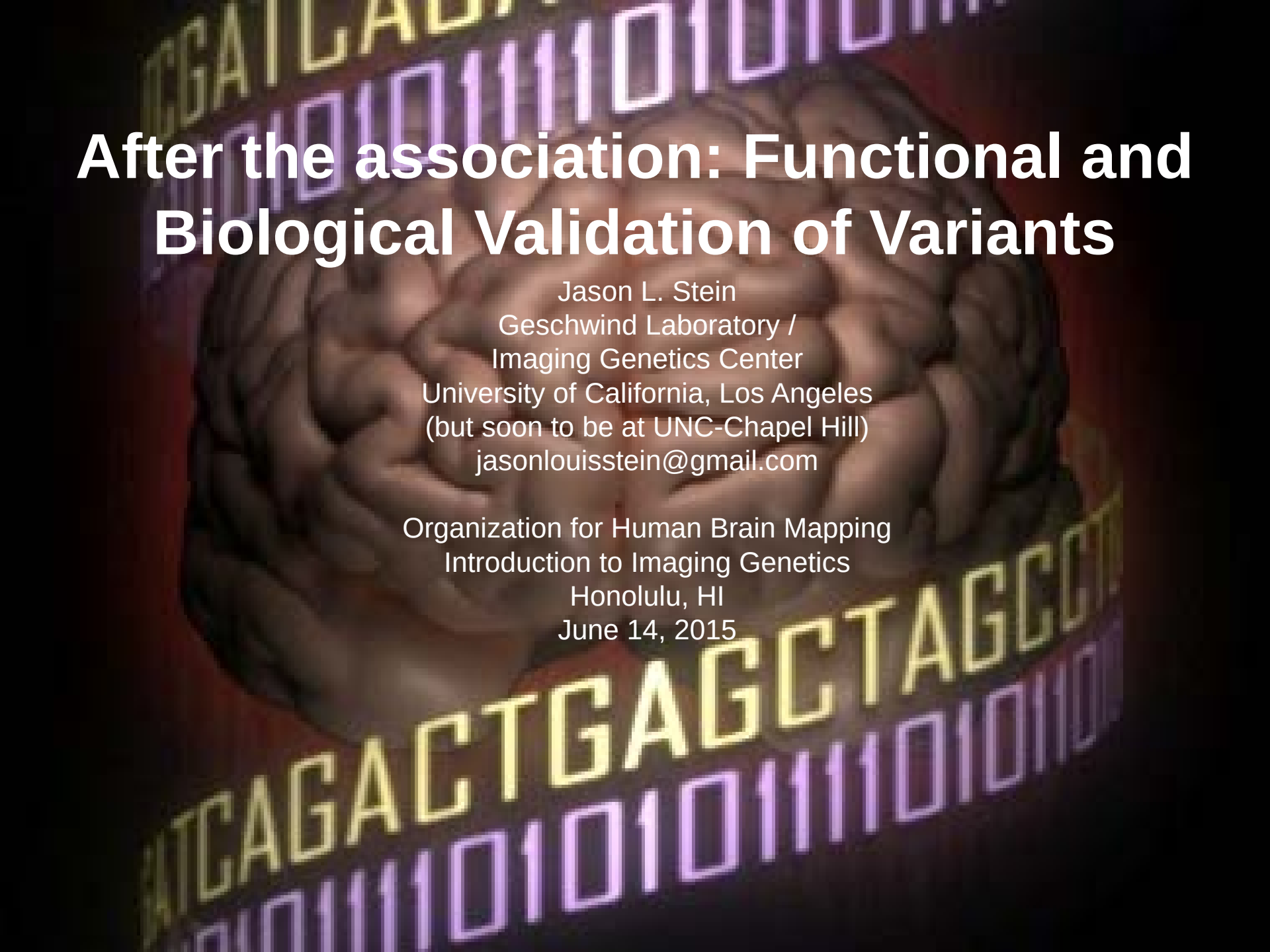
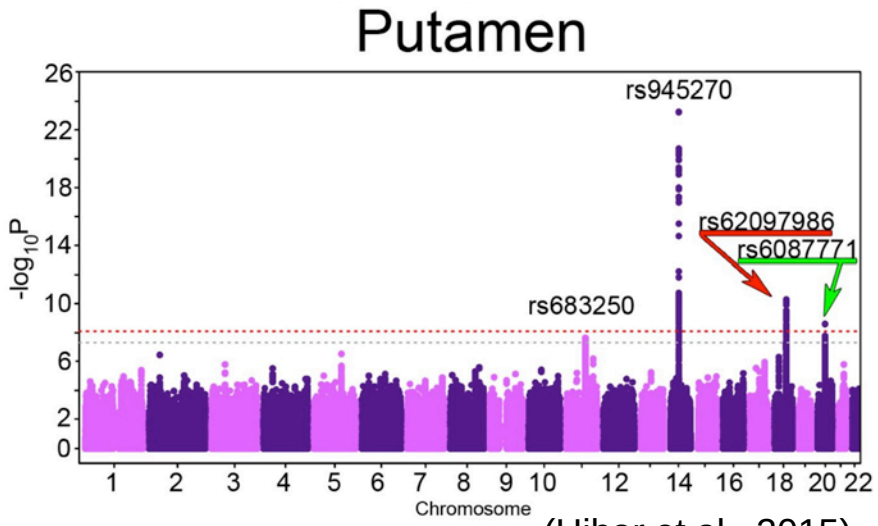
After the association: Functional and Biological Validation of Variants

Jason L. Stein
Geschwind Laboratory /
Imaging Genetics Center
University of California, Los Angeles
(but soon to be at UNC-Chapel Hill)
jasonlouisstein@gmail.com

Organization for Human Brain Mapping
Introduction to Imaging Genetics
Honolulu, HI
June 14, 2015



A hit is just the beginning…



(Hibar et al., 2015)

<u>What you have found</u>

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

<u>What you want to know</u>

- 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[1], Andrew Allen[1,2], Jonathan Keebler[1], Elliott H. Margulies[3], Steven Petrou[4,5], Slavé Petrovski[1,6] and Shamil Sunyaev[7]

"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[1,2], Guoping Feng[1,3,4] & Steven E Hyman[1,5]


ORIGINAL ARTICLES

Spurious Genetic Associations

Patrick F. Sullivan

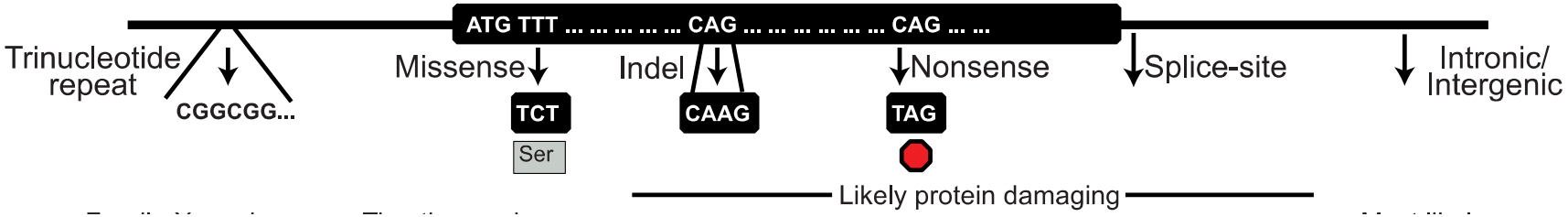
"Findings from single association studies constitute '**tentative knowledge**' and must be interpreted with exceptional caution.

<u>Biological plausibility is not a substitute for statistical significance</u>

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

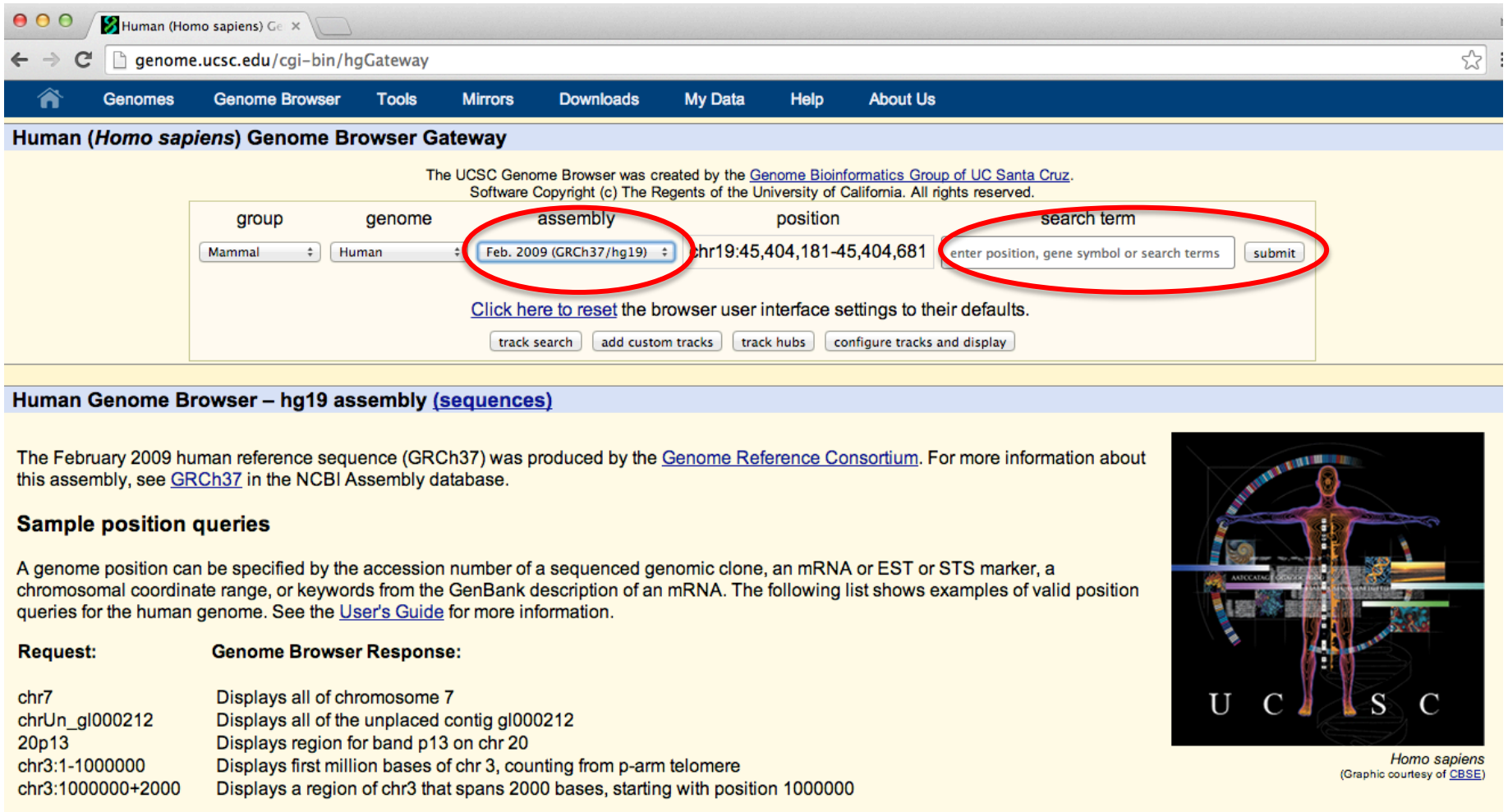


- 88% of significant GWAS (common variant) loci are often found in intergenic or intronic regions with no clear gene(s) of action (Hindorff et al., *PNAS*, 2009)

- 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)

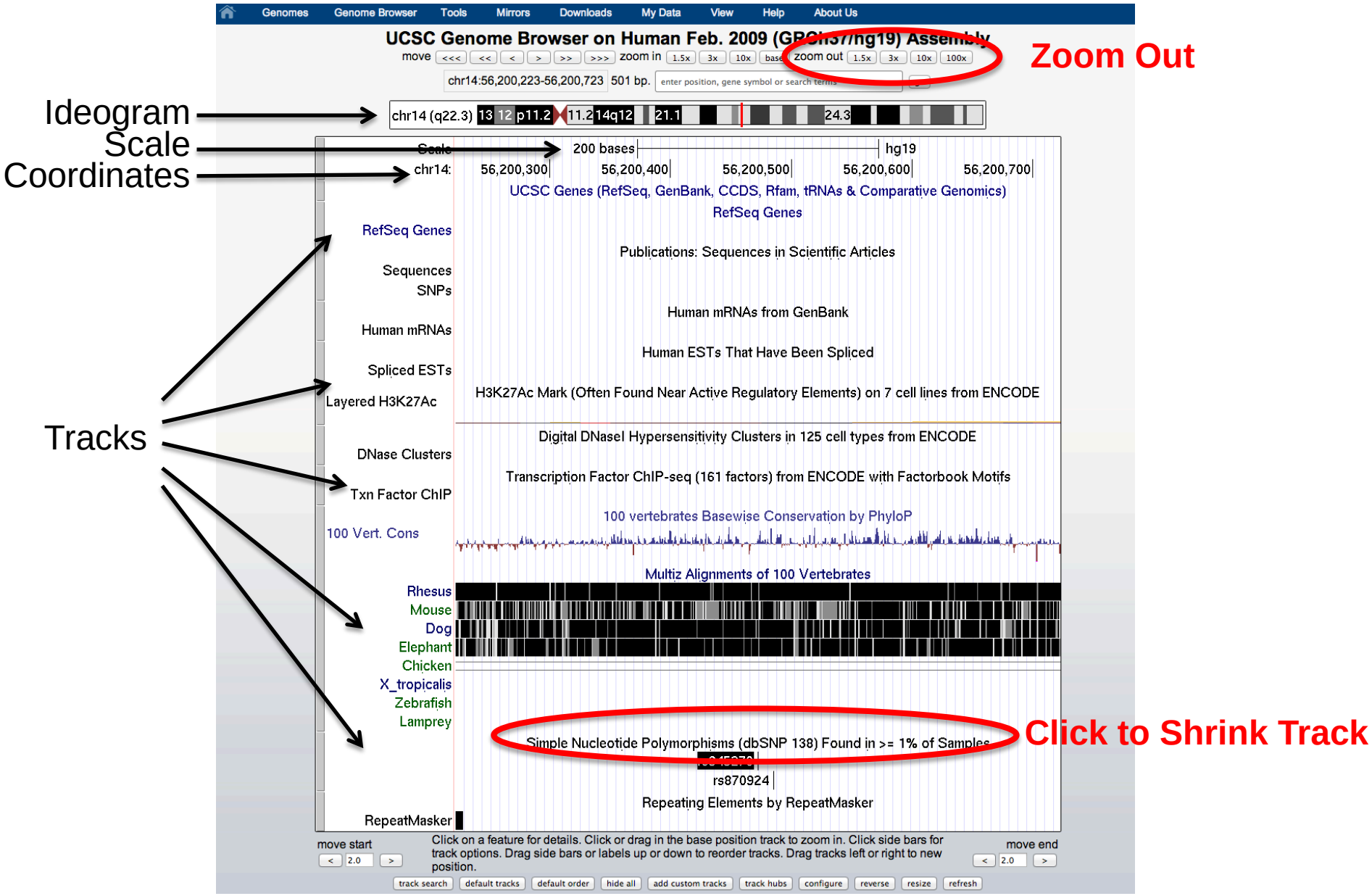
- Rare variant mutations including missense or nonsense mutations have a clear gene of action, but are rare.

UCSC Genome Browser

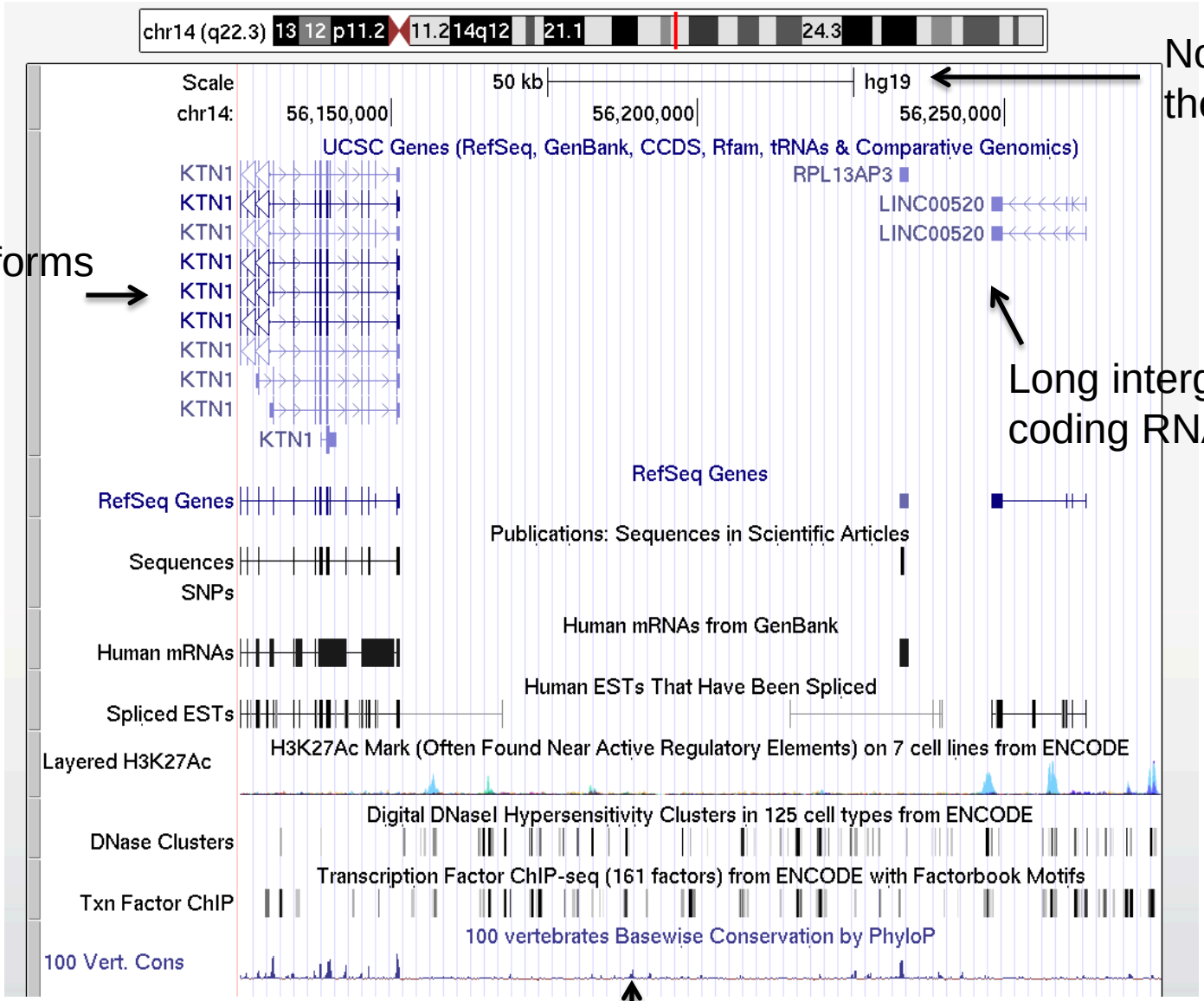


http://genome.ucsc.edu/cgi-bin/hgGateway

Exploring a hit (rs945270)



Exploring a hit (rs945270)



Notice the scale

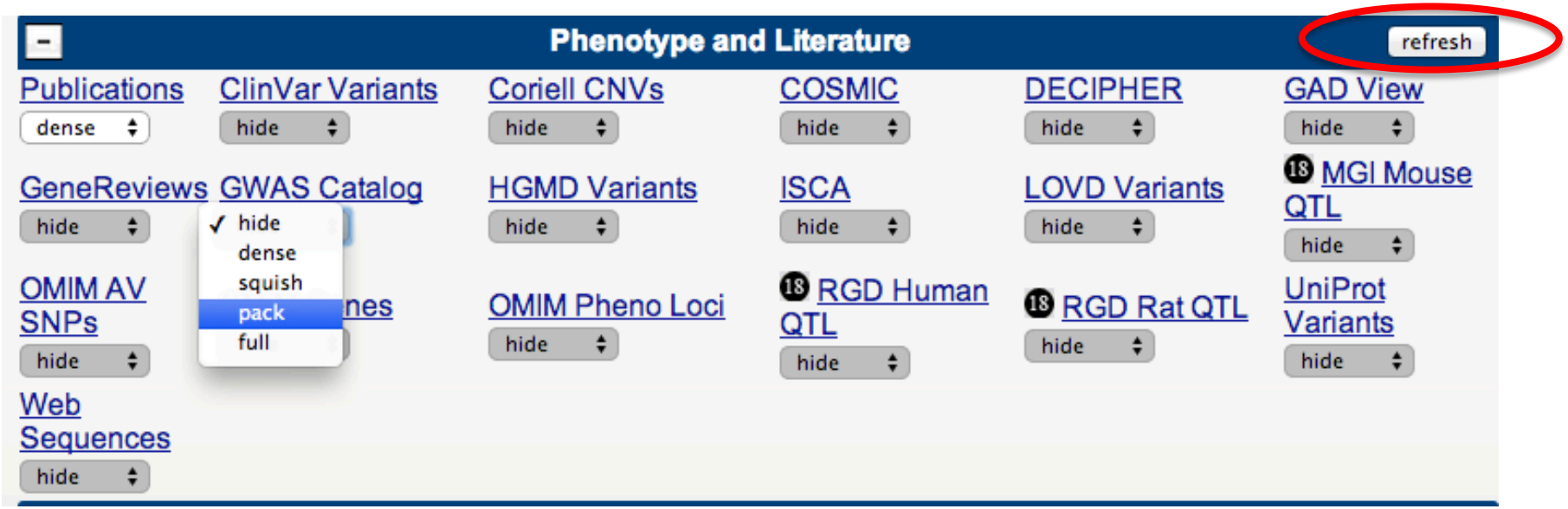
Multiple isoforms of genes

Long intergenic non-coding RNAs

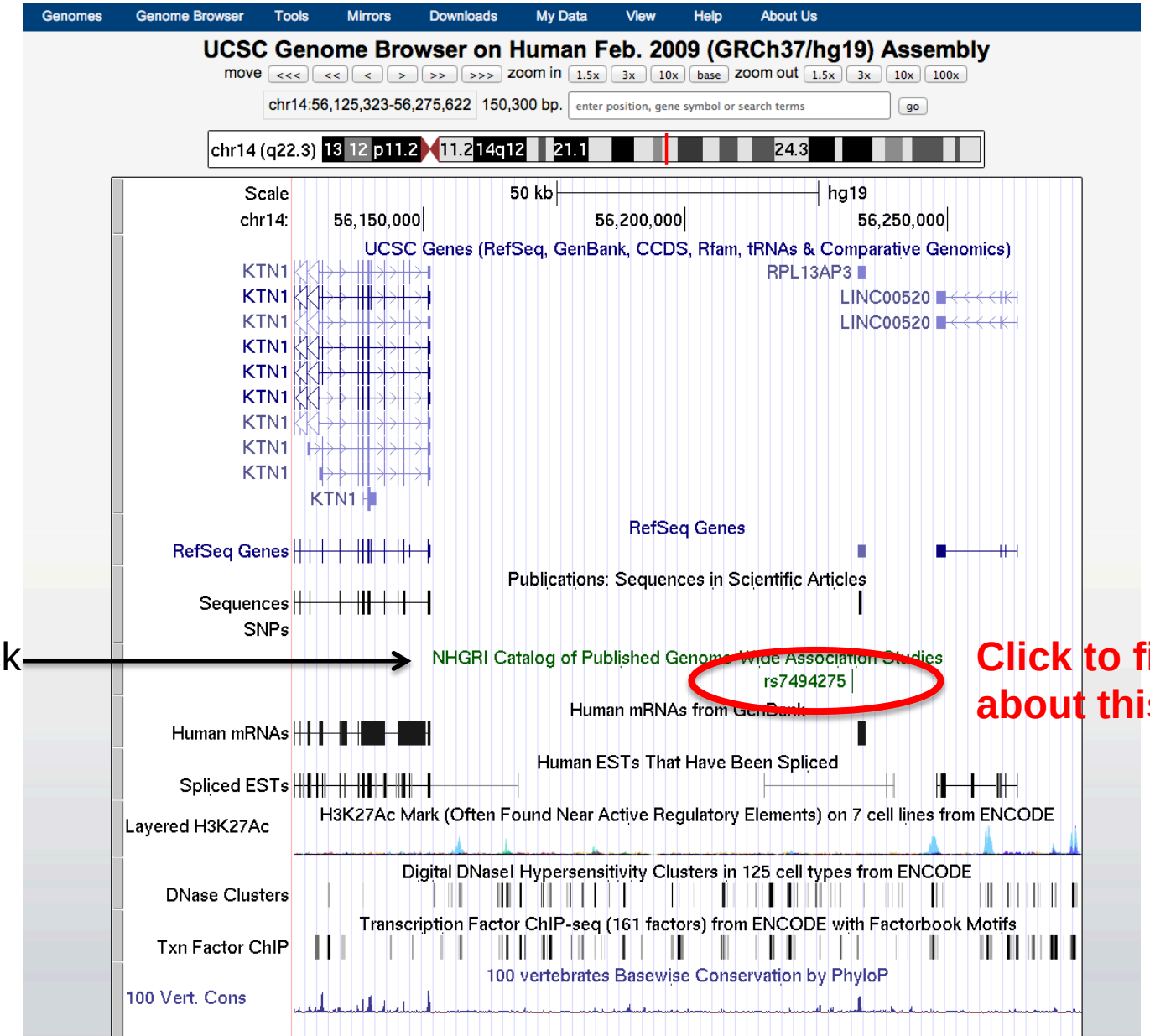
The variant initially entered stays in the center

Exploring a hit (rs945270)

Add a Track



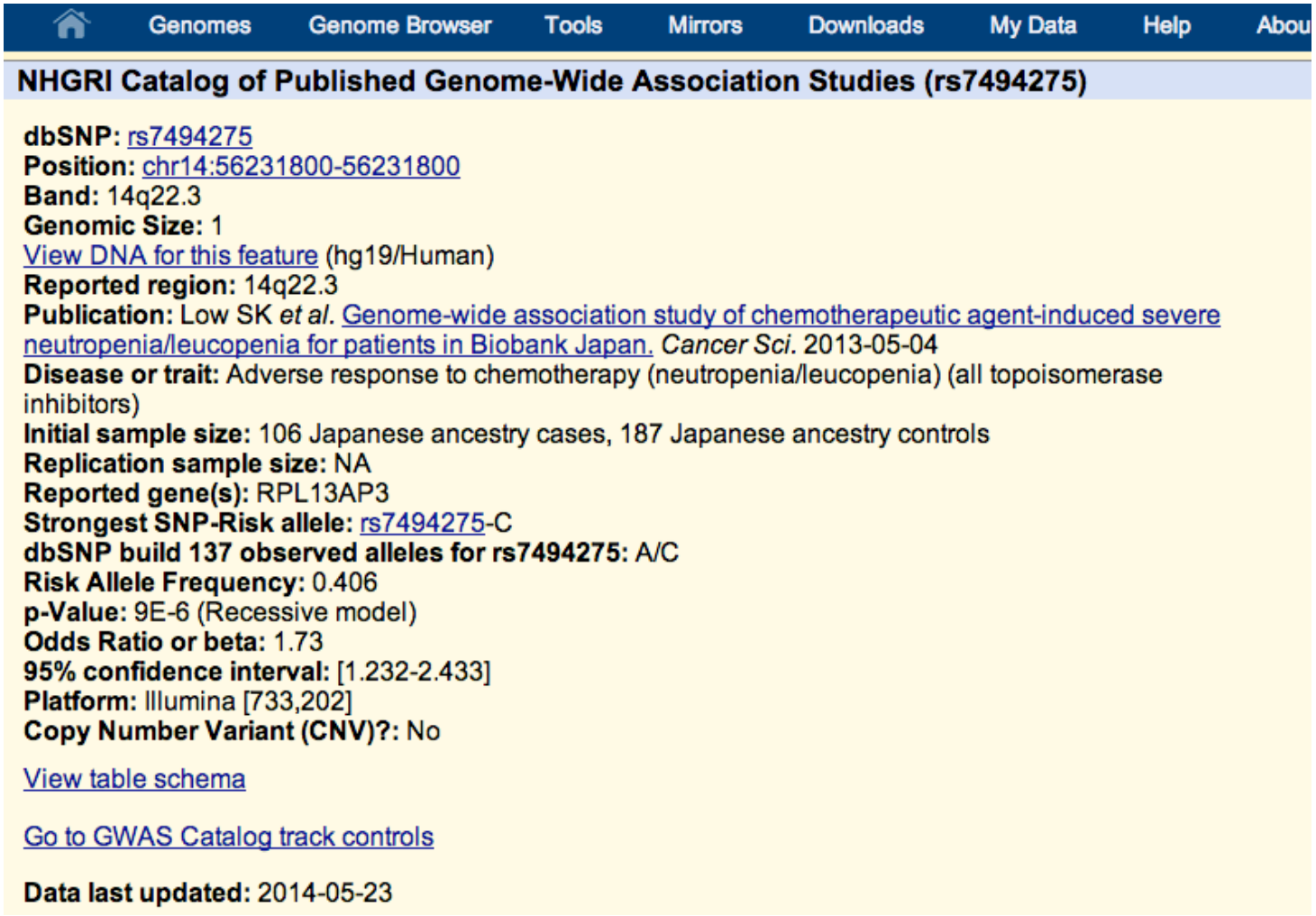
Exploring a hit (rs945270)



New Track

Click to find out about this variant

Exploring a hit (rs945270)



Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | Abou

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

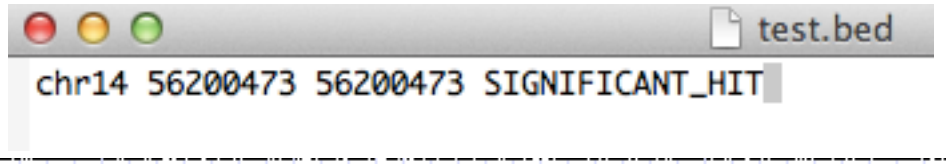
View table schema

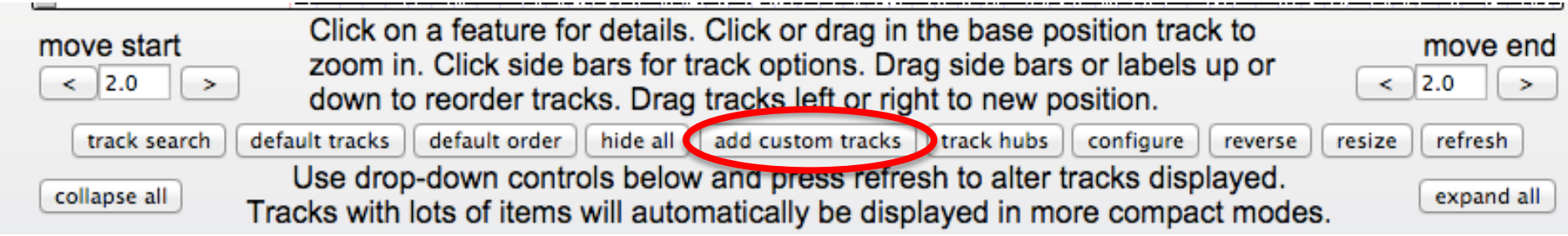
Go to GWAS Catalog track controls

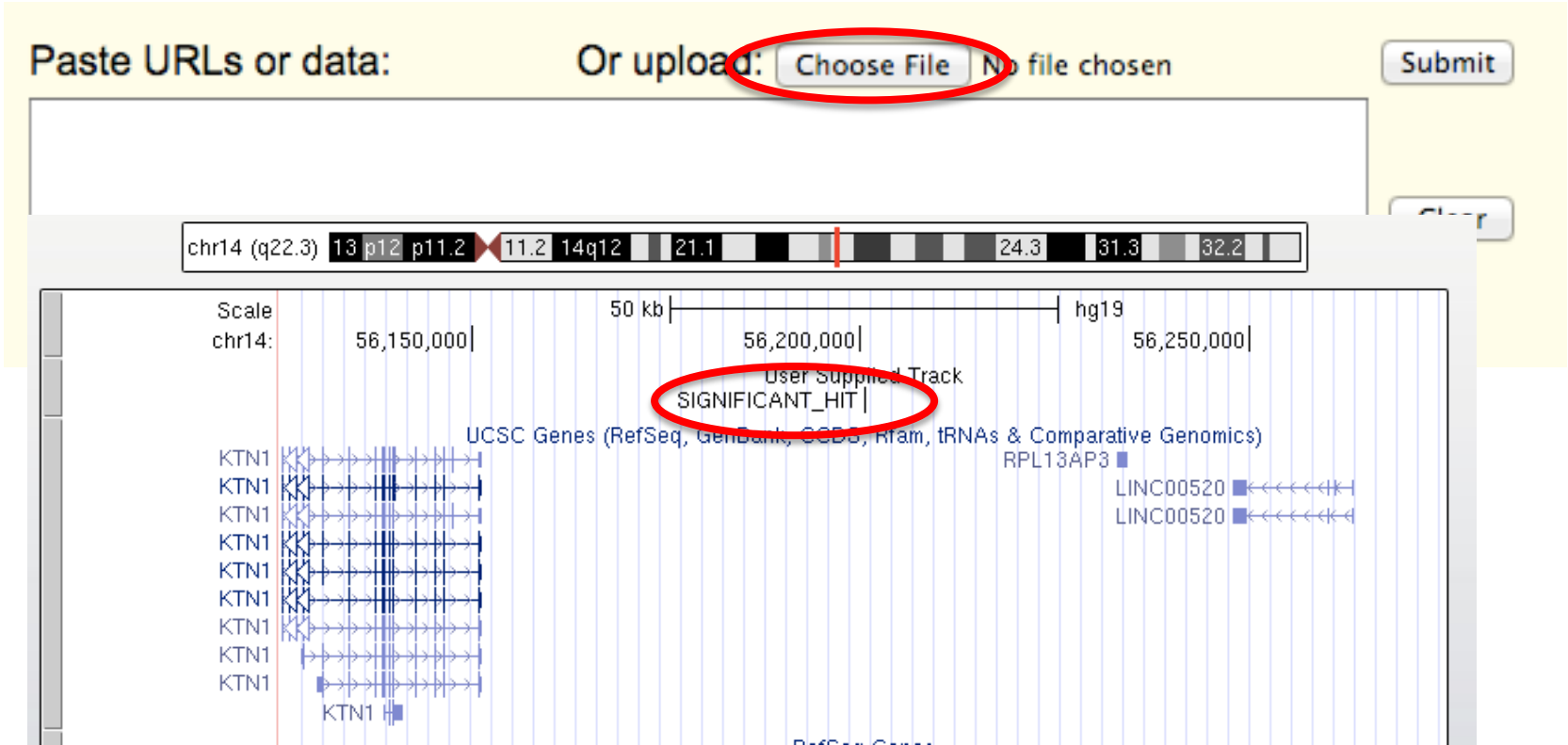
Data last updated: 2014-05-23

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

Uploading a user track

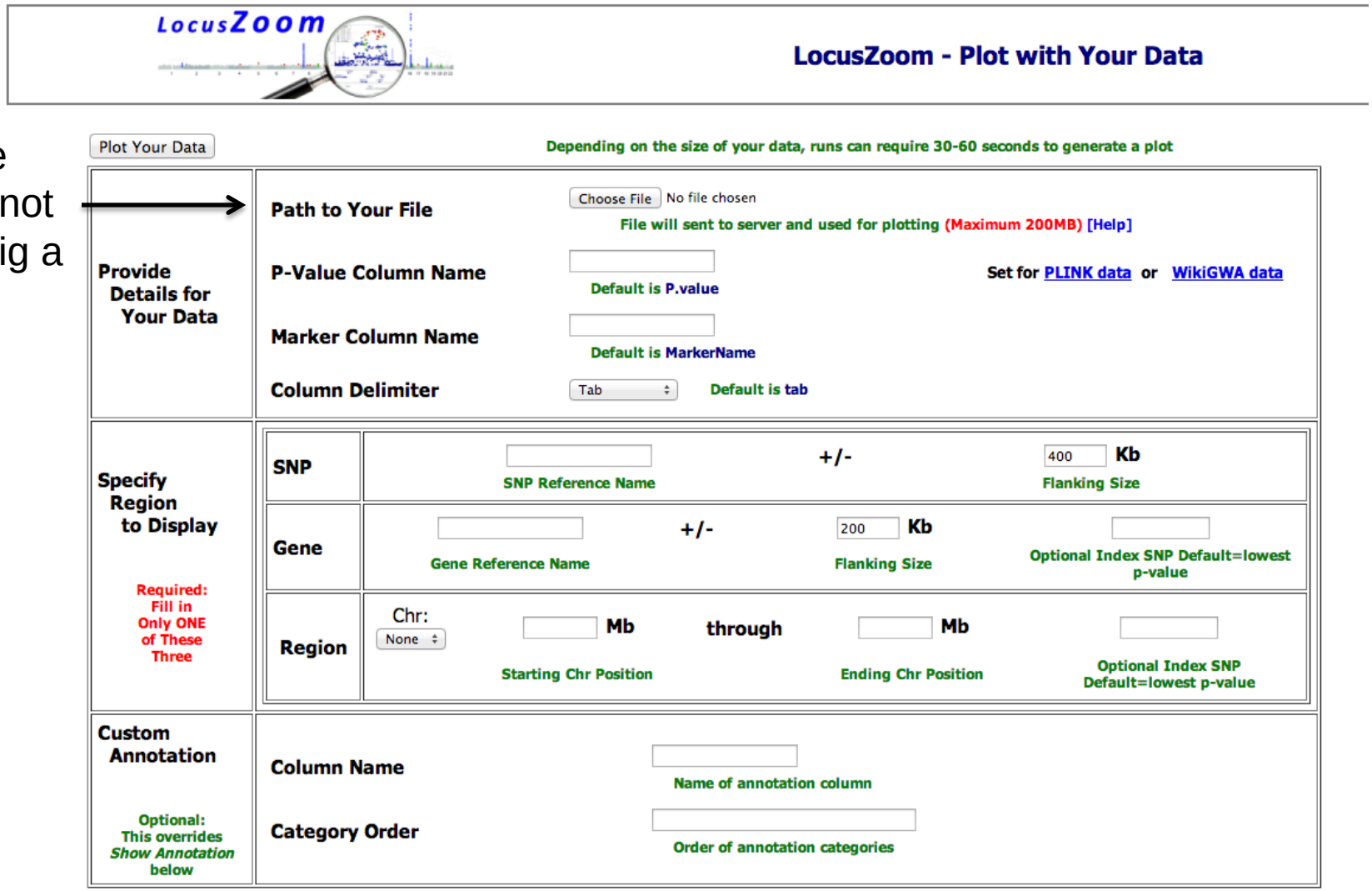






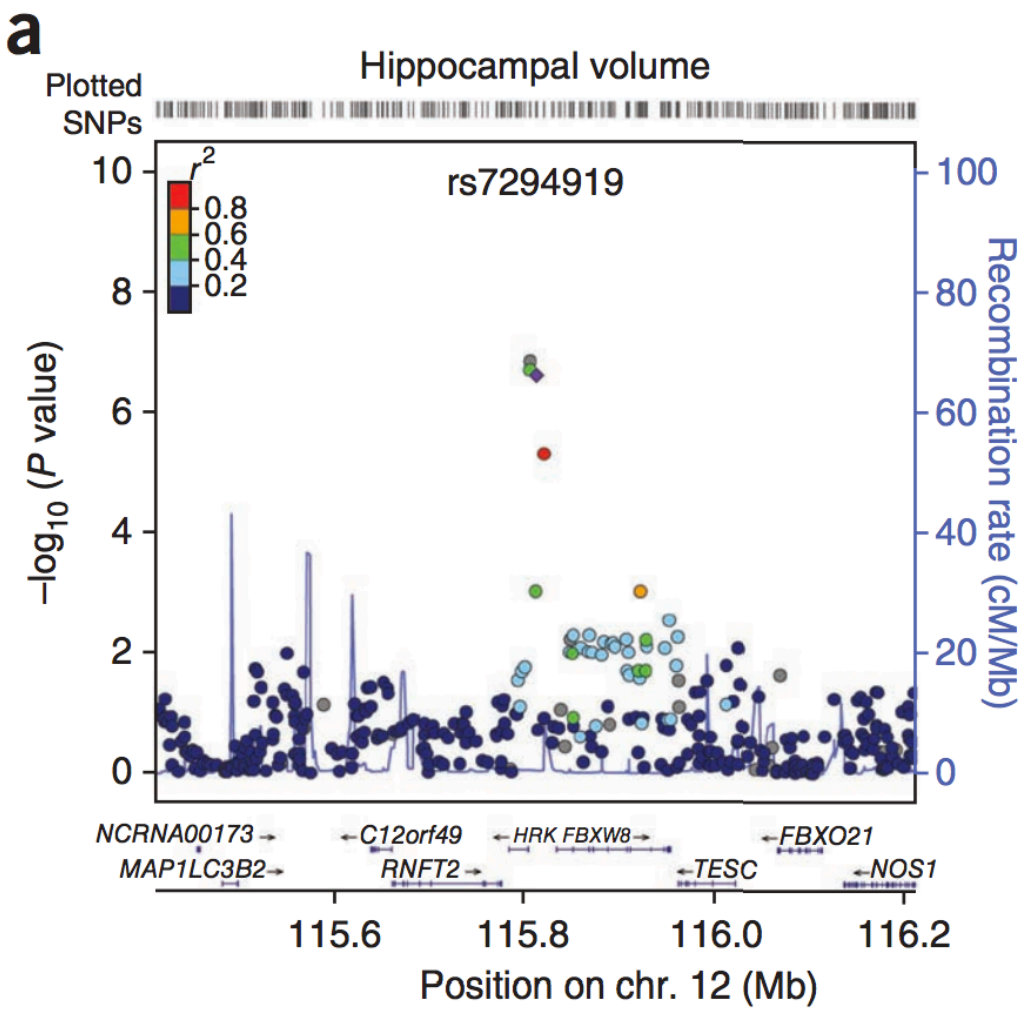
LocusZoom: Making Prettier Pictures

Make sure not too big a file

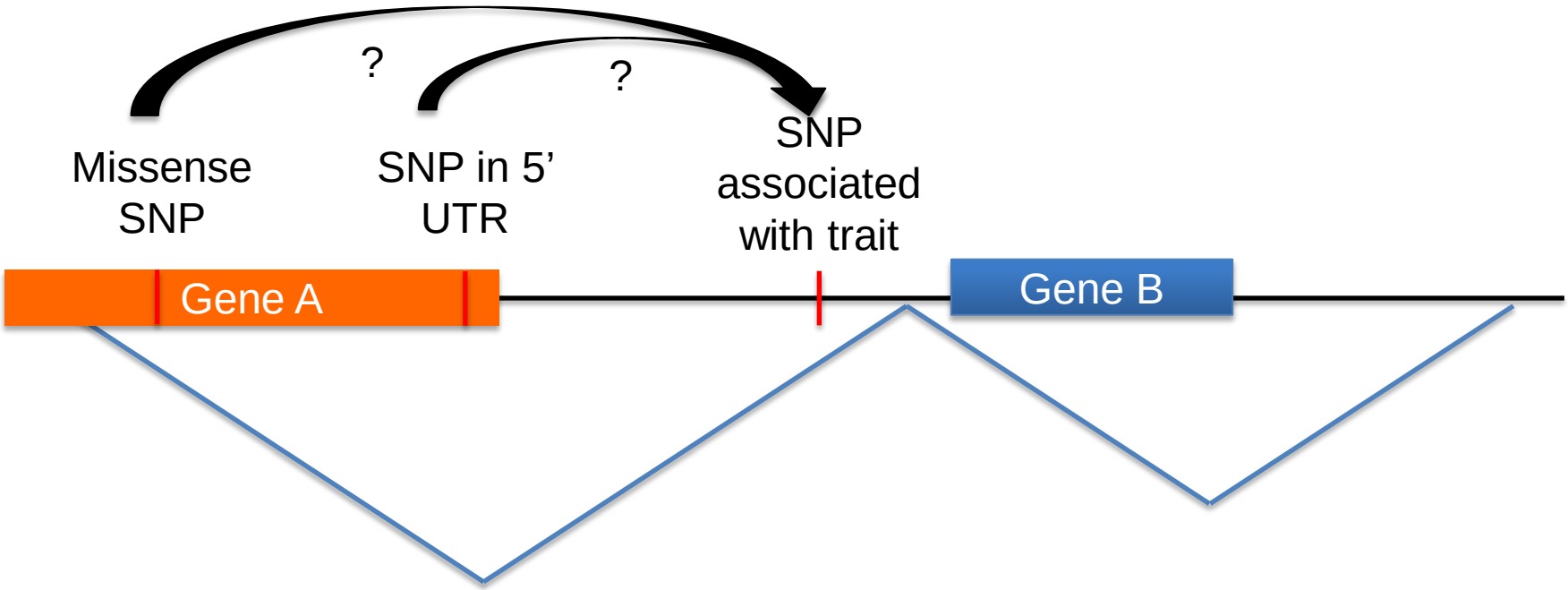


https://statgen.sph.umich.edu/locuszoom/genform.php?type=yourdata

LocusZoom: Making Prettier Pictures

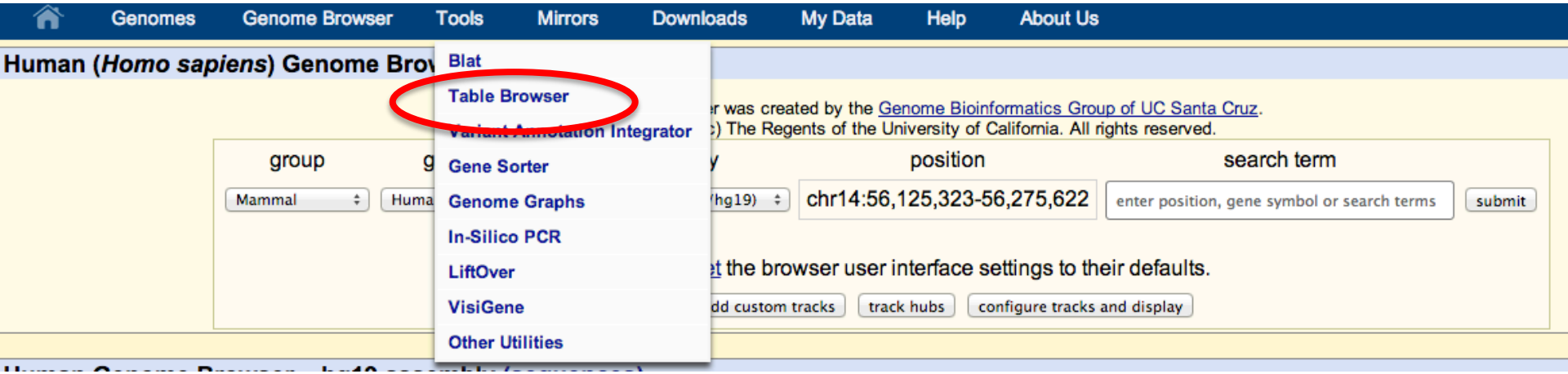


Trying to find possible gene function



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

Finding Variants of Known Functionality



Go to the Table Browser in UCSC Genome Browser

Finding Variants of Known Functionality

Select interpretable functional Variants

Click to create a filter

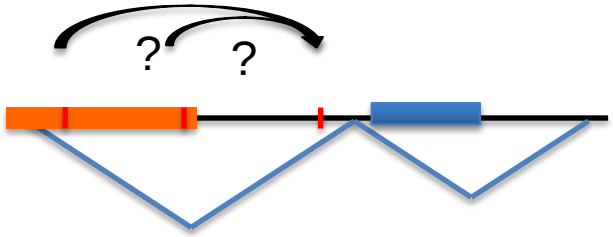
Get output spreadsheet

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

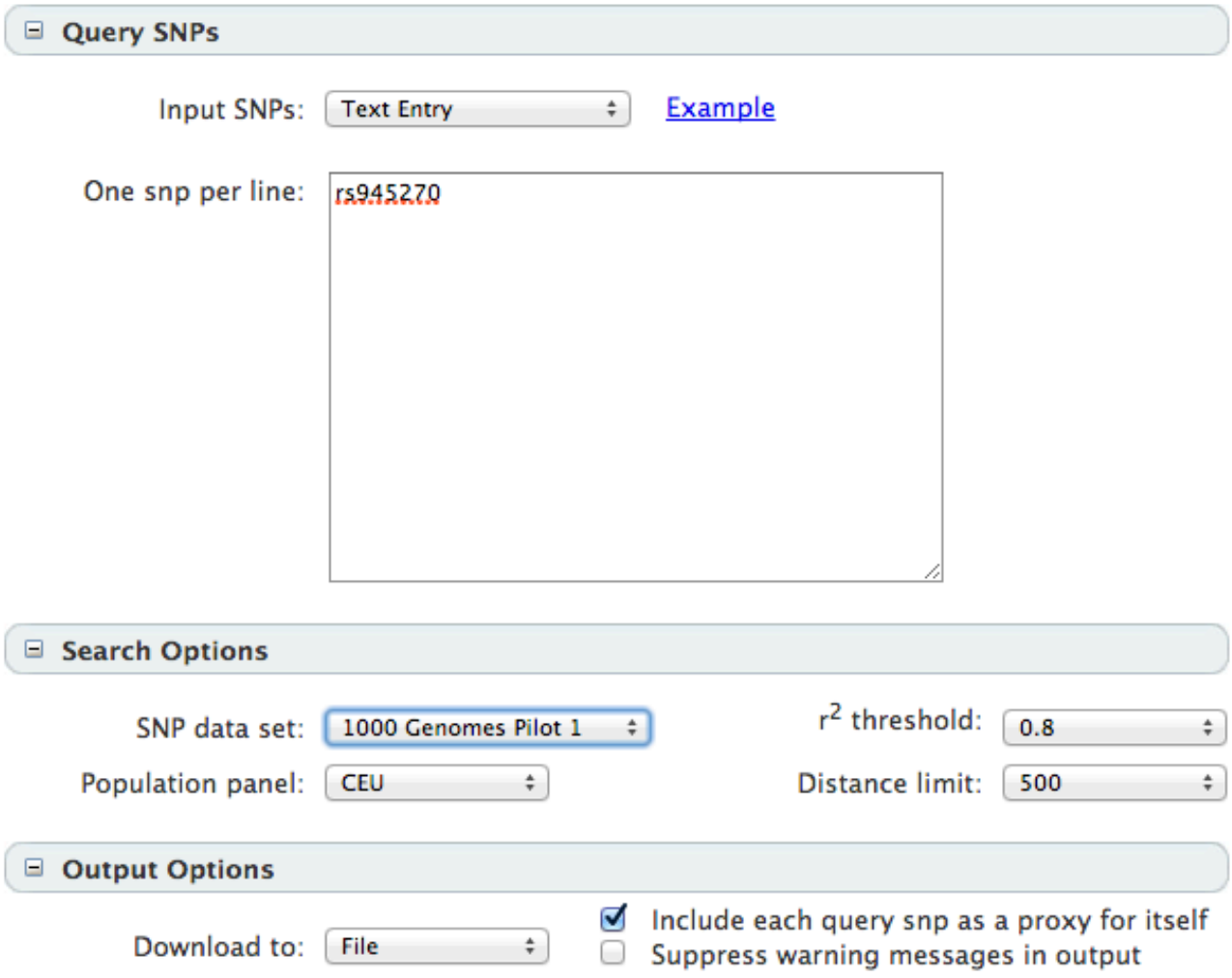
#bin	chrom	chromStart	chromEnd	name	refNCBI	refUCSC	observed	func
1012	chr14	56068483	56068484	rs116289145	C	C	C/T	intron,ncRNA,untranslated-5
1012	chr14	56068520	56068521	rs10083303	T	T	C/T	intron,ncRNA,untranslated-5
1012	chr14	56078738	56078739	rs1138345	T	T	G/T	ncRNA,untranslated-5
1012	chr14	56079038	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



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

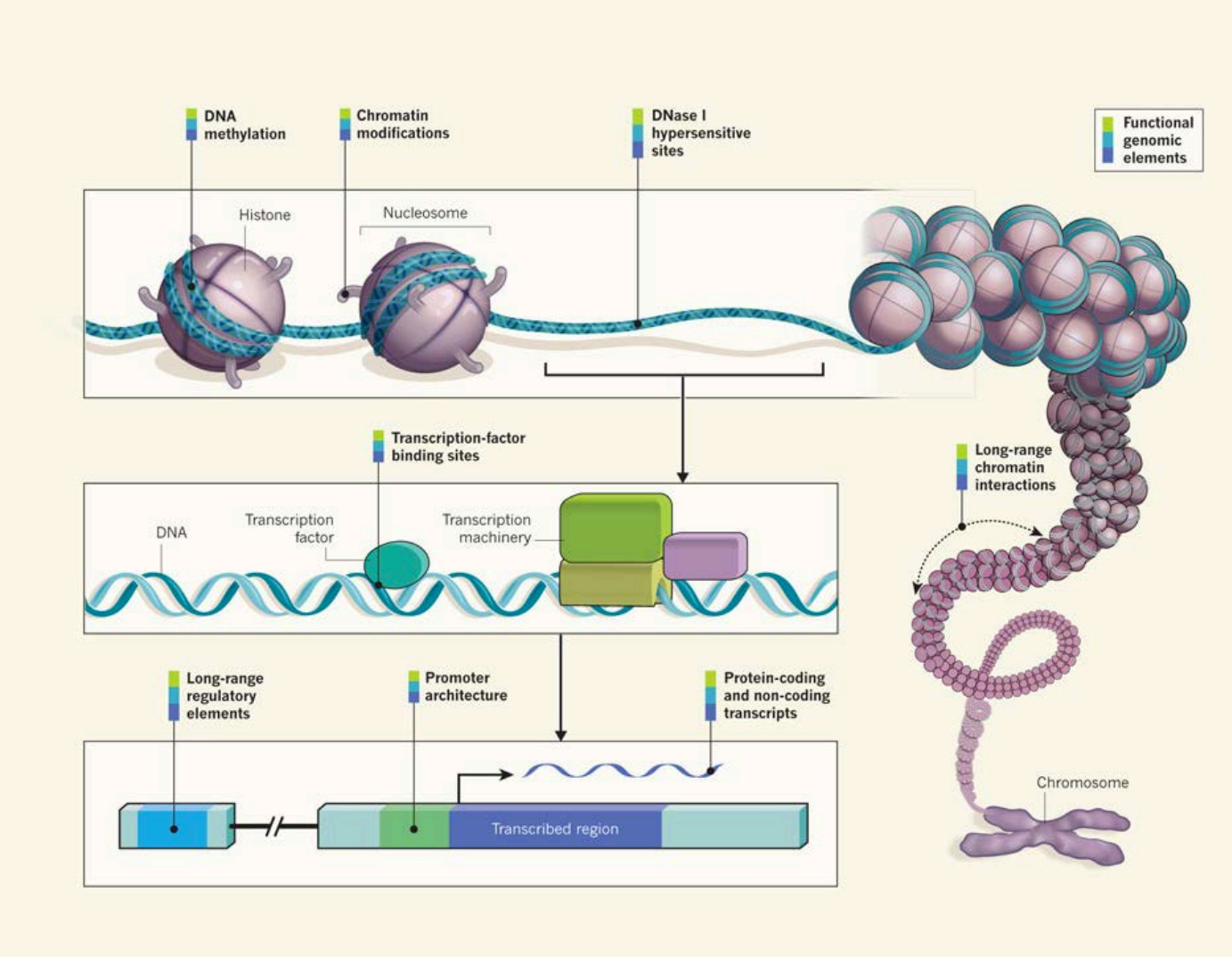
Finding Variants of Known Functionality

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.

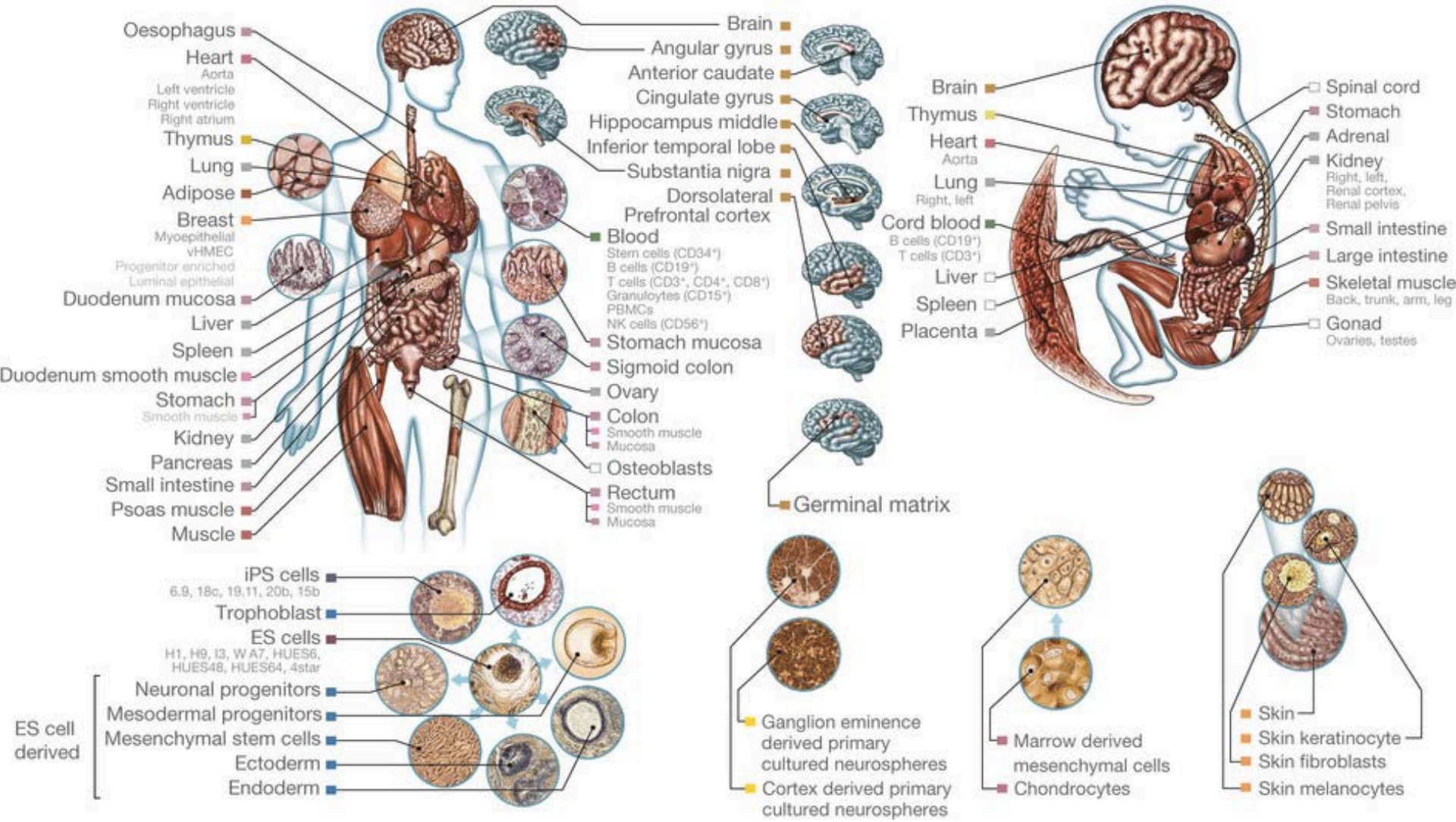
Our hit cannot be explained by known functional variants

Epigenetics



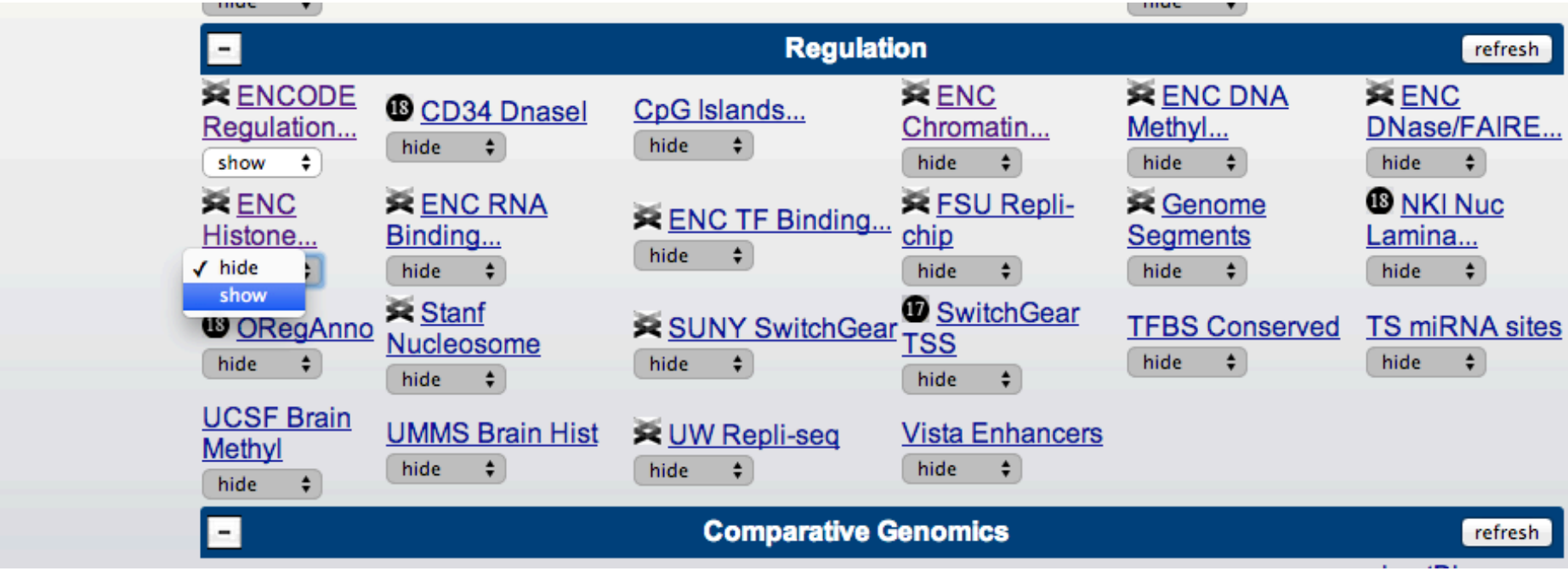
(Ecker et al., 2012)

Epigenetics Roadmap



(Roadmap Epigenetics Consortium et al., 2015)

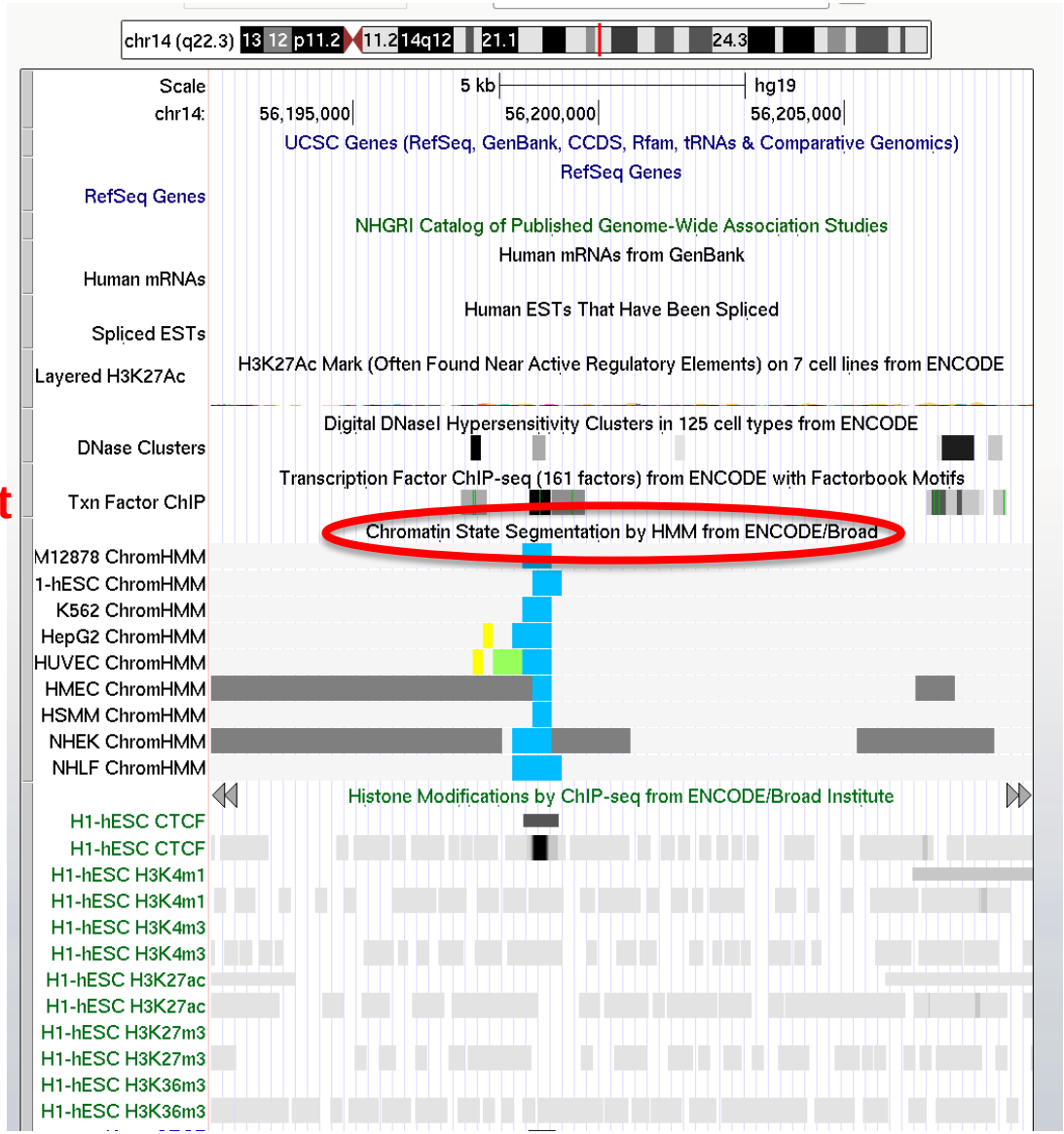
ENCODE in UCSC Genome Browser



Add some tracks that may help us explore function

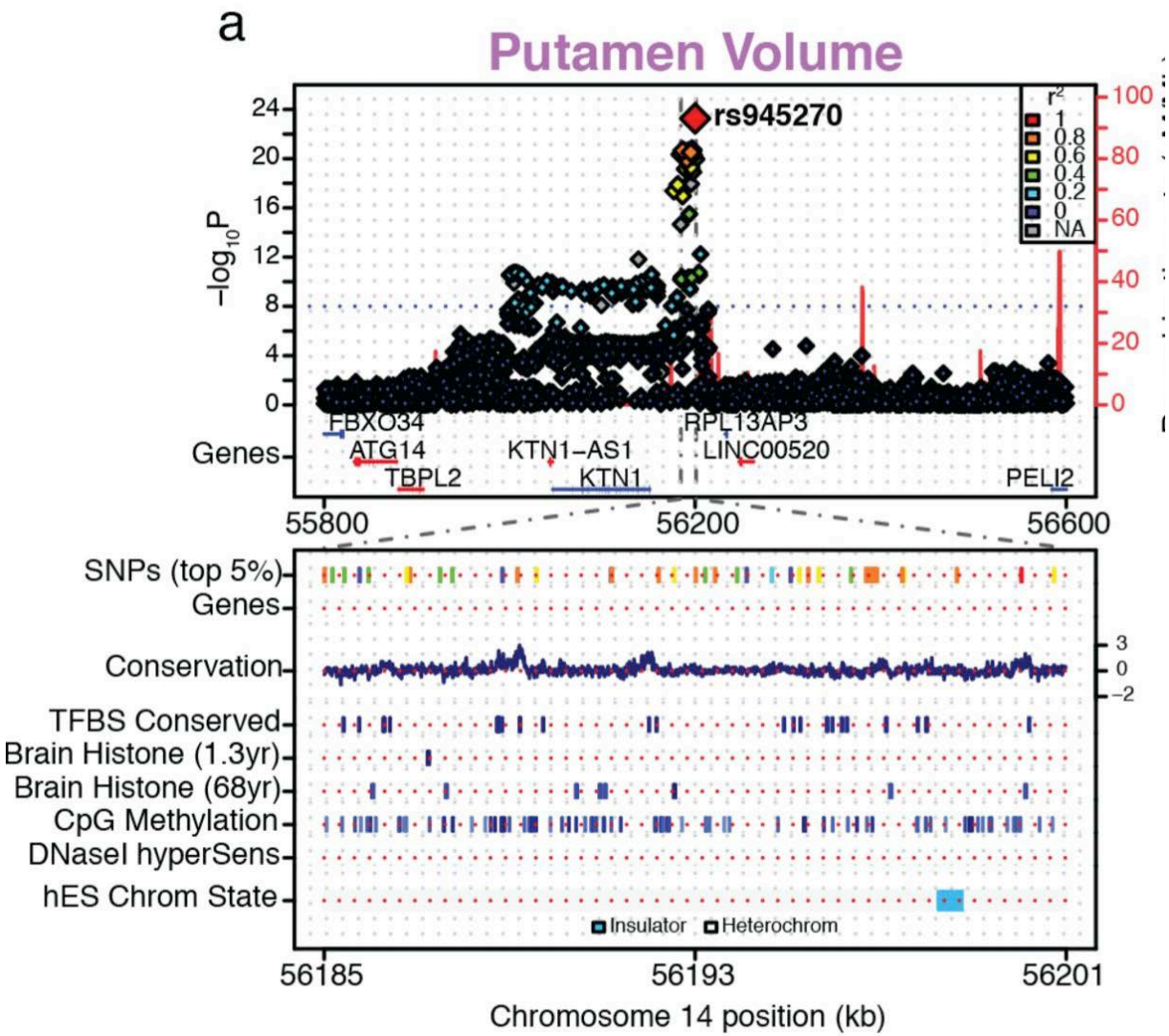
Epigenetics & ENCODE

Click to see what the colors mean



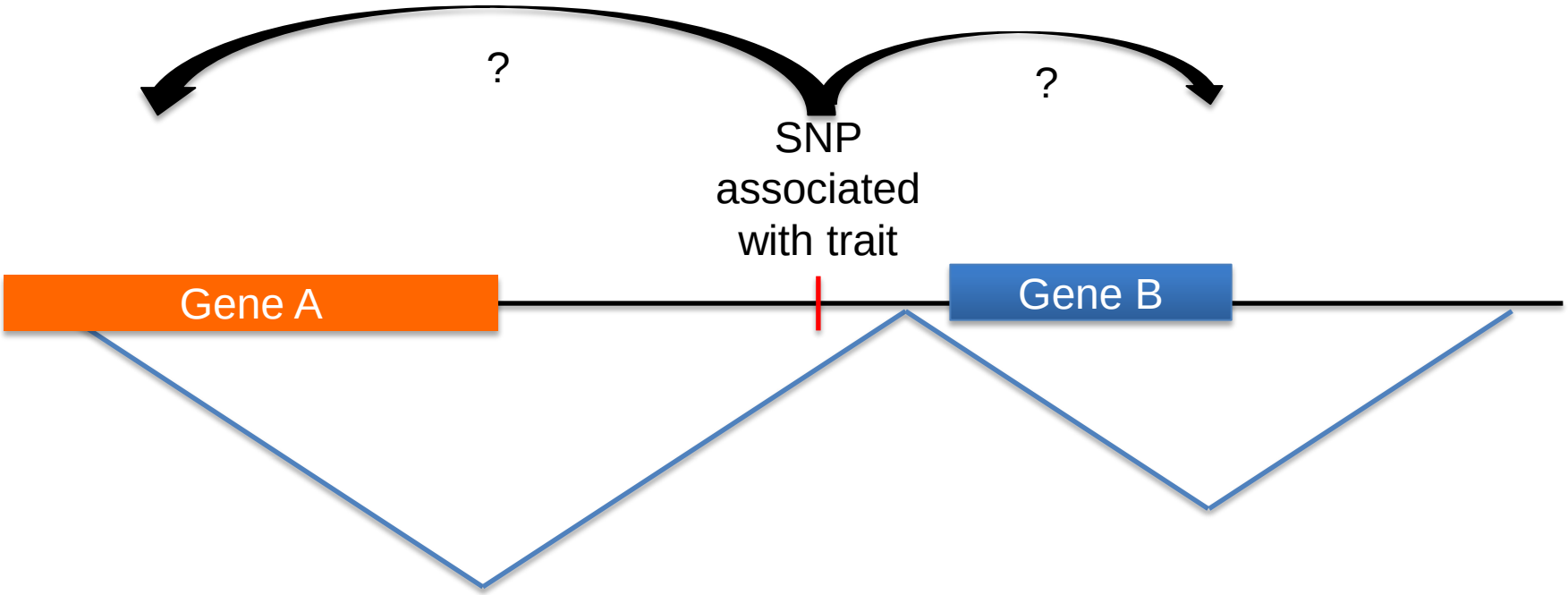
Appears to be a CTCF binding site (insulator) very close to the locus!

LocusTrack: Other ways to visualize



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

expression QTLs (eQTLs)

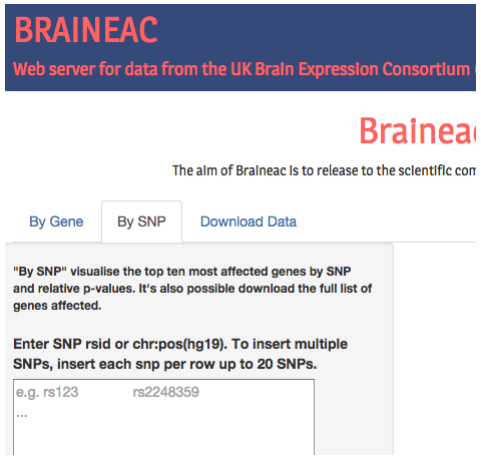


Do the alleles at the risk SNP affect the expression of any nearby genes?

A way to move from variant to gene.

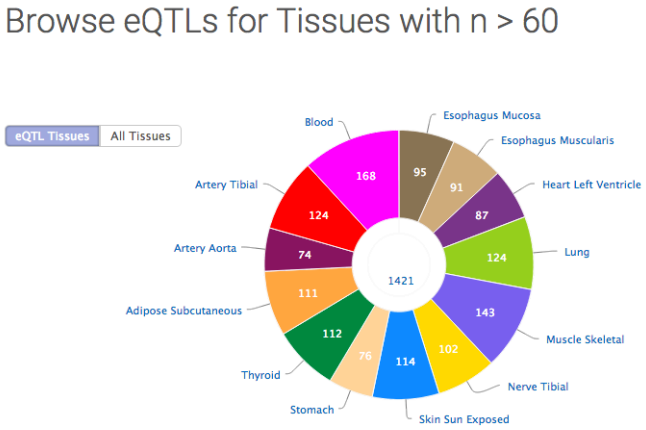
eQTL databases

BRAINEAC



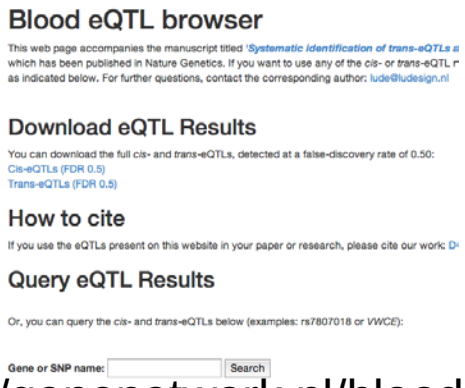
http://www.braineac.org/

GTEx



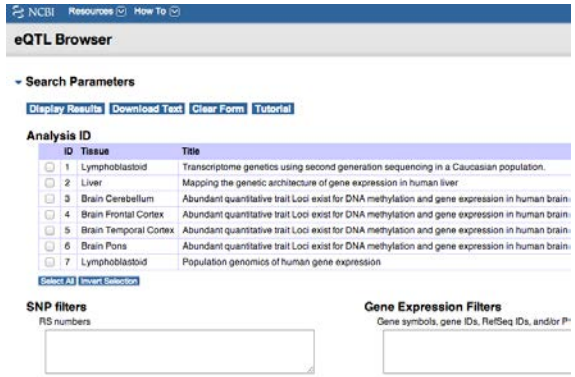
http://www.gtexportal.org/home/
(Brain on its way)

Blood eQTL browser



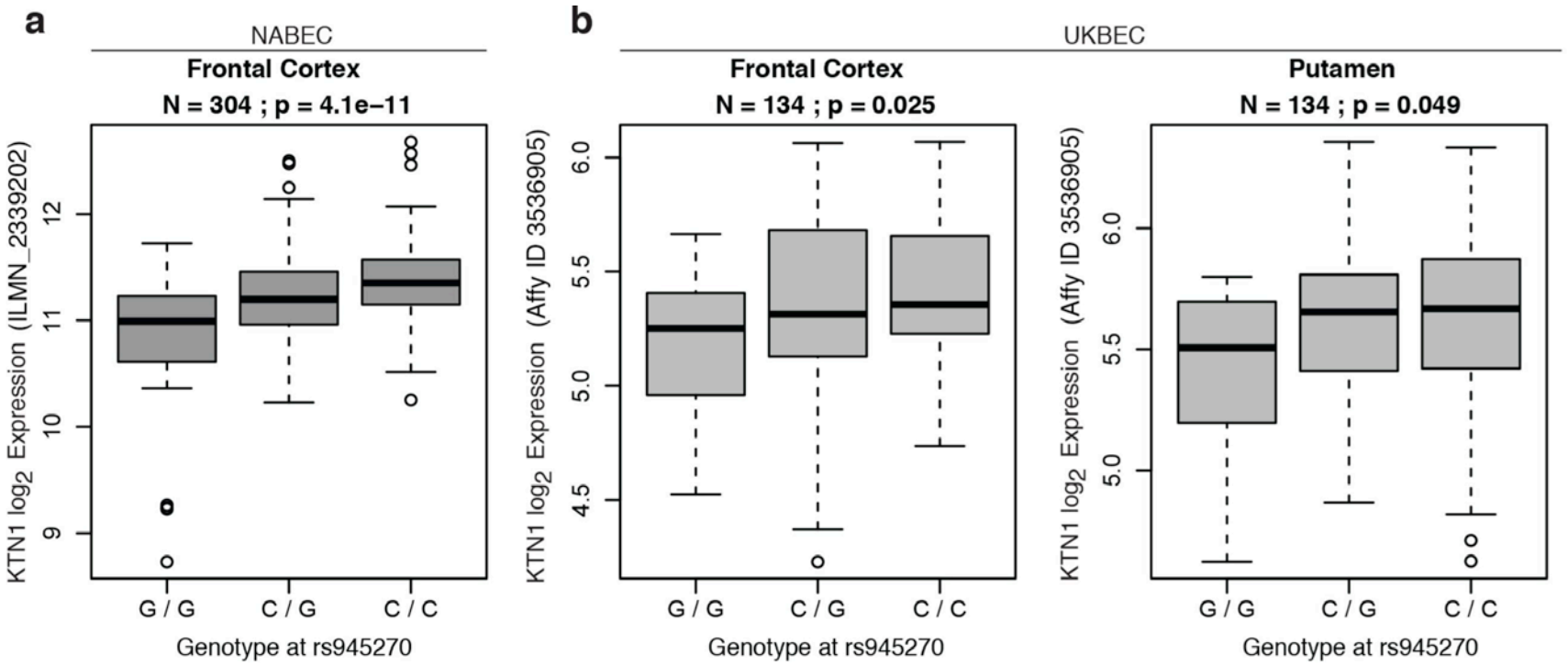
http://genenetwork.nl/bloodeqtlbrowser/

NCBI eQTL Browser



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!

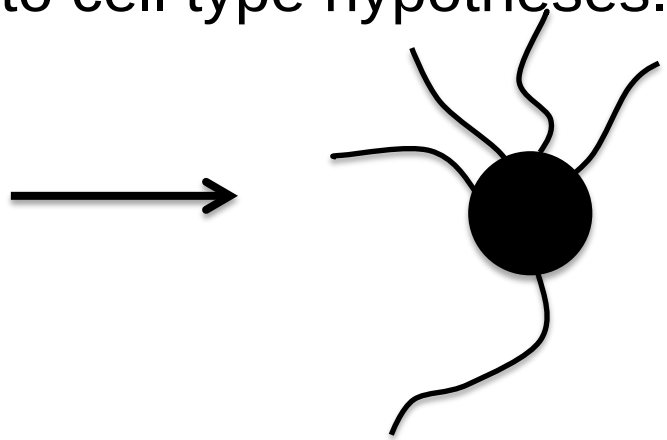
When and Where is Gene Expressed?

- 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).

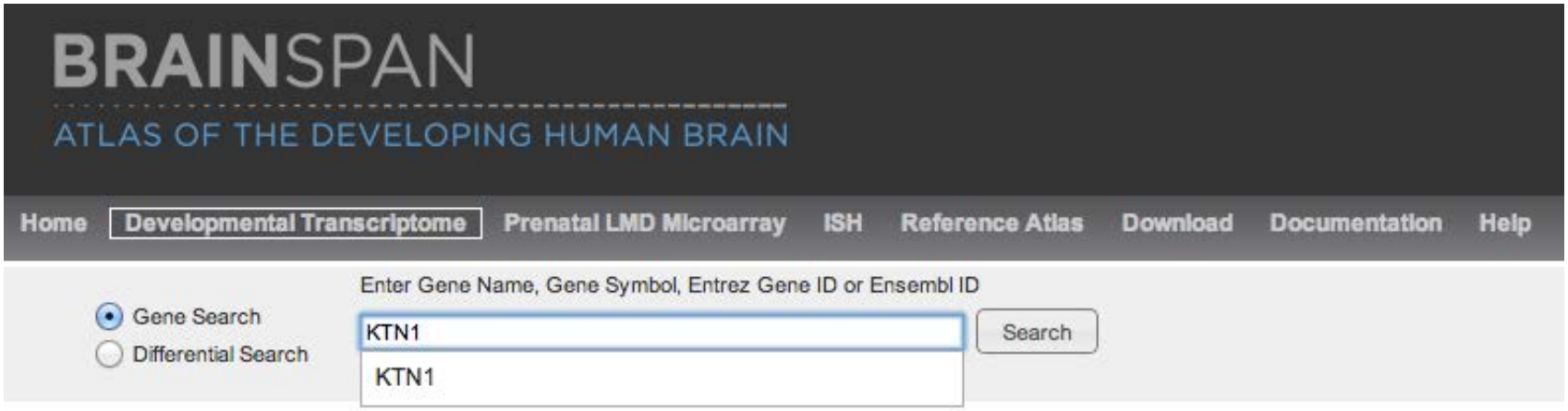
- Expression of a gene in brain does little to implicate it as causal.

- However, finding the time period or region of gene expression may lead us to cell type hypotheses.

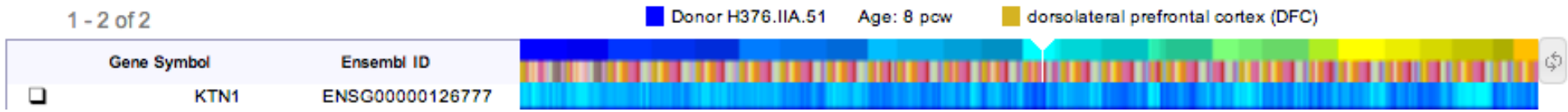
Gene A



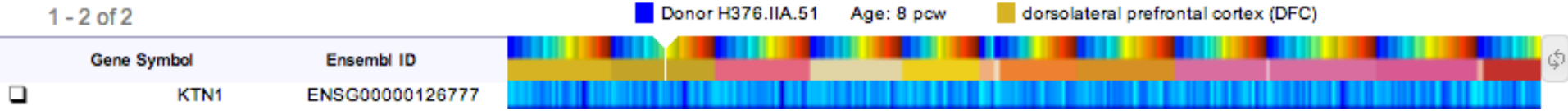
When and Where is Gene Expressed?



Expression by time →

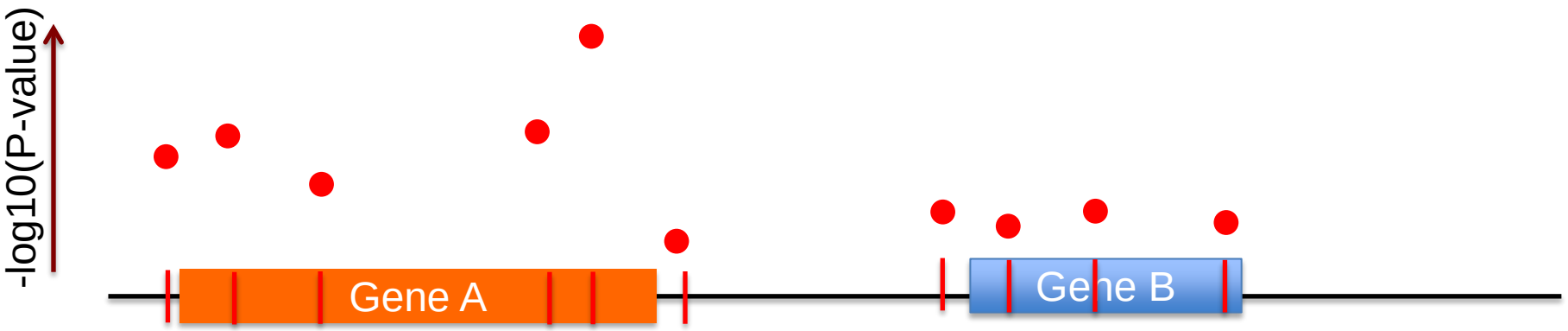


Expression by region →



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

Gene-based tests



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

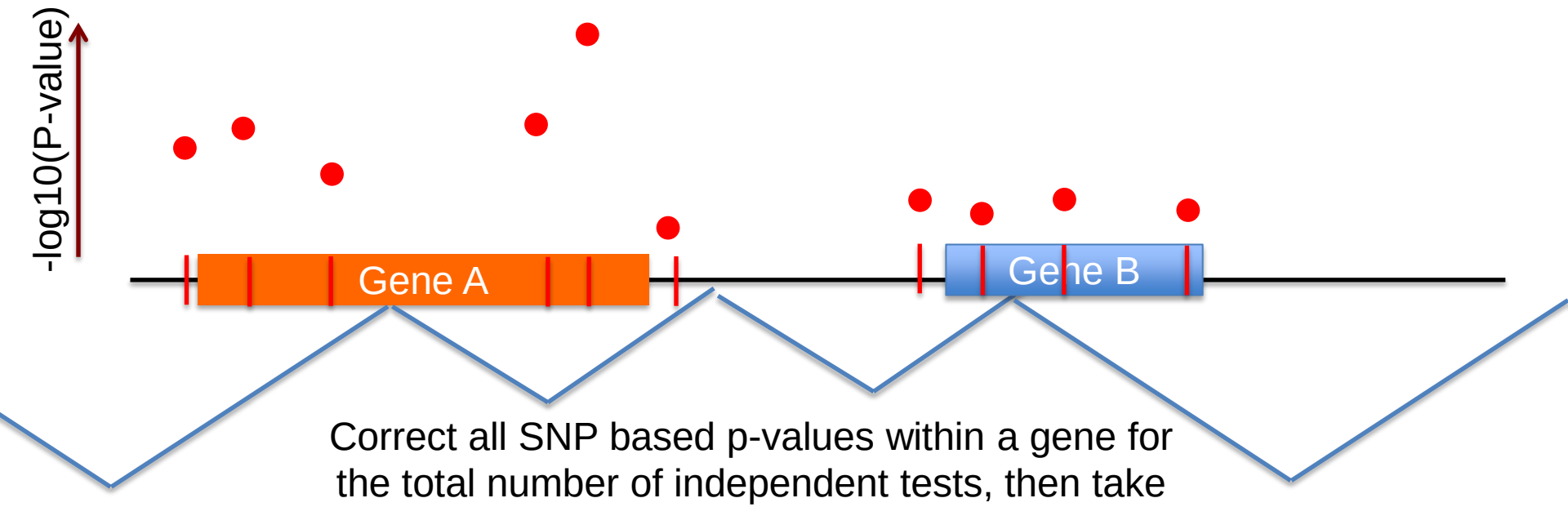
Advantages

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

Disadvantages

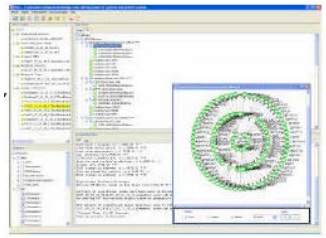
- Ignores intergenic variation (most of genome)
- Generally ignores direction of association
- Ignores that a variant within the intron of one gene may be affecting a totally different gene

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.

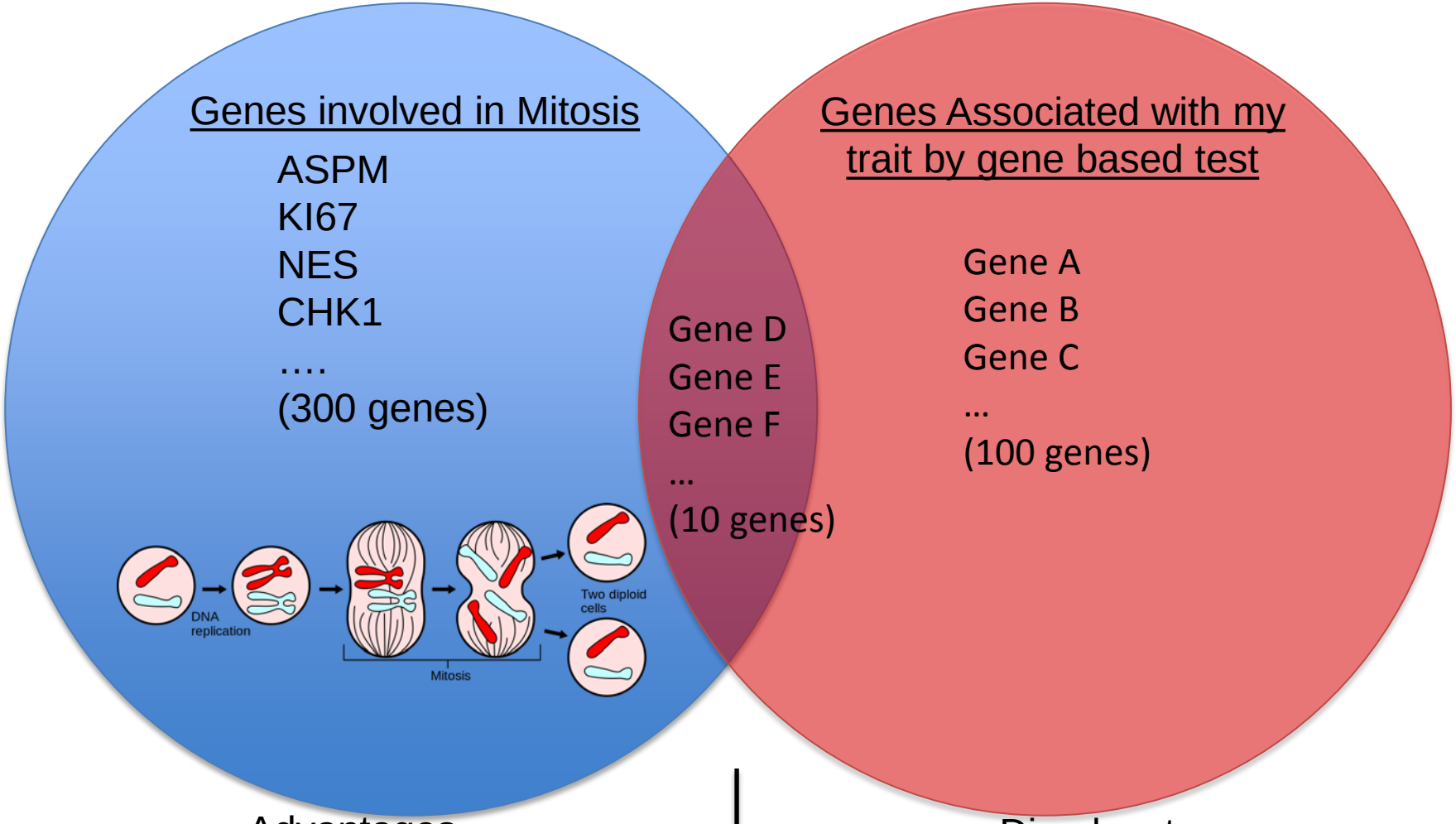
See GATES algorithm implemented in KGG toolbox (Li et al., 2011)



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

Pathway Analysis

Look for enrichment of your associated genes in known pathways



Advantages

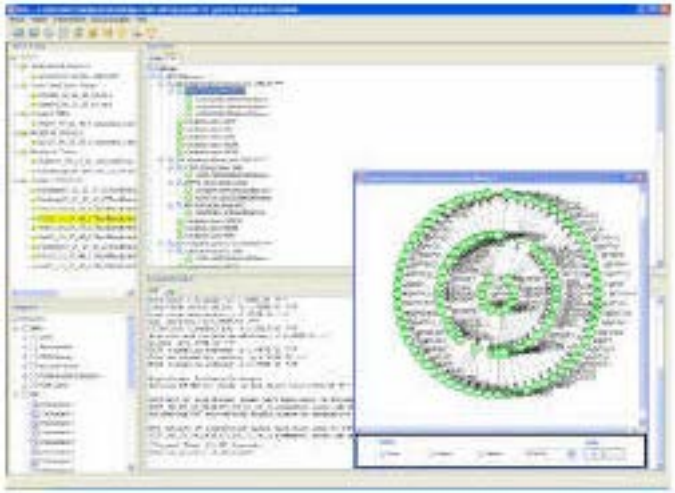
- Amazing interpretability – exactly what you're looking for

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