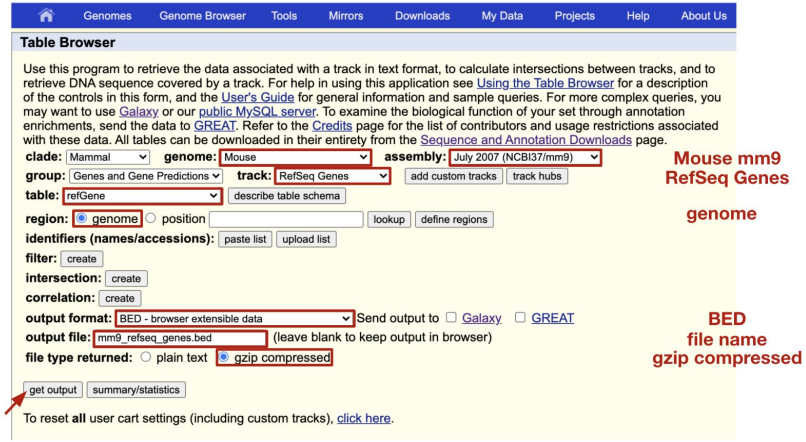
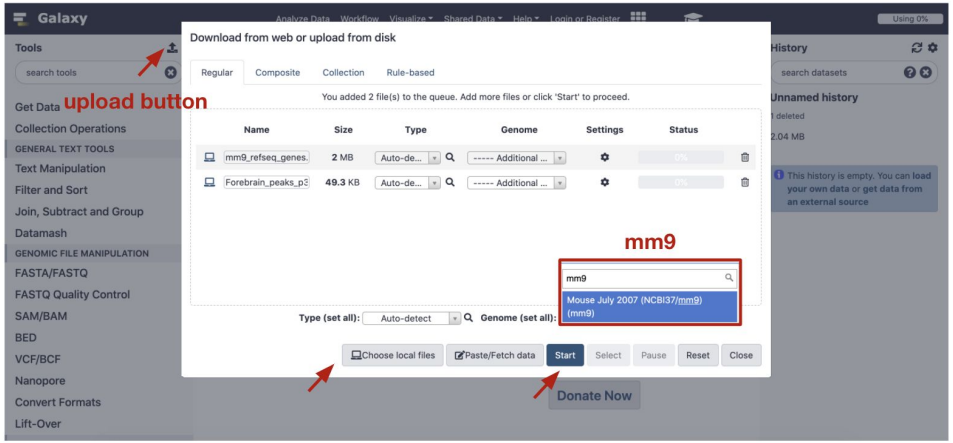
Before we start, please complete **steps 1 & 2** for the assignment

(sometimes Galaxy is cranky and takes a while to run)





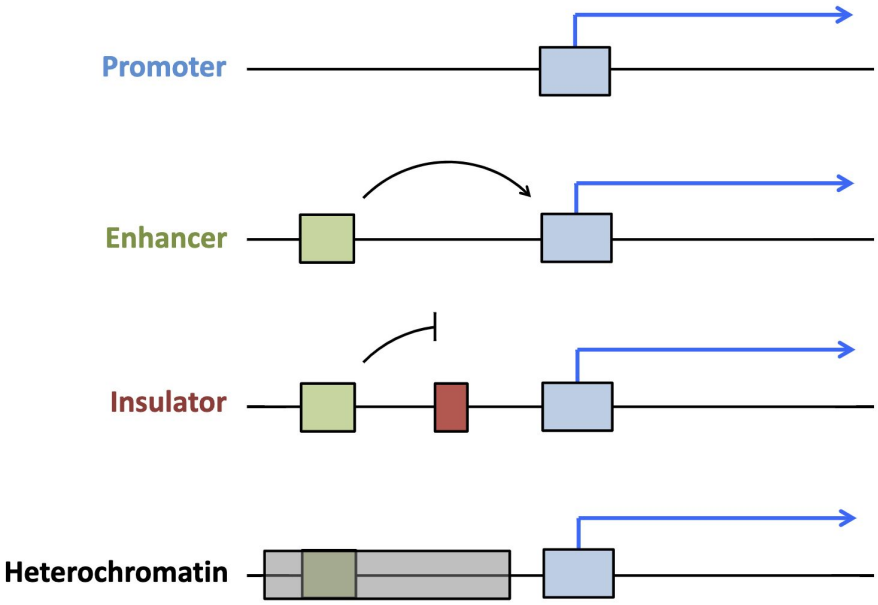
MMG1001 Genomics
Week 4 Tutorial

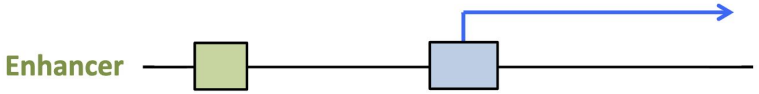
ChIP-seq accurately predicts tissue-specific activity of enhancers

TA: Heather Gibling

*Many slides borrowed and/or modified from Laura Campitelli

Enhancers are cis-acting regulatory elements





Experimental tests ("Functional analysis")	• Increases transcription in reporter assays
Genomic tests ("Biochemical analysis")	• Histone marks • Open chromatin • Associates with RNA Pol II • Associates with regulatory proteins
Evolutionary constraint ("Conservation analysis")	• Conservation enriches for enhancers...but, many enhancers (possibly most) are not conserved

Note that many promoters can function as enhancers.

**Nomenclature from Kellis et al., PNAS 2014*

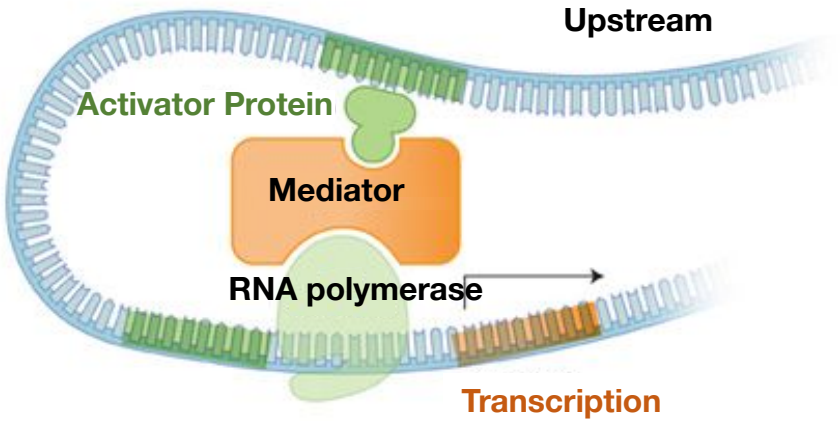
Tim Hughes

Enhancers are cis-acting regulatory elements

Enhancer: DNA sequence bound by activator proteins to increase gene transcription

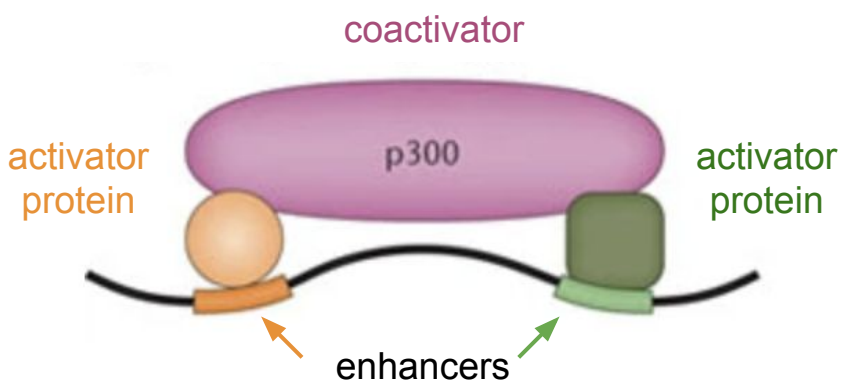
Can be various distances from target promoters

Active enhancers vary by cell type (cell-specific transcriptional profiles)



nature.com/scitable/content/modulation-of-transcription-14711116

p300 is a coactivator protein required for embryonic development



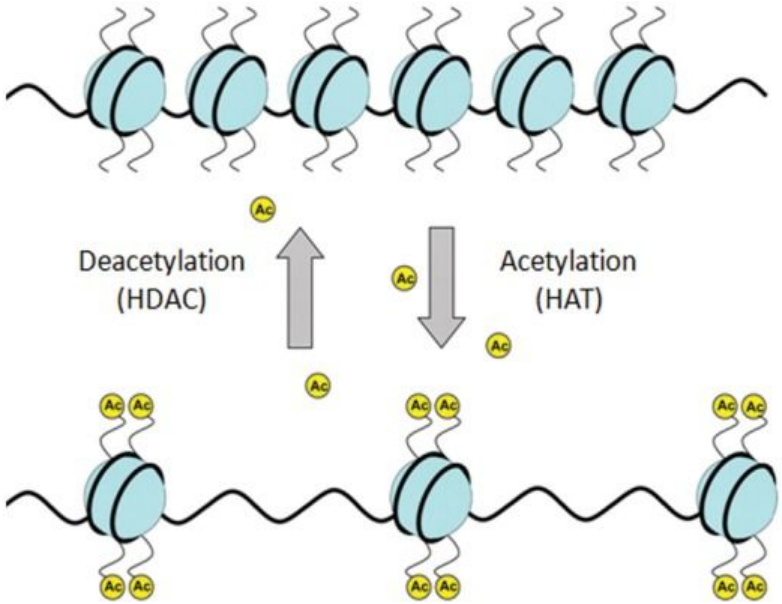
Coactivator: transcriptional coregulator that binds to transcription factors

p300 regulates transcription via **chromatin remodeling**

Acts as a **histone acetyltransferase** (has HAT domain)

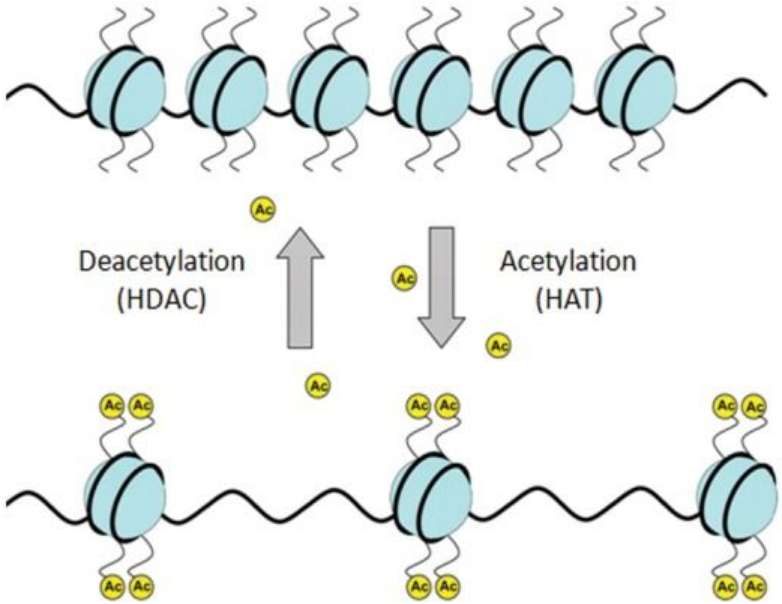
Spitz (2012) Nature Reviews Genetics

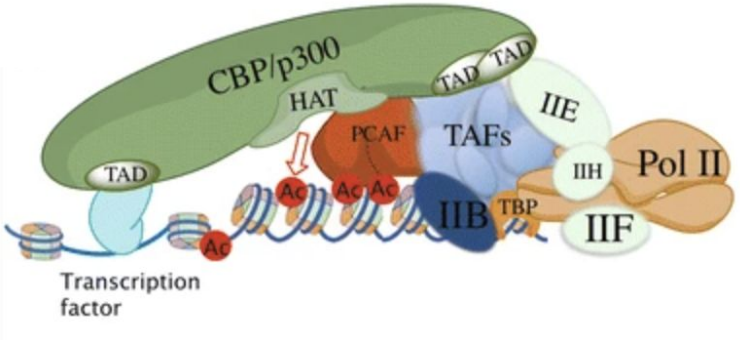
Histone acetylation relaxes chromatin and allows for transcription



Eslaminejad *et al*. (2013) Cell Journal

Histone acetylation relaxes chromatin and allows for transcription



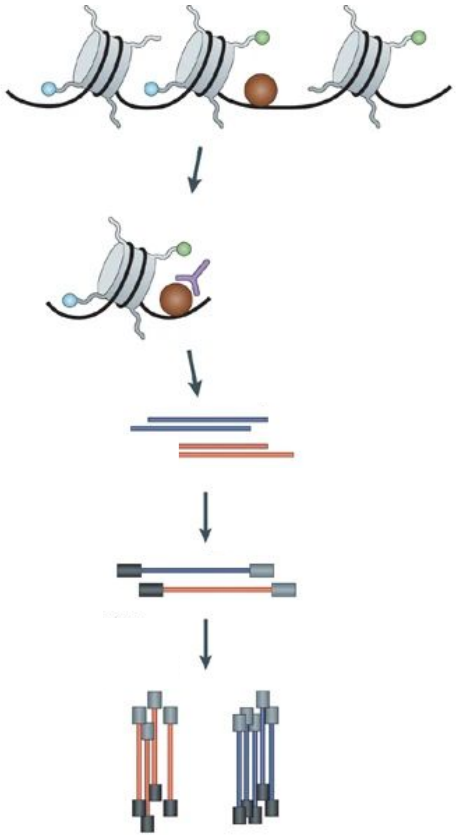


p300 opens chromatin and bridges DNA-bound transcription factors to transcription machinery

Eslaminejad *et al*. (2013) Cell Journal

Wang *et al*. (2013) Cellular and Molecular Life Sciences

ChIP-Seq: Chromatin Immunoprecipitation Sequencing



Crosslink DNA and proteins

Fragment samples
(sonication, endonucleases)

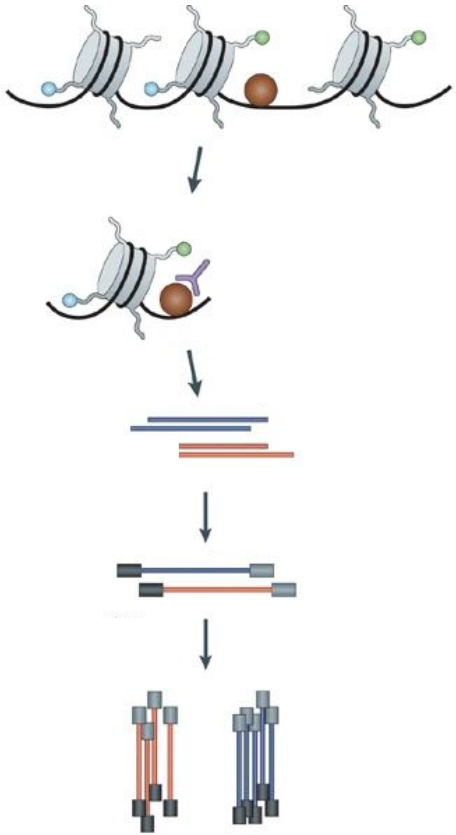
Immunoprecipitate
target protein

Reverse crosslinks
and purify DNA

Prepare and sequence
remaining DNA

Modified from Park (2009) Nature Reviews Genetics

ChIP-Seq: Chromatin Immunoprecipitation Sequencing



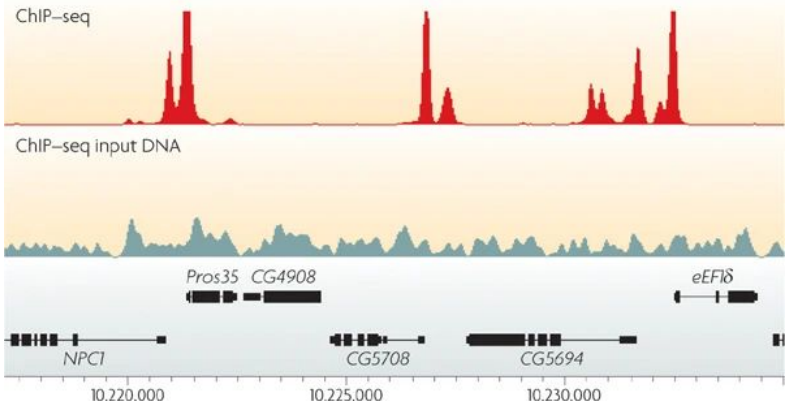
Crosslink DNA and proteins

Fragment samples
(sonication, endonucleases)

Immunoprecipitate
target protein

Reverse crosslinks
and purify DNA

Prepare and sequence
remaining DNA



Alignment of reads to reference genome results in **peaks** that reflect where target protein was bound

More reads at a locus ≃ higher peaks (compared to control input DNA)

Modified from Park (2009) Nature Reviews Genetics

nature Vol 457|12 February 2009|doi:10.1038/nature07730

ARTICLES

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]

What are the main goals of the paper?

What are the main goals of the paper?

- Identify location and timing of enhancers in different tissues
  - “Evolutionary constraint of non-coding sequences can predict the location of enhancers in the genome, but does not reveal when and where these enhancers are active *in vivo*.”
  - “...a substantial proportion of regulatory elements is not sufficiently conserved to be detectable by comparative genomic methods.”
- Expand on *in vitro* results that show p300 associates with enhancers
- **See if ChIP-seq can identify tissue-specific location and activity of p300 in embryonic mouse forebrain, midbrain, and limb**

Why p300?

How did they do it?

How did they do it?

- **Tissue dissections** of forebrain, midbrain, and limb from E11.5 mouse embryos
- **ChIP-seq** targeting **p300** on DNA from the tissues
- Compare locations of peaks between tissues and to available **conservation**-identified locations
- Compare locations of peaks to **microarray gene expression results** to determine proximity of enhancers to expressed genes
- **Transgenic mouse reporter assays** to validate ChIP-seq predictions of p300 activity

Why E11.5 (and not E11 or E12)?

What were the major findings?

Figure 1

ChIP-seq against p300 in 3 embryonic tissues

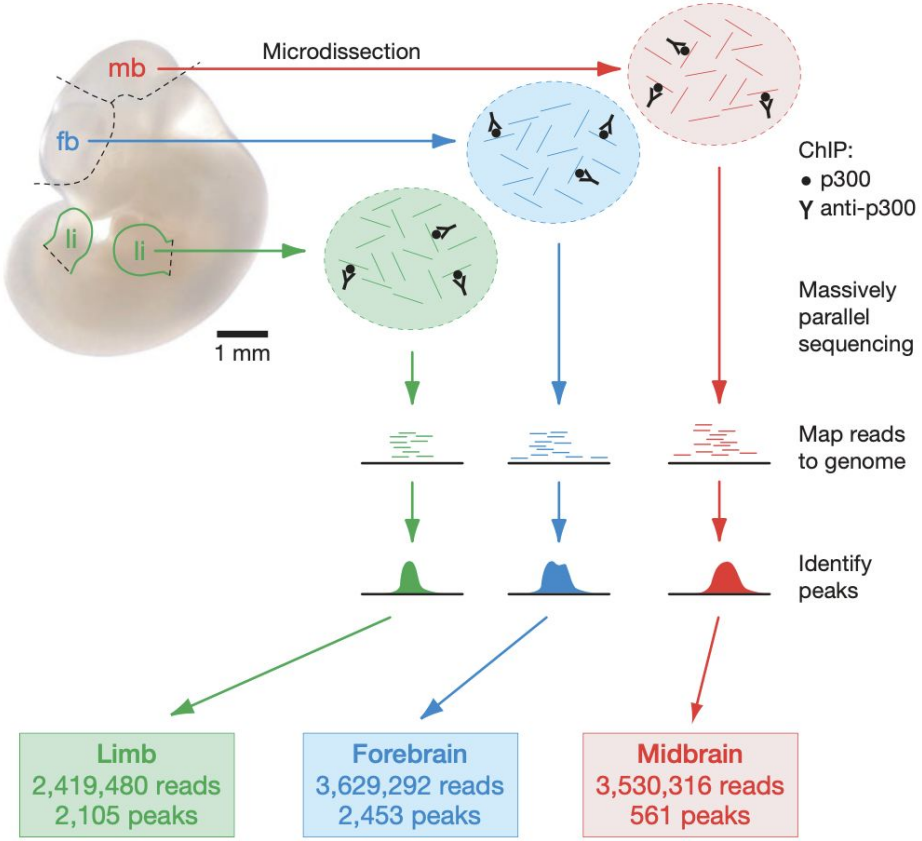
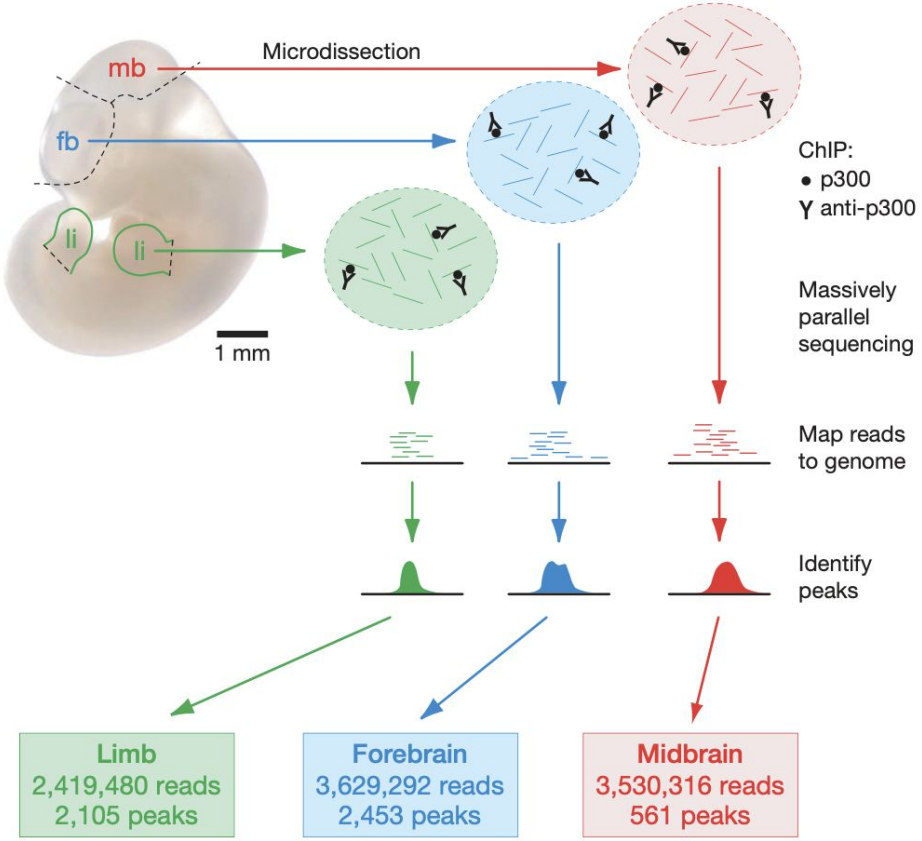


Figure 1

ChIP-seq against p300 in 3 embryonic tissues

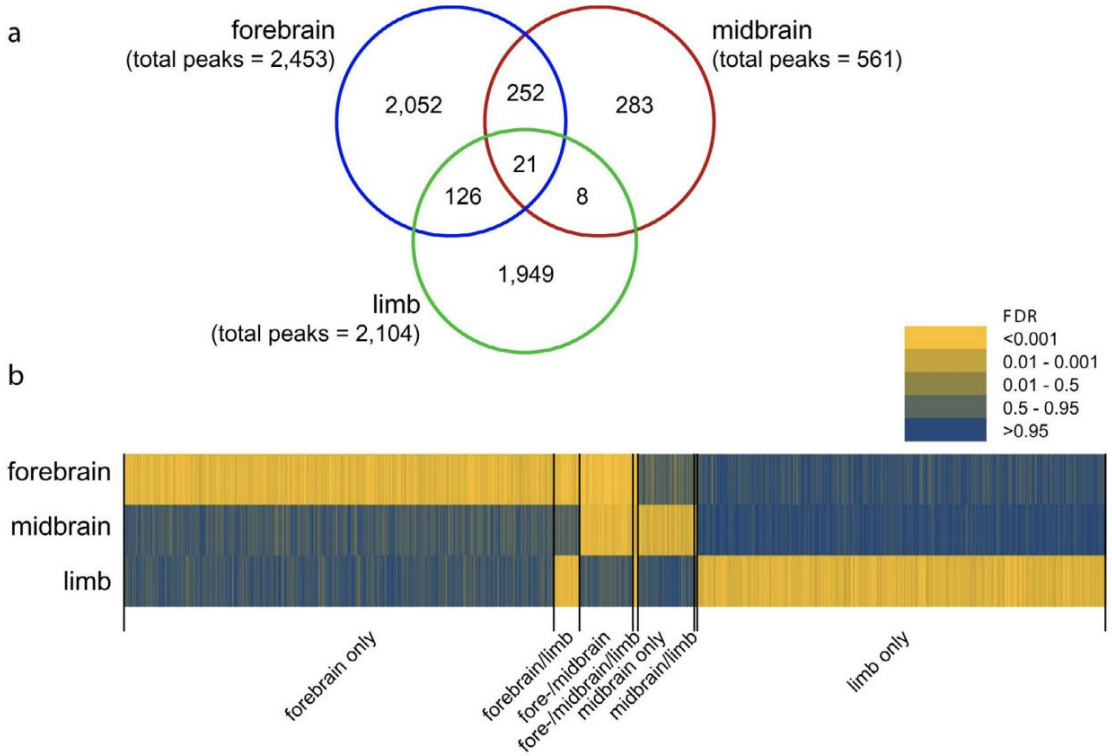


Dissections for forebrain, midbrain, and limbs in E11.5 mouse embryos

Each sample involved pooled tissue from more than 150 embryos

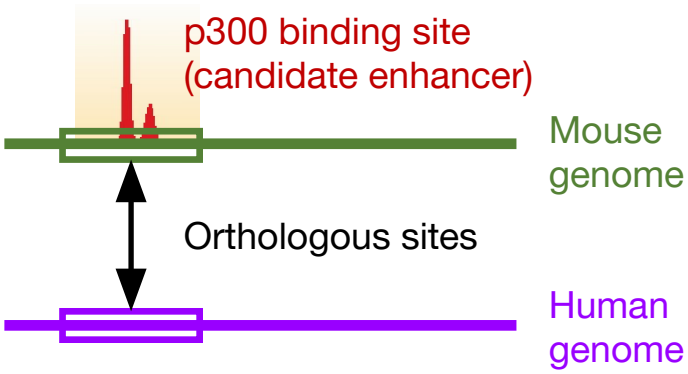
Why were there so few midbrain peaks compared to limb and forebrain peaks?

Majority of peaks were tissue-specific

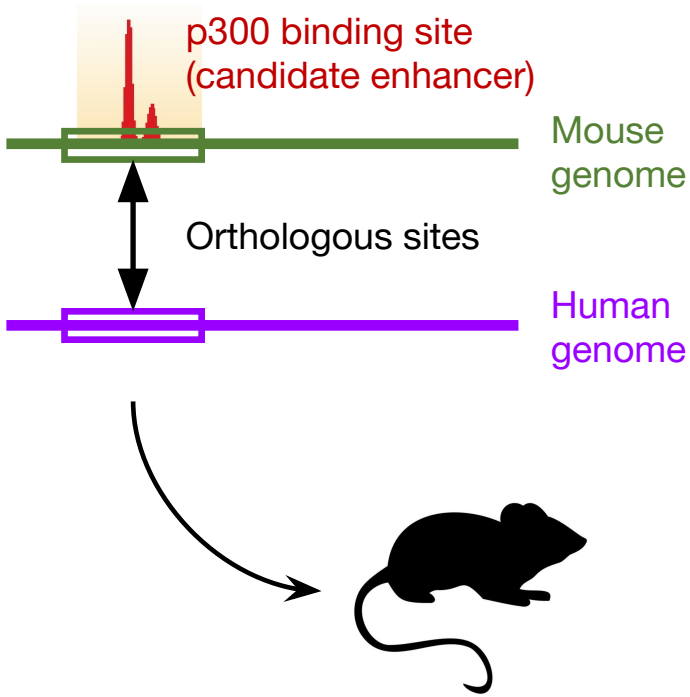


Does it make sense that there were more shared peaks between forebrain and midbrain than forebrain and limb?

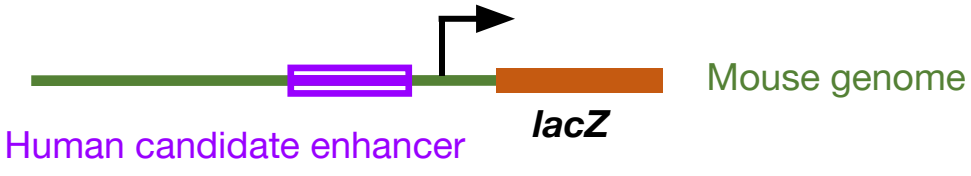
Candidate enhancers were tested in a reporter assay using orthologous human sequence



Candidate enhancers were tested in a reporter assay using orthologous human sequence

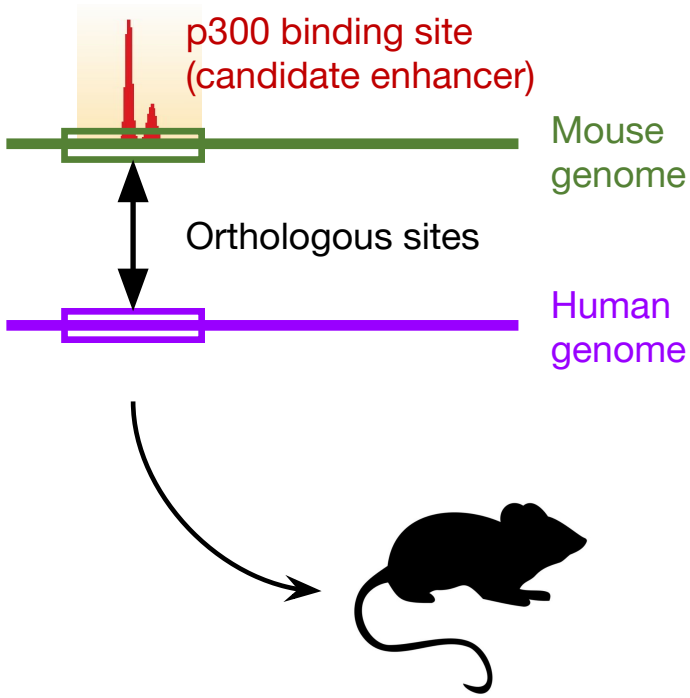


Generate transgenic mouse that expresses **lacZ** if the **human** region orthologous to the mouse p300 binding site is a sufficient enhancer (at another locus)

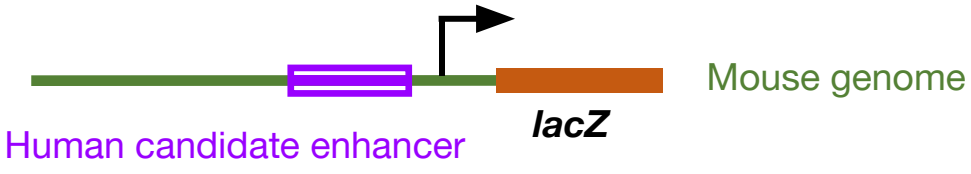


Laura Campitelli

Candidate enhancers were tested in a reporter assay using orthologous human sequence



Generate transgenic mouse that expresses **lacZ** if the **human** region orthologous to the mouse p300 binding site is a sufficient enhancer (at another locus)



Blue = lacZ expressed
= enhancer is **active** in this tissue

Laura Campitelli

Figure 2

p300 binding accurately predicts enhancers and their tissue-specific activity

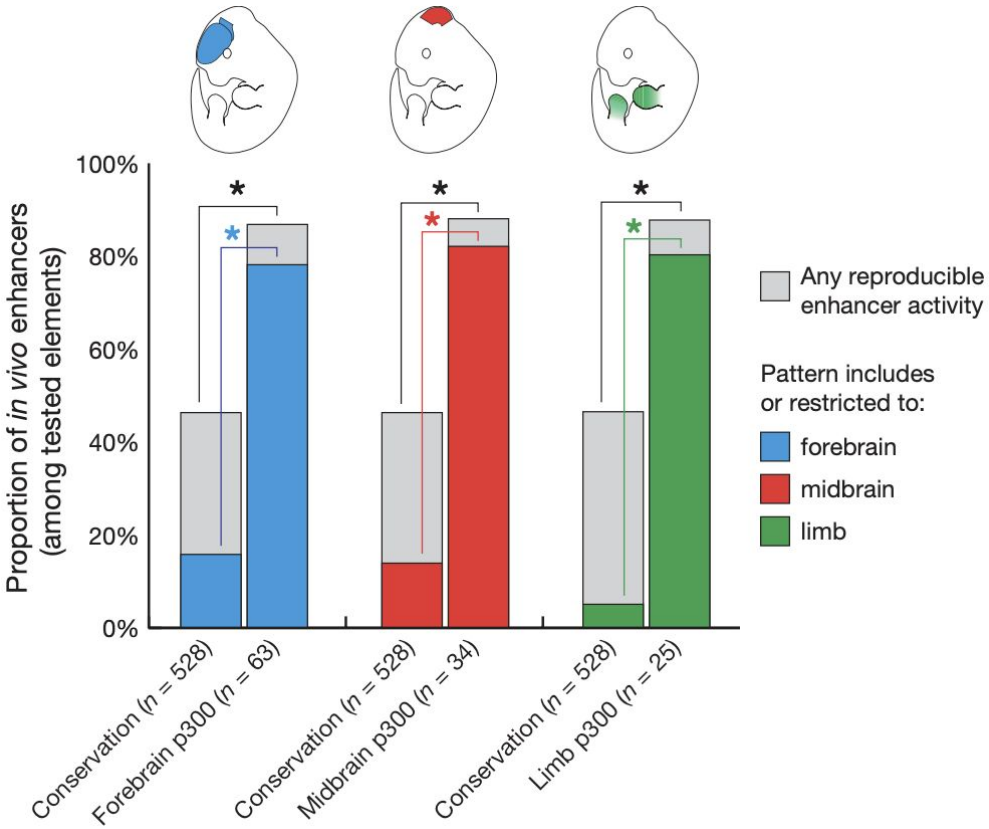


Figure 2

p300 binding accurately predicts enhancers and their tissue-specific activity

528 previously tested sequences identified through **conservation**

87% of p300 predicted enhancers **reproducible** in transgenic embryos, compared to 47% of conservation predicted enhancers

69% of tested sequences **perfectly demonstrated** the **tissue-specific activity** predicted by p300 binding

Why doesn't conservation predict tissue-specific patterns?

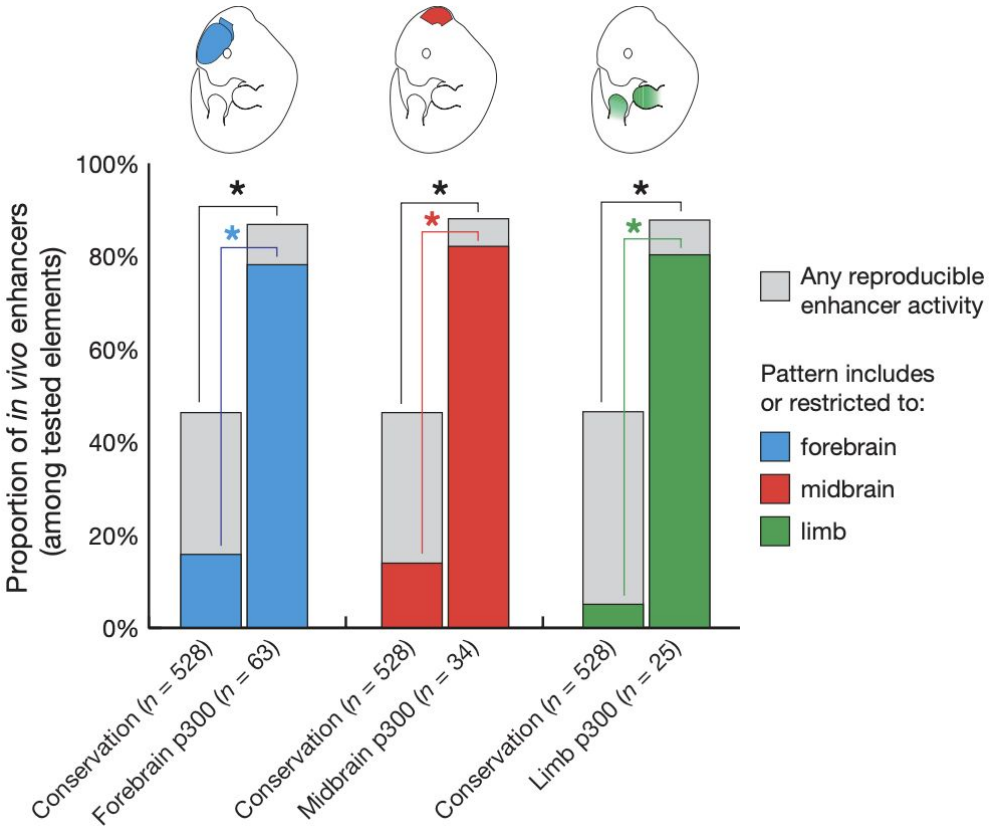


Figure 3

p300 binding successfully predicted many tissue-specific enhancers (validated *in vivo*)

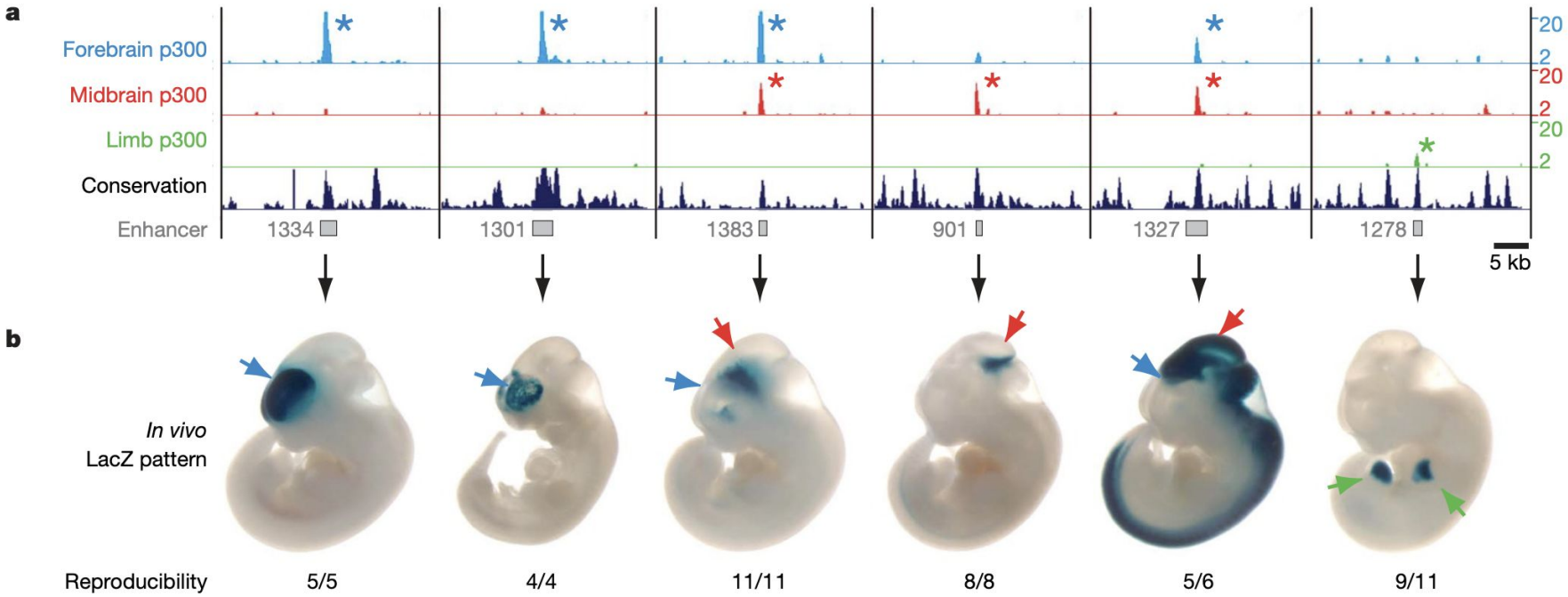
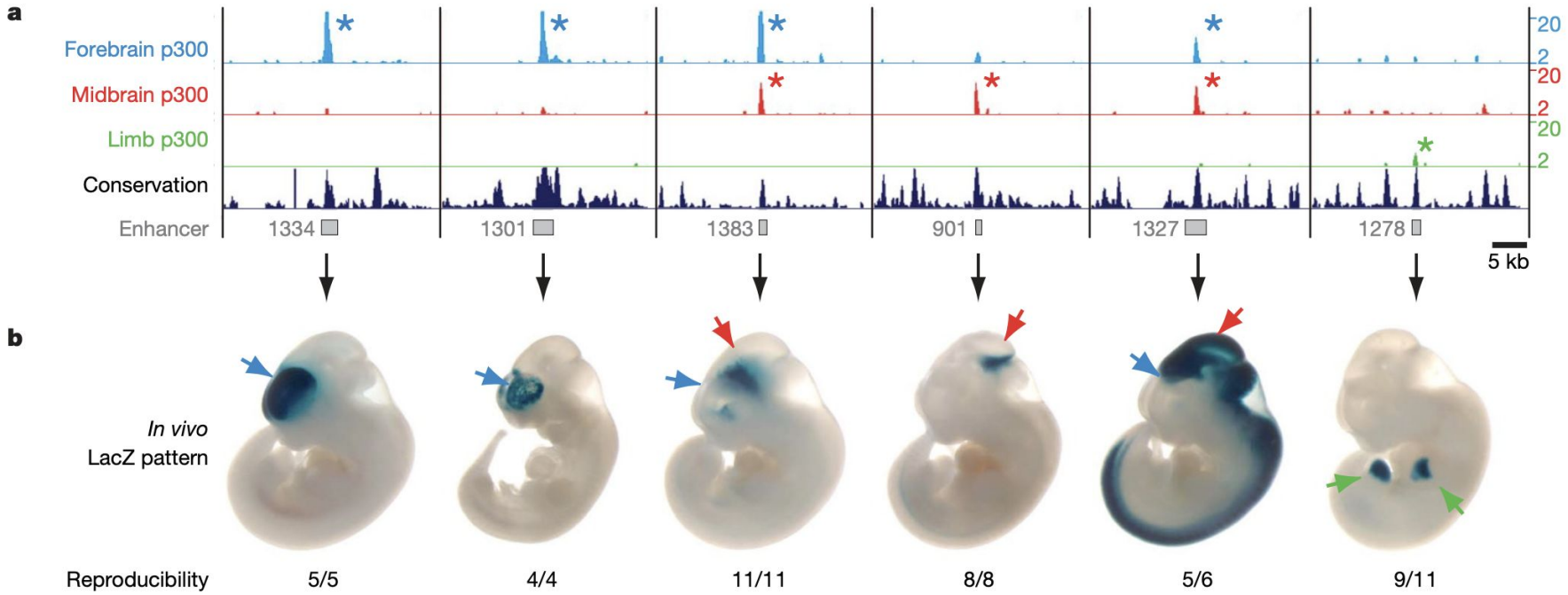


Figure 3

p300 binding successfully predicted many tissue-specific enhancers (validated *in vivo*)



Are these examples of enhancers that can be predicted by conservation?

Figure 4

p300 binding sites are enriched in highly conserved non-coding regions

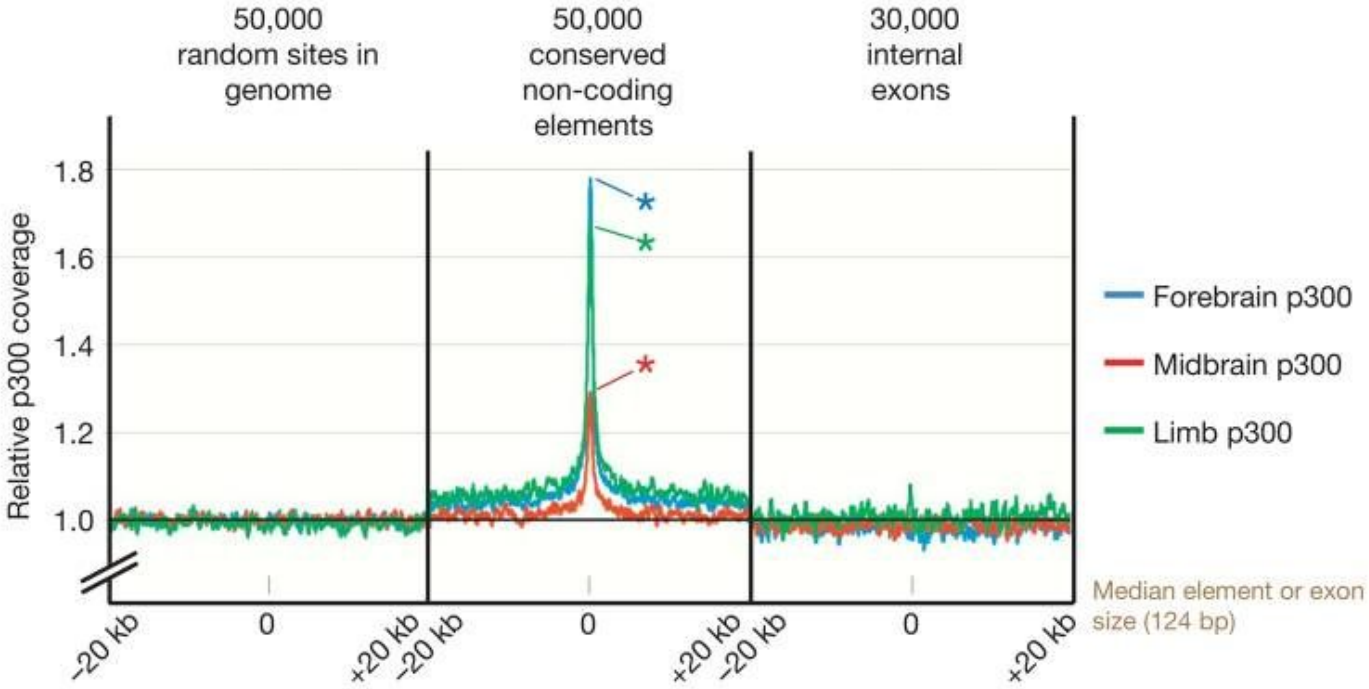
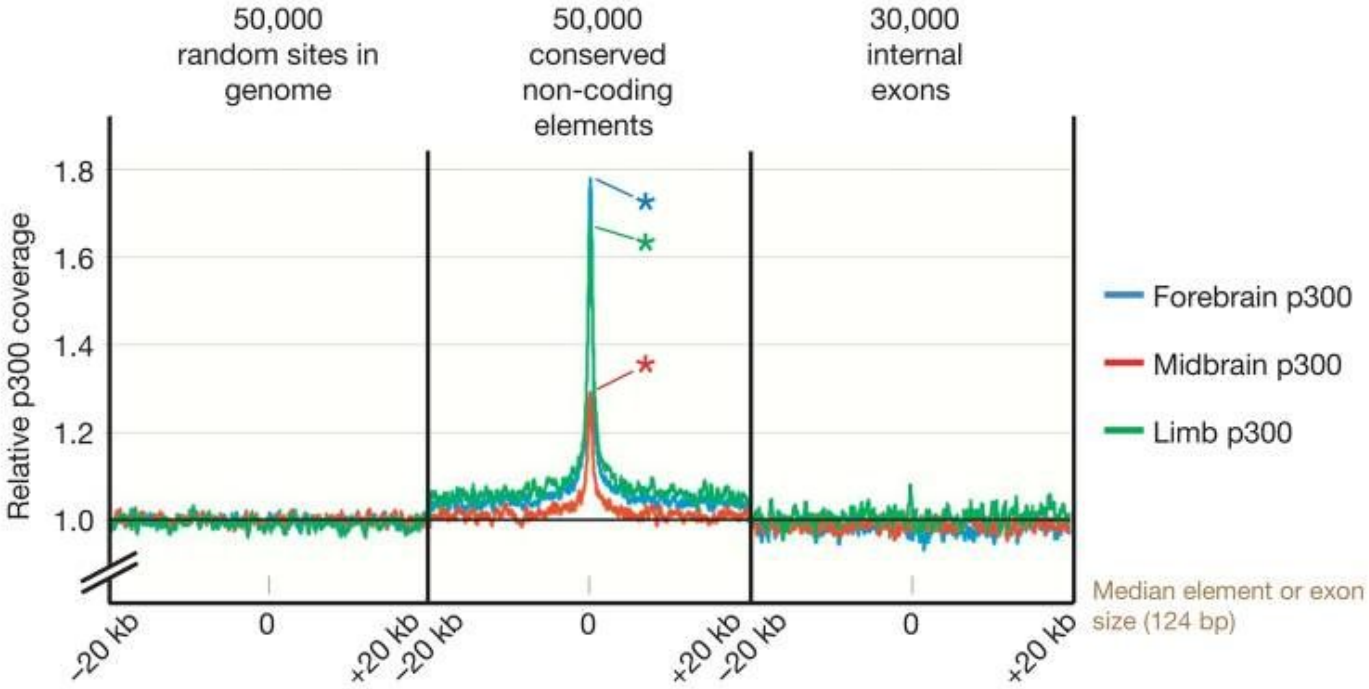


Figure 4

p300 binding sites are enriched in highly conserved non-coding regions

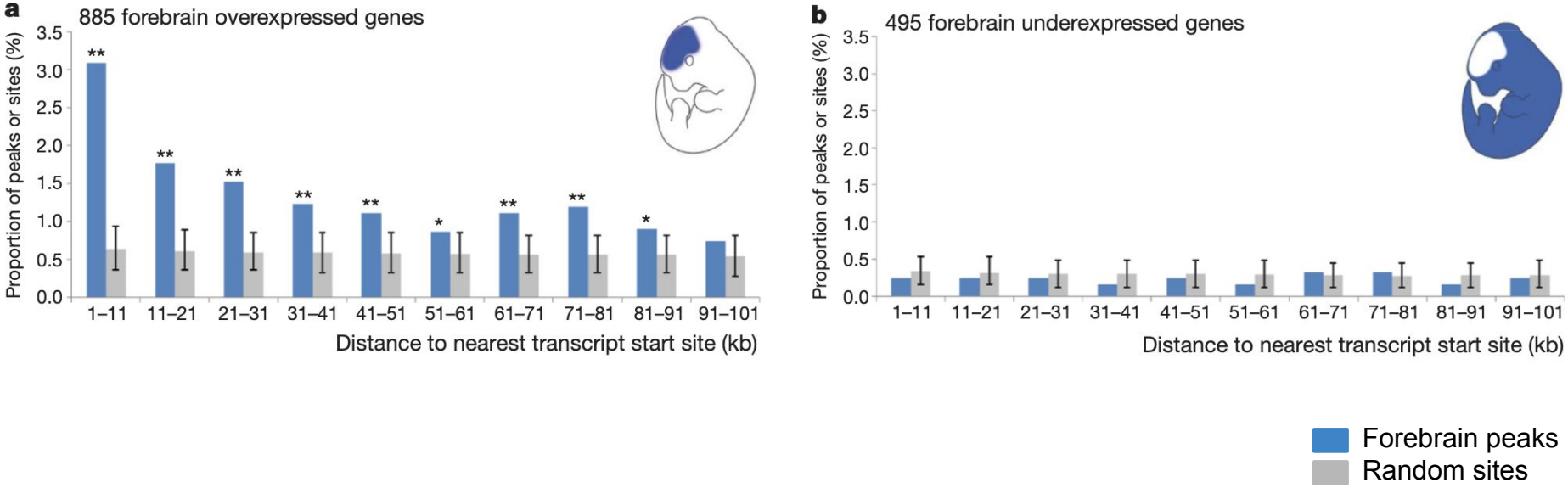


Not all active enhancers are under detectable evolutionary constraint, but most p300 sites detected are (86-91%, compared to 30% of random sites)

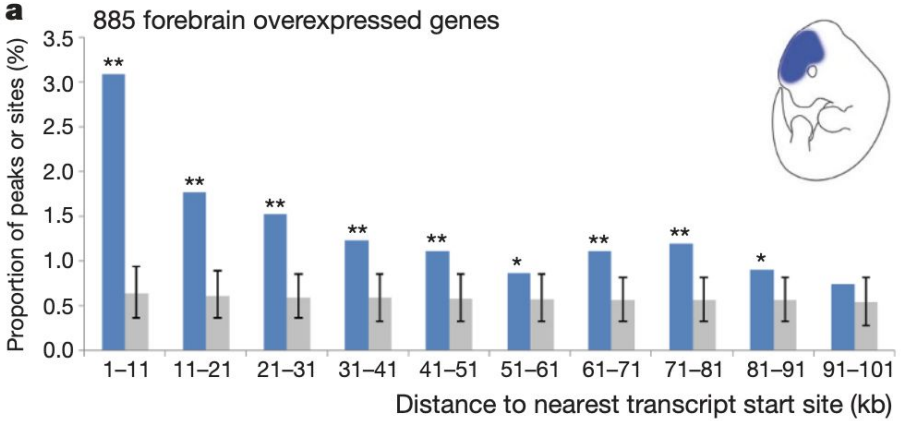
How many peaks were at loci that are highly constrained?

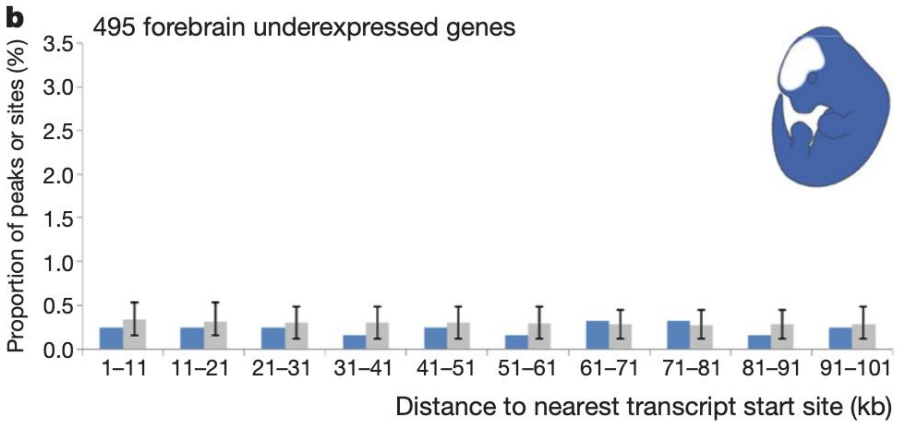
Figure 5

p300 peaks are enriched near genes expressed in the same tissue



p300 peaks are enriched near genes expressed in the same tissue





Compared p300 peaks with forebrain **microarray gene expression** results (885 overexpressed genes compared to whole embryos)

Forebrain peaks
Random sites

Majority of peaks occur **within 10kb** of overexpressed genes

What does this suggest about enhancer locations in the genome?

Summary of results

- Profiled p300 occupancy using ChIP-seq in 3 embryonic mouse tissues
- Confirmed hypothesis that bound sites generally represent **enhancers**, and confirmed many of them as having **tissue-specific enhancer activity** using a transgenic mouse lacZ expression assay
- p300 occupancy turned out to be a better predictor of enhancers than non-coding sequence conservation
    - p300 occupancy can make **more accurate** predictions of enhancers than conservation
    - p300 occupancy can be used to make **tissue-specific** predictions

Why does it matter?

Why does it matter?

- Provided insights into tissue-specific p300 enhancer activity not then available with *in vitro* cell culture experiments
- Paved the way for more *in vitro* regulatory element experimentation in other tissues

Assignment: Working with BED Files in Galaxy and Conducting GO Enrichment Analysis

Assignment: Working with BED Files in Galaxy and Conducting GO Enrichment Analysis

- ...what is a BED file? **Browser Extensible Data**
- Stores genomic coordinate information for specific features (SNPs, genes, etc)
- Must have at least 3 columns: **Chromosome name**, **start position**, **end position**

Supplementary Table 2: forebrain p300 peak information

	A	B	C	D	E	F
1	**Forebrain p300 Peaks (mm9)**					
2	**Chromosome**	**Start**	**End**	**Maximum Peak Height**	**Total Overlapping Reads**	**FDR**
3	chr1	5222650	5223201	7	9	2.6E-03
4	chr1	6719650	6720626	7	16	2.6E-03
5	chr1	11990575	11991426	9	13	1.6E-05
6	chr1	12045650	12046551	19	30	<1.0E-10
7	chr1	12400025	12400476	7	8	2.6E-03
8	chr1	13851050	13851676	7	10	2.6E-03
9	chr1	17106600	17107201	8	10	2.2E-04

BED file of forebrain p300 peaks

```
chr1    5222650    5223201
chr1    6719650    6720626
chr1    11990575   11991426
chr1    12045650   12046551
chr1    12400025   12400476
chr1    13851050   13851676
chr1    17106600   17107201
```

Assignment: Working with BED Files in Galaxy and Conducting GO Enrichment Analysis

- ...what is a BED file? **Browser Extensible Data**
- Stores genomic coordinate information for specific features (SNPs, genes, etc)
- Must have at least 3 columns: **Chromosome name**, **start position**, **end position**
- Can also have specific columns with more information (usually used for genome browsers like UCSC)

BED file for mouse genes in the mm9 reference genome

chr1	174056885	174066628	NM_001356514	0	+	174056909	174064421	0	7	94,110,166,86,162,177,2291, 0,1876,2400,3147,3857,6375,7452,
chr1	174056885	174066628	NM_001356513	0	+	174056909	174064413	0	8	94,110,166,86,162,177,45,2239, 0,1876,2400,3147,3857,6375,7245,7504,
chr1	174056885	174066628	NM_023041	0	+	174056909	174064421	0	8	94,110,166,86,162,177,45,2291, 0,1876,2400,3147,3857,6375,7245,7452,
chr1	174056885	174066628	NM_001159525	0	+	174056909	174064421	0	6	94,86,162,177,45,2291, 0,3147,3857,6375,7245,7452,
chr1	171902437	172019075	NM_022563	0	−	171907985	171966125	0	17	5683,150,235,189,128,224,211,131,63,244,184,106,148,232,103,109,150,

name score strand genome browser properties

Assignment: Working with BED Files in Galaxy and Conducting GO Enrichment Analysis

- ...what is a BED file? **Browser Extensible Data**
- Stores genomic coordinate information for specific features (SNPs, genes, etc)
- Must have at least 3 columns: **Chromosome name**, **start position**, **end position**
- Can also have specific columns with more information (usually used for genome browsers like UCSC)
- We will be working with two BED files:
  - **mm9_refseq_genes.bed** (from UCSC Table Browser)
  - **forebrain_peaks_p300.bed** (modified from Supplementary Table 2)

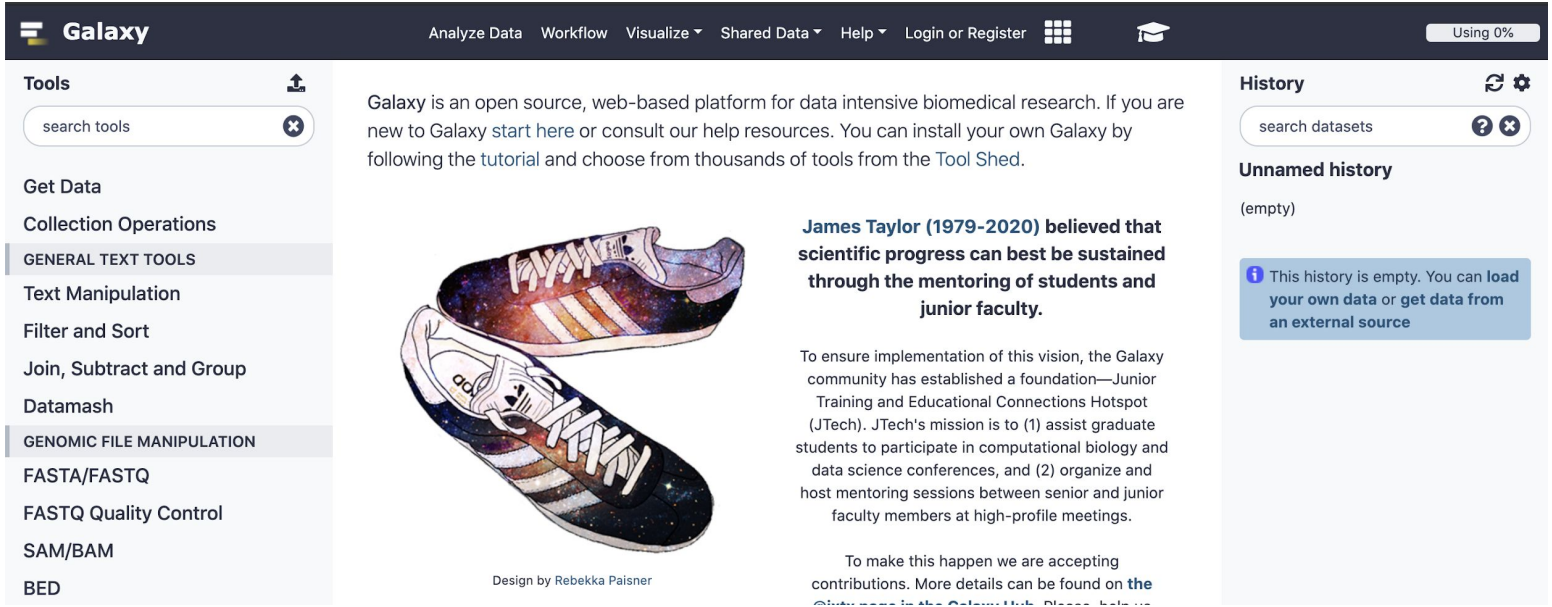
Assignment: Working with BED Files in Galaxy and Conducting GO Enrichment Analysis

- ...what are we doing exactly?
- The authors of the paper found that **forebrain p300 peaks were particularly enriched 10kb up- or downstream of genes expressed** in E11.5 forebrain tissue
- We will follow a workflow to identify:
  - what **genes are within this 10kb region of the p300 peaks**
  - and perform a **GO enrichment analysis** to see **what molecular functions these genes have**

- In other words, **predict what genes are under the influence of p300 enhancer activity and what they do**

Assignment: Working with BED Files in Galaxy and Conducting GO Enrichment Analysis

- ...what is Galaxy? **Free public server for bioinformatics analyses**

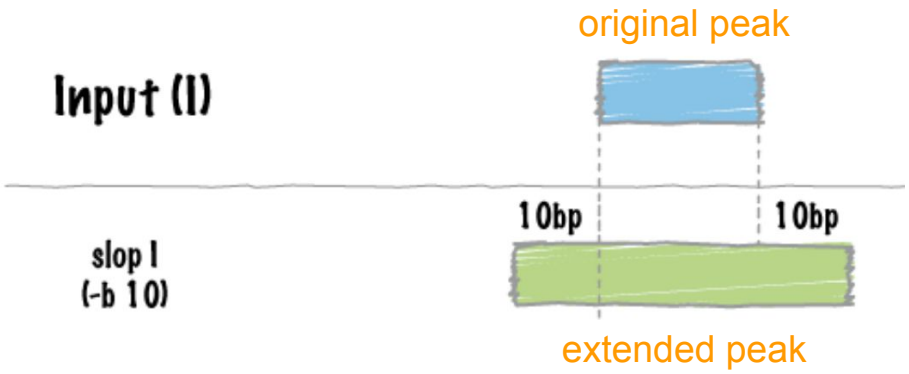


tools

data

Assignment: Working with BED Files in Galaxy and Conducting GO Enrichment Analysis

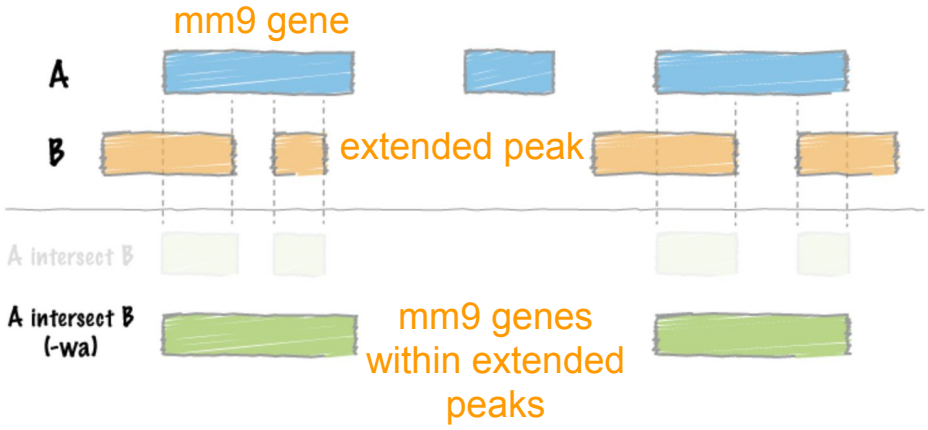
- ...what is bedtools? **Software for viewing and manipulating BED files**
- Normally command-line based, but Galaxy gives you that pointy-clicky experience
- We will be using two commands:
  - **bedtools slop**
    - Extends feature coordinates
    - We will extend peak coordinates by 10kbp in either direction
    - **This roughly corresponds to the most likely "reach" of p300 as an enhancer-binding coactivator**



https://bedtools.readthedocs.io

Assignment: Working with BED Files in Galaxy and Conducting GO Enrichment Analysis

- ...what is bedtools? **Software for viewing and manipulating BED files**
- Normally command-line based, but Galaxy gives you that pointy-clicky experience
- We will be using two commands:
  - **bedtools slop** (extends coordinates)
  - **bedtools intersect**
    - Finds overlaps between two BED files
    - We will compare the extended peak coordinates and the mm9 gene coordinates
    - We will keep the mm9 genes that overlap with the extended peaks
    - **This roughly corresponds to the mm9 genes affected by p300**

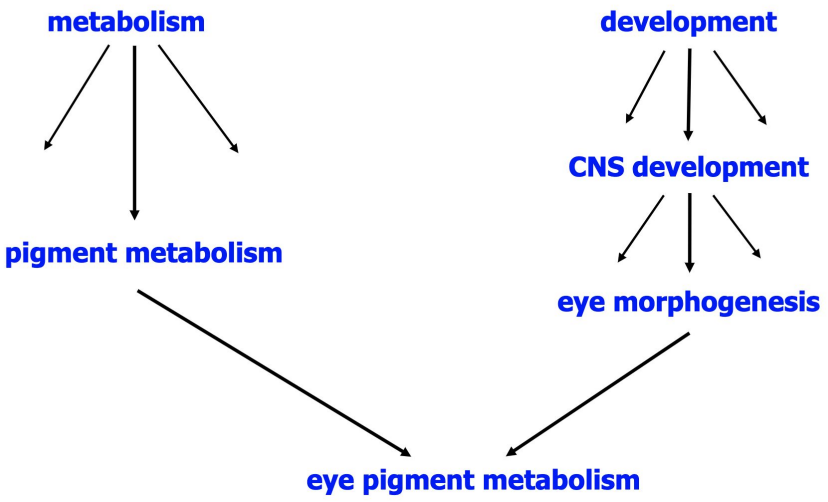


https://bedtools.readthedocs.io

Assignment: Working with BED Files in Galaxy and Conducting GO Enrichment Analysis

- ...what is GO enrichment analysis? **Analysis for enrichment of Gene Ontology (GO) terms in a list of genes**
- We will be looking at **GO biological function** enrichment for the list of mm9 genes within the reach of p300 that we **predict will be enhanced in developing forebrain tissue**
- This gives us an idea of what **functions** the p300-activated genes in developing forebrain have

GO-Biological Process hierarchy



Assignment: Working with BED Files in Galaxy and Conducting GO Enrichment Analysis

Caveats of our approach:

- Not filtering for actually overexpressed genes (might have false positives)
- Excluding genes beyond 10kb of the peaks (false negatives)
- Our BED coordinates are actually 1bp off (but this most likely will not affect our results)
    - BED coordinates start counting at 0, whereas the supplementary table likely uses coordinates that start counting at 1
    - Off-by-1 errors are a very annoying part of bioinformatics!