DR1
HNF1
GATA2
AP4
HNF4_DR1
LMO2COM
GATA1
CAAT
MEIS1AHOXA9
FOX
Skeletal muscle
occupancy
importance
30% 20% 10% 0 7.26 14.52 21.78 29.04 36.3
MEF2
MTATA
OCT1
SRF
E47
MYOD
EGR
EVI1
CEBPA
MMEF2
Smooth muscle
occupancy
importance
30% 20% 10% 0 6.64 13.28 19.92 26.56 33.2
SRF
OCT1
HMGIY
CETS1P54
MYOGNF1
E2F
MEIS1BHOXA9
RSRFC4
IRF2
XFD2
PB-CD8+ Tcells
occupancy
importance
30% 20% 10% 0 15.56 31.12 46.68 62.24 77.8
ETS
ELF1
GABP
CETS1P54
NFKB
STAT5A
PU1
OCT1
ELK1
ETS1
PB-CD19+ Bcells
occupancy
importance
30% 20% 10% 0 16.62 33.24 49.86 66.48 83.1
ELK1
ETS
STAT5A
CREB
PU1
AP4
ETS2
TAL1BETAE47
OCT1
FOXJ2

Figure 6.

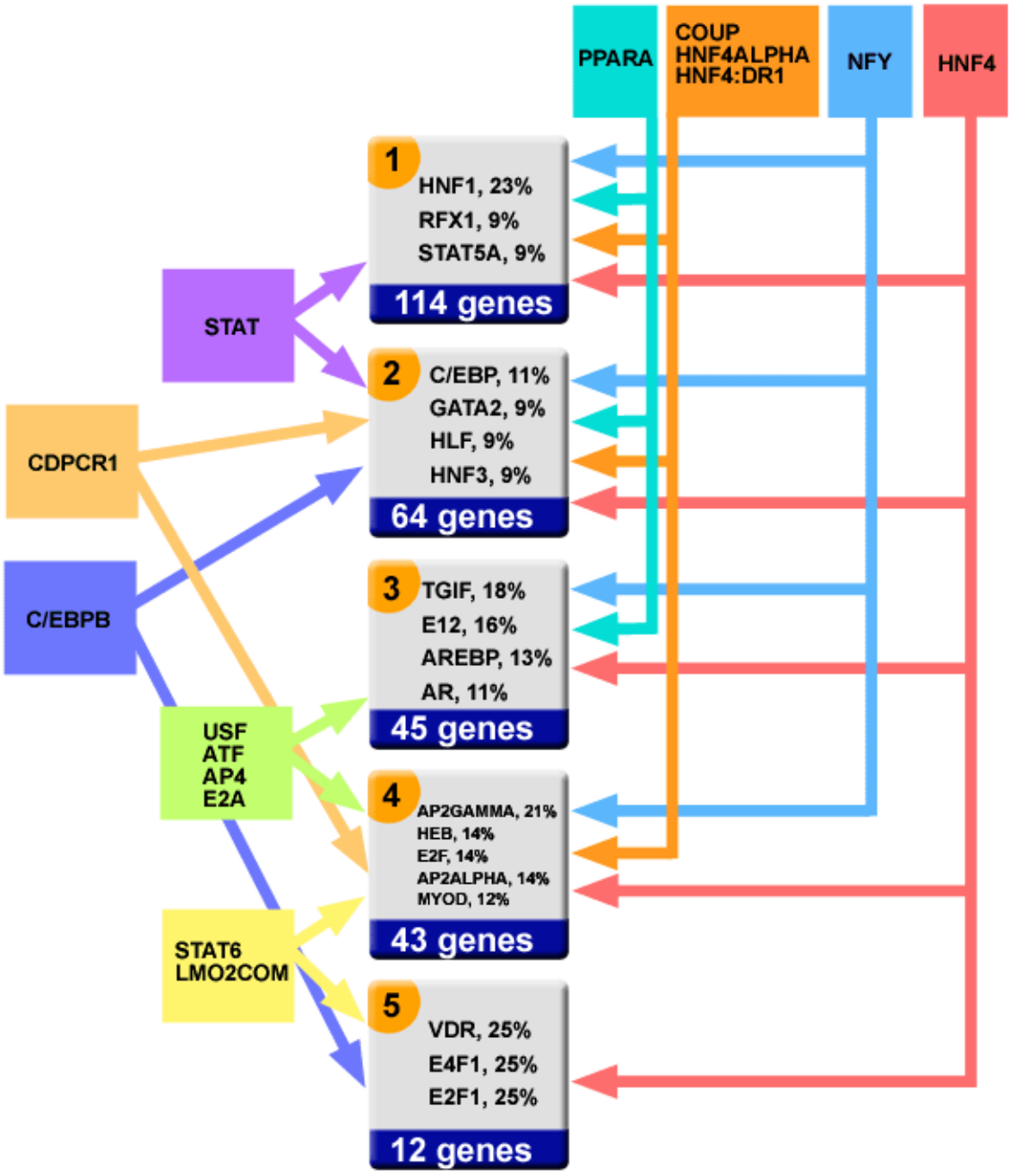
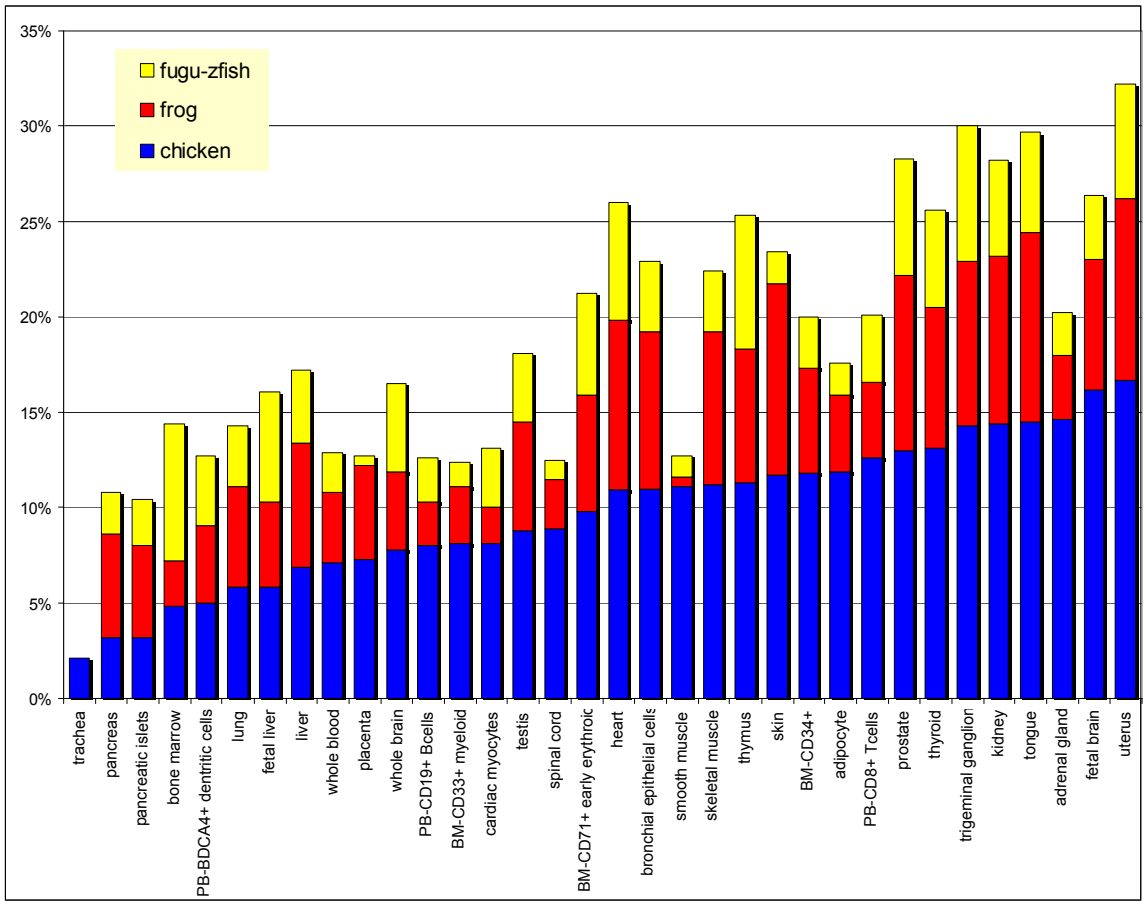

PPARA
COUP
HNF4ALPHA
HNF4:DR1
NFY
HNF4
1
HNF1, 23%
RFX1, 9%
STAT5A, 9%
114 genes
2
C/EBP, 11%
GATA2, 9%
HLF, 9%
HNF3, 9%
64 genes
3
TGIF, 18%
E12, 16%
AREBP, 13%
AR, 11%
45 genes
4
AP2GAMMA, 21%
HEB, 14%
E2F, 14%
AP2ALPHA, 14%
MYOD, 12%
43 genes
5
VDR, 25%
E4F1, 25%
E2F1, 25%
12 genes
STAT
CDPCR1
C/EBPB
USF
ATF
AP4
E2A
STAT6
LMO2COM

Figure 7.


35%
30%
25%
20%
15%
10%
5%
0%
fugu-zfish
frog
chicken
trachea
pancreas
pancreatic islets
bone marrow
PB-BDCA4+ dentritic cells
lung
fetal liver
liver
whole blood
placenta
whole brain
PB-CD19+ Bcells
BM-CD33+ myeloid
cardiac myocytes
testis
spinal cord
BM-CD71+ early erythroic
heart
bronchial epithelial cells
smooth muscle
skeletal muscle
thymus
skin
BM-CD34+
adipocyte
PB-CD8+ Tcells
prostate
thyroid
trigeminal ganglion
kidney
tongue
adrenal gland
fetal brain
uterus