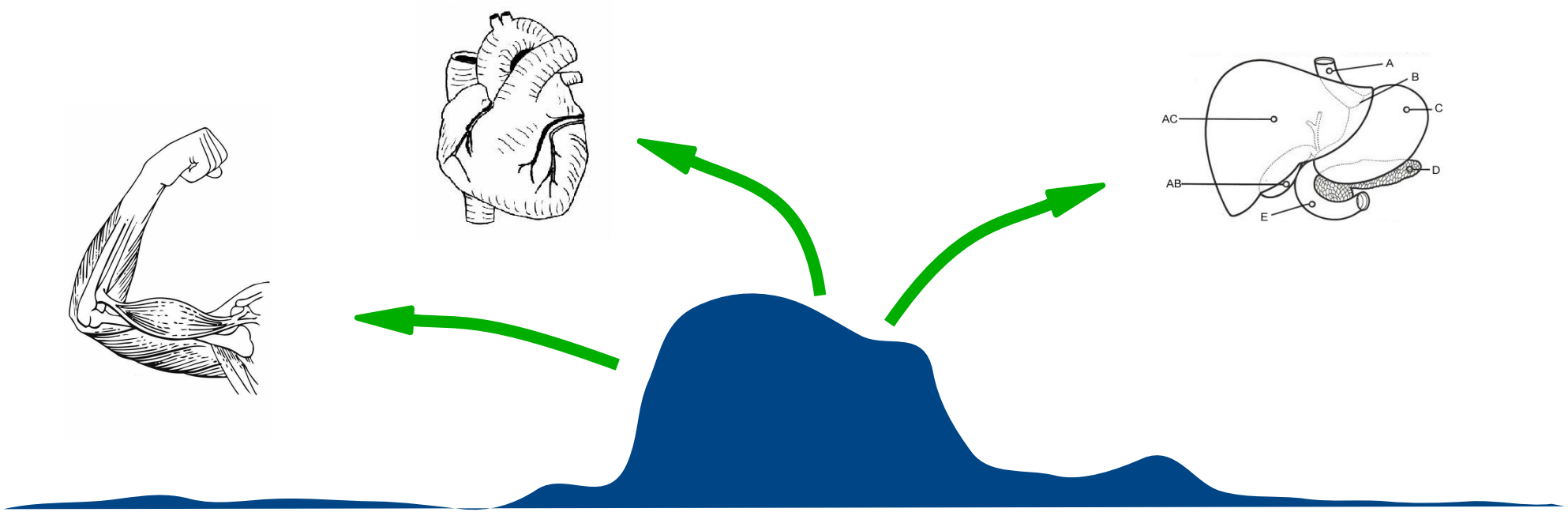
Functional annotation of ChIP-peaks



how do we make biological sense out of the peaks ?

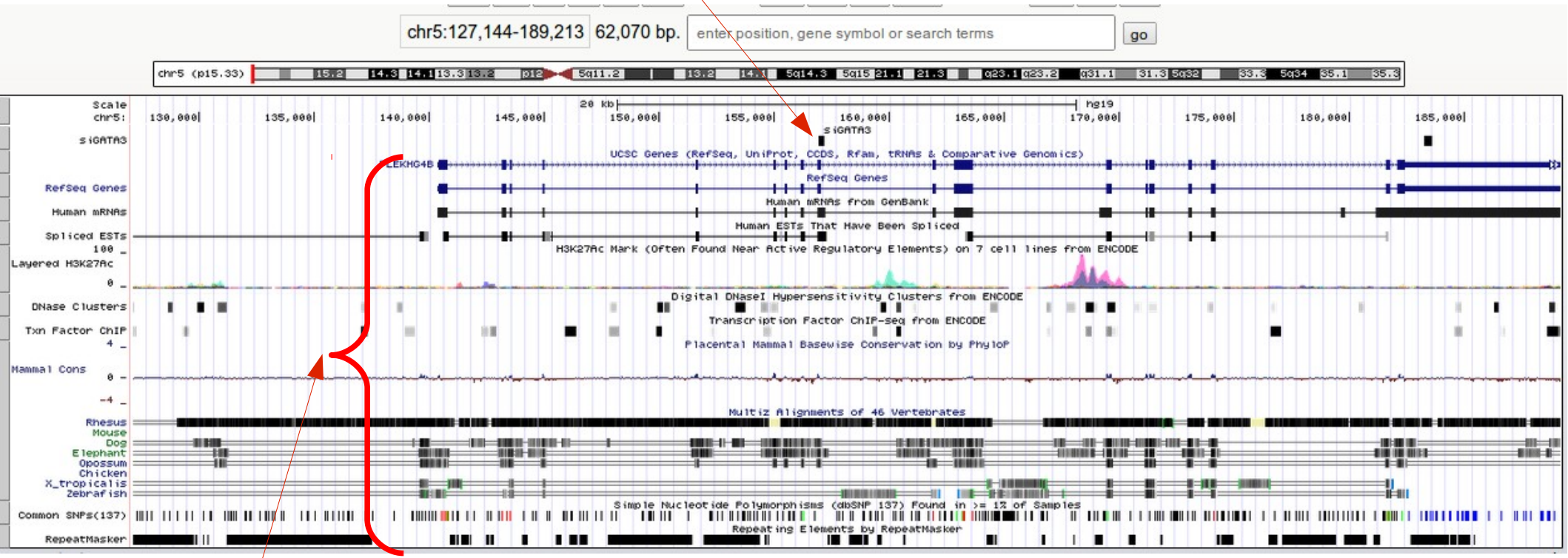
Our data in the context

- We can display our datasets (peaks locations) in the **UCSC genome browser**, to take advantage of published datasets (e.g. ENCODE)
- These datasets can be displayed as **tracks** to see coincidences with our datasets
- Track control options allow to display (in many modes) or hide specific datasets
- Personal datasets can be uploaded as **Custom Tracks**
  these datasets remain confidential, and do not appear publicly!

http://genome.ucsc.edu

Our data in the context

Custom track uploded by the user (here ESR1 peaks in siGATA3 context)

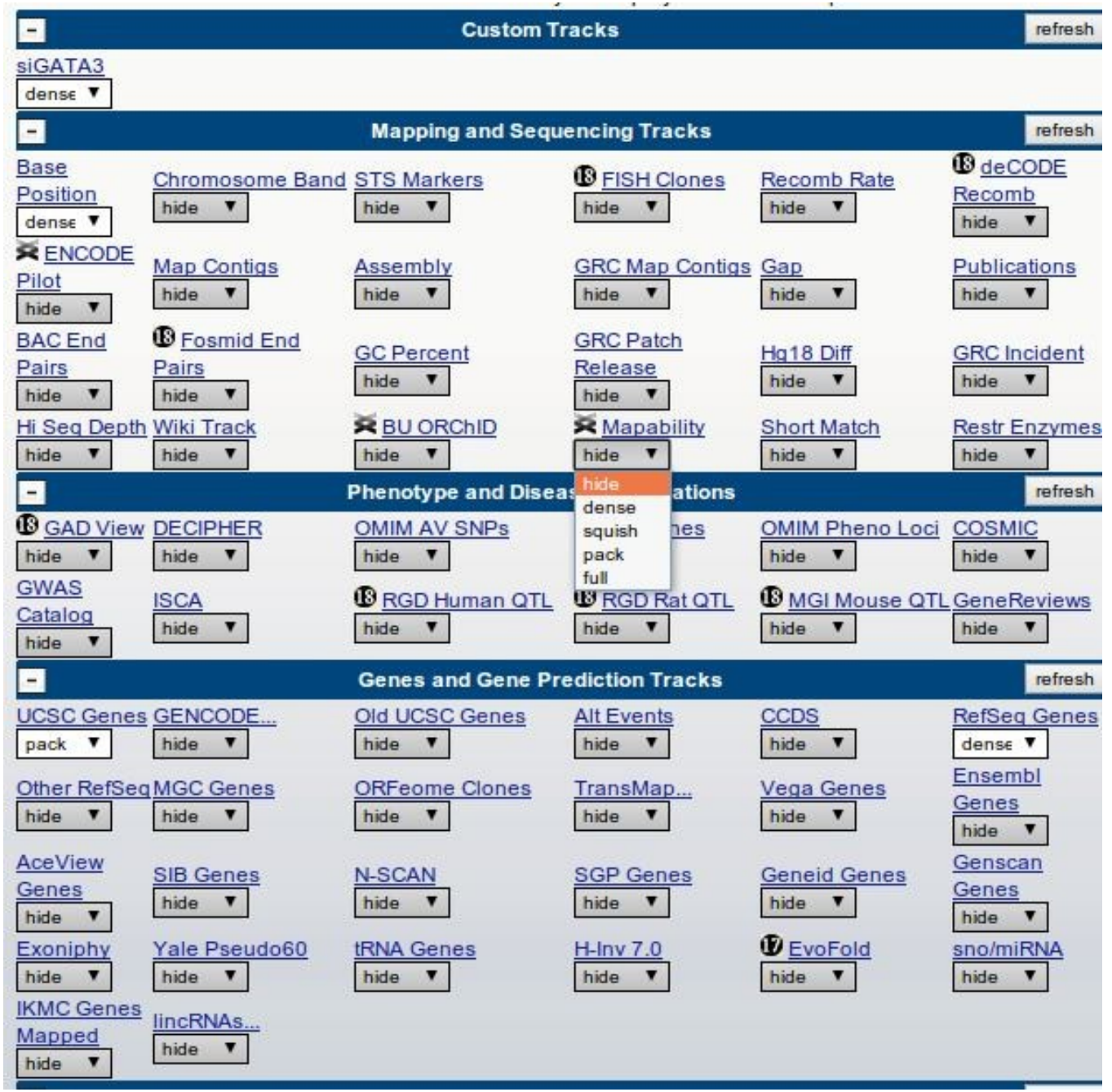


public UCSC annotation/data tracks

Our data in the context

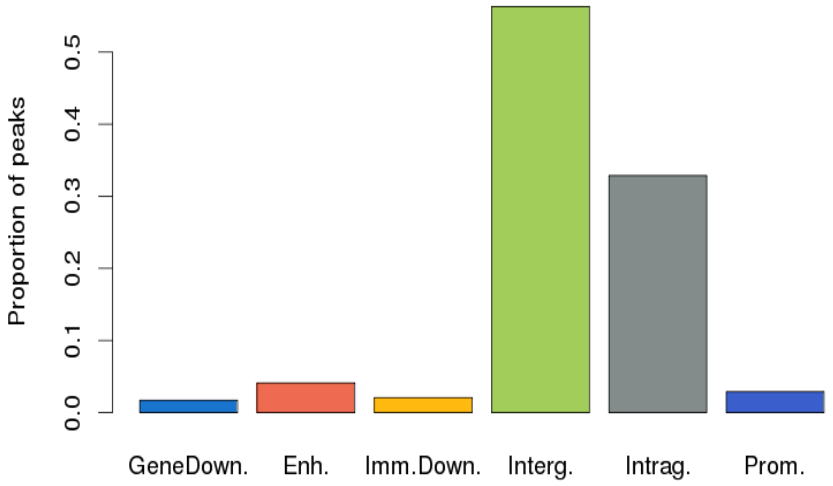
Track control panel;

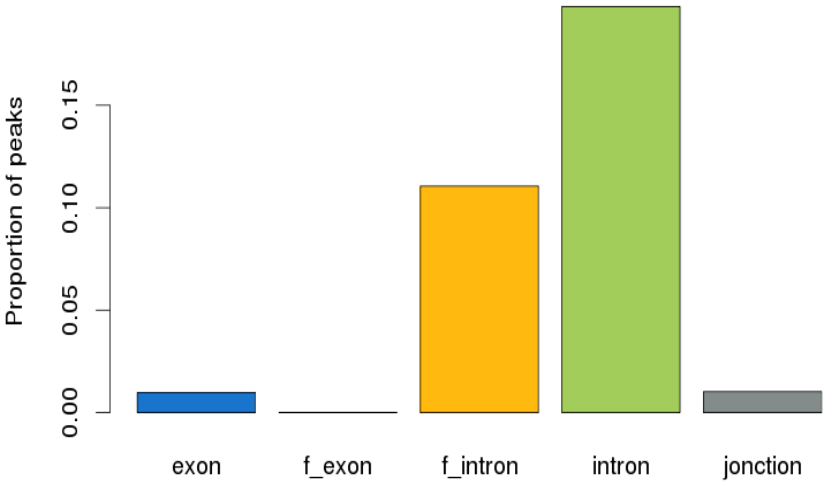
- hide
- display (dense)
- display (squish)
- display (pack)
- display (full)



Positional biases

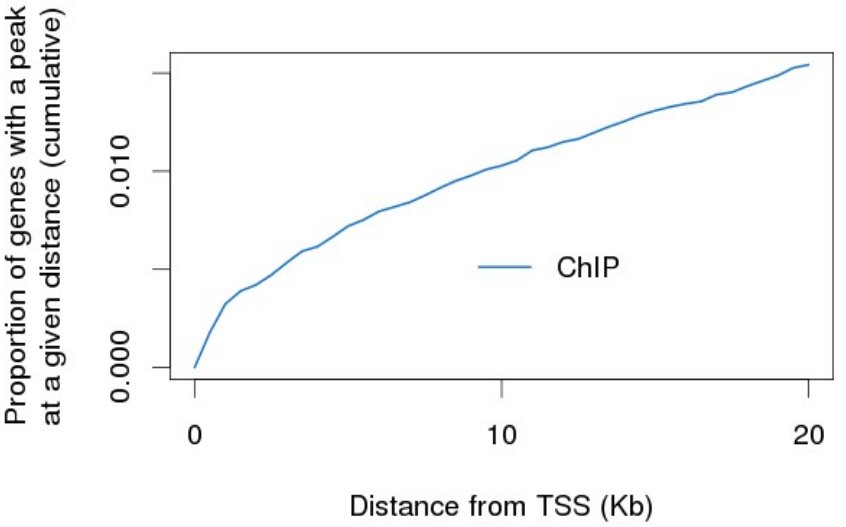
- where do peaks localize ? Proximal promoter ? Intergenic regions ? Intronic regions ?

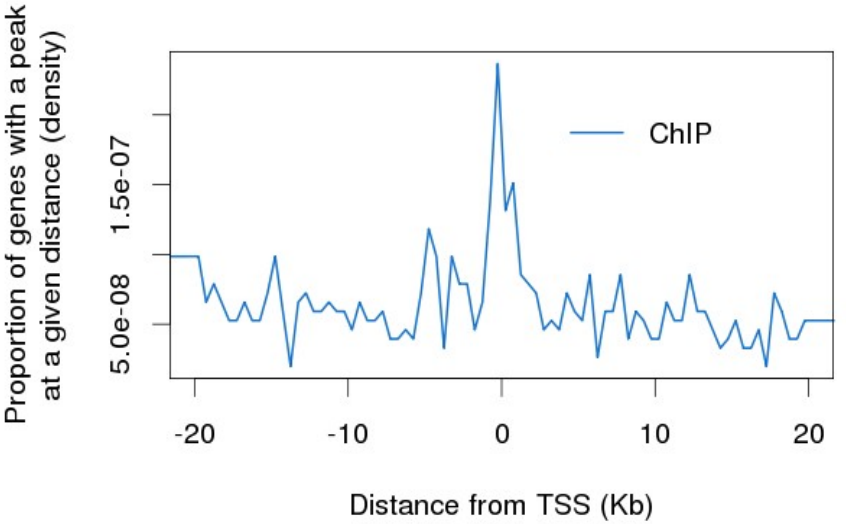




Positional biases

- Distance to TSS :





Peaks → **Genes → Functions**

- collect sets of genes
- compute over-represented functional annotations
  - Gene Ontology
  - Phenotypic annotations
  - Biological Pathways

- Typical tools
  - **DAVID** [Huang et al., NAR 2009]
  - **Babelomics** [Medina et al., NAR 2010]

Peaks → Genes → Functions

- **Drawbacks**
  - restricting to proximal regions **discards** a large number of binding events
  - "nearest gene" approach introduces **bias** towards genes with large intergenic regions
    e.g. : "multicellular organism development" : 14% of the genes, but 33% of the genome associated

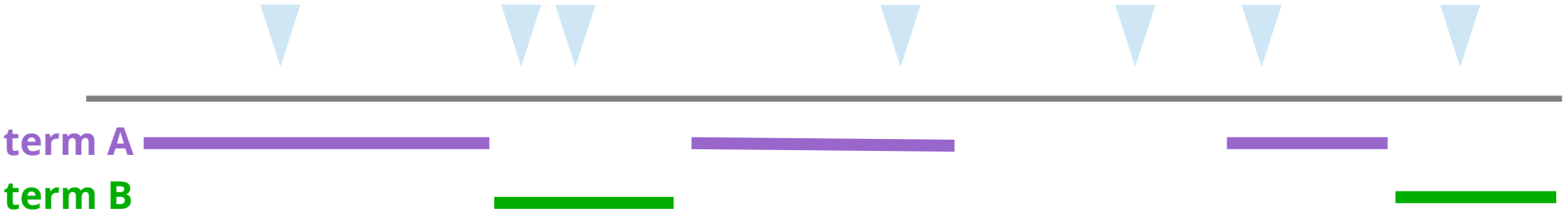
Genes → Regions ← Peaks

- **Idea** :
  - assign functional annotation to genomic regions
  - use statistics to avoid biases
- assign to each gene a regulatory domain
  - basal (-5kb/+1kb from TSS)
  - extended (up to nearest basal region ; max 1Mb)

- each domain is annotated to the functional terms of the corresponding gene
  → "Functional domains"

"GREAT improves functional interpretation of cis-regulatory regions"
McLean et al. Nat. Biotech. (2010)

Genes → **Regions** ← **Peaks**



*Given that **60%** of the genome is annotated to A, would I randomly expect 3 or more peaks to fall into region A ?*

p > 0.5

*Given that **15%** of the genome is annotated to B, would I randomly expect 3 or more peaks to fall into region B ?*

p = 0.07

"GREAT improves functional interpretation of cis-regulatory regions"
McLean et al. Nat. Biotech. (2010)

Job description

Job ID: 20111012-public-uXBYVG

Display name: GSM348066_limb_p300_peaks.NEW

Test set: GSM348066_limb_p300_peaks.NEW.bed (3,839 genomic regions)
Show in UCSC genome browser. *What is this?*

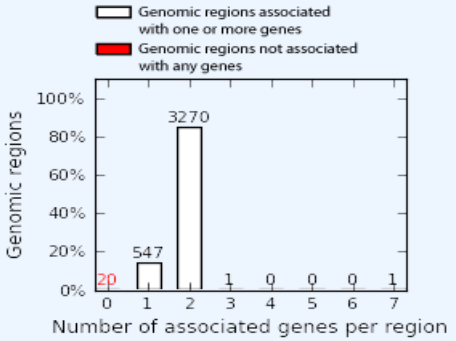
Background: Whole genome background

Assembly: Mouse: NCBI build 37 (UCSC mm9, Jul 2007) *What gene set does GREAT use?*

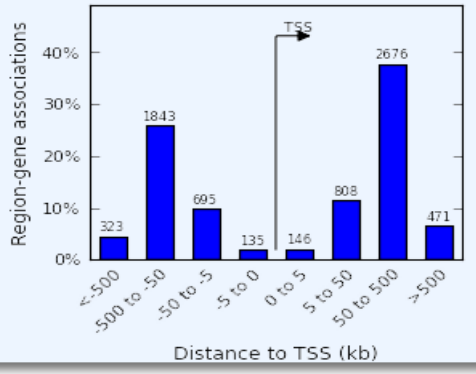
Associated genomic regions: Basal+extension (constitutive 5.0 kb upstream and 1.0 kb downstream, up to 1000.0 kb max extension). Curated regulatory domains are included.
20 of all 3,839 genomic regions (0.5%) are not associated with any genes.
View genomic region-gene associations. *What is this?*
Revise the region-gene association rule. *What is a region-gene association rule?*

Region-gene association graphs:

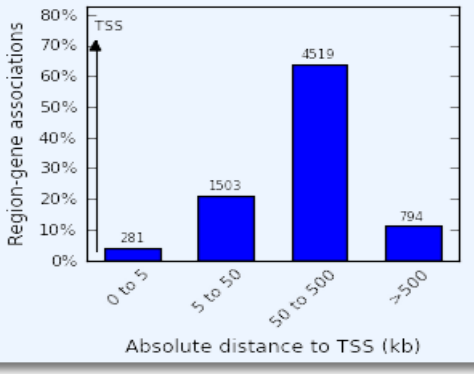
Number of associated genes per region
Download as PDF.



Binned by orientation and distance to TSS
Download as PDF.



Binned by absolute distance to TSS
Download as PDF.



Mouse Phenotype

Global Controls

Table controls: Export ▾ Shown top rows in this table: 20 Set Term annotation count: Min: 1 Max: Inf Set

Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
abnormal limbs/digits/tail morphology	2	**2.0559e-91**	**6.6837e-88**	2.1465	780	20.32%	6	**2.5295e-40**	2.2020	278	681	8.31%
abnormal craniofacial morphology	3	**9.3822e-91**	**2.0334e-87**	2.0082	887	23.10%	10	**8.9231e-36**	2.0382	297	786	8.88%
abnormal limb morphology	5	**2.4990e-80**	**3.2497e-77**	2.3077	604	15.73%	9	**7.4787e-37**	2.4541	202	444	6.04%
abnormal appendicular skeleton morphology	10	**3.0255e-70**	**1.9672e-67**	2.3450	517	13.47%	17	**3.9549e-30**	2.4098	172	385	5.14%
abnormal skeleton extremities morphology	12	**3.2687e-69**	**1.7711e-66**	2.3724	499	13.00%	21	**7.0557e-29**	2.4222	163	363	4.87%
abnormal paw/hand/foot morphology	13	**4.0300e-69**	**2.0156e-66**	2.6813	404	10.52%	23	**5.4918e-28**	2.7186	126	250	3.77%
abnormal head morphology	14	**6.4657e-67**	**3.0029e-64**	2.0134	672	17.50%	25	**2.9042e-27**	2.0562	223	585	6.67%
abnormal digit morphology	18	**1.0543e-61**	**3.8084e-59**	2.6982	358	9.33%	36	**1.2033e-25**	2.7998	109	210	3.26%
abnormal cartilage morphology	23	**7.3728e-58**	**2.0843e-55**	2.3432	430	11.20%	29	**1.1337e-26**	2.5089	140	301	4.19%
abnormal skeleton development	24	**3.5769e-56**	**9.6904e-54**	2.0833	530	13.81%	38	**5.2377e-25**	2.1414	185	466	5.53%
abnormal long bone morphology	25	**4.6593e-56**	**1.2118e-53**	2.3374	419	10.91%	43	**4.9983e-24**	2.3823	140	317	4.19%

GREAT vs. proximal peaks

	GREAT				Proximal 2kb peaks			
	Best GO term	*P-val*	*MGI expression*	*P-val*	*Best GO-term*	*P-val*	*MGI expression*	*P-val*
p300 limb	Embryonic limb morphogenesis	1E-27	TS19 limb	7E-49	Skeletal system development	4E-06	TS19 limb	3E-05
p300 forebrain	CNS development	8E-36	TS17 forebrain	6E-41	Forebrain development	2E-04	TS22 forebrain	3E-07
p300 midbrain	CNS development	1E-12	TS 15 CNS	1E-14	none		none	

- **more specific terms with higher significance**
- **more peaks/genes taken into account**