OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

# Automated Discovery of Tissue-Targeting Enhancers and Transcription Factors from Binding Motif and Gene Function Data

**Geetu Tuteja[1], Karen Betancourt Moreira[1], Tisha Chung[1], Jenny Chen[2], Aaron M. Wenger[3], Gill Bejerano[1,3]***

**1** Department of Developmental Biology, Stanford University, Stanford, California, United States of America, **2** Biomedical Informatics Program, Stanford University, Stanford, California, United States of America, **3** Department of Computer Science, Stanford University, Stanford, California, United States of America

## Abstract

Identifying enhancers regulating gene expression remains an important and challenging task. While recent sequencing-based methods provide epigenomic characteristics that correlate well with enhancer activity, it remains onerous to comprehensively identify all enhancers across development. Here we introduce a computational framework to identify tissue-specific enhancers evolving under purifying selection. First, we incorporate high-confidence binding site predictions with target gene functional enrichment analysis to identify transcription factors (TFs) likely functioning in a particular context. We then search the genome for clusters of binding sites for these TFs, overcoming previous constraints associated with biased manual curation of TFs or enhancers. Applying our method to the placenta, we find 33 known and implicate 17 novel TFs in placental function, and discover 2,216 putative placenta enhancers. Using luciferase reporter assays, 31/36 (86%) tested candidates drive activity in placental cells. Our predictions agree well with recent epigenomic data in human and mouse, yet over half our loci, including 7/8 (87%) tested regions, are novel. Finally, we establish that our method is generalizable by applying it to 5 additional tissues: heart, pancreas, blood vessel, bone marrow, and liver.

**Citation:** Tuteja G, Moreira KB, Chung T, Chen J, Wenger AM, et al. (2014) Automated Discovery of Tissue-Targeting Enhancers and Transcription Factors from Binding Motif and Gene Function Data. PLoS Comput Biol 10(1): e1003449. doi:10.1371/journal.pcbi.1003449

**Editor:** Marcelo A. Nobrega, University of Chicago, United States of America

**Received** September 23, 2013; **Accepted** December 9, 2013; **Published** January 30, 2014



**Funding:** This work was supported by the A. P. Giannini Foundation Postdoctoral Research Fellowship to GT, Bio-X Stanford Interdisciplinary Graduate Fellowship to AMW, a BioX IIP award and a Burroughs Wellcome Preterm Disease Planning grant to GB. GB is a Packard Fellow and a Microsoft Faculty Fellow. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: bejerano@stanford.edu

## Introduction

Transcriptional regulation in mammals is a highly orchestrated process directed in part by the binding of sequence-specific transcription factors (TFs) to genomic regulatory elements, such as enhancers. Enhancers contain binding sites for sequence-specific TFs that recognize particular DNA motifs. The combined input of the multiple TFs that bind to a single enhancer region results in tissue- and time-point- specific gene activation [1]. Identification of active enhancers, particularly those enhancers that are most relevant to a developmental process, is a challenging task that is the subject of intense investigation.

The ENCODE project and Roadmap Epigenomics project have recently provided DNase I hypersensitive sites (DHSs), which can mark enhancers, promoters, silencers, insulators, and locus control regions in many human cell and tissue samples [2,3]. Additionally, the mouse ENCODE project has provided ChIP-Seq data for enhancer-associated chromatin marks in multiple mouse tissues and cell types [4]. While highly valuable, these data provide only indirect evidence of cis-regulatory activity. Chromatin must be open for most trans-factors to bind, but not all open chromatin must be active. Other epigenomic marks are highly correlated with characterized cis-regulatory elements, but they are not confined to demarcate only these elements, nor do they mark all of them. Computational analysis can provide valuable complementary information: it can predict the identity of the trans-factors binding to putative cis-regulatory elements, it can highlight enhancers under active purifying selection, and it can be used to provide enhancer predictions in spatio-temporal contexts that have yet to be assayed.

Many computational screens have been carried out in an attempt to identify enhancers that are active in a particular tissue [5,6]. Previous computational methods often rely heavily on manual curation of TFs that are known to have a role in a particular tissue, or manual curation of lists of known active enhancers [5]. Known enhancers can be used to build a training set that will allow the identification of patterns that are enriched in the training set compared to a background set. Any region across the genome with the same (binding site) patterns are putative enhancers in the tissue from which the training set was built [5]. Because these methods rely on manual curation of data sets, they either do not allow discovery of TFs that are important for a process but have not yet been characterized, or are often limited by the enhancer regions they were trained on.

Here, we introduce an integrated computational framework to identify enhancers in a specific tissue by searching for clusters of TF binding sites (TFBS) with a related function. Our framework first uses a recently published approach to predict high-confidence

## Author Summary

Enhancers are distal gene regulatory elements that can activate tissue- and time-point specific gene expression. Identification of active enhancers is challenging, and is the subject of intense investigation. We developed an automated computational framework to predict transcription factors (TFs) and enhancers that target a tissue of interest by combining two growing resources: TF binding motifs and target gene function annotations. We applied our framework to the placenta, and confirmed our enhancer predictions are more active in placental cell types than others. To demonstrate generalizability, we applied our approach to 5 additional tissues. The combination of experimental sampling with computational prediction approaches will aid in the identification of those enhancers that are most likely active in a particular tissue, as well as the characterization of groups of TFs associated with these enhancers.


binding sites across the genome [7]. Then, each TF is associated with biological functions by taking the set of predictions and analyzing them with GREAT, a functional enrichment analysis tool that assigns biological meaning to a set of putative cis-regulatory genomic regions [8]. We use this approach to first identify TFs with functions related to a particular tissue, which solves the constraint of manual literature curation of TFs and allows identification of TFs with previously uncharacterized roles in the tissue. Because transcription factors generally work in concert through discrete enhancer modules [1], we then search for clusters of binding site predictions for TFs with a related function. These clusters of binding site predictions represent putative enhancers in the tissue of interest.

We applied the above method to discover active enhancers in the mammalian placenta, a tissue that is understudied despite its critical role in human development. Placenta development is a complex, step-wise process, where spatio-temporal control of gene expression must be tightly regulated to ensure proper embryonic and fetal growth [9]. In early stages of gestation, trophoblast cells that surround the developing embryo are directly involved in implantation by attaching the blastocyst to the uterine epithelium [9]. As the placenta continues to develop, it contributes to establishing blood flow between mother and fetus, transporting nutrients, and eliminating waste products [9,10]. Therefore, distinct genetic programs are activated at various times and locations throughout placenta development. Defects in placenta development have also been associated with human disorders such as preeclampsia, and while many SNPs have been identified in association with preeclampsia [11], the function of these SNPs remains unknown.

Our screen identified 2,216 putative placenta enhancers, or TFBS clusters. Of these putative enhancers, 36 were tested using luciferase reporter assays in two placental cell types: mouse trophoblast stem cells (TSCs), and mouse trophoblast giant cells (TGCs) differentiated in culture from TSCs. We also tested the candidates in a primary non-placental cell type as a negative control. We found that 31 (86%) of the candidates were able to drive activity in at least one of TSCs and TGCs, the bulk of which had significantly higher activity in trophoblast cells compared to the other cell type. These results show that our method is able to accurately predict evolutionarily conserved placenta enhancers, which likely function in the development of the human placenta. Because our approach is fully integrated with existing gene ontology databases, we demonstrate it can be easily adapted to well-annotated tissue types by running it on 5 additional tissues.

## Results

### Identification of TFs associated with placenta development

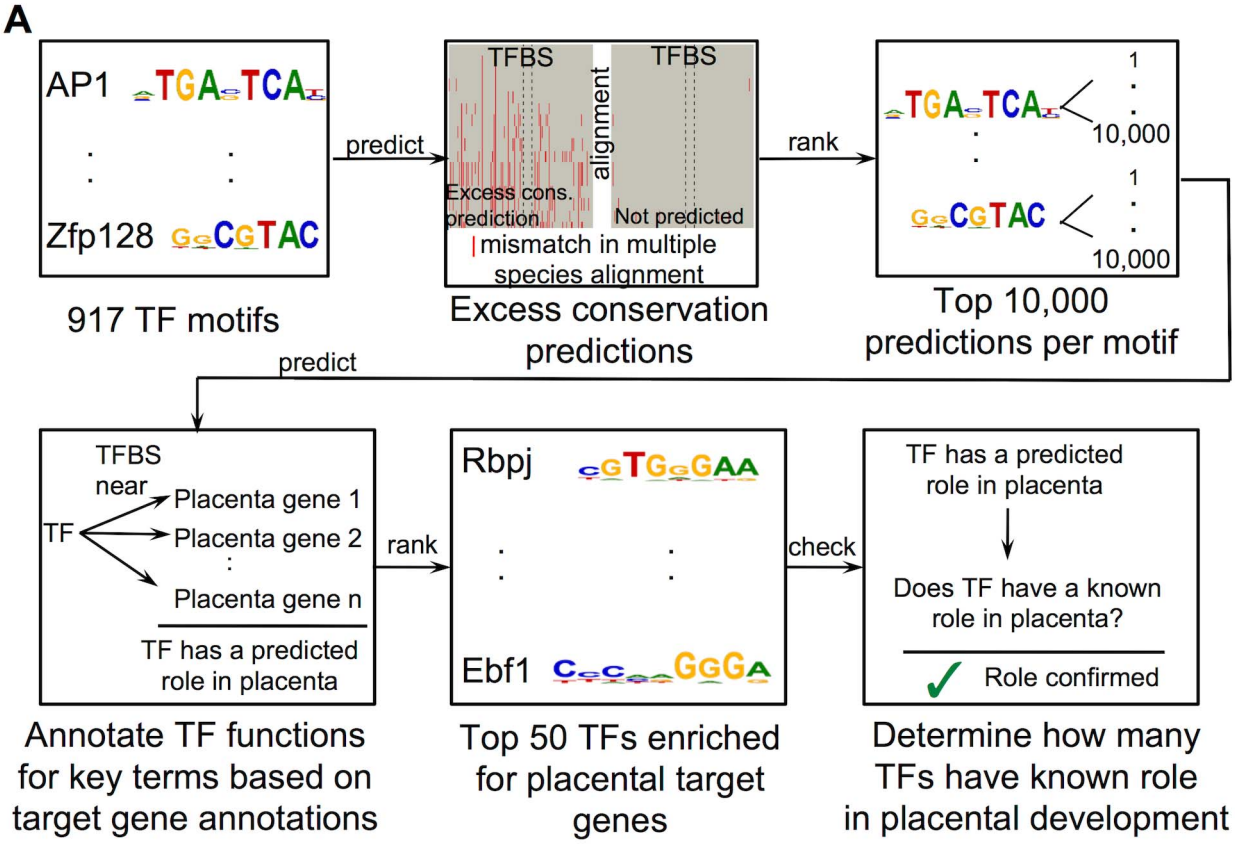
We have previously shown that we can predict TFBS with high accuracy across the genome using an excess conservation metric [7] (Figure 1A). This metric, which improves state of the art TFBS predictions, measures the likelihood for a binding site to be conserved to the observed phylogenetic depth in a particular region of the genome, and favors binding sites that are conserved more strongly than the surrounding sequence [7]. We have also shown that binding site predictions for each TF can be analyzed using a functional enrichment analysis tool, GREAT (the Genomic Regions Enrichment of Annotations Tool) [8], to predict functions for the TFs [7]. GREAT contains *terms*, or lists of genes that have functional commonalities (e.g. placenta development). Given a particular term, GREAT computes the fraction of the genome covered by the regulatory domains of the genes in the list, and the number of binding site predictions hitting these regulatory domains. These data can be used to calculate a p-value for each term using the binomial test, thereby providing a statistic for the enrichment of TFBS near genes annotated for the particular term.

Here we focus the above approach on a particular tissue, the placenta, and search for TFs (motifs) most associated with specific GREAT terms related to the placenta, such as ''abnormal placenta morphology'', ''placenta development'', and ''abnormal placenta labyrinth morphology'' (see Materials and Methods). We used 917 non-redundant motifs curated from UniPROBE [12], JASPAR [13], and TRANSFAC [14], and used the excess conservation metric to identify high confidence matches to each of the motifs genome-wide, that are conserved in mouse and human. By requiring conservation of predictions between mouse and human, we focus on similarities between the species, which aids the study of human development by use of a mouse model. We then obtained functional enrichments for each of the TFs by analyzing the top 10,000 predictions for each motif using GREAT, which allowed identification of the TFs that have the most enrichment for placenta terms (Figure 1A). We next collapsed similar binding motifs (see Materials and Methods), such that each distinct motif was assigned both a q-value (corrected p-value), based on the placenta term that was most enriched, and a most likely TF, based on placenta gene annotations. The 50 TFs with the most significant q-value for a placenta term are shown in Figure 1B.

### Assessment of TFs with known and predicted roles in placenta development

To assess the quality of the 50 TFs predicted to be important for placenta development, we first used an automated method to determine which TFs in the entire ranked list have known roles in placenta development. We predict a TF is involved in placenta development if it has multiple binding sites near genes involved in placenta development, and our prediction can be confirmed when the TF itself is already associated with a placenta term (Figure 1A). To obtain a list of TFs annotated for placenta function, we combined gene lists from two placenta terms in GREAT (see Materials and Methods). We found that a significant number ($p = 0.013$) of TFs in the top 50 appeared in the known placenta gene list (Supplementary Figure S1). We further assessed the ranked list by manually annotating the top 50 TFs. While gene ontologies can be used to identify many genes associated with a process, they may not identify all genes associated with a process. We classified each TF as either previously known to have a role in placenta function, based on the literature, or predicted to have a

A




B

| Motif rank | Most likely gene | q-value | Known in placenta | DNA-binding domain | Motif rank | Most likely gene | q-value | Known in placenta | DNA-binding domain |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Rbpj | 1.30E-30 | ✓ | Lag1 | 26 | Elf5 | 9.63E-19 | ✓ | Ets |
| 2 | Nfkb | 1.73E-30 | ✓ | Rel | 27 | Twist1 | 1.38E-18 | ✓ | bHLH |
| 3 | Ets2 | 3.83E-30 | ✓ | Ets | 28 | Arnt | 2.34E-18 | ✓ | bHLH |
| 4 | Zbtb7b | 6.13E-26 | | C2H2 | 29 | Zbtb6 | 5.43E-18 | | C2H2 |
| 5 | Zscan10 | 1.27E-25 | | C2H2 | 30 | Ikzf2 | 2.36E-17 | ✓ | C2H2 |
| 6 | Tcf4 | 5.55E-25 | ✓ | bHLH | 31 | Ltf | 9.05E-17 | | Lactoferrin |
| 7 | JunB | 6.93E-24 | ✓ | bZIP | 32 | Smad1 | 9.26E-17 | ✓ | MH1 |
| 8 | Gli2 | 3.88E-23 | ✓ | C2H2 | 33 | Ehf | 1.12E-16 | ✓ | Ets |
| 9 | Nfe2 | 7.83E-23 | ✓ | bZIP | 34 | Zfp148 | 2.08E-16 | | C2H2 |
| 10 | Tead4 | 1.30E-22 | ✓ | Homeo | 35 | Plagl1 | 6.48E-16 | | C2H2 |
| 11 | Nr2f2 | 2.46E-22 | ✓ | NR | 36 | Hivep1 | 1.37E-15 | ✓ | C2H2 |
| 12 | Foxo1 | 3.59E-22 | ✓ | Forkhead | 37 | Tfap2c | 2.98E-15 | ✓ | bHLH |
| 13 | Mash2 | 4.58E-22 | ✓ | bHLH | 38 | Nfatc2 | 9.39E-15 | | Rel |
| 14 | Ikzf3 | 5.53E-22 | ✓ | C2H2 | 39 | Rxra | 9.47E-15 | ✓ | NR |
| 15 | Runx1 | 6.80E-22 | ✓ | Runt | 40 | Dmtf1 | 1.16E-14 | | Myb |
| 16 | Gata2/3 | 2.11E-21 | ✓ | GATA | 41 | Zfp219 | 1.20E-14 | | C2H2 |
| 17 | Pparb/d | 8.57E-21 | ✓ | NR | 42 | Tcf3 | 1.25E-14 | ✓ | bHLH |
| 18 | Sp100 | 8.87E-21 | | Sand | 43 | Erg | 1.41E-14 | ✓ | Ets |
| 19 | Zfp691 | 2.40E-20 | | C2H2 | 44 | Tcf12 | 1.48E-14 | | bHLH |
| 20 | Klf6 | 5.12E-20 | ✓ | C2H2 | 45 | Trp53 | 1.90E-14 | ✓ | LSH |
| 21 | Nfatc1 | 5.12E-20 | | Rel | 46 | Klf5 | 2.32E-14 | ✓ | C2H2 |
| 22 | Tfcp2 | 5.18E-20 | | CP2 | 47 | Tcfcp2l1 | 2.42E-14 | | CP2 |
| 23 | Stat3 | 5.67E-20 | ✓ | Stat | 48 | Srf | 4.43E-14 | ✓ | MADS |
| 24 | Insm1 | 1.45E-19 | | C2H2 | 49 | Spdef | 5.88E-14 | ✓ | Ets |
| 25 | Sox2 | 2.14E-19 | ✓ | HMG | 50 | Ebf1 | 9.13E-14 | | bHLH |

**Figure 1. Prediction of transcription factors likely to have a role in placenta development.** (A) For a library of 917 motifs, genome-wide binding site predictions were generated using the excess conservation method [7]. The top 10,000 predictions for each motif were analyzed using

GREAT [8] and TFs were ranked by significance of association with a placenta term. The top 50 TFs were further analyzed to determine if their role in placenta development has already been characterized. (B) The top 50 TFs most enriched for placenta terms, the TF DNA-binding domains, whether the TF is known to have a role in placenta development and the corresponding placenta term q-values.
doi:10.1371/journal.pcbi.1003449.g001

role in placenta function, based on our current approach (Figure 1B, Supplementary Table S1). We determined that 33 (66%) of the top 50 TFs have a known role in the placenta. At the top of our prediction list is *Rbpj*, for which knockout mice show a number of abnormalities, including defects in placenta development [15]. Other well-known TFs in the top 50 that have knockout mice showing placental defects include *Ets2*, *Junb*, *Ascl2*, and *Foxo1* [16–20].

Interestingly, 17 (34%) of the top 50 TFs do not currently have a well-characterized role in placenta development. We determined the expression levels of the predicted TFs, using published human RNA-Seq data in three placental components: amnion, chorion, and decidua [21], as well as published mouse RNA-Seq data in TSCs [22]. Fifteen out of the seventeen predicted TFs (88%) are highly expressed in at least one of the four placental cell types (Supplementary Table S2), which is significant ($p = 5\times10^{-3}$) when compared to the number of TFs highly expressed when TFs are chosen randomly from all TFs with expression values (1,000 simulations), providing further evidence for their role in placenta development.

## Identification of potential placenta enhancers

To identify placenta enhancers, we developed an algorithm to discern clusters of binding sites using predictions from the top 50 TFs annotated with a placenta function. The algorithm first uses spatial hierarchical clustering based on distance between predicted TFBS, and then segments the cluster hierarchy based on cluster score (see Materials and Methods). The clustering process groups the initial set of 485,038 binding site predictions (~10,000 per TF in top 50) to 255,209 regions (average length 76 bp, maximum 544 bp) with 1 or more binding site prediction from the top 50 TFs. The 255,209 regions had a low fold enrichment for placenta terms, as determined through GREAT. The fold enrichment reported by GREAT measures the number of regions associated with a term compared to the expected region hits, given the size of the input set and the fraction of the genome covered by the term. Because functional enhancers often harbor binding sites for multiple TFs [1], we enriched for regions that are likely to function as placenta enhancers by keeping only those regions with 5 or more non-overlapping binding site predictions for one or more of the top 50 TFs, giving 3,014 potential placenta enhancers that had an average size of 279 bp. This filtering step increased the fold enrichment for multiple placenta terms well beyond the default GREAT two fold significance threshold (Figure 2A).

We further quantified the enrichment obtained by choosing regions with ≥5 TFBS. We show that when randomly choosing 3,014 regions from the set of 255,209 clusters that contain one or more binding site predictions, the observed q-value (when regions with ≥5 TFBS are chosen) of $3.99\times10^{-24}$ for the "placenta development" term is highly significant ($\log_{10}$ of Z-score = 19.08), as the best q-value observed in 1,000 simulations was $3.85\times10^{-9}$ (Figure 2B).

We also conducted a sensitivity analysis to determine the effect of varying the thresholds we used in this process. We wanted to choose a threshold that would result in a large number of TFBS clusters (>1,000) and a high GREAT fold enrichment for a representative GREAT term (>3). First, we determined the effect on the GREAT fold enrichment for placenta terms when, instead of clustering predictions from the top 50 TFs, we clustered predictions from the top 25, 75, or 100 TFs. As shown in Supplementary Figure S2, using the top 50 TFs provides both the best fold enrichment and best balance between quantity and purity. We next varied the number of non-overlapping TFBS used to identify placenta TFBS clusters from at least 3 non-overlapping placenta TFBS to at least 7 non-overlapping placenta TFBS. Requiring at least 6 non-overlapping placenta TFBS slightly increases the fold enrichment, but provides less than 1,000 predictions, while dropping the threshold to at least 4 non-overlapping placenta TFBS lowers the fold enrichment to below 3 (Supplementary Figure S2). Finally, we tested the effect of including non-placenta TFBS when counting non-overlapping binding sites in a TFBS cluster to accommodate for both general purpose and specific TFBS in the same enhancer. For example, for each threshold of at least 4 non-overlapping TFBS to at least 7 non-overlapping TFBS in a region, we required that ≥3 of the TFBS be from the placenta (top 50) TFs. As expected, we see that reducing the number of TFBS that are required to be from placenta TFs increases the number of TFBS clusters identified. However, it also reduces the GREAT placenta term fold enrichment (Supplementary Figure S2), indicating the TFBS clusters identified are less likely to be involved in placenta functions.

To further enrich for enhancers likely specific to placenta development, we filtered regions that contain a high number of non-placenta TFBS clusters. While these regions may be active enhancers in the placenta, it is difficult to claim that they are specific to the placenta based on TFBS composition alone, as the number of non-placenta TFBS clusters in these regions is high. To this end, we ran the binding site clustering approach using 50 random TFs that were unlikely to have a role in placenta development (rank below 100 from the ordered list of TFs). We carried out this process a total of 1,000 times to obtain a set of background clusters with ≥5 non-overlapping TFBS. We then determined the number of times each of the 3,014 putative placenta enhancers overlapped a background cluster. Finally, we removed from the putative placenta enhancer set those regions that were identified in at least 5% of the background runs. This process further enriched our set for placenta functions, increasing the range of GREAT fold enrichments for placenta terms to 3.17–4.56, and left us with our final set of 2,216 placenta TFBS clusters (Figure 2A, Supplementary Table S3).

Because of the initial conservation criteria we used for binding site predictions, each of the placenta TFBS clusters is conserved in human. A roughly balanced 1.4% to 8.2% of the total predictions we started with for each of the top 50 TFs before clustering ends up in a placenta TFBS cluster; placenta TFBS clusters are generally heterotypic, made up of binding site predictions for multiple different TFs; and all of the top 50 TFs contribute to the different clusters, irrespective of their motif information content, or whether they are already known or not in placental contexts (Supplementary Figure S3).

## Functional validation of putative placenta enhancers

To functionally test the placenta TFBS clusters, we performed enhancer reporter assays in a mouse placenta cell culture system. This system allows one to maintain trophoblast stem cells (TSCs) derived from early mouse development, which are the precursor cells of the differentiated cells of the placenta [23]. TSCs can also

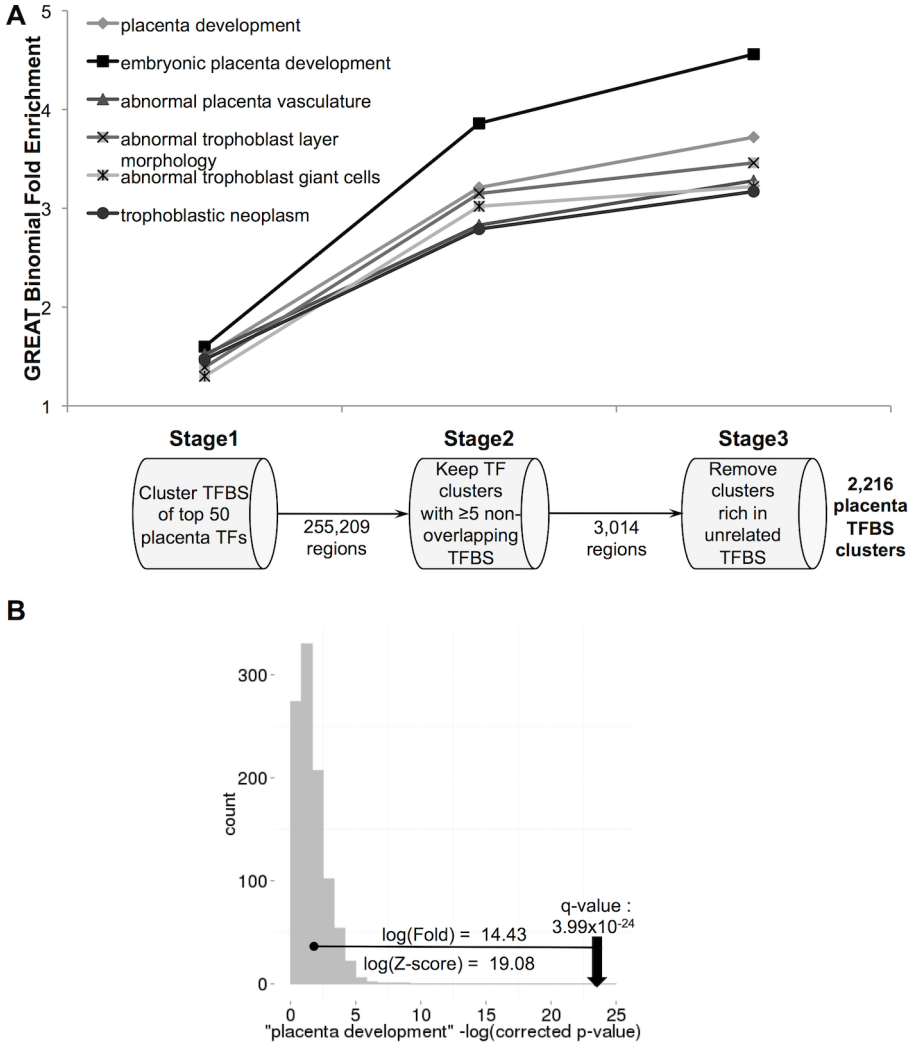



**Figure 2. Identification of placenta TFBS clusters.** (A) Three stages of data processing are used to identify placenta TFBS clusters. In stage 1, the top 10,000 predictions for the top 50 TFs were compiled, and binding site predictions in close proximity were clustered. In the second stage, clusters containing less than 5 non-overlapping binding sites were removed. In stage 3, regions that are rich in binding site predictions for TFs below rank 100 were removed. The filtering steps strongly enrich for placenta terms in GREAT, as shown by plotting the fold enrichment in each stage for six different terms. (B) Null model showing that choosing clusters with ≥5 non-overlapping TFBS enriches for clusters in the regulatory domain of genes involved in placenta development. Gray bars represent distribution of $-\log_{10}$(q-value) for "placenta development" term when 3,014 (size-matched to stage 2 in (A)) clusters are selected randomly from stage 1 in (A). Arrow points to $-\log_{10}$(q-value) when only clusters with ≥5 non-overlapping TFs are selected.
doi:10.1371/journal.pcbi.1003449.g002

be differentiated into trophoblast giant cells (TGCs), which are the placental cell type that invade maternal tissue to help establish maternal-fetal blood flow [23]. This cell culture system therefore allows testing of enhancer elements using luciferase reporter assays in two different placental cell types (Supplementary Figure S4). Because TSCs often spontaneously differentiate into TGCs, we ensured the purity of our transfected cell populations by using non-overlapping transfection conditions that were optimized for each cell type (see Materials and Methods), and by transfecting the mouse placental lactogen II (*mPL-II*) enhancer as a control for each set of TSC or TGC transfections. PL-II is known to be specific to TGCs, and the enhancer region we tested has been shown to be active in rat trophoblast giant cell-like cells [24]. Indeed, we found consistently high *mPL-II* enhancer activity in our TGCs and little to no activity in our TSCs (Supplementary Figure S4).

We tested 36 placenta TFBS clusters upstream of a minimal promoter driving luciferase activity in TSCs and TGCs (Supplementary Table S4). The candidates were selected to cover a range of distances from gene transcription start sites (TSSs), to cover a range of non-overlapping TFBS predictions, and to contain different proportions of binding sites for known to predicted TFs. Luciferase fold activity for all candidates was calculated compared to an empty vector control. Of the 36 candidates, 31 (86%) showed more than 2-fold activity in at least one of TSCs and TGCs, 26 (72%) showed more than 2-fold activity in TSCs, and 28 (78%) showed more than 2-fold activity in TGCs (Supplementary Figure S5). 19 of the 36 candidates consisted of more binding sites for TFs we predicted to have a role in placenta development rather than previously characterized placenta TFs. Of these 19 candidates, 16 (84%) had activity in at least one of TSCs and TGCs. Additionally, there was no strong correlation between fold activity and

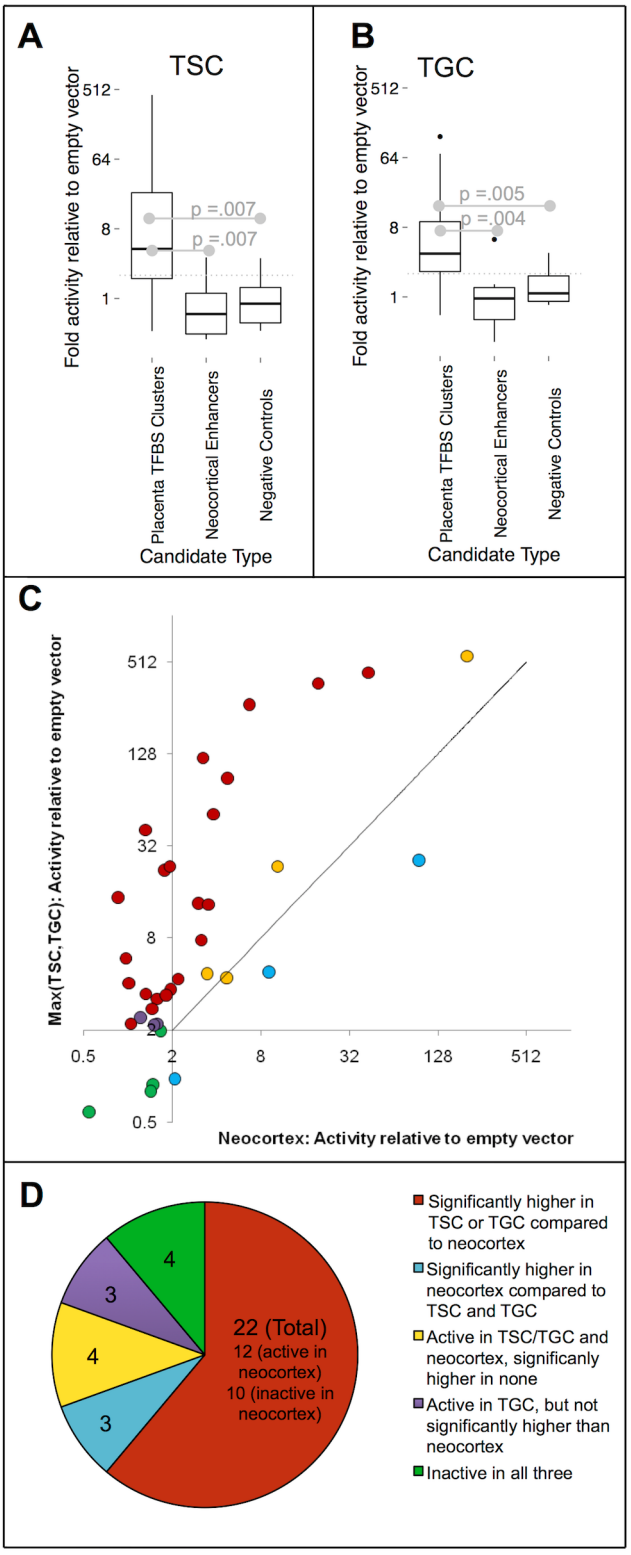



**Figure 3. Luciferase activity of placenta TFBS cluster candidates.** Box plots summarizing activity of placenta TFBS cluster candidates and negative controls in TSCs (A) and TGCs (B). Significance is calculated using the unpaired *t*-test. (C) Comparison of enhancer activity relative to empty vector for the higher value between TSCs and TGCs versus neocortical cells. In general, enhancers were much more active in at least one of TSCs and TGCs. Colors correspond to groups in (D). (D) Pie chart showing that the number of enhancers that had significantly higher activity (unpaired *t*-test, p-value $<0.05$) in placental cell types versus neocortical cells was greater than the number of enhancers in any other category. Candidates with $\geq$2-fold activity are considered active, and those with $<$2-fold activity are considered inactive.

doi:10.1371/journal.pcbi.1003449.g003

proportion of known placenta TFBS in either of the two cells types for the candidates (TSC $r^2 = 0.19$, TGC $r^2 = 0.17$), further implicating the predicted TFs in placenta development.

We then tested the specificity of the transfection system in placental cells using 15 negative controls. Because our lab has interest in the neocortex [25], we chose 6 negative controls that are robust enhancers in neocortical cells (Supplementary Figure S6). We also chose 9 negative controls within the same range of GC content, length, and level of conservation to the placenta TFBS cluster candidates. For both sets of negative controls, we observed significantly lower activity in TSCs (unpaired *t*-test p-value $= 7\times10^{-3}$) and TGCs (unpaired *t*-test p-value $\leq 5\times10^{-3}$) compared to the placenta TFBS cluster candidates, demonstrating the robustness of the transfection system (Figure 3A–B, Supplementary Figure S6).

## Assessing tissue specificity of placenta TFBS clusters

To assess the specificity of the 36 placenta TFBS cluster candidates tested in placental cell types, we also performed luciferase reporter assays on neocortical cells isolated from e14.5 mice [25]. Of the 31 candidates that had more than 2-fold activity in TSCs or TGCs, 22 (71%) had significantly higher (unpaired *t*-test, p-value $<0.05$) activity in TSCs or TGCs compared to neocortical cells, including 12 candidates (39%) that were inactive ($<$2-fold activity) in neocortical cells (Figure 3C–D and Supplementary Figure S7). To ensure the difference we see is not due to high basal activity of the promoter in placental cell types compared to neocortex, we show that the basal activity of the promoter is low in each cell type and should have no impact on our results (Supplementary Figure S7). These data demonstrate that our approach is able to identify placenta enhancers with high accuracy, the bulk of which are more active in trophoblast cells compared to neocortical cells.

To compare the predicted placenta TFBS clusters more globally to putative enhancers in other tissues, we overlapped them with epigenomic enhancer-associated marks in 18 other mouse tissues and cell types [4]. Each cell type we compared to has, on average, 50,166 putative enhancers. We see that 793 (36%) of the placenta TFBS clusters overlap with putative enhancers in zero other tissues. 1,108 (50%) of the placenta TFBS clusters overlap with putative enhancers in $\leq$2 other tissues, and only one placenta TFBS cluster overlaps with putative enhancers in all other tissues (Supplementary Figure S8). This comparison further suggests that the placenta TFBS clusters are less likely to be active in many other tissues.

## Discovery of functional enhancers near genes involved in placenta development

Our method has led to the identification of previously uncharacterized enhancers in the regulatory domains of target genes that play an important role in placenta development. For example, *Hand1* has been shown to be essential for placenta development, as the knockout mice arrest by e7.5 and have defective trophoblast giant cell differentiation [26]. Our analysis reveals multiple putative enhancers in the regulatory domain of *Hand1*, two of which are over 30 kb upstream of the transcriptional start site and are conserved amongst placental mammals

(Figure 4A). Our luciferase reporter assays show that both candidates have significantly higher activity in the placental cell types compared to neocortical cells (Supplementary Figure S5). Another example is a putative enhancer ~13 kb upstream of *Dll4*, a gene that is involved in the development of the placenta vasculature (Figure 4B) [27,28]. This candidate was also found to have significantly higher activity in placental cell types than in neocortical cells (Supplementary Figure S5). In total, 14 of the 36 candidates tested are found in the regulatory domain of genes that are known to be involved in placenta development. Both TSCs and TGCs showed no significant difference in activity for candidates near target genes known to be involved in placenta development compared to those that have unknown functions in placenta (unpaired *t*-test, TSC p-value = 0.27, TGC p-value = 0.32), suggesting that our method can be used not only to predict TFs but also to predict target genes that were not previously known to have a function in placenta.

Because genes important in a particular context can be regulated by multiple enhancers [25,29–32], we searched for genes with the most predicted placenta TFBS clusters in their regulatory domains. To carry out this analysis, we determined a q-value (using GREAT) for all the genes in the genome based on the likelihood associated with the observed number of placenta TFBS clusters per gene, normalized to the length of the individual gene's regulatory domain. The ten genes with the most significant q-values are shown in Supplementary Table S5. Of the top ten genes, four of them, *Pdgfb*, *Junb*, *Epha2*, and *Socs3* have previously characterized roles in placenta development. Interestingly, *Zbtb7b*, a TF we predict to have a role in placenta development based on our approach, is also within the top ten, with five placenta TFBS clusters within its regulatory domain.

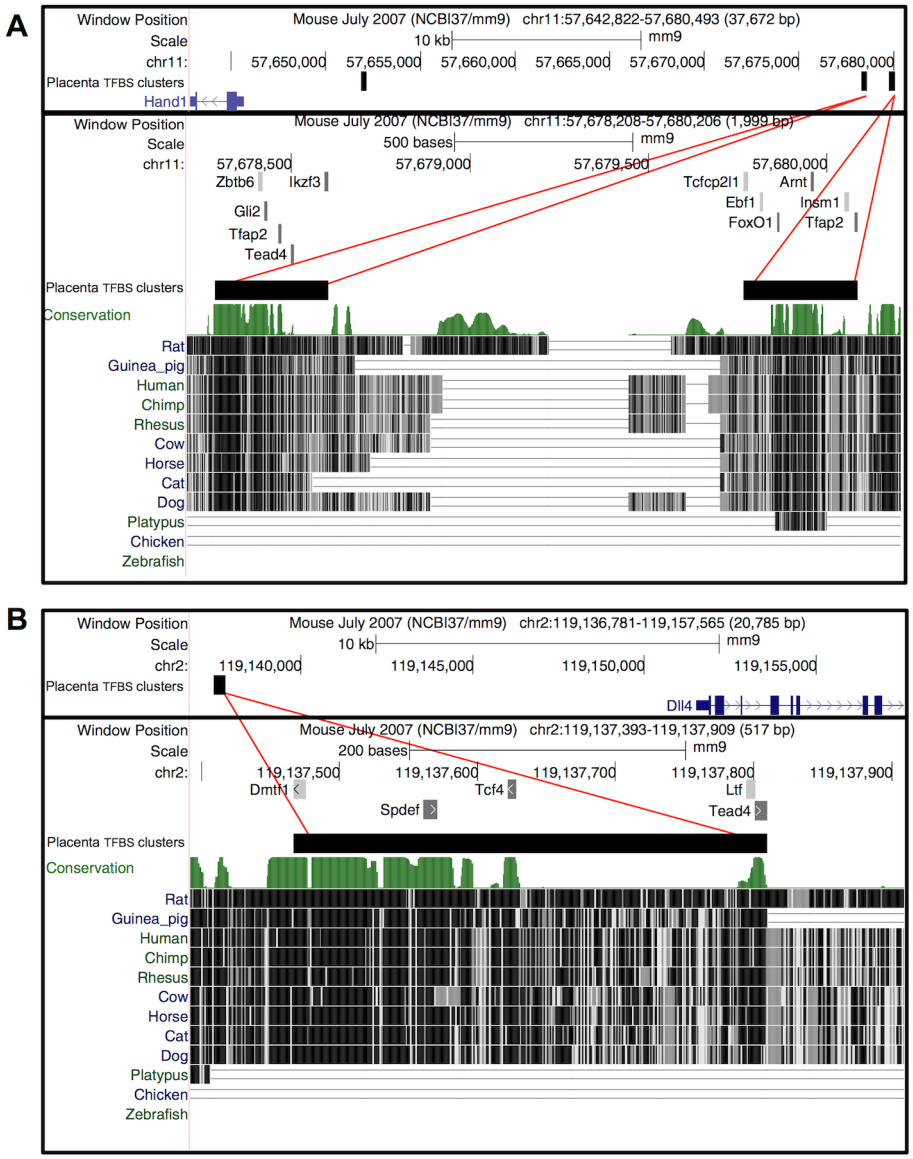



**Figure 4. Placenta TFBS clusters in the regulatory domain of genes with important roles in placenta development.** *Hand1* (A) and *Dll4* (B) contain placenta TFBS clusters in their regulatory domains. Placenta TFBS clusters that were tested are shown in the lower panel of each figure along with representative binding sites that were predicted over them. Binding sites for TFs that have a known role in placenta development are shaded dark gray, whereas bindings sites for TFs that have a predicted role in placenta development are shaded in light gray.
doi:10.1371/journal.pcbi.1003449.g004

## Comparison to mouse and human placenta data

We next sought to further compare our computational enhancer predictions to published datasets that use biochemical assays to predict enhancers genome-wide. The mouse ENCODE project has recently generated ChIP-Seq data for enhancer-associated chromatin marks in mouse term placenta [4], and the ENCODE and Roadmap Epigenomics projects have generated DNase-Seq data in human placenta tissue at 85–113 days gestation to assay for open chromatin [2,3]. To enrich for enhancer-associated chromatin from the mouse ChIP-Seq data, we combined regions marked by enhancer-associated marks H3K27ac and H3K4me1 that do not contain the promoter-associated mark H3K4me3. For the human DNase-Seq set, we combined data from six biological replicates (see Materials and Methods). Because we were interested specifically in comparing enhancer-associated regions, we removed regions within 1 kb of gene transcriptional start sites from all three sets before comparison. This brought the number of: placenta TFBS clusters to 1,847 (covering 0.02% of the genome), mouse histone mark based data to 70,951 (covering 10.22% of the genome), and human open chromatin data to 80,922 (covering 0.38% of the genome).

We first determined the enrichment for placenta terms using GREAT for all three sets. As expected the placenta TFBS clusters have the highest enrichment for several placenta terms, partly because the process used to define them specifically enriches for a subset of these terms. However, significance for the other two sets was quite low, with most values between 1.7–1.9 fold enrichment, below GREAT's standard significance cut-off of 2-fold (Table 1). These low values suggest that it might be more difficult to identify functional placenta enhancers from these large sets.

Second, we wanted to determine if any of the placenta TFBS clusters overlapped with the mouse or human enhancer-associated regions. To compare mouse and human data, we first converted human region coordinates to mouse coordinates (see Materials and Methods). We found that 880 (48%) of the placenta TFBS clusters overlapped with the mouse experimental set, and 390 (21%) of the placenta TFBS clusters overlapped with the human experimental set (Figure 5A–B). To determine if the overlaps we observed were significant, we chose 1,847 (the set size of the placenta TFBS clusters that are not within 1 kb of a TSS) size-matched random genomic regions and checked the overlap with the mouse and human experimental sets. We ran this process a total of 10,000 times. The simulation demonstrates that the placenta TFBS clusters have very significant overlap with both the mouse ($p<10^{-4}$, Z-score = 50.25) and human ($p<10^{-4}$, Z-score = 77.71) experimental data (Figure 5C–D). The simulation also demonstrates that while the human experimental set has less overlap with the placenta TFBS clusters, the Z-score for the overlap is more significant. The difference in overlap and Z-score is likely due to the difference in peak width between the datasets: the peaks in the mouse experimental set are 3,000 bp wide whereas the peaks in the human experimental set are, on average, 125 bp wide. To check this, we padded each peak in the human experimental data set such that the peak widths were 3,000 bp (with the full set now covering 9.26% of the genome). This increased the overlap with the placenta TFBS clusters to 796 (43%), and brought the Z-score (53.61) closer to the Z-score of the mouse experimental data set.

We next wanted to check that the placenta TFBS clusters that do not overlap with the mouse and human experimental data are likely to be involved in placenta development. We first identified TFBS clusters that do not overlap with either the mouse or the human experimental data. 831 (45%) of the placenta TFBS clusters that are not within 1 kb of a TSS do not overlap with either experimental dataset (Figure 5E). Additionally, of the candidates that were tested for enhancer activity in TSCs and TGCs, 8 are unique to the placenta TFBS clusters, and 7 of these had more than 2-fold activity compared to the empty vector in at least one of TSCs and TGCs. GREAT analysis of the 831 elements unique to placenta TFBS clusters shows they still have strong enrichment for placenta functions; for example, "abnormal trophoblast layer morphology" has a q-value of $2.49\times10^{-8}$ and a fold enrichment of 3.53 (Supplementary Table S6). We next determined if the fraction of regions associated with a placenta term in GREAT was higher for unique placenta TFBS clusters, compared to regions that were only identified in the mouse experimental set, or regions that were only identified in the human experimental enhancer set. We found that the unique placenta TFBS clusters have between 1.21-fold and 2.60-fold more regions associated with placenta terms (Table 2). The GREAT terms are somewhat incomplete, in that every gene involved in placenta development has not been characterized. Nevertheless, this test suggests that if choosing a candidate randomly from the sets that are unique to each method, a candidate chosen from the placenta TFBS clusters is more likely to function in the placenta.

Because our pipeline filters for regions of the genome that contain an abnormally high number of non-placenta TFBS clusters, we also compared our data to a filtered version of the mouse experimental data, containing putative placenta-specific enhancers. This set contains 4,326 regions (>1 kb from a gene TSS) and was generated using a tissue-specificity index based on H3K4me1 occupancy in the 18 tissue and cell types described above [4]. GREAT analysis shows that while the putative

**Table 1.** GREAT enrichments for placenta TFBS clusters, mouse placenta epigenomic set, human placenta epigenomic set.

| Ontology | Term | Placenta TFBS clusters fold enrichment | Mouse epigenomic set fold enrichment | Human epigenomic set fold enrichment |
|---|---|---|---|---|
| GO Biological Process | embryonic placenta development | **5.12** | 1.98 | 1.92 |
| GO Biological Process | placenta development | **4.13** | 1.86 | 1.87 |
| Disease ontology | trophoblastic neoplasm | **3.8** | 1.74 | 1.73 |
| Mouse Phenotype Single KO | abnormal trophoblast layer morphology | **3.79** | **2.01** | 1.93 |
| Mouse Phenotype Single KO | abnormal trophoblast giant cells | **3.58** | **2.09** | 1.87 |
| Mouse Phenotype Single KO | abnormal placenta vasculature | **3.46** | 1.83 | 1.71 |

GREAT fold enrichment for various placenta terms are shown for the three putative enhancer datasets. The placenta TFBS clusters have consistently higher fold enrichment compared to the other datasets. Bold numbers are above the default fold enrichment cutoff for a term to be considered enriched in GREAT.
doi:10.1371/journal.pcbi.1003449.t001

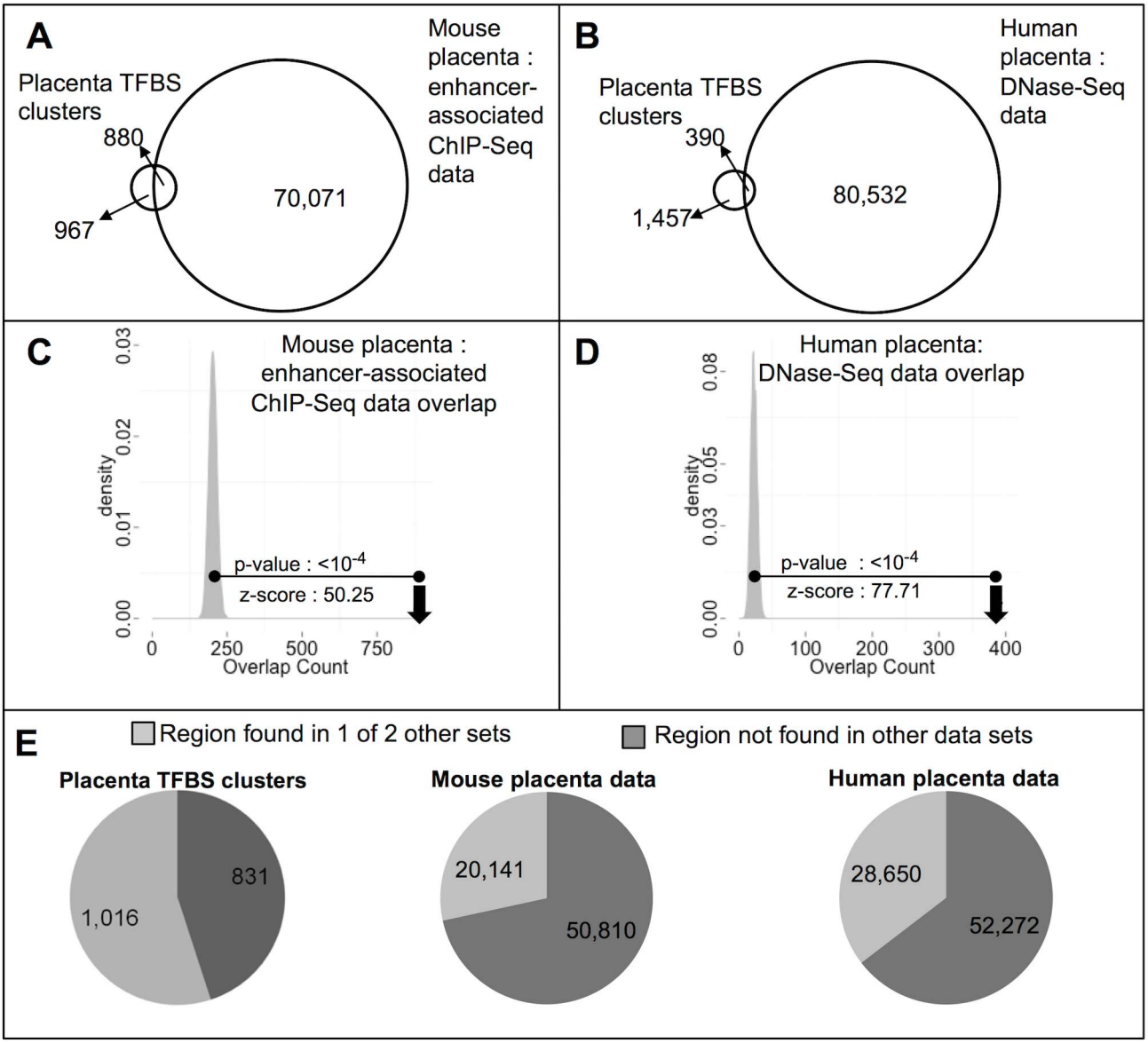



**Figure 5. Overlap with large-scale experimental data.** Overlap of placenta TFBS clusters with mouse placenta enhancer-associated ChIP-Seq data (A) and human placenta DNase-Seq data (B). Placenta TFBS clusters had a significant overlap with both the mouse (C) and human (D) experimental data. (E) Each dataset had a large proportion of regions that were not identified in the other two datasets. Over 800 placenta TFBS clusters were not found in either the mouse or human experimental data. Counts for human data are after conversion to mouse mm9.
doi:10.1371/journal.pcbi.1003449.g005

**Table 2.** Fraction of unique placenta TFBS clusters associated with placenta terms compared to placenta experimental sets.

| Ontology | Term | Placenta TFBS clusters : Mouse | Placenta TFBS clusters : Human |
|---|---|---|---|
| GO Biological Process | embryonic placenta development | 1.68 | 1.95 |
| GO Biological Process | placenta development | 1.54 | 1.67 |
| Disease ontology | trophoblastic neoplasm | 2.59 | 2.60 |
| Mouse Phenotype Single KO | abnormal trophoblast layer morphology | 1.92 | 2.29 |
| Mouse Phenotype Single KO | abnormal trophoblast giant cells | 1.72 | 2.23 |
| Mouse Phenotype Single KO | abnormal placenta vasculature | 1.21 | 1.53 |

For all terms, the fraction of unique placenta TFBS clusters (not found in the other two sets) found next to a placenta annotated target gene was higher than the fraction of unique mouse/human experimentally annotated placenta enhancer sets associated with the same placenta term. First two columns show the term, the third column shows the ratio of the fraction of unique placenta TFBS clusters to the fraction of unique mouse placenta experimental enhancers, and the fourth column shows the ratio of the fraction of unique placenta TFBS clusters to the fraction of unique human placenta experimental enhancers.
doi:10.1371/journal.pcbi.1003449.t002

placenta-specific enhancers from [4] are more enriched for placenta terms than the full set of mouse experimental data, 5 out of 6 placenta terms have higher fold enrichment in GREAT analysis of our placenta TFBS clusters (Supplementary Table S7). There are only 12 regions shared between the placenta TFBS clusters and mouse putative placenta-specific enhancers, and for the same 5 out of 6 placenta terms, regions unique to the placenta TFBS clusters have a higher fraction of regions associated with the placenta terms (Supplementary Table S7).

Finally, we compared our data to recently published ChIP-Seq data for enhancer marks and repressor marks in mouse TSCs [22]. One would hope that our placenta TFBS clusters have significant overlap with enhancer mark ChIP-Seq data, and little overlap with repressor mark ChIP-Seq data in this cell type, and this is indeed the case. While 299 placenta TFBS clusters overlap with TSC ChIP-Seq enhancer-associated peaks (enrichment $p<10^{-4}$, Z-score = 21.93), only 1 placenta TFBS cluster overlaps with TSC ChIP-Seq repressor associated peaks (H3K9me3: depletion $p<10^{-4}$, Z-score = 8.09; H3K27me3: depletion $p<10^{-4}$, Z-score = 4.13) (Supplementary Figure S9).

Together, these data show that while the large-scale experimental data are valuable, they can be strengthened with complementary computational analysis. The method we describe provides a smaller, more focused set with additional candidates that are near genes functioning in placenta development, and that very likely drive activity in placental cell types.

### Generalizing our method to other tissues

To demonstrate that our approach can be generalized, we applied our method to identify TFs and TFBS clusters in five additional tissues: heart, pancreas, blood vessel, bone marrow, and liver. The top 50 motifs and all TFBS clusters for each tissue are provided as Supplementary Tables S8, S9, S10, S11, S12. We first examined whether TFs annotated by GREAT as contributing to tissue development are indeed enriched in our predicted top 50 TFs per tissue (Supplementary Figure S10). The enrichment is strong in four of five tissues (heart, blood vessel, bone marrow and liver; all $p<10^{-3}$), but not in pancreas ($p = 0.1$). When we examine the top GREAT enrichment for the TFBS clusters for each tissue we obtain a similar picture: clusters for the same four tissues yield a top prediction that matches the tissue identity (e.g. "abnormal ventricle myocardium morphology" for heart), while the top term for pancreas is a mismatched brain related term (Table 3).

Because our method identifies the most relevant TFs for a tissue based on enrichment of the TFBS near target genes already annotated to have a role in the tissue, we expected that our approach would be most suitable for tissues with well-annotated terms. If a GREAT term is not well annotated for a tissue, then it is more difficult to determine if the binding site predictions for a particular TF are enriched near genes involved in development of the tissue, because the number of genes that are known to be relevant for the tissue is low. Indeed, we see that the GREAT term used to identify TFs involved in pancreas development is only annotated with 75 target genes, compared to 340–1,047 genes associated with terms used for the other tissues investigated (Supplementary Table S13). These data confirm that our approach works best when genes involved in the tissues or processes are well annotated, likely because of increased ability to predict the TFs that are the most relevant to the tissue.

## Discussion

Here we describe a novel method to identify previously uncharacterized TFs that may have a role in a tissue of interest, as well as active enhancers particularly relevant to the tissue of interest. The automated identification of TFs relevant to a tissue overcomes limitations of current methods for computational identification of enhancers.

We first used our method to implicate 50 TFs in placenta development, 33 of which were confirmed to have roles in placenta development in the literature. We predict that the 17 remaining TFs have a role in placenta development, and binding sites for these TFs would not be included as input for previous computational methods that rely heavily on manual curation of TFs. There are multiple lines of evidence supporting our prediction that the remaining 17 TFs have a role in placenta development, both through our analysis and the literature. For example, the highest ranking TF ($4^{th}$) we predict to have a role in placenta development is *Zbtb7b*. *Zbtb7b* knockout mice have defects in the hematopoietic system, and defects in T-cell development and differentiation [33]. Furthermore, intercrosses of mice with a mutation in *Zbtb7b* produce small litters, whereas wild-type females crossed with mutant males have normal litter sizes [33]. This suggests impairment of female fertility in mutant mice, perhaps due to defects in the decidua, the maternal component of the placenta. *ZBTB7B* is also highly expressed in three components of the human placenta, including the decidua (top quartile for expression value) [21]; is highly expressed in mouse TSCs [22]; and has been shown to be up-regulated in placentas from pregnancies that resulted in intrauterine growth restriction (IUGR), or placentas from pregnancies that resulted in pre-eclampsia and IUGR [34]. To further investigate and confirm the importance of the TFs we predict to be involved in placenta development, placenta-specific mouse misregulation models could be generated [35,36].

**Table 3.** TFBS clusters for other tissues.

| Tissue | # TFBS clusters | Most enriched term | q-value | Fold enrichment |
|---|---|---|---|---|
| Heart | 2251 | abnormal ventricle myocardium morphology | 1.06E-24 | 3.43 |
| *Pancreas* | *1697* | *abnormal brain commissure morphology* | *1.88E-22* | *3.36* |
| Blood Vessel | 1674 | abnormal dorsal aorta morphology | 2.16E-22 | 4.04 |
| Bone Marrow | 1184 | abnormal CD4-positive T cell physiology | 2.05E-12 | 3.64 |
| Liver | 566 | abnormal liver size | 7.49E-11 | 3.31 |

We ran our pipeline on 5 additional tissues: heart, pancreas, blood vessel, bone marrow, and liver (Supplementary Tables S8, S9, S10, S11, S12), and show the top enriched term for the MGI Phenotype Single KO ontology for each. Pancreas (in italics) is the only tissue for which the top term reported through GREAT was not relevant to the tissue analyzed (see text).

doi:10.1371/journal.pcbi.1003449.t003

We next used our method to identify placenta enhancers by searching for regions of the genome containing clusters of five or more non-overlapping placenta TFBS. We found that the set of placenta TFBS clusters was highly enriched for placenta terms in GREAT, and we generated a null model to ensure the enrichment was due to choosing regions with ≥5 placenta TFBS. We validated our approach using luciferase reporter assays for two placental cell types: TSCs, and TGCs differentiated from TSCs. 31 (86%) of the candidates we tested were active in at least one of TSCs and TGCs. It remains possible that the 5 (14%) that did not show activity are active in a different placenta cell type. Of the 5 candidates that do not show activity in TSCs or TGCs, 4 (80%) are annotated as biochemically active in the mouse or human experimental placenta datasets [2–4]. Many of the candidates we tested are in the regulatory domain of genes that are well studied and have a known role in placenta development. The enhancers we identify near these genes have not been characterized, and could be important regulators of the placenta genes. We also show that of the 19 placenta TFBS clusters we tested that consist of more binding sites for TFs we predict to have a role in placenta development, 16 (84%) are active in at least one of TSCs and TGCs. These enhancers would not be identified in computational screens that first identify relevant TFs based on manual curation and then search for clusters of binding sites for only those TFs.

We also showed that we can add value to the experimental assays used to generate large-scale data sets through the ENCODE and Roadmap Epigenomics projects. These projects have provided tens of thousands of regions, consisting of enhancer-associated chromatin marks and open chromatin that may be functional during a specific time of placenta development. Our approach allows focusing on regions that are likely more specific for placenta functionality, and identifies many additional putative placenta enhancers, unique to those identified in the experimental datasets. The comparison between putative enhancers identified through our computational approach and putative enhancers identified through epigenomic approaches demonstrates that both approaches likely result in numerous false negatives. While our approach identifies putative enhancers that are missed at specific time-points assayed, we only capture a subset of the putative enhancers identified through experimental approaches. Therefore, it is the combination of computational and experimental approaches that will allow us to more comprehensively understand the enhancers that govern embryo development, through its many tissue and time point combinations.

To show that our approach can be generalized, we applied it to generate TFBS clusters in 5 other tissues. Of these tissues, 4 yielded a relevant top enrichment in GREAT. The tissue for which TFBS clusters did not result in a relevant GREAT enrichment was limited by the low number of genes annotated for the relevant GREAT term. These results show that the computational framework described can be easily adapted to other tissues, developmental processes, or across different environmental conditions for which functional annotations are available. This can help overcome the burden associated with carrying out biochemical assays on every tissue and experimental condition and will become even more powerful as gene ontologies for various tissues and processes continue to improve. Gene ontologies are also becoming more specific, and terms in the ontologies more often relate to a particular cell type of a tissue. Tissues are generally not made up of homogenous cell populations, so as the more specific terms become better annotated, our approach is expected to provide enhancer sets for these different cell types.

Comparison of mouse and human placentas has shown that gene expression patterns and pathways are often conserved [37,38]. The conserved placenta TFBS clusters we identify are likely functional in both species. Using a conservation metric provides confidence in the functionality of the elements we identify, as it has been shown that conserved binding sites are more likely to lie within active enhancers [39,40]. Additionally, our method of identifying commonalities between mouse and human is beneficial as it allows us to learn about the human condition by using insights from the mouse model.

Transcriptional regulation is a complex process, and identification of enhancers that regulate each developmental process is a challenging task. We have shown that the method we described can be used to accurately predict enhancers in the placenta. This method can be generalized to other tissues, and can complement tissue and time-point restricted data coming from projects such as ENCODE and the Epigenomics Roadmap, highlighting enhancers that have been under purifying selection through mammalian evolution, and therefore are more likely to contribute to phenotypic and disease susceptibility differences.

## Materials and Methods

### Transcription factor motif library curation

A transcription factor motif library was curated as described previously [7], resulting in a non-redundant set of 917 motifs from UniPROBE [12], JASPAR [13], and TransFac [14].

### Binding site predictions

The excess conservation method used for binding site prediction was described previously [7]. Binding site predictions for our motif library were carried out in mouse mm9, ensuring that all predictions were conserved in human hg18, using the PRISM pipeline for binding site prediction and scoring with the following parameters: binding site prediction threshold was 800, binding sites were allowed to shift by 20 base pairs relative to the reference, binding sites were required to have a minimum branch length of 2 substitutions per site, the binding site had to be present in at least 5 species, one of those species had to be human, and the p-value of the observed motif score against motif shuffles in similarly conserved windows had to be ≤0.05.

### Identification of TFs functioning in placenta

The top 10,000 predictions, ranked by p-value, were obtained for all 917 motifs. If a motif did not have 10,000 predictions satisfying the prediction criteria, then the lower number of predictions for that motif was used. If the motif had additional predictions after the top 10,000 with the same score (ties) as the last prediction included, they were also included for further analysis. Each set of predictions was run through GREAT [8] using default parameters and a binomial fold enrichment cutoff ≥ 1.5. GREAT results were filtered for "placenta" and "trophoblast", using only GO Biological Process, MGI Phenotype, and MGI Phenotype Single Knockout (KO) ontologies. The MGI Phenotype Single KO ontology is a version of the MGI Phenotype ontology that only includes single mutant gene to phenotype associations. Each motif was assigned a q-value, based on the best q-value for all placenta terms. Motifs were then sorted by these q-values.

### Clustering similar motifs

Similar Position Weight Matricies (PWMs) were grouped in order to remove redundant binding site predictions as described previously [41]. The similarity of two motifs was defined as the maximum pairwise alignment score achieved when all alignments of the two motifs were assessed by shifting the motifs relative to

each other for both orientations of the motifs. The alignment score was defined as the sum of column scores for all aligning columns normalized by the geometric mean of the self-alignment of each motif to itself. The column scoring function used was:

(1) $\Pr(\text{Match}) = f_{A1} * f_{A2} + f_{C1} * f_{C2} + f_{G1} * f_{G2} + f_{T1} * f_{T2}$

(2) $\text{matchScore} = \Pr(\text{Match}) - (1 - \Pr(\text{Match}))/3$

The column score function attempts to capture the probability that a specific column will select the same base and is then penalized by the chance that it will not select the same base. A similarity threshold of 0.85 was used.

## Automated assessment of enrichment of known TFs

To generate a list of known placenta TFs using an automated approach, we combined gene lists from GREAT for the most general placenta terms, abnormal placenta morphology (GO Biological Process ontology), and placenta development (Mouse Phenotype Single KO ontology). This resulted in a list of 349 genes. We then associated PWMs with gene names to classify each group of similar PWMs as 'known' or 'unknown' depending on whether any single PWM in a group of similar motifs mapped to the list of 349 genes. After each motif was classified using this method, we determined whether a significant number of TFs in the top 50 appeared in the known placenta gene list. We did so by comparing the number of 'known' TFs in the top 50 to the number of TFs that appear in the known placenta gene list when 50 random TFs (below rank 100) were chosen from the list, a total of 1,000 times. We also calculated a Wilcoxon rank-sum p-value to determine if known TFs were enriched toward the top of the list. For each of the other tissues analyzed, we generated a list of known TFs using the same approach, but with the following GREAT terms: heart: heart development, abnormal heart development; pancreas: pancreas development, abnormal pancreas development; blood vessel: blood vessel development, abnormal blood vessel morphology; bone marrow: abnormal bone marrow cell morphology/development; liver: liver development, abnormal liver size.

## Predicted versus known placenta TFs

To more carefully determine if the top 50 TFs in our list were known to be involved in placenta development, we searched the literature. For a TF to be considered 'known' or 'well-studied' in the placenta, the literature must show strong experimental evidence, such as placenta abnormalities upon gene knockout, or defects in trophoblast function upon gene knockout/knockdown/overexpression in relevant placenta cell lines. TFs that have structurally similar family members involved in placenta development for which we do not have motifs were also considered 'known'. Additionally, TFs with PWMs that closely resemble PWMs that map to genes involved in placenta development, but were just below the 0.85 grouping threshold, were considered 'known'. For a TF to be considered 'predicted', single studies may have implicated the TF in placenta development, but the relationship has not been well characterized. Additionally, if the only evidence for a TF's involvement in placenta was gene expression, the TF was considered 'predicted'.

## Identifying potential placenta enhancers (placenta TFBS clusters)

To identify potential placenta enhancers, we used a hierarchical clustering approach (UPGMA) over the genome to search for binding site predictions in close proximity of each other for the 50 TFs described above. We first placed each binding site prediction in its own TFBS cluster, mapped to its own centroid. We then iteratively agglomerated the two TFBS clusters with the nearest centroids. Each TFBS cluster was scored based on the number of non-overlapping binding site predictions (TFBS that share $\leq$3 bp) falling within it. To reward TFBS density, the scores for regions longer than 250 bp were weighted by a penalty function: $exp(-0.5(regionLength-250)^2/250^2)$. TFBS clusters were ranked by score, and then the ranked list was traversed, outputting only those TFBS clusters that did not overlap a previously output TFBS cluster. We discarded TFBS clusters that overlapped with exons.

## Filtering GREAT results

GREAT analysis was performed using default GREAT filters for significant terms: region-based fold enrichment $\geq$2 and false discovery rate (FDR) q-value $\leq$0.05, with the additional requirement that at least 25 genes in the term were hit. Unless specifically noted, GREAT results were filtered by fold enrichment.

## TSC and TGC cell culture

TSCs (a kind gift from Dr. Emin Maltepe at UCSF) were grown, differentiated, and passaged according to standard protocols [42]. Passage 2 mouse embryonic fibroblasts (MEFs) (Applied Stem Cell) were expanded and treated with Mytomycin C as previously described [42], aliquoted and frozen to use as feeder cells for TSCs. TSCs were split once a week at a 1:50 dilution onto a plate of fresh MEFs. For differentiation into trophoblast giant cells (TGCs), a 1:10 dilution of confluent TSCs was plated onto a 10 cm plate. Five days later, the differentiating TSCs were split into a 24-well plate at a 1:8 dilution.

## Cloning

Inserts were amplified from mouse genomic DNA (Clonetech Laboratories, Inc.) using Phusion High Fidelity DNA Polymerase (NEB, Inc.) and cloned into the 5′ KpnI and 3′ HindIII sites of pGL4.23 (Promega, Corp.). A second reporter vector was constructed, pGL4.23 LIC, by introducing a Ligation Independent Cloning (LIC) linker into the 5′KpnI and 3′ HindIII sites of pGL4.23. The LIC forward site was: 5′-cGCTCTTCGGGATGGAGGGATATCCACCTTACCCGAAGAGCa-3′ and the LIC reverse site was: 5′-agcttGCTCTTCGGGTAAGGTGGATATCCCTCCATCCCGAAGAGCggtac-3′.

The genomic inserts were cloned into the pGL4.23 LIC vector using an LIC method described previously [43]. All positive clones were identified by colony PCR and sequenced. Primers used to amplify genomic regions are listed in Supplementary Table S4.

## TSC/TGC transfection assays

Transfections were done according to the Invitrogen protocol for Lipofectamine LTX & Plus reagent. For TSCs, a confluent 10 cm plate was split 1:4 into a 24 well plate. We used a 1 μg:4 μl ratio of DNA to reagent, and transfected 1 μg of reporter construct and 20 ng of pRL-TK vector (used as a transfection efficiency control vector, Promega Corp.) per well. Cells were lysed 24 hours post-transfection and frozen until luciferase assays were performed. For TGCs, plates were transfected 12 days after starting differentiation. We used a 1 μg:3 μl ratio of DNA to reagent, and transfected 750 ng reporter construct and 15 ng of pRL-TK vector (used as a control vector, Promega Corp.). Cells were lysed 48 hours post-transfection and frozen until luciferase assays were performed. Each candidate was tested in triplicate within a single plate (technical replicates), and on at least 3 different days (biological replicates).

### Neocortex nucleofection assays

Primary neocortical cells were prepared from e14.5 mice and transfected with pGL4.23 or pGL4.23 LIC containing the genomic regions and pRL-CMV (as a control vector) using the Amaxa 96-well Shuttle Protocol for Primary Mammalian Neurons (Lonza) [25]. We used $2.5\times10^5$ cells, 100 ng plasmid DNA, and 60 ng pRL-CMV per nucleofection sample. Transfected cells were resuspended in 120 µl supplemented PNBM media (Lonza) and plated on a 96-well plate treated with Poly-D-Lysine. 40 µl of this suspension were added to 160 µl PNBM per well. Cells were lysed 48 hours post-transfection and frozen until luciferase assays were performed.

### Luciferase assays

Luciferase assays were done using the DLR kit (Promega) according to manufacturer's instructions and read using a Promega Glomax luminometer using the "Dual-Luciferase 2 injectors" program with a 50 µl injection volume for both LAR II and Stop & Glo Reagent.

### Mouse placenta data

Mouse placenta data from [4] was downloaded from the Ren Lab website: http://chromosome.sdsc.edu/mouse/download.html

We took the union of the regions in the placenta.enhancer.txt file with the regions in the placenta.h3k27ac.peak.txt file, after padding each given coordinate by $\pm1500$ bp, as recommended by the Ren lab. We then removed regions within 1 kb of gene transcriptional start sites, resulting in 70,951 peaks. Placenta-specific regions were downloaded from the same website, and were similarly padded by $\pm1500$ bp.

### Mouse ChIP-Seq data from 18 tissues

Mouse tissue data from [4] was downloaded from the Ren Lab website: http://chromosome.sdsc.edu/mouse/download.html

We downloaded enhancer files for 18 tissues and cell types, and padded each given coordinate by $\pm1500$ bp before determining the overlap with the placenta TFBS clusters. We analyzed the following tissues and cell types: cortex, MEFs, bone marrow, cerebellum, e14.5 liver, e14.5 brain, e14.5 heart, e14.5 limb, liver, heart, intestine, kidney, spleen, lung, mESCs, olfactory bulb, testes, and thymus.

### Human placenta data

DNAse I hypersensitive sites in human placenta were provided by the Stamatoyannopoulos lab [2,3]. Data from 6 samples were provided, aged at 113 days gestation, 108 days gestation, 105 days gestation, 91 days gestation, and two at 85 days gestation. Peaks from samples below 100 days old were intersected and peaks from replicates above 100 days old were intersected. The union of the two sets was then taken and converted to mm9 coordinates using UCSC's liftover tool with default parameters. We then removed regions within 1 kb of gene transcriptional start sites, resulting in 80,922 peaks.

### TSC RNA-Seq and ChIP-Seq data

TSC data are from [22] and were downloaded from GEO. RNA-Seq data were ranked according to average tag count between biological replicates, normalized to the 3′ UTR length for reported genes. TSC H3K27ac and H3K4me1 data were intersected to generate a TSC putative enhancer set. For each set of peaks we used (H3K27ac, H3K4me1, H3K27me3, H3K9me3), regions within 1 kb of gene TSS were removed.

## Supporting Information

**Figure S1 Enrichment of known placenta TFs.** Gray distribution shows the number of TFs that are annotated by GREAT as being involved in placenta development when 50 random TFs (below rank 100; See Figure 1B) are chosen from the ranked list a total of 1,000 times. Our chosen list of top 50 (black arrow) has a p-value of 0.013. Overall, the ranked list is also enriched for placenta TFs towards the top, as indicated by a Wilcoxon rank-sum test.
(PDF)

**Figure S2 Sensitivity analysis for thresholds used to identify TFBS clusters.** Each panel shows the GREAT fold enrichment for a representative GREAT term versus the number of TFBS clusters identified at various thresholds. In the top panel, we vary the number of top TFs used to obtain predictions for binding site clustering. In the middle panel, using the top 50 TFs, we varied the minimal number of non-overlapping placenta TFBS required to call a placenta TFBS cluster. In the bottom panel, we allowed a mixed TFBS threshold. The label above each point in the bottom panel indicates the number of non-overlapping TFBS required, at least 3 of which are from the top 50 TFs (placental TFs). In each panel, the threshold that we chose, based on wanting to maximize GREAT fold enrichment for placenta terms while keeping the number of TFBS clusters identified relatively high, is circled in red.
(PDF)

**Figure S3 Details on TF composition of placenta TFBS clusters.** Top panel shows the % of the total predictions we started with for each TF (before clustering) that end up in placenta TFBS clusters. Middle panel (left) shows the total number of TFBS in each cluster, and (right) the number of unique TF motifs (appearing once or more) in each cluster. Bottom panel shows that predictions for each of the top 50 TFs occurs frequently within the placenta TFBS clusters, regardless of whether they have a known role or a predicted role in placenta development. Predicted placenta TFs with the most binding site predictions have high information content, indicating predictions are not due to weak motif matches to the genome.
(PDF)

**Figure S4 Differentiation of mouse trophoblast stem cells into trophoblast giant cells.** TSCs that were isolated from early stages of mouse development were provided by Emin Maltepe (UCSF). We show that we can differentiate the TSCs into TGCs by removal of FGF4 and Heparin from the media (left), using protocols described previously [23]. PL-II is known to be specific to TGCs, and an enhancer region for the gene has been shown to be active in rat trophoblast giant cell-like cells [24]. Therefore, we demonstrated the purity of our cell populations by transfecting the corresponding mouse (*mPL-II*) enhancer region in both of our cell types. We see activity is consistently high in our TGCs, and low in our TSCs (right). Error bars represent the standard deviation of the mean. Significance was calculated using the unpaired *t*-test. We always used the *mPL-II* enhancer as a control in TSC and TGC transfections.
(PDF)

**Figure S5 Placenta TFBS clusters activity over empty vector.** Chart showing enhancer activity relative to empty vector for all 36 candidates in TSCs, TGCs, and neocortical cells. Error bars represent the standard deviation of the mean. The x-axis shows predicted target genes of the placenta TFBS clusters that were tested for enhancer activity.
(PDF)

**Figure S6 Activity over empty vector for negative controls.** Top panel shows negative controls that were length, GC, and conservation matched to placenta TFBS clusters. Activity is shown for TSCs and TGCs. Bottom panel shows enhancers that are active in neocortical cells. Activity is shown for TSCs, TGCs, and neocortical cells. Error bars represent the standard deviation of the mean.
(PDF)

**Figure S7 Comparing activity of placenta TFBS clusters in TSCs vs. neocortex and TGCs vs. neocortex.** When comparing activity of placenta TFBS clusters in TSCs (top panel) and TGCs (middle panel) to neocortex, we see that TFBS clusters generally have higher activity in placental cell types. Candidates with $\geq$2-fold activity are considered active, and those with $<$2-fold activity are considered inactive. Very few candidates have higher activity in neocortex. Basal promoter activity (bottom panel), as measured by activity (luciferase/renilla activity) of the empty vector, is low in each of the cell types tested.
(PDF)

**Figure S8 Comparing placenta TFBS clusters to putative enhancers in 18 other tissues and cell types.** We obtained putative enhancer data from [4], where each tissue has on average 50,166 putative enhancers. For each placenta TFBS cluster we determined how many other tissues have overlapping putative enhancers. 50% of placenta TFBS clusters overlap with putative enhancers in $\leq$2 other tissues.
(PDF)

**Figure S9 Overlap with mouse TSC ChIP-Seq data.** ChIP-Seq data from [22] were compared to the placenta TFBS clusters. Significantly more TFBS clusters overlap with TSC enhancer data compared to random regions of the genome (top panel) whereas significantly less TFBS clusters overlap with TSC repressor data compared to random regions of the genome (middle and bottom panels).
(PDF)

**Figure S10 Enrichment of top 50 TFs in 5 additional tissues tested.** For each panel, the gray distribution shows the number of TFs that are annotated as being involved in development of the tissue when 50 random TFs (below rank 100) are chosen from the ranked list for each tissue a total of 1,000 times. Arrow points to the number of TFs in the top 50 that are annotated as being involved in development of the tissue, along with matching p-value. Wilcoxon rank-sum p-value is also shown at the bottom of each panel.
(PDF)

**Table S1 Top 50 TFs.** The motif name and most likely TF for the top 50 motifs ranked by their association with a placenta term is shown. Each TF is categorized as known to have a role in placenta development, or predicted to have a role in placenta development based on the PMIDs (PubMed Identifiers) provided.
(XLSX)

**Table S2 Expression levels of predicted TFs.** Human expression data was obtained from [21] and ranked according to Fragments Per Kilobase of transcript per Million mapped reads (FPKM) values provided. TSC data was obtained from [22], and ranked according to average tag count between biological replicates, normalized to the 3′ UTR length for reported genes. It was then determined which quartile of expression each predicted TF was in, where 1 is the top quartile and 4 is the bottom quartile. 15 out of 17 predicted TFs in Figure 1B were within the top two quartiles and considered expressed in at least one of the 4 data sets.
(DOCX)

**Table S3 Placenta TFBS clusters.** For each of the final set of 2,216 placenta TFBS clusters, the chromosome, start coordinate, end coordinate, and number of TFBS in each placenta TFBS cluster are listed.
(XLSX)

**Table S4 List of primers used in to clone candidate regions.** For each region that was tested, forward and reverse primers, as well as the vector that was used are listed. Sequences listed are for mm9.
(XLSX)

**Table S5 Genes with the most significant number of placenta TFBS clusters in their regulatory domains.** Genes are ranked by q-value, which indicates the likelihood associated with the observed number of placenta TFBS clusters per gene versus the length of the individual gene's regulatory domain. Genes in bold have well characterized roles in placenta development. PMID: PubMed Identifier.
(DOCX)

**Table S6 GREAT results for placenta TFBS clusters that do not overlap experimentally annotated placenta enhancers.** Placenta TFBS clusters that do not overlap the mouse or human experimentally annotated placenta enhancers are still enriched for placenta functions in GREAT.
(DOCX)

**Table S7 GREAT comparisons for placenta TFBS clusters and putative placenta-specific enhancers from mouse experimental data.** GREAT fold enrichments for various placenta terms are shown for the placenta TFBS clusters (column 3) and mouse placenta-specific regions (column 4). The placenta TFBS clusters have higher fold enrichment in 5 out of 6 terms. Column 5 shows that the fraction of unique placenta TFBS clusters associated with a placenta term was higher for 5 out of 6 terms than the fraction of unique mouse placenta-specific regions associated with the same placenta term.
(DOCX)

**Table S8 Heart TFs and TFBS clusters.** For each of the final set of heart TFBS clusters, the chromosome, start coordinate, end coordinate, and number of TFBS in each heart TFBS cluster are listed. The top 50 motifs and TF names are also shown.
(XLSX)

**Table S9 Pancreas TFs and TFBS clusters.** For each of the final set of pancreas TFBS clusters, the chromosome, start coordinate, end coordinate, and number of TFBS in each pancreas TFBS cluster are listed. The top 50 motifs and TF names are also shown.
(XLSX)

**Table S10 Blood vessel TFs and TFBS clusters.** For each of the final set of blood vessel TFBS clusters, the chromosome, start coordinate, end coordinate, and number of TFBS in each blood vessel TFBS cluster are listed. The top 50 motifs and TF names are also shown.
(XLSX)

**Table S11 Bone marrow TFs and TFBS clusters.** For each of the final set of bone marrow TFBS clusters, the chromosome, start coordinate, end coordinate, and number of TFBS in each bone marrow TFBS cluster are listed. The top 50 motifs and TF names are also shown.
(XLSX)

**Table S12 Liver TFs and TFBS clusters.** For each of the final set of liver TFBS clusters, the chromosome, start coordinate,

end coordinate, and number of TFBS in each liver TFBS cluster are listed. The top 50 motifs and TF names are also shown.
(XLSX)

**Table S13 Gene counts associated with terms used to identify TFs relevant for various tissues.** The primary term used to identify TF enrichment for each tissue is shown in column 2, and column 3 shows the number of genes annotated with the term. For bone barrow, blood vessel, placenta, and liver, the MGI Mouse Phenotype ontology is shown, and for heart and pancreas the GO Biological Process term is shown. The term used for pancreas has a much lower gene count than other tissues investigated.
(DOCX)

## Acknowledgments

We thank Emin Maltepe (UCSF) for providing us with TSCs and for advice on cell culture, Julie Baker and her lab (Roberta Hannibal and Edward Chuong) for advice on cell culture, Feng Yue (Bing Ren Lab) for advice on mouse placenta data processing, Bob Thurman (John Stamatoyannopoulos Lab) for human placenta DNase-Seq data, Michael Hiller for the MGI Phenotype Single KO ontology, Harendra Guturu for the tool to cluster similar motifs and for manuscript advice, J. Gray Camp for manuscript advice, and the Bejerano Lab members for data analysis advice.

## Author Contributions

Conceived and designed the experiments: GT GB. Performed the experiments: GT KBM TC. Analyzed the data: GT JC. Contributed reagents/materials/analysis tools: AMW. Wrote the paper: GT GB.

## References

1. Spitz F, Furlong EE (2012) Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 13: 613–626. doi:10.1038/nrg3207.
2. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. Science 337: 1190–1195. doi:10.1126/science.1222794.
3. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, et al. (2012) The accessible chromatin landscape of the human genome. Nature 489: 75–82. doi:10.1038/nature11232.
4. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, et al. (2012) A map of the cis-regulatory sequences in the mouse genome. Nature 488: 116–120. doi:10.1038/nature11243.
5. Hardison RC, Taylor J (2012) Genomic approaches towards finding cis-regulatory modules in animals. Nat Rev Genet 13: 469–483. doi:10.1038/nrg3242.
6. Yanez-Cuna JO, Kvon EZ, Stark A (2013) Deciphering the transcriptional cis-regulatory code. Trends Genet 29: 11–22. doi:10.1016/j.tig.2012.09.007.
7. Wenger AM, Clarke SL, Guturu H, Chen J, Schaar BT, et al. (2013) PRISM offers a comprehensive genomic approach to transcription factor function prediction. Genome Res 23: 889–904. doi:10.1101/gr.139071.112.
8. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, et al. (2010) GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28: 495–501. doi:10.1038/nbt.1630.
9. Cross JC, Werb Z, Fisher SJ (1994) Implantation and the placenta: key pieces of the development puzzle. Science 266: 1508–1518.
10. John R, Hemberger M (2012) A placenta for life. Reprod Biomed Online 25: 5–11. doi:10.1016/j.rbmo.2012.03.018.
11. Tuteja G, Cheng E, Papadakis H, Bejerano G (2012) PESNPdb: a comprehensive database of SNPs studied in association with pre-eclampsia. Placenta 33: 1055–1057. doi:10.1016/j.placenta.2012.09.016.
12. Newburger DE, Bulyk ML (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res 37: D77–82. doi:10.1093/nar/gkn660.
13. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, et al. (2012) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res 38: D105–10. doi:10.1093/nar/gkp950.
14. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34: D108–10. doi:10.1093/nar/gkj143.
15. Oka C, Nakano T, Wakeham A, de la Pompa JL, Mori C, et al. (1995) Disruption of the mouse RBP-J kappa gene results in early embryonic death. Development 121: 3291–3301.
16. Guillemot F, Nagy A, Auerbach A, Rossant J, Joyner AL (1994) Essential role of Mash-2 in extraembryonic development. Nature 371: 333–336. doi:10.1038/371333a0.
17. Yamamoto H, Flannery ML, Kupriyanov S, Pearce J, McKercher SR, et al. (1998) Defective trophoblast function in mice with a targeted mutation of Ets2. Genes Dev 12: 1315–1326.
18. Wen F, Tynan JA, Cecena G, Williams R, Munera J, et al. (2007) Ets2 is required for trophoblast stem cell self-renewal. Dev Biol 312: 284–299. doi:10.1016/j.ydbio.2007.09.024.
19. Ferdous A, Morris J, Abedin MJ, Collins S, Richardson JA, et al. (2011) Forkhead factor FoxO1 is essential for placental morphogenesis in the developing embryo. Proc Natl Acad Sci U S 108: 16307–16312. doi:10.1073/pnas.1107341108.
20. Schorpp-Kistner M, Wang ZQ, Angel P, Wagner EF (1999) JunB is essential for mammalian placentation. EMBO J 18: 934–948. doi:10.1093/emboj/18.4.934.
21. Kim J, Zhao K, Jiang P, Lu Z, Wang J, et al. (2012) Transcriptome landscape of the human placenta. BMC Genomics 13: 115. doi:10.1186/1471-2164-13-115.
22. Chuong EB, Rumi MAK, Soares MJ, Baker JC (2013) Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat Genet 45: 325–329. doi:10.1038/ng.2553.
23. Tanaka S, Kunath T, Hadjantonakis AK, Nagy A, Rossant J (1998) Promotion of trophoblast stem cell proliferation by FGF4. Science 282: 2072–2075.
24. Sun Y, Duckworth ML (1999) Identification of a placental-specific enhancer in the rat placental lactogen II gene that contains binding sites for members of the Ets and AP-1 (activator protein 1) families of transcription factors. Mol Endocrinol 13: 385–399.
25. Wenger AM, Clarke SL, Notwell JH, Chung T, Tuteja G, et al. (2013) The Enhancer Landscape during Early Neocortical Development Reveals Patterns of Dense Regulation and Co-option. PLoS Genet 9: e1003728. doi:10.1371/journal.pgen.1003728.
26. Riley P, Anson-Cartwright L, Cross JC (1998) The Hand1 bHLH transcription factor is essential for placentation and cardiac morphogenesis. Nat Genet 18: 271–275. doi:10.1038/ng0398-271.
27. Duarte A, Hirashima M, Benedito R, Trindade A, Diniz P, et al. (2004) Dosage-sensitive requirement for mouse Dll4 in artery development. Genes Dev 18: 2474–2478. doi:10.1101/gad.1239004.
28. Gale NW, Dominguez MG, Noguera I, Pan L, Hughes V, et al. (2004) Haploinsufficiency of delta-like 4 ligand results in embryonic lethality due to major defects in arterial and vascular development. Proc Natl Acad Sci U S 101: 15949–15954. doi:10.1073/pnas.0407290101.
29. Montavon T, Soshnikova N, Mascrez B, Joye E, Thevenet L, et al. (2011) A regulatory archipelago controls Hox genes transcription in digits. Cell 147: 1132–1145. doi:10.1016/j.cell.2011.10.023.
30. Jeong Y, El-Jaick K, Roessler E, Muenke M, Epstein DJ (2006) A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. Dev Camb Engl 133: 761–772. doi:10.1242/dev.02239.
31. Werner T, Hammer A, Wahlbuhl M, Bösl MR, Wegner M (2007) Multiple conserved regulatory elements with overlapping functions determine Sox10 expression in mouse embryogenesis. Nucleic Acids Res 35: 6526–6538. doi:10.1093/nar/gkm727.
32. Frankel N (2012) Multiple layers of complexity in cis-regulatory regions of developmental genes. Dev Dyn Off Publ Am Assoc Anat 241: 1857–1866. doi:10.1002/dvdy.23871.
33. Dave VP, Allman D, Keefe R, Hardy RR, Kappes DJ (1998) HD mice: a novel mouse mutant with a specific defect in the generation of CD4(+) T cells. Proc Natl Acad Sci U S 95: 8187–8192.
34. Chelbi ST, Wilson ML, Veillard AC, Ingles SA, Zhang J, et al. (2012) Genetic and epigenetic mechanisms collaborate to control SERPINA3 expression and its association with placental diseases. Hum Mol Genet 21: 1968–1978. doi:10.1093/hmg/dds006.
35. Fan X, Ren P, Dhal S, Bejerano G, Goodman SB, et al. (2011) Noninvasive monitoring of placenta-specific transgene expression by bioluminescence imaging. PloS One 6: e16348. doi:10.1371/journal.pone.0016348.
36. Fan X, Petitt M, Gamboa M, Huang M, Dhal S, et al. (2012) Transient, inducible, placenta-specific gene expression in mice. Endocrinology 153: 5637–5644. doi:10.1210/en.2012-1556.
37. Cross JC, Baczyk D, Dobric N, Hemberger M, Hughes M, et al. (2003) Genes, development and evolution of the placenta. Placenta 24: 123–130.
38. Cox B, Kotlyar M, Evangelou AI, Ignatchenko V, Ignatchenko A, et al. (2009) Comparative systems biology of human and mouse as a tool to guide the modeling of human placental pathology. Mol Syst Biol 5: 279. doi:10.1038/msb.2009.37.
39. Cheng Y, King DC, Dore LC, Zhang X, Zhou Y, et al. (2008) Transcriptional enhancement by GATA1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif. Genome Res 18: 1896–1905. doi:10.1101/gr.083089.108.
40. Rada-Iglesias A, Bajpai R, Prescott S, Brugmann SA, Swigut T, et al. (2012) Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. Cell Stem Cell 11: 633–648. doi:10.1016/j.stem.2012.07.006.
41. Guturu H, Doxey AC, Wenger AM, Bejerano G (2013) Structure-aided prediction of mammalian transcription factor complexes in conserved non-coding elements. Philos Trans R Soc Lond B Biol Sci 368: 20130029. doi:10.1098/rstb.2013.0029.

42. Himeno E, Tanaka S, Kunath T (2008) Isolation and manipulation of mouse trophoblast stem cells. Curr Protoc Stem Cell Biol Chapter 1: Unit 1E 4. doi:10.1002/9780470151808.sc01e04s7.

43. Du R, Li S, Zhang X (2011) A modified plasmid vector pCMV-3Tag-LIC for rapid, reliable, ligation-independent cloning of polymerase chain reaction products. Anal Biochem 408: 357–359. doi:10.1016/j.ab.2010.08.042.