Software Considerations for Processing, Analyzing and Interpreting Exome & Genome Sequence Data in Clinical Settings

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 accept salary from anyone other than my current employer, CSHL.
- Any revenue that I earn from providing medical care is donated to UFBR for genetics research.
- I worked on the Clarity Challenge as an unpaid medical consultant to: Omicia Unlocking Individualized Medicine

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]

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.

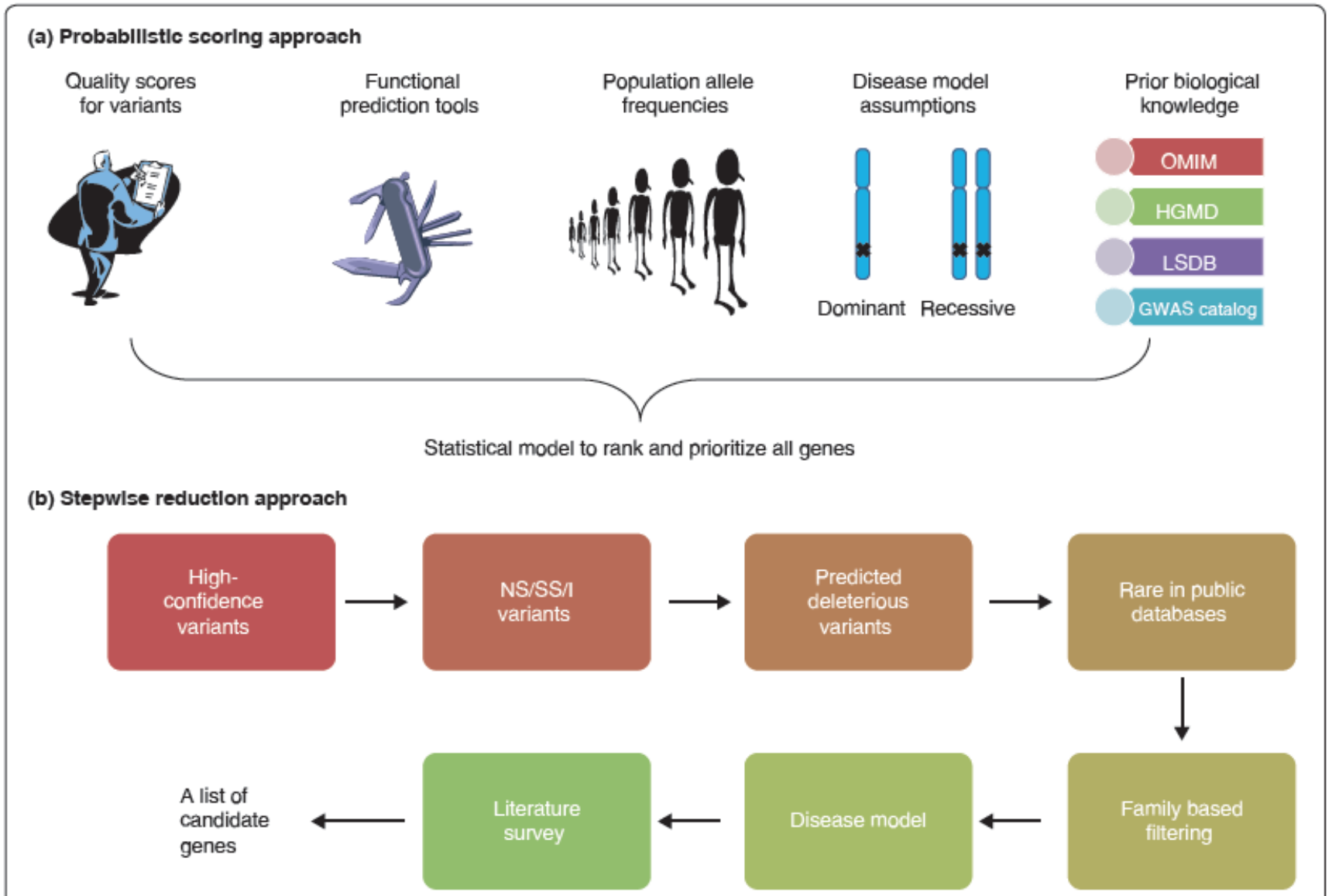


Figure 1. Two approaches for prioritizing disease causal genes from whole-genome or exome sequencing data. **(a)** The probabilistic scoring approach collects relevant information from multiple data sources, and compiles a statistical model that ranks all genes in the genome by their likelihood of being disease causal. **(b)** The stepwise reduction approach removes variants that are unlikely to be disease causal based on a series of filtering criteria, until a small set of candidate genes is found. The first approach may be more effective and rigorous, yet the second approach may be easier for non-specialists to understand and interpret. GWAS, genome-wide association study; HGMD, Human Gene Mutation Database; I, indel; LSDB, locus-specific database; NS, non-synonymous; OMIM, Online Mendelian Inheritance in Man; SS, splice acceptor or donor site.

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.

How do we get to “whole” genome sequencing for everyone?

- Tool Building for Human Genetics
- Can we reliably detect a comprehensive, and accurate, set of variants using more than one pipeline, or even more than one sequencing platform?
- How much data is enough, and how reliable and reproducible are variant calls?

Moving Exome and WGS into a Clinical Setting 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?

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.

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

Method

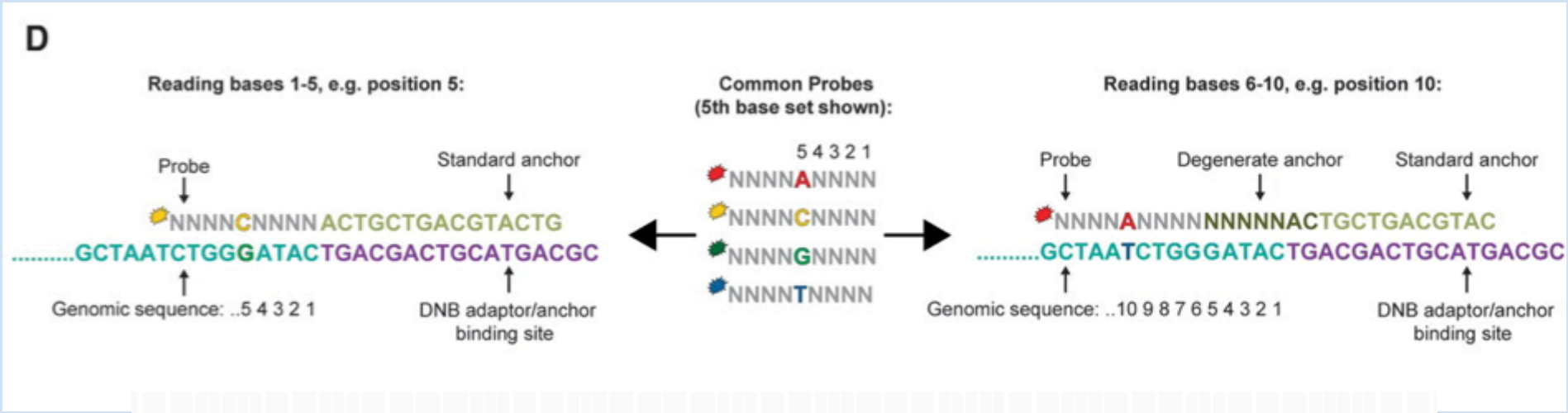
Accurate and comprehensive sequencing of personal genomes

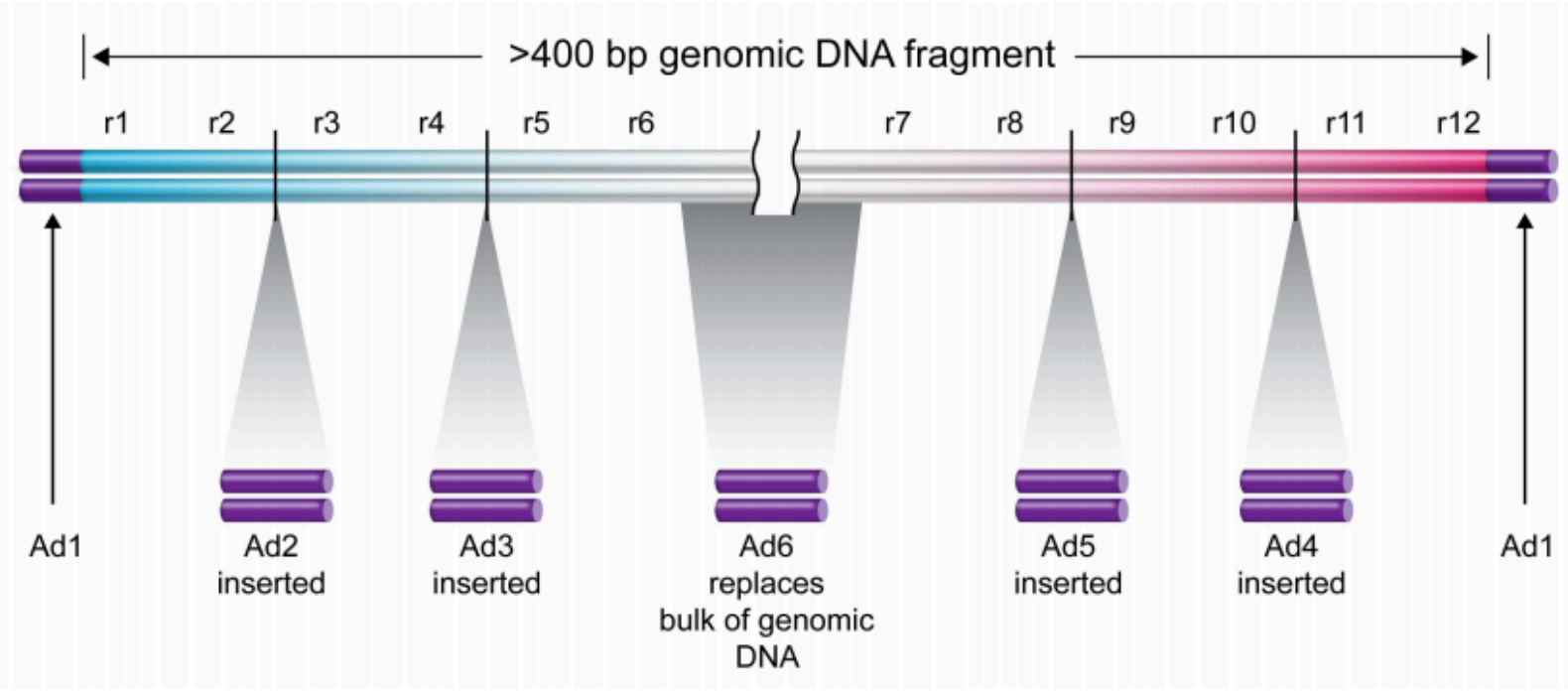
Subramanian S. Ajay,[1] Stephen C.J. Parker,[1] Hatice Ozel Abaan,[1] Karin V. Fuentes Fajardo,[2] and Elliott H. Margulies[1,3,4]

[1]*Genome Informatics Section, Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA;* [2]*Undiagnosed Diseases Program, Office of the Clinical Director, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA*

As whole-genome sequencing becomes commoditized and we begin to sequence and analyze personal genomes for clinical and diagnostic purposes, it is necessary to understand what constitutes a complete sequencing experiment for determining genotypes and detecting single-nucleotide variants. Here, we show that the current recommendation of ~30× coverage is not adequate to produce genotype calls across a large fraction of the genome with acceptably low error rates. Our results are based on analyses of a clinical sample sequenced on two related Illumina platforms, $GAII_x$ and HiSeq 2000, to a very high depth (126×). We used these data to establish genotype-calling filters that dramatically increase accuracy. We also empirically determined how the callable portion of the genome varies as a function of the amount of sequence data used. These results help provide a "sequencing guide" for future whole-genome sequencing decisions and metrics by which coverage statistics should be reported.

Complete Genomics chemistry - combinatorial probe anchor ligation (cPAL)





Accuracy of Complete Genomics Whole Human Genome Sequencing Data

Analysis Pipeline v2.0

	FALSE POSITIVES	EST FPs	FALSE NEGATIVES	TOTAL DISCORDANCES	CONCORDANCE
Discordant SNVs per called MB	1.56 x 10-6	4,450	1.67 x 10-6	3.23 x 10-6	99.9997% of bases

Table 2. *Concordance of Technical Replicates.*

COMPLETE GENOMICS CALL	OTHER PLATFORM	PLATFORM-SPECIFIC SNVs	VALIDATION RATE	EST FPs	FPR
Het or Hom SNV	No SNV Reported	99K	17/18 = 94.4%	5,577	0.16%
No-call or Hom-Ref	SNV Reported	345K	2/15 = 13.3%	299,115	8.2%

Table 3. *False Positive Rate.*

Complete Genomics – LFR technology

Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells

Brock A. Peters[1]*, Bahram G. Kermani[1]*, Andrew B. Sparks[1]†, Oleg Alferov[1], Peter Hong[1], Andrei Alexeev[1], Yuan Jiang[1], Fredrik Dahl[1]†, Y. Tom Tang[1], Juergen Haas[1], Kimberly Robasky[2,3], Alexander Wait Zaranek[2], Je-Hyuk Lee[2,4], Madeleine Price Ball[2], Joseph E. Peterson[1], Helena Perazich[1], George Yeung[1], Jia Liu[1], Linsu Chen[1], Michael I. Kennemer[1], Kaliprasad Pothuraju[1], Karel Konvicka[1], Mike Tsoupko-Sitnikov[1], Krishna P. Pant[1], Jessica C. Ebert[1], Geoffrey B. Nilsen[1], Jonathan Baccash[1], Aaron L. Halpern[1], George M. Church[2] & Radoje Drmanac[1]

NATURE | VOL 487 | 12 JULY 2012

"Substantial error rates (1 single nucleotide variants (SNV) in 100–1,000 called kilobases) are a common attribute of all current massively parallelized sequencing technologies. These rates are probably too high for diagnostic use and complicate many studies searching for new mutations."

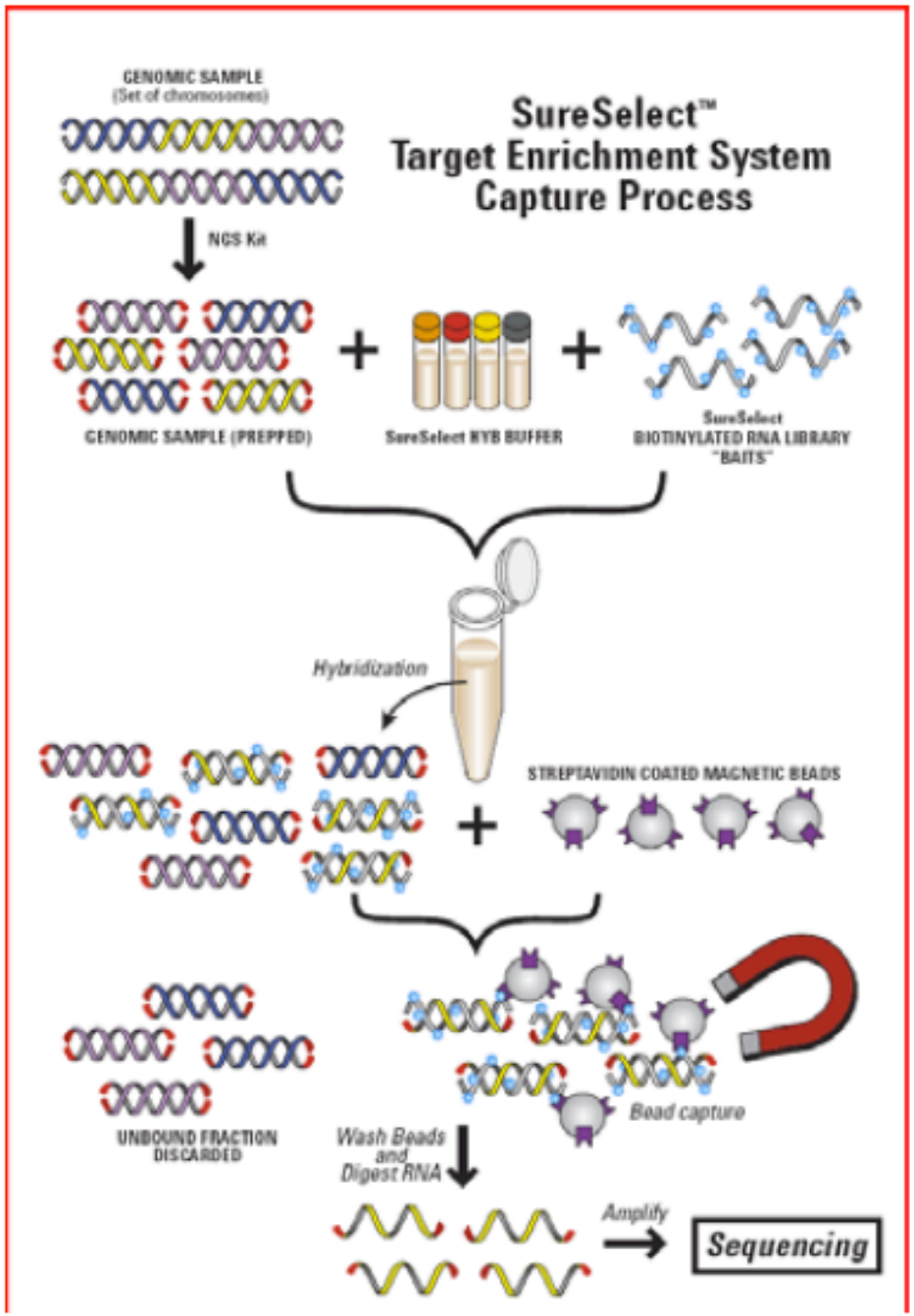
Much Higher Accuracy with LFR data

"To test LFR reproducibility we compared haplotype data between the two NA19240 replicate libraries. In general, the libraries were very concordant, with only 64 differences per library in 2.2 million heterozygous SNPs phased by both libraries or **1 of this error type in 44 Mb**."

- ~$3000 for 30x “whole” genome as part of Illumina Genome Network on a research basis only, but ~$5,000 for whole genome performed in a CLIA lab at Illumina.

Agilent Technologies
SureSelect method

Whole-exome kit
38Mb and 50Mb



http://cp.literature.agilent.com/litweb/pdf/5990-3532EN.pdf

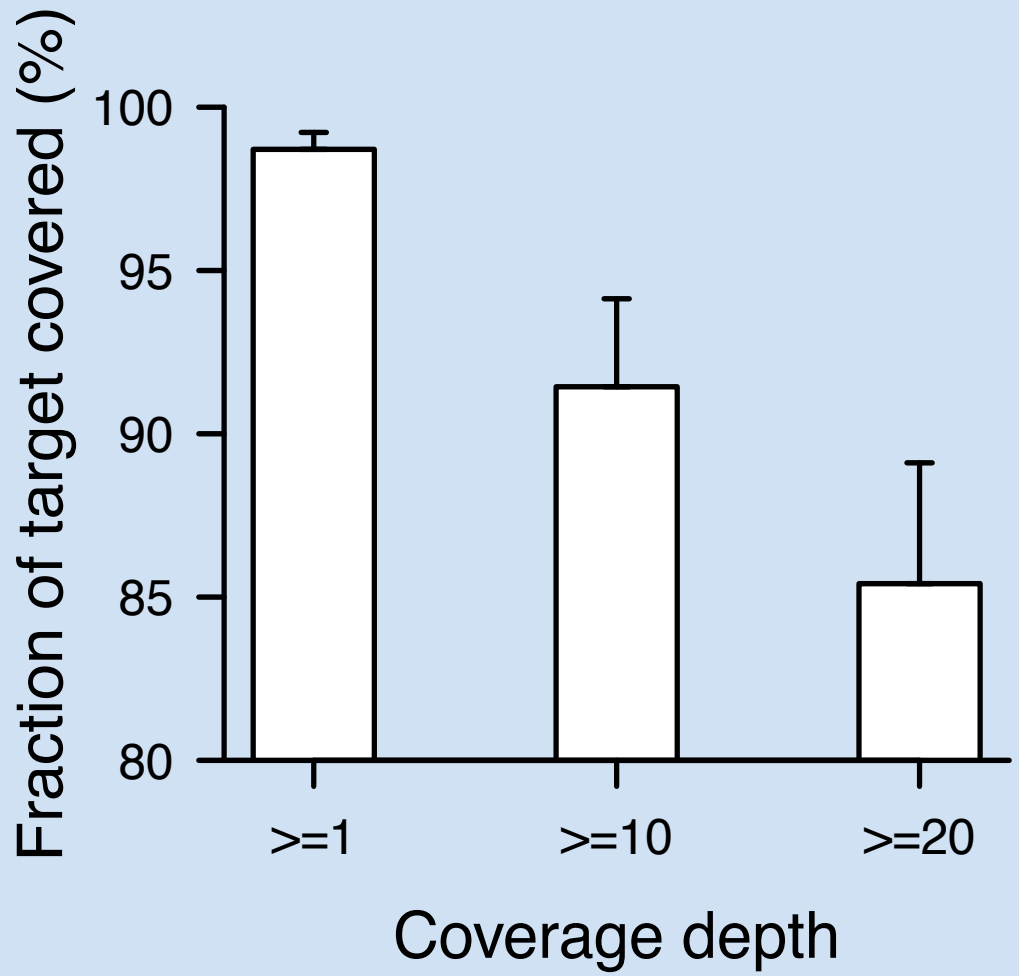
Optimizing Variant Calling in Exomes at BGI in 2011

- 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

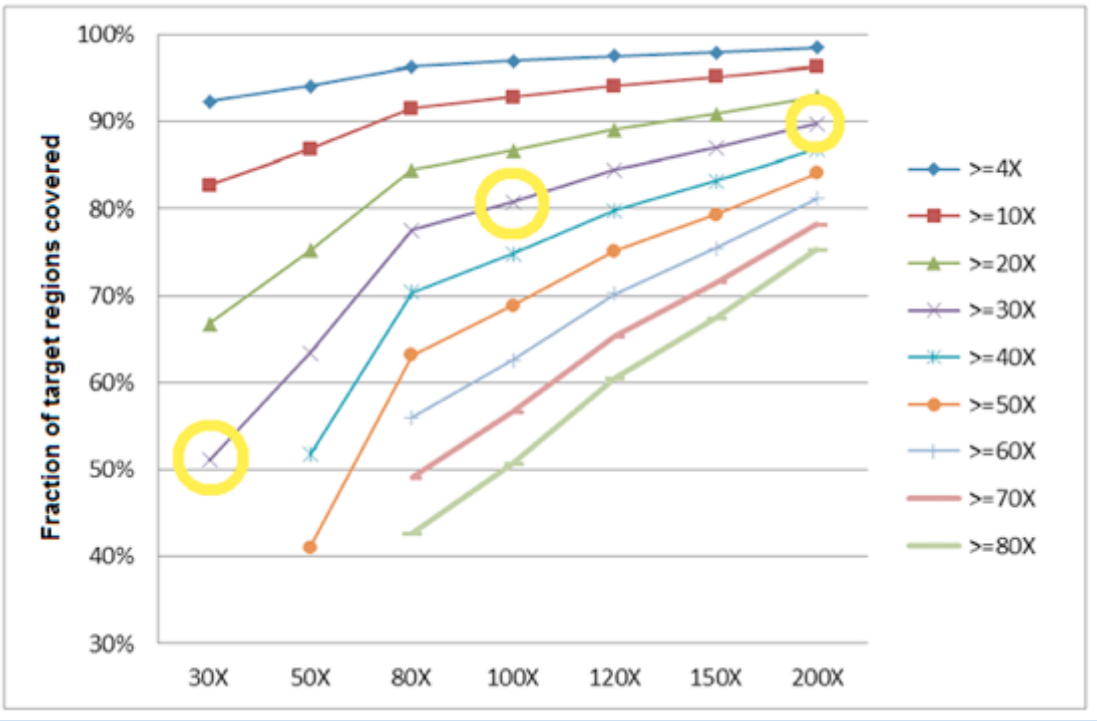


Figure from BGI website: http://bgiamericas.com/news-events/why-deep-exome-sequencing/

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.

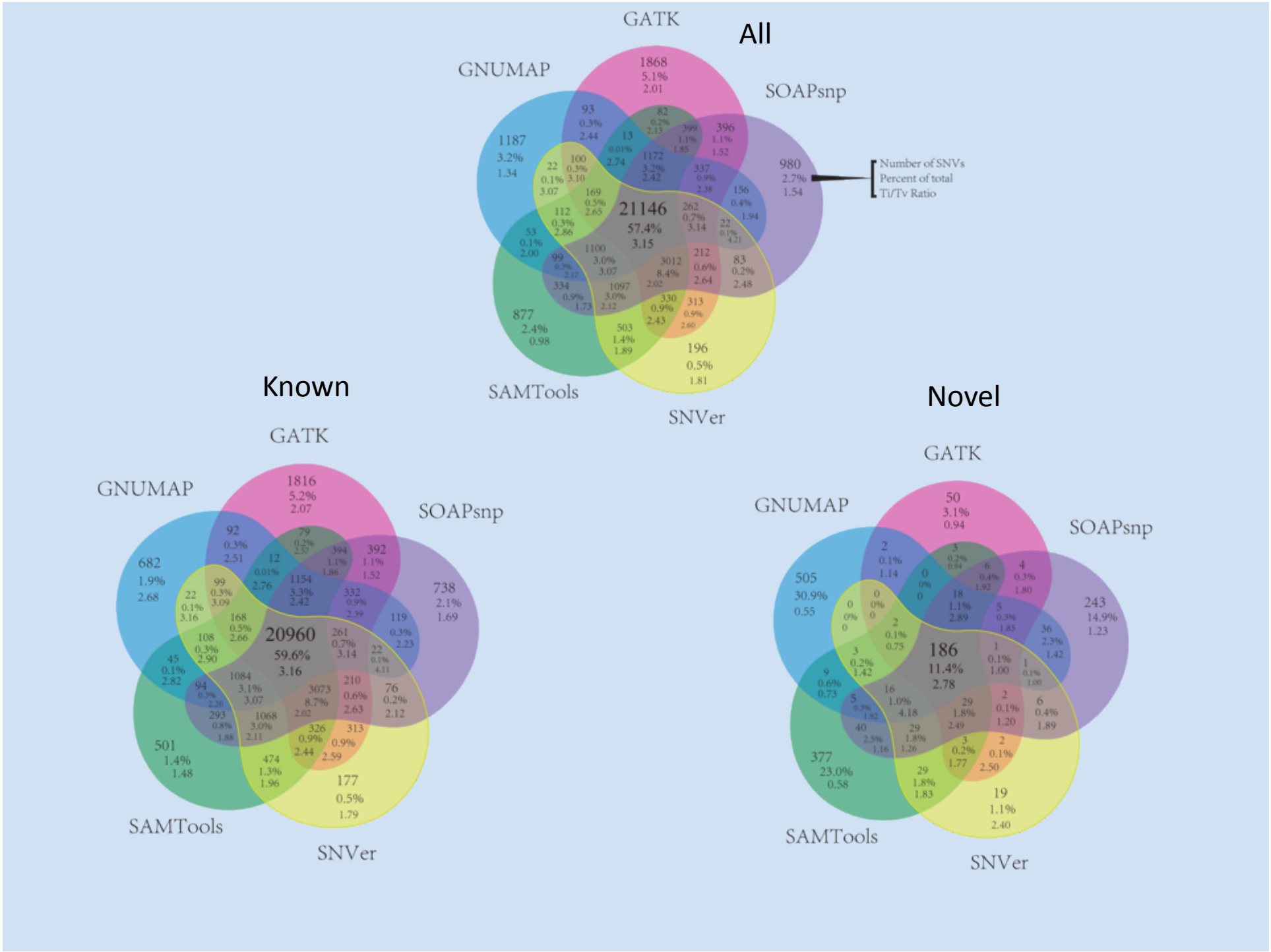
GWAS has statistical rigor with a threshold p value

- 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?

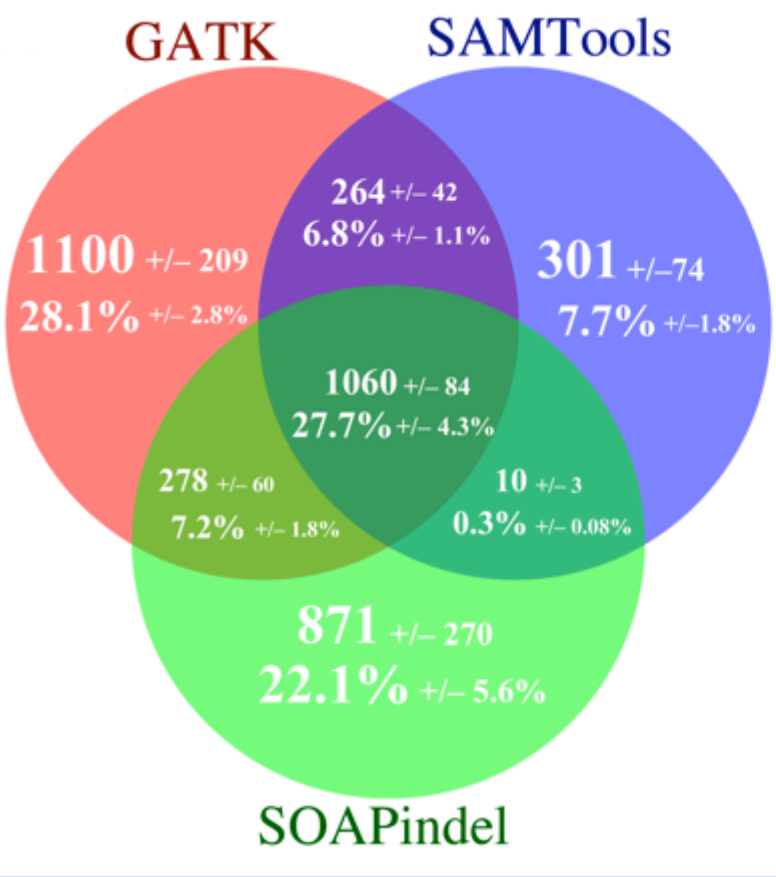
5 Pipelines Used on Same Set of Seq Data by Different Analysts, using Hg19 Reference Genome

Pipeline name	Alignment method	Variant calling module	Description of variant calling algorithm
SOAP	SOAPaligner/BWA	SOAPsnp/SOAPindel	SOAP uses a method based on Bayes' theorem to call consensus genotype by carefully considering the data quality, alignment, and recurring experimental errors [22].
GATK	BWA	GATK	GATK employs a general Bayesian framework to distinguish and call variants. Error correction models are guided by expected characteristics of human variation to further refine variant calls [19].
SNVer	BWA	SNVer	SNVer uses a more general frequentist framework and formulates variant calling as a hypothesis testing problem [25].
GNUMAP	GNUMAP	GNUMAP	GNUMAP incorporates the base uncertainty of the reads into mapping analysis using a Probabilistic Needleman-Wunsch algorithm [24].
SAMTools	BWA	mpileup	SAMTools [20] calls variants by generating a consensus sequence using the MAQ model framework which uses a general Baysean framework for picking the base which maximizes the posterior probability with the highest phred quality score.

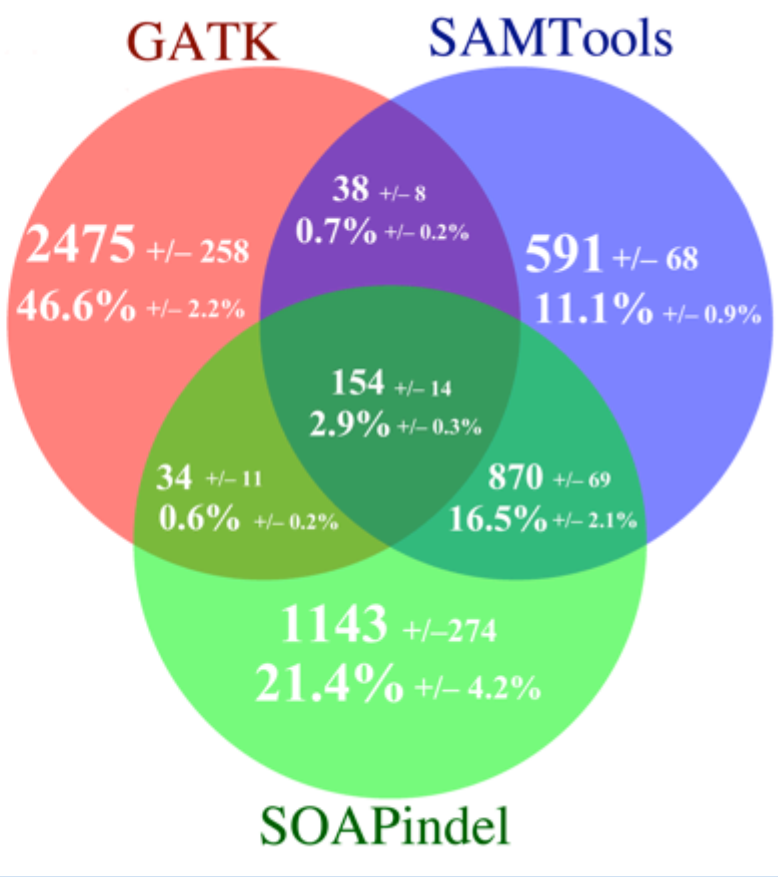

All
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1868 5.1% 2.01
1187 3.2% 1.34
980 2.7% 1.54
877 2.4% 0.98
196 0.5% 1.81
21146 57.4% 3.15
Number of SNVs
Percent of total
Ti/Tv Ratio
Known
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1816 5.2% 2.07
682 1.9% 2.68
738 2.1% 1.69
501 1.4% 1.48
177 0.5% 1.79
20960 59.6% 3.16
Novel
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
50 3.1% 0.94
505 30.9% 0.55
243 14.9% 1.23
377 23.0% 0.58
19 1.1% 2.40
186 11.4% 2.78

INDELS

Indels- Overlap by Base Position only



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

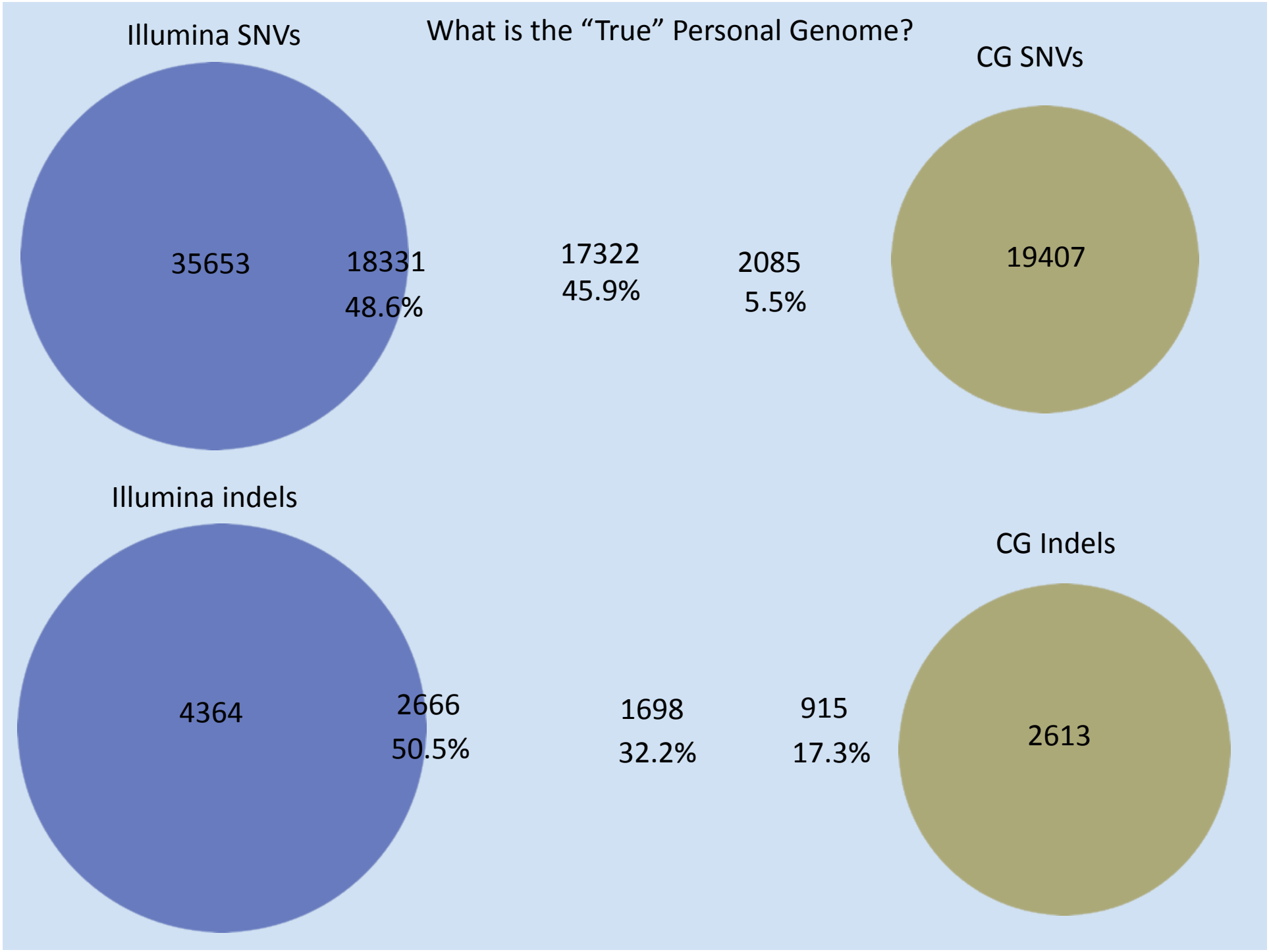


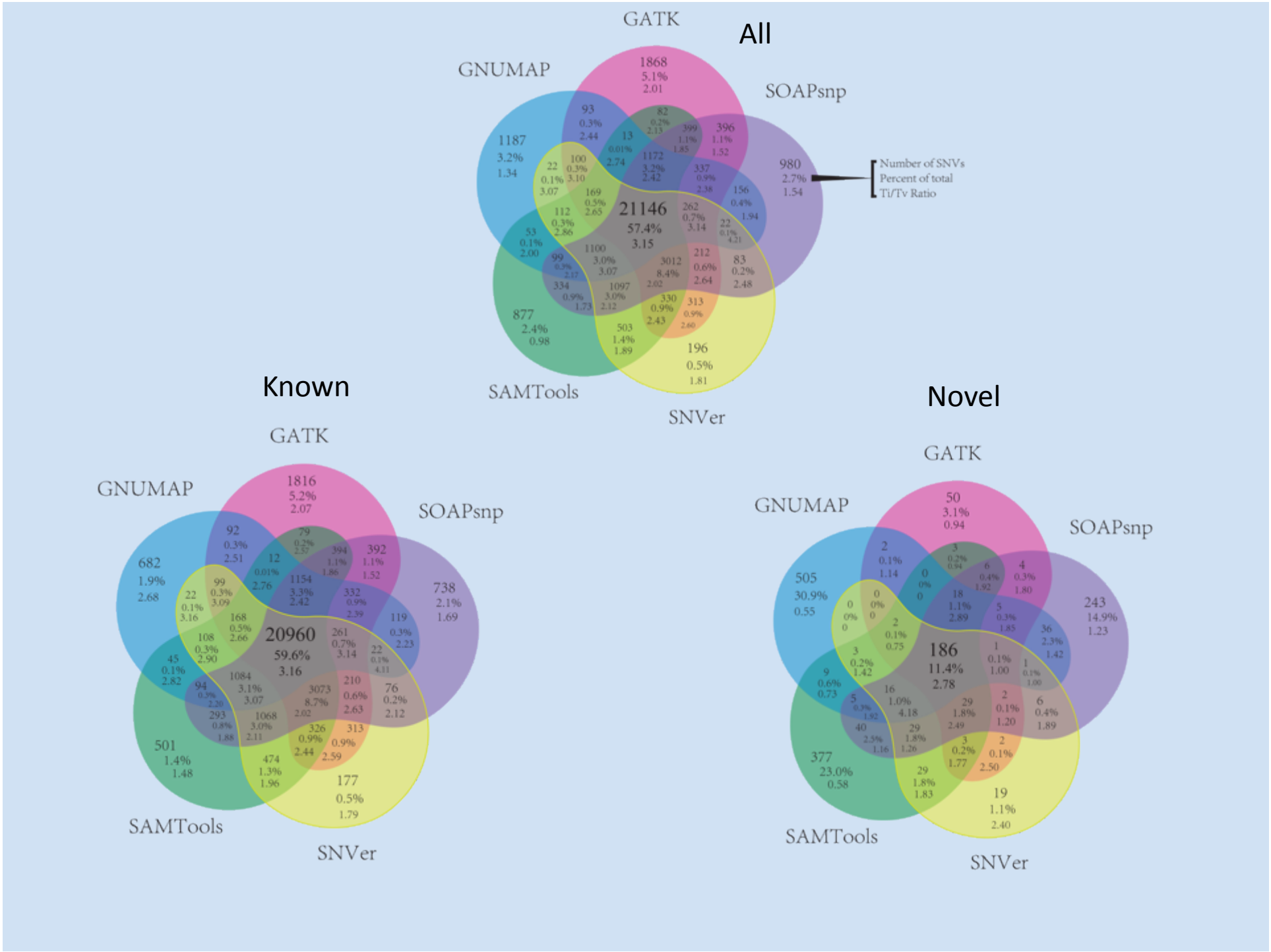
Total mean overlap, plus or minus one standard deviation, observed between three indel calling pipelines: GATK, SOAP-indel, and SAMTools. a) Mean overlap when indel position was the only necessary agreement criterion. **b)** Mean overlap when indel position, base length and base composition were the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

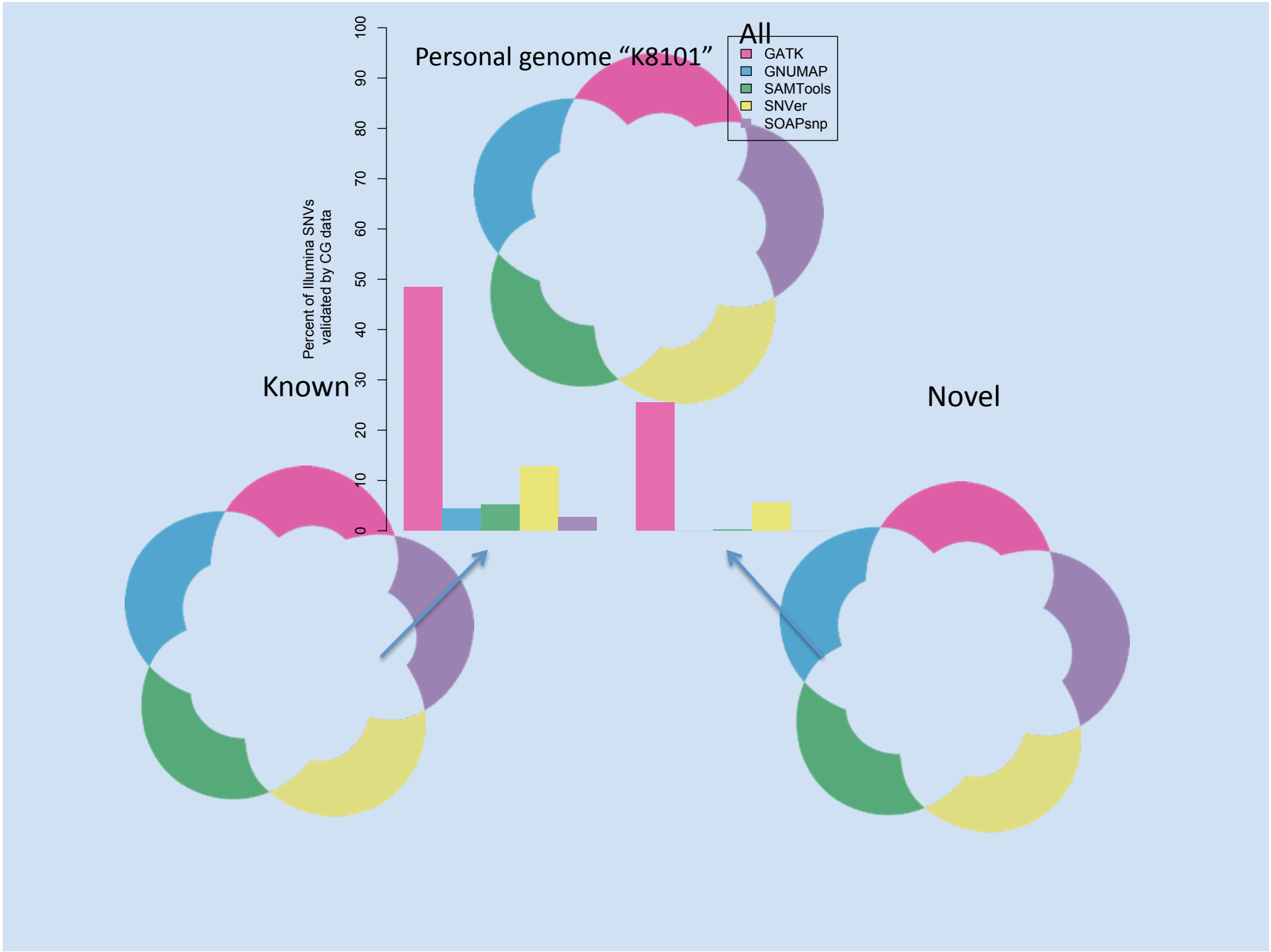
- How reliable are variants that are uniquely called by individual pipelines?

- Are some pipelines better at detecting rare, or novel variants than others?

Cross validation using orthogonal sequencing technology (Complete Genomics)


What is the “True” Personal Genome?
Illumina SNVs
CG SNVs
35653
18331
48.6%
17322
45.9%
2085
5.5%
19407
Illumina indels
CG Indels
4364
2666
50.5%
1698
32.2%
915
17.3%
2613

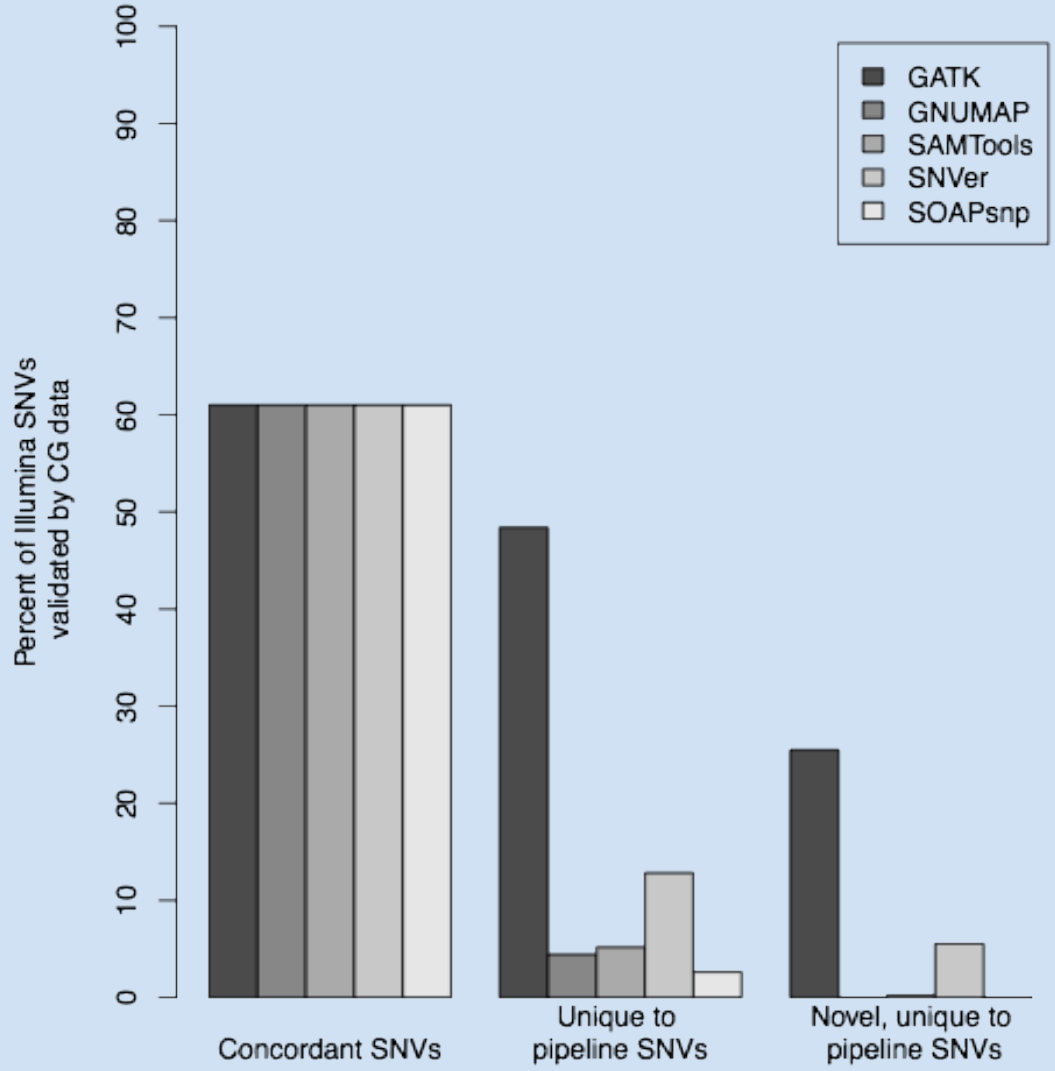

All
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1868 5.1% 2.01
1187 3.2% 1.34
980 2.7% 1.54
877 2.4% 0.98
196 0.5% 1.81
21146 57.4% 3.15
Number of SNVs
Percent of total
Ti/Tv Ratio
Known
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1816 5.2% 2.07
682 1.9% 2.68
738 2.1% 1.69
501 1.4% 1.48
177 0.5% 1.79
20960 59.6% 3.16
Novel
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
50 3.1% 0.94
505 30.9% 0.55
243 14.9% 1.23
377 23.0% 0.58
19 1.1% 2.40
186 11.4% 2.78


Personal genome “K8101”
All
GATK
GNUMAP
SAMTools
SNVer
SOAPsnp
Percent of Illumina SNVs validated by CG data
0
10
20
30
40
50
60
70
80
90
100
Known
Novel

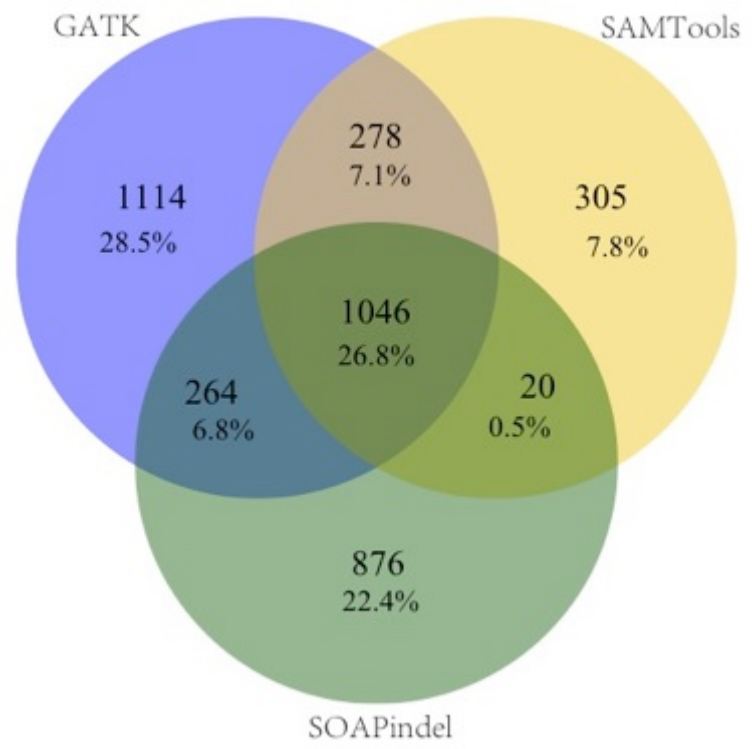
Higher Validation of SNVs with the BWA-GATK pipeline

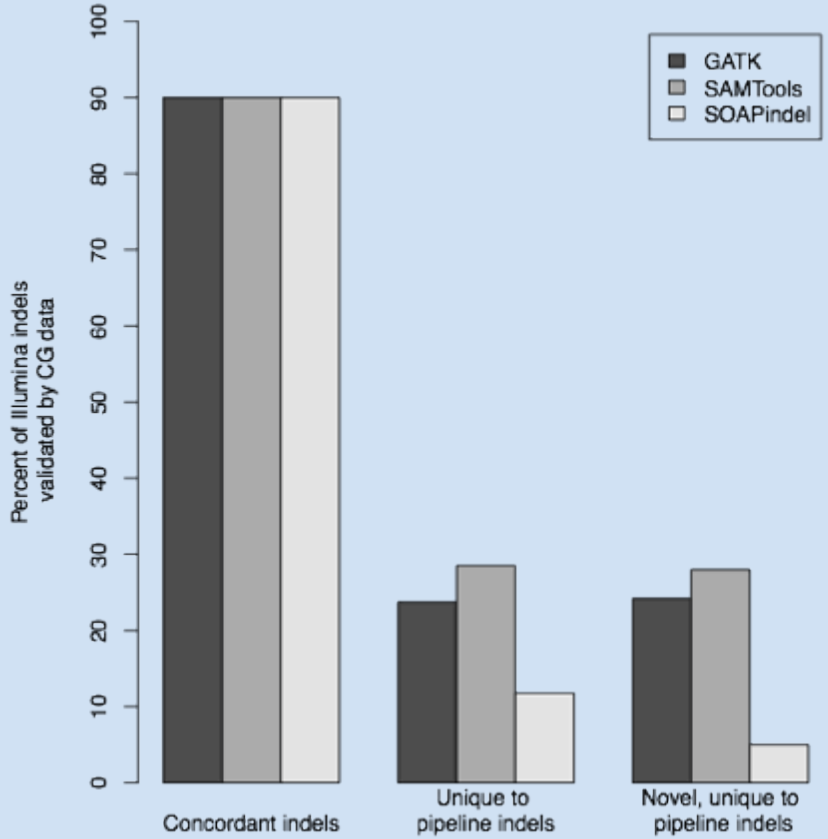
- Reveals higher validation rate of unique-to-pipeline variants, as well as uniquely discovered novel variants, for the variants called by BWA-GATK, in comparison to the other 4 pipelines (including SOAP).

Much Higher Validation of the Concordantly Called Variants (by the CG data)



Validating Indels with Complete Genomics Data for the 3 pipelines

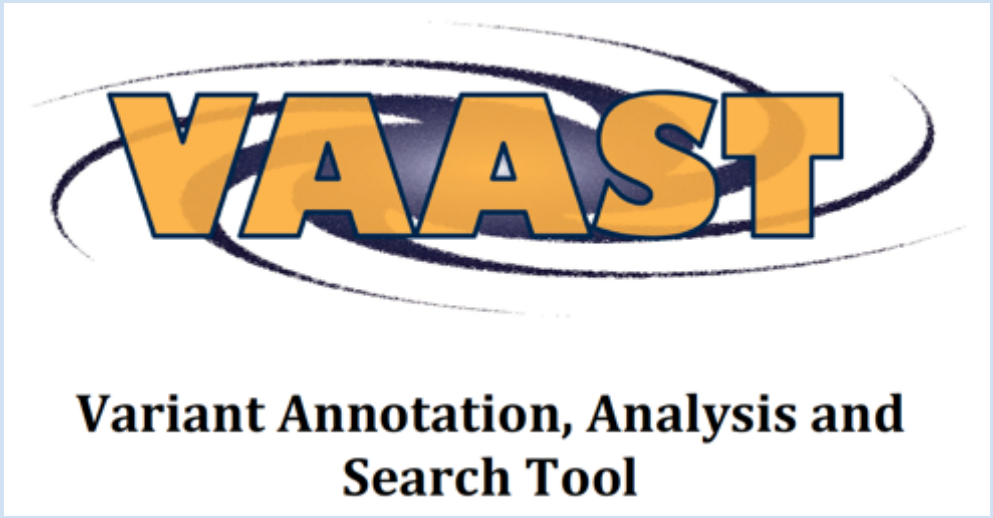




Tools sensitivity for longer indels

- Standard read mapping and scanning algorithms, such as **BWA**, **GATK**, and **SAMTools**, are suitable for detecting mutations only for a few nucleotides.
  - The sensitivity drops significantly for indels larger than 10bp
  - Large insertions (> read length), are hard to detect.
  - As a result, variants > 15 bp have rarely been reported in exome studies

Gavin R. Oliver, **F1000 Research**, 2012



Yandell M, Huff C, Hu H, Singleton M, Moore B, Xing J, Jorde LB, Reese MG. A probabilistic disease-gene finder for personal genomes. Genome Res. 2011 Sep;21(9): 1529-42.

ARTICLE

Using VAAST to Identify an X-Linked Disorder Resulting in Lethality in Male Infants Due to N-Terminal Acetyltransferase Deficiency

Alan F. Rope,[1] Kai Wang,[2,19] Rune Evjenth,[3] Jinchuan Xing,[4] Jennifer J. Johnston,[5] Jeffrey J. Swensen,[6,7] W. Evan Johnson,[8] Barry Moore,[4] Chad D. Huff,[4] Lynne M. Bird,[9] John C. Carey,[1] John M. Opitz,[1,4,6,10,11] Cathy A. Stevens,[12] Tao Jiang,[13,14] Christa Schank,[8] Heidi Deborah Fain,[15] Reid Robison,[15] Brian Dalley,[16] Steven Chin,[6] Sarah T. South,[1,7] Theodore J. Pysher,[6] Lynn B. Jorde,[4] Hakon Hakonarson,[2] Johan R. Lillehaug,[3] Leslie G. Biesecker,[5] Mark Yandell,[4] Thomas Arnesen,[3,17] and Gholson J. Lyon[15,18,20,*]

VAAST

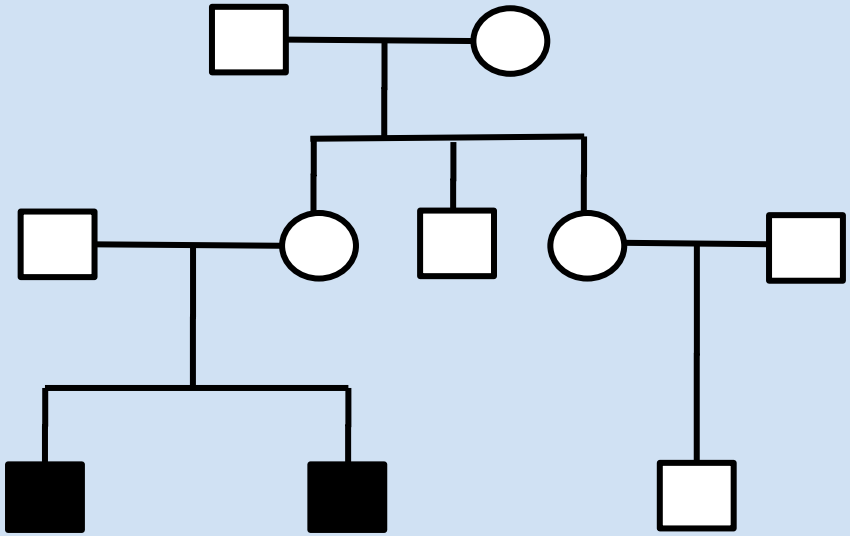
- A probabilistic disease-gene finder for personal genomes
- Rapidly search personal genome sequences for damaged genes by identifying significant differences in variant frequencies in cases vs. controls
- Integrates both allele & AAS frequencies into a single probabilistic framework
- Can score both coding and non-coding variants
- Leverage phase and pedigree data
- Can be used to hunt for both rare and common disease genes and their causative alleles
- Determines the statistical significance of candidate genes

Omicia THE UNIVERSITY OF UTAH

VAAST integrates AAS & Variant frequencies in a single probabilistic framework

- non-coding variants scored using allele frequency differences
- n_i : frequency of variant type among all variants observed in Background and Target genomes
- a_i: frequency of variant type among disease causing mutations in OMIM
- This approach means that *every* variant can be scored, non-synonymous, synonymous, coding, and non-coding. Phylogenetic conservation not required.

New Syndrome with Dysmorphology, Mental Retardation, “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

7 years old

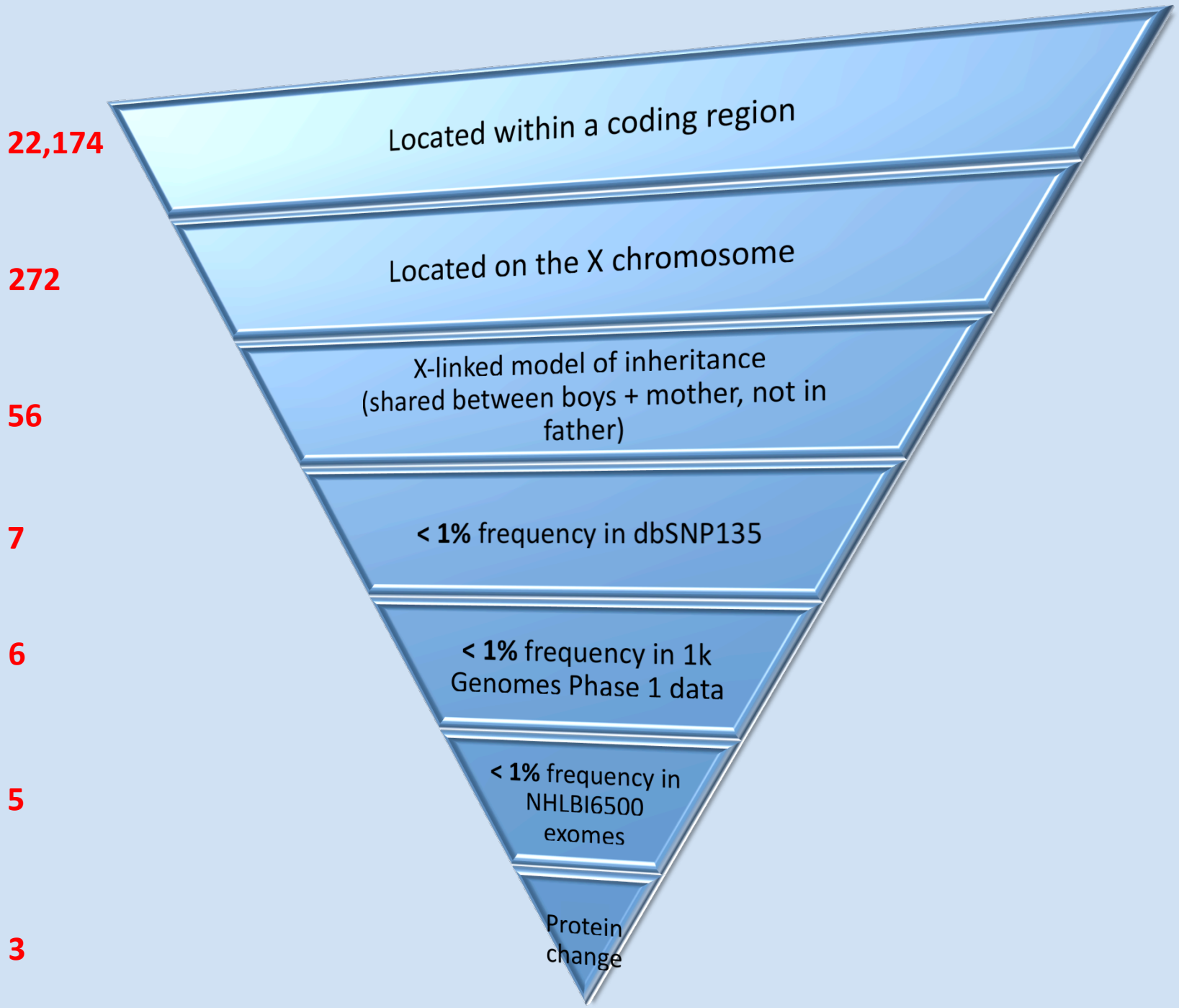
3 years old

5 years old

9 years old

Workup Ongoing for past 10 years

- Numerous genetic tests negative, including negative for Fragile X and many candidate genes.
- No obvious pathogenic CNVs – microarrays normal.
- Sequenced whole genomes of Mother, Father and Two Boys, using Complete Genomics, obtained data in June of this year, i.e. version 2.0 CG pipeline.


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

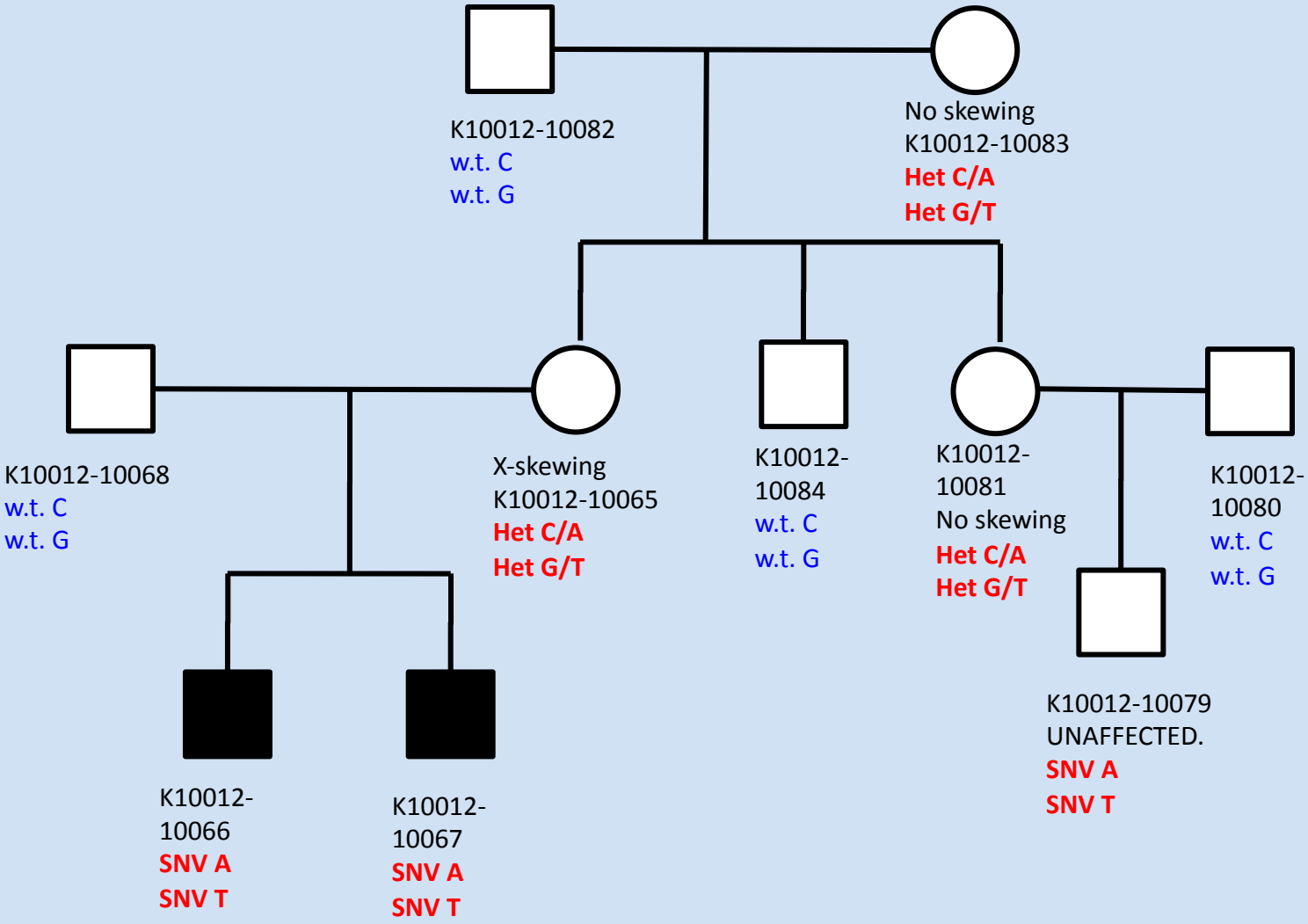
Mutations in the *ZNF41* Gene Are Associated with Cognitive Deficits: Identification of a New Candidate for X-Linked Mental Retardation

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

Sanger validation: ASB12 and ZNF41 mutations



The mutation in ZNF41 may **NOT** be necessary, and it is certainly **NOT** sufficient to cause the phenotype.

So, of course we need baseline whole genome sequencing on everyone to at least understand the DNA genetic background in each pedigree or clan.

Ancestry Matters!

Some are calling for technical replicates of exomes for higher accuracy.

2426–2431 *Nucleic Acids Research, 2012, Vol. 40, No. 6* *Published online 29 November 2011*
doi:10.1093/nar/gkr1073

The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process

Verena Heinrich[1], Jens Stange[2], Thorsten Dickhaus[2], Peter Imkeller[2], Ulrike Krüger[1], Sebastian Bauer[1], Stefan Mundlos[1], Peter N. Robinson[1], Jochen Hecht[3] and Peter M. Krawitz[1,*]

[1]Institute for Medical and Human Genetics, Charité Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, [2]Department of Mathematics, Humboldt-University Berlin, Unter den Linden 6, 10099 Berlin and [3]Berlin-Brandenburg Center for Regenerative Therapies (BCRT), Charité Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany

Received July 20, 2011; Revised October 19, 2011; Accepted October 28, 2011

"In a usual exome, one expects between 10 000 and 15 000 heterozygous variants. Our results indicate that one will miss around a hundred heterozygous variants by sequencing an exome only once simply due to the stochastic fluctuation of the allele frequencies after amplification…. Additionally for a sequencing depth above 30x, the false negative rate does not decrease further. Thus, once a sufficient sequencing depth has been reached, only technical replication is able to further reduce the total error rates substantially."

Some argue that exon capture should complement WGS sequencing….

Performance comparison of exome DNA sequencing technologies

Michael J Clark[1,4], Rui Chen[1,4], Hugo Y K Lam[1], Konrad J Karczewski[1], Rong Chen[2], Ghia Euskirchen[1,3], Atul J Butte[2] & Michael Snyder[1,3]

VOLUME 29 NUMBER 10 OCTOBER 2011 NATURE BIOTECHNOLOGY

"It may be argued that the importance of targeted sequencing is transient and will diminish as WGS becomes less expensive. However, we found that exome sequencing can identify variants that are not evident in WGS because of greater base coverage after enrichment. Even at equivalent coverage levels, specific regions had higher read depth in exome sequencing resulting in greater sensitivity in those regions. Target capture by exome sequencing unambiguously identified some of these difficult regions through preferential selection and observation at higher local read depth. "

And, others are calling for potentially biological replicates with WGS sequencing on two platforms

Performance comparison of whole-genome sequencing platforms

Hugo Y K Lam[1,8], Michael J Clark[1], Rui Chen[1], Rong Chen[2,8], Georges Natsoulis[3], Maeve O'Huallachain[1], Frederick E Dewey[4], Lukas Habegger[5], Euan A Ashley[4], Mark B Gerstein[5–7], Atul J Butte[2], Hanlee P Ji[3] & Michael Snyder[1]

"both methods clearly call variants missed by the other technology. Many of these lie in exons and thus can affect coding potential. In fact, 1,676 genes have platform-specific SNVs in exons … We demonstrated that the best approach for comprehensive variant detection is to sequence genomes with both platforms if budget permits. We assessed the cost effectiveness of sequencing on both platforms and found that on average it costs about four cents per additional variant (Online Methods). Alternatively, supplementing with exome sequencing can assess the most interpretable part of the genome at higher depth of coverage and accuracy and fill in the gaps in the detection of coding variants."

"Genomic Dark Matter"

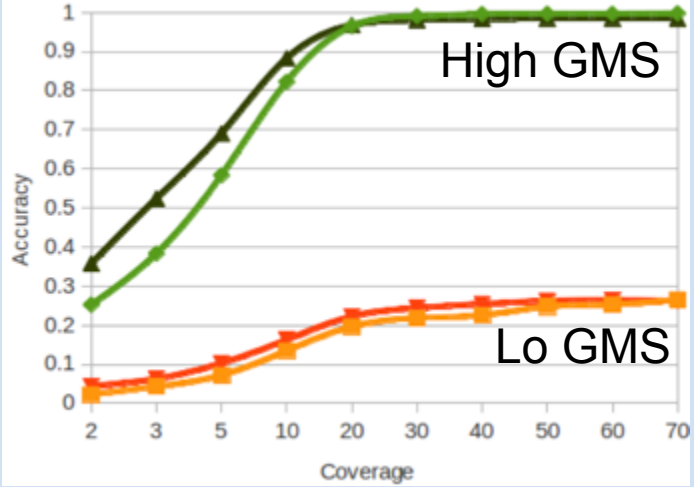
Short read mapping is a widely used for identifying mutations in the genome

- Not every base of the genome can mapped equally well, because repeats may obscure where the reads originated

Species (build)	size	paired/single	whole (%)	transcription (%)
yeast (sc2)	12 Mbp	paired	94.85	95.04
		single	94.25	94.62
fly (dm3)	130 Mbp	paired	90.52	96.14
		single	89.70	95.94
mouse (mm9)	2.7 Gbp	paired	89.39	96.03
		single	87.47	94.75
human (hg19)	3.0 Gbp	paired	89.02	97.40
		single	87.79	96.38

Introduced a new probabilistic metric - the Genome Mappability Score - that quantifies how reliably reads can be mapped to every position in the genome

- We have little power to measure 11-13% of the human genome, including of known clinically relevant variations
- Errors in variation discovery are dominated by false negatives in low GMS regions



Genomic Dark Matter: The reliability of short read mapping illustrated by the GMS.
Lee, H., Schatz, M.C. (2012) *Bioinformatics*. 10.1093/bioinformatics/bts330

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.
- 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.

Summary

- Next Gen Sequencing Technology constantly improving, with longer read lengths and higher accuracy of base calling.
- Variant-calling for SNVs, indels and CNVs is also constantly improving.
- Downstream filtering and probabilistic ranking algorithms depend on a highly accurate and comprehensive list of variant calls.
- Ancestry, i.e. genetic background, matters! So, we need to collect large families and move to whole genome sequencing as much as possible.

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

The VAAST DEVELOPMENT GROUP

www.yandell-lab.org

- Hao Hu[+]
- Barry Moore[+]
- Steve Chervitz[+!]
- Chad Huff[x]
- Jinchuan Xing[x+]
- Marc Singleton[+]
- Edward Kiruluta[!]
- Archie Russell[!]
- Fidel Salas[!]
- Ginger Guozhen Fan[+]

[+]Yandell lab, [!]Omicia, [x]Jorde Lab

Omicia

THE UNIVERSITY OF UTAH