Published in final edited form as:
Nature. 2009 February 12; 457(7231): 854–858. doi:10.1038/nature07730.

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], and **Len A. Pennacchio**[1,2]

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

[2]US Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA.

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

Abstract

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 because they are scattered among 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 present the results of chromatin immunoprecipitation with the enhancer-associated protein p300 followed by massively parallel sequencing, and 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 demonstrated reproducible enhancer activity in the tissues that were 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 data sets 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], has 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 remain 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].

Correspondence and requests for materials should be addressed to L.A.P. (LAPennacchio@lbl.gov)..
*These authors contributed equally to this work.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

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

Reprints and permissions information is available at www.nature.com/reprints. All ChIP-seq data sets 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.

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 the 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 micro-dissected forebrain, midbrain and limb tissue from more than 150 embryonic day 11.5 (E11.5) mouse embryos and performed ChIP directly from these tissue samples using a p300 antibody (Fig. 1). Immunoprecipitated DNA fragments were analysed using massively parallel sequencing and the resulting 36 base pair (bp) 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] (Supplementary Table 1). Using an estimated false-discovery rate (FDR) threshold of <0.01, we identified 2,543, 561 and 2,105 peaks from forebrain, midbrain and limb respectively (Supplementary Tables 2-4). Most peaks were located at least 10 kb from transcript start sites (Supplementary Fig. 1). The smaller number of peaks from midbrain is probably due to variability in the efficiency of enrichment by ChIP (Supplementary 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 further binding sites can be identified that are occupied only in smaller subsets of cells within each tissue (Supplementary Fig. 3). Although most 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 (Supplementary Fig. 4).

p300 predicts enhancer activity patterns

To test directly whether 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 (Supplementary Table 5). These elements were selected blind to the identity of genes near which they are located, showed 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 (Supplementary 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 of 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 present on average in more than 80% of the embryos obtained per construct (Supplementary 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 data sets separately, 55 out of 63 (87%) forebrain predictions, 30 out of 34 (88%) midbrain predictions and 22 out of 25 (88%) limb predictions were active enhancers *in vivo* at E11.5 as defined by reproducible LacZ staining (Fig. 2, grey and coloured bars). Overall, 87% (75 out of 86) of the tested elements were reproducible enhancers at E11.5. This compares with a success rate for predicting enhancers of 47% (246 out of 528) from our previous studies in which elements were identified on the basis of 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 fourfold lower than that with extreme evolutionary conservation (13% compared to 53% previously; $P = 4.2 \times 10^{-10}$, 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). Similarly, 28 out of 34 (82%) tested elements identified by midbrain p300 enrichment (Fig. 2, red), and 20 out 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 out of 32 (84%) were active in at least one of the predicted tissues, and 22 sequences (69%) perfectly recapitulated the predicted expression patterns (Supplementary 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 on the basis of extreme sequence constraint that were not associated with a prior tissue specificity prediction[10,11]. For example, whereas forebrain enhancers account for only 16% (86 out of 528) of the tested elements identified through comparative approaches, 78% (49 out of 63) of elements predicted by forebrain p300 peaks were found to be active enhancers in the forebrain (Fig. 2). Forebrain predictions are therefore fivefold enriched in forebrain enhancers compared with enhancers identified through comparative approaches ($P < 1 \times 10^{-22}$). Similarly, we observed a sixfold enrichment of midbrain enhancers ($P < 1 \times 10^{-11}$) and a 16-fold enrichment of limb enhancers ($P < 1 \times 10^{-18}$) 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 (ref. [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 whether 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 and Supplementary Table 7). Vice versa, between 86% and 91% of the p300 peaks overlap sequences that are under evolutionary constraint in vertebrates[34], compared to less than 30% of size-matched random regions ($P < 1 \times 10^{-172}$, Fisher's exact test; Supplementary 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 ($P < 1 \times 10^{-82}$). These results indicate that most p300 peaks in the investigated tissues are under 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 neighbouring 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 overexpressed in fore-brain at E11.5 compared to whole embryos (Supplementary 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 90 kb 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 101 kb from a promoter of a forebrain-overexpressed gene. The most pronounced enrichment (4.8-fold, $P < 0.01$) was observed within 10 kb 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, Supplementary Table 9). Near genes that were overexpressed fivefold or more in the fore-brain, even higher enrichment of forebrain peaks was observed (11-fold enrichment within 10 kb from promoters, data not shown). We found similar enrichment of limb-derived p300 peaks near limb-overexpressed genes (Supplementary Fig. 6 and Supplementary 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 (ref. [23]) using ChIP-seq[17] directly from developing mouse tissues. Notably, 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 (Supplementary 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 markedly improved specificity for locating enhancers in the human genome compared to conservation-based methods[10,11] and also predicts their *in vivo* activity patterns with higher accuracy than motif-based computational methods available at present (for example, refs [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 of 113 bp) 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 data sets may be produced in the future, covering additional embryonic stages, anatomical regions and subregions, and perhaps considering extra 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 immuno-precipitation 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 500 bp 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 300 bp in the 3′ direction and used to determine the read coverage at individual nucleotides at 25-bp intervals throughout the mouse genome. p300-enriched regions (peaks) with an estimated FDR 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 probable artefacts.

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 E11.5 whole embryos and forebrain tissue was hybridized to GeneChip Mouse Genome 430 2.0 arrays (Affymetrix) and analysed 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 data set 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 (Gene Expression Omnibus (GEO) series GSE10516, samples GSM264689, GSM264690 and GSM264691)[49].

METHODS

Tissue dissection and chromatin immunoprecipitation

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 (1% formaldehyde, 10 μM NaCl, 100 mM EDTA, 50 μM EGTA, 5 mM HEPES, pH 8.0) for 15 min at room temperature. Cross-linking was terminated by the addition of 125 mM glycine and cells were dissociated in a glass douncer. Chromatin isolation, sonication and immunoprecipitation were performed as previously described[40,46].In brief, 1 mg of sonicated chromatin (OD260) was incubated with 10 μg of antibody (rabbit polyclonal anti-p300 (C-20), 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, 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 buffer (10 mM Tris, pH 8.0, 1 mM EDTA). After washing, bound DNA was eluted at 65 °C in elution buffer (10 mM Tris, pH 8.0, 1 mM EDTA and 1% SDS) for 10 min and incubated at 65 °C overnight to reverse cross-links. After the reversal of cross-linking, immunoprecipitated DNA was treated sequentially with proteinase K and 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 min at 55% power output and 90% amplitude. Sheared ChIP DNA extract was end-repaired using the End-It DNA End-Repair Kit (Epicentre). Illumina adapters (56 bp and 34 bp) were ligated using T4 DNA ligase (5 U μl^{-1}, Fermentas) and recovered using a MinElute Reaction Cleanup Kit (Qiagen). Linker ligated ChIP DNA was amplified by emulsion PCR for 40 cycles as previously described[47]. Amplified ChIP DNA between 300 and 500 bp 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 adaptor 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 36-bp 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 >1 bp 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 artefact 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 300 bp 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 25 bp 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 previously[17]. In brief, the probability of observing a peak with a coverage depth of at least H reads is given by a sum of Poisson probabilities as:

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

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

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 (FDR ≤ 0.01). Candidate peaks were identified as sites in which the 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 two reads were merged into a single peak. Candidate peaks mapping to chr_random contigs, centromeric regions, telomeric regions, segmental duplications, satellite repeats, ribosomal RNA repeats or regions of >70% repeat sequence, and those coinciding with enriched regions in the control sample (input DNA) were removed as probable artefacts due to misalignment of heterochromatic sequences that are not at present 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 data sets with respect to nearby genes (UCSC known genes[50]), internal exons (mouse RefSeq51 exons >30 kb 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 $R = 50$; refs [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 of 457 for all peaks versus 626 for tested peaks). Enhancer candidate regions (average size of 2.4 kb) 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 Supplementary Table 5. Transgenic mouse embryos were generated by pronuclear injection and F_0 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 Supplementary 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 forebrain/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 complementary RNA, 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 complementary DNA 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 (Supplementary Tables 8 and 9). The same procedure was used to analyse 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 hind-limb gene expression data (GEO series GSE10516, samples GSM264689, GSM264690 and GSM264691)[49], with the whole embryo gene expression data generated in the present study (Supplementary 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We wish to thank R. Hosseini and S. Phouanenavong for technical support, and J. Rubenstein, J. Long, J. Choi and Y. 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. and 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. B.R. was supported by grants from the National Human Genome Research Institute and the Ludwig Institute for Cancer Research.

References

1. Venter JC, et al. The sequence of the human genome. Science 2001;291:1304–1351. [PubMed: 11181995]
2. Lander ES, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921. [PubMed: 11237011]
3. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol 1997;268:78–94. [PubMed: 9149143]
4. Okazaki Y, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 2002;420:563–573. [PubMed: 12466851]
5. Marshall H, et al. A conserved retinoic acid response element required for early expression of the homeobox gene *Hoxb-1*. Nature 1994;370:567–571. [PubMed: 7914354]
6. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. Scanning human gene deserts for long-range enhancers. Science 2003;302:413. [PubMed: 14563999]
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 2005;15:1061–1072. [PubMed: 16024824]
8. Woolfe A, et al. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 2005;3:e7. [PubMed: 15630479]
9. Prabhakar S, et al. Close sequence comparisons are sufficient to identify human cis-regulatory elements. Genome Res 2006;16:855–863. [PubMed: 16769978]
10. Pennacchio LA, et al. *In vivo* enhancer analysis of human conserved non-coding sequences. Nature 2006;444:499–502. [PubMed: 17086198]
11. Visel A, et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nature Genet 2008;40:158–160. [PubMed: 18176564]
12. Holland LZ, et al. The amphioxus genome illuminates vertebrate origins and cephalochordate biology. Genome Res 2008;18:1100–1111. [PubMed: 18562680]
13. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007;447:799–816. [PubMed: 17571346]
14. Margulies EH, et al. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res 2007;17:760–774. [PubMed: 17567995]
15. Cooper GM, Brown CD. Qualifying the relationship between sequence conservation and molecular function. Genome Res 2008;18:201–205. [PubMed: 18245453]
16. McGaughey DM, et al. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. Genome Res 2008;18:252–260. [PubMed: 18071029]
17. Robertson G, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods 2007;4:651–657. [PubMed: 17558387]
18. Mikkelsen TS, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 2007;448:553–560. [PubMed: 17603471]
19. Robertson AG, et al. Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. Genome Res 2008;18:1906–1917. [PubMed: 18787082]
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 2009;19:24–32. [PubMed: 19056695]
21. Wederell ED, et al. Global analysis of *in vivo* Foxa2-binding sites in mouse adult liver using massively parallel sequencing. Nucleic Acids Res 2008;36:4549–4564. [PubMed: 18611952]
22. Valouev A, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nature Methods 2008;5:829–834. [PubMed: 19160518]
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 1994;8:869–884. [PubMed: 7523245]

24. Eckner R, Yao TP, Oldread E, Livingston DM. Interaction and functional collaboration of p300/CBP and bHLH proteins in muscle and B-cell differentiation. Genes Dev 1996;10:2478–2490. [PubMed: 8843199]

25. Yao TP, et al. Gene dosage-dependent embryonic development and proliferation defects in mice lacking the transcriptional integrator p300. Cell 1998;93:361–372. [PubMed: 9590171]

26. Merika M, Williams AJ, Chen G, Collins T, Thanos D. Recruitment of CBP/p300 by the IFNb enhanceosome is required for synergistic activation of transcription. Mol. Cell 1998;1:277–287. [PubMed: 9659924]

27. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. Annu. Rev. Genomics Hum. Genet 2006;7:29–59. [PubMed: 16719718]

28. Heintzman ND, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genet 2007;39:311–318. [PubMed: 17277777]

29. Xi H, et al. Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. PLoS Genet 2007;3:e136. [PubMed: 17708682]

30. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature 2002;420:520–562. [PubMed: 12466850]

31. Kothary R, et al. A transgene containing lacZ inserted into the dystonia locus is expressed in neural tube. Nature 1988;335:435–437. [PubMed: 3138544]

32. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic Acids Res 2007;35:D88–D92. [PubMed: 17130149]

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 2008;18:1896–1905. [PubMed: 18818370]

34. Siepel A, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 2005;15:1034–1050. [PubMed: 16024819]

35. Hallikas O, et al. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell 2006;124:47–59. [PubMed: 16413481]

36. Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I. Predicting tissue-specific enhancers in the human genome. Genome Res 2007;17:201–211. [PubMed: 17210927]

37. Kim TH, et al. A high-resolution map of active promoters in the human genome. Nature 2005;436:876–880. [PubMed: 15988478]

38. Boyle AP, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell 2008;132:311–322. [PubMed: 18243105]

39. Schones DE, Zhao K. Genome-wide approaches to studying chromatin modifications. Nature Rev. Genet 2008;9:179–191. [PubMed: 18250624]

40. Barrera LO, et al. Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. Genome Res 2008;18:46–59. [PubMed: 18042645]

41. Chen X, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 2008;133:1106–1117. [PubMed: 18555785]

42. Kwok RP, et al. Nuclear protein CBP is a coactivator for the transcription factor CREB. Nature 1994;370:223–226. [PubMed: 7913207]

43. Ogryzko VV, Schiltz RL, Russanova V, Howard BH, Nakatani Y. The transcriptional coactivators p300 and CBP are histone acetyltransferases. Cell 1996;87:953–959. [PubMed: 8945521]

44. Agalioti T, et al. Ordered recruitment of chromatin modifying and general transcription factors to the IFN-β promoter. Cell 2000;103:667–678. [PubMed: 11106736]

45. Ge K, et al. Transcription coactivator TRAP220 is required for PPARγ2-stimulated adipogenesis. Nature 2002;417:563–567. [PubMed: 12037571]

46. Li Z, et al. A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. Proc. Natl Acad. Sci. USA 2003;100:8164–8169. [PubMed: 12808131]

47. Blow MJ, et al. Identification of the source of ancient remains through genomic sequencing. Genome Res 2008;18:1347–1353. [PubMed: 18426903]

48. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res 2002;12:656–664. [PubMed: 11932250]

49. Krawchuk D, Kania A. Identification of genes controlled by LMX1B in the developing mouse limb bud. Dev. Dyn 2008;237:1183–1192. [PubMed: 18351676]

50. Karolchik D, et al. The UCSC Genome Browser Database: 2008 update. Nucleic Acids Res 2008;36:D773–D779. [PubMed: 18086701]

51. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2007;35:D61–D65. [PubMed: 17130148]

52. Giardine B, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res 2005;15:1451–1455. [PubMed: 16169926]

53. Mikkelsen TS, et al. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. Nature 2007;447:167–177. [PubMed: 17495919]

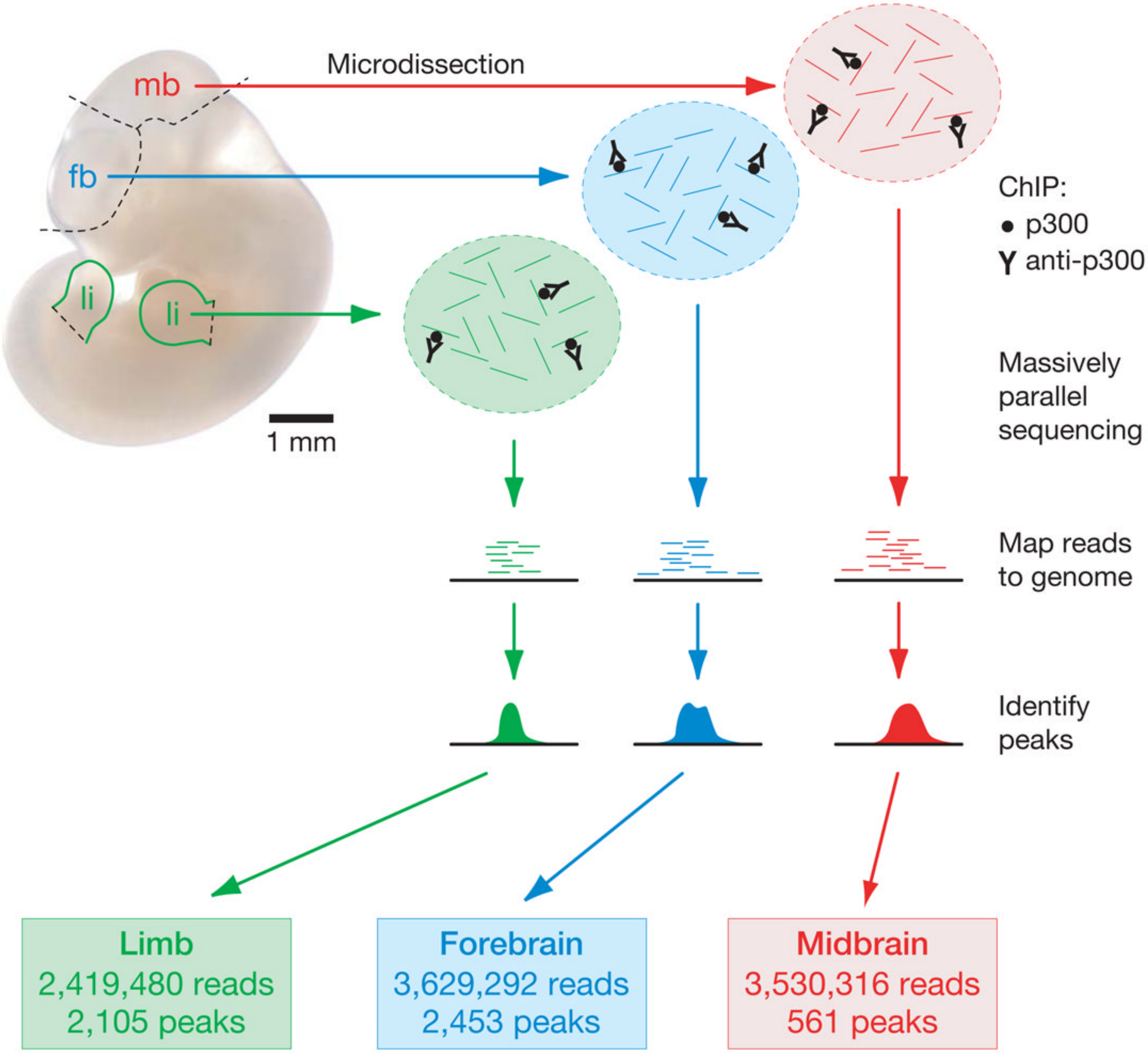

Figure 1. Tissue dissection boundaries, overview of the ChIP-seq approach and summary of p300 results

Tissue dissection boundaries are indicated in a representative unstained E11.5 mouse embryo. For each sample, tissue was pooled from more than 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 Supplementary Table 1. fb, forebrain; li, limb; mb, midbrain.

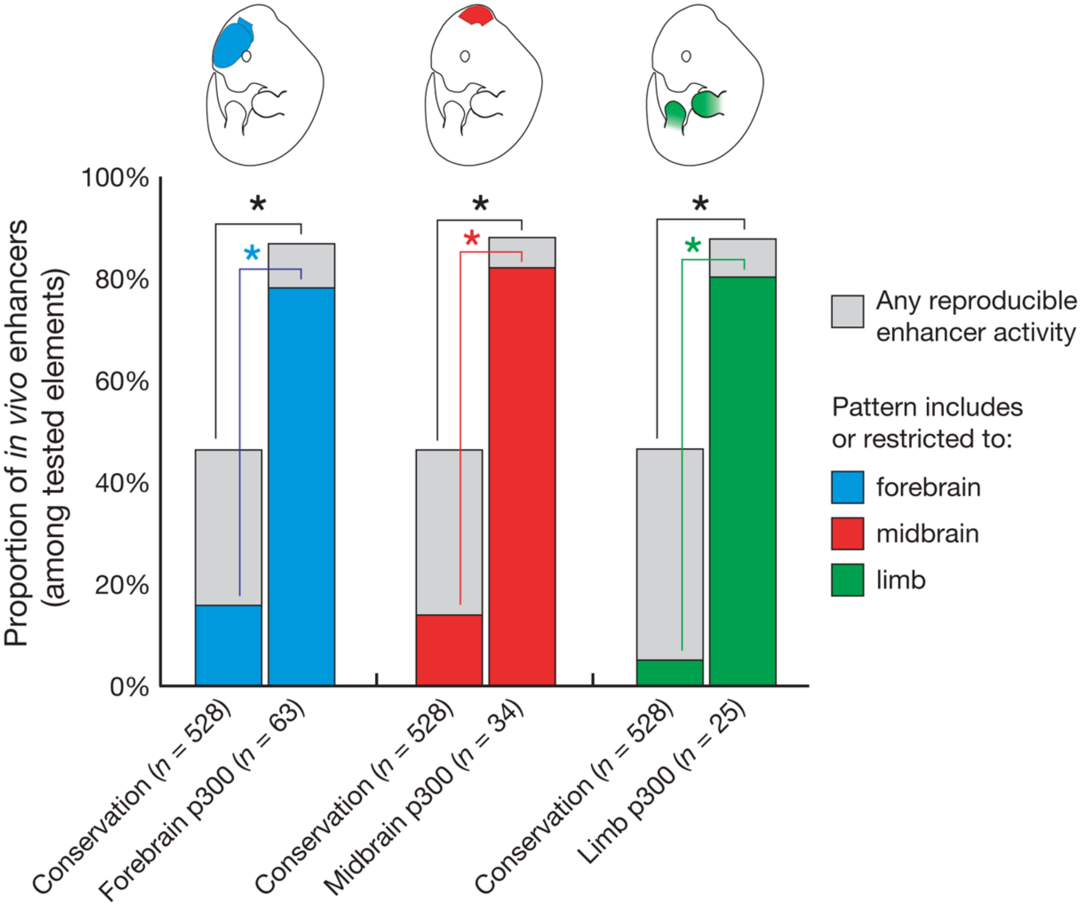

Figure 2. p300 binding accurately predicts enhancers and their tissue-specific activity patterns

Bar height indicates the frequency of *in vivo* enhancers (reproducible at E11.5) that are active in any tissue (grey bars and coloured bars) and the fraction of enhancers in which the pattern includes or is restricted to reproducible forebrain (blue bars), midbrain (red bars) or limb activity (green bars). 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 data sets from refs [10] and [11]). The component activities of elements predicted to be active in several tissues were counted separately. *$P < 0.00005$, Fisher's exact test, one-tailed.

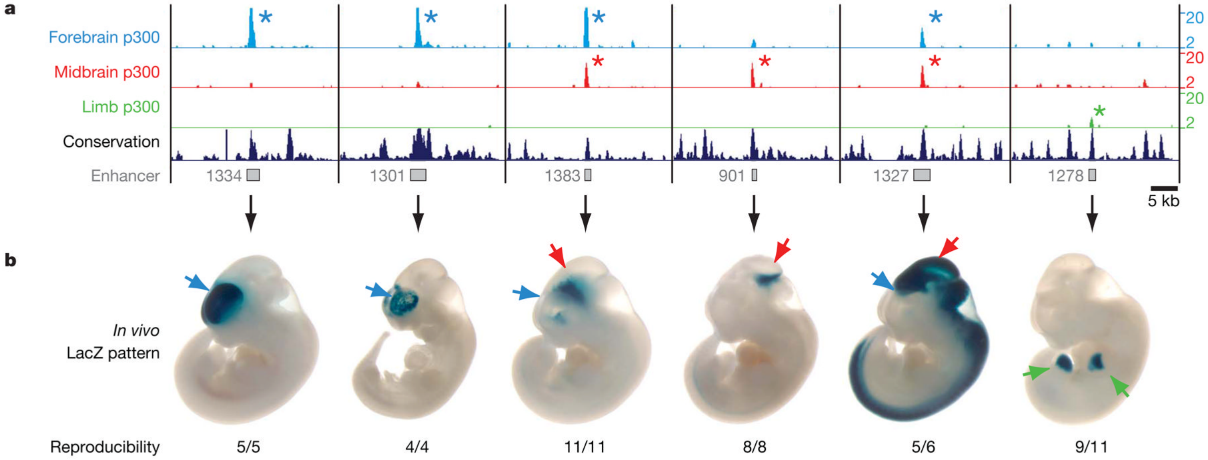

Figure 3. Examples of successful prediction of *in vivo* enhancers by p300 binding in embryonic tissues

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. Multispecies vertebrate conservation plots (black) were obtained from the UCSC genome browser[50]. Grey boxes correspond to candidate enhancer regions. Numbers at 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 the reproducibility of LacZ reporter staining. Additional embryos obtained with each construct and genomic coordinates are available using the enhancer ID in the bottom portion of **a** at the Vista Enhancer Browser[32].

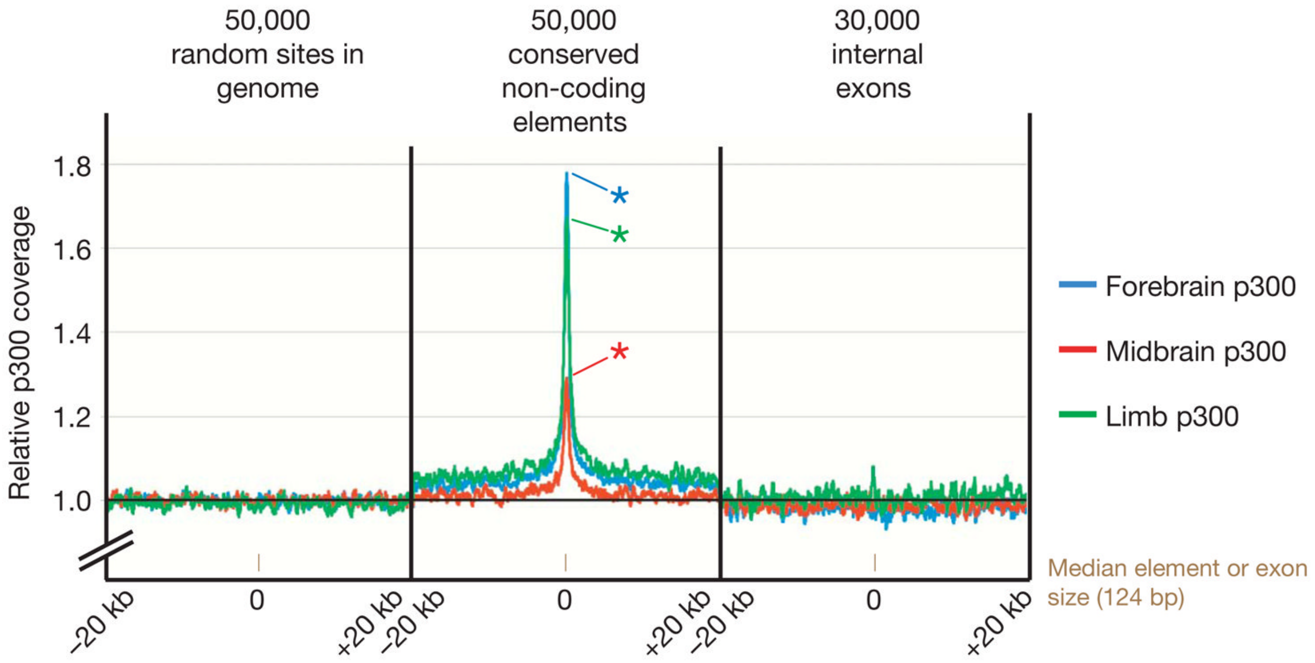

Figure 4. In all tissues examined, p300 is enriched at highly conserved non-coding regions

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. $*P < 1 \times 10^{-100}$, Fisher's exact test. Relative p300 coverage near random sites and internal exons is shown for comparison. Brown bars indicate the median sizes of conserved elements or exons (124 bp in both cases). For further details, see Supplementary Table 7.

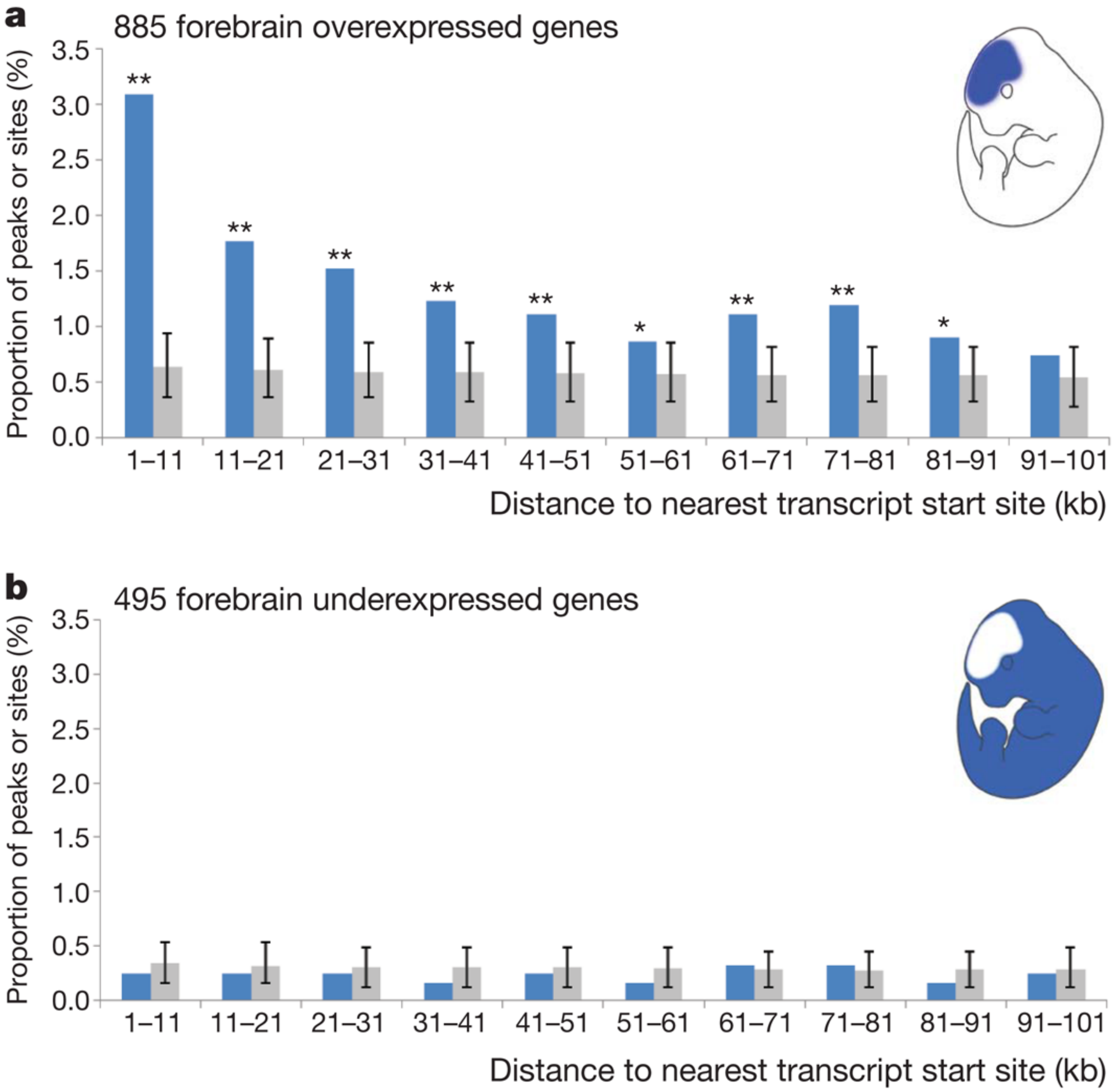

Figure 5. p300 peaks are enriched near genes that are expressed in the same tissue

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. Eight-hundred-and-eighty-five genes were forebrain-specifically overexpressed, and 495 genes were underexpressed relative to whole embryo RNA at the selected thresholds. Promoters (defined as 1 kb upstream and downstream of transcription start sites) were excluded from the analysis. Blue bars denote comparison to 2,435 forebrain-derived peaks, grey bars denote comparison to 2,435 random sites. **a,** Ten-kilobase bins up to 91 kb away from forebrain-overexpressed genes were significantly enriched in forebrain p300 peaks. **b,** No peak enrichment was observed for forebrain-underexpressed genes. Error bars indicate the 90% confidence interval on the basis of 1,000 iterations of randomized distribution. *$P < 0.05$, **$P < 0.01$, both one-tailed.