Taking NGS into the Clinic

Gholson J. Lyon, M.D. Ph.D.

CSH
STANLEY INSTITUTE FOR COGNITIVE GENOMICS
COLD SPRING HARBOR LABORATORY

UTAH FOUNDATION FOR BIOMEDICAL RESEARCH

@GholsonLyon

Conflicts of Interest

- I do not receive salary compensation or any other “donations” from anyone other than my current employer, CSHL .

- Any revenue that I earn from providing medical care in Utah is donated to UFBR for genetics research.

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

Genome Medicine

REVIEW

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

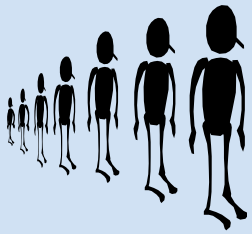
Gholson J Lyon[*1,2] and Kai Wang[*2,3]

A. Probabilistic scoring approach

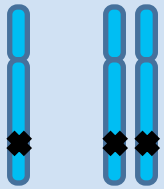
Quality scores for variants

Functional prediction tools

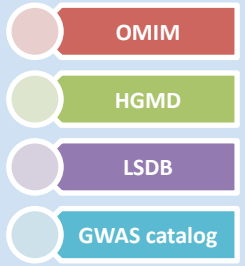
Population allele frequencies



Disease model assumptions



Prior biological knowledge



Statistical model to rank and prioritize all genes

B. Stepwise reduction approach

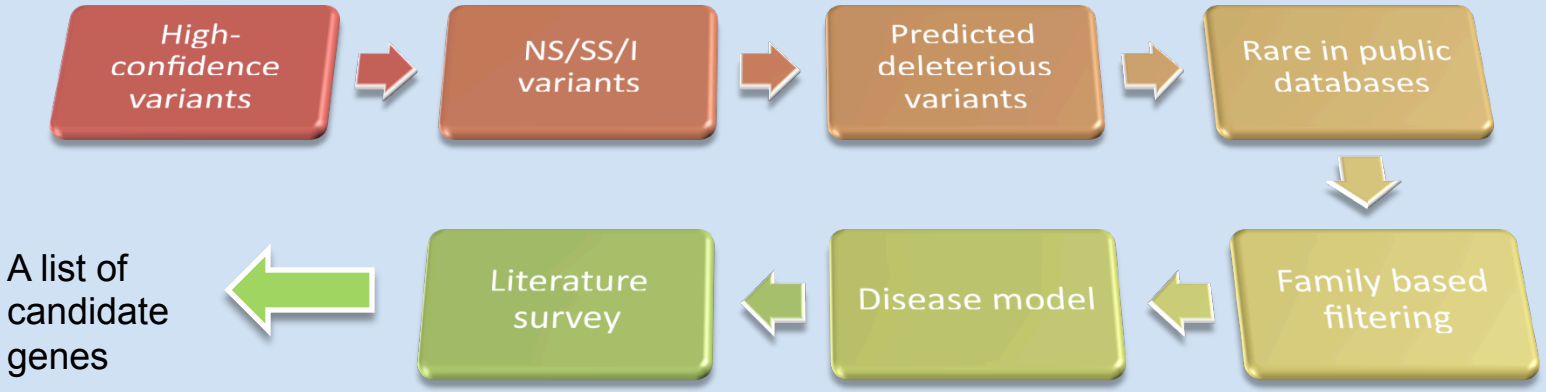


Table 1. Considerations and challenges for the identification of disease causal mutations

Considerations		Challenges	Solutions
Mutation detection	Platform selection	Different sequencing platforms have variable error rates	Increased sequencing coverage for platforms with high error rates
	Sequencing target selection	Exome sequencing may miss regulatory variants that are disease causal	Use whole genome sequencing when budget is not a concern, or when diseases other than well-studied classical Mendelian diseases are encountered
	Variant generation	Genotype calling algorithms differ from each other and have specific limitations	Use multiple alignment and variant calling algorithms and look for concordant calls. Use local assembly to improve indel calls
	Variant annotation	Multiple gene models and multiple function prediction algorithms are available	Perform comprehensive set of annotations and make informed decisions; use probabilistic model for ranking genes/variants
	Variant validation	Predicted disease causal mutations may be false positives	Secondary validation by Sanger sequencing or capture-based sequencing on specific genes/regions
Type of mutations	Coding and splice variants	Many prediction algorithms are available	Evaluate all prediction algorithms under different settings. Develop consensus approaches for combining evidence from multiple algorithms
	Untranslated region, synonymous and non-coding variants	Little information on known causal variants in databases such as HGMD	Improved bioinformatics predictions using multiple sources of information (ENCODE data, multispecies conservation, RNA structure, and so on)
Specific application areas	Somatic mutations in cancer	Tissues selected for sequencing may not harbor large fractions of cells with causal mutations due to heterogeneity; variant calling is complicated by stromal contamination; current databases on allele frequencies do not apply to somatic mutations; current function prediction algorithms focus on loss-of-function mutations	Sample several tissue locations for sequencing; utilize algorithms specifically designed for tumor with consideration for heterogeneity; use somatic mutation databases such as COSMIC; develop function prediction algorithms specifically for gain-of-function mutations in cancer-related genes/pathways
	Non-invasive fetal sequencing	Variants from fetal and maternal genomes need to be teased apart; severe consequences when variants are incorrectly detected and predicted to be highly pathogenic	Much increased sequence depth and more sophisticated statistical approaches that best leverage prior information for inferring fetal alleles; far more stringent criteria to predict pathogenic variants
Inheritance pattern	Inherited from affected parents	Rare/private mutations may be neutral	Evaluate extended pedigrees and 'clans' to assess the potential role of private variants
	De novo mutations from unaffected parents	Every individual is expected to carry three *de novo* mutations, including about one amino acid altering mutation per newborn	Detailed functional analysis of the impacted genes
Biological validation	Known disease causal genes	Difficult to conclude causality when a mutation is found in a well-known disease causal gene	Examine public databases such as locus-specific databases
	Previously characterized genes not known to cause the disease of interest	Relate known molecular function to phenotype of interest	Evaluate loss of function by biochemical assays where available
	Genes without known function	Difficult to design functional follow-up assays	Evaluate gene expression data. Use model organisms to recapitulate the phenotype of interest
Statistical validation	Rare diseases	Limited power to declare association	Sequence candidate genes in unrelated patients to identify additional causal variants
	Idiopathic diseases	Lack of additional unrelated patients	Comprehensive functional follow-up of the biospecimens from patients to prove causality
	Mendelian diseases or traits	Finding rare, unrelated individuals with same phenotype and same mutation to help prove causality	Networking of science through online databases can help find similarly affected people with same phenotype and mutation
Type of phenotypes	Mendelian forms of complex diseases or traits	Several major-effect mutations may work together to cause disease	Statistical models of combined effects (additive and epistatic) of multiple variants within each individual
	Complex diseases or traits	Many variants may contribute to disease risk, each with small effect sizes	Refrain from making predictions unless prior evidence suggested that such predictive models are of practical utility (for example, receiver operating characteristic >0.8)

HGMD, Human Gene Mutation Database.

Table 2. A list of open-access bioinformatics software tools or web servers that can perform batch annotation of genetic variants from whole-exome/genome sequencing data*

Tool	URL	Description	Features	Limitations
ANNOVAR	[http://www.openbioinformatics.org/annovar/]	A software tool written in Perl to perform gene-based, region-based and filter-based annotation	Rapid and up-to-date annotations for multiple species; thousands of annotation types are supported	Requires format conversion for VCF files; command line interface cannot be accessed by many biologists
AnnTools	[http://anntools.sourceforge.net/]	A software tool written in Python to annotate SNVs, indels and CNVs	Fast information retrieval by MySQL database engine; output in VCF format for easy downstream processing	Only supports human genome build 37; does not annotate variant effect on coding sequence
Mu2a	[http://code.google.com/p/mu2a/]	A Java web application for variant annotation	Web interface for users with limited bioinformatics expertise; output in Excel and text formats	Does not allow annotation of indels or CNVs
SeattleSeq	[http://snp.gs.washington.edu/SeattleSeqAnnotation/]	A web server that provides annotation on known and novel SNPs	Multiple input formats are supported; users can customize annotation tasks	Limited annotation on indels or CNVs
Sequence Variant Analyzer	[http://www.svaproject.org/]	A graphical Java software tool to annotate, visualize and organize variants	Intuitive graphical user interface; ability to prioritize candidate genes from multiple patients	Functionality is not very customizable; depends on ENSEMBL database for annotations
snpEff	[http://snpeff.sourceforge.net]	A command-line software tool to calculate the effects of variants on known genes such as amino acid changes	Rapid annotation on multiple species and genome builds; supports multiple codon table	Only supports gene-based annotation
TREAT	[http://ndc.mayo.edu/mayo/research/biostat/stand-alone-packages.cfm]	A command-line software tool with rich integration of publicly available and in-house developed annotations	An Amazon Cloud Image is available for users with limited bioinformatics infrastructure; offers a complete set of pipelines to process FASTQ files and generates annotation outputs	Only supports ENSEMBL gene definition and with limited sets of annotations
VAAST	[http://www.yandell-lab.org/software/vaast.html]	A command-line software tool implementing a probabilistic disease-gene finder to rank all genes	Prioritize candidate genes for Mendelian and complex diseases	Main focus is disease gene finding with limited set of annotations
VARIANT	[http://variant.bioinfo.cipf.es]	A Java web application for variant annotation and visualization	Intuitive interface with integrated genome viewer	Highly specific requirement for internet browser; slow performance
VarSifter	[http://research.nhgri.nih.gov/software/VarSifter/]	A graphical Java program to display, sort, filter and sift variation data	Nice graphical user interface; allows interaction with Integrative Genomics Viewer	Main focus is variant filtering and visualization with limited functionality in variant annotation
VAT	[http://vat.gersteinlab.org/]	A web application to annotate a list of variants with respect to genes or user-specified intervals	Application can also be deployed locally; can generate image for genes to visualize variant effects	Requires multiple other packages to work; only supports gene-based annotation by GENCODE
wANNOVAR	[http://wannovar.usc.edu/]	A web server to annotate user-supplied list of whole genome or whole exome variants with a set of pre-defined annotation tasks	Easy-to-use interface for users with limited bioinformatics skills	Limited set of annotation types are available

*Tools that are only commercially available (such as CLC Bio, Omicia, Golden Helix, DNANexus and Ingenuity) or are designed for a specific type of variant (such as SIFT server and PolyPhen server) are not listed here. CNV, copy number variation; SNP, single nucleotide polymorphism; SNV, single nucleotide variation; VCF, variant call format.

Some Definitions …

- The words “penetrance” and “expressivity” are throwbacks to the era of Drosophila genetics, defined classically as:
- Penetrance: whether someone has ANY symptoms of a disease, i.e. all or none, 0% or 100%. **Nothing in between.**
- Expressivity: how much disease (or how many symptoms) someone with 100% penetrance has.
- This has led to endless confusion!
- Some just use the word “penetrance” to mean the expressivity of disease, i.e. incomplete penetrance, and I agree with combining the two terms into ONE word with the full expression from 0-100% of phenotypic spectrum.

Definitions. It is unknown what portion of "complex" disease will be oligogenic vs. polygenic

- **Oligogenic** – multiple mutations together contributing to aggregate disease, BUT with only 1-2 mutations of ~ >10% penetrance (or "effect size) in EACH person in that clan.
- **Polygenic** – Dozens to hundreds of mutations in different genes in the SAME person, together contributing to the disease in the SAME person, hence **additive** and/or **epistatic** contribution with ~0.01-1% penetrance for each mutation.

Results from Exome and WGS requires both Analytic and Clinical Validity

- Analytical Validity: the test is accurate with high sensitivity and specificity.

- Clinical Validity: Given an accurate test result, what impact and/or outcome does this have on the individual person?

Penetrance Issues

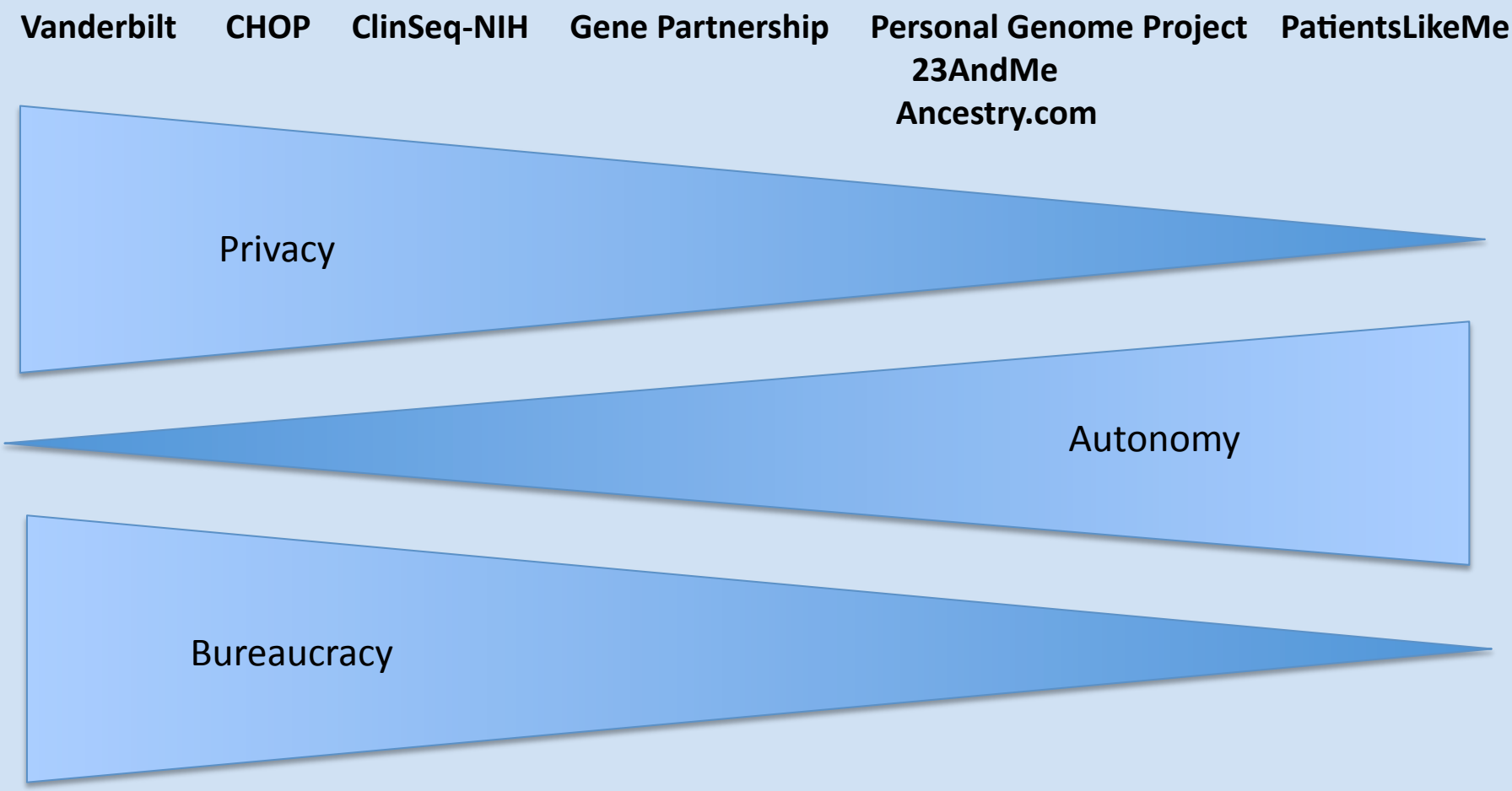
- We do not really know the penetrance of pretty much ALL mutations in **humans**, as we have not systematically sequenced or karyotyped any genetic alteration in **Thousands to Millions** of **randomly** selected people, nor categorized into ethnic classes, i.e. clans.

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

Analytical Validity of Exome and WGS?

- Minimal Standard: exomes and genomes ought to be performed in a CLIA-certified environment for germline genomic DNA from live humans .
- Easier said than done in academia, but some companies offer this now: Illumina, 23andMe, Ambry Genetics, and some academic places do offer this now: UCLA, Baylor, Emory and WashU for exomes.
- I do NOT think the FDA should get involved to regulate this, nor do the results have to go through a physician, i.e. DTC is fine as long as CLIA-certified. This is genetic INFORMATION, not cyanide, some other drug, or surgery.

Autonomy vs. Privacy vs. Bureaucracy



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.

PatientsLikeMe

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."

VAAST shows that probabilistic ranking will be very useful going forward

- But, VAAST is currently dependent on the variant lists provided to it, as there is still a heuristic threshold with input of variant data, i.e. no probabilistic weighting of SNV or indel “true positive likelihood”.

- Therefore, currently need to optimize variant-calling to make sure variants provided are correct. Plus, VAAST chokes if background genomes are full of false positives.

- Thus, focused now on comprehensive comparison of NGS variant-calling on deep exome sequencing data

CLIA-certified exomes and WGS

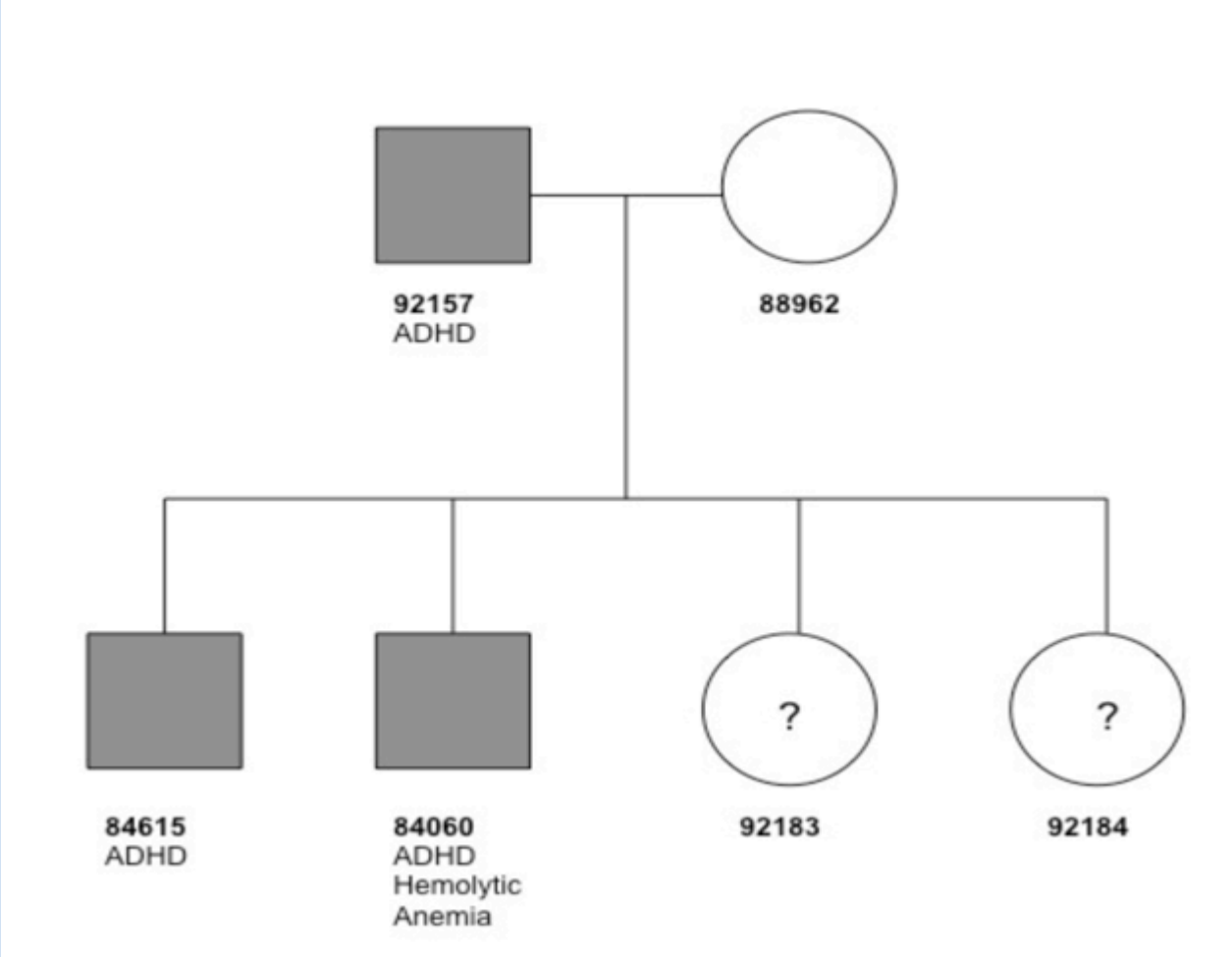
- The CLIA-certified pipelines attempt to minimize false positives with increased depth of sequencing, although there can still be many no-calls and other areas of uncertainty, which should be reported as No-Call Regions.
- This will minimize false positives and also tend to prevent false negatives.

Exome Sequencing and Unrelated Findings in the Context of Complex Disease Research: Ethical and Clinical Implications

Gholson J. Lyon, Tao Jiang, Richard Van Wijk, Wei Wang, Paul Mark Bodily, Jinchuan Xing, Lifeng Tian, Reid J. Robison, Mark Clement, Lin Yang, Peng Zhang, Ying Liu, Barry Moore, Joseph T. Glessner, Josephine Elia, Fred Reimherr, Wouter W. van Solinge, Mark Yandell, Hakon Hakonarson, Jun Wang, William Evan Johnson, Zhi Wei, and Kai Wang

Discov Med. 2011 Jul;12(62):41-55.

Exome sequencing of one pedigree in a research setting.



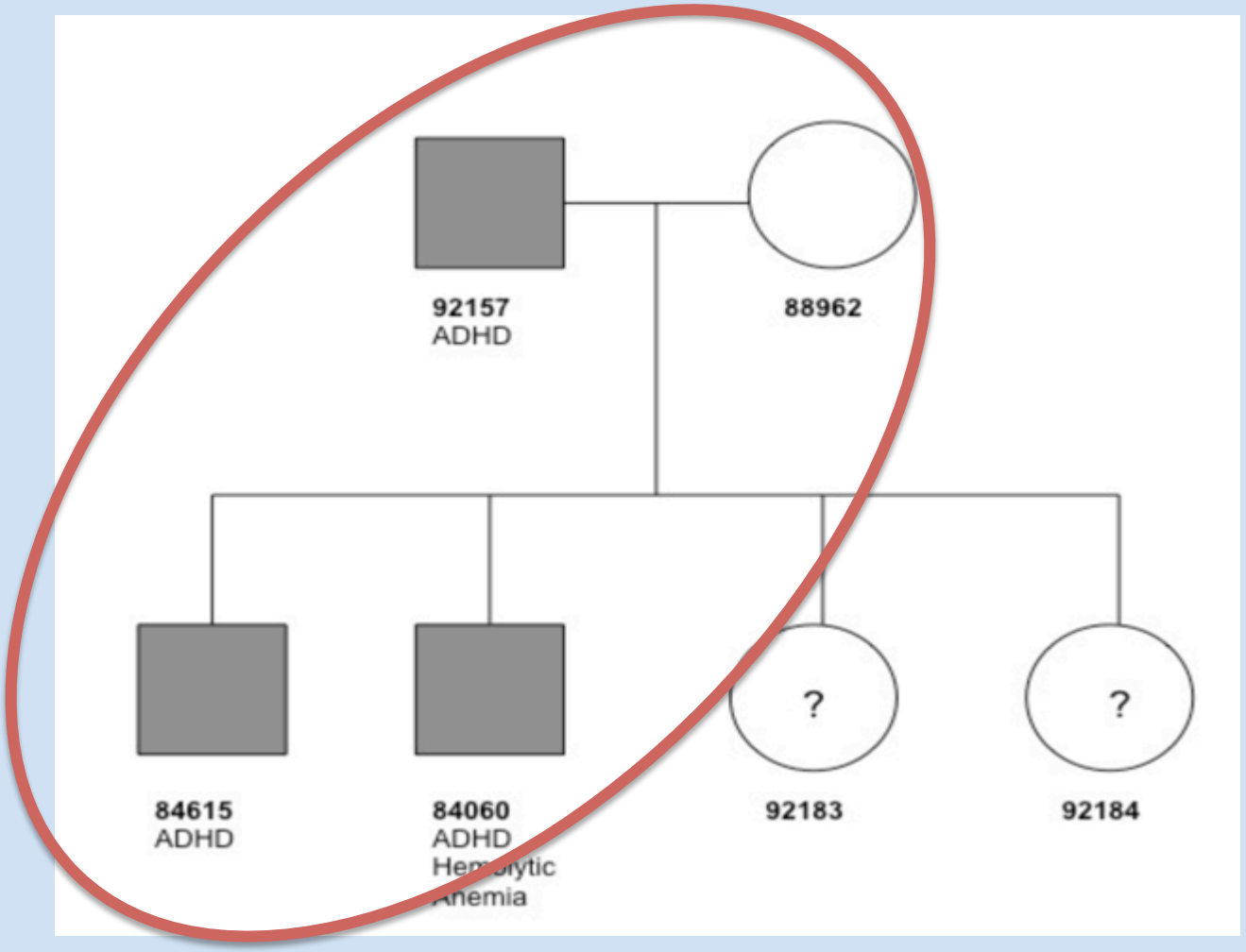
Phenotyping is Critically Important in Neuropsychiatric Disorders!

Supplementary Table 1. ADHD measures during a clinical trial of methylphenidate transdermal system.

		92157	84060	84615
Baseline				
	WRAADDS	16	22	16
	ODD	1	11	7
	CAARS	40	55	38
	CGI-S	4	4	4
Active Medication				
	WRAADDS	0	4	3
	ODD	0	1	3
	CAARS	10	0	13
	CGI-I	1	1	1
	CGI-S	1	3	2
Placebo				
	WRAADDS	15	24	20
	ODD	6	8	7
	CAARS	33	51	42
	CGI-I	4	4	N/A
	CGI-S	4	5	N/A

WRAADDS: Total score on the Wender Reimherr Adult ADD Scale
ODD: Oppositional Defiant Disorder scaore on the WRAADDS ODD subscale
CAARS: Total score Connor's Adult ADHD Rating Scale
CGI-S: Clinical Global Impression, Severity score.

Exome sequencing of one pedigree in a research setting.



Exome method used ~January 2010 with BGI

- Exome capture for the three males was carried out in January 2010 using the commercially available Agilent SureSelect Human All Exon v1 38 MB in solution method as per the manufacturer guidelines (Agilent).
- The DNA from the unaffected mother was obtained at a later date, allowing us to use the newly released SureSelect Human All Exon v. 2 Kit, which targets approximately 44 Mb, covering 98.2% of the CCDS database.
- Paired end sequencing was performed using the Illumina Genome Analyzer IIx platform with read lengths of 76 base pairs, providing at least 20x average coverage at the targeted region. The unaffected mother was sequenced with read lengths of 90 base pairs due to technological advancements during the course of the study, at an average coverage of 30x at the targeted region.

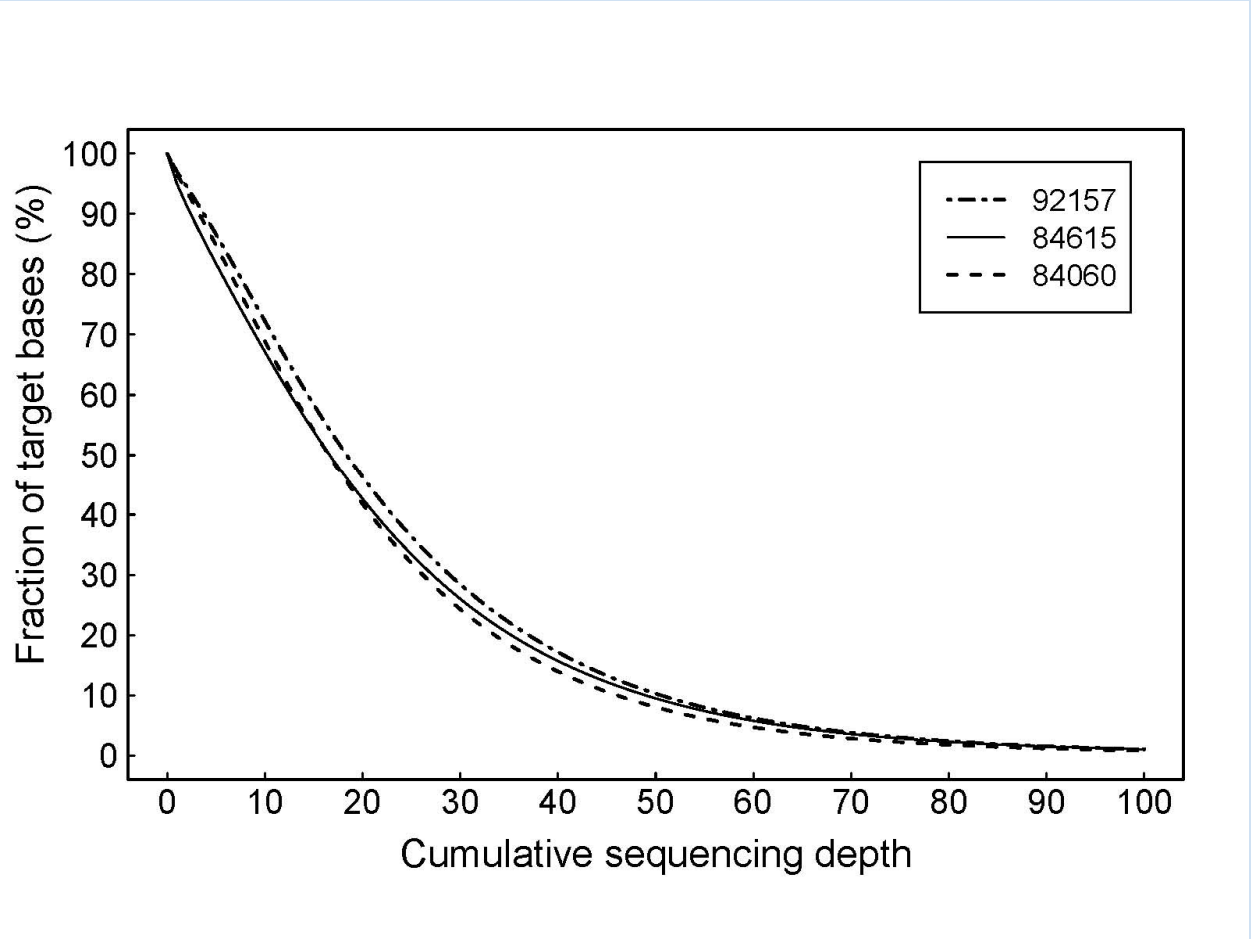
Supplementary Table 2. Summary of data production and evenness for samples.

Exon Capture	84615	84060	92157
Initial bases on target	37,806,033	37,806,033	37,806,033
*Initial bases near target	126,431,894	126,431,894	126,431,894
Initial bases on or near target	164,237,927	164,237,927	164,237,927
**Total effective reads	18,578,623	18,978,287	19,437,592
Total effective yield (Mb)	1,374.80	1,394.45	1,428.19
Average read length (bp)	74.00	73.48	73.48
Effective sequence on target(Mb)	831.55	807.17	890.49
Effective sequence near target(Mb)	259.93	290.95	240.09
Effective sequence on or near target(Mb)	1,091.48	1,098.12	1,130.57
Fraction of effective bases on target	60.50%	57.90%	62.4%
Fraction of effective bases on or near target	79.40%	78.70%	79.2%
Average sequencing depth on target	22.00	21.35	23.55
Average sequencing depth near target	2.06	2.30	1.90
Mismatch rate in target region	0.28%	0.27%	0.28%
Mismatch rate in all effective sequence	0.29%	0.28%	0.30%
Base covered on target	35,919,196	36,523,196	36,676,340
Coverage of target region	95.00%	96.60%	97.0%
Base covered near target	44,578,612	50,837,058	44,482,108
Coverage of flanking region	35.30%	40.20%	35.2%
Fraction of target covered with at least 20X	42.60%	41.80%	46.3%
Fraction of target covered with at least 10X	67.20%	68.90%	72.3%
Fraction of target covered with at least 4X	84.90%	87.90%	89.4%
Fraction of flanking region covered with at least 20X	1.90%	2.10%	1.6%
Fraction of flanking region covered with at least 10X	6.50%	7.20%	5.7%
Fraction of flanking region covered with at least 4X	15.90%	18.10%	14.8%

Supplementary Table 3. Exome sequencing for mother, K24510-88962

Exome Capture Statistics	K24510-88962
Target region (bp)	46,401,121
Raw reads	33,218,260
Raw data yield (Mb)	2,990.00
Reads mapped to genome	28,985,053
Reads mapped to target region	21,076,479
Data mapped to target region (Mb)	1,585.28
Mean depth of target region	**34.16**
Coverage of target region (%)	**95.51**
Average read length (bp)	89.57
Rate of nucleotide mismatch (%)	0.42
Fraction of target covered >=4X	86.58
Fraction of target covered >=10X	75.02
Fraction of target covered >=20X	58.39
Fraction of target covered >=30X	43.35
Capture specificity (%)	72.97
Reads mapped to flanking region	3,915,627
Mean depth of flanking region	9.29
Coverage of flanking region (%)	81.53
Fraction of flanking covered >=4X	54.69
Fraction of flanking covered >=10X	30.11
Fraction of flanking covered >=20X	13
Fraction of flanking covered >=30X	6.74
Fraction of unique mapped bases on or near target	85.42
Duplication rate	7.30
Mean depth of chrX	47.98
Mean depth of chrY	5.36
GC rate	48.28
Gender test result	F

Note:

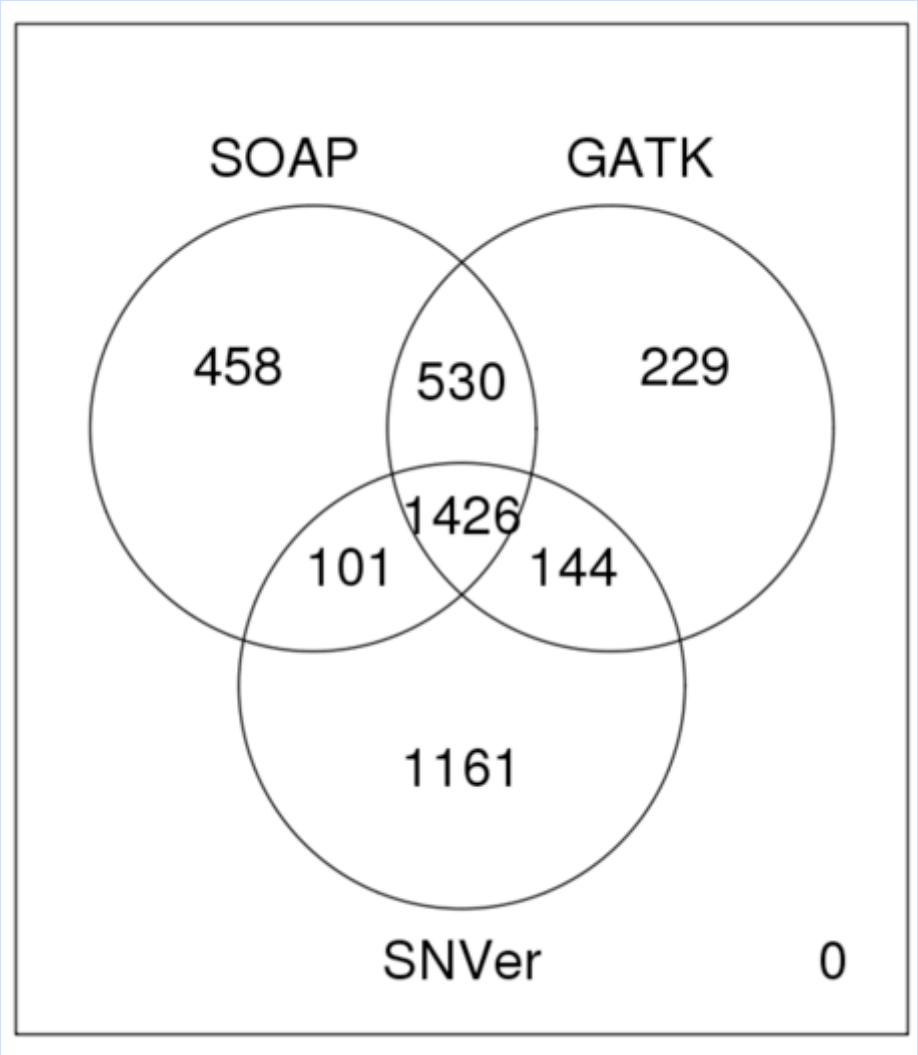


Suppl. Figure 2. Cumulative depth distribution in target regions for three samples. X-axis denotes sequencing depth, and y-axis indicated the fraction of bases that achieves at or above a given sequencing depth. From the figure above, we can see at least 67% of target region bases obtain at least 10x fold coverage in three exomes and more than 85% of target region achieved at least 4x, which shows that the three exomes have similar enrichment uniformity.

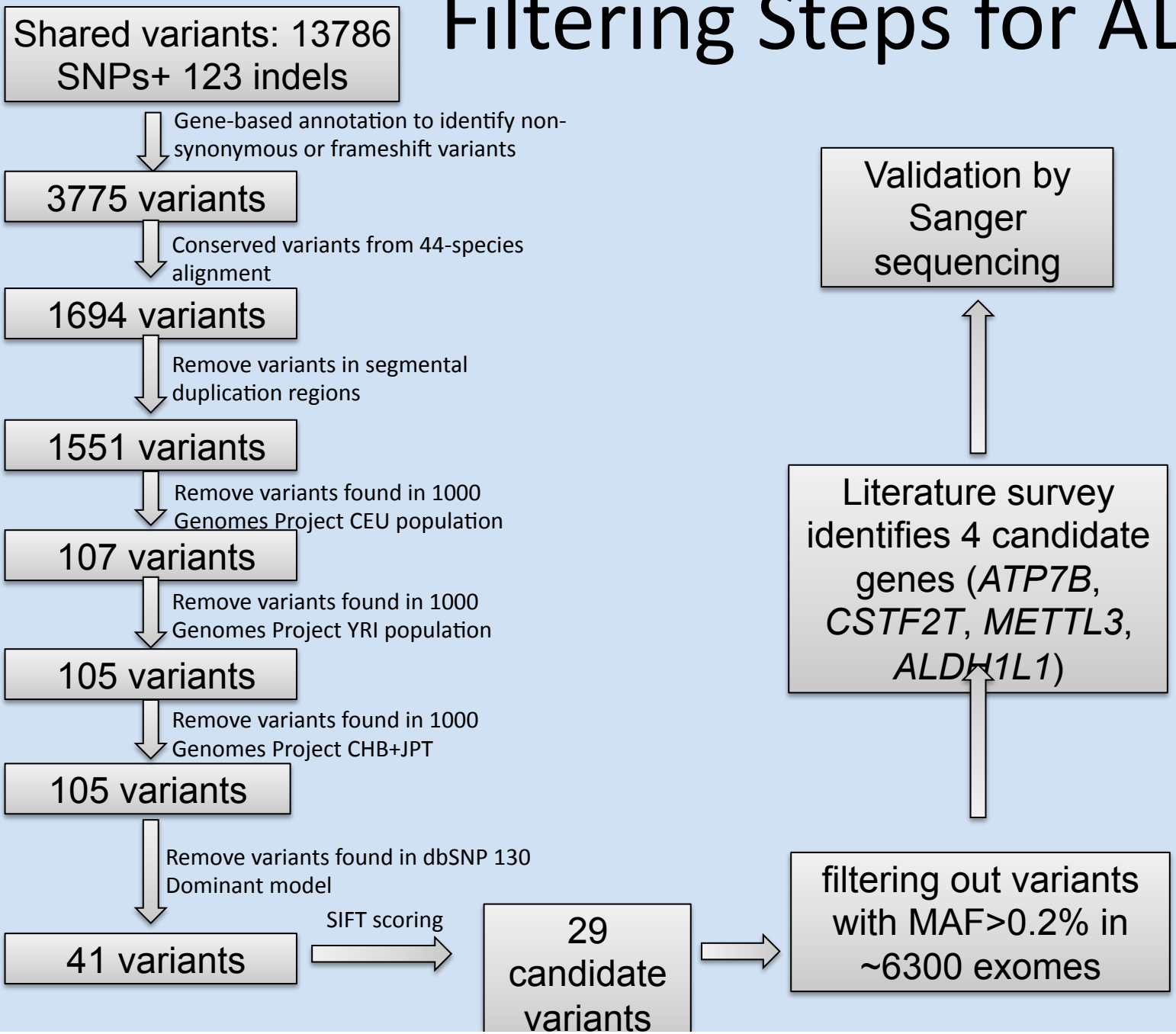
Bioinformatics Analysis for ADHD pedigree

Table 1. Summary of SNVs for exome capture samples				
ExomeCapture	84060 (child 1)	84615 (child 2)	92157 (father)	88962 (mother)
Sequencing platform	GA IIX	GAIIX	GAIIX	HiSeq 2000
Reads property	76bp PE	76bp PE	76bp PE	90bp PE
Number of SNVs (Method 1: SOAP)	19825	19270	20430	22294
Ti/Tv ratio	2.8	2.7	2.9	2.8
Number of SNVs+indels (Method 2: BWA+GATK)	19655+947	18892+955	20100+916	21572+513
Ti/Tv ratio	2.9	2.9	3.0	2.9
Number of SNVs (Method 3: Shrimp2+SNVer)	16063	16704	18253	23917
Ti/Tv ratio	2.7	2.6	2.7	2.4
*We have not yet analyzed the mother's exome with the 4th method (GNUMAP), so we have omitted this method from the table.				

Poor concordance: Intersection of variants. We show here the variants identified by the three main pipelines as being present in the three males with ADHD, but not present in the unaffected mother.



Filtering Steps for ADHD



Supplementary Table 6. Validated variants for ADHD and their population frequency in 5,680 and ~600 deep-sequenced exomes at BGI and Baylor, respectively.

# Chrom.	Position in HG19	Reference allele	Mutant allele	Gene	Type of Mutation	Amino acid change	# variants in BGI exomes[1]	% in BGI exomes	# variants in ~600 Baylor exomes	% in Baylor exomes
chr17	66872692	A	G	ABCA8	Nonsynonymous	C1387R	0	0.0%	0	0.0%
chr11	68566802	G	A	CPT1A	Nonsynonymous	L193F	0	0.0%	0	0.0%
chr8	100994274	A	G	RGS22	Nonsynonymous	I1084T	0	0.0%	0	0.0%
chr18	61654247	G	T	SERPINB8	Nonsynonymous	G287V	0	0.0%	0	0.0%
chr1	207200877	-	T	C1orf116	frameshift insertion		34	1.4%	0	0.0%
chr18	29101156	T	G	DSG2	Nonsynonymous	V158G	1	0.0%	1	0.2%
chr3	125877290	**G**	**A**	**ALDH1L1**	**Nonsynonymous**	**P107L**	**2**	**0.0%**	**0**	**0.0%**
chr13	52542680	**A**	**G**	**ATP7B**	**Nonsynonymous**	**V536A**	**1**	**0.0%**	**1**	**0.2%**
chr10	53458646	**A**	**C**	**CSTF2T**	**Nonsynonymous**	**C222G**	**4**	**0.1%**	**1**	**0.2%**
chr14	21972019	**G**	**A**	**METTL3**	**Nonsynonymous**	**R36W**	**9**	**0.2%**	**1**	**0.2%**
chr11	76954790	-	A	GDPD4	frameshift insertion		36	1.5%	6	1.0%
chr7	87160618	A	T	ABCB1	Nonsynonymous	S893T	815	14.3%[1]	9	1.5%
chr11	134128923	C	G	ACAD8	Nonsynonymous	S171C	112	2.0%	20	3.3%
chr20	17956347	C	T	C20orf72	Nonsynonymous	R178W	23	0.4%	8	1.3%
chr8	33318891	T	C	FUT10	Nonsynonymous	Q27R	15	0.3%	3	0.5%
chr13	20797025	A	T	GJB6	Nonsynonymous	S199T	68	1.2%	4	0.7%
chr16	71015329	G	T	HYDIN	Nonsynonymous	P1491H	77	1.4%	dozens	>5.0%
chr10	22019855	G	A	MLLT10	Nonsynonymous	R713H	15	0.3%	6	1.0%
chr17	10415269	A	G	MYH1	Nonsynonymous	Y435H	99	1.7%	14	2.3%
chr1	145015877	G	T	PDE4DIP	Nonsynonymous	L142I	1256	22.1%	hundreds	>30.0%
chr2	98809432	T	C	VWA3B	Nonsynonymous	I513T	15	0.3%	16	2.7%
chr5	115202418	AAGA	-	AP3S1	frameshift deletion		185	7.8%	19	3.2%

1. The indels were only measured thus far in 2,360 exomes at BGI, whereas the SNPs were measured in 5,680 exomes.

Supplementary Table 6. Validated variants for ADHD and their population frequency in 5,680 and ~600 deep-sequenced exomes at BGI and Baylor, respectively.

# Chrom.	Position in HG19	Reference allele	Mutant allele	Gene	Type of Mutation	Amino acid change	# variants in BGI exomes[1]	% in BGI exomes	# variants in ~600 Baylor exomes	% in Baylor exomes
chr17	66872692	A	G	ABCA8	Nonsynonymous	C1387R	0	0.0%	0	0.0%
chr11	68566802	G	A	CPT1A	Nonsynonymous	L193F	0	0.0%	0	0.0%
chr8	100994274	A	G	RGS22	Nonsynonymous	I1084T	0	0.0%	0	0.0%
chr18	61654247	G	T	SERPINB8	Nonsynonymous	G287V	0	0.0%	0	0.0%
chr1	207200877	-	T	C1orf116	frameshift insertion		34	1.4%	0	0.0%
chr18	29101156	T	G	DSG2	Nonsynonymous	V158G	1	0.0%	1	0.2%
chr3	125877290	**G**	**A**	**ALDH1L1**	**Nonsynonymous**	**P107L**	**2**	**0.0%**	**0**	**0.0%**
chr13	52542680	**A**	**G**	**ATP7B**	**Nonsynonymous**	**V536A**	**1**	**0.0%**	**1**	**0.2%**
chr10	53458646	**A**	**C**	**CSTF2T**	**Nonsynonymous**	**C222G**	**4**	**0.1%**	**1**	**0.2%**
chr14	21972019	**G**	**A**	**METTL3**	**Nonsynonymous**	**R36W**	**9**	**0.2%**	**1**	**0.2%**
chr11	76954790	-	A	GDPD4	frameshift insertion		36	1.5%	6	1.0%
chr7	87160618	A	T	ABCB1	Nonsynonymous	S893T	815	14.3%[1]	9	1.5%
chr11	134128923	C	G	ACAD8	Nonsynonymous	S171C	112	2.0%	20	3.3%
chr20	17956347	C	T	C20orf72	Nonsynonymous	R178W	23	0.4%	8	1.3%
chr8	33318891	T	C	FUT10	Nonsynonymous	Q27R	15	0.3%	3	0.5%
chr13	20797025	A	T	GJB6	Nonsynonymous	S199T	68	1.2%	4	0.7%
chr16	71015329	G	T	HYDIN	Nonsynonymous	P1491H	77	1.4%	dozens	>5.0%
chr10	22019855	G	A	MLLT10	Nonsynonymous	R713H	15	0.3%	6	1.0%
chr17	10415269	A	G	MYH1	Nonsynonymous	Y435H	99	1.7%	14	2.3%
chr1	145015877	G	T	PDE4DIP	Nonsynonymous	L142I	1256	22.1%	hundreds	>30.0%
chr2	98809432	T	C	VWA3B	Nonsynonymous	I513T	15	0.3%	16	2.7%
chr5	115202418	AAGA	-	AP3S1	frameshift deletion		185	7.8%	19	3.2%

1. The indels were only measured thus far in 2,360 exomes at BGI, whereas the SNPs were measured in 5,680 exomes.

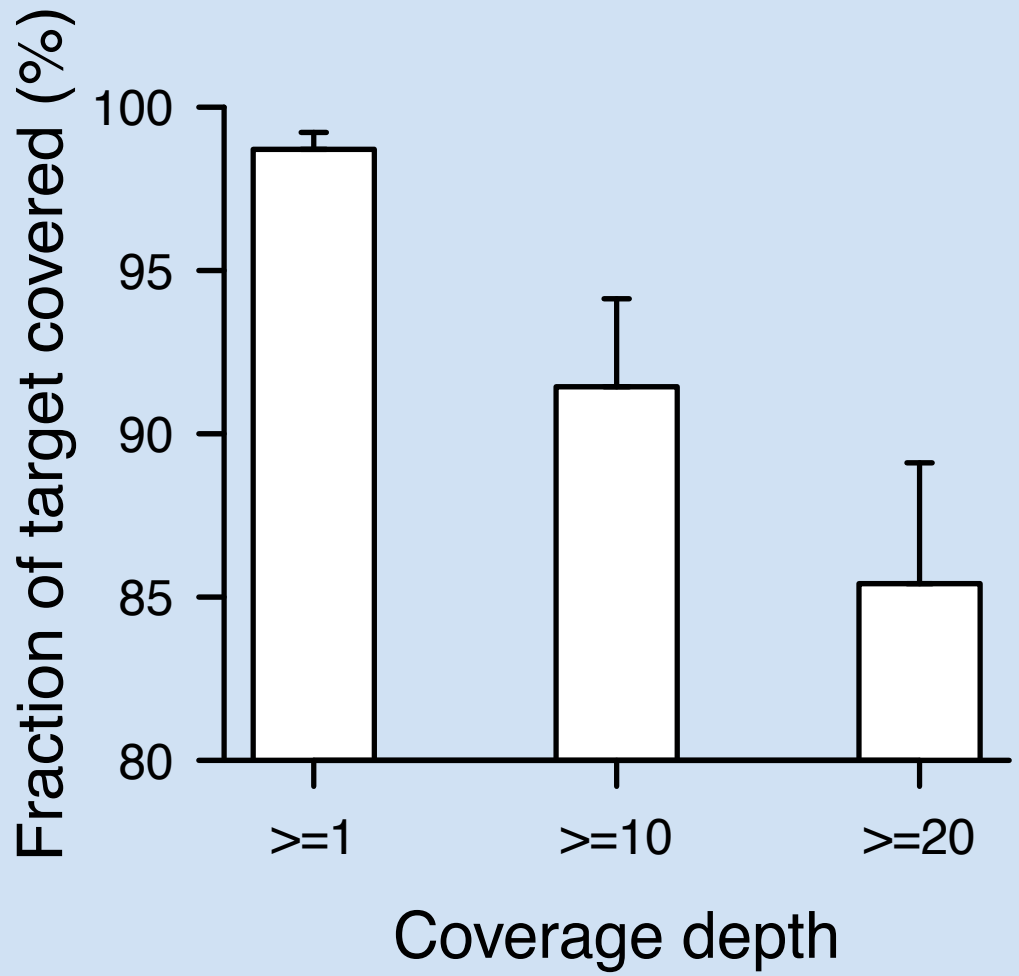
Optimizing Variant Calling in Exomes at BGI in 2011

- Agilent v2 44 MB exome kit
- Illumina Hi-Seq for sequencing.
- Average coverage ~100-150x.
- Depth of sequencing of >80% of the target region with >20 reads or more per base pair.
- Comparing various pipelines for alignment and variant-calling.

2-3 rounds of sequencing at BGI to attain goal of >80% of target region at >20 reads per base pair

Exome Capture Statistics	K24510-84060	K24510-92157-a	K24510-84615	K24510-88962
Target region (bp)	46,401,121	46,401,121	46,401,121	46,257,379
Raw reads	138,779,950	161,898,170	156,985,870	104,423,704
Raw data yield (Mb)	12,490	14,571	14,129	9,398
Reads mapped to genome	110,160,277	135,603,094	135,087,576	83,942,646
Reads mapped to target region	68,042,793	84,379,239	80,347,146	61,207,116
Data mapped to target region (Mb)	5,337.69	6,647.18	6,280.01	4,614.47
Mean depth of target region	**115.03**	**143.25**	**135.34**	**99.76**
Coverage of target region (%)	**0.9948**	**0.9947**	**0.9954**	**0.9828**
Average read length (bp)	89.91	89.92	89.95	89.75
Fraction of target covered >=4X	98.17	98.38	98.47	94.25
Fraction of target covered >=10X	95.18	95.90	95.97	87.90
Fraction of target covered >=20X	**90.12**	**91.62**	**91.75**	**80.70**
Fraction of target covered >=30X	84.98	87.42	87.67	74.69
Capture specificity (%)	61.52	62.12	59.25	73.16
Fraction of unique mapped bases on or near target	65.59	65.98	63.69	85.46
Gender test result	M	M	M	F

Depth of Coverage in 15 exomes > 20 reads per bp in target region



Deep Exome sequencing

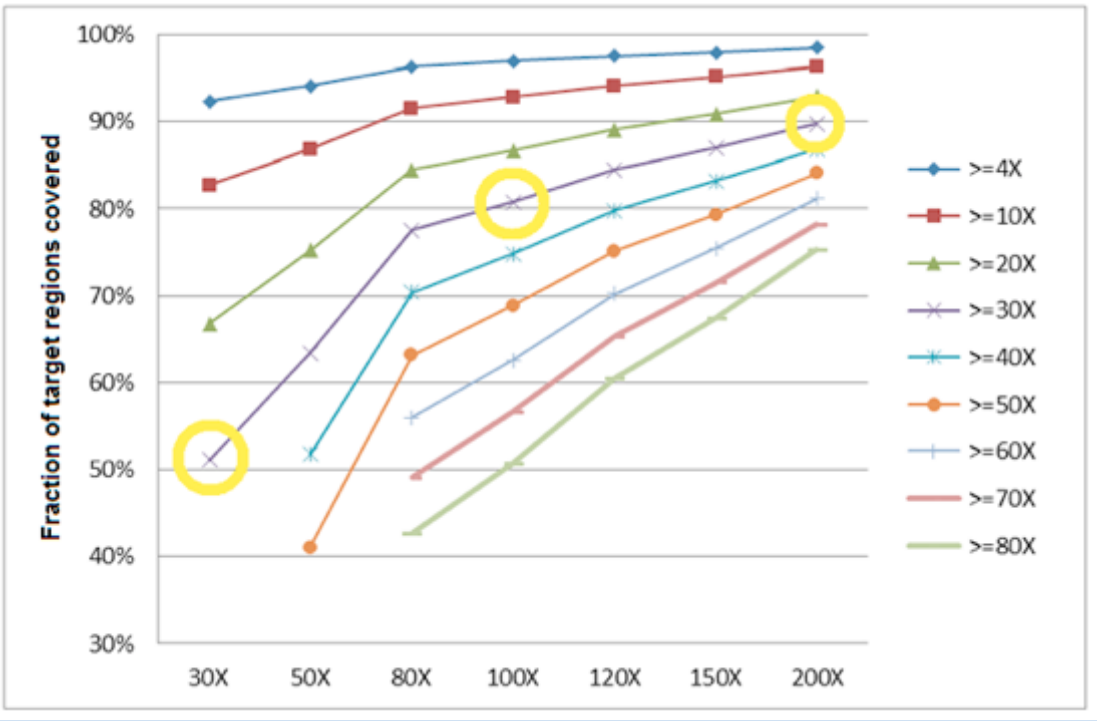


Figure from BGI website: http://bgiamericas.com/news-events/why-deep-exome-sequencing/

Fig.1 Correlation between the percentage of target regions covered and the sequencing depth in human exome sequencing. Take >=30X series (the purple line) for example: when the sequencing depth is 30X, only half of the target regions (51%) are covered at above 30X. While at the 100X and 200X sequencing depths, a much higher percentage (81% and 90%, respectively) of the target regions is covered at above 30X.

GWAS has statistical rigor with a threshold p value

- Should exome sequencing also have a threshold level of rigor, such as >80% of target region with 20 reads or more per base pair?

- This is accepted practice at major genome sequencing centers (Baylor, WashU, Broad), but apparently not everywhere else…. Shouldn't this be required?

“Methods” should really mean something

- Papers should include detailed methods, allowing reproduction of analyses.
- Or, better yet, “papers” should be simply analyses published online, connected to datasets, updateable in Wiki fashion..
- Data should be made available as well, with standardized analyses in place.
- At least there is now some movement toward “open science”.

In a prior project on a new, rare disorder, that we named Ogden Syndrome, the X-chromosome Exon Capture and Coverage was high depth with Average Base Coverage of 214x …

Table 2. Coverage Statistics in Family 1. Based on GNUMAP

Region	RefSeq Transcripts	Unique Exons	Percent Exon Coverage ≥1X	Percent Exon Coverage ≥10X	Unique Genes	Average Base Coverage	VAAST Candidate SNVs
X-chromosome	1,959	7,486	97.8	95.6	913	214.6	1 (*NAA10*)
chrX: 10054434-40666673	262	1,259	98.1	95.9	134	213.5	0
chrX: 138927365-153331900	263	860	97.1	94.9	132	177.1	1 (*NAA10*)

* On chromosome X, there are 8,222 unique RefSeq exons. Of these exons, 736 were excluded from the SureSelect X-Chromosome Capture Kit because they were designated as pseudoautosomal or repetitive sequences (UCSC genome browser).

Using VAAST to Identify an X-Linked Disorder Resulting in Lethality in Male Infants Due to N-Terminal Acetyltransferase Deficiency. Am J Hum Genet. 2011 Jul 15;89(1):28-43. Epub 2011 Jun 23.

Replication is so critically important: “To show that 'A' is true, you don't do 'B'. You do 'A' again.”

Ed Yong, Nature 485, 298–300 (17 May 2012)

- Gave Ogden Syndrome data to Omicia, Golden Helix and Synapse for replication and data upload.
- Replicated already by Omicia and Golden Helix.
- Anyone can download data from Synapse Portal – just email me to gain access to the data.

2-3 rounds of sequencing at BGI to attain goal of >80% of target region at >20 reads per base pair

Exome Capture Statistics	K24510-84060	K24510-92157-a	K24510-84615	K24510-88962
Target region (bp)	46,401,121	46,401,121	46,401,121	46,257,379
Raw reads	138,779,950	161,898,170	156,985,870	104,423,704
Raw data yield (Mb)	12,490	14,571	14,129	9,398
Reads mapped to genome	110,160,277	135,603,094	135,087,576	83,942,646
Reads mapped to target region	68,042,793	84,379,239	80,347,146	61,207,116
Data mapped to target region (Mb)	5,337.69	6,647.18	6,280.01	4,614.47
Mean depth of target region	**115.03**	**143.25**	**135.34**	**99.76**
Coverage of target region (%)	**0.9948**	**0.9947**	**0.9954**	**0.9828**
Average read length (bp)	89.91	89.92	89.95	89.75
Fraction of target covered >=4X	98.17	98.38	98.47	94.25
Fraction of target covered >=10X	95.18	95.90	95.97	87.90
Fraction of target covered >=20X	**90.12**	**91.62**	**91.75**	**80.70**
Fraction of target covered >=30X	84.98	87.42	87.67	74.69
Capture specificity (%)	61.52	62.12	59.25	73.16
Fraction of unique mapped bases on or near target	65.59	65.98	63.69	85.46
Gender test result	M	M	M	F

Pipeline Used on Same Set of Seq Data by Different Analysts

1) BWA-Sam format to Bam format-Picard to remove duplicates- **GATK** (version 1.5) with recommended parameters (GATK IndelRealigner, base quality scores were re-calibrated by GATK Table Recalibration tool. Genotypes called by GATK UnifiedGenotyper.

2) BWA-Sam format to Bam format-Picard to remove duplicates- **SamTools** version 0.1.18 to generate genotype calls -- The "mpileup" command in SamTools were used for identify SNPs and indels.

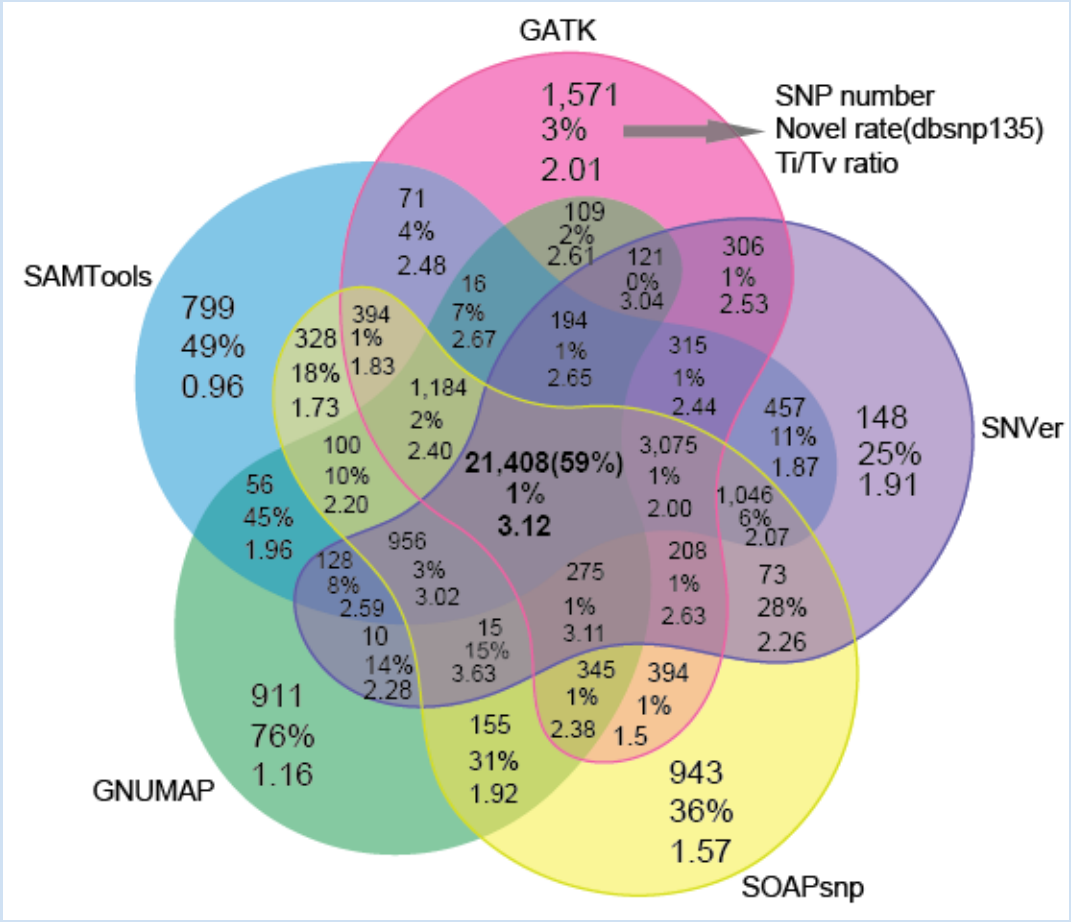
3) SOAP-Align – SOAPsnp – then BWA-SOAPindel (adopts local assembly based on an extended de Bruijn graph)

4) GNUMAP-SNP (probabilistic Pair-Hidden Markov which effectively accounts for uncertainty in the read calls as well as read mapping in an unbiased fashion)

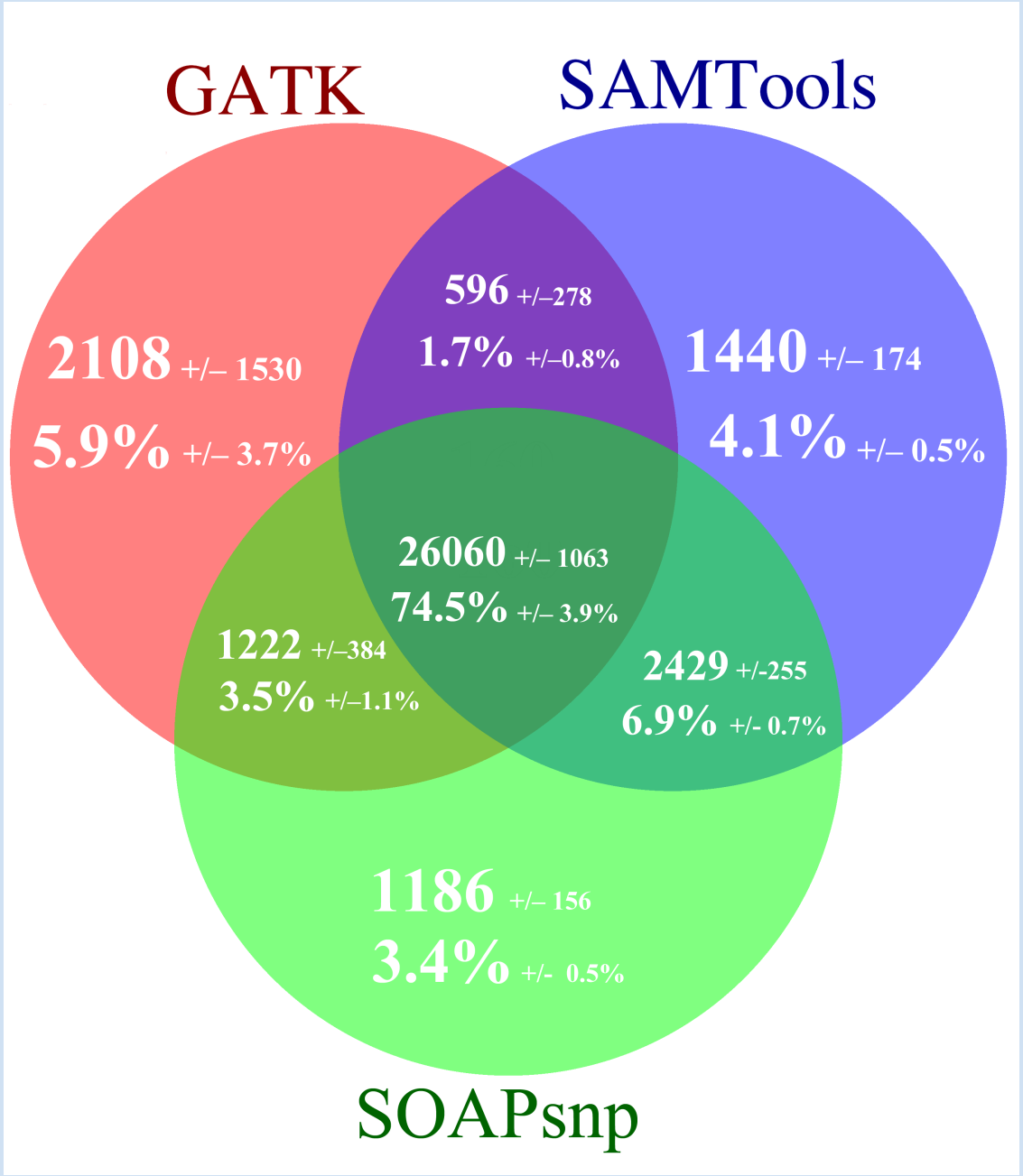
5) BWA-Sam format to Bam format-Picard to remove duplicates- **SNVer**

Total SNVs

A)

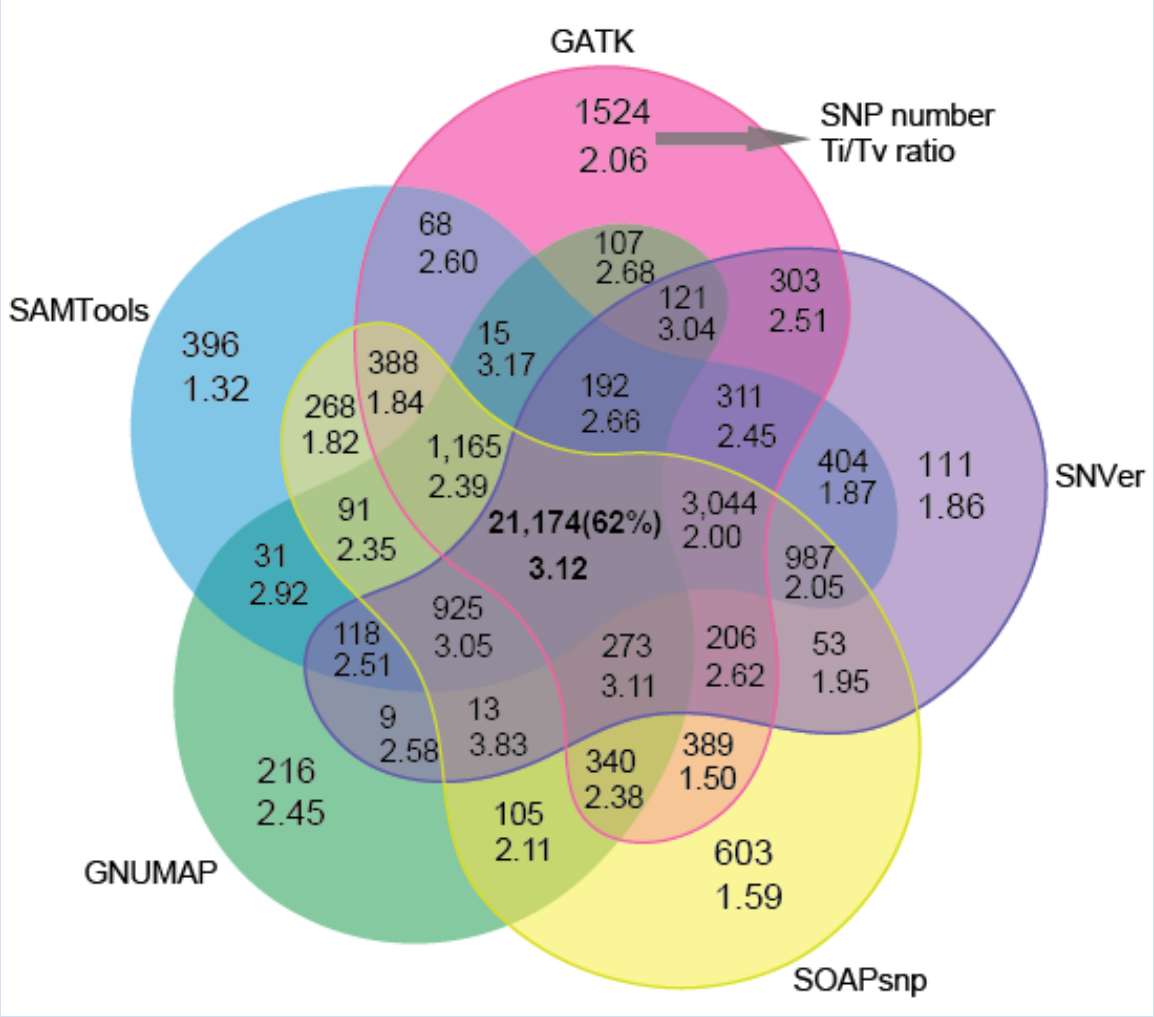


Mean # of total SNVs across 15 exomes, called by 5 pipelines. The percentage in the center of the the Venn diagram(Parenthesis) is the percent of total SNVs called by all five pipelines.


GATK
SAMTools
SOAPsnp
2108 +/– 1530
5.9% +/– 3.7%
596 +/–278
1.7% +/–0.8%
1440 +/– 174
4.1% +/– 0.5%
26060 +/– 1063
74.5% +/– 3.9%
1222 +/–384
3.5% +/–1.1%
2429 +/-255
6.9% +/- 0.7%
1186 +/– 156
3.4% +/- 0.5%

Known SNVs

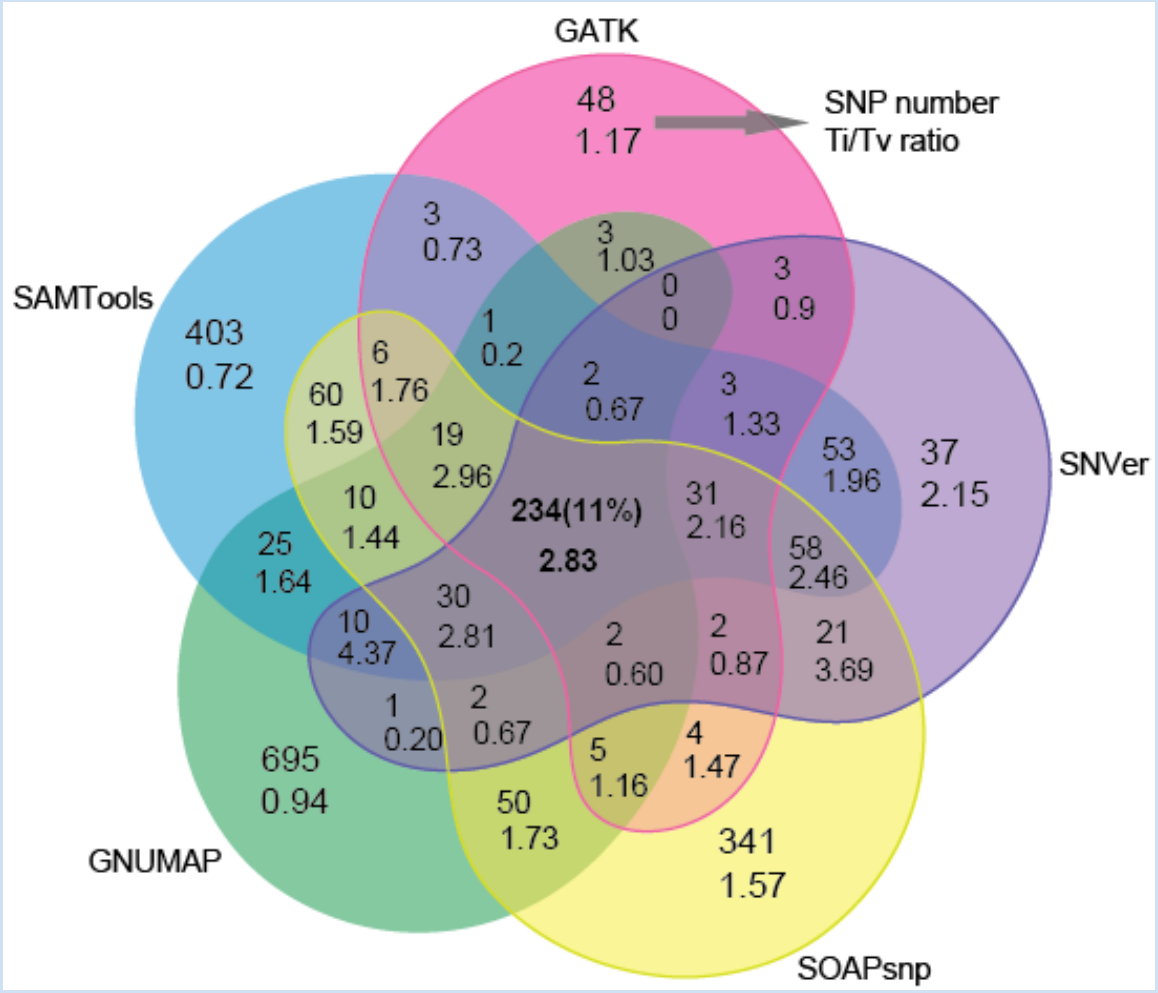
B)



B) Mean # of known SNVs (present in dbSNP135) found by 5 pipelines across 15 exomes. The percentage in the center of the the Venn diagram is the percent of known SNVs called by all five pipelines.

Novel SNVs

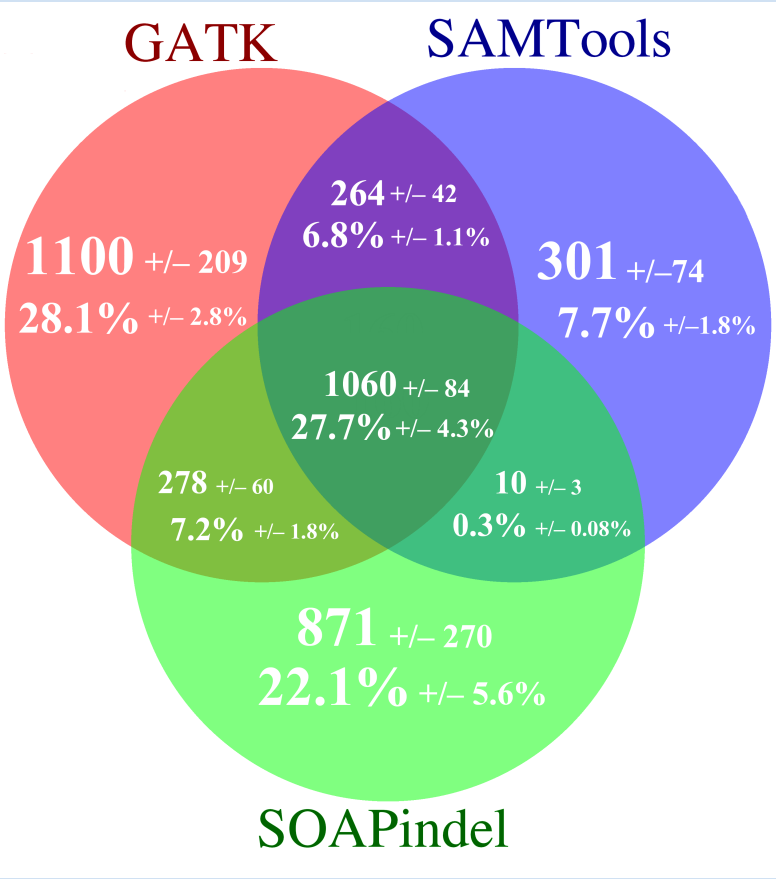
C)



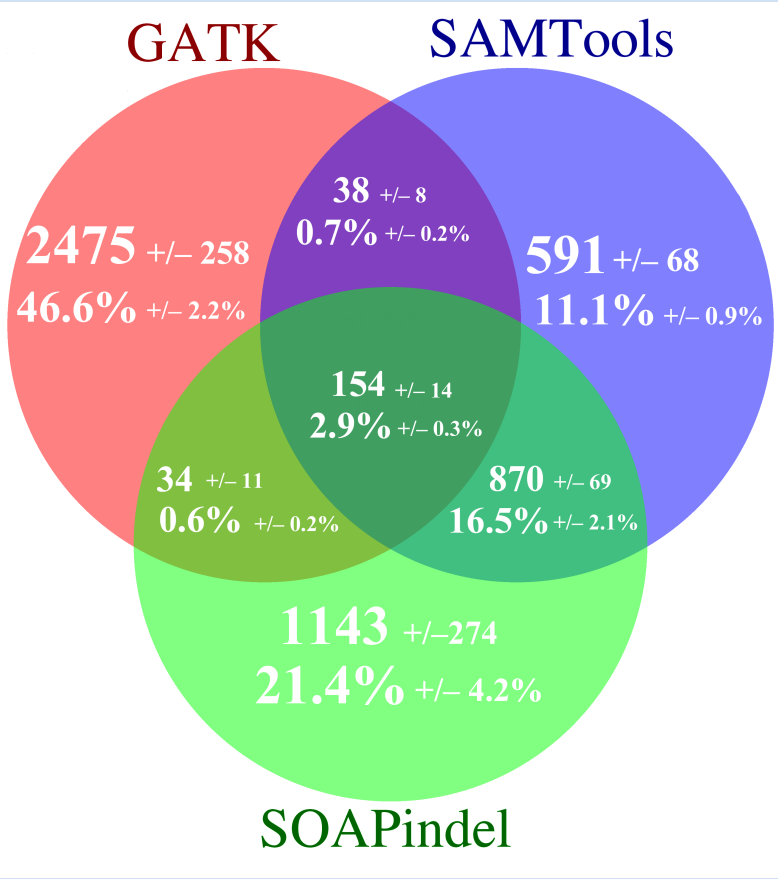
- **C)** Mean # of novel SNVs (not present in dbSNP135) found by 5 pipelines across 15 exomes. The percentage in the center of the Venn diagram is the percent of novel SNVs called by all five pipelines.

INDELS

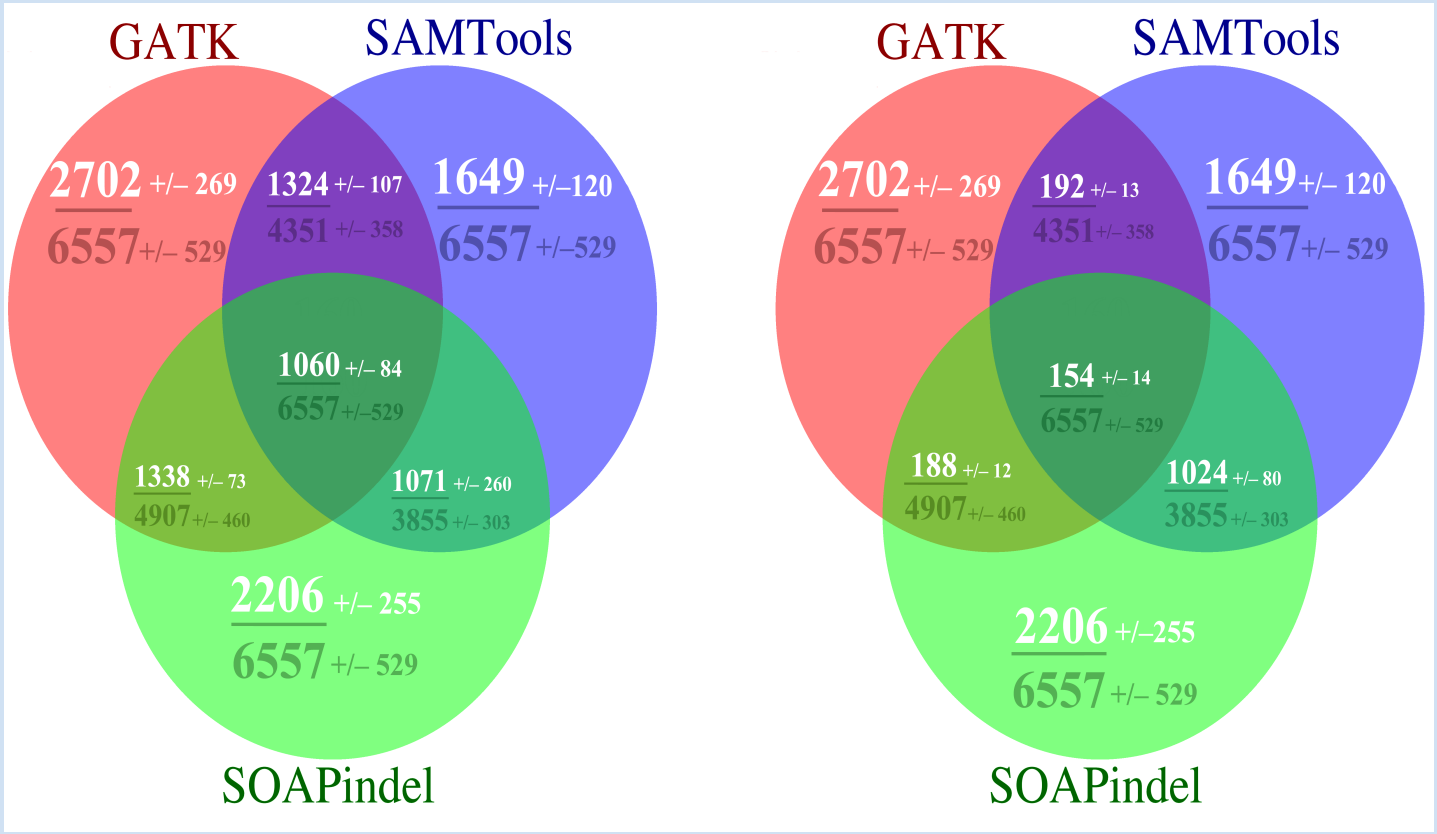
Indels- Overlap by Base Position only



Indels- Overlap by Base Position, Length **and** Composition



Total mean overlap, plus or minus one standard deviation, observed between three indel calling pipelines: GATK, SOAP-indel, and SAMTools. a) Mean overlap when indel position was the only necessary agreement criterion. **b)** Mean overlap when indel position, base length and base composition were the necessary agreement criteria.


GATK
SAMTools
2702 +/– 269
6557 +/– 529
1324 +/– 107
4351 +/– 358
1649 +/–120
6557 +/–529
1060 +/– 84
6557 +/–529
1338 +/– 73
4907 +/– 460
1071 +/– 260
3855 +/– 303
2206 +/– 255
6557 +/– 529
SOAPindel
GATK
SAMTools
2702 +/– 269
6557 +/– 529
192 +/– 13
4351 +/– 358
1649 +/– 120
6557 +/– 529
154 +/– 14
6557 +/– 529
188 +/– 12
4907 +/– 460
1024 +/– 80
3855 +/– 303
2206 +/–255
6557 +/– 529
SOAPindel

Optimizing the Variant Calling Pipeline Using Family Relationships

We looked for SNVs that were detected in children but not in parents using 3 different strategies:

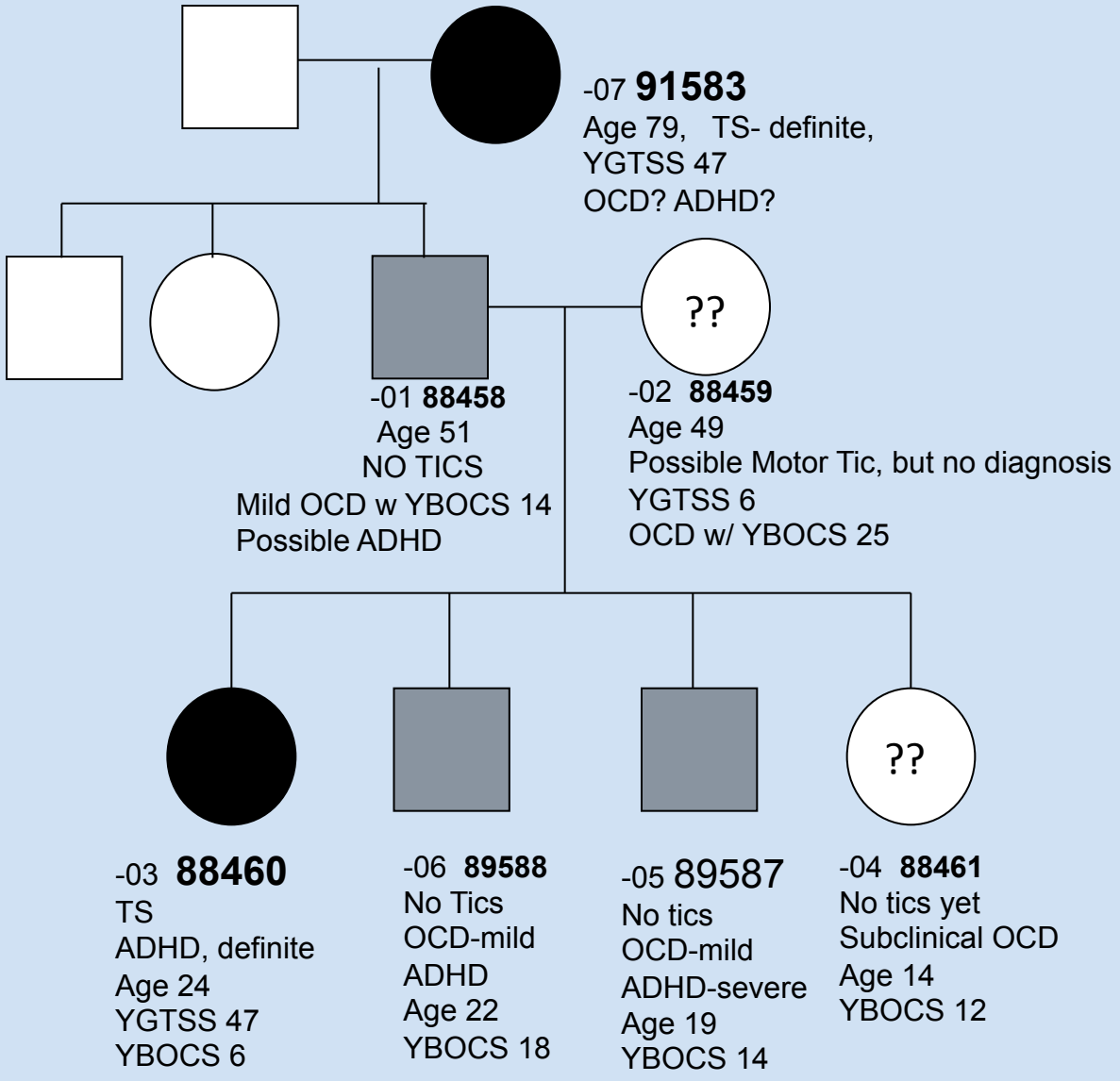
1. We used all of the SNVs that were detected by all 5 pipelines for both parents and children
2. We used all of the detected SNVs for parents, but only the concordant SNVs between the 5 different pipelines for children.
3. We used SNVs concordant between the 5 different pipelines for children and parents.

Optimizing pipeline based on literature value of ~1 true de novo protein-altering mutation per exome

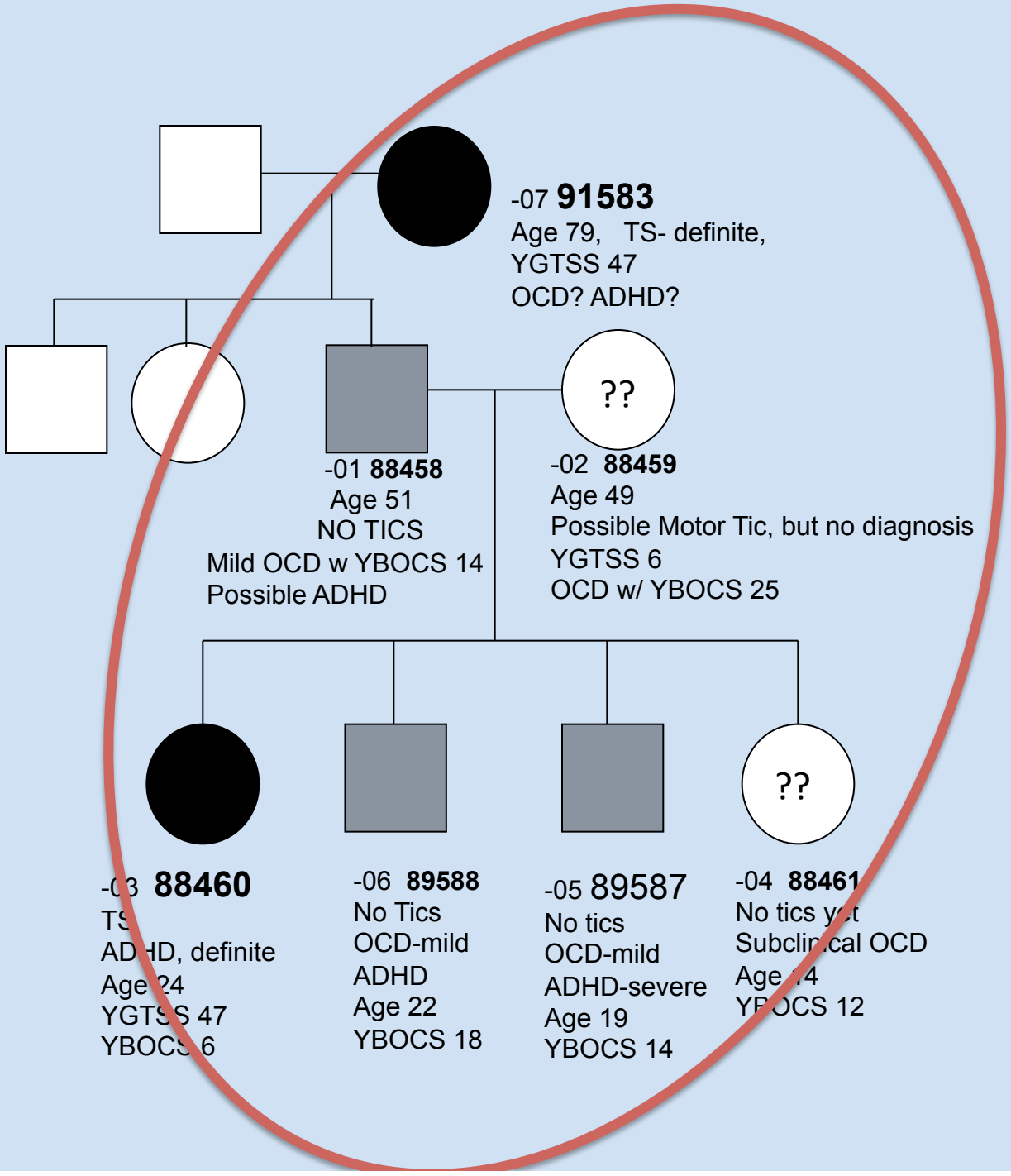
	All SNVs, both for parents and children, were considered	All parental SNVs that were detected were considered. Only SNVs concordant between the 5 pipelines were considered for children	SNVs concordant between 5 pipelines for children and parents
Number of SNVs found in child A but not in parents	1308	186	1795
Number of SNVs found in child B but not in parents	1332	161	1762
Number of nonsyn SNVs in child A but not in parents	381	52	420
Number of nonsyn SNVs in child B but not in parents	392	42	394
Number of shared nonsyn SNVs in the children, but not in parents	98	14	171

The result is that using all of the detected SNVs for both parents and children should minimize the false negative rate but similarly show a relatively high false positive rate. Using all of the SNVs detected for parents but only the SNVs concordant among the five pipelines shows mutation rates similar to those reported by the literature and is expected to have moderate false positive rates and moderate false negative rates. Using only the SNVs concordant among the 5 different pipelines for both parents and children should minimize the false positive rate but similarly show a relatively high false negative rate.

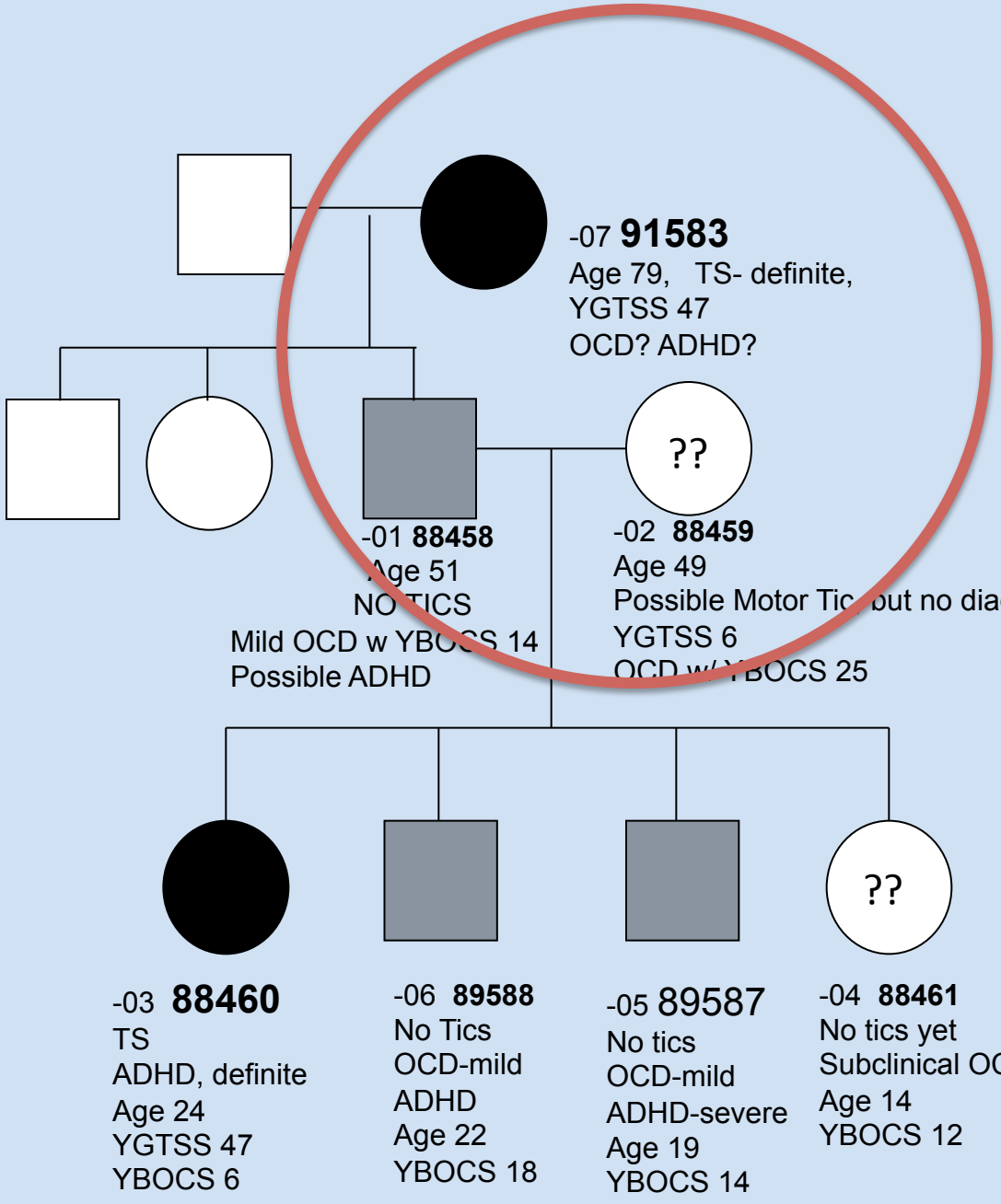
TDT- 09 -1018
K26679



TDT- 09 -1018
K26679



TDT- 09 -1018
K26679

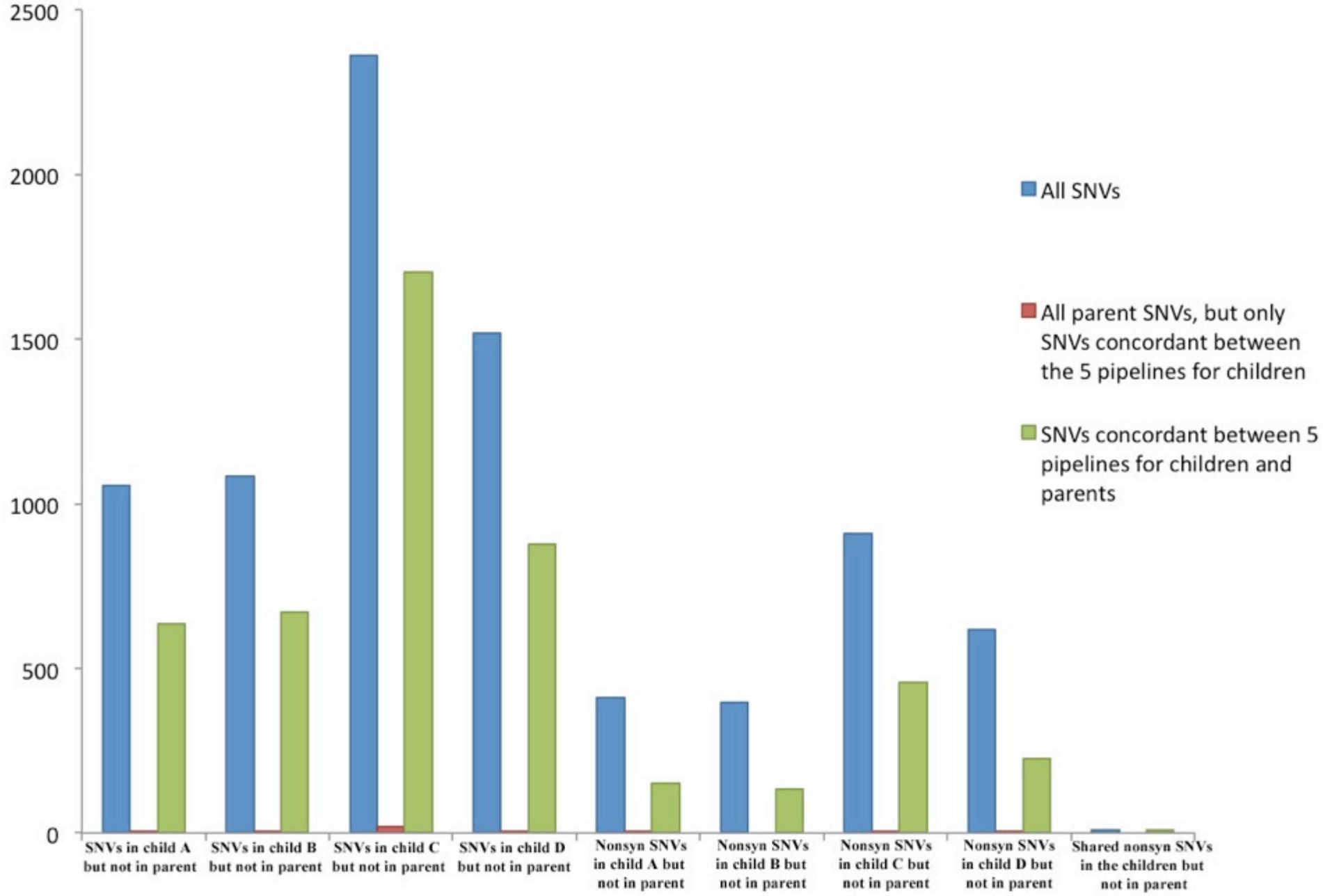


"Parents"

Analysis based on various pipelines

- “Parents” in this case means the mother, father AND grandmother.
- Taking the **Union** of SNVs from all 5 pipelines from “Parents”, and subtract that from the **Union** of all SNVs in each child.
- Or Subtract the **Union** of these “Parents” from the SNVs in the child **concordant** between 5 pipelines.
- Or, subtract the **concordant** variants from 5 pipelines in “Parents” from the **concordant** variants for 5 pipelines in each child.

	All SNVs, both for parents and children, were considered	All parental SNVs that were detected were considered. Only SNVs concordant between the 5 pipelines were considered for children	SNVs concordant between 5 pipelines for children and parents
Number of SNVs found in child A but not in parents	1057	2	637
Number of SNVs found in child B but not in parents	1084	1	672
Number of SNVs found in child C but not in parents	2363	20	1703
Number of SNVs found in child D but not in parents	1518	5	876
Number of nonsyn SNVs in child A but not in parents	411	1	150
Number of nonsyn SNVs in child B but not in parents	396	0	135
Number of nonsyn SNVs in child C but not in parents	911	6	459
Number of nonsyn SNVs in child D but not in parents	619	3	225
Number of shared nonsyn SNVs in the children, but not in parents	8	0	9


2500
2000
1500
1000
500
0
All SNVs
All parent SNVs, but only SNVs concordant between the 5 pipelines for children
SNVs concordant between 5 pipelines for children and parents
SNVs in child A but not in parent
SNVs in child B but not in parent
SNVs in child C but not in parent
SNVs in child D but not in parent
Nonsyn SNVs in child A but not in parent
Nonsyn SNVs in child B but not in parent
Nonsyn SNVs in child C but not in parent
Nonsyn SNVs in child D but not in parent
Shared nonsyn SNVs in the children but not in parent

Preliminary Conclusions

- Sequencing a grandparent seems to help eliminate errors derived from the current depth of sequencing coverage in the mother and father.
- An alternative might be just deeper depth of sequencing in the parents, although still investigating errors that might be overcome by sequencing a grandparent.
- Need to decide on whether to proceed with the concordance of 2 or more pipelines, like SOAP + GATK, or just accept (with everybody else it seems!) that GATK is somehow the “de facto standard”.

For now, more effort should be placed on the following:

- Implementing Standards for a “clinical-grade” exome, and promoting the “networking of science” model.
- Focusing on rare, highly penetrant mutations running in families, with cascade carrier testing of even more relatives as needed.
- The genomic background is much more constant in families.
- The environmental background is sometimes more constant in families.
- This allows one to figure out penetrance of rare variants in these families, along with other issues, such as somatic mosaicism.

Please Read and Email me with Any Questions or Comments!
Email: GholsonJLyon@gmail.com

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]

Acknowledgments

Alan Rope
John C. Carey
Chad D. Huff
W. Evan Johnson
Lynn B. Jorde
Barry Moore
Jeffrey J Swensen
Jinchuan Xing
Mark Yandell

Golden Helix
Gabe Rudy

Sage Bionetworks
Stephen Friend
Lara Mangravite

Reid Robison
Edwin Nyambi

Kai Wang

Zhi Wei
Lifeng Tian
Hakon Hakonarson

our study families

Thomas Arnesen
Rune Evjenth
Johan R. Lillehaug

Jason O'Rawe
Michael Schatz
Giuseppe Narzisi

Tao Jiang
Guangqing Sun
Jun Wang

Genomic Dark Matter: The reliability of short read mapping illustrated by the Genome Mappability Score

Hayan Lee[1,2*] and Michael C. Schatz [1,2]

[1]Department of Computer Science, Stony Brook University, Stony Brook, NY
[2]Simons Center for Quantitive Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Bioinformatics Advance Access published June 4, 2012

- Genome Mappability Score (GMS) -- measure of the complexity of resequencing a genome = a weighted probability that any read could be unambiguously mapped to a given position, and thus measures the overall composition of the genome itself.
- The detection failure errors are dominated by false negatives, which means the SNP calling program fails to find such variations. In particular, among all 5022 false negatives, 3505 (70%) are located in low GMS region, and only 1517 (30%) are in high GMS region. Considering only 13-14% of human genome is low GMS region, variations in low GMS regions are clearly and substantially overrepresented. It is not surprising that errors are dominated by false negatives, as the SNP-calling algorithm will use the mapping quality score to filter out low confidence mapping. What is surprising is the extent of false negatives and the concentration of false negatives almost entirely within low GMS regions.
- The GMS should be considered in every resequencing project to pinpoint the dark matter of the genome, including of known clinically relevant variations in these regions.

Genomic Dark Matter, cont....

- That means that unlike typical false negatives, increasing coverage will not help identify mutations in low GMS regions, even with 0% sequencing error.
- Instead this is because the SNP-calling algorithms use the mapping quality scores to filter out unreliable mapping assignments, and low GMS regions have low mapping quality score (by definition). Thus even though many reads may sample these variations, the mapping algorithms cannot ever reliably map to them.
- Since about 14% of the genome has low GMS value with typical sequencing parameters, it is expected that about 14% of all variations of all resequencing studies will not be detected.
- To demonstrate this effect, we characterised the SNP variants identified by the 1000 genomes pilot project, and found that 99.99% of the SNPs reported were in high GMS regions of the genome, and in fact 99.95% had GMS over 90.