



# After the association: Functional and Biological Validation of Variants

Jason L. Stein

Geschwind Laboratory /  
Imaging Genetics Center

University of California, Los Angeles  
(but soon to be at UNC-Chapel Hill)

[jasonlouisstein@gmail.com](mailto:jasonlouisstein@gmail.com)

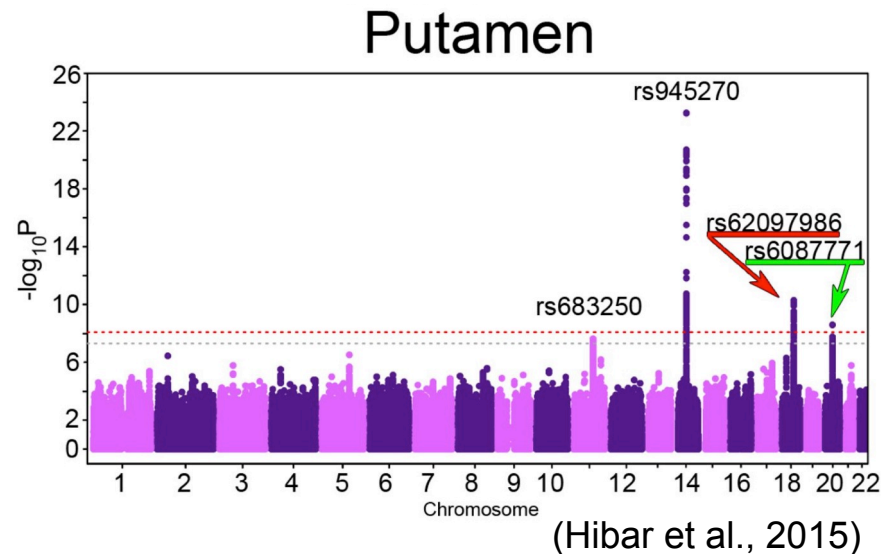
Organization for Human Brain Mapping

Introduction to Imaging Genetics

Honolulu, HI

June 14, 2015

# A hit is just the beginning...



## What you have found

- Variation in a **locus** of the genome which significantly influences your **trait** (brain structure/disease)

## What you want to know

- A mechanism by which genetic variation influences brain structure or function and risk for disease
- Causal variant(s)
- Causal gene(s)
- Causal biological pathway(s)
- Causal brain region(s)

# But be wary....

## APPLICATIONS OF NEXT-GENERATION SEQUENCING

### Sequencing studies in human genetics: design and interpretation

David B. Goldstein<sup>1</sup>, Andrew Allen<sup>1,2</sup>, Jonathan Keebler<sup>1</sup>, Elliott H. Margulies<sup>5</sup>,  
Steven Petrou<sup>4,5</sup>, Slavé Petrovski<sup>1,6</sup> and Shamil Sunyaev<sup>7</sup>

“Human genomes have a high level of **‘narrative potential’** to provide compelling but statistically poorly justified connections between mutations and phenotypes.”

“A critical challenge for biologists [...] will be avoiding premature hypotheses born of biological plausibility and **‘Just So’** stories.”

### Genome-scale neurogenetics: methodology and meaning

Steven A McCarroll<sup>1,2</sup>, Guoping Feng<sup>1,3,4</sup> & Steven E Hyman<sup>1,5</sup>

## ORIGINAL ARTICLES

### Spurious Genetic Associations

Patrick F. Sullivan

“Findings from single association studies constitute **‘tentative knowledge’** and must be interpreted with exceptional caution.

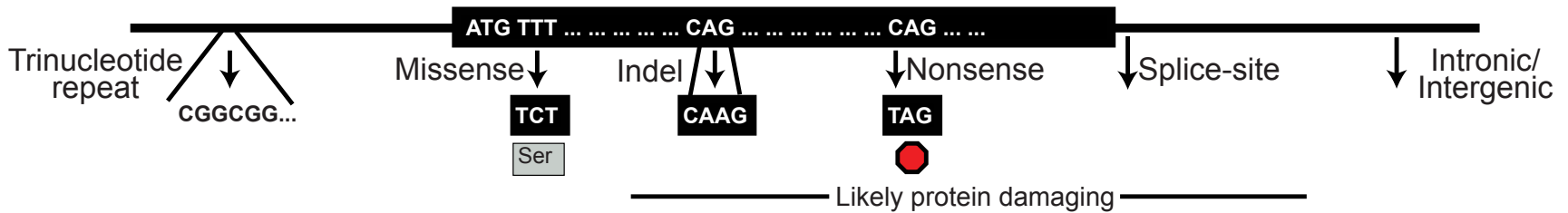
**Biological plausibility is not a substitute for statistical significance**

# Exploring biological mechanisms

---

- Exploring the genetic locus
- Epigenetics
- Move from locus to gene
- Exploring the expression of the gene
- Enrichment in biological pathways

# Genetic hit locations



- 88% of significant GWAS (common variant) loci are often found in intergenic or intronic regions with no clear gene(s) of action (Hindorff et al., *PNAS*, 2009)
- GWAS loci tag very large regions with multiple functional elements including genes
  - For the SCZ GWAS loci: max 800kb (SZ working group, *Nature*, 2014)
  - For the SCZ GWAS loci: mean 171 kb ( $r^2 > 0.6$ )
  - Mean gene size: 29kb (Gencode v19)
- Rare variant mutations including missense or nonsense mutations have a clear gene of action, but are rare.

# UCSC Genome Browser

Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the [Genome Bioinformatics Group of UC Santa Cruz](#).  
Software Copyright (c) The Regents of the University of California. All rights reserved.

group genome assembly position search term

Mammal Human Feb. 2009 (GRCh37/hg19) chr19:45,404,181-45,404,681 enter position, gene symbol or search terms submit

[Click here to reset](#) the browser user interface settings to their defaults.

[track search](#) [add custom tracks](#) [track hubs](#) [configure tracks and display](#)

## Human Genome Browser – hg19 assembly ([sequences](#))

The February 2009 human reference sequence (GRCh37) was produced by the [Genome Reference Consortium](#). For more information about this assembly, see [GRCh37](#) in the NCBI Assembly database.

### Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the [User's Guide](#) for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gl000212	Displays all of the unplaced contig gl000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000



*Homo sapiens*  
(Graphic courtesy of [CBSE](#))

<http://genome.ucsc.edu/cgi-bin/hgGateway>

# Exploring a hit (rs945270)

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr14:56,200,223-56,200,723 501 bp.

chr14 (q22.3) 13 12 p11.2 11.2 14q12 21.1 24.3

Scale: 200 bases hg19

chr14: 56,200,300 | 56,200,400 | 56,200,500 | 56,200,600 | 56,200,700

UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)

RefSeq Genes

Publications: Sequences in Scientific Articles

Human mRNAs from GenBank

Human ESTs That Have Been Spliced

Layered H3K27Ac

H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE

DNase Clusters

Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE

Txn Factor ChIP

Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs

100 Vert. Cons

100 vertebrates Basewise Conservation by PhyloP

Multiz Alignments of 100 Vertebrates

Rhesus

Mouse

Dog

Elephant

Chicken

X\_tropicalis

Zebrafish

Lamprey

Simple Nucleotide Polymorphisms (dbSNP 138) Found in >= 1% of Samples

rs870924

Repeating Elements by RepeatMasker

track search default tracks default order hide all add custom tracks track hubs configure reverse resize refresh

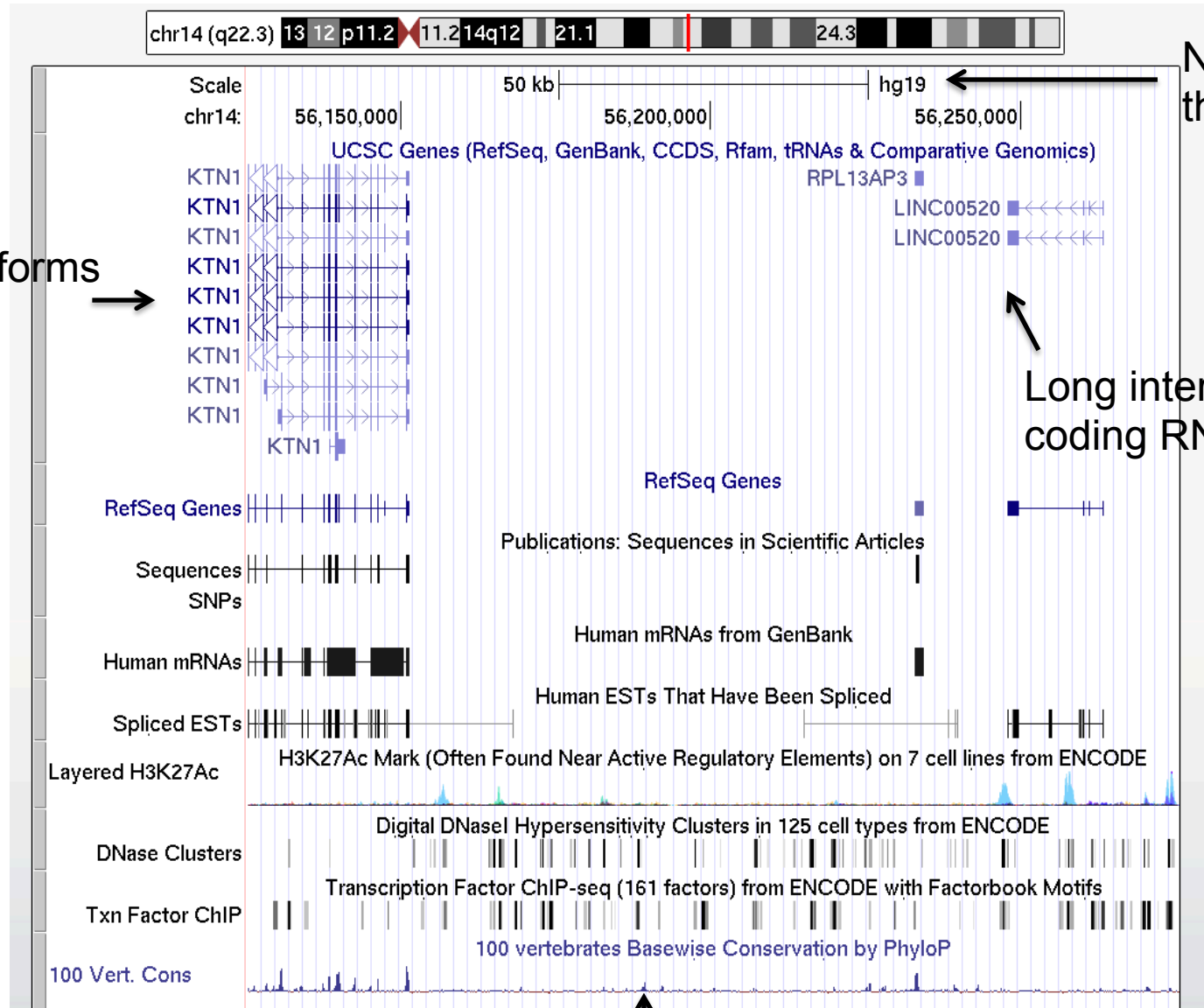
Zoom Out

Click to Shrink Track

Ideogram  
Scale  
Coordinates

Tracks

# Exploring a hit (rs945270)



Multiple isoforms of genes



Notice the scale

Long intergenic non-coding RNAs



The variant initially entered stays in the center





# Exploring a hit (rs945270)

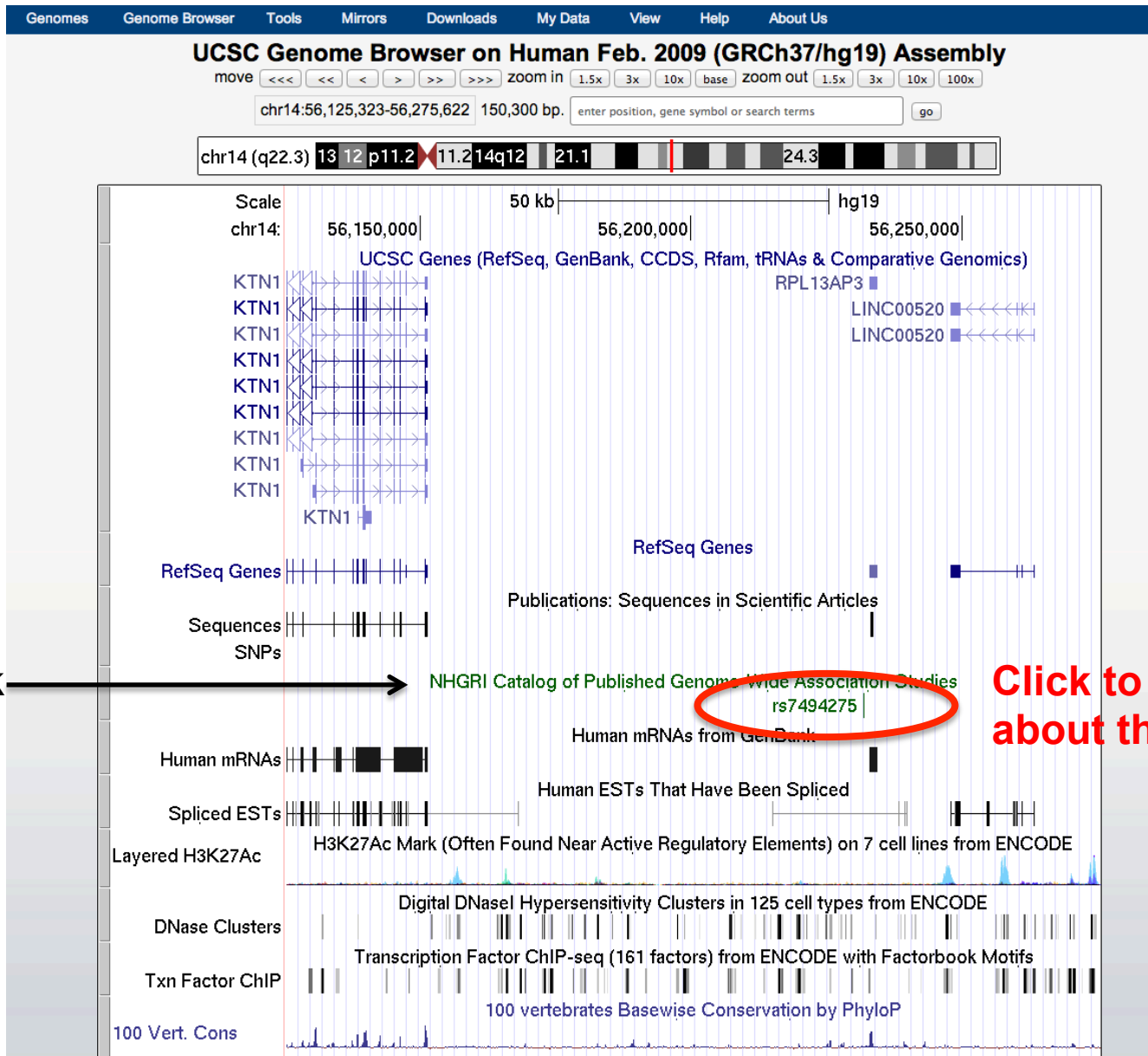
Add a Track

The screenshot displays a genomic data interface with a dark blue header bar labeled "Phenotype and Literature". On the right side of this bar, a "refresh" button is circled in red. Below the header, there are several tracks, each with a title and a "hide" button:

- Publications**: dense (dropdown)
- ClinVar Variants**: hide (dropdown)
- Coriell CNVs**: hide (dropdown)
- COSMIC**: hide (dropdown)
- DECIPHER**: hide (dropdown)
- GAD View**: hide (dropdown)
- GeneReviews**: hide (dropdown)
- GWAS Catalog**: hide (dropdown)
- HGMD Variants**: hide (dropdown)
- ISCA**: hide (dropdown)
- LOVD Variants**: hide (dropdown)
- 18 MGI Mouse QTL**: hide (dropdown)
- OMIM AV SNPs**: hide (dropdown)
- OMIM Pheno Loci**: hide (dropdown)
- 18 RGD Human QTL**: hide (dropdown)
- 18 RGD Rat QTL**: hide (dropdown)
- UniProt Variants**: hide (dropdown)
- Web Sequences**: hide (dropdown)

A dropdown menu is open for the "OMIM AV SNPs" track, showing the following options: hide (checked), dense, squish, pack (highlighted), and full.

# Exploring a hit (rs945270)



# Exploring a hit (rs945270)

[Home](#) [Genomes](#) [Genome Browser](#) [Tools](#) [Mirrors](#) [Downloads](#) [My Data](#) [Help](#) [About](#)

## NHGRI Catalog of Published Genome-Wide Association Studies (rs7494275)

**dbSNP:** [rs7494275](#)  
**Position:** [chr14:56231800-56231800](#)  
**Band:** 14q22.3  
**Genomic Size:** 1  
[View DNA for this feature](#) (hg19/Human)  
**Reported region:** 14q22.3  
**Publication:** Low SK *et al.* [Genome-wide association study of chemotherapeutic agent-induced severe neutropenia/leucopenia for patients in Biobank Japan](#). *Cancer Sci.* 2013-05-04  
**Disease or trait:** Adverse response to chemotherapy (neutropenia/leucopenia) (all topoisomerase inhibitors)  
**Initial sample size:** 106 Japanese ancestry cases, 187 Japanese ancestry controls  
**Replication sample size:** NA  
**Reported gene(s):** RPL13AP3  
**Strongest SNP-Risk allele:** [rs7494275-C](#)  
**dbSNP build 137 observed alleles for rs7494275:** A/C  
**Risk Allele Frequency:** 0.406  
**p-Value:** 9E-6 (Recessive model)  
**Odds Ratio or beta:** 1.73  
**95% confidence interval:** [1.232-2.433]  
**Platform:** Illumina [733,202]  
**Copy Number Variant (CNV)?:** No

[View table schema](#)

[Go to GWAS Catalog track controls](#)

**Data last updated:** 2014-05-23

Not super convincing given low sample size and non-genome wide significant P-value.

# Uploading a user track

```
test.bed  
chr14 56200473 56200473 SIGNIFICANT_HIT
```

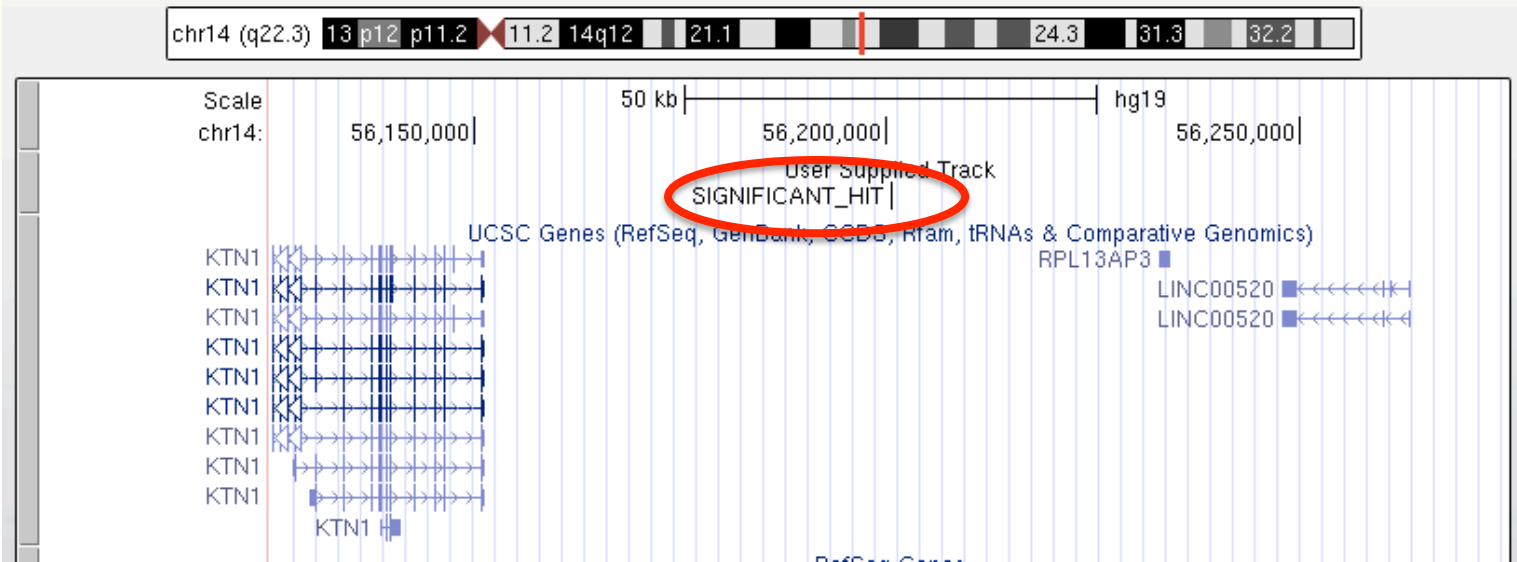
move start < 2.0 > move end < 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

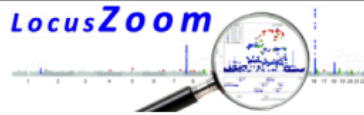
track search default tracks default order hide all **add custom tracks** track hubs configure reverse resize refresh

collapse all Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes. expand all

Paste URLs or data: Or upload: **Choose File** No file chosen Submit



# LocusZoom: Making Prettier Pictures



## LocusZoom - Plot with Your Data

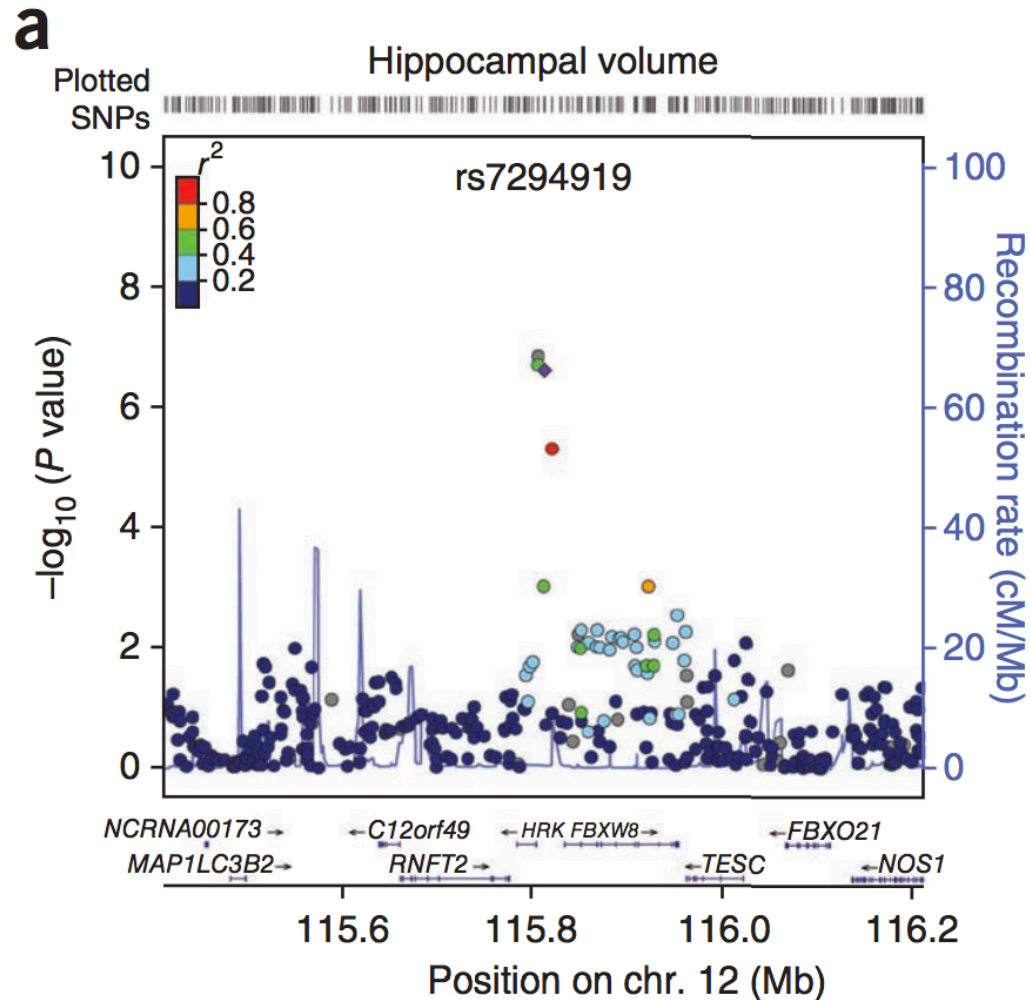
Make sure not too big a file

Plot Your Data Depending on the size of your data, runs can require 30-60 seconds to generate a plot

Provide Details for Your Data	Path to Your File	<input type="button" value="Choose File"/> No file chosen File will sent to server and used for plotting (Maximum 200MB) [Help]		
	P-Value Column Name	<input type="text"/>	Set for <a href="#">PLINK data</a> or <a href="#">WikiGWA data</a> Default is P.value	
	Marker Column Name	<input type="text"/>	Default is MarkerName	
	Column Delimiter	<input type="text" value="Tab"/>	Default is tab	
Specify Region to Display  Required: Fill in Only ONE of These Three	SNP	<input type="text"/>	+/-	<input type="text" value="400"/> Kb Flanking Size
	Gene	<input type="text"/>	+/-	<input type="text" value="200"/> Kb Flanking Size Optional Index SNP Default=lowest p-value
	Region	Chr: <input type="text" value="None"/>	<input type="text"/> Mb	through <input type="text"/> Mb Optional Index SNP Default=lowest p-value
Custom Annotation  Optional: This overrides Show Annotation below	Column Name	<input type="text"/> Name of annotation column		
	Category Order	<input type="text"/> Order of annotation categories		

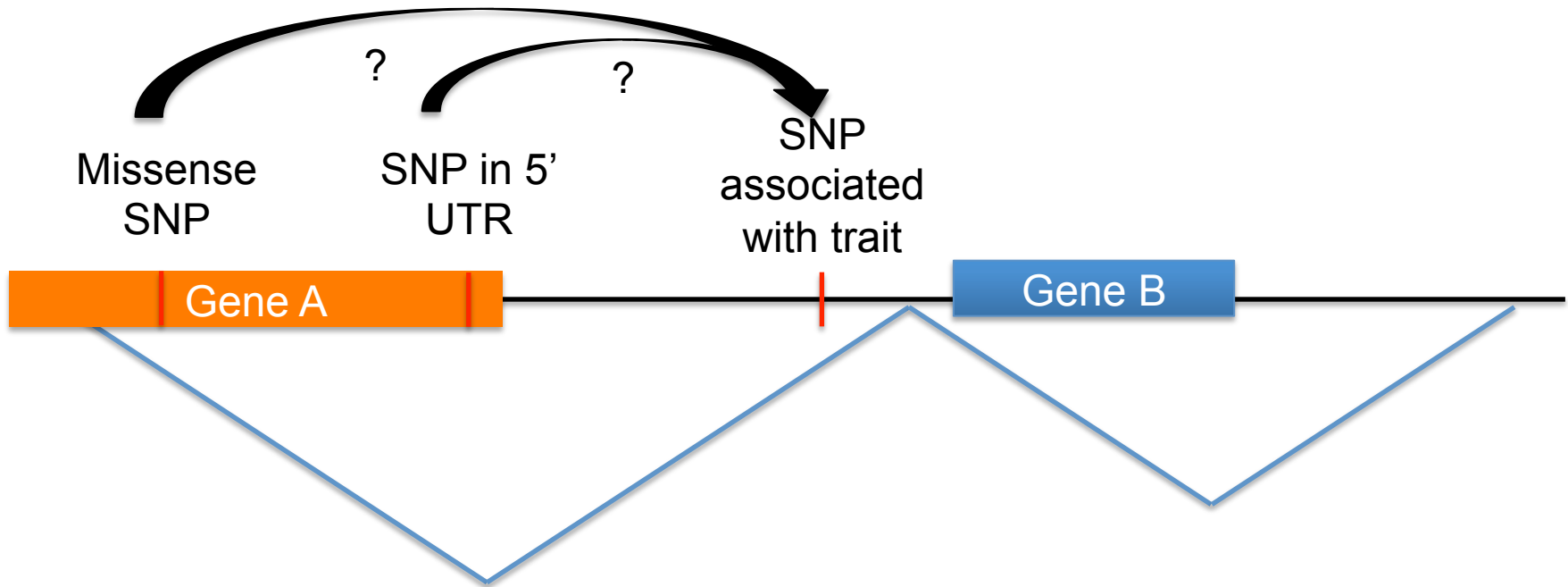
<https://statgen.sph.umich.edu/locuszoom/genform.php?type=yourdata>

# LocusZoom: Making Prettier Pictures



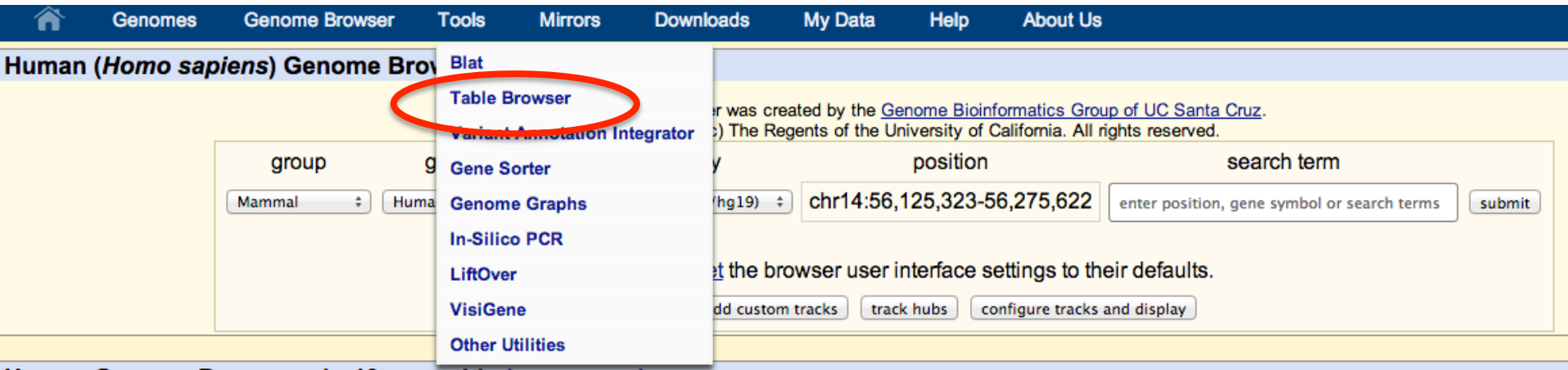
# Trying to find possible gene function

---



We found a genetic variant, but is it in LD with anything of known functionality?

# Finding Variants of Known Functionality



The screenshot shows the top navigation bar of the UCSC Genome Browser with the following items: Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Help, and About Us. A dropdown menu is open under 'Tools', listing: Blat, **Table Browser** (circled in red), Variant Annotation Integrator, Gene Sorter, Genome Graphs, In-Silico PCR, LiftOver, VisiGene, and Other Utilities. Below the menu, a search box is visible with 'hg19' selected in the genome dropdown, 'chr14:56,125,323-56,275,622' in the position field, and a 'submit' button.

Go to the Table Browser in UCSC Genome Browser



The screenshot shows the UCSC Table Browser interface. The 'filter: create' button is circled in red. To its right, the text 'Click to create a filter' is written in red. The interface includes the following fields and buttons:

- clade: Mammal
- genome: Human
- assembly: Feb. 2009 (GRCh37/hg19)
- group: Variation
- track: Common SNPs(138)
- add custom tracks
- track hubs
- table: snp138Common
- describe table schema
- region:  genome  ENCODE Pilot regions  position
- chr14:56031386-56369561
- lookup
- define regions
- identifiers (names/accessions): paste list upload list
- filter: create
- intersection: create
- correlation: create
- output format: all fields from selected table
- Send output to  Galaxy  GREAT
- output file: chr14.txt (leave blank to keep output in browser)
- file type returned:  plain text  gzip compressed
- get output
- summary/statistics



# Finding Variants of Known Functionality

## Select interpretable functional Variants

func does include  \*  unknown  coding-synon  intron  near-gene-3  
 near-gene-5  ncRNA  nonsense  missense  stop-loss  
 frameshift  cds-indel  untranslated-3  untranslated-5  splice-3  
 splice-5

Click to create a filter

clade: Mammal genome: Human assembly: Feb. 2009 (GRCh37/hg19)  
group: Variation track: Common SNPs(138)    
table: snp138Common   
region:  genome  ENCODE Pilot regions  position chr14:56,031,386-56,369,    
identifiers (names/accessions):    
filter:    
intersection:   
correlation:   
output format: all fields from selected table Send output to  Galaxy  GREAT  
output file: chr14.txt (leave blank to keep output in browser)  
file type returned:  plain text  gzip compressed

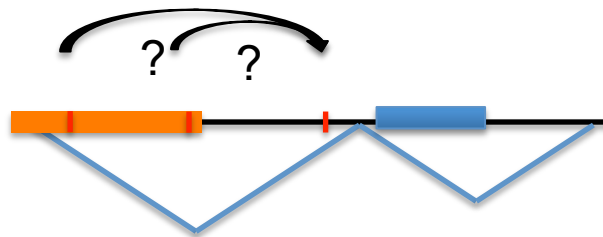
Get output spreadsheet

# Finding Variants of Known Functionality

#filter: (FIND\_IN\_SET('splice-5', snp138Common.func)>0 OR FIND\_IN\_SET('splice-3', snp138Common.func)>0 OR FIND\_IN\_SET('untranslat

#bin	chrom	chromStart	chromEnd	name	refNCBI	refUCSC	observed	func
1012	chr14	56068483	56068484	rs116289145	C	C	C/T	intron,ncRNA,untranslated-5
1012	chr14	56068520	56068521	rs10083303	T	T	C/T	intron,ncRNA,untranslated-5
1012	chr14	56078738	56078739	rs1138345	T	T	G/T	ncRNA,untranslated-5
1012	chr14	56079038	56079039	rs34879854	A	A	A/T	coding-synon,ncRNA
1012	chr14	56094725	56094726	rs17128636	C	C	A/C	coding-synon,ncRNA
1012	chr14	56096685	56096686	rs74053638	A	A	A/C	coding-synon,ncRNA
1012	chr14	56096730	56096731	rs2274075	A	A	A/G	coding-synon,ncRNA
1013	chr14	56146356	56146357	rs11546	G	G	A/G	coding-synon,ncRNA

Variants of known function



Are variants of known function in LD with our top hit?

# Finding Variants of Known Functionality

**Query SNPs**

Input SNPs:  [Example](#)

One snp per line:

**Search Options**

SNP data set:   $r^2$  threshold:

Population panel:  Distance limit:

**Output Options**

Download to:   Include each query snp as a proxy for itself  
 Suppress warning messages in output

<http://www.broadinstitute.org/mpg/snap/ldsearch.php>

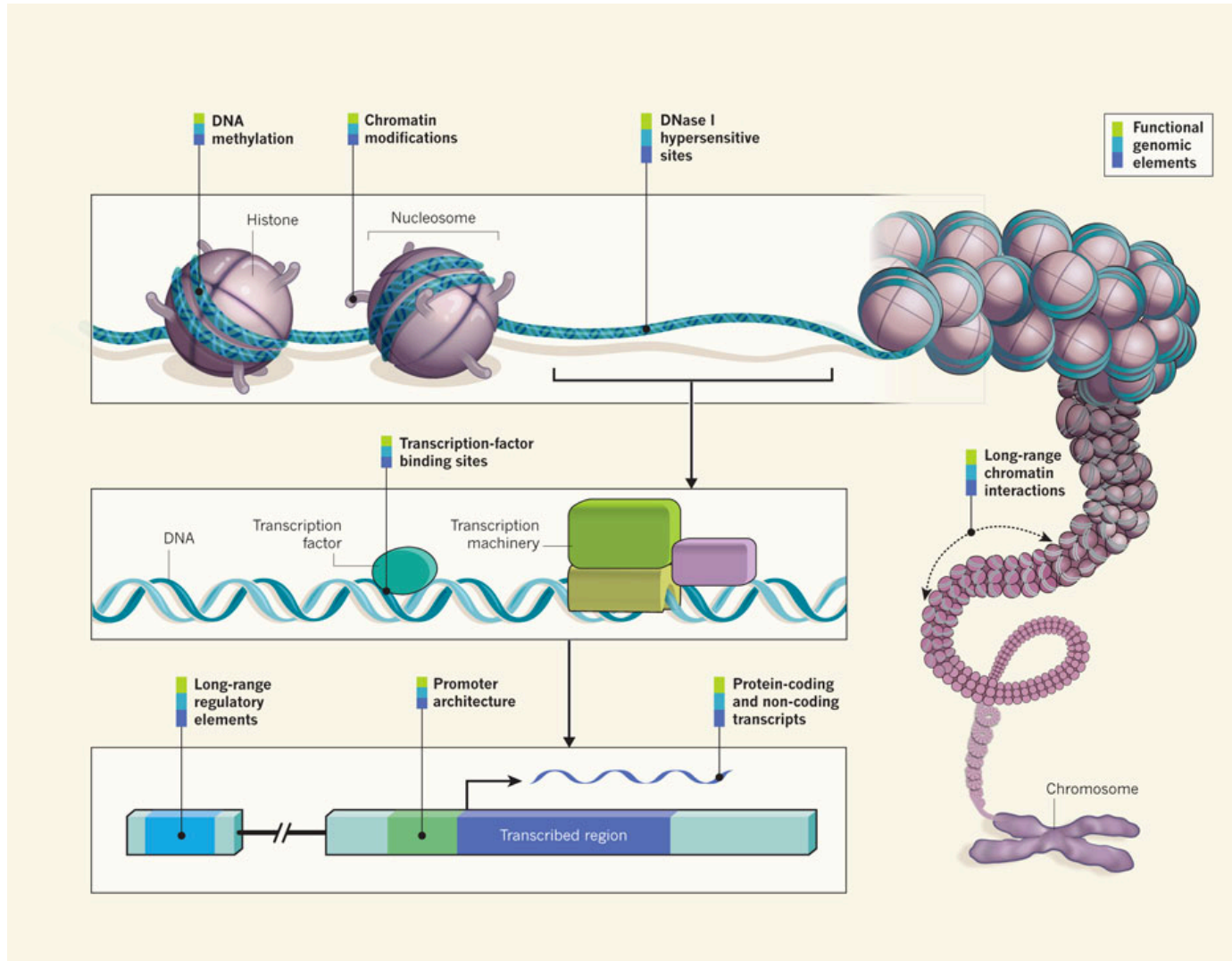
# Finding Variants of Known Functionality

SNP	Proxy	Distance	RSquared	DPrime	Arrays	Chromosome	Coordinate_HG18
rs945270	rs945270	0	1	1	None	chr14	55270226
rs945270	rs8017172	1425	1	1	I2,I5,I6,I6	chr14	55268801
rs945270	rs1959089	2636	0.875	1	None	chr14	55267590
rs945270	rs1953350	3199	0.875	1	None	chr14	55267027
rs945270	rs1953351	3248	0.875	1	None	chr14	55266978
rs945270	rs1953352	3314	0.875	1	I3,I5,I6,I6	chr14	55266912
rs945270	rs2342589	3405	0.875	1	None	chr14	55266821
rs945270	rs2342588	3434	0.875	1	None	chr14	55266792
rs945270	rs868202	4711	0.875	1	None	chr14	55265515
rs945270	rs10129414	7201	0.875	1	None	chr14	55263025
rs945270	rs8012377	9060	0.875	1	AN,A5,A6	chr14	55261166
rs945270	rs10145631	11143	0.875	1	A6,OQ	chr14	55259083
rs945270	rs8014725	13520	0.875	1	None	chr14	55256706
rs945270	rs7157327	8023	0.84	0.964	None	chr14	55262203
rs945270	rs8021018	22965	0.807	0.929	I2,I5,I6,I6	chr14	55247261

Are any of the proxy SNPs in the list of functional SNPs?... Nope.

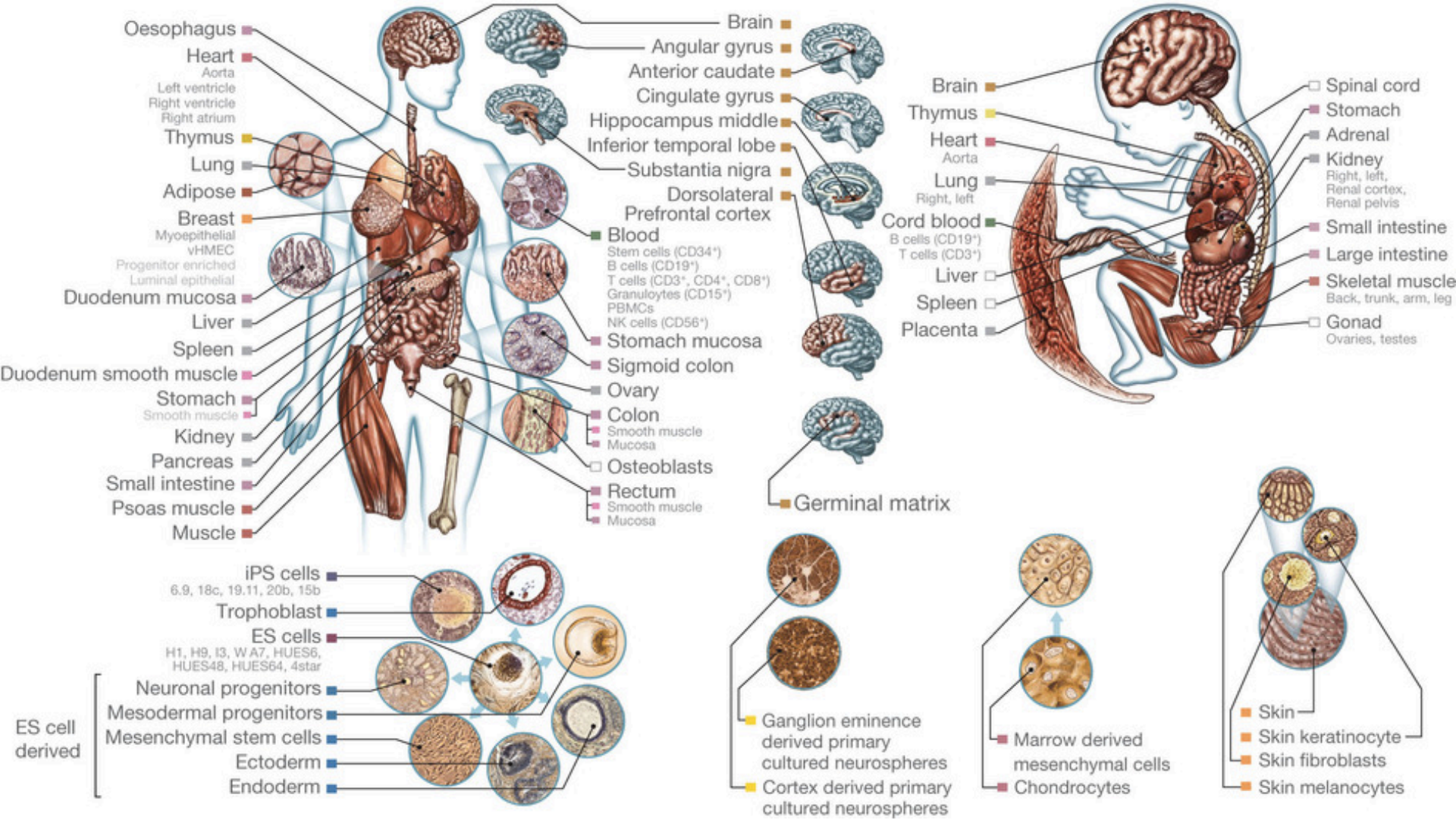
Our hit cannot be explained by known functional variants

# Epigenetics



(Ecker et al., 2012)

# Epigenetics Roadmap



(Roadmap Epigenetics Consortium et al., 2015)

# ENCODE in UCSC Genome Browser

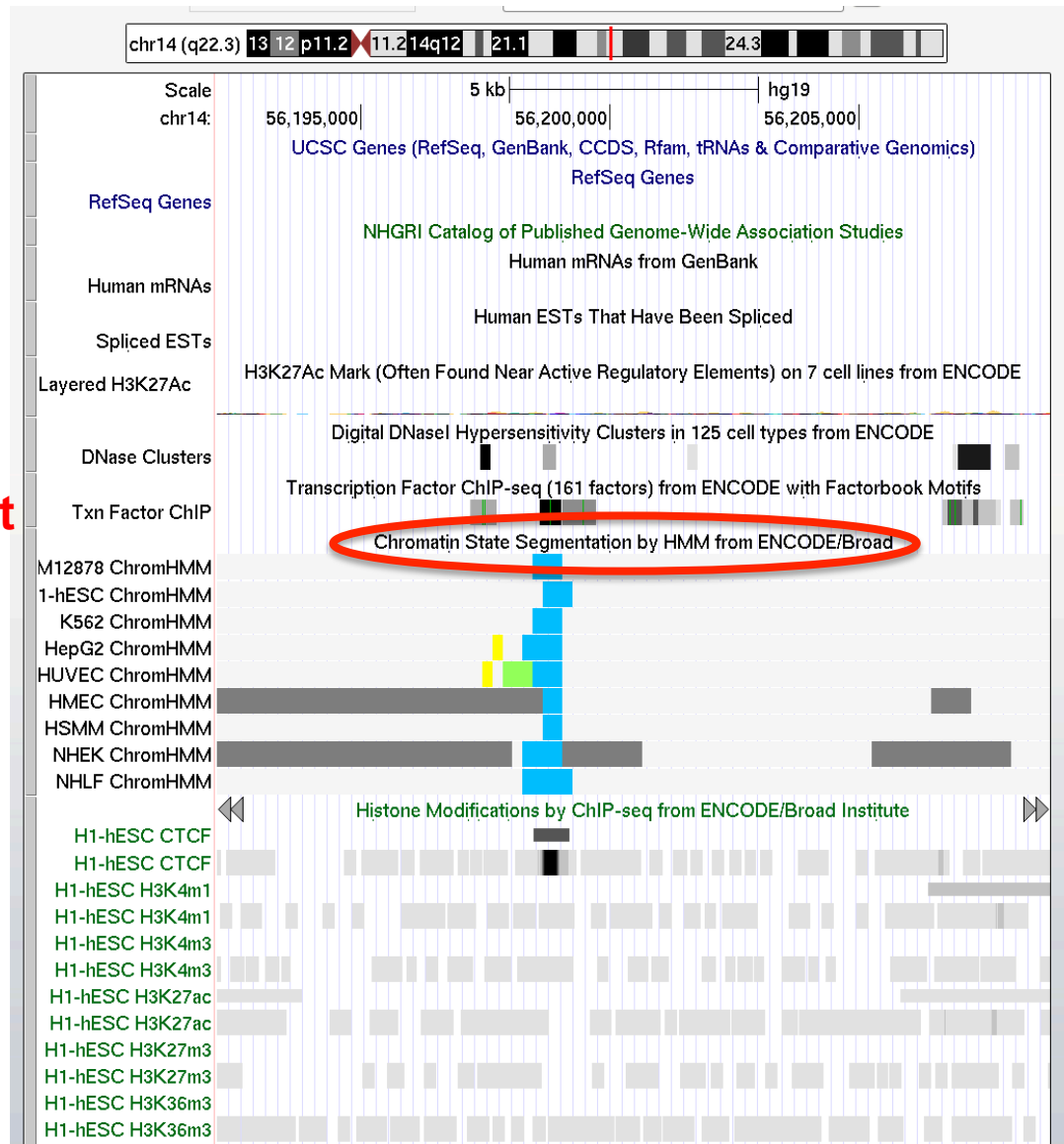
**Regulation** refresh

<input checked="" type="checkbox"/> <a href="#">ENC Histone...</a>	<input type="checkbox"/> <a href="#">ENC RNA Binding...</a>	<input type="checkbox"/> <a href="#">ENC TF Binding...</a>	<input type="checkbox"/> <a href="#">FSU Repli-chip</a>	<input type="checkbox"/> <a href="#">Genome Segments</a>	<input type="checkbox"/> <a href="#">NKI Nuc Lamina...</a>
<input type="checkbox"/> <a href="#">ENC Regulation...</a>	<input type="checkbox"/> <a href="#">CD34 DnaseI</a>	<input type="checkbox"/> <a href="#">CpG Islands...</a>	<input type="checkbox"/> <a href="#">ENC Chromatin...</a>	<input type="checkbox"/> <a href="#">ENC DNA Methyl...</a>	<input type="checkbox"/> <a href="#">ENC DNase/FAIRE...</a>
<input type="checkbox"/> <a href="#">OREgAnno</a>	<input type="checkbox"/> <a href="#">Stanf Nucleosome</a>	<input type="checkbox"/> <a href="#">SUNY SwitchGear TSS</a>	<input type="checkbox"/> <a href="#">SwitchGear TSS</a>	<input type="checkbox"/> <a href="#">TFBS Conserved</a>	<input type="checkbox"/> <a href="#">TS miRNA sites</a>
<input type="checkbox"/> <a href="#">UCSF Brain Methyl</a>	<input type="checkbox"/> <a href="#">UMMS Brain Hist</a>	<input type="checkbox"/> <a href="#">UW Repli-seq</a>	<input type="checkbox"/> <a href="#">Vista Enhancers</a>		

**Comparative Genomics** refresh

Add some tracks that may help us explore function

# Epigenetics & ENCODE

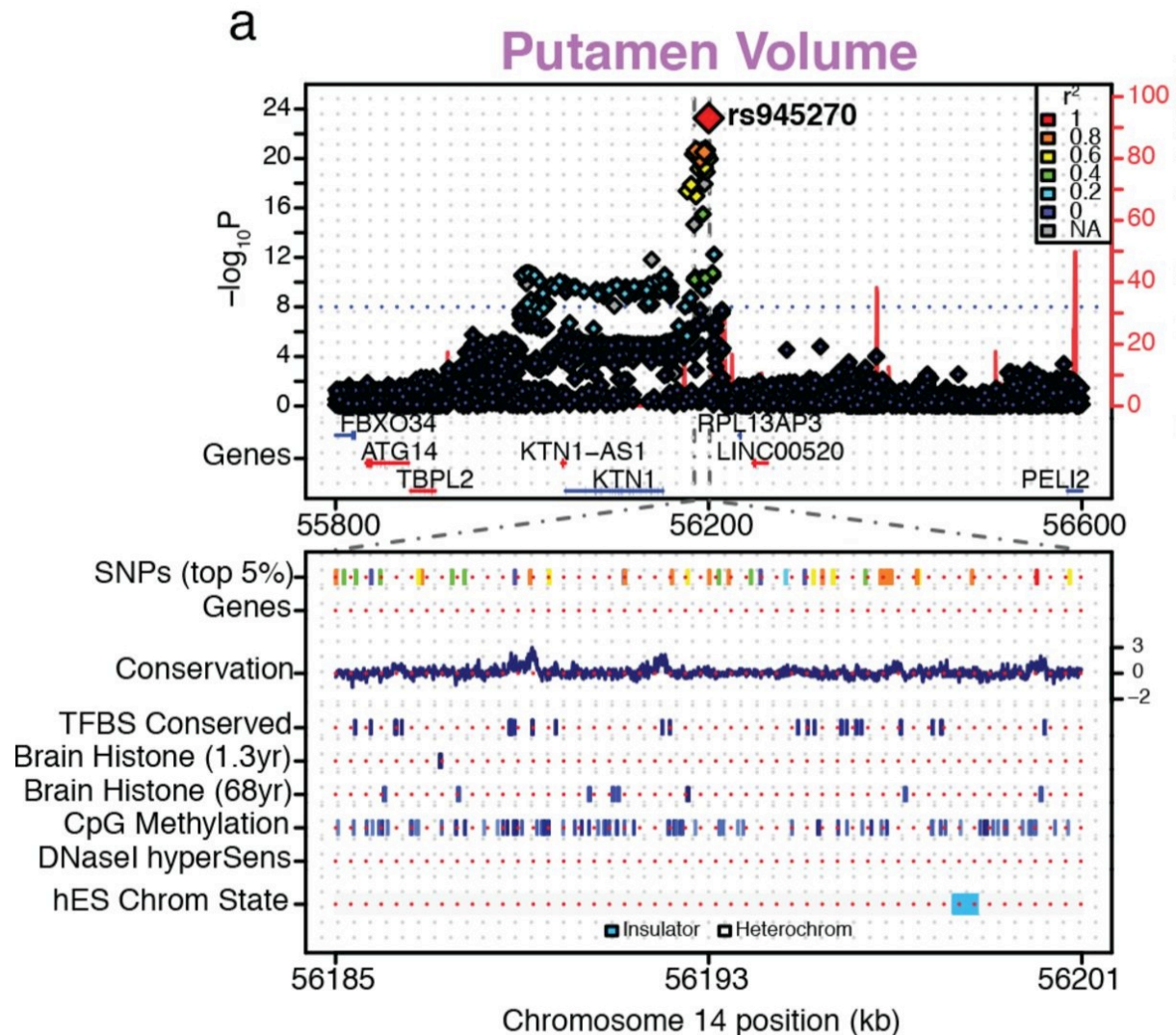


Click to see what the colors mean

Appears to be a CTCF binding site (insulator) very close to the locus!



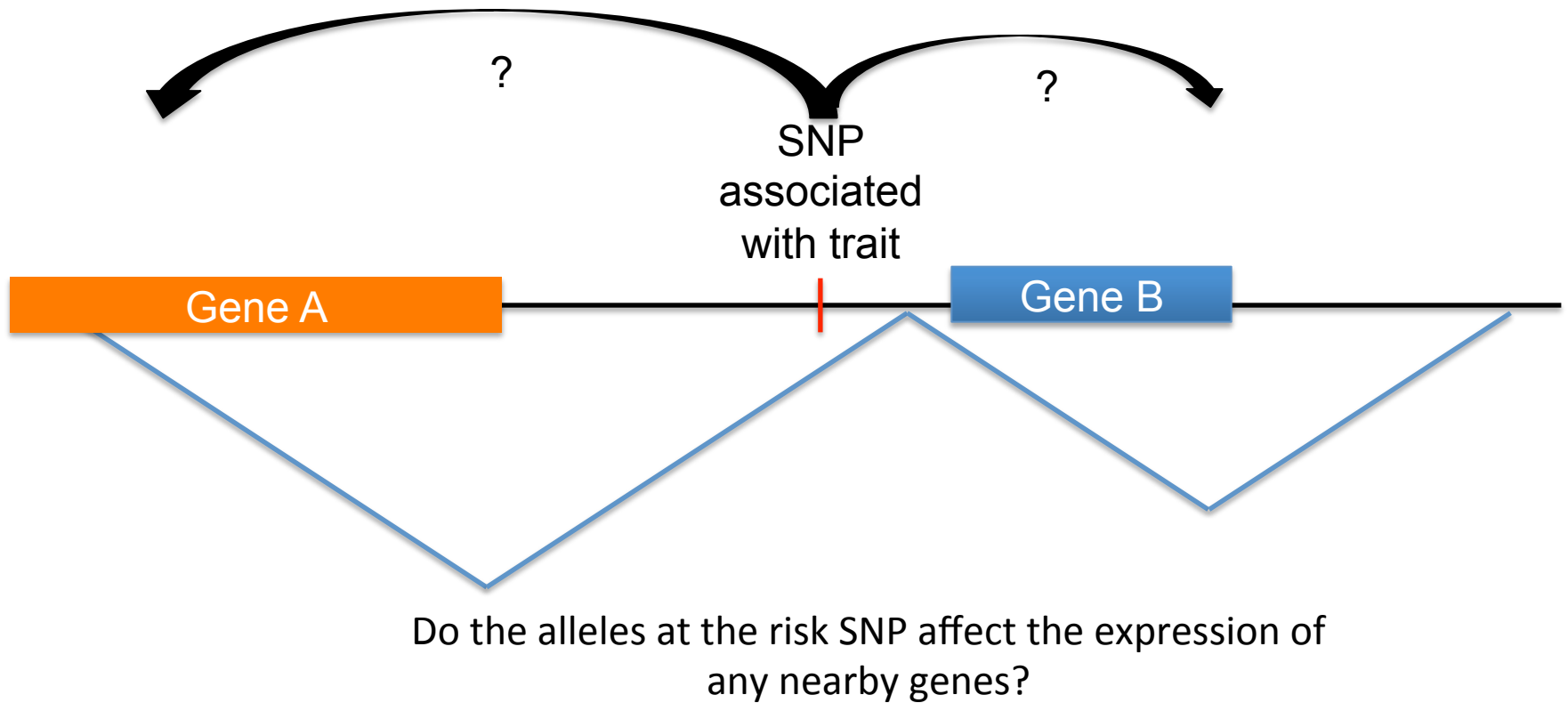
# LocusTrack: Other ways to visualize



<http://gump.qimr.edu.au/general/gabrieC/LocusTrack/index.html>

# expression QTLs (eQTLs)

---



A way to move from variant to gene.

# eQTL databases

## BRAINEAC

**BRAINEAC**  
Web server for data from the UK Brain Expression Consortium

Braineac

The aim of Braineac is to release to the scientific com

By Gene By SNP Download Data

"By SNP" visualise the top ten most affected genes by SNP and relative p-values. It's also possible to download the full list of genes affected.

Enter SNP rsid or chr:pos(hg19). To insert multiple SNPs, insert each snp per row up to 20 SNPs.

e.g. rs123	rs2248359
...	

<http://www.braineac.org/>

## Blood eQTL browser

### Blood eQTL browser

This web page accompanies the manuscript titled *Systematic identification of trans-eQTLs* which has been published in Nature Genetics. If you want to use any of the cis- or trans-eQTLs as indicated below. For further questions, contact the corresponding author: lude@ludesign.nl

### Download eQTL Results

You can download the full cis- and trans-eQTLs, detected at a false-discovery rate of 0.50:  
Cis-eQTLs (FDR 0.5)  
Trans-eQTLs (FDR 0.5)

### How to cite

If you use the eQTLs present on this website in your paper or research, please cite our work: D:

### Query eQTL Results

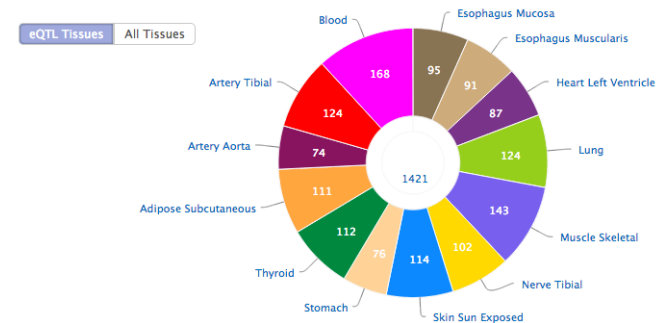
Or, you can query the cis- and trans-eQTLs below (examples: rs7807018 or VWCE):

Gene or SNP name:  Search

<http://www.genenetwork.nl/bloodeqtlbrowser/>

## GTEx

Browse eQTLs for Tissues with n > 60



<http://www.gtexportal.org/home/>  
(Brain on its way)

## NCBI eQTL Browser

NCBI Resources How To

**eQTL Browser**

Search Parameters

Display Results Download Text Clear Form Tutorial

ID	Tissue	Title
<input type="checkbox"/>	1 Lymphoblastoid	Transcriptome genetics using second generation sequencing in a Caucasian population.
<input type="checkbox"/>	2 Liver	Mapping the genetic architecture of gene expression in human liver
<input type="checkbox"/>	3 Brain Cerebellum	Abundant quantitative trait Loci exist for DNA methylation and gene expression in human brain
<input type="checkbox"/>	4 Brain Frontal Cortex	Abundant quantitative trait Loci exist for DNA methylation and gene expression in human brain
<input type="checkbox"/>	5 Brain Temporal Cortex	Abundant quantitative trait Loci exist for DNA methylation and gene expression in human brain
<input type="checkbox"/>	6 Brain Pons	Abundant quantitative trait Loci exist for DNA methylation and gene expression in human brain
<input type="checkbox"/>	7 Lymphoblastoid	Population genomics of human gene expression

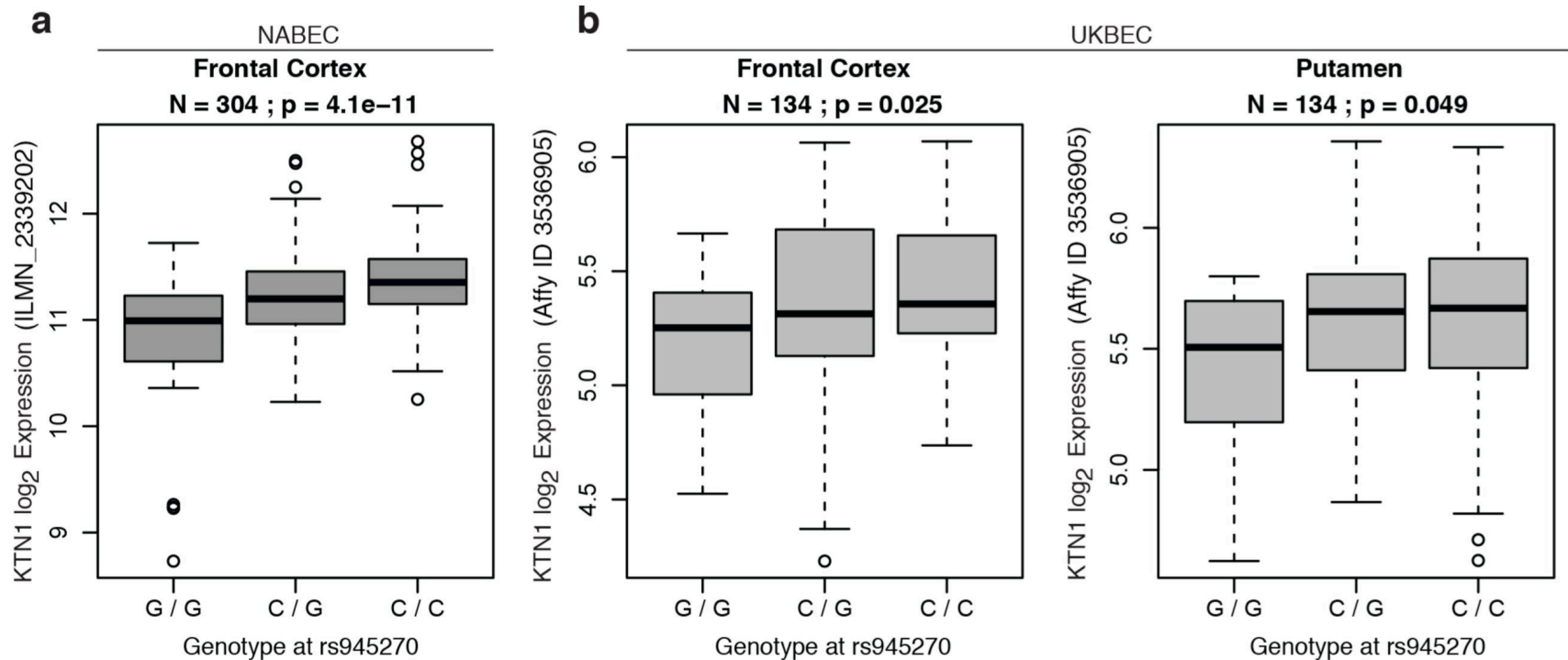
Select All Invert Selection

SNP filters  
RS numbers

Gene Expression Filters  
Gene symbols, gene IDs, RefSeq IDs, and/or Pr

<http://www.ncbi.nlm.nih.gov/projects/gap/eqtl/index.cgi>

# eQTL phone a friend



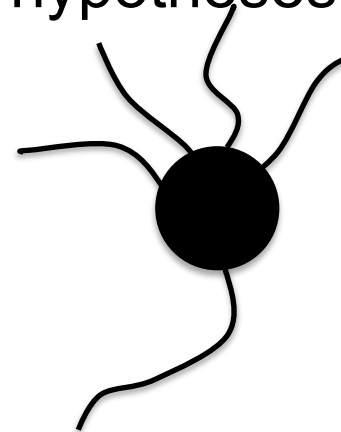
This SNP affects a gene (replicated in brain),  
we now have a gene!

# When and Where is Gene Expressed?

---

- 86-95% of genes are expressed in the brain at some point during the lifespan and 90% of those were differentially regulated across region or time (Kang et al., 2011; Miller et al., 2014).
- Expression of a gene in brain does little to implicate it as causal.
- However, finding the time period or region of gene expression may lead us to cell type hypotheses.

Gene A



# When and Where is Gene Expressed?

**BRAINSPAN**  
ATLAS OF THE DEVELOPING HUMAN BRAIN

Home **Developmental Transcriptome** Prenatal LMD Microarray ISH Reference Atlas Download Documentation Help

Enter Gene Name, Gene Symbol, Entrez Gene ID or Ensembl ID

Gene Search  Differential Search

## Expression by time →

1 - 2 of 2

■ Donor H376.IIA.51 Age: 8 pcw ■ dorsolateral prefrontal cortex (DFC)



## Expression by region →

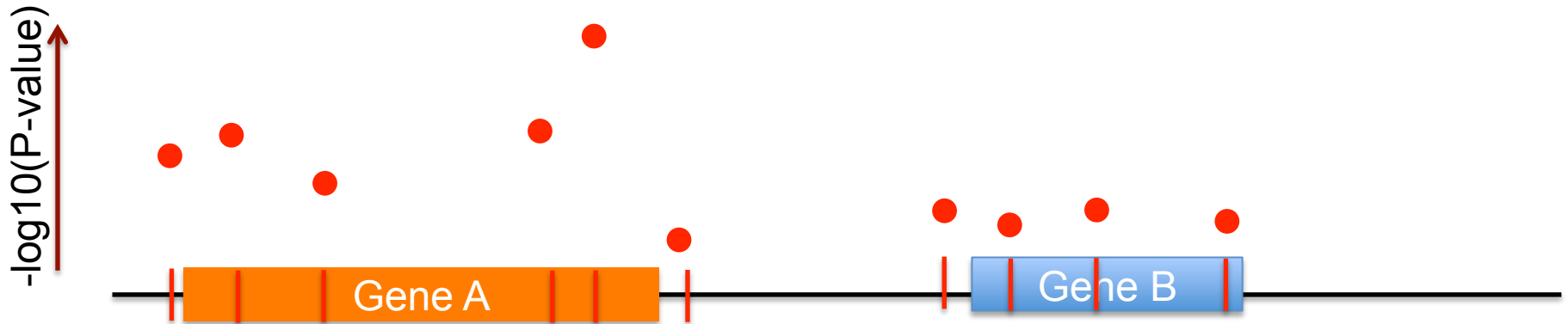
1 - 2 of 2

■ Donor H376.IIA.51 Age: 8 pcw ■ dorsolateral prefrontal cortex (DFC)



<http://brainspan.org/rnaseq/search/index.html>

# Gene-based tests



Combines SNP associations across genes to form a gene based p-value

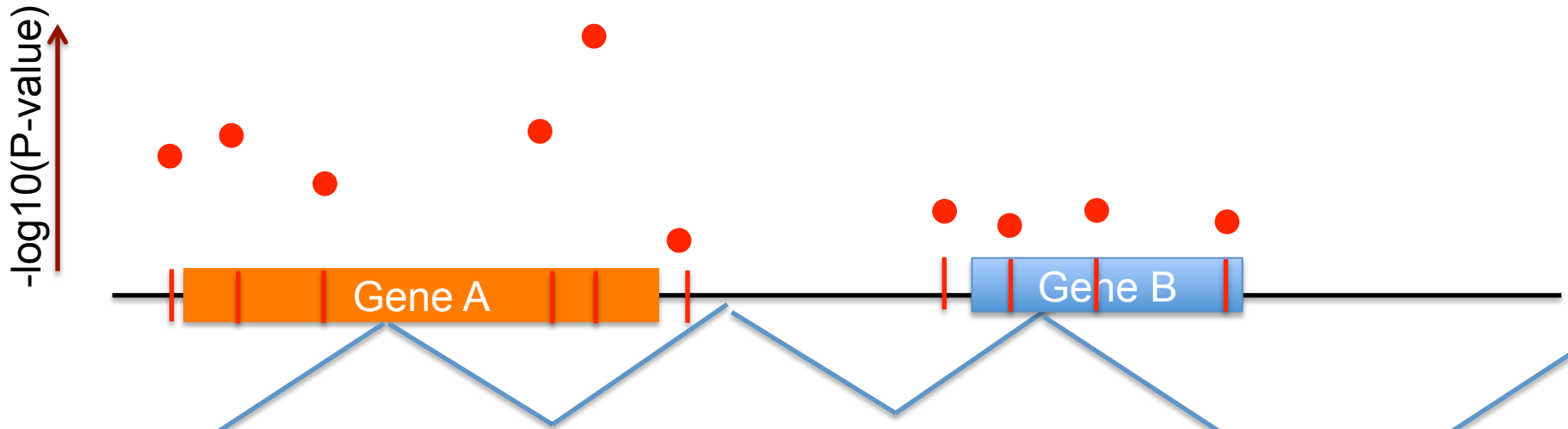
## Advantages

- Greater interpretability
- Fewer multiple comparisons
- Can feed into pathway based approaches

## Disadvantages

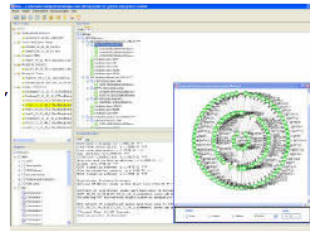
- Ignores intergenic variation (most of genome)
- Generally ignores direction of association
- Ignores that a variant within the intron of one gene may be affecting a totally different gene

# Tools for Gene Based Analyses



Correct all SNP based p-values within a gene for the total number of independent tests, then take the minimum p-value.

See GATES algorithm implemented in KGG toolbox (Li et al., 2011)

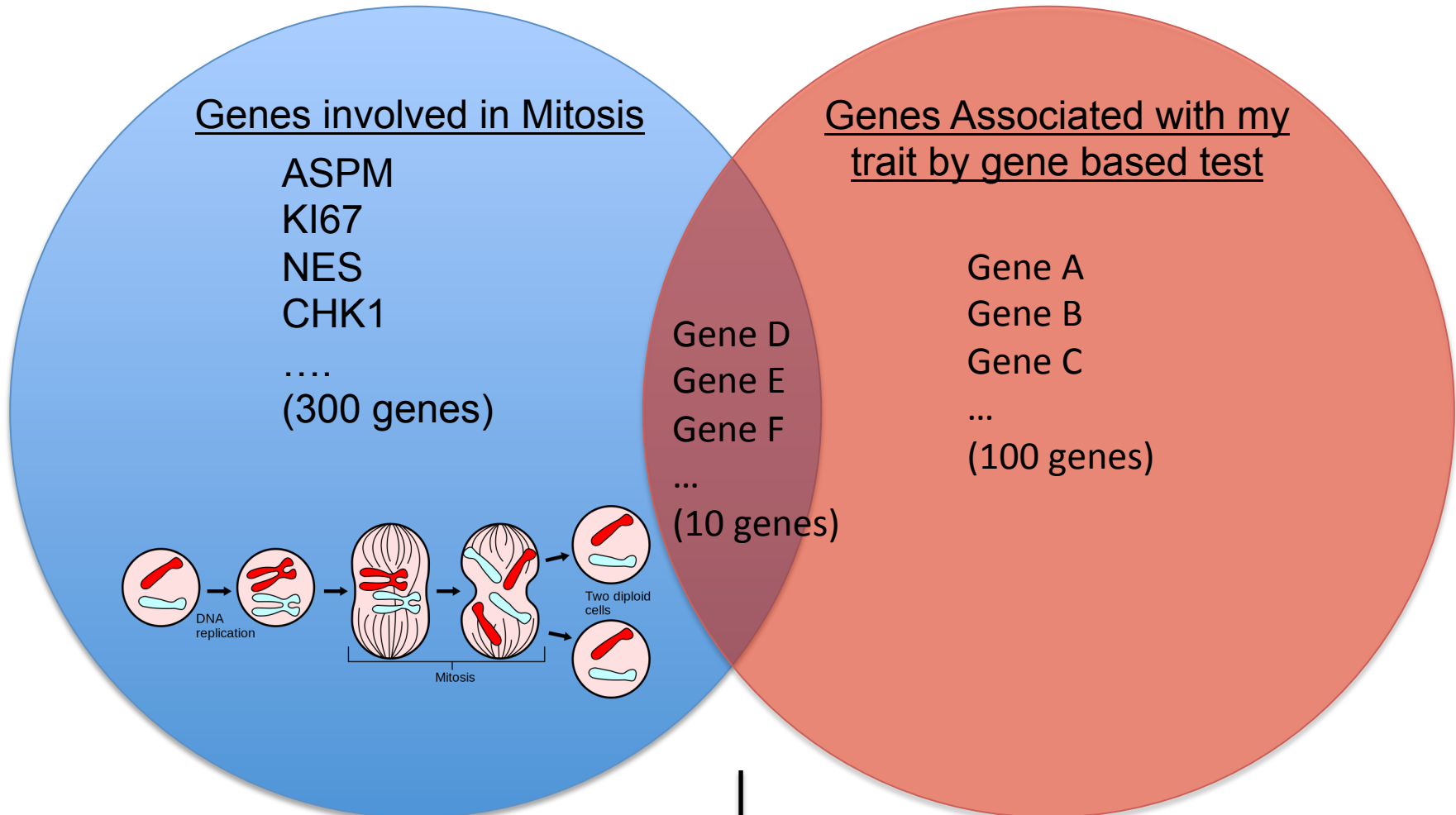


<http://statgenpro.psychiatry.hku.hk/limx/kgg/index.html>



# Pathway Analysis

Look for enrichment of your associated genes in known pathways



## Advantages

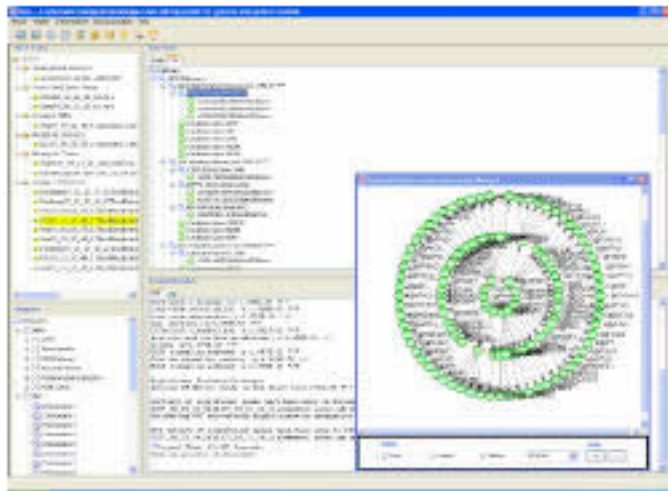
- Amazing interpretability – exactly what you're looking for

## Disadvantages

- Uses gene based tests
- Pathway gene lists are generally not well known

# Tools for Pathway Based Analyses

## Knowledge-Based Mining System for Genome-Wide Genetic Studies (KGG)



<http://statgenpro.psychiatry.hku.hk/limx/kgg/index.html>

## MAGENTA

**MAGENTA:** Meta-Analysis Gene-set Enrichment of variant Associations



<http://www.broadinstitute.org/mpg/magenta/>

# Conclusions

---

- Identifying the genetic locus is a causal foothold into understanding novel biological mechanisms.
- There are many databases and tools that will allow you to form hypotheses about the biological mechanisms.
- It's easy to make a story! Let the evidence guide you.

# Acknowledgements

---



**Derrek Hibar, IGC**  
**Sarah Medland, QIMR**  
**Miguel Renteria, QIMR**

**Paul Thompson, IGC**