ChIP-seq Accurately Predicts Tissue-Specific Activity of Enhancers

Axel Visel[*,1], Matthew J. Blow[*,1,2], Zirong Li[3], Tao Zhang[2], Jennifer A. Akiyama[1], Amy Holt[1], Ingrid Plajzer-Frick[1], Malak Shoukry[1], Crystal Wright[2], Feng Chen[2], Veena Afzal[1], Bing Ren[3], Edward M. Rubin[1,2], Len A. Pennacchio[**,1,2]

[1] Genomics Division, MS 84-171, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 USA.

[2] U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598 USA.

[3] Ludwig Institute for Cancer Research, University of California San Diego (UCSD) School of Medicine, La Jolla, CA 92093 USA.

* These authors contributed equally to this work.

** To whom correspondence should be addressed: Len A. Pennacchio, Genomics Division, One Cyclotron Road, MS 84-171, Lawrence Berkeley National Laboratory, Berkeley, CA 94720. Email: LAPennacchio@lbl.gov, Phone: (510) 486-7498, Fax: (510) 486-4229.

Summary

A major yet unresolved quest in decoding the human genome is the identification of the regulatory sequences that control the spatial and temporal expression of genes. Distant-acting transcriptional enhancers are particularly challenging to uncover since they are scattered amongst the vast non-coding portion of the genome. Evolutionary sequence constraint can facilitate the discovery of enhancers, but fails to predict when and where they are active *in vivo*. Here, we performed chromatin immunoprecipitation with the enhancer-associated protein p300, followed by massively-parallel sequencing, to map several thousand *in vivo* binding sites of p300 in mouse embryonic forebrain, midbrain, and limb tissue. We tested 86 of these sequences in a transgenic mouse assay, which in nearly all cases revealed reproducible enhancer activity in those tissues predicted by p300 binding. Our results indicate that *in vivo* mapping of p300 binding is a highly accurate means for identifying enhancers and their associated activities and suggest that such datasets will be useful to study the role of tissue-specific enhancers in human biology and disease on a genome-wide scale.

The initial sequencing of the human genome [1,2], complemented by effective computational and experimental strategies for mammalian gene discovery [3,4], have resulted in a virtually complete list of protein-coding sequences. In contrast, the genomic location and function of regulatory elements that orchestrate gene expression in the developing and adult body remains more obscure, hindering studies of their contribution to developmental processes and human disease. Evolutionary constraint of non-coding sequences can predict the location of enhancers in the genome [5-12], but does not reveal when and where these enhancers are active *in vivo*. Furthermore, it has been suggested that a substantial proportion of regulatory elements is not sufficiently conserved to be detectable by comparative genomic methods [13-16].

Chromatin immunoprecipitation coupled to massively-parallel sequencing (ChIP-seq) has been shown to enable genome-wide mapping of protein binding and epigenetic marks [17-22]. The ChIP-seq approach is dependent on the cross-linking of proteins to specific DNA elements, followed by antibody enrichment of the protein:DNA complexes, and high-throughput sequencing of the recovered DNA fragments. In principle, ChIP-seq using an antibody specific for an enhancer-binding protein could provide a conservation-independent approach for the identification of candidate enhancer sequences.

The acetyltransferase and transcriptional coactivator p300 is a near-ubiquitously expressed component of enhancer-associated protein assemblies and is critically required for embryonic development [23-27]. In homogeneous cell preparations, p300 has been shown to be associated with enhancers [28,29], but these *in vitro* studies provided access only to subsets of enhancers that are active in a given cell type under culture conditions, providing limited insight into their *in vivo* function. In the present study, we have determined the genome-wide

occupancy of p300 in forebrain, midbrain, and limb tissue isolated directly from developing mouse embryos. Using a transgenic mouse reporter assay, we show that p300 binding in these embryonic tissues predicts with high accuracy not only where enhancers are located in the genome, but also in what tissues they are active *in vivo*. Depending on tissue type, the success rate of predicting forebrain, midbrain, and limb enhancers was between 5- and 16-fold increased compared to previous studies in which such enhancers were discovered by comparative genomics [10,11].

Genome-wide Mapping of p300 in Tissues

To generate genome-wide maps of p300 binding *in vivo*, we microdissected forebrain, midbrain, and limb tissue from over 150 embryonic day 11.5 (e11.5) mouse embryos and performed ChIP with an antibody for p300 directly from these tissue samples (Fig. 1). Immunoprecipitated DNA fragments were analyzed using massively-parallel sequencing and the resulting 36bp sequence reads were aligned to the reference mouse genome [17,30].

After appropriate quality filtering, between 2.4 and 3.6 million aligned reads obtained from each of the tissue samples were used to identify regions of the genome with significant enrichment in p300-associated DNA sequences, hereafter referred to as 'peaks' owing to their appearance in genome-wide density plots [17] (Suppl. Table 1). Using an estimated false-discovery rate (FDR) threshold of < 0.01, we identified 2543, 561 and 2105 peaks from forebrain, midbrain and limb respectively (Suppl. Tables 2-4). The majority of peaks were located at least 10kb from transcript start sites (Suppl. Fig. 1). The smaller number of peaks from midbrain is likely due to variability in the efficiency of enrichment by ChIP (Suppl. Fig. 2). Re-sampling of subsets of data suggests that the major p300-binding sites from these

three tissues have been discovered, whereas with increased sequencing coverage it is anticipated that additional binding sites can be identified that are occupied only in smaller subsets of cells within each tissue (Suppl. Fig. 3). While the majority of genomic regions with *in vivo* p300 binding were identified by peaks in a single tissue, there were 386 regions at which peaks were observed in two tissues, and 21 regions at which p300 peaks were observed in all three tissues (Suppl. Fig. 4).

p300 Predicts Enhancer Activity Patterns

To directly test if p300 binding in developing mouse tissues is indicative of enhancer activity *in vivo*, we selected 86 regions with a p300 peak in at least one of the tissues for analysis in transgenic mice, comprising a total of 122 individual predictions of enhancer activity in specific tissues (Suppl. Table 5). These elements were selected blind to the identity of genes near which they are located, exhibited a wide range of evolutionary conservation with other vertebrate species (see Methods) and approximately reflect the genome-wide distribution properties of p300 peaks among intronic and intergenic regions, as well as their distances relative to known genes (Suppl. Fig. 1).

We cloned the human genomic sequences orthologous to these enhancer candidate regions into an enhancer reporter vector and generated transgenic mice as previously described [10,31]. For each of the 86 candidate enhancers, several independent transgenic embryos (average: n=8) were assessed for reproducible reporter gene expression. A pattern was considered reproducible if the same anatomical structure was stained in three or more embryos. In almost all cases, this minimum threshold was exceeded and reproducible

reporter staining in forebrain, midbrain or limb was on average present in over 80% of the embryos obtained per construct (Suppl. Table 5).

First, we determined whether p300 binding was predictive of reproducible *in vivo* enhancer activity regardless of their tissue specificity. Considering peaks from each of the three p300 datasets separately, 55 of 63 (87%) forebrain predictions, 30 of 34 (88%) midbrain predictions and 22 of 25 (88%) limb predictions were active enhancers *in vivo* at e11.5 as defined by reproducible LacZ staining (Fig 2, gray + colored bars). Overall, 87% (75 of 86) of the tested elements were reproducible enhancers at e11.5. This compares with a success rate for predicting enhancers of 47% (246 of 528) from our previous studies in which elements were identified based on their extreme evolutionary conservation and tested using the same transgenic mouse assay [10,11]. Thus, the rate of false-positive predictions using p300 ChIP-seq was more than 4-fold lower than with extreme evolutionary conservation (13% compared to 53% previously; $P = 4.2e\text{-}10$, Fisher's Exact test).

We next determined the accuracy with which p300 binding predicts the tissue in which enhancer activity will occur. Of the 63 tested elements that overlapped a forebrain p300 peak, 49 (78%) were found to have reproducible enhancer activity in the developing forebrain (Fig. 2, blue). Likewise, 28 of 34 (82%) tested elements identified by midbrain p300 enrichment (Fig. 2, red) and 20 of 25 (80%) tested elements identified by limb p300 enrichment (Fig. 2, green) were confirmed to be active in the predicted tissue. The 86 tested elements included 32 sequences that were identified by p300 binding in more than one tissue. Of these, 27 of 32 (84%) were found to be active in at least one of the predicted tissues while 22 sequences (69%) perfectly recapitulated the predicted expression patterns (Suppl. Table 6).

To assess the degree of enrichment of enhancer activities in predicted tissues, we compared the relative frequency of enhancers for each of the three tissues examined here with a background set of 528 previously tested sequences predicted to be developmental enhancers based on extreme sequence constraint that were not associated with *a priori* tissue specificity predictions [10,11]. For example, whereas forebrain enhancers account for only 16% (86 of 528) of the tested elements identified through comparative approaches, 78% (49 of 63) of elements predicted by forebrain p300 peaks were found to be active enhancers in the forebrain (Fig. 2). Forebrain predictions are therefore 5-fold enriched in forebrain enhancers compared with enhancers identified through comparative approaches ($P < 1e-22$). Similarly, we observed a 6-fold enrichment of midbrain enhancers ($P < 1e-11$) and a 16-fold enrichment of limb enhancers ($P < 1e-18$) at midbrain and limb p300 peaks, respectively. Representative examples of enhancers identified by ChIP-seq are shown in Fig. 3 and detailed annotations and reproducibility across transgenic mice for all elements tested in this study can be found at http://enhancer.lbl.gov [32]. Taken together, these results indicate that p300 peaks are a highly accurate predictor of *in vivo* enhancers and their spatial activity patterns.

Most p300-Bound Regions are Conserved

Previous studies have indicated a positive correlation between enhancer activity during development and non-coding sequence conservation [6,8-11,33], but it has also been suggested that not all regulatory elements in vertebrate genomes are under detectable evolutionary constraint [13-16]. To test if p300-binding in e11.5 tissues is generally associated with evolutionarily constrained non-coding sequences, we determined if ChIP-seq reads are

overall enriched at previously identified extremely conserved non-coding sequences [9,11]. We observed strong enrichment of p300 ChIP-seq reads at these conserved sequences, but not at random sites or exons (Fig. 4, Suppl. Table 7). Vice versa, between 86% and 91% of the significant p300 peaks overlap sequences that are under evolutionary constraint in vertebrates [34], compared to less than 30% of size-matched random regions (P < 1e-172, Fisher's Exact test, Suppl. Fig. 5). Using a more stringent constraint threshold score, we observed that between 10% and 21% of peaks are highly constrained, compared to 1% of random regions (P < 1e-82). These results indicate that the majority of p300 peaks in the investigated tissues are under significant evolutionary constraint and support a global enrichment of p300 in highly conserved non-coding regions of the genome previously correlated with developmental enhancers.

Correlation with Gene Expression Patterns

To examine the correlation of p300-enriched regions in embryonic tissues with the transcriptional regulation of neighboring genes, we compared the genomic distribution of p300 peaks in e11.5 forebrain with gene expression data from this tissue. Using high-density microarrays, we identified a set of 885 genes that are over-expressed in forebrain at e11.5 compared to whole embryos (Suppl. Table 8). When we compared the genomic position of these forebrain genes to the genome-wide distribution of 2,453 forebrain-derived p300 peaks, we observed that the intervals 90kb up- and downstream of their promoters are 2.4-fold enriched overall in p300-binding sites (P < 0.05, Fig. 5a). In total, 14% of all forebrain p300 peaks are located within 101kb from a promoter of a forebrain-overexpressed gene. The most pronounced enrichment (4.8-fold, P < 0.01) was observed within 10kb up- and

downstream of promoters of forebrain-specifically expressed genes. In contrast, forebrain peaks are not enriched near genes over-expressed in other parts of the body (Fig. 5b, Suppl. Table 9). Near those genes over-expressed 5-fold or more in the forebrain, even higher enrichment of forebrain peaks was observed (11-fold enrichment within 10kb from promoters, data not shown). We found similar enrichment of limb-derived p300 peaks near limb-overexpressed genes (Suppl. Fig. 6, Suppl. Tables 10 and 11). These observations are consistent with the sequences bound by p300 in the forebrain or limb of day 11.5 embryos being enhancers that drive the expression of adjacent genes in these tissues at this time point.

Discussion

In the present study, we have determined the genome-wide distribution of the transcriptional coactivator protein p300 [23] by ChIP-seq [17] directly from developing mouse tissues. Remarkably, enrichment of p300 in different mouse tissues correctly predicted the spatial enhancer activities of human non-coding sequences in 80% of cases tested in a transgenic mouse assay, whereas absence of p300 enrichment correlated in 93% of cases with absence of enhancer activity in the respective tissue (Suppl. Table 5). The few elements that did not drive reporter gene expression in the tissue predicted by p300 ChIP-seq may represent cases in which the function of regulatory elements has diverged between the mouse sequences identified by ChIP-seq and the human orthologous regions tested in the transgenic mouse assay. In support of this hypothesis, we observed several cases in which the non-coding p300-bound region from mouse, but not the orthologous human sequence had reproducible enhancer activities as predicted by p300 ChIP-seq from mouse tissues (data

not shown). Taken together, the present approach provides a dramatically improved specificity for locating enhancers in the human genome compared to conservation-based methods [10,11] and, importantly, predicts their *in vivo* activity patterns with higher accuracy than currently available motif-based computational methods (e.g., ref. [35,36]).

Most p300-binding regions identified in developing mouse tissues are under detectable evolutionary constraint. They typically overlap conserved non-coding sequences whose length (median: 113bp) far exceeds that of an individual transcription factor binding site, suggesting the presence of larger functional modules. In cell culture-based chromatin studies, a sizeable fraction of non-coding regions in the human genome was found to be functional yet not constrained [13,14]. This apparent discrepancy might be due to differences in evolutionary constraint between enhancers active in developing tissues compared to those in individual cell types, but highlights the intrinsic challenge of inferring *in vivo* functionality from studies in cell culture.

A generalized picture of the epigenetic marks and proteins associated with different types of functional non-coding elements has started to emerge from genome-wide chromatin studies [13,18,28,37-41]. We can now begin to use these signatures to unravel gene regulation on a genomic scale in the context of living organisms. The highly specific approach for identification of developmental enhancers and their activity patterns presented here represents a step in this direction. Complementary *in vivo*-derived genomic datasets may be produced in the future, covering additional embryonic stages, anatomical regions and subregions, and perhaps considering additional molecular markers [28,42-45]. Focused experiments informed by such insights will expedite studies of the genome-wide activity dynamics of enhancers in developmental, physiological and pathological processes.

Methods Summary

Embryonic forebrain, midbrain and limb tissue was isolated from mouse embryos at e11.5. Cross-linking, chromatin isolation, sonication and immunoprecipitation using an anti-p300 antibody were performed as previously described [40,46]. ChIP DNA was further sheared by sonication, end-repaired, ligated to sequencing adapters and amplified by emulsion PCR as previously described [47]. Gel purified amplified ChIP DNA between 300 and 500bp was sequenced on the Illumina Genome Analyzer II platform to generate 36 bp reads.

Sequence reads were aligned to the mouse reference genome (mm9) using BLAT [48]. Uniquely aligned reads were extended to 300bp in the 3' direction and used to determine the read coverage at individual nucleotides at 25bp intervals throughout the mouse genome. p300-enriched regions (peaks) with an estimated false discovery rate of ≤0.01 were identified by comparison with a random distribution of the same number of reads. Candidate peaks mapping to repetitive regions were removed as likely artifacts.

Candidate regions for transgenic testing were selected based on ChIP-seq results and cover a wide spectrum of conservation. Enhancer candidate regions were amplified by PCR from human genomic DNA and cloned into an Hsp68-promoter-LacZ reporter vector as previously described [6,31]. Transgenic mouse embryos were generated and evaluated for reproducible LacZ activity at e11.5 as previously described [6].

Total RNA from embryonic day 11.5 whole embryos and forebrain tissue was hybridized to GeneChip Mouse Genome 430 2.0 arrays (Affymetrix) and analyzed according to the manufacturer's recommendations. Forebrain- and whole embryo-enriched genes were identified as having at least 2.5-fold greater expression in one dataset compared with the

other, and a minimum signal intensity of 100. Limb-enriched genes were identified by comparison with publicly available wild-type e11.5 proximal hindlimb gene expression data (GEO Series GSE10516, samples GSM264689, GSM264690, GSM264691) [49].

Acknowledgements

The authors wish to thank Roya Hosseini and Sengthavy Phouanenavong for technical support; John Rubenstein, Jason Long, Justin Choi and Yiwen Zhu for help with microarray experiments. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396. L.A.P./E.M.R. were supported by the Berkeley-PGA, under the Programs for Genomic Applications, funded by National Heart, Lung, & Blood Institute, and L.A.P. by the National Human Genome Research Institute. A.V. was supported by an American Heart Association postdoctoral fellowship.

Supplementary information is linked to the online version of the paper at www.nature.com/nature.

All ChIP-seq datasets described in this study have been deposited at the National Center for Biotechnology Information (NCBI) in the Gene Expression Omnibus (GEO) database under accession number GSE13845.

The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.A.P. (lapennacchio@lbl.gov).

References

1. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304-51 (2001).

2. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

3. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78-94 (1997).

4. Okazaki, Y. et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563-73 (2002).

5. Marshall, H. et al. A conserved retinoic acid response element required for early expression of the homeobox gene Hoxb-1. *Nature* **370**, 567-71 (1994).

6. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).

7. de la Calle-Mustienes, E. et al. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res* **15**, 1061-72 (2005).

8. Woolfe, A. et al. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**, e7 (2005).

9. Prabhakar, S. et al. Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res* **16**, 855-63 (2006).

10. Pennacchio, L. A. et al. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499-502 (2006).

11. Visel, A. et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40**, 158-60 (2008).

12. Holland, L. Z. et al. The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res* **18**, 1100-11 (2008).

13. ENCODE Project Consortium et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).

14. Margulies, E. H. et al. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res* **17**, 760-74 (2007).

15. Cooper, G. M. & Brown, C. D. Qualifying the relationship between sequence conservation and molecular function. *Genome Res* **18**, 201-5 (2008).

16. McGaughey, D. M. et al. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. *Genome Res* **18**, 252-60 (2008).

17. Robertson, G. et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**, 651-7 (2007).

18. Mikkelsen, T. S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-60 (2007).

19. Robertson, A. G. et al. Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res* **18**, 1906-1917 (2008).

20. Cuddapah, S. et al. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res*, (Epub ahead of print) (2008).

21. Wederell, E. D. et al. Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res* **36**, 4549-64 (2008).

22. Valouev, A. et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* (2008).

23. Eckner, R. et al. Molecular cloning and functional analysis of the adenovirus E1A-associated 300-kD protein (p300) reveals a protein with properties of a transcriptional adaptor. *Genes Dev* **8**, 869-84 (1994).

24. Eckner, R., Yao, T. P., Oldread, E. & Livingston, D. M. Interaction and functional collaboration of p300/CBP and bHLH proteins in muscle and B-cell differentiation. *Genes Dev* **10**, 2478-90 (1996).

25. Yao, T. P. et al. Gene dosage-dependent embryonic development and proliferation defects in mice lacking the transcriptional integrator p300. *Cell* **93**, 361-72 (1998).

26. Merika, M., Williams, A. J., Chen, G., Collins, T. & Thanos, D. Recruitment of CBP/p300 by the IFN beta enhanceosome is required for synergistic activation of transcription. *Mol Cell* **1**, 277-87 (1998).

27. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7**, 29-59 (2006).

28. Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-8 (2007).

29. Xi, H. et al. Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet* **3**, e136 (2007).

30. Mouse Genome Sequencing Consortium et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).

31. Kothary, R. et al. A transgene containing lacZ inserted into the dystonia locus is expressed in neural tube. *Nature* **335**, 435-7 (1988).

32. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**, D88-92 (2007).

33. Cheng, Y. et al. Transcriptional enhancement by GATA1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif. *Genome Res* **18**, 1896-1905 (2008).

34. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-50 (2005).

35. Hallikas, O. et al. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47-59 (2006).

36. Pennacchio, L. A., Loots, G. G., Nobrega, M. A. & Ovcharenko, I. Predicting tissue-specific enhancers in the human genome. *Genome Res* **17**, 201-11 (2007).

37. Kim, T. H. et al. A high-resolution map of active promoters in the human genome. *Nature* **436**, 876-80 (2005).

38. Boyle, A. P. et al. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-22 (2008).

39. Schones, D. E. & Zhao, K. Genome-wide approaches to studying chromatin modifications. *Nat Rev Genet* **9**, 179-91 (2008).

40. Barrera, L. O. et al. Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. *Genome Res* **18**, 46-59 (2008).

41. Chen, X. et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-17 (2008).

42. Kwok, R. P. et al. Nuclear protein CBP is a coactivator for the transcription factor CREB. *Nature* **370**, 223-6 (1994).

43. Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell* **87**, 953-9 (1996).

44. Agalioti, T. et al. Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. *Cell* **103**, 667-78 (2000).

45. Ge, K. et al. Transcription coactivator TRAP220 is required for PPAR gamma 2-stimulated adipogenesis. *Nature* **417**, 563-7 (2002).

46. Li, Z. et al. A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc Natl Acad Sci U S A* **100**, 8164-9 (2003).

47. Blow, M. J. et al. Identification of the source of ancient remains through genomic sequencing. *Genome Res* (2008).

48. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).

49. Krawchuk, D. & Kania, A. Identification of genes controlled by LMX1B in the developing mouse limb bud. *Dev Dyn* **237**, 1183-92 (2008).

50. Karolchik, D. et al. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36**, D773-9 (2008).

51. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-5 (2007).

52. Giardine, B. et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**, 1451-5 (2005).

53. Mikkelsen, T. S. et al. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature* **447**, 167-77 (2007).

(Note: Reference 50 is cited in a figure legend, references 51-53 are only cited in the Online Methods).

Figures and Figure Legends

(High-resolution figures are provided as separate files)

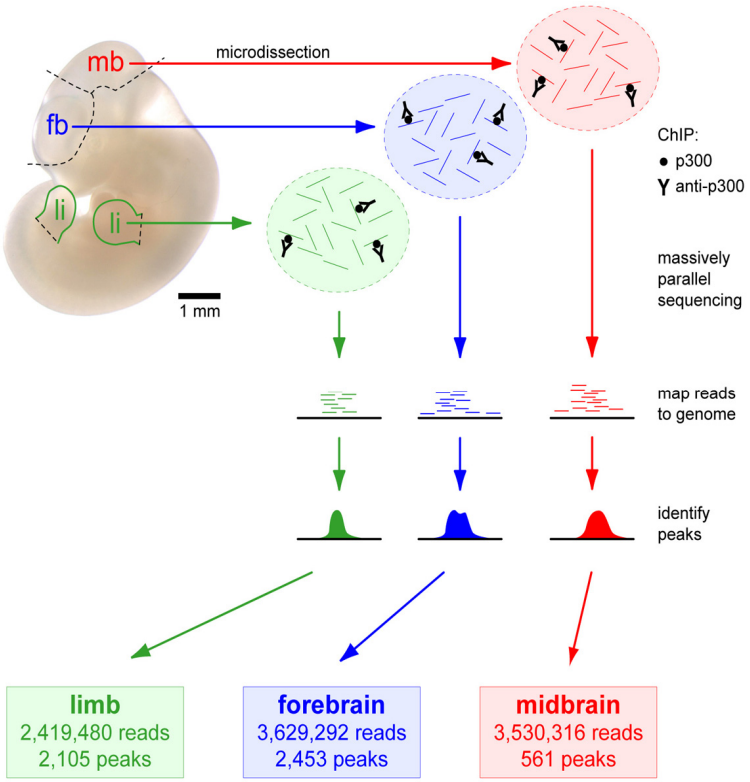


Figure 1: Tissue dissection boundaries, overview of the ChIP-seq approach and summary of p300 results. Tissue dissection boundaries are indicated in a representative unstained e11.5 mouse embryo. For each sample, tissue was pooled from over 150 embryos and ChIP-seq was performed with a p300-antibody. Reads obtained for each of the three tissues that unambiguously aligned to the reference mouse genome were used to define peaks (FDR<0.01). A more comprehensive overview of sequencing and mapping results is provided in Suppl. Table 1. fb, forebrain; mb, midbrain; li, limb.

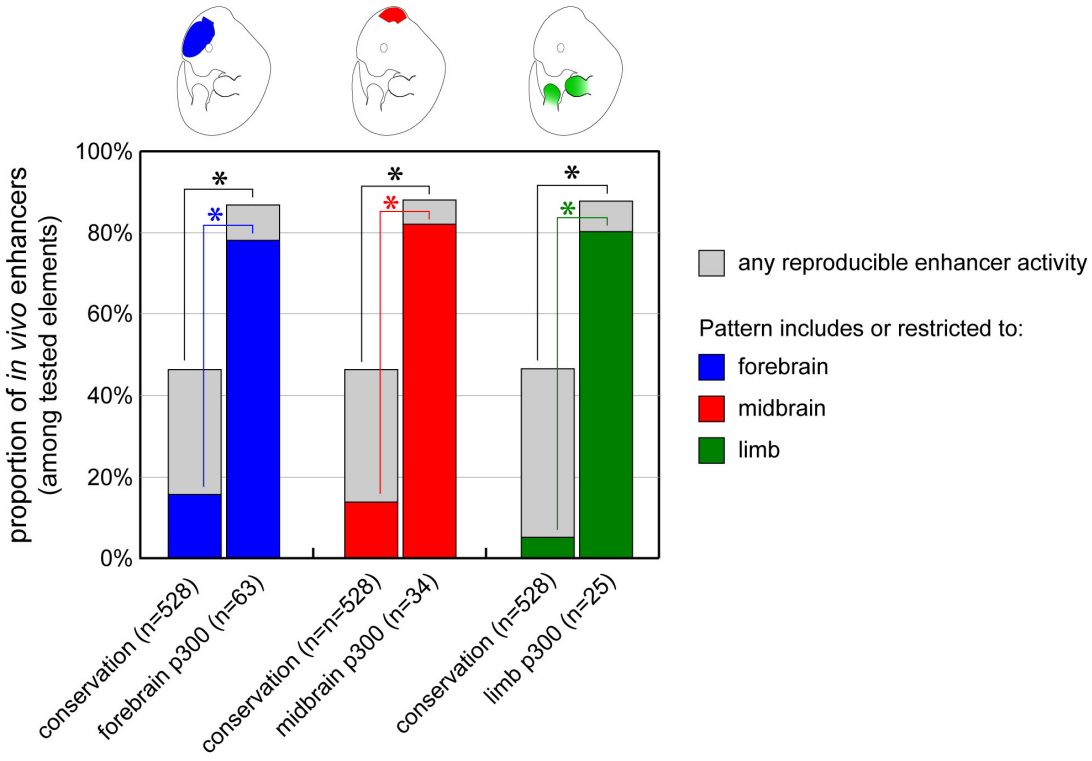


Figure 2: p300 binding accurately predicts enhancers and their tissue-specific activity patterns. Bar height indicates frequency of *in vivo* enhancers (reproducible at e11.5) that are active in any tissue (gray bars + colored bars) and the fraction of enhancers whose pattern includes reproducible forebrain, midbrain or limb activity (colored boxes only). In each case, candidate elements predicted by p300 peaks in forebrain, midbrain or limb were compared to the frequency of the respective pattern in a background set of 528 previously tested sequences identified through extreme evolutionary conservation (combined datasets from references [10] and [11]). The component activities of elements predicted to be active in multiple tissues were counted separately. *, $P < 0.00005$; Fisher's Exact test, one-tailed.

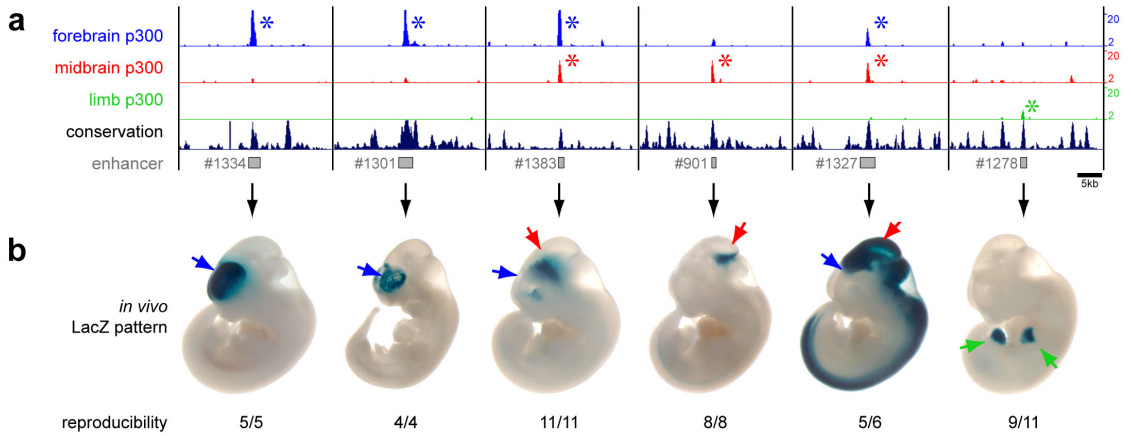


Figure 3: Examples of successful prediction of *in vivo* enhancers by p300 binding in embryonic tissues. a) Coverage by extended p300 reads in forebrain (blue), midbrain (red), and limb (green). Asterisks indicate significant (FDR < 0.01) p300-enrichment in chromatin isolated from the respective tissue. Multi-species vertebrate conservation plots (black) were obtained from UCSC genome browser [50]. Gray boxes correspond to candidate enhancer regions. Numbers to the right indicate overlapping extended reads. b) Representative LacZ-stained embryos with *in vivo* enhancer activity at e11.5. Reproducible staining in forebrain, midbrain, and limb is indicated by arrows. Numbers show reproducibility of LacZ reporter staining. Additional embryos obtained with each construct and genomic coordinates are available at the Vista Enhancer Browser [32].

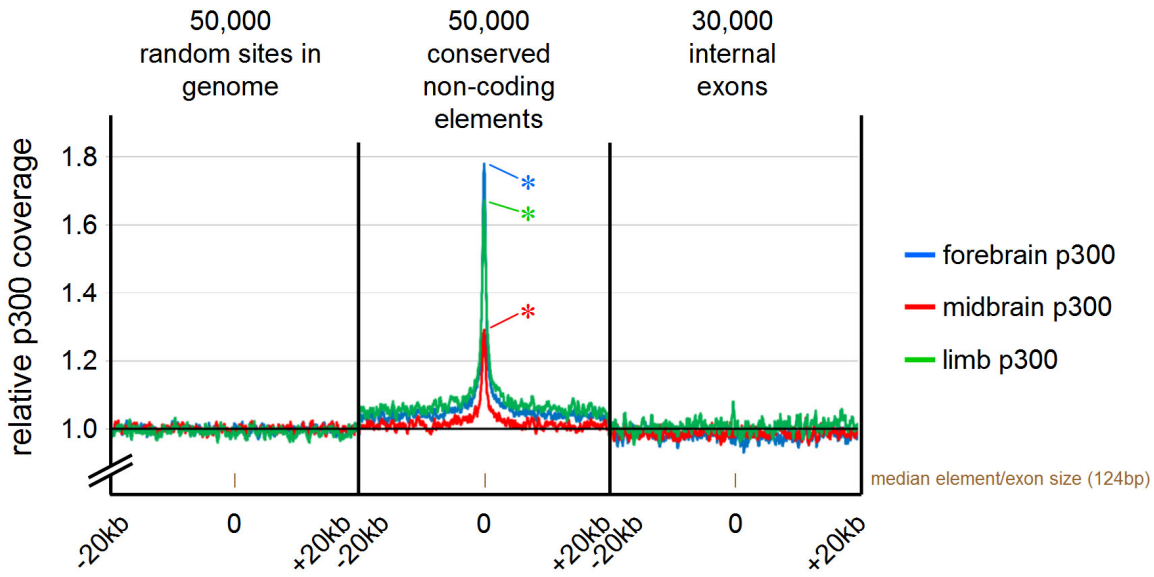


Figure 4: In all tissues examined, p300 is enriched at highly conserved non-coding regions. We used a genome-wide set of 50,000 extremely constrained non-coding sequences identified in human-mouse-rat genome alignments [11] to assess the correlation between p300 enrichment and non-coding sequence conservation. Even though only subsets of the constrained non-coding elements are expected to be active enhancers in any given embryonic tissue, we observe strong enrichment in p300 binding in all three tissues compared to input DNA (*, $P < 1e\text{-}100$, Fisher's Exact test). Relative p300 coverage near random sites and internal exons is shown for comparison. Brown bars indicate median sizes of conserved elements/exons (124bp in both cases). For additional details, see Suppl. Table 7.

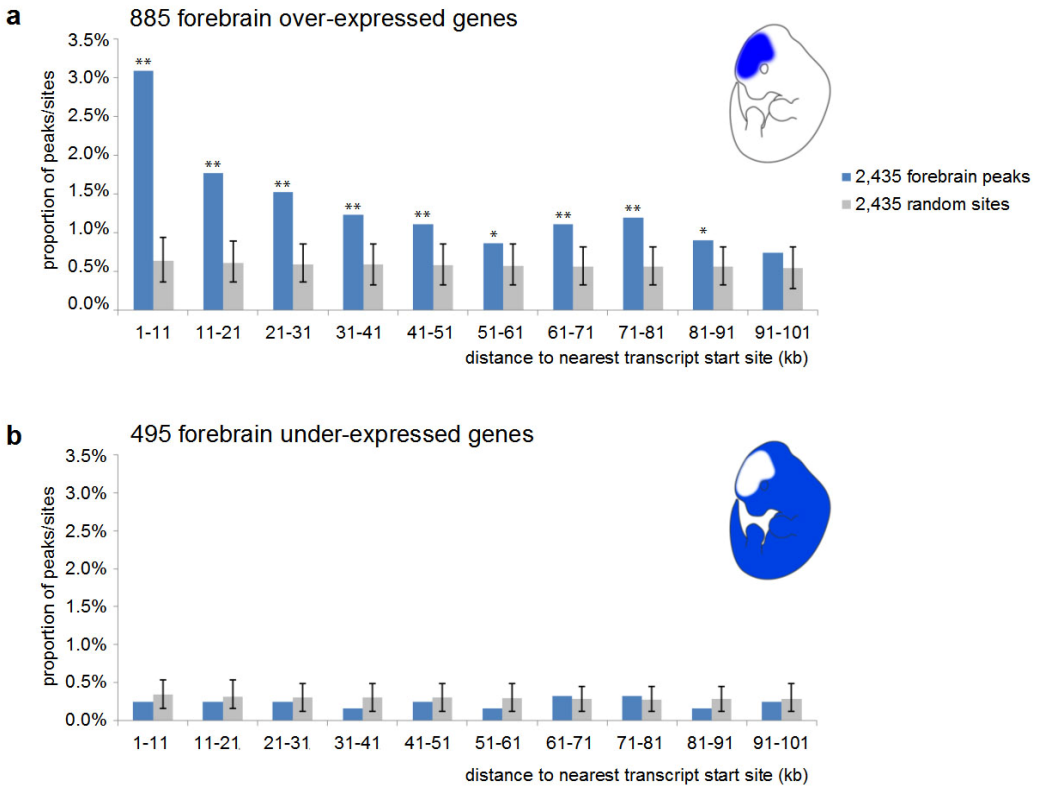


Figure 5: p300 peaks are enriched near genes that are expressed in the same tissue. We compared the genome-wide distribution of p300-enriched regions in forebrain tissue at e11.5 with microarray expression data for forebrain at the same stage. 885 genes were forebrain-specifically over-expressed and 495 genes were under-expressed relative to whole embryo RNA at the selected thresholds. Promoters (defined as 1kb upstream and downstream of transcription start sites) were excluded from the analysis. a) 10kb bins up to 91kb away from forebrain over-expressed genes were significantly enriched in forebrain p300 peaks. b) No peak enrichment was observed for forebrain under-expressed genes. Error bars indicate the 90% confidence interval based on 1000 iterations of randomized distribution (*, $P < 0.05$; **, $P < 0.01$; both one-tailed).

Online Methods

Tissue Dissection and Chromatin Immunoprecipitation (ChIP)

Embryonic forebrain, midbrain and limb tissue was isolated from timed-pregnant CD-1 strain mouse embryos at e11.5 by microdissection in cold PBS along the anatomical boundaries indicated in Fig. 1. Tissue samples were cross-linked for 15 min (1% formaldehyde, 10mM NaCl, 100uM EDTA, 50uM EGTA, 5mM HEPES pH8.0) at room temperature. Cross-linking was terminated by the addition of 125mM glycine and cells were dissociated in a glass douncer. Chromatin isolation, sonication and immunoprecipitation were performed as previously described [40,46]. Briefly, 1mg of sonicated chromatin (OD260) was incubated with 10ug of antibody (rabbit polyclonal anti-p300; SC-585, Santa Cruz Biotechnology) coupled to IgG magnetic beads (Dynal Biotech) overnight at 4°C. The magnetic beads were washed eight times with RIPA buffer (50 mM HEPES at pH 8.0, 1 mM EDTA, 1% NP-40, 0.7% DOC, and 0.5 M LiCl, supplemented with Complete protease inhibitors from Roche Applied Science), and washed once with TE (10 mM Tris at pH 8.0, 1 mM EDTA). After washing, bound DNA was eluted at 65°C in elution buffer (10 mM Tris at pH 8.0, 1 mM EDTA, and 1% SDS) for 10 minutes and incubated at 65°C overnight to reverse cross-links. After reversal of cross-linking, immunoprecipitated DNA was treated sequentially with Proteinase K, RNase A and desalted using the QIAquick PCR purification kit (Qiagen).

Amplification and Illumina Sequencing of ChIP DNA

ChIP DNA was quantified by Qubit assay HS kit. Approximately 0.1 ng of each ChIP DNA sample was sheared using Sonicator XL2020 (Misonix) with a microplate horn for 10 minutes at 55% power output and 90% amplitude. Sheared ChIP DNA extract was end-repaired using the End-It™ DNA End-Repair Kit (ER0720, Epicentre, Madison, WI). Illumina adapters (56 bp and 34 bp) were ligated using T4 DNA ligase (5U/ul, Fermentus, #EL0334) and recovered using a MiniElute Reaction Cleanup Kit (#28204, Qiagen). Linker ligated ChIP DNA was amplified by emulsion PCR for 40 cycles as previously described [47]. Amplified ChIP DNA between 300 and 500bp was gel purified on 2% agarose and sequenced on the Illumina Genome Analyzer II according to the manufacturer's instructions except that emulsion PCR amplified DNA containing the GA2 sequencing adapter was applied directly to the cluster station for bridge amplification. The resulting flow-cell was sequenced for 36 cycles to generate 36 bp reads.

Processing of Illumina sequence data

Unfiltered 36bp Illumina sequence reads were aligned to the mouse reference genome (NCBI build 37, mm9) using BLAT [48] with optional parameters (-minScore=20 -minIdentity=80 -stepSize=5). BLAT was performed in parallel on a sge-cluster. For each read, the two highest-scoring alignments were compared and reads were rejected as repetitive unless the score of the best alignment was at least two greater than that of the second best alignment. The remaining reads were further filtered to reject those with a BLAT alignment score <21, with >1bp insertion or deletion, or with >2 unaligned bases at the start of the read. Finally, reads with identical start sites in the mouse genome were considered likely to

be duplicate sequences arising as an artifact of sample amplification or sequencing, and were counted only once. The remaining reads were classed as uniquely aligned to the mouse genome.

Uniquely aligned reads were extended to 300bp in the 3' direction to account for the average length of size-selected p300 ChIP fragments used for sequencing. These extended read coordinates were used to determine the read coverage at individual nucleotides at 25bp intervals throughout the mouse genome. This data was used to produce coverage plots for visualization in the UCSC genome browser.

To identify p300-enriched regions (peaks), we compared the observed frequency of coverage depths with those expected from a random distribution of the same number of reads generated computationally as described by Robertson *et al.* [17] Briefly, the probability of observing a peak with a coverage depth of at least H reads is given by a sum of Poisson probabilities as :

$$1 - \sum_{k=0}^{H-1} \frac{e^{-\lambda} \lambda^k}{k!}$$

Where λ is the average genome-wide coverage of extended reads given by (read length x number of aligned reads) / alignable genome length. To estimate the alignable genome length, one million randomly selected 36mers from the mouse genome were realigned to the mouse genome using the same alignment and filtering scheme as for reads. 77.3% of 36mers were uniquely mapped back to the mouse genome, resulting in an alignable genome length of 2.107Gb.

For each sample, we determined the read coverage depth at which the observed frequency of sites with that coverage exceeded the expected frequency by a factor of 100 (false discovery rate, FDR ≤ 0.01). Candidate peaks were identified as sites where coverage exceeded this threshold, and peak boundaries were extended to the nearest flanking positions at which read coverage fell below two reads. All consecutive regions of enrichment separated by regions of continuous coverage greater than 2 reads were merged into a single peak. Candidate peaks mapping to chr_random contigs, centromeric regions, telomeric regions, segmental duplications, satellite repeats, rRNA repeats or regions of >70% repeat sequence, and those coinciding with enriched regions in the control sample (input DNA) were removed as likely artifacts due to mis-alignment of heterochromatic sequences that are not currently represented in the mouse reference genome sequence. The remaining peaks represent high-confidence p300 enriched regions and putative enhancers with activity in specific tissues.

Annotation of p300 ChIP-seq read datasets with respect to nearby genes (UCSC known genes [50]), internal exons (mouse RefSeq [51] exons >30kb from the ends of transcripts) and conserved non-coding sequences (top 50,000 constrained non-coding human-mouse-rat conserved elements identified using GUMBY with R-ratio parameter R = 50 [9,11]) was performed using Galaxy [52] and custom Perl scripts. Annotation of p300-enriched regions with respect to UCSC known genes and vertebrate phastCons elements [34] was performed using custom Perl scripts.

Transgenic mouse enhancer assay

Candidate regions for transgenic testing were selected based on ChIP-seq results. Peaks for which human orthologous regions could not be unambiguously established and those without detectable conservation in opossum [53] were excluded from transgenic testing. Thus, the tested peaks cover a wide spectrum of conservation, but are overall more constrained than all peaks identified genome-wide (median score 457 for all peaks vs. 626 for tested peaks). Enhancer candidate regions (average size: 2.4kb) were amplified by PCR from human genomic DNA (Clontech) and cloned into an Hsp68-promoter-LacZ reporter vector upstream of an Hsp68-promoter coupled to a LacZ reporter gene as previously described [6,31]. Candidate sequences were not cloned in any particular orientation, effectively resulting in randomized insert orientation among the test constructs. Genomic coordinates of amplified regions are reported in Suppl. Table 5. Transgenic mouse embryos were generated by pronuclear injection and F0 embryos were collected at e11.5 and stained for LacZ activity as previously described [6]. Only patterns that were observed in at least three different embryos resulting from independent transgenic integration events of the same construct were considered reproducible (see Suppl. Table 5). To account for minor variation in separating forebrain from midbrain during tissue dissection, forebrain and midbrain p300 peaks were also considered correct predictions if the reproducible *in vivo* pattern was located in the fore-/midbrain boundary region, whereas absence of a p300 peak was only considered a false-negative prediction if the reproducible *in vivo* pattern clearly extended beyond the boundary region.

Microarrays

Tissue was isolated from timed-pregnant CD-1 strain mouse embryos at e11.5. Forebrains were further subdivided into basal telencephalon (subpallium), dorsal telencephalon (pallium), and diencephalon, which were processed separately in subsequent steps. For comparison, whole embryos (littermates) were collected. All samples were collected, processed and hybridized in duplicate. Total RNA was extracted using Trizol reagent (Invitrogen). Synthesis of cRNA, hybridization to GeneChip Mouse Genome 430 2.0 arrays (Affymetrix) and analysis of hybridization results was performed according to the manufacturer's recommendations. For each sample, the average expression value from duplicates was used for downstream analyses. Forebrain-enriched genes were defined as those with expression at least 2.5-fold greater expression in at least one of the three forebrain regions compared with the whole embryo, and with a minimum signal intensity of 100. Whole embryo-enriched genes are defined as those with at least 2.5-fold greater expression in the whole embryo than in each of the three forebrain regions, and a minimum signal intensity of 100. Distances between p300 peaks and the 5' end of Affymetrix consensus cDNA sequences from mouse MOE430 (A and B) aligned to the mouse reference genome (mm9) were used to determine the closest forebrain-enriched and whole embryo-enriched genes (Suppl. Tables 8 and 9). The same procedure was used to analyze the correlation of limb p300 peaks with limb gene expression, except that limb expressed genes were identified by comparison of publicly available wild-type e11.5 proximal hindlimb gene expression data (GEO Series GSE10516, samples GSM264689, GSM264690, GSM264691) [49], with the whole embryo gene expression data generated in the present study (Suppl. Tables 10 and 11).

Animal Work

All animal work was performed in accordance with protocols reviewed and approved by the Lawrence Berkeley National Laboratory Animal Welfare and Research Committee.

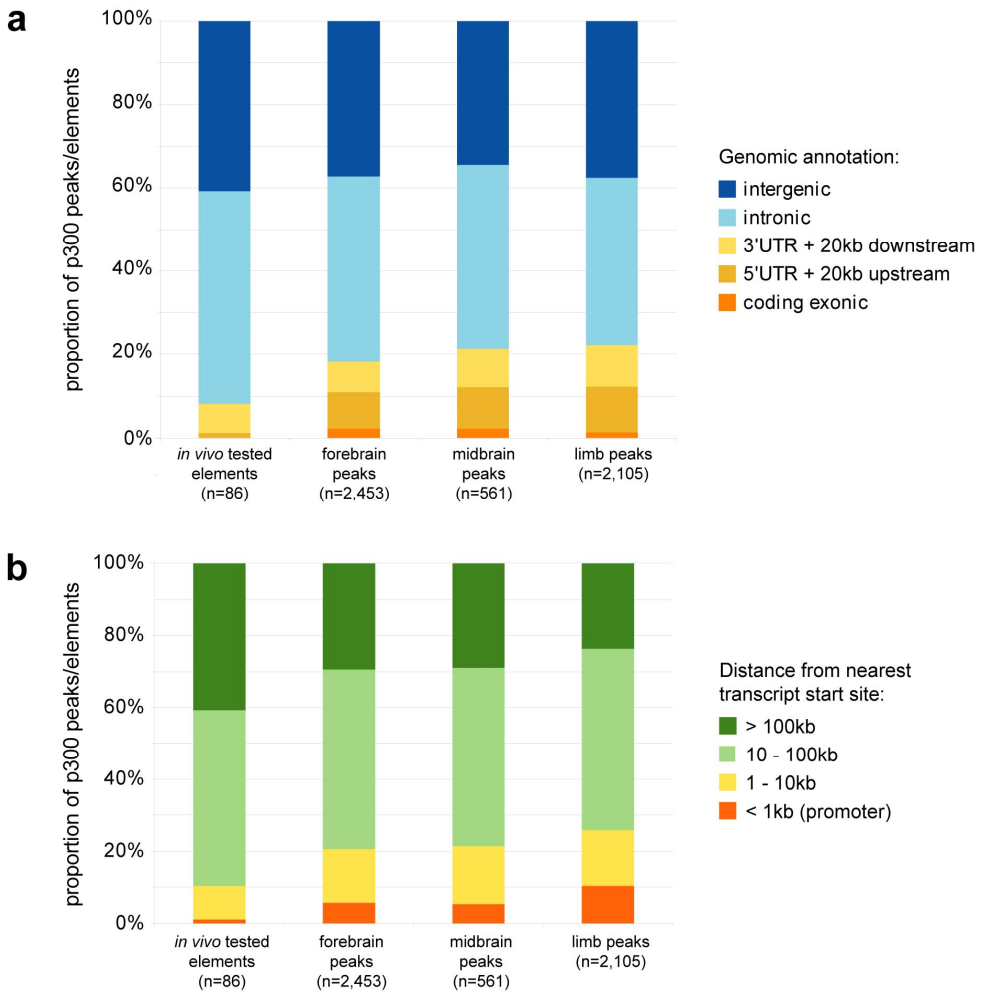
Additional References (Online Methods Only)

51. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-5 (2007).

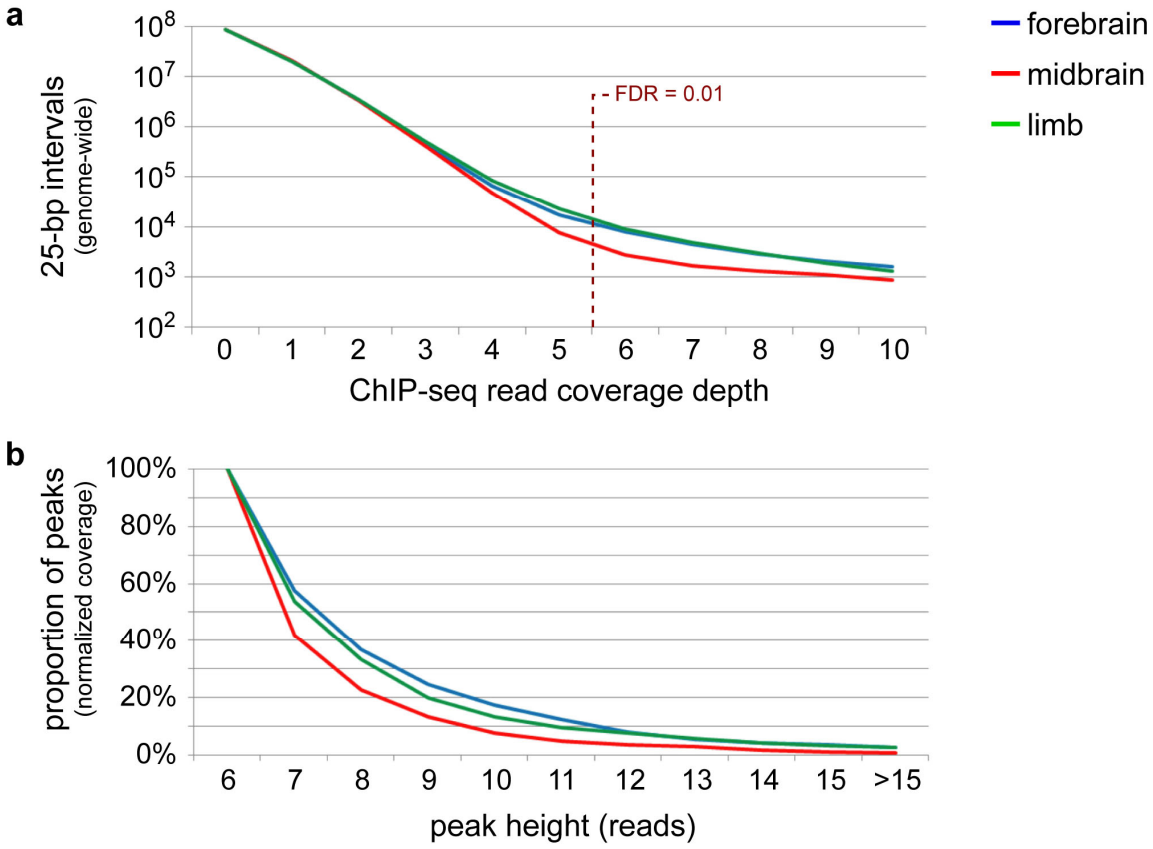
52. Giardine, B. et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**, 1451-5 (2005).

53. Mikkelsen, T. S. et al. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature* **447**, 167-77 (2007).

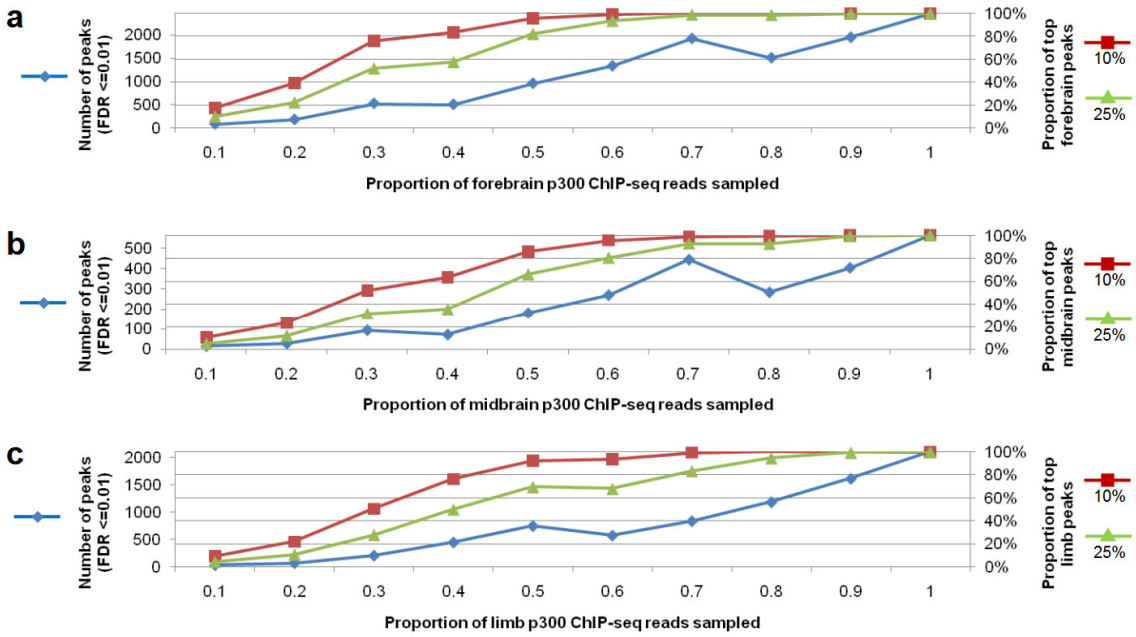
Supplementary Information



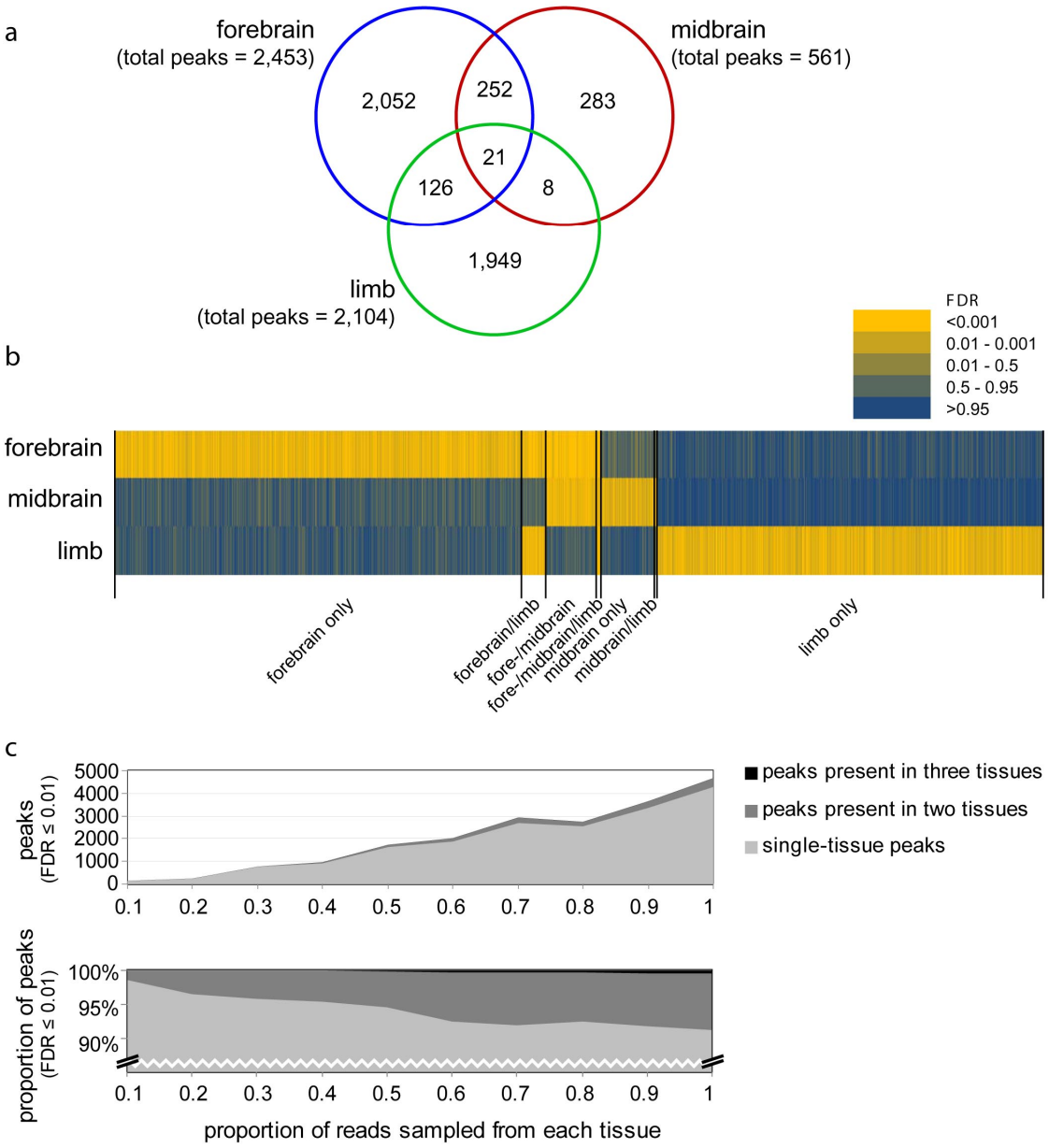
Suppl. Fig. 1: Genome-wide distribution of p300 peaks relative to annotated genes. a) Peaks were classified according to their position relative to UCSC Known Genes. Upstream regions were defined as 20kb upstream of an annotated transcript start site, downstream was defined as 20kb downstream of an annotated transcript end site. Intergenic, intronic and downstream regions were tested in the transgenic mouse assay at an approximately representative ratio, whereas exonic and upstream regions were excluded from transgenic testing. **b)** The distance from the midpoint of each p300 peak to the respectively nearest UCSC Known Genes annotated transcript start site was determined. Peaks within 1kb of the nearest known transcript start site were considered potential promoters and excluded from *in vivo* testing. One tested non-coding element (#1331) is located in an intron of the *Pou2f1* gene, but is here classified as promoter-proximal due to its proximity to an alternative internal transcription start site. Compared to the genome-wide distribution of p300 peaks, the *in vivo* tested elements are mildly skewed towards putative medium-distance (10-100kb) and long-distance (>100kb) enhancers.



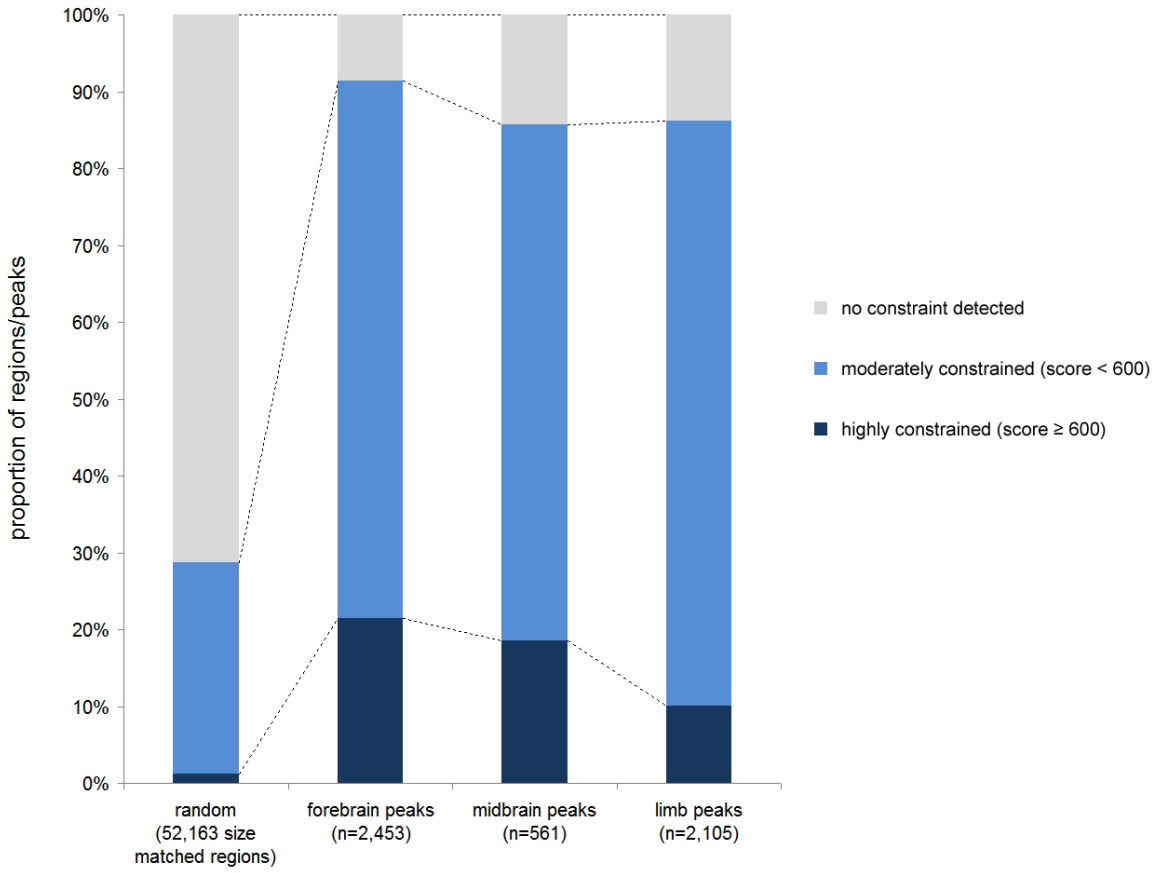
Suppl. Fig. 2: Read clustering and peak coverage properties of p300 ChIP-seq reads from forebrain, midbrain and limb. a) A randomly sampled equal number of reads (n=2,419,480) from each dataset was mapped to the mouse genome, followed by calculation of read coverage in 25bp intervals throughout the genome. The number of peaks with an FDR ≤ 0.01 (i.e., six or more overlapping reads) identified from these identically sized subsets of reads is smaller for midbrain (408) than for forebrain (1,731) and limb (2,105). A smaller fraction of midbrain-derived reads overlap each other compared to forebrain and limb, suggesting less efficient ChIP enrichment prior to massively-parallel sequencing. For example, 7,500 of the 25bp-intervals in the genome are covered by five reads from p300-enriched DNA from midbrain, compared to 17,000 and 22,000 sites in the forebrain and limb datasets, respectively. **b)** The read coverage (peak height) for all significantly enriched regions in each of the three coverage-normalized datasets was determined (cumulative plot). Midbrain-derived p300 peaks are on average covered by fewer reads than forebrain- and limb-derived p300 peaks (average forebrain 7.8 reads, midbrain 7.0 reads, limb 7.6 reads). Consistent with this observation and the overall lower number of peaks from the midbrain sample, the proportion of reads overlapping p300 peaks in the coverage-normalized datasets is lower for midbrain (0.14%) than for forebrain (0.83%) and limb (0.68%).



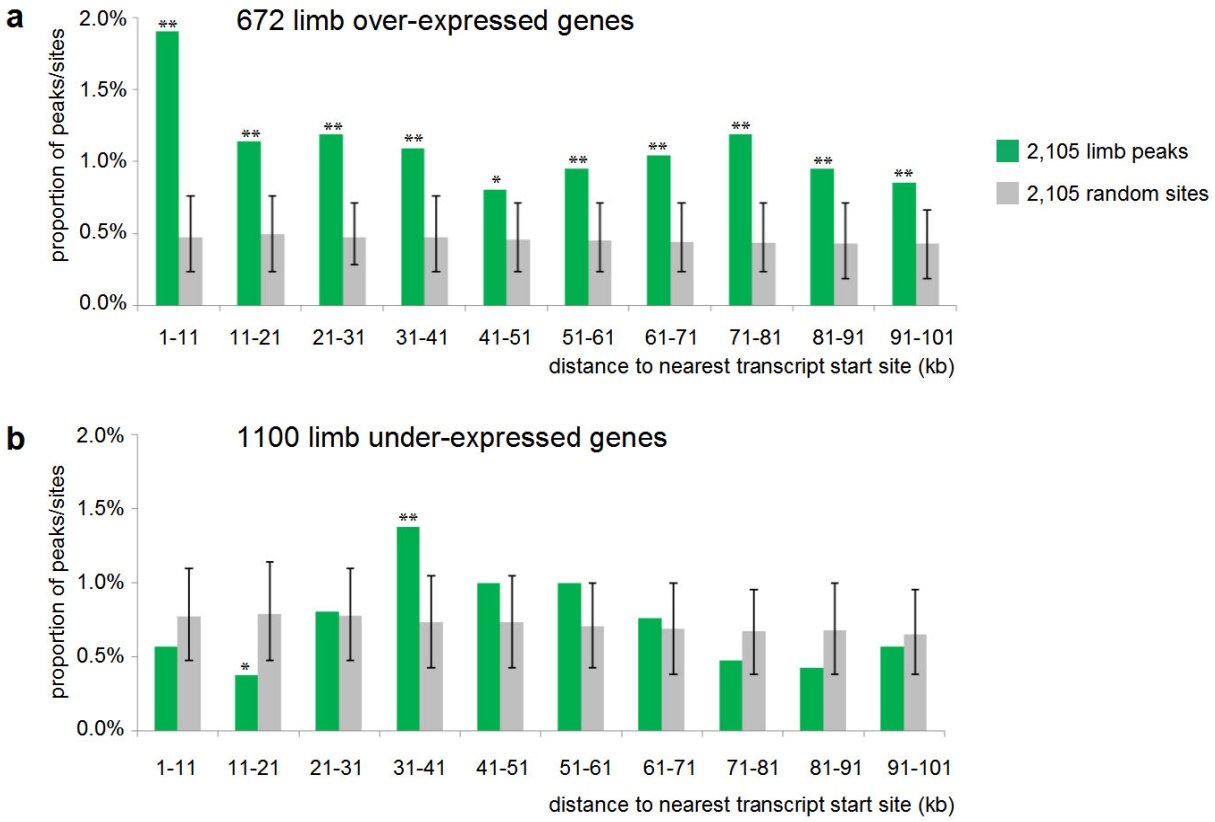
Suppl. Fig. 3: Sensitivity of p300 peak discovery by ChIP-seq increases with sampling depth. For each p300 ChIP-seq dataset, between 10% and 90% of reads were sampled at random and used to calculate peaks at a threshold of FDR ≤ 0.01 (blue, left axis). The number of peaks identified at this threshold does not monotonically increase with increasing sampling depth due to discrete changes in the number or overlapping reads that define FDR = 0.01. To estimate if saturation was obtained for the most significant peaks in each tissue, the resultant peaks for each sample were intersected with the top 10% (red) and top 25% (green) most highly covered peaks from the complete dataset of the respective tissues. In all three tissues, 86%-96% of the eventually most read-enriched peaks (top 10%) were discovered if only half of the data was sampled, suggesting that additional peaks discovered with increasing sampling depth are overall less significantly p300 ChIP-enriched. Each data point represents the average value from 5 individual random samples.



Suppl. Fig. 4: Most p300-enriched regions are tissue-specific. a) Less than 10% of the regions are significantly (FDR<0.01) p300-enriched in more than one of the three tissues, while over 90% are significantly enriched in only one of the three tissues. Most overlap was observed between forebrain and the anatomically adjacent midbrain. **b)** Heat-map representation of all genome regions that are significantly (FDR<0.01) p300-enriched in at least one of the three tissues, clustered into the 7 categories as defined in a). **c)** Re-sampling of subsets of the data indicates that with increasing sequencing depth, the total number of peaks present in only one of the three tissues overall increases (top) while there is a moderate increase in the proportion of regions with significant p300 ChIP-seq read coverage in two or all three tissues examined (bottom).



Suppl. Figure 5: Most *in vivo* p300-binding regions are significantly constrained in vertebrates. Genome-wide sets of p300-enriched regions in e11.5 forebrain, midbrain and limb were intersected with vertebrate constrained elements [34] and the score (transformed log-odds) of the most constrained element overlapping each of the peaks was considered. For comparison, results for a random set of genome regions (size-matched to forebrain peaks) are shown.



Suppl. Fig. 6: p300 limb peaks are enriched near genes expressed in the limb. We compared the genome-wide distribution of p300-enriched regions in limb tissue at e11.5 with microarray expression data from limb at the same stage [49]. 672 genes were limb-specifically over-expressed and 1100 genes were under-expressed relative to whole embryo RNA at the selected thresholds. Analysis was performed as described for forebrain p300 and microarray data (see Fig. 5 and Methods). a) 10kb bins up to 101kb away from limb over-expressed genes were significantly enriched in limb p300 peaks. b) Consistent with forebrain data, no overall peak enrichment or depletion was observed near limb under-expressed genes except for the 11-21kb bin (depletion) and the 31-41kb bin (enrichment). Note that higher background variation and weaker enrichment in this data compared to forebrain (Fig. 5) may result from differences in the regions of the limb sampled for the gene expression dataset [49] compared with those used for p300 ChIP-seq (Fig. 1). Error bars indicate the 90% confidence interval based on 1000 iterations of randomized distribution (*, $P < 0.05$; **, $P < 0.01$; both one-tailed).

Overview of Supplementary Tables:

Suppl. Table 1: Summary of ChIP-seq and mapping results – *embedded in this file, see below*

Suppl. Tables 2-4: p300 peaks in forebrain, midbrain, and limb – *provided as separate Excel files*

Suppl. Table 5: Results of *in vivo* enhancer assays in transgenic mice – *provided as separate Excel file*

Suppl. Table 6: Predicted and observed *in vivo* enhancer activities – *embedded in this file, see below*

Suppl. Table 7: Genome-wide distribution of ChIP-seq reads – *embedded in this file, see below*

Suppl. Tables 8-11: Forebrain and limb over- and underexpressed genes – *provided as separate Excel files*

Tissue	Total Reads	Alignable to Mouse	Unambiguously Aligned	Unique Unambiguously Aligned	Peak Coverage Threshold (FDR<0.01)	Peaks Genome-Wide (FDR<0.01)
forebrain	26,759,420	13,728,898 (51%)	8,659,420	3,629,292	7 reads	2,453
midbrain	24,340,547	14,517,733 (60%)	9,431,268	3,530,316	7 reads	561
limb	11,888,250	6,426,526 (54%)	3,950,854	2,419,480	6 reads	2,105
input DNA	39,965,419	13,481,520 (34%)	8,390,111	5,621,346	9 reads	N / A

Suppl. Table 1: Summary of ChIP-seq and mapping results.

			Observed (*in vivo* activity)								
		total transgenic experiments	forebrain	midbrain	limb	forebrain + midbrain	forebrain + limb	midbrain + limb	forebrain + midbrain + limb	positive in other only	negative
Predicted (p300 peaks)	forebrain	*31*	**18**	0	0	6	0	0	0	3	4
	midbrain	*4*	0	**3**	0	1	0	0	0	1	0
	limb	*19*	0	0	**14**	0	0	1	0	1	3
	forebrain + midbrain	*26*	0	0	0	**19**	0	0	2	1	4
	forebrain + limb	*2*	0	0	2	0	**0**	0	0	0	0
	midbrain + limb	*0*	0	0	0	0	0	**0**	0	0	0
	forebrain + midbrain + limb	*4*	1	0	0	0	0	0	**3**	0	0
	Total	***86***									

Correct predictions (Observed pattern exactly matches predicted pattern)	57	66.3%
Partial Predictions (Observed pattern matches prediction in at least one tissue, but missing or unpredicted patterns observed in other tissues)	13	15.1%
Full+partial predictions (combination of the above two categories)	70	81.4%
False positives (At least one predicted pattern was not observed)	19	22.1%
False negatives (At least one observed pattern was not predicted)	10	11.6%

Suppl. Table 6: Predicted and observed *in vivo* enhancer activities.

			feature			
			mappable portion of mouse genome (mm9)	**coding exons***	**introns**	**conserved non-coding sequences****
genome-wide			2,107,016,717bp	33,482,525bp	826,871,215bp	7,926,536bp
proportion of mappable genome			100%	1.6%	39.2%	0.38%
p300 ChIP-seq dataset	**forebrain**	total read bases	1,088,781,607bp	15,305,984bp	378,100,278bp	8,005,450bp
		proportion of read bases	100.0%	1.4%	34.7%	0.74%
		enrichment+	**1.0**	**0.9**	**0.9**	**2.0**
	midbrain	total read bases	1,059,090,633bp	15,802,623bp	368,194,852bp	5,302,923bp
		proportion of read bases	100.0%	1.5%	34.8%	0.5%
		enrichment+	**1.0**	**0.9**	**0.9**	**1.3#**
	limb	total read bases	725,839,855bp	13,215,833bp	260,851,034bp	4,563,299bp
		proportion of read bases	100.0%	1.8%	35.9%	0.63%
		enrichment+	**1.0**	**1.1**	**0.9**	**1.7**

Suppl. Table 7: Genome-wide distribution of ChIP-seq reads. *Coding exons are all complete exons or portions of exons that are translated. ** see methods for definition of conserved non-coding sequences. +Enrichment is calculated as [(fraction of ChIP-seq bases overlapping feature / bases of feature in genome) x bases in mappable genome]. The mappable genome length is estimated as 77.3% of the length of the mouse reference genome sequence (mm9, see methods). # Note that the relatively low enrichment of midbrain p300 reads at conserved non-coding sequences is consistent with fewer peaks and lower ChIP enrichment in this sample.