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...



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....

C APPLICATIONS OF NEXT-GENERATION SEQUENCING

Sequencing studies in human genetics: design and interpretation

David B. Goldstein¹, Andrew Allen^{1,2}, Jonathan Keebler¹, Elliott H. Margulies⁵, Steven Petrou^{4,5}, Slavé Petrovski^{1,6} and Shamil Sunyaev⁷

"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.

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 (r² > 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

000	🔗 Human (Ho	mo sapiens) Ge ×									
← → C	🗋 genome	e.ucsc.edu/cgi-bin/h	gGateway								\$
^	Genomes	Genome Browser	Tools	Mirrors	Downloads	My Data	Help	About Us	3		
Human (H	lomo sap	<i>iens</i>) Genome Br	rowser Ga	ateway							
		group	Th genome	Feb. 200 Click he	me Browser was c Copyright (c) The R assembly 19 (GRCh37/hg19) re to reset the b search add custo	reated by the <u>Ge</u> tegents of the Ur thr19:45, rowser user i m tracks trac	nome Bioinf niversity of C position 404,181-4 nterface se	ormatics Gro california. All 5,404,681 ettings to th nfigure tracks	up of UC Santa Cruz. rights reserved. search term enter position, gene symbol or search terms neir defaults.	submit	

Human Genome Browser - hg19 assembly (sequences)

The February 2009 human reference sequence (GRCh37) was produced by the <u>Genome Reference Consortium</u>. For more information about this assembly, see <u>GRCh37</u> 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 <u>User's Guide</u> for more information.

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



Homo sapiens (Graphic courtesy of <u>CBSE</u>)

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





The variant initially entered stays in the center

Add a Track

-		Phenotype and	Literature		refresh
Publications dense 🗘	ClinVar Variants	Coriell CNVs hide \$	COSMIC hide \$	DECIPHER hide \$	GAD View hide \$
GeneReviews	GWAS Catalog ✓ hide dense	HGMD Variants	ISCA hide \$	LOVD Variants	MGI Mouse QTL hide \$
OMIMAV SNPs hide \$	squish pack <u>NES</u> full	OMIM Pheno Loci	B <u>RGD Human</u> QTL hide \$	RGD Rat QTL hide	UniProt Variants hide \$
Web Sequences hide +					





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

Uploading a user track



LocusZoom: Making Prettier Pictures



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?

â (Genomes	Genome Browser	Tools	Mirrors	Downl	oads	My Data	Help	About Us	
Human (H	lomo sap	iens) Genome Bro	Blat							
		6	Table E	Browser		r was cre	eated by the Ge	enome Bioinf	ormatics Grou	ip of UC Santa Cruz.
			Variant	Annotation In	tegrator) The Re	egents of the Ur	niversity of C	California. All ri	ghts reserved.
		group g	Gene S	orter		y		position		search term
		Mammal ‡ Huma	Genom	e Graphs		(hg19) ‡	chr14:56,	125,323-5	6,275,622	enter position, gene symbol or search terms submit
			In-Silic	o PCR						
			LiftOve	r		the br	rowser user i	nterface se	ettings to the	eir defaults.
			VisiGe	ne		dd custor	m tracks trac	k hubs co	nfigure tracks a	and display
	L		Other L	Jtilities						

Go to the Table Browser in UCSC Genome Browser

clade: Mammal
group: Variation
table: snp138Common
region: _ genome _ ENCODE Pilot regions • position chr14:56031386-56369561 lookup define regions
identifiers (names/accessions): paste list upload list
tilter: create Click to create a
intersection: create filter
correlation: create
output format: all fields from selected table
output file: chr14.txt (leave blank to keep output in browser)
file type returned: plain text gzip compressed
get output summary/statistics

Select interpretable functional Variants

func	does include
submit ance	Click to create a filter
	clade: Mammal
	group: Variation
	table: snp138Common
	region: Ogenome OENCODE Pilot regions Oposition Chr14:56,031,386-56,369, lookup define regions
	identifiers (names/accessions): paste list upload list
	filter: edit clear
	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 spreadsheet	get output simmary/statistics

#filter: (FIND	_IN_SET('splic	ce-5', snp1380	Common.func)>0 OR FIND_	IN_SET('splic	e-3', snp1380	Common.func)>0 OR FIND_IN_SET('untranslat
#bin	chrom	chromStart	chromEnd	name	refNCBI	refUCSC	observed	func
1012	chr14	56068483	56068484	rs116289145	С	С	C/T	intron, ncRNA, untranslated-5
1012	chr14	56068520	56068521	rs10083303	Т	Т	C/T	intron, ncRNA, untranslated-5
1012	chr14	56078738	56078739	rs1138345	Т	Т	G/T	ncRNA, untranslated-5
1012	chr14	56079038	56079039	rs34879854	Α	Α	A/T	coding-synon,ncRNA
1012	chr14	56094725	56094726	rs17128636	С	С	A/C	coding-synon,ncRNA
1012	chr14	56096685	56096686	rs74053638	Α	Α	A/C	coding-synon,ncRNA
1012	chr14	56096730	56096731	rs2274075	Α	Α	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?

Query SNPs						
Input SNPs:	Text Entry	÷ <u>E</u>	<u>xample</u>			
One snp per line:	rs945270					
E Search Ontions	L					
Search Options						
SNP data set:	1000 Genomes Pilot 1	÷	r ² thre	eshold:	0.8	\$
Population panel:	CEU ‡		Distanc	e limit:	500	\$
Output Options						
Download to:	File \$	✓ Inc □ Su	lude each query ppress warning i	snp as message	a proxy for i es in output	tself

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

SNP	Proxy	Distance	RSquared	DPrime	Arrays	Chromosome	Coordinate_HG18	
rs945270	rs945270	0	1	1	None	chr14	55270226	
rs945270	rs8017172	1425	1	1	12,15,16,16	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	13,15,16,16	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	A6,0Q	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	12,15,16,16	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

Click to see what the colors mean



Epigenetics & ENCODE

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 Brainea The aim of Braineac is to release to the scientific com By SNP

Download Data

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

Enter SNP rsid or chr:pos(hq19). To insert multiple SNPs, insert each snp per row up to 20 SNPs rs2248359

http://www.braineac.org/

Blood eQTL browser

Blood eQTL browser

This web page accompanies the manuscript titled 'Systematic identification of trans-eQTLs a which has been published in Nature Genetics. If you want to use any of the cis- or trans-eQTL r 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

By Gene

e.a. rs123

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: http://genenetwork.nl/bloodeqtlb rowser/

GTEx

Browse eQTLs for Tissues with n > 60



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

Searc	h P	arameters	
Displa	y Re	eults Download Tex	t Clear Form Tutorial
Analy	ele	ID	
Analy	ID	Tissue	Title
0	1	Lymphoblastoid	Transcriptome genetics using second generation sequencing in a Caucasian population.
	2	Liver	Mapping the genetic architecture of gene expression in human liver
	3	Brain Cerebellum	Abundant quantitative trait Loci exist for DNA methylation and gene expression in human brain
	4	Brain Frontal Cortex	Abundant quantitative trait Loci exist for DNA methylation and gene expression in human brain
	5	Brain Temporal Cortex	Abundant quantitative trait Loci exist for DNA methylation and gene expression in human brain
0	6	Brain Pons	Abundant quantitative trait Loci exist for DNA methylation and gene expression in human brain
	7	Lymphoblastoid	Population genomics of human gene expression
Selec	e Al	Invert Selection	
SNP f	ilter	r6 ers	Gene Expression Filters Gene symbols, gene IDs, RetSeq IDs, and/or P

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?



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.

Gene A

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



Gene B

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

Pathway Analysis

Look for enrichment of your associated genes in known pathways



Tools for Pathway Based Analyses

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



http://statgenpro.psychiatry.hku.hk/lim x/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