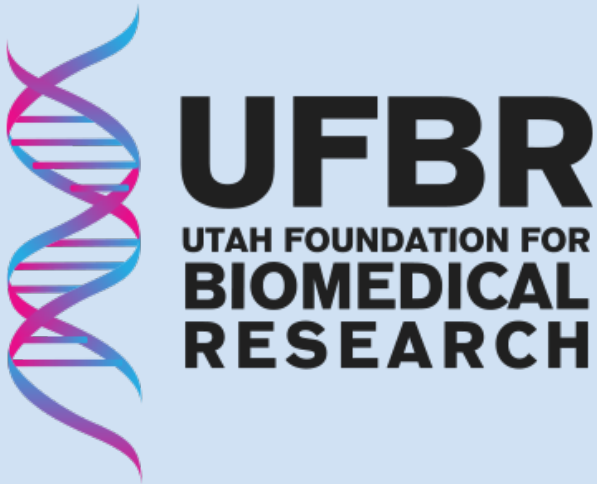
Toward more accurate variant calling for personal genomes

Gholson Lyon, M.D. Ph.D.



Acknowledgments

Omicia

Martin Reese
Edward Kiruluta

UFBR
UTAH FOUNDATION FOR BIOMEDICAL RESEARCH

Reid Robison

VBI VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

David Mittelman

THE UNIVERSITY OF UTAH

Barry Moore
Alan Rope
Jeffrey J Swensen
Lynn Jorde
Mark Yandell

CSH STANLEY INSTITUTE FOR COGNITIVE GENOMICS
COLD SPRING HARBOR LABORATORY

Jason O'Rawe
Yiyang Wu
Michael Schatz
Giuseppe Narzisi

USC TROJANS

Kai Wang

illumina®

Tina Hambuch
Erica Davis
Dawn Barry

our study families

O'Rawe *et al. Genome Medicine* 2013, **5**:28
http://genomemedicine.com/content/5/3/28

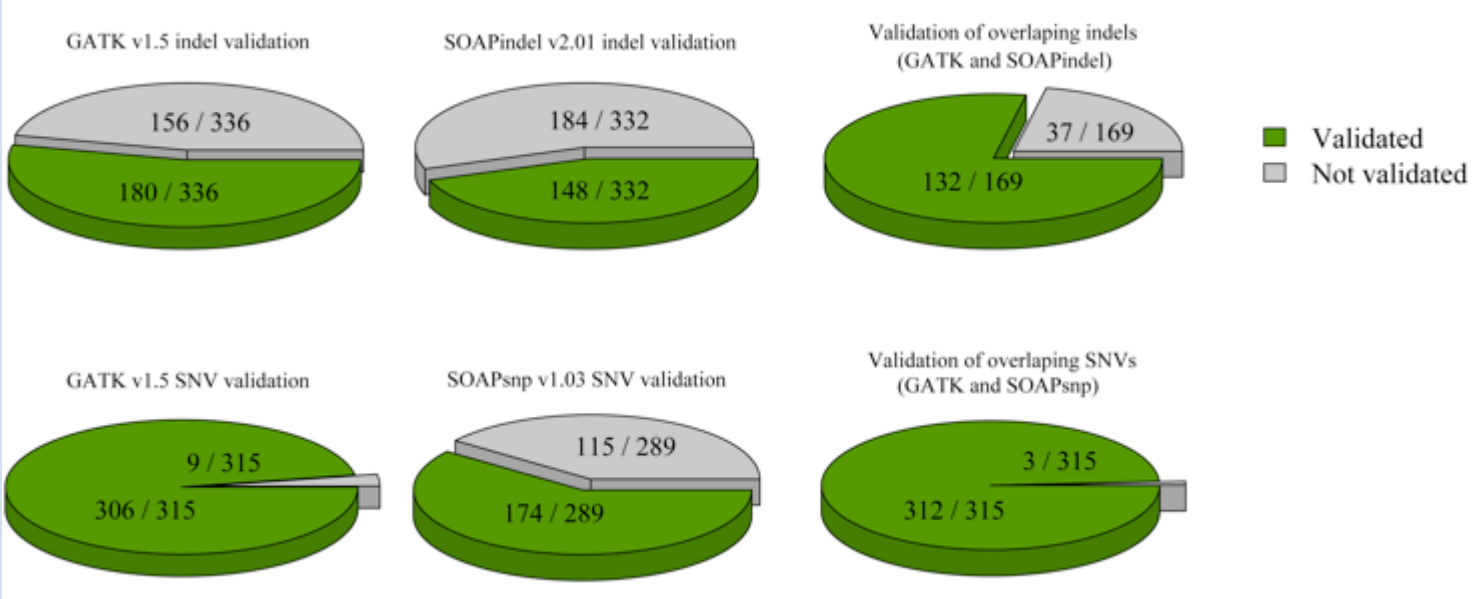
Genome Medicine

RESEARCH Open Access

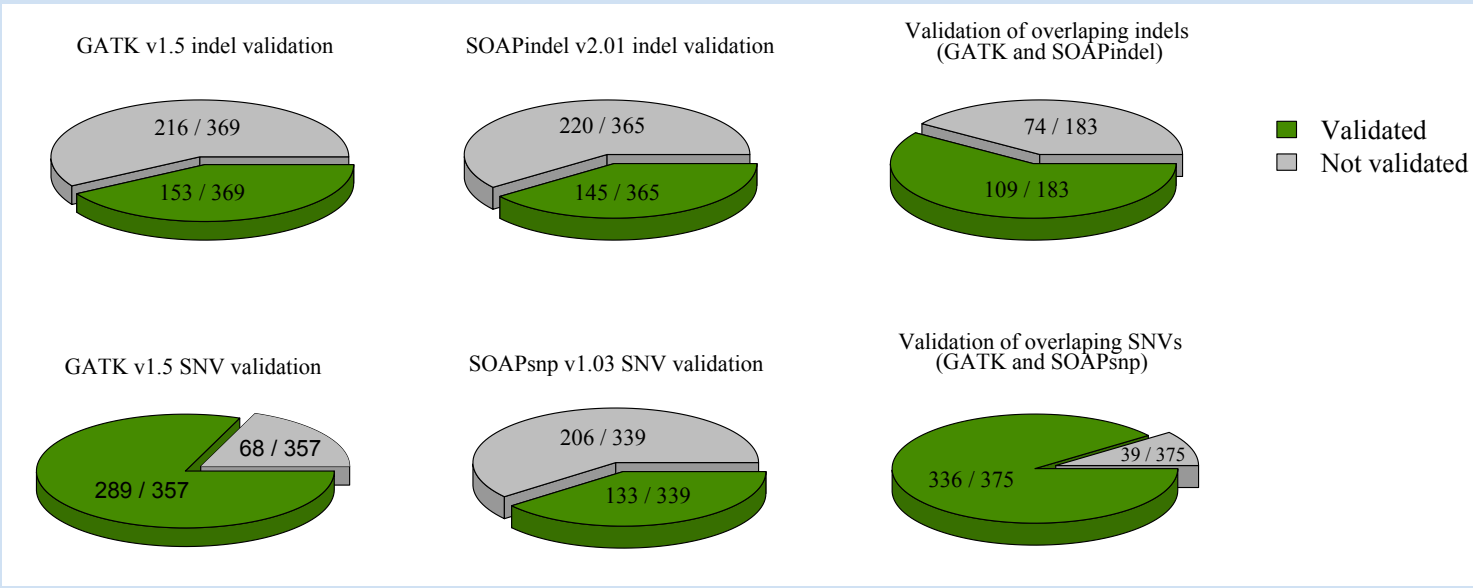
Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing

Jason O'Rawe[1,2], Tao Jiang[3], Guangqing Sun[3], Yiyang Wu[1,2], Wei Wang[4], Jingchu Hu[3], Paul Bodily[5], Lifeng Tian[6], Hakon Hakonarson[6], W Evan Johnson[7], Zhi Wei[4], Kai Wang[8,9*] and Gholson J Lyon[1,2,9*]

MiSeq



PacBio



- “If you want something hard, try clinical medicine” - Bruce Korf

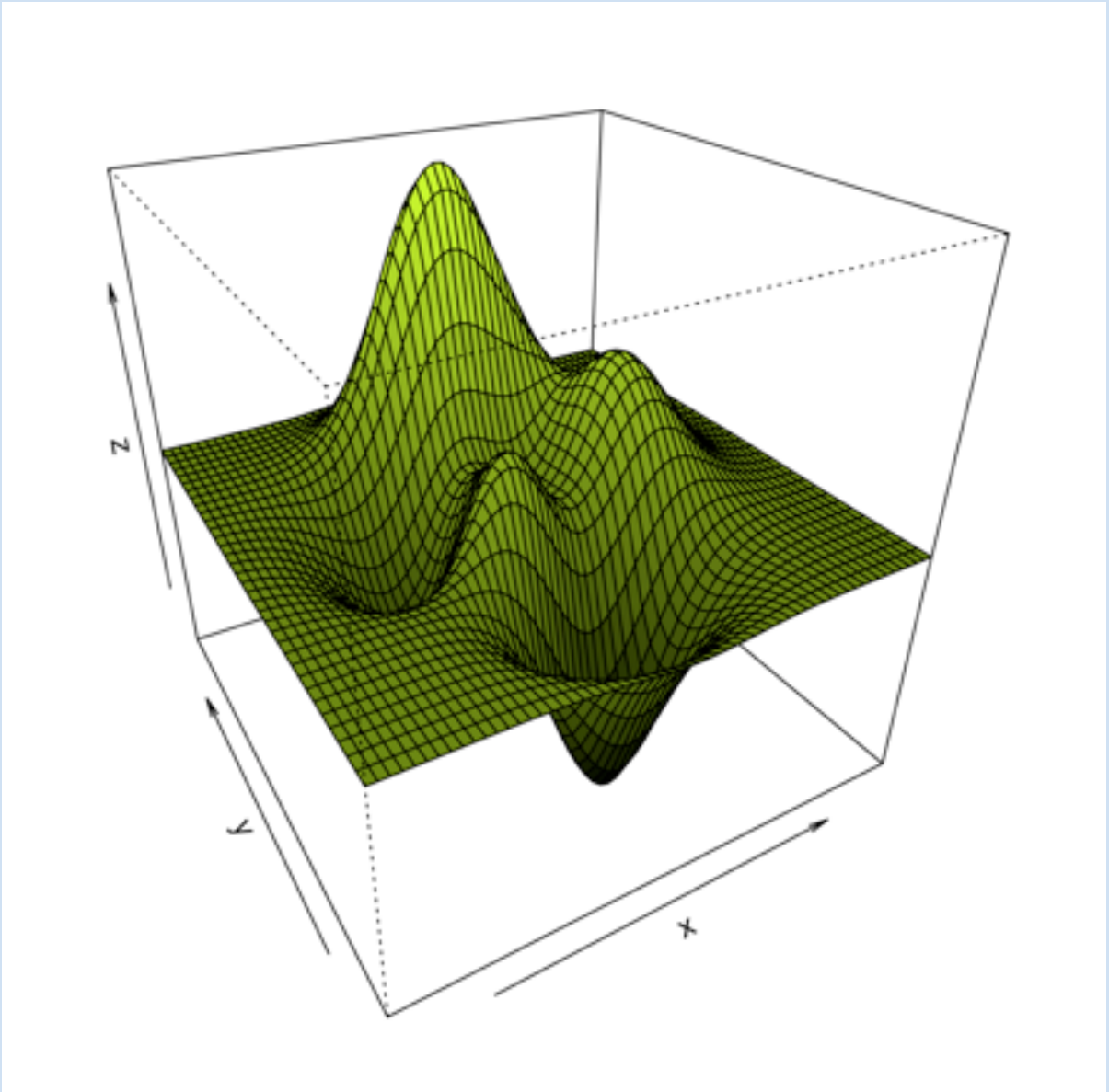
yet

- “proprietary databases” = “walled gardens”

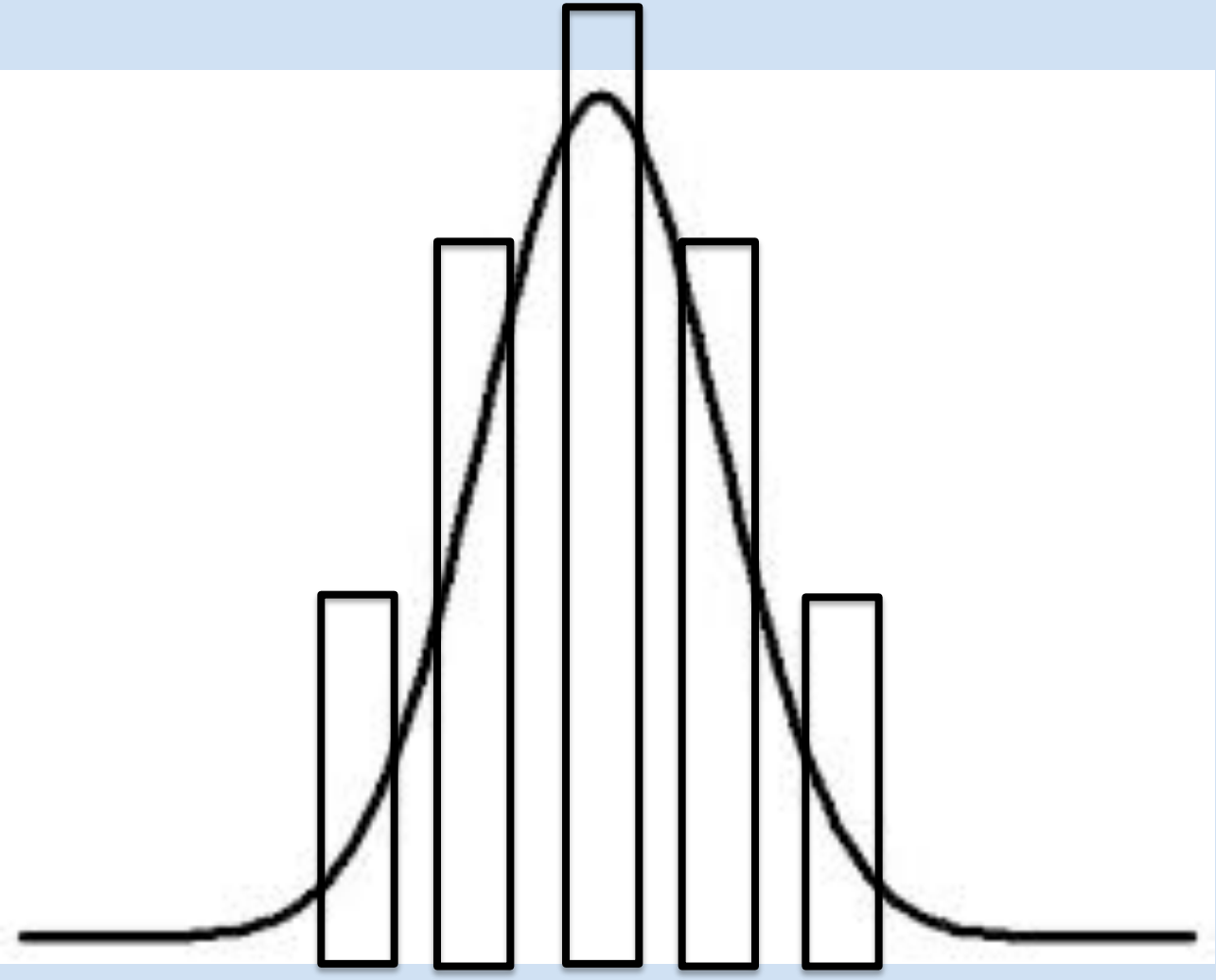
Biometry

Mutation ≠ Phenotype

- Ancestry matters!
- Context matters!
- Other mutations matter!
- Environment matters!
- Longitudinal course matters!


z
y
x

Categorical Thinking Misses Complexity



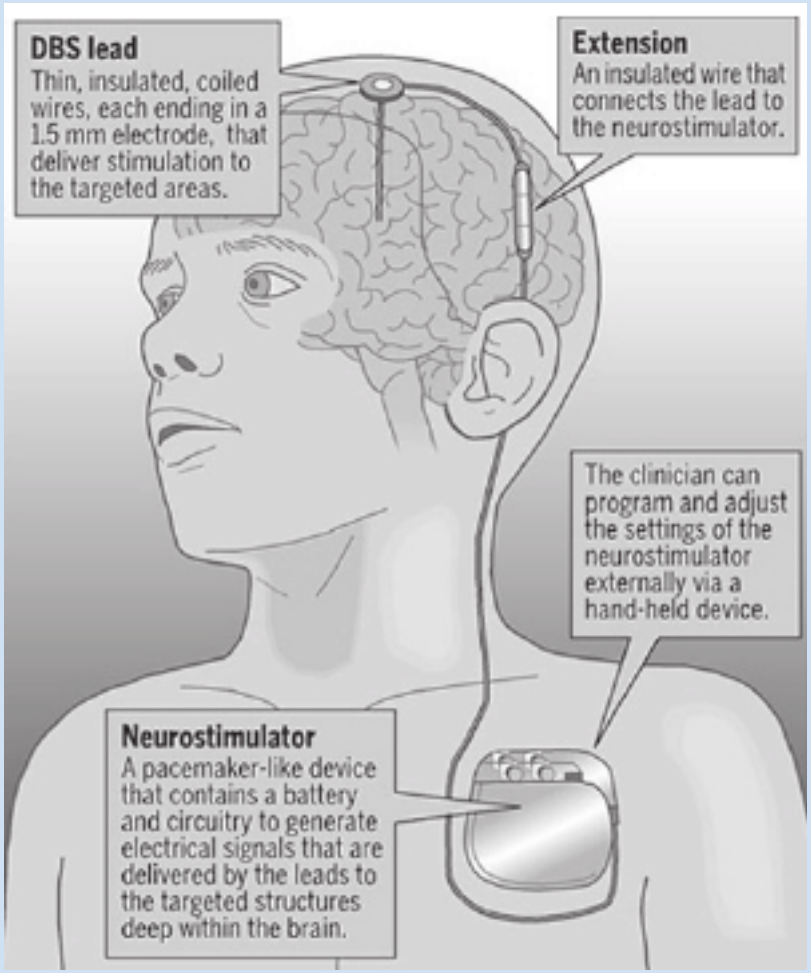
Expression Issues

- We do not really know the expression of pretty much ALL mutations in **humans**, as we have not systematically sequenced or karyotyped any genetic alteration in **Thousands to Millions** of **randomly** selected people, nor categorized into ethnic classes, i.e. clans.

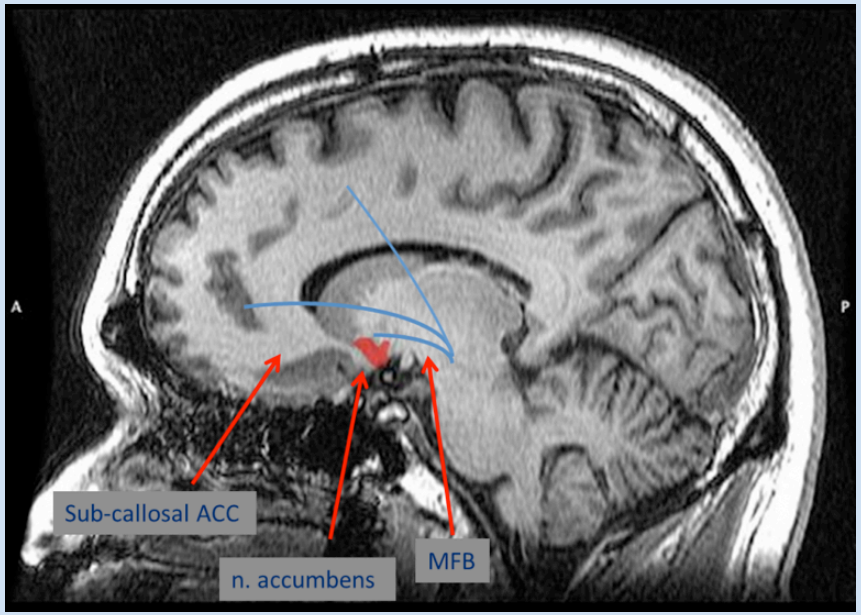
- There is a **MAJOR** clash of world-views, i.e. do single mutations drive outcome predominately, or are the results modified substantially by genetic background and/or environment? i.e. is there really such a thing as genetic determinism for **MANY** mutations?

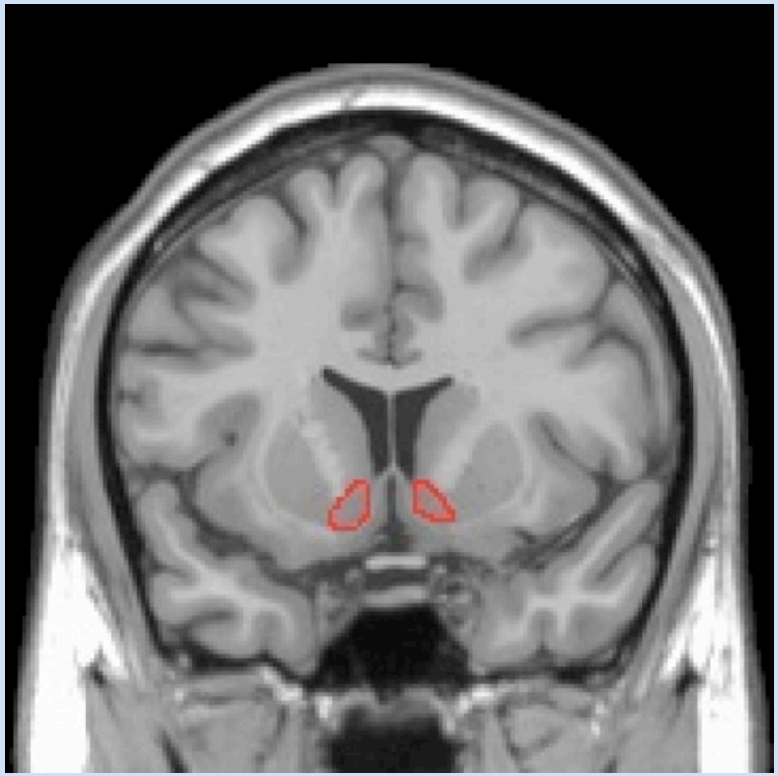
Vignette #1: One person with very severe obsessive compulsive disorder, severe depression and intermittent psychoses

40 year old Caucasian man from Utah


DBS lead
Thin, insulated, coiled wires, each ending in a 1.5 mm electrode, that deliver stimulation to the targeted areas.
Extension
An insulated wire that connects the lead to the neurostimulator.
The clinician can program and adjust the settings of the neurostimulator externally via a hand-held device.
Neurostimulator
A pacemaker-like device that contains a battery and circuitry to generate electrical signals that are delivered by the leads to the targeted structures deep within the brain.

Nucleus accumbens





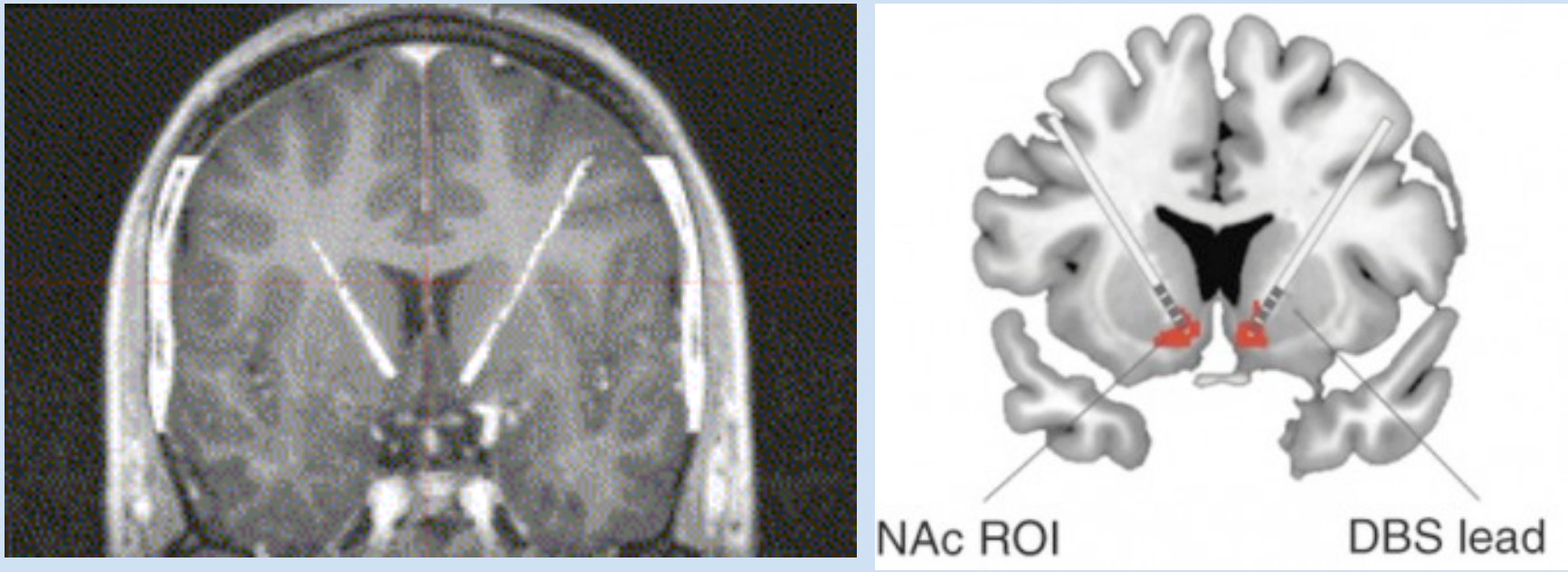
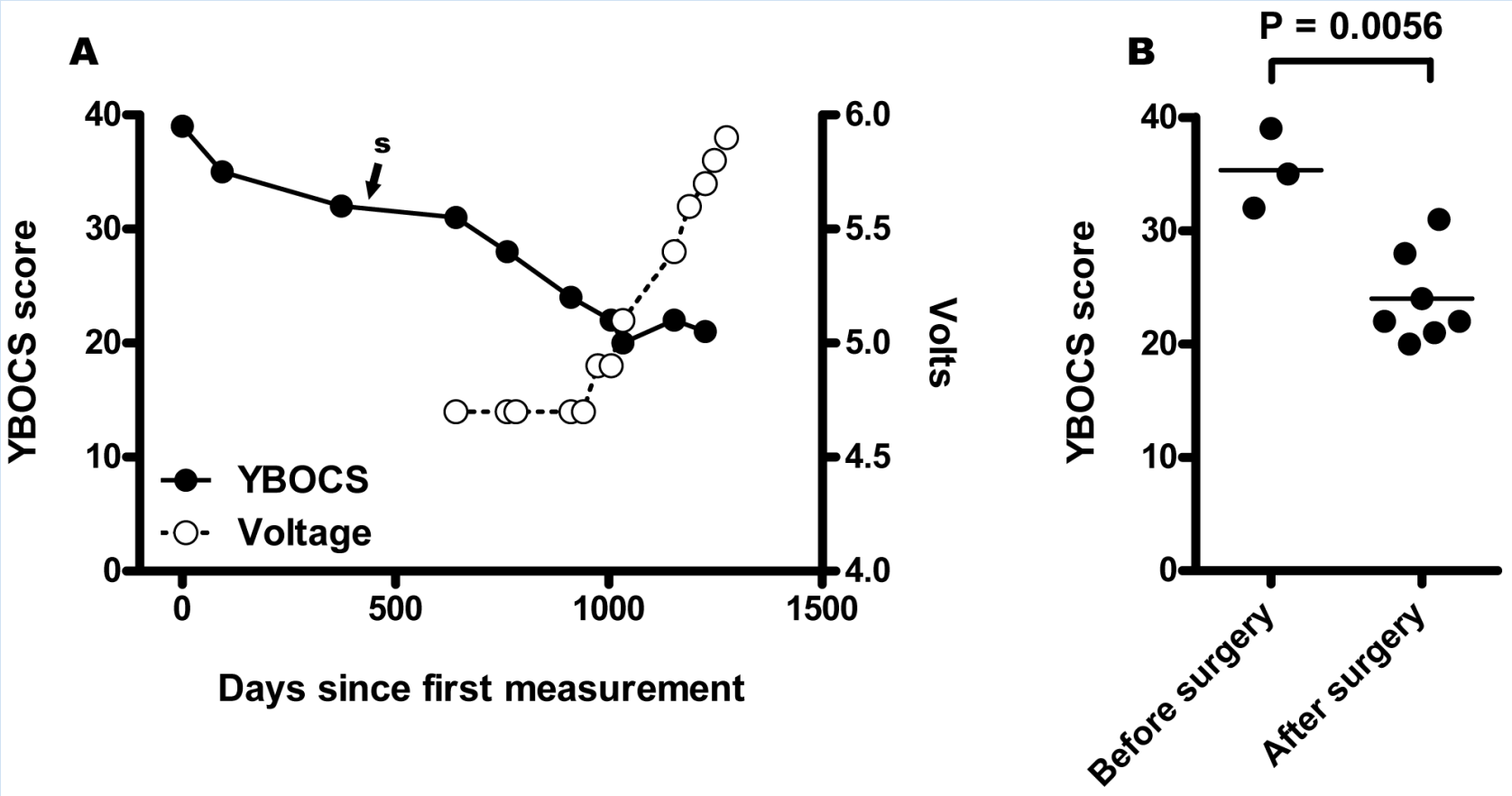


Fig. 1. Coronal section of the brain near the nucleus accumbens with the track of the electrodes on the left and right side.

Two year follow-up



Contents lists available at SciVerse ScienceDirect

Applied & Translational Genomics

journal homepage: www.elsevier.com/locate/atg

Practical, ethical and regulatory considerations for the evolving medical and research genomics landscape

Gholson J. Lyon [a,b,*], Jeremy P. Segal [c,**]

[a] Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, NY, United States
[b] Utah Foundation for Biomedical Research, Salt Lake City, UT, United States
[c] New York Genome Center, New York City, NY, United States

Table 1
Processes involved in a CLIA-certified genetic test.

Preanalytic system
1) Test request and specimen collection criteria
2) Specimen submission, handling and referral procedures
3) Preanalytic systems assessment
Analytic system
1) A detailed step-by-step procedure manual
2) Test systems, equipment, instruments, reagents, materials and supplies
3) Establishment and verification of performance specifications
4) Maintenance and function checks
5) Calibration and calibration verification procedures
6) Control procedures, test records, and corrective actions
7) Analytic systems assessment
Post-analytic system
1) Test report, including (among other things):
a) interpretation
b) reference ranges and normal values
2) Post-analytic systems assessment

1. Sample Collection and handling

2. Sequencing/Analytics

3. Interpretation

Individual Genome Sequencing Service

Available from Illumina's CLIA-certified laboratory.

"This laboratory test was developed, and its performance characteristics were determined by the Illumina Clinical Services Laboratory (CLIA-certified, CAP-accredited). Consistent with laboratory-developed tests, it has not been cleared or approved by the U.S. Food and Drug Administration. If you have any questions or concerns about what you might learn through your genome sequence information, you should contact your doctor or a genetic counselor. Please note that Illumina does not accept orders for Individual Genome Sequencing services from Florida and New York."

Sample Collection and Handling

The Sample Collection kit includes barcoded collection tubes, a Test Requisition form, an Informed Patient Consent form, and a pre-paid shipping envelope. All paperwork must be completed and returned for sample processing. Requests for Sample Collection kits must be submitted by a physician.

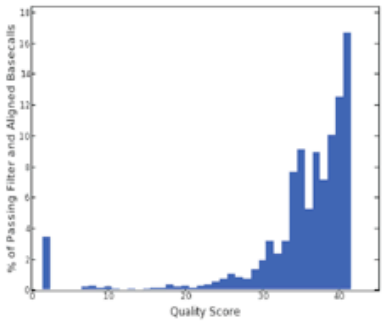
http://www.illumina.com/clinical/illumina_clinical_laboratory/igs_for_doctors/how_to_order.ilmn

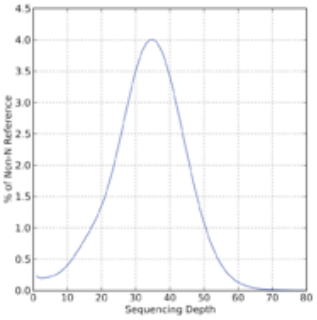
Sequencing and Analytics

Data Volume and Quality

	Yield (Gigabases)	% Bases ≥ Q30	% Bases Aligned
Passing Filter	113.10	87.10%	87.80%

	% Callable	% ≥ 5x depth	% ≥ 10x depth	% ≥ 20x depth	Mean depth(x)
Non-N Reference	93.28%	97.57%	96.22%	88.54%	33.35



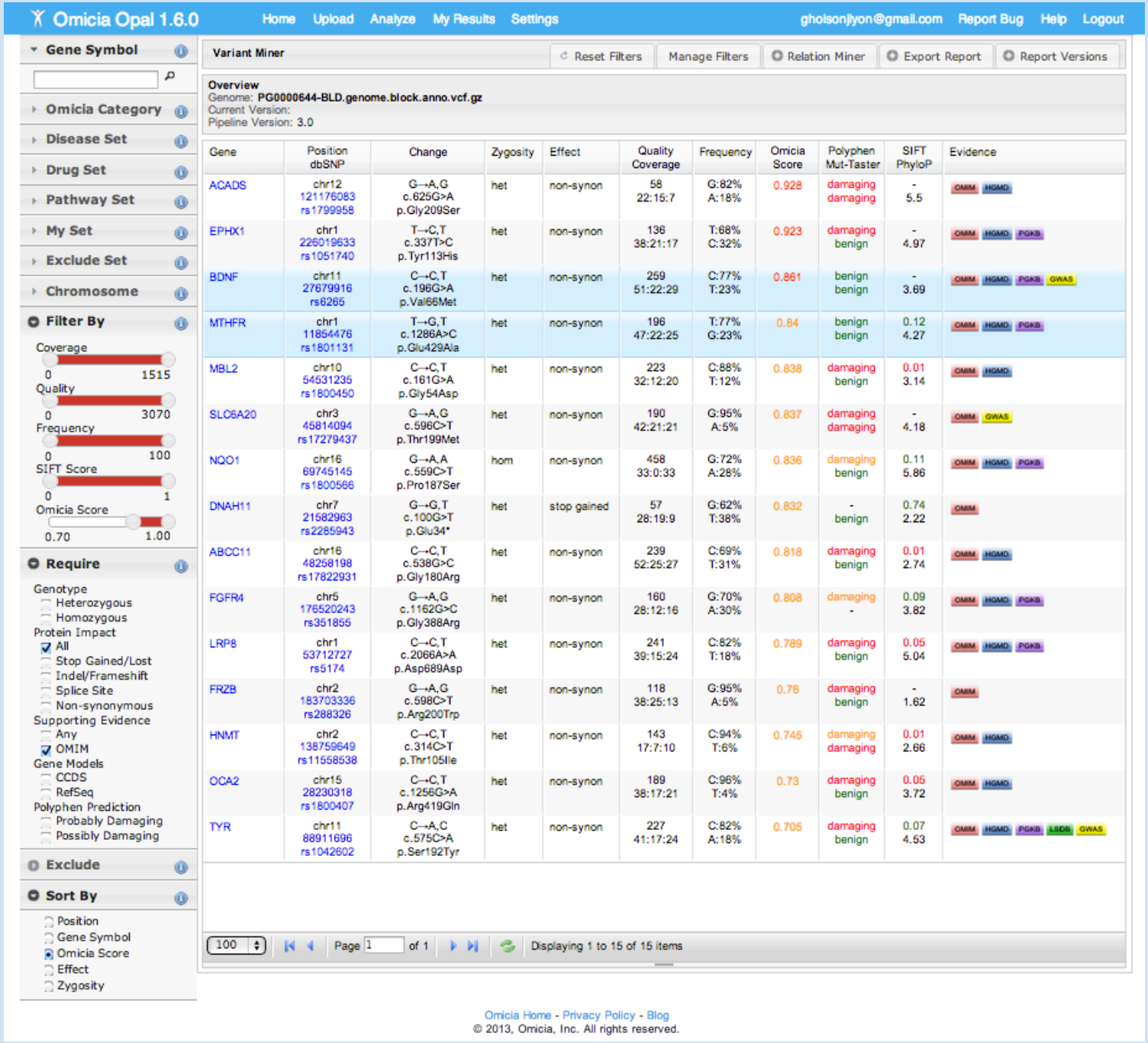


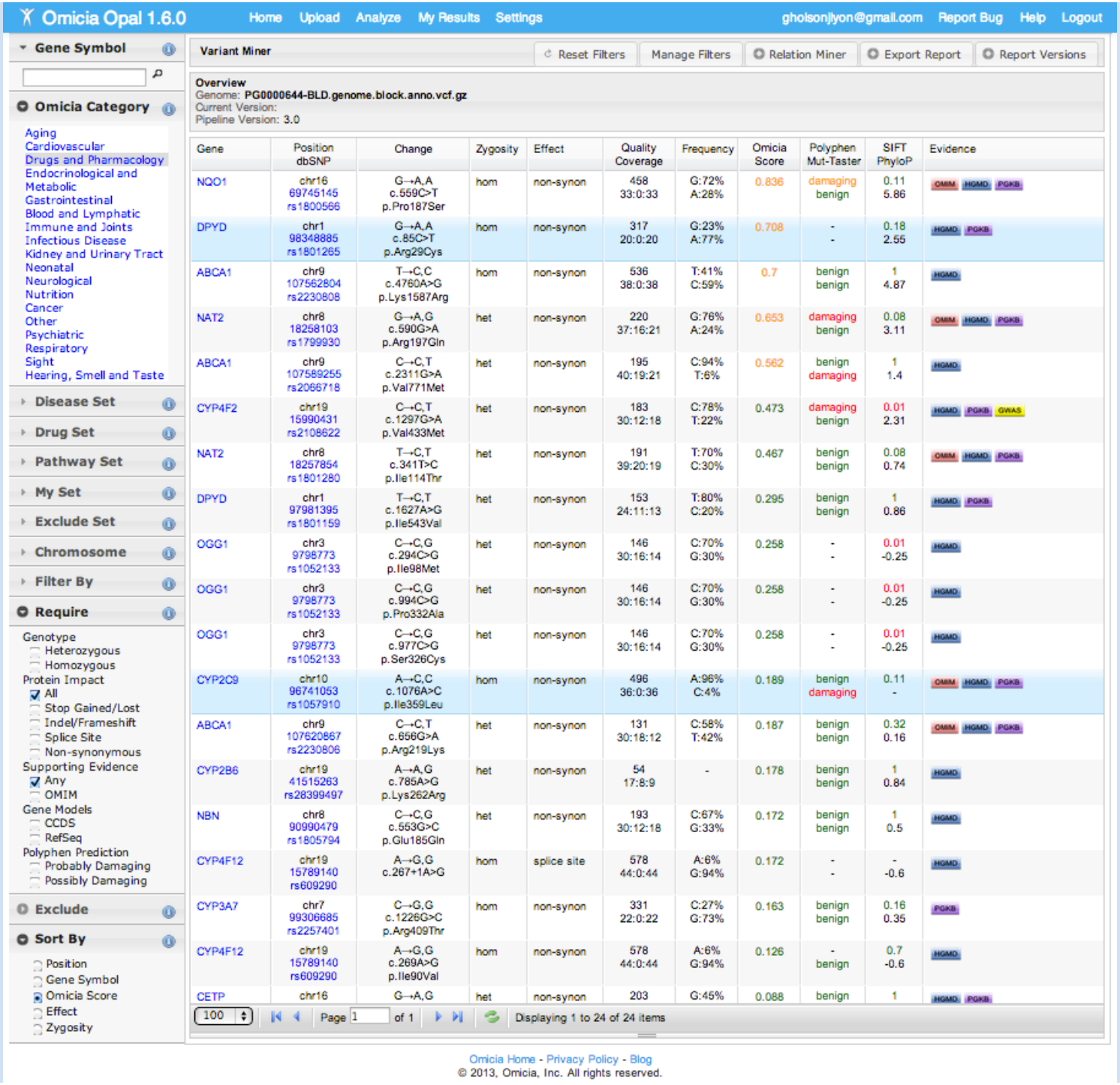
SNP Assessment

Total	Het/Hom	% in dbSNP	% in Genes	% in Coding
3,308,246	1.61	98.13%	45.47%	0.63%

Variant Statistics

	SNVs
Total Number	3,308,246
Number in Genes	1,504,121
Number in Coding Regions	20,879
Number in UTRs	24,946
Splice Site Region	2,917
Stop Gained	72
Stop Lost	16
Non-synonymous	9,884
Synonymous	10,907
Mature miRNA	36

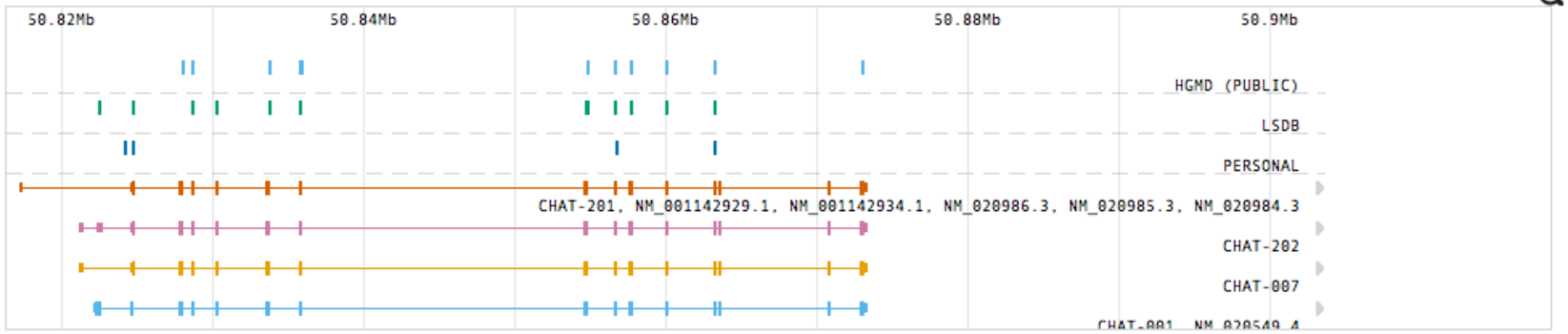

Omicia Opal 1.6.0
Home
Upload
Analyze
My Results
Settings
gholsonjlyon@gmail.com
Report Bug
Help
Logout
Gene Symbol
Omicia Category
Disease Set
Drug Set
Pathway Set
My Set
Exclude Set
Chromosome
Filter By
Coverage
0
1515
Quality
0
3070
Frequency
0
100
SIFT Score
0
1
Omicia Score
0.70
1.00
Require
Genotype
Heterozygous
Homozygous
Protein Impact
All
Stop Gained/Lost
Indel/Frameshift
Splice Site
Non-synonymous
Supporting Evidence
Any
OMIM
Gene Models
CCDS
RefSeq
Polyphen Prediction
Probably Damaging
Possibly Damaging
Exclude
Sort By
Position
Gene Symbol
Omicia Score
Effect
Zygosity
Variant Miner
Reset Filters
Manage Filters
Relation Miner
Export Report
Report Versions
Overview
Genome: PG0000644-BLD.genome.block.anno.vcf.gz
Current Version:
Pipeline Version: 3.0
Gene
Position dbSNP
Change
Zygosity
Effect
Quality Coverage
Frequency
Omicia Score
Polyphen Mut-Taster
SIFT PhyloP
Evidence
ACADS chr12 121176083 rs1799958 G→A,G c.625G>A p.Gly209Ser het non-synon 58 22:15:7 G:82% A:18% 0.928 damaging damaging - 5.5 OMIM HGMD
EPHX1 chr1 226019633 rs1051740 T→C,T c.337T>C p.Tyr113His het non-synon 136 38:21:17 T:68% C:32% 0.923 damaging benign - 4.97 OMIM HGMD PGKB
BDNF chr11 27679916 rs6265 C→C,T c.196G>A p.Val66Met het non-synon 259 51:22:29 C:77% T:23% 0.861 benign benign - 3.69 OMIM HGMD PGKB GWAS
MTHFR chr1 11854476 rs1801131 T→G,T c.1286A>C p.Glu429Ala het non-synon 196 47:22:25 T:77% G:23% 0.84 benign benign 0.12 4.27 OMIM HGMD PGKB
MBL2 chr10 54531235 rs1800450 C→C,T c.161G>A p.Gly54Asp het non-synon 223 32:12:20 C:88% T:12% 0.838 damaging benign 0.01 3.14 OMIM HGMD
SLC6A20 chr3 45814094 rs17279437 G→A,G c.596C>T p.Thr199Met het non-synon 190 42:21:21 G:95% A:5% 0.837 damaging damaging - 4.18 OMIM GWAS
NQO1 chr16 69745145 rs1800566 G→A,A c.559C>T p.Pro187Ser hom non-synon 458 33:0:33 G:72% A:28% 0.836 damaging benign 0.11 5.86 OMIM HGMD PGKB
DNAH11 chr7 21582963 rs2285943 G→G,T c.100G>T p.Glu34* het stop gained 57 28:19:9 G:62% T:38% 0.832 - benign 0.74 2.22 OMIM
ABCC11 chr16 48258198 rs17822931 C→C,T c.538G>C p.Gly180Arg het non-synon 239 52:25:27 C:69% T:31% 0.818 damaging benign 0.01 2.74 OMIM HGMD
FGFR4 chr5 176520243 rs351855 G→A,G c.1162G>C p.Gly388Arg het non-synon 160 28:12:16 G:70% A:30% 0.808 damaging - 0.09 3.82 OMIM HGMD PGKB
LRP8 chr1 53712727 rs5174 C→C,T c.2066A>A p.Asp689Asp het non-synon 241 39:15:24 C:82% T:18% 0.789 damaging benign 0.05 5.04 OMIM HGMD PGKB
FRZB chr2 183703336 rs288326 G→A,G c.598C>T p.Arg200Trp het non-synon 118 38:25:13 G:95% A:5% 0.76 damaging benign - 1.62 OMIM
HNMT chr2 138759649 rs11558538 C→C,T c.314C>T p.Thr105Ile het non-synon 143 17:7:10 C:94% T:6% 0.745 damaging damaging 0.01 2.66 OMIM HGMD
OCA2 chr15 28230318 rs1800407 C→C,T c.1256G>A p.Arg419Gln het non-synon 189 38:17:21 C:96% T:4% 0.73 damaging benign 0.05 3.72 OMIM HGMD
TYR chr11 88911696 rs1042602 C→A,C c.575C>A p.Ser192Tyr het non-synon 227 41:17:24 C:82% A:18% 0.705 damaging benign 0.07 4.53 OMIM HGMD PGKB LSDB GWAS
100
Page 1 of 1
Displaying 1 to 15 of 15 items
Omicia Home - Privacy Policy - Blog
© 2013, Omicia, Inc. All rights reserved.


Omicia Opal 1.6.0
Home Upload Analyze My Results Settings
gholsonjlyon@gmail.com Report Bug Help Logout
Gene Symbol
Omicia Category
Aging
Cardiovascular
Drugs and Pharmacology
Endocrinological and Metabolic
Gastrointestinal
Blood and Lymphatic
Immune and Joints
Infectious Disease
Kidney and Urinary Tract
Neonatal
Neurological
Nutrition
Cancer
Other
Psychiatric
Respiratory
Sight
Hearing, Smell and Taste
Disease Set
Drug Set
Pathway Set
My Set
Exclude Set
Chromosome
Filter By
Require
Genotype
Heterozygous
Homozygous
Protein Impact
All
Stop Gained/Lost
Indel/Frameshift
Splice Site
Non-synonymous
Supporting Evidence
Any
OMIM
Gene Models
CCDS
RefSeq
Polyphen Prediction
Probably Damaging
Possibly Damaging
Exclude
Sort By
Position
Gene Symbol
Omicia Score
Effect
Zygosity
Variant Miner
Reset Filters
Manage Filters
Relation Miner
Export Report
Report Versions
Overview
Genome: PG0000644-BLD.genome.block.anno.vcf.gz
Current Version:
Pipeline Version: 3.0
Gene
Position dbSNP
Change
Zygosity
Effect
Quality Coverage
Frequency
Omicia Score
Polyphen Mut-Taster
SIFT PhyloP
Evidence
NQO1 chr16 69745145 rs1800566 G→A,A c.559C>T p.Pro187Ser hom non-synon 458 33:0:33 G:72% A:28% 0.836 damaging benign 0.11 5.86 OMIM HGMD PGKB
DPYD chr1 98348885 rs1801265 G→A,A c.85C>T p.Arg29Cys hom non-synon 317 20:0:20 G:23% A:77% 0.708 - - 0.18 2.55 HGMD PGKB
ABCA1 chr9 107562804 rs2230808 T→C,C c.4760A>G p.Lys1587Arg hom non-synon 536 38:0:38 T:41% C:59% 0.7 benign benign 1 4.87 HGMD
NAT2 chr8 18258103 rs1799930 G→A,G c.590G>A p.Arg197Gln het non-synon 220 37:16:21 G:76% A:24% 0.653 damaging benign 0.08 3.11 OMIM HGMD PGKB
ABCA1 chr9 107589255 rs2066718 C→C,T c.2311G>A p.Val771Met het non-synon 195 40:19:21 C:94% T:6% 0.562 benign damaging 1 1.4 HGMD
CYP4F2 chr19 15990431 rs2108622 C→C,T c.1297G>A p.Val433Met het non-synon 183 30:12:18 C:78% T:22% 0.473 damaging benign 0.01 2.31 HGMD PGKB GWAS
NAT2 chr8 18257854 rs1801280 T→C,T c.341T>C p.Ile114Thr het non-synon 191 39:20:19 T:70% C:30% 0.467 benign benign 0.08 0.74 OMIM HGMD PGKB
DPYD chr1 97981395 rs1801159 T→C,T c.1627A>G p.Ile543Val het non-synon 153 24:11:13 T:80% C:20% 0.295 benign benign 1 0.86 HGMD PGKB
OGG1 chr3 9798773 rs1052133 C→C,G c.294C>G p.Ile98Met het non-synon 146 30:16:14 C:70% G:30% 0.258 - - 0.01 -0.25 HGMD
OGG1 chr3 9798773 rs1052133 C→C,G c.994C>G p.Pro332Ala het non-synon 146 30:16:14 C:70% G:30% 0.258 - - 0.01 -0.25 HGMD
OGG1 chr3 9798773 rs1052133 C→C,G c.977C>G p.Ser326Cys het non-synon 146 30:16:14 C:70% G:30% 0.258 - - 0.01 -0.25 HGMD
CYP2C9 chr10 96741053 rs1057910 A→C,C c.1076A>C p.Ile359Leu hom non-synon 496 36:0:36 A:96% C:4% 0.189 benign damaging 0.11 - OMIM HGMD PGKB
ABCA1 chr9 107620867 rs2230806 C→C,T c.656G>A p.Arg219Lys het non-synon 131 30:18:12 C:58% T:42% 0.187 benign benign 0.32 0.16 OMIM HGMD PGKB
CYP2B6 chr19 41515263 rs28399497 A→A,G c.785A>G p.Lys262Arg het non-synon 54 17:8:9 - 0.178 benign benign 1 0.84 HGMD
NBN chr8 90990479 rs1805794 C→C,G c.553G>C p.Glu185Gln het non-synon 193 30:12:18 C:67% G:33% 0.172 benign benign 1 0.5 HGMD
CYP4F12 chr19 15789140 rs609290 A→G,G c.267+1A>G hom splice site 578 44:0:44 A:6% G:94% 0.172 - - - -0.6 HGMD
CYP3A7 chr7 99306685 rs2257401 C→G,G c.1226G>C p.Arg409Thr hom non-synon 331 22:0:22 C:27% G:73% 0.163 benign benign 0.16 0.35 PGKB
CYP4F12 chr19 15789140 rs609290 A→G,G c.269A>G p.Ile90Val hom non-synon 578 44:0:44 A:6% G:94% 0.126 - benign 0.7 -0.6 HGMD
CETP chr16 G→A,G het non-synon 203 G:45% 0.088 benign 1 HGMD PGKB
100
Page 1 of 1
Displaying 1 to 24 of 24 items
Omicia Home - Privacy Policy - Blog
© 2013, Omicia, Inc. All rights reserved.

3 common SNVs in this person that have been implicated in the literature as predisposing to mental illness.

Gene name	Genomic coordinates	Amino acid change	Zygosity	Mutation type	Population Frequency	Clinical significance
MTHFR	chr1: 11854476	Glu>Ala	heterozygous	non-synon	T:77% G:23%	Susceptibility to psychoses, schizophrenia, occlusive vascular disease, neural tube defects, colon cancer, acute leukemia, and methylenetetra-hydrofolate reductase deficiency
BDNF	chr11: 27679916	Val>Met	heterozygous	non-synon	C:77% T:23%	Susceptibility to OCD, psychosis, and diminished response to exposure therapy
CHAT	chr10: 50824117	Asp>Asn	heterozygous	non-synon	G:85% A:15%	Susceptibility to schizophrenia and other psychopathological disorders.

Gene Summary for CHAT



Gene Overview	
Symbol	CHAT
Name	choline O-acetyltransferase
Location	10q11.2
Summary	This gene encodes an enzyme which catalyzes the biosynthesis of the neurotransmitter acetylcholine. This gene product is a characteristic feature of cholinergic neurons, and changes in these neurons may explain some of the symptoms of Alzheimer's disease. Polymorphisms in this gene have been associated with Alzheimer's disease and mild cognitive impairment. Mutations in this gene are associated with congenital myasthenic syndrome associated with episodic apnea. Multiple transcript variants encoding different isoforms have been found for this gene, and some of these variants have been shown to encode more than one isoform. [provided by RefSeq, May 2010]

Relevant Reference Resources	
NCBI Gene	http://www.ncbi.nlm.nih.gov/gene/1103
GeneTests	http://www.ncbi.nlm.nih.gov/sites/GeneTests/lab/gene/CHAT
Ensembl	http://www.ensembl.org/human/Gene/Summary?g=ENSG00000070748
UCSC Gene Browser	http://genome.ucsc.edu/cgi-bin/hgTracks?org=human&db=hg19&singleSearch=knownCanonical&position=CHAT
Genetics Home Reference	http://ghr.nlm.nih.gov/gene/CHAT

Associated Disease Categories

Category	Disease	Citation
DRUGS, CLINICAL PHARMACOLOGY AND ENVIRONMENT	Drug toxicity	Roden et al., 2002

Associated Knowledge Sets

Name	Type	Description
ODG - Alzheimers	disease	Omicia Disease Genes (ODG) Top 10 Neurological - Alzheimers
TruSight Exome	disease	Illumina's targeted rare genetic conditions exome test containing 2,761 genes covered in the HGMD database.
MitoGO	myset	
Longo - Phenomizer Fatty Acid Big	myset	A list of genes from phenomizer build from Patient Features HP:0004359. Long List ~3000 genes

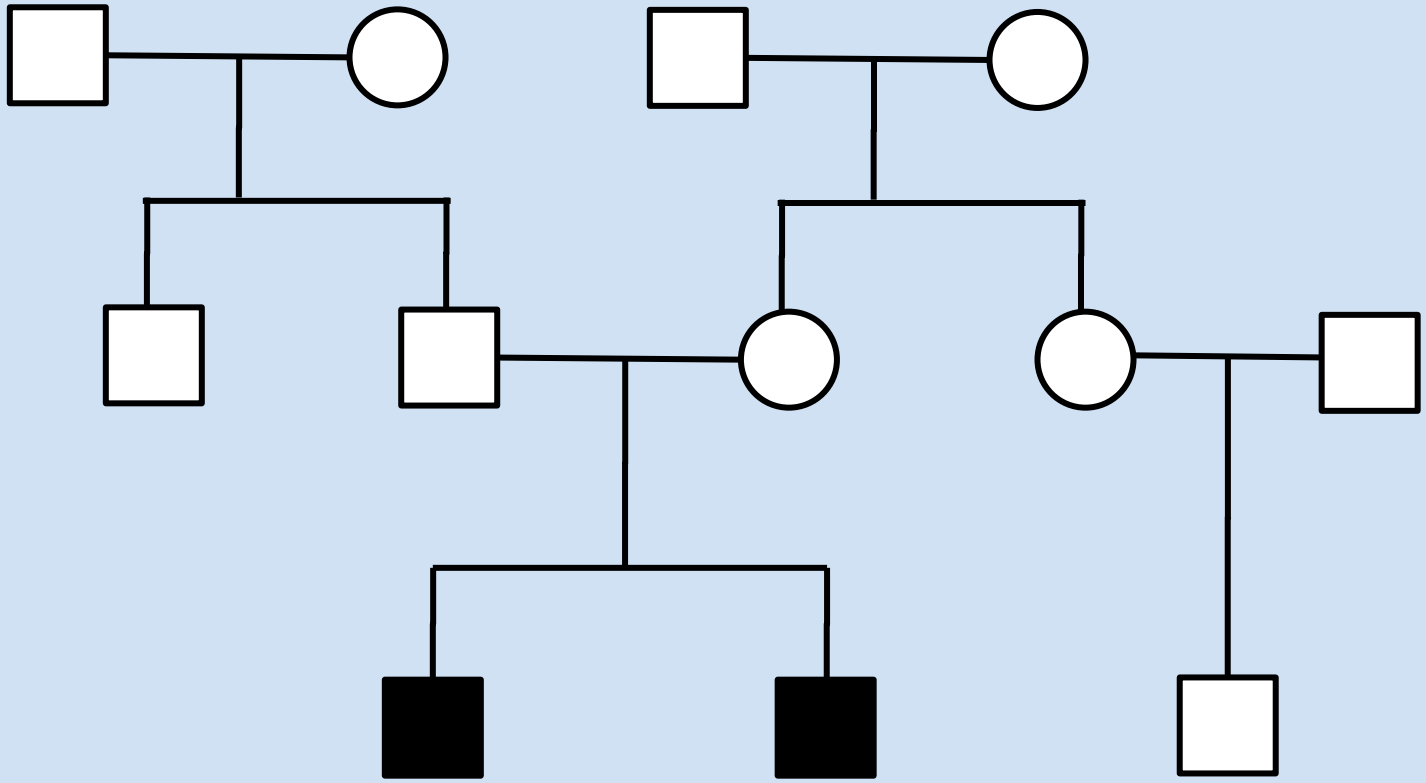
Personal Variants in this Gene

Position	Transcript	Transcript HGVS	Protein	Protein HGVS	Zyg	Effect
50824117	NM_001142933.1	c.19G>A	NP_001136405	p.Asp7Asn	het	non-synon
50824619	NM_001142933.1	c.112G>A	NP_001136405	p.Ala38Thr	het	non-synon
50856652	NM_020549	c.1382G>A	NP_065574	p.Val461Met	hom	non-synon
50863147	NM_020549	c.1642T>C	NP_065574	p.His548His	hom	synonymous

Lessons

- Should look at entire genome.
- Should archive and store genomic data for re-analysis for many years to come.

Vignette #2: New Syndrome with Intellectual Disability, “Autism”, “ADHD”



Likely X-linked or Autosomal Recessive, with X-linked being supported by extreme X-skewing in the mother

1.5 years old

3.5 years old

3 years old

5 years old

Dysmorphic
Mental Retardation
“autism”
“ADHD”
Hearing difficulties

Workup Ongoing for past 10 years

- Numerous genetic tests negative, including negative for Fragile X and MANY candidate genes.

- Whole genome sequencing was performed using :
  - Complete Genomics sequencing and analysis pipeline v2.0
  - Ilumina HiSeq 2000 sequencing platform.
    - Illumina reads were mapped to the hg19 reference genome using BWA v. 0.6.2-r126
    - Variant detection was performed using the GATK v. 2.4-9.
    - A second analytical pipeline was used to map reads to the hg19 reference genome using Novoalign, and variants were also detected using the FreeBayes caller.
- For each sequenced individual, a union of the calls made by both sequencing platforms, as well as all bioinformatics analysis platforms, was taken.

- Standard approaches can then be used to identify potentially deleterious mutations conforming to classical disease models for genetic disorders.

- We subset the full dataset to evaluate differences between raw numbers of mutations detected between different data sets:

  – WGS data from the nuclear family,
  – WGS from a larger portion of the family.

Using only nuclear family:

55195 Variants were found to be *de-novo* in the two affected boys

122 were coding :

107 non-synonymous missense

4 splicing

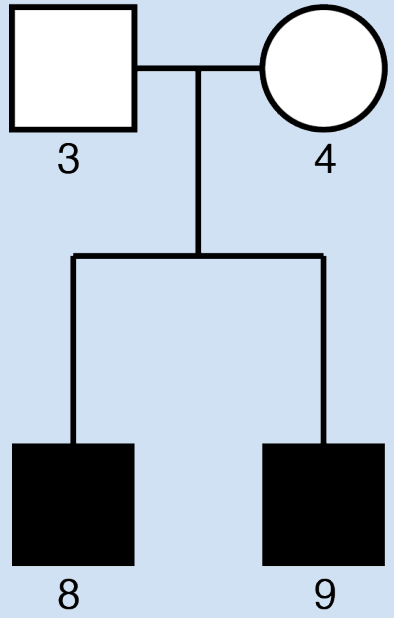
3 frame-shift deletions

3 frame-shift insertions

2 frame-shift substitutions

2 stop-gain

1 stop-loss



26514 Variants were found to conform to an X-linked disease model

28 were coding:

27 non-synonymous missense

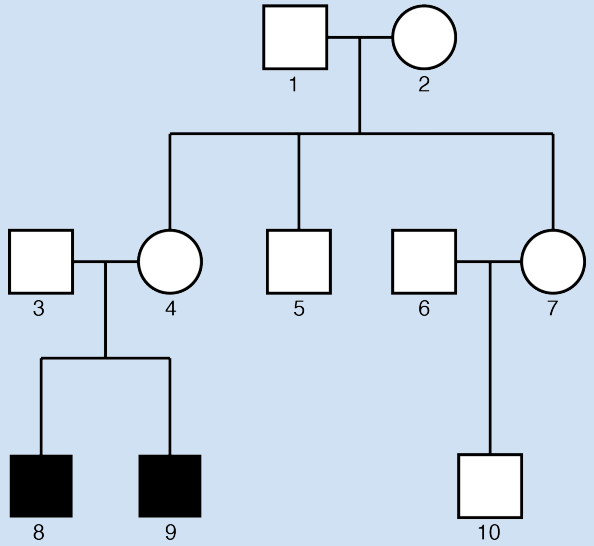
1 splicing

Using information from a greater portion of the family structure:

17726 Variants were found to be *de-novo* in the two affected boys

40 were coding :

- 32 non-synonymous missense
- 3 splicing
- 2 frame-shift deletions
- 1 stop-loss
- 1 frame-shift insertion
- 1 frame-shift substitution

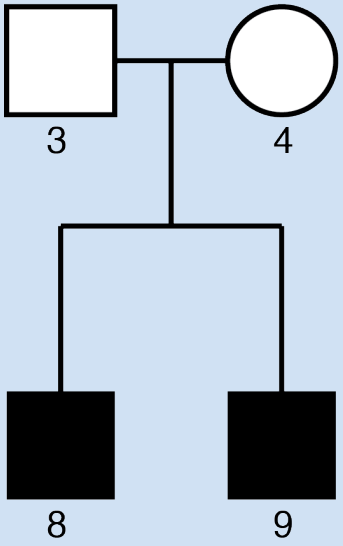


2824 Variants were found to conform to an X-linked disease model

4 were coding:

- 3 non-synonymous missense
- 1 splicing

- The numbers of mutations differ as expected between these two sets of analyses:
  - More mutations are filtered when a greater portion of the family is incorporated into the analysis.
  - This is likely due to false positive and false negative rates across sequencing and informatics platforms.



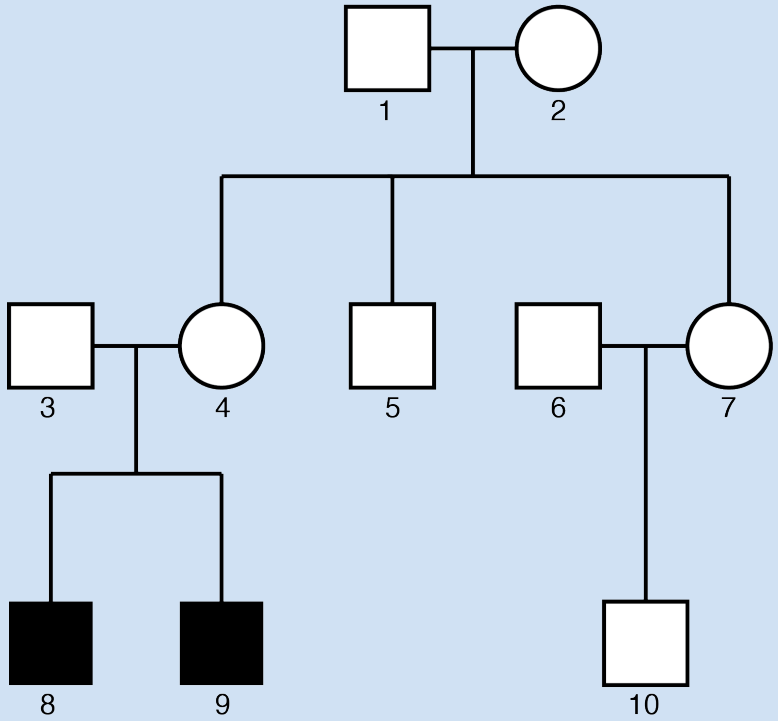
Using only nuclear family:

De-novo ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
1	PRAMEF4	0.00192	0.00144,0.00265	13.13	chr1:12939476;13.13;G->C;N->K;0,1
2	PRAMEF10	0.00318	0.00243,0.00417	20.77	chr1:12954852;20.77;T->C;H->R;3,2
3	LOC440563	0.00523	0.00416,0.00653	9.89	chr1:13183056;9.89;T->C;N->D;0,1

X-linked ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
1	ASB12	0.000898	0.000898,0.00119	18.7	chrX:63444792;18.70;C->A;G->C;0,1
2	TAF1	0.00153	0.00117,0.00214	14.59	chrX:70621541;14.59;T->C;I->T;0,1
3	ZNF41	0.002	0.0015,0.00275	12.9	chrX:47307978;12.90;G->T;D->E;0,1



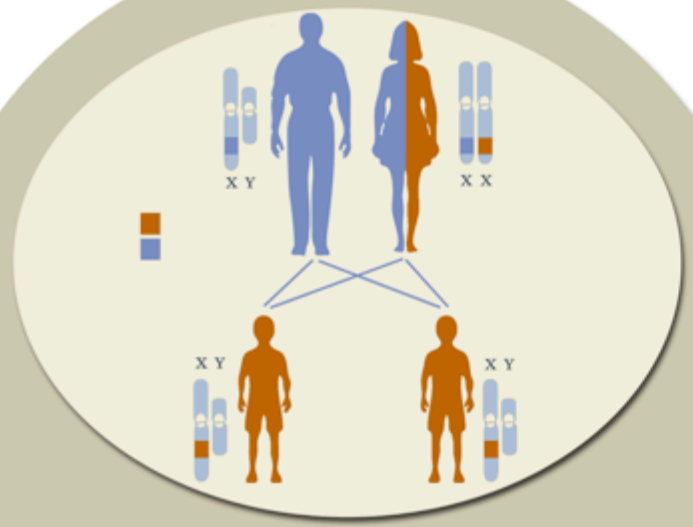
Using information from a greater portion of the family structure:

De-novo ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
1	PRAMEF10	0.00342	0.00262,0.00445	20.77	chr1:12954852;20.77;T->C;H->R;3,2

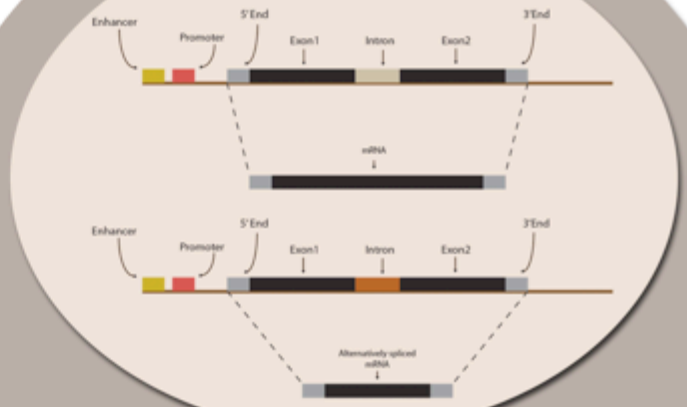
X-linked ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
1	TAF1	0.002	0.0015,0.00275	14.59	chrX:70621541;14.59;T->C;I->T;0,1



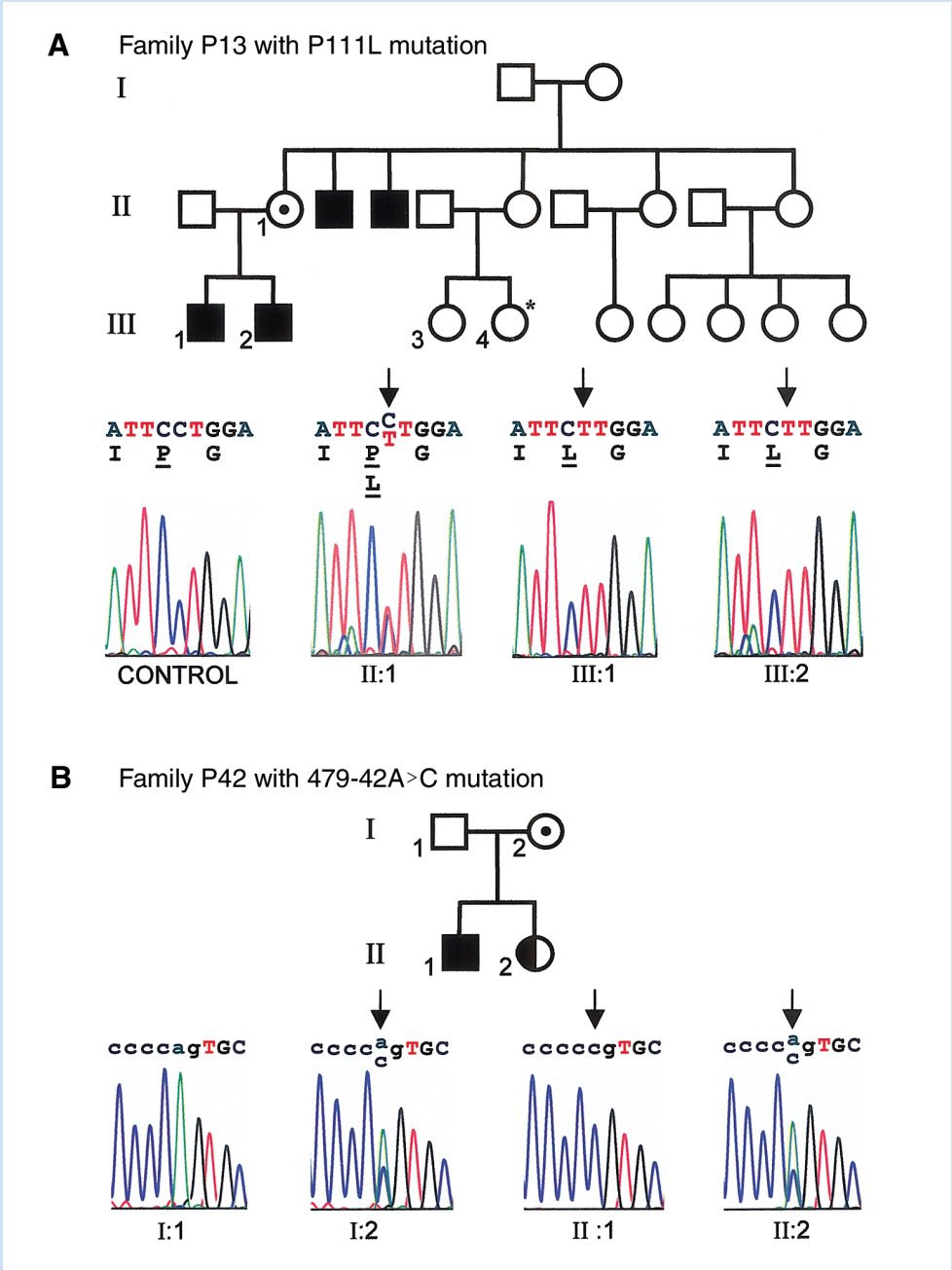
X-linked

Gene	Locus	Exon	Protein
ZNF41	X:47307978	5	p.Asp397Glu
ASB12	X:63444792	2	p.Gly247Cys
TAF1	X:70621541	25	p.Ile1337Thr



Non-coding

Gene	Locus	Exon	Protein
UTR3 AR	X:66945414	-------	-------
FAM155B (*dist=271971*)	X:68453113	-------	-------
MIR221 (*dist=35606*)	X:45569979	-------	-------
DMD-AS2 *intronic*	X:31284835	-------	-------
MID1 (*dist=30252*)	X:10383096	-------	-------



The two brothers with the P111L mutations reported in the prior paper do have mental deficiency, hyperkinesia, no motor or neurologic sign except for the delay, and slight dysmorphic facial anomalies: large low-set ears, thin upper lip, slight downward palpebral slants, but no upturned nose, and a short philtrum. The mother was normal in appearance.

- Previously reported P111L change in the ZNF41 protein has now also been found in two "male controls" (EVS server, ESP6500), and furthermore, there are two rare, likely heterozygous ZNF41 frameshift mutations and one heterozygous stop-gained mutation reported in control individuals (ESP6500) (personal communication from Dr. Vera Kalscheuer).

Major Conclusion: Clinical Validity?

This is so complex that the only solid way forward is with a “networking of science” model, i.e. online database with genotype and phenotype longitudinally tracked for thousands of volunteer families.

We need a 1 Million Genome/ Phenome Project!

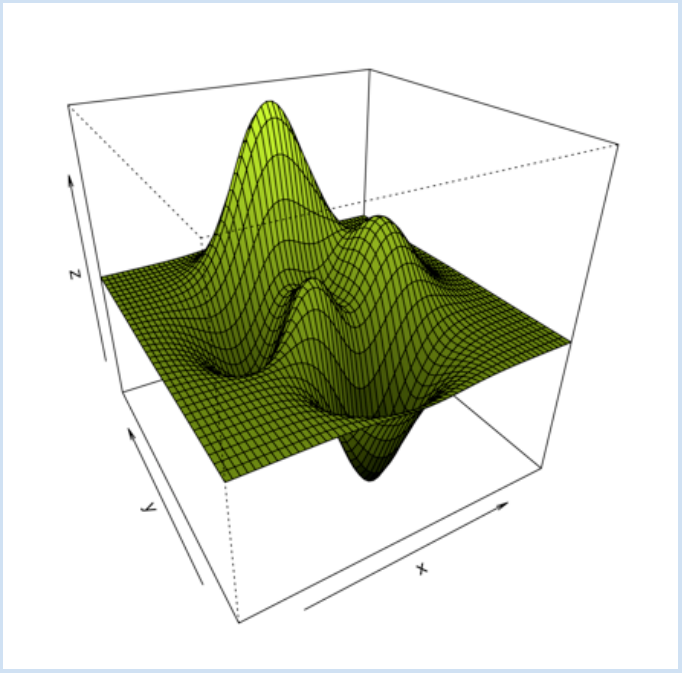
Creating a Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data

June 3, 2013

Organizing Committee

David Altshuler	Broad Institute of Harvard and MIT, MGH
Peter Goodhand	Ontario Institute for Cancer Research
David Haussler	HHMI/University of California, Santa Cruz
Thomas Hudson	Ontario Institute for Cancer Research
Brad Margus	A-T Children's Project
Betsy Nabel*	Brigham and Women's Hospital
Charles Sawyers	HHMI / Memorial Sloan-Kettering
Michael Stratton*	Wellcome Trust Sanger Institute

Need sharing of thousands to millions of “individuated genomes” – credit Nathaniel Pearson
http://www.slideshare.net/NathanielPearson/pearsontcgc2013


z
y
x

The End

Extra slides appear hereafter, not shown in the main talk

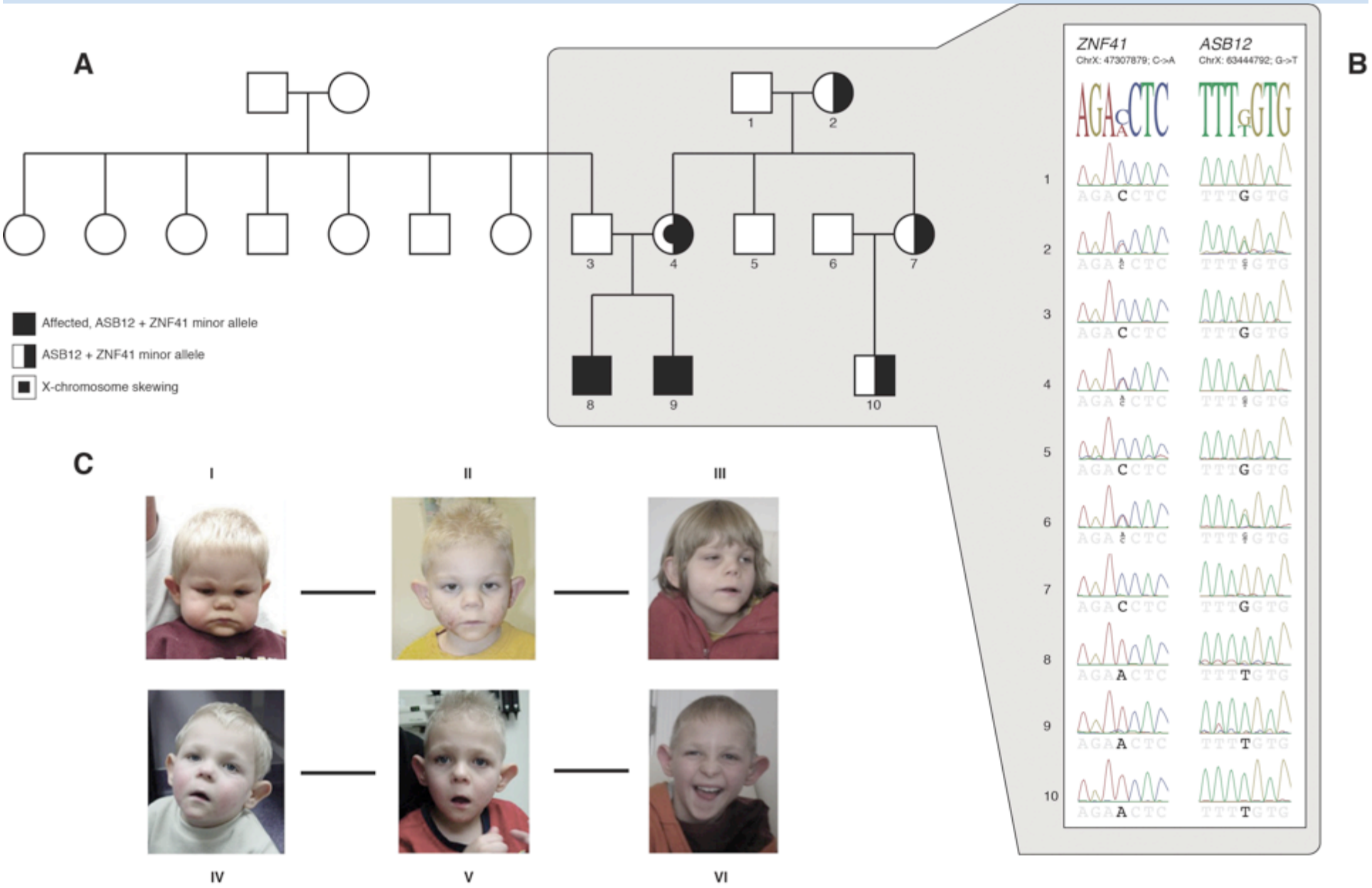
"We'd now like to open the floor to shorter speeches disguised as questions."

Mutations in the *ZNF41* Gene Are Associated with Cognitive Deficits: Identification of a New Candidate for X-Linked Mental Retardation

Sarah A. Shoichet,[1] Kirsten Hoffmann,[1] Corinna Menzel,[1] Udo Trautmann,[2] Bettina Moser,[1] Maria Hoeltzenbein,[1] Bernard Echenne,[3] Michael Partington,[4] Hans van Bokhoven,[5] Claude Moraine,[6] Jean-Pierre Fryns,[7] Jamel Chelly,[8] Hans-Dieter Rott,[2] Hans-Hilger Ropers,[1] and Vera M. Kalscheuer[1]

[1]Max-Planck-Institute for Molecular Genetics, Berlin; [2]Institute of Human Genetics, University of Erlangen-Nuremberg, Erlangen-Nuremberg; [3]Centre Hospitalier Universitaire de Montpellier, Hôpital Saint-Eloi, Montpellier, France, [4]Hunter Genetics and University of Newcastle, Waratah, Australia; [5]Department of Human Genetics, University Medical Centre, Nijmegen, The Netherlands; [6]Services de Génétique–INSERM U316, CHU Bretonneau, Tours, France; [7]Center for Human Genetics, Clinical Genetics Unit, Leuven, Belgium; and [8]Institut Cochin de Génétique Moleculaire, Centre National de la Recherche Scientifique/INSERM, CHU Cochin, Paris

Am. J. Hum. Genet. 73:1341–1354, 2003


A
1
2
3
4
5
6
7
8
9
10
Affected, ASB12 + ZNF41 minor allele
ASB12 + ZNF41 minor allele
X-chromosome skewing
B
ZNF41
ChrX: 47307879; C->A
ASB12
ChrX: 63444792; G->T
C
I
II
III
IV
V
VI

- Whole genome sequencing was performed using :
  - Complete Genomics sequencing and analysis pipeline v2.0
  - Ilumina HiSeq 2000 sequencing platform.
    - Illumina reads were mapped to the hg19 reference genome using BWA v. 0.6.2-r126
    - Variant detection was performed using the GATK v. 2.4-9.
    - A second analytical pipeline was used to map reads to the hg19 reference genome using Novoalign, and variants were also detected using the FreeBayes caller.
- For each sequenced individual, a union of the calls made by both sequencing platforms, as well as all bioinformatics analysis platforms, was taken.

- Standard approaches can then be used to identify potentially deleterious mutations conforming to classical disease models for genetic disorders.

- We subset the full dataset to evaluate differences between raw numbers of mutations detected between different data sets:

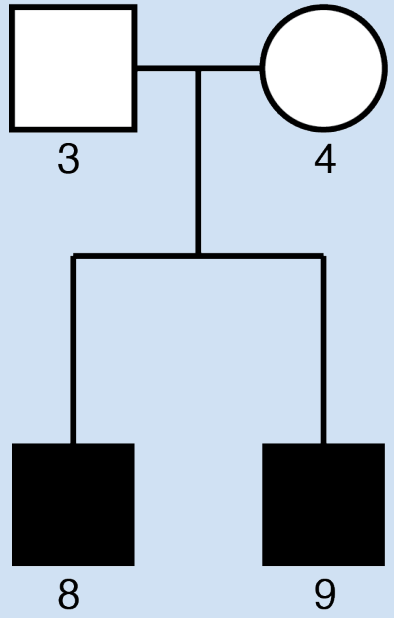
  – WGS data from the nuclear family,
  – WGS from a larger portion of the family.

Using only nuclear family:

55195 Variants were found to be *de-novo* and shared in the two affected boys

122 were coding :

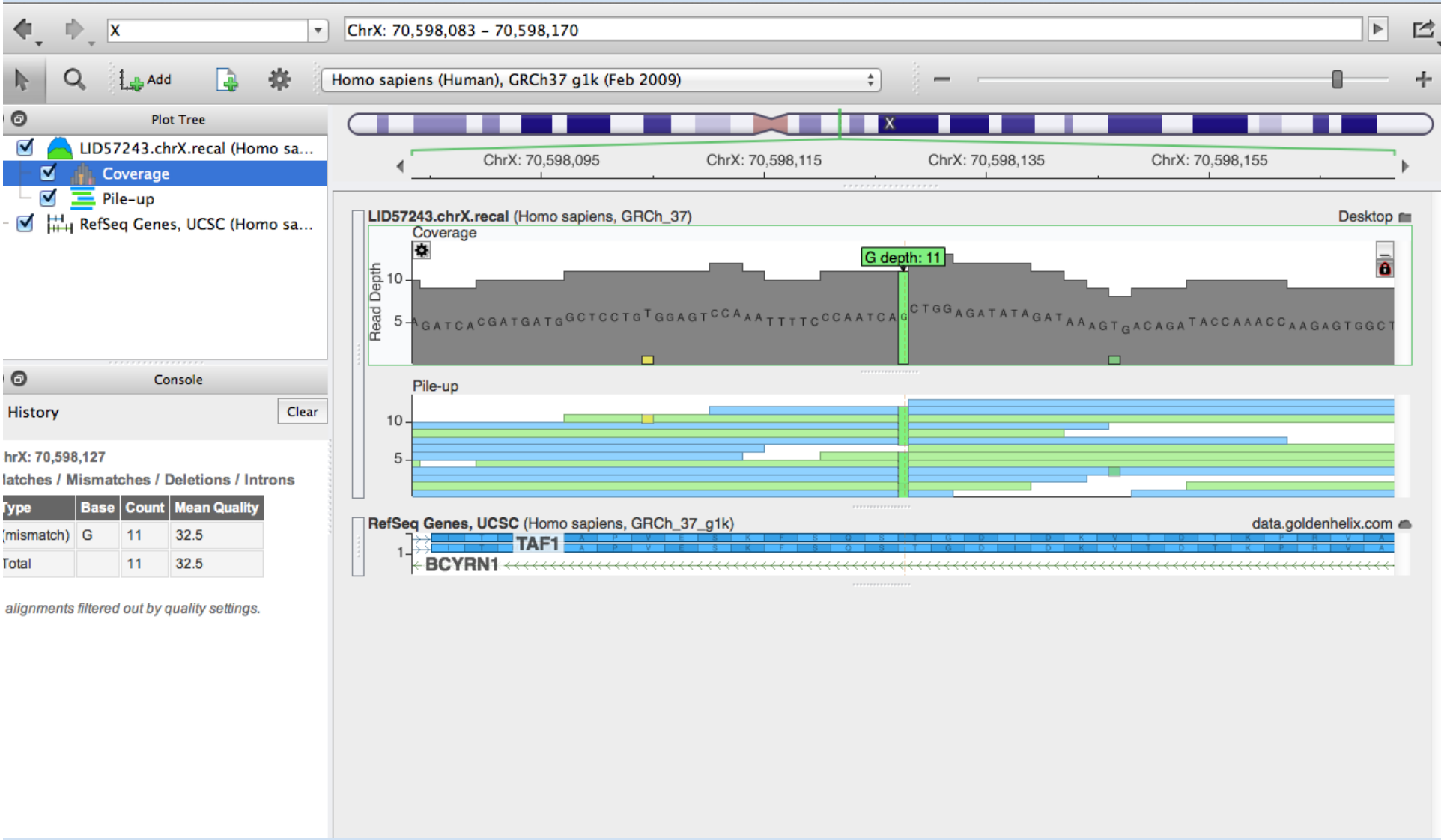
- 107 non-synonymous missense
- 4 splicing
- 3 frame-shift deletions
- 3 frame-shift insertions
- 2 frame-shift substitutions
- 2 stop-gain
- 1 stop-loss



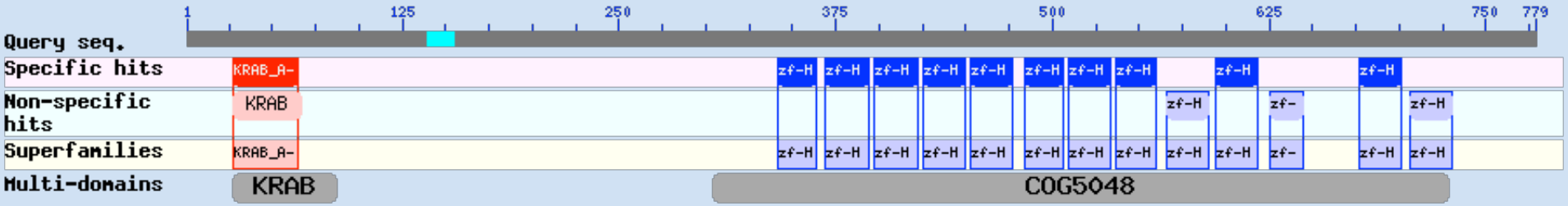
26514 Variants were found to conform to an X-linked disease model

28 were coding:

- 27 non-synonymous missense
- 1 splicing


X
ChrX: 70,598,083 - 70,598,170
Add
Homo sapiens (Human), GRCh37 g1k (Feb 2009)
Plot Tree
LID57243.chrX.recal (Homo sa...
Coverage
Pile-up
RefSeq Genes, UCSC (Homo sa...
ChrX: 70,598,095
ChrX: 70,598,115
ChrX: 70,598,135
ChrX: 70,598,155
LID57243.chrX.recal (Homo sapiens, GRCh_37)
Desktop
Coverage
G depth: 11
Read Depth
Pile-up
RefSeq Genes, UCSC (Homo sapiens, GRCh_37_g1k)
data.goldenhelix.com
TAF1
BCYRN1
Console
History
Clear
hrX: 70,598,127
latches / Mismatches / Deletions / Introns
Type
Base
Count
Mean Quality
(mismatch)
G
11
32.5
Total
11
32.5
alignments filtered out by quality settings.

ZNF41



- KRAB (Kruppel-associated box) domain -A box.
- The KRAB domain is a transcription repression module, found in a subgroup of the zinc finger proteins (ZFPs) of the C2H2 family, KRAB-ZFPs. KRAB-ZFPs comprise the largest group of transcriptional regulators in mammals, and are only found in tetrapods.
- The KRAB domain is a protein-protein interaction module which represses transcription through recruiting corepressors. The KAP1/ KRAB-AFP complex in turn recruits the heterochromatin protein 1 (HP1) family, and other chromatin modulating proteins, leading to transcriptional repression through heterochromatin formation.

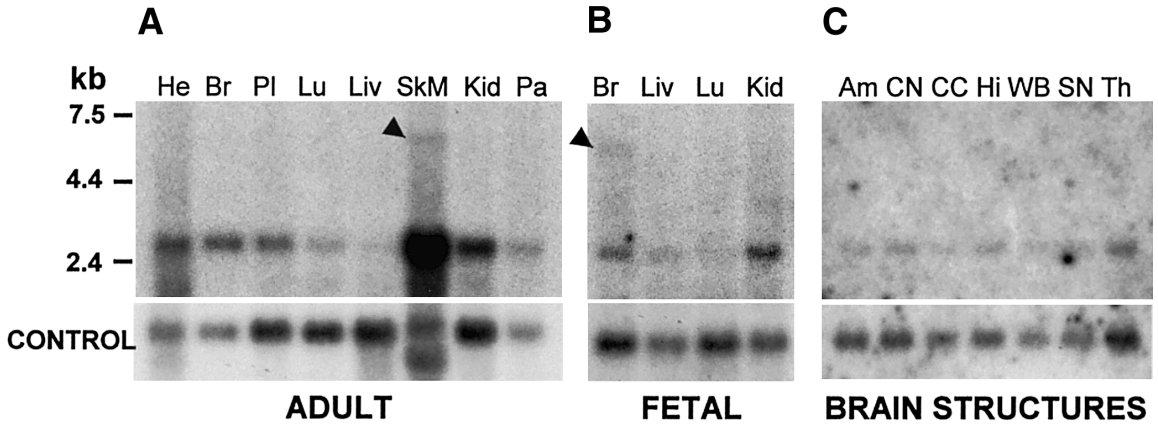


Figure 6 Northern blot hybridization of *ZNF41*, by use of a probe corresponding to nucleotides 621–1099 of *ZNF41* transcript variant 1. *A*, Adult tissues (left to right): heart, brain, placenta, lung, liver, skeletal muscle, kidney, and pancreas. *B*, Fetal tissues (left to right): brain, lung, liver, and kidney. *C*, Adult brain structures (left to right): amygdala, caudate nucleus, corpus callosum, hippocampus, whole brain, substantia nigra, and thalamus. Black arrowheads highlight the presence of a novel 6-kb transcript. *Actin* (*A* and *C*) or *GAPDH* (*B*) served as controls for RNA loading.

Proving Causality

- Will need to find a second, unrelated family with same exact mutation and similar phenotype.
- Can also perform in vitro/in vivo studies and structural modeling, and make knock-in mice and/or test in zebrafish, etc… for biological function.

Nuclear family, using only Illumina data

50715 Variants were found to be *de-novo* in the two affected boys
107 were coding :

2 frame-shift substitutions
2 frame-shift deletions
3 frame-shift insertions
94 non-synonymous missense
4 splicing
2 stop-gain

25157 Variants were found to conform to an X-linked disease model
29 were coding:

1 frame-shift deletions
28 non-synonymous missense

De-novo ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
1	PRAMEF4	0.00216	0.00162,0.00294	13.13	chr1:12939476;13.13;G->C;N->K;0,1
2	LOC440563	0.00538	0.00429,0.00671	9.89	chr1:13183056;9.89;T->C;N->D;0,1

X-linked ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
1	ASB12	0.000898	0.000898,0.00119	18.7	chrX:63444792;18.70;C->A;G->C;0,1
2	VCX3A	0.00145	0.00112,0.00203	13	chrX:6451809;13.00;T->C;M->V;0,1
3	TAF1	0.00169	0.00128,0.00235	14.59	chrX:70621541;14.59;T->C;I->T;0,1
4	ZNF41	0.00231	0.00174,0.00314	12.9	chrX:47307978;12.90;G->T;D->E;0,1

Nuclear family, using only CG data

42072 Variants were found to be *de-novo* in the two affected boys
75 were coding :

3 frame-shift deletions
4 frame-shift insertions
62 non-synonymous missense
5 splicing
1 stop-loss

21397 Variants were found to conform to an X-linked disease model
25 were coding:

1 splicing
24 non-synonymous missense

De-novo ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
1	PRAMEF10	0.00318	0.00243,0.00417	20.77	chr1:12954852;20.77;T->C;H->R;3,2

X-linked ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
1	ASB12	0.000898	0.000898,0.00119	18.7	chrX:63444792;18.70;C->A;G->C;0,1
2	TAF1	0.00145	0.00112,0.00203	14.59	chrX:70621541;14.59;T->C;I->T;0,1
3	ZNF41	0.00271	0.00205,0.00361	12.9	chrX:47307978;12.90;G->T;D->E;0,1

A greater portion of the family, using only Illumina data

11675 Variants were found to be *de-novo* in the two affected boys

18 were coding :

15 non-synonymous missense
2 splicing
1 frame-shift substitution

1773 Variants were found to conform to an X-linked disease model

3 were coding:

3 non-synonymous missense

De-novo ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
-	-	-	-	-	-

X-linked ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
1	TAF1	0.00184	0.00139,0.00255	14.59	chrX:70621541;14.59;T->C;I->T;0,1

Genotype First, Phenotype Second AND Longitudinally

Human Molecular Genetics, 2010, Vol. 19, Review Issue 2 ***R176–R187***
doi:10.1093/hmg/ddq366
Advance Access published on August 31, 2010

Phenotypic variability and genetic susceptibility to genomic disorders

Santhosh Girirajan and Evan E. Eichler*

Department of Genome Sciences, Howard Hughes Medical Institute, University of Washington School of Medicine, PO Box 355065, Foege S413C, 3720 15th Avenue NE, Seattle, WA 98195, USA

Genome-Wide Association Study of Multiplex Schizophrenia Pedigrees

Am J Psychiatry Levinson et al.; AiA:1–11

"Rare CNVs were observed in regions with strong previously documented association with schizophrenia, but with variable patterns of segregation. This should serve as a reminder that we still know relatively little about the distribution of these CNVs in the entire population (e.g., in individuals with no or only mild cognitive problems) or about the reasons for the emergence of schizophrenia in only a minority of carriers, so great caution is required in genetic counseling and prediagnosis."

Lyon and Wang *Genome Medicine* 2012, 4:58
http://genomemedicine.com/content/4/7/58

REVIEW

Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress

Gholson J Lyon[*1,2] and Kai Wang[*2,3]

Contents lists available at SciVerse ScienceDirect

Applied & Translational Genomics

journal homepage: www.elsevier.com/locate/atg

Practical, ethical and regulatory considerations for the evolving medical and research genomics landscape

Gholson J. Lyon [a,b,*], Jeremy P. Segal [c,**]

[a] *Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, NY, United States*
[b] *Utah Foundation for Biomedical Research, Salt Lake City, UT, United States*
[c] *New York Genome Center, New York City, NY, United States*

2 mutations present in mother and two boys, on X-chromosome, not in father, not in dbSNP135, not in 1000Genomes April 2012 release, and not in NHLBI 6500 Exomes

- Nonsyn SNV ZNF41 c.1191C>A p.Asp397Glu
- Nonsyn SNV TAF1 c.4010T>C p.Ile1337Thr

TAF1 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 250kDa

Lyon and Wang *Genome Medicine* 2012, 4:58
http://genomemedicine.com/content/4/7/58

REVIEW

Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress

Gholson J Lyon[*1,2] and Kai Wang[*2,3]

Contents lists available at SciVerse ScienceDirect

Applied & Translational Genomics

journal homepage: www.elsevier.com/locate/atg

Practical, ethical and regulatory considerations for the evolving medical and research genomics landscape

Gholson J. Lyon [a,b,*], Jeremy P. Segal [c,**]

[a] *Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, NY, United States*
[b] *Utah Foundation for Biomedical Research, Salt Lake City, UT, United States*
[c] *New York Genome Center, New York City, NY, United States*

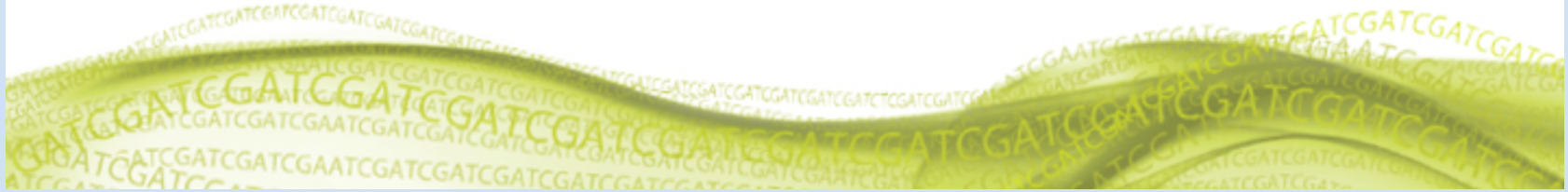
De novo **variants were detected using WGS data from 10 individuals in one family.**

	Number of putative *de novo* variants that were detected	Number of putative *de novo* coding non-synonymous or splicing variants detected
Using WGS data only from nuclear family	**68241**	**163**
Using WGS data including extended family (which includes 6 additional members)	**21168**	**59**

Whole genomes for all 10 individuals were sequenced using the Illumia HiSeq 2000 platform and then analyzed using the BWA-GATK and Novalign-FreeBayes informatics pipelines. The mother and father of the affected, as well as the two affected, were additionally sequenced using the Complete Genomics sequencing and analysis pipeline, version 2.0. For each individual, variants detected by each sequencing platform and by each analysis platform were unioned to include all detected variants. Using this combination of data, de novo variants were detected in two affected children.

Understand Your Genome Symposium

During this two-day educational event, industry experts will discuss the clinical implementation of whole-genome next-generation sequencing (NGS) technology.



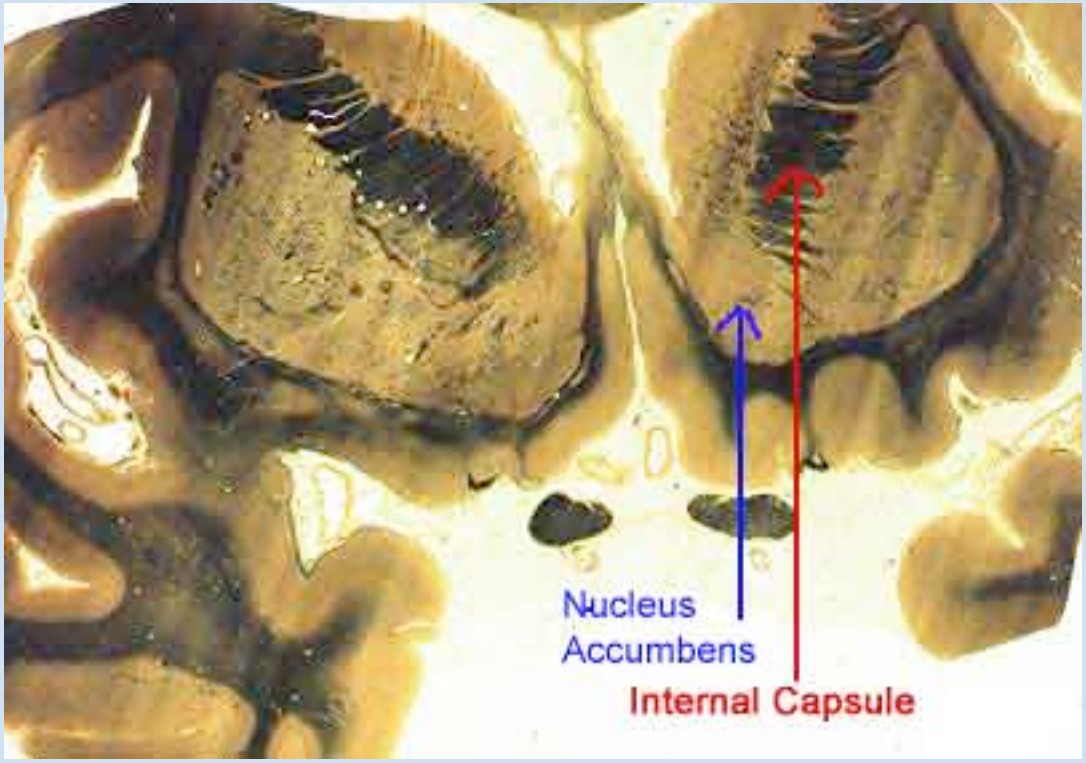
illumina®

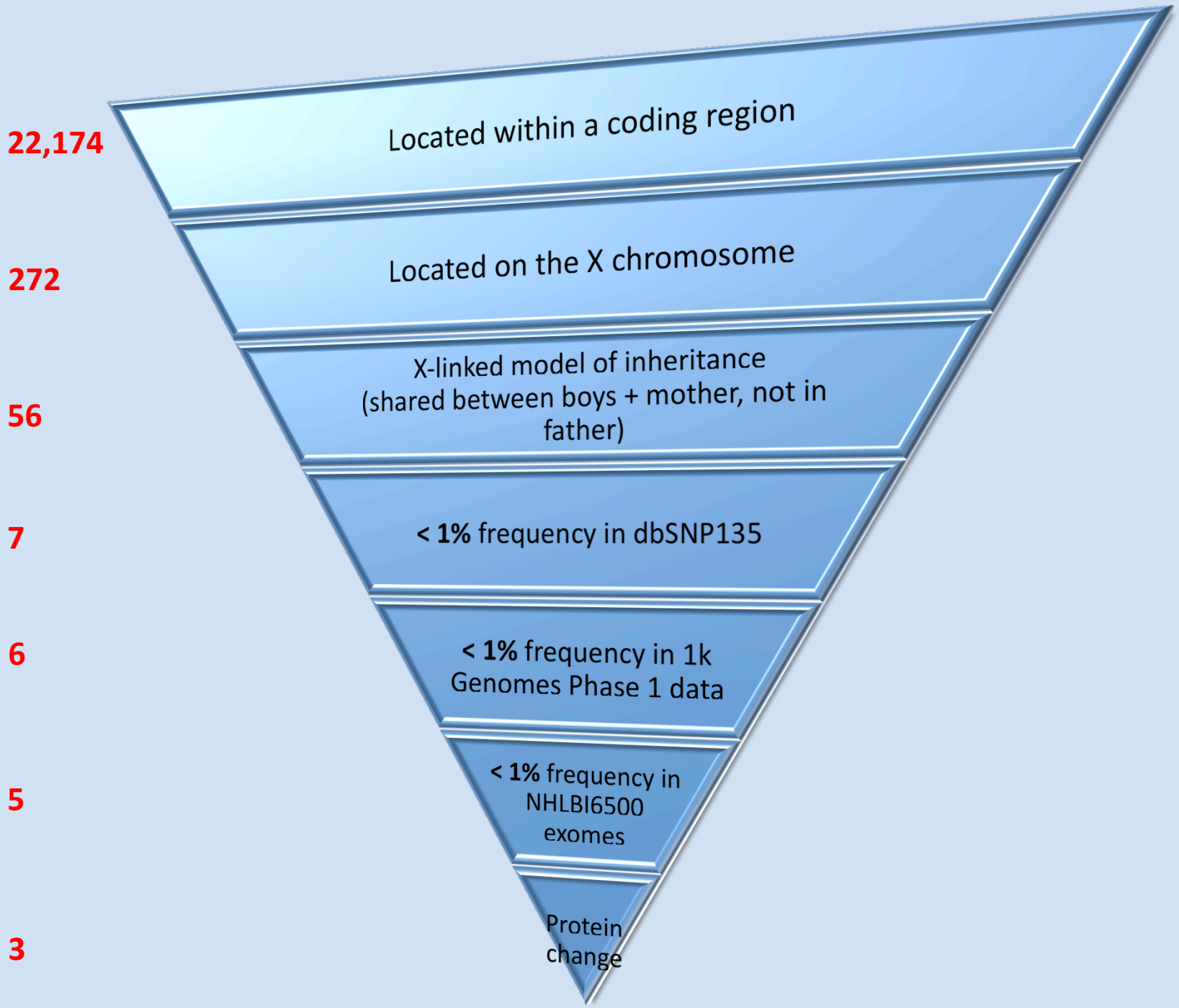
Ordering Physician:
Gholson Lyon, MD
Steinmann Institute
10 West Broadway, Suite #820
Salt Lake City, UT 84101

Individual Genome Sequence Results

Clinical Report

www.everygenome.com
CLIA#: 05D1092911


Nucleus
Accumbens
Internal Capsule


22,174
Located within a coding region
272
Located on the X chromosome
56
X-linked model of inheritance (shared between boys + mother, not in father)
7
< 1% frequency in dbSNP135
6
< 1% frequency in 1k Genomes Phase 1 data
5
< 1% frequency in NHLBI6500 exomes
3
Protein change

Variant classification

Variant	Reference	Alternate	Classification	Gene 1	Transcript 1	Exon 1	HGVS Coding 1	HGVS Protein 1
X:47307978-SNV	G	T	Nonsyn SNV	ZNF41	NM_007130	5	c.1191C>A	p.Asp397Glu
X:63444792-SNV	C	A	Nonsyn SNV	ASB12	NM_130388	2	c.739G>T	p.Gly247Cys
X:70621541-SNV	T	C	Nonsyn SNV	TAF1	NM_004606	25	c.4010T>C	p.Ile1337Thr

SIFT classification

Chromosome	Position	Reference	Coding?	SIFT Score	Score <= 0.05	Ref/Alt Alleles
X	47307978	G	YES	0.649999976	0	G/T
X	63444792	C	YES	0	1	C/A
X	70621541	T	YES	0.009999999776	1	T/C

VAAST score

RANK	Gene	p-value	p-value-ci	Score	Variants
1	ASB12	1.56E-11	1.55557809307134e-11,0.000290464582480396	38.63056297	chrX:63444792;38.63;C->A;G->C;0,3
2	TAF1	1.56E-11	1.55557809307134e-11,0.000290464582480396	34.51696816	chrX:70621541;34.52;T->C;I->T;0,3
3	ZNF41	1.56E-11	1.55557809307134e-11,0.000290464582480396	32.83011803	chrX:47307978;32.83;G->T;D->E;0,3