Summary of Efforts to Achieve and Evaluate High-Quality Exomes and Genomes

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

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?

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.

CLIA-certified exomes and WGS

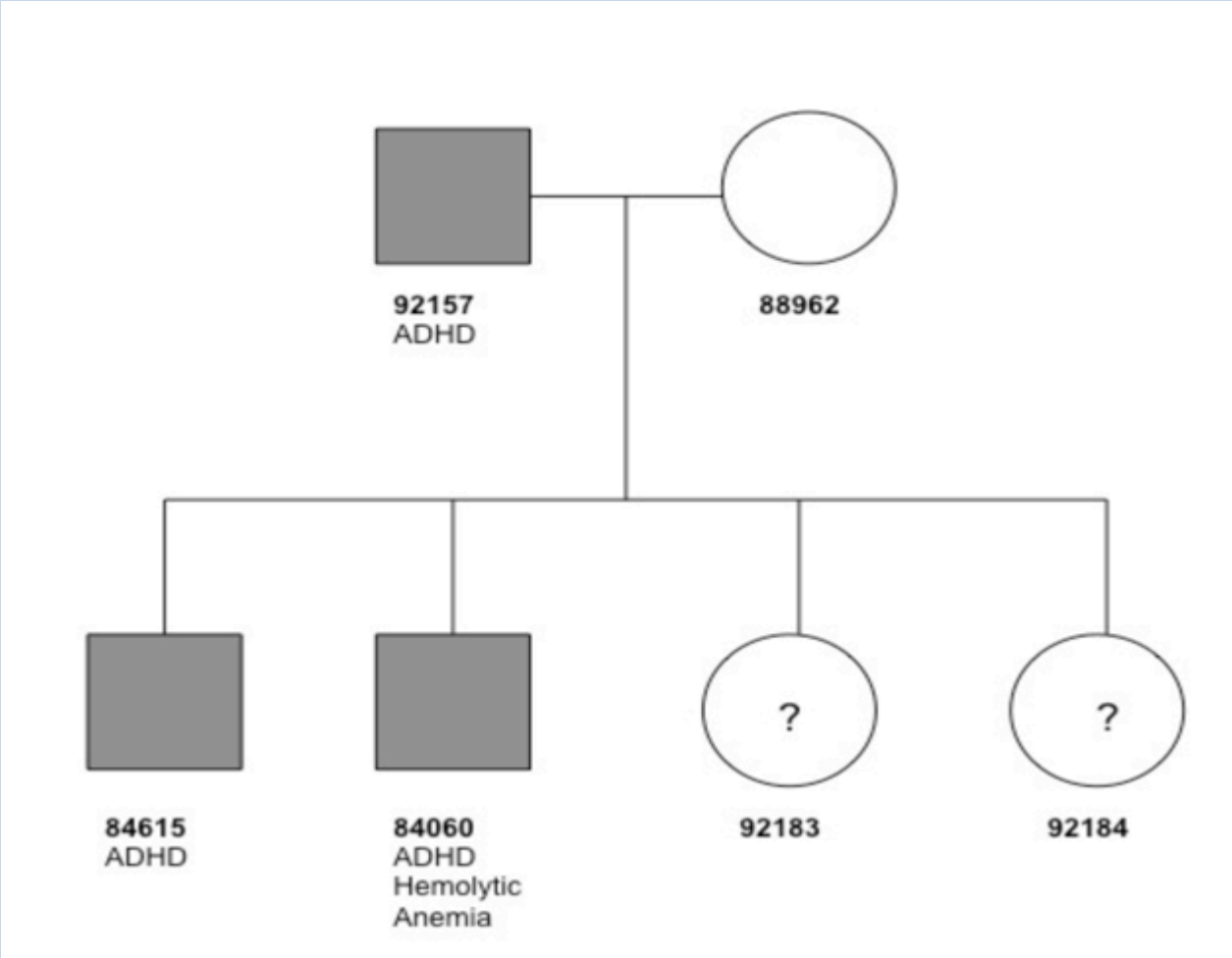
- The CLIA-certified pipelines attempt to minimize false positives with increased depth of sequencing, although there can still be many no-calls and other areas of uncertainty, which should be reported as No-Call Regions.
- This will minimize false positives and also tend to prevent false negatives.

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

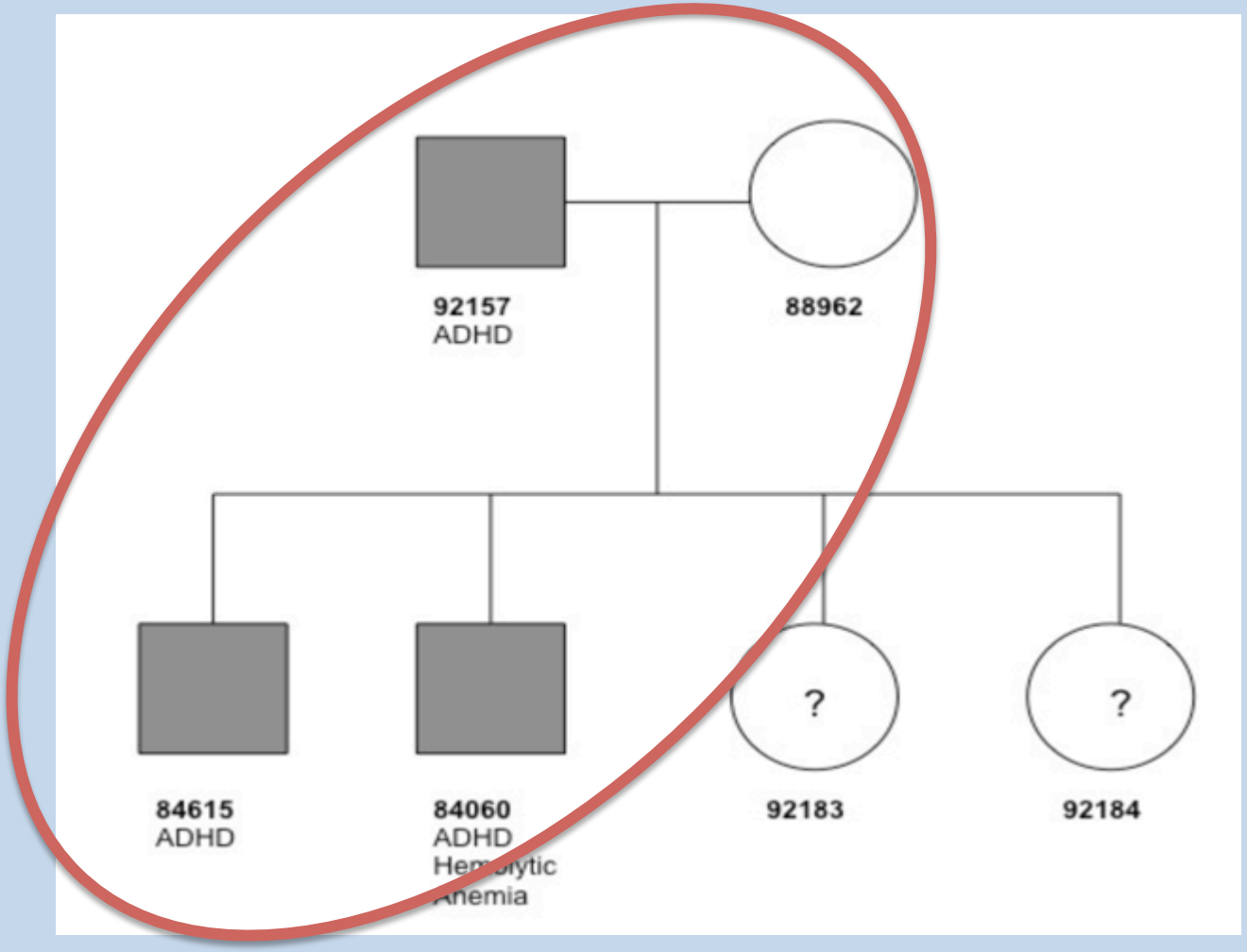
Gholson J. Lyon, Tao Jiang, Richard Van Wijk, Wei Wang, Paul Mark Bodily, Jinchuan Xing, Lifeng Tian, Reid J. Robison, Mark Clement, Lin Yang, Peng Zhang, Ying Liu, Barry Moore, Joseph T. Glessner, Josephine Elia, Fred Reimherr, Wouter W. van Solinge, Mark Yandell, Hakon Hakonarson, Jun Wang, William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

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

Exome sequencing of one pedigree in a research setting.



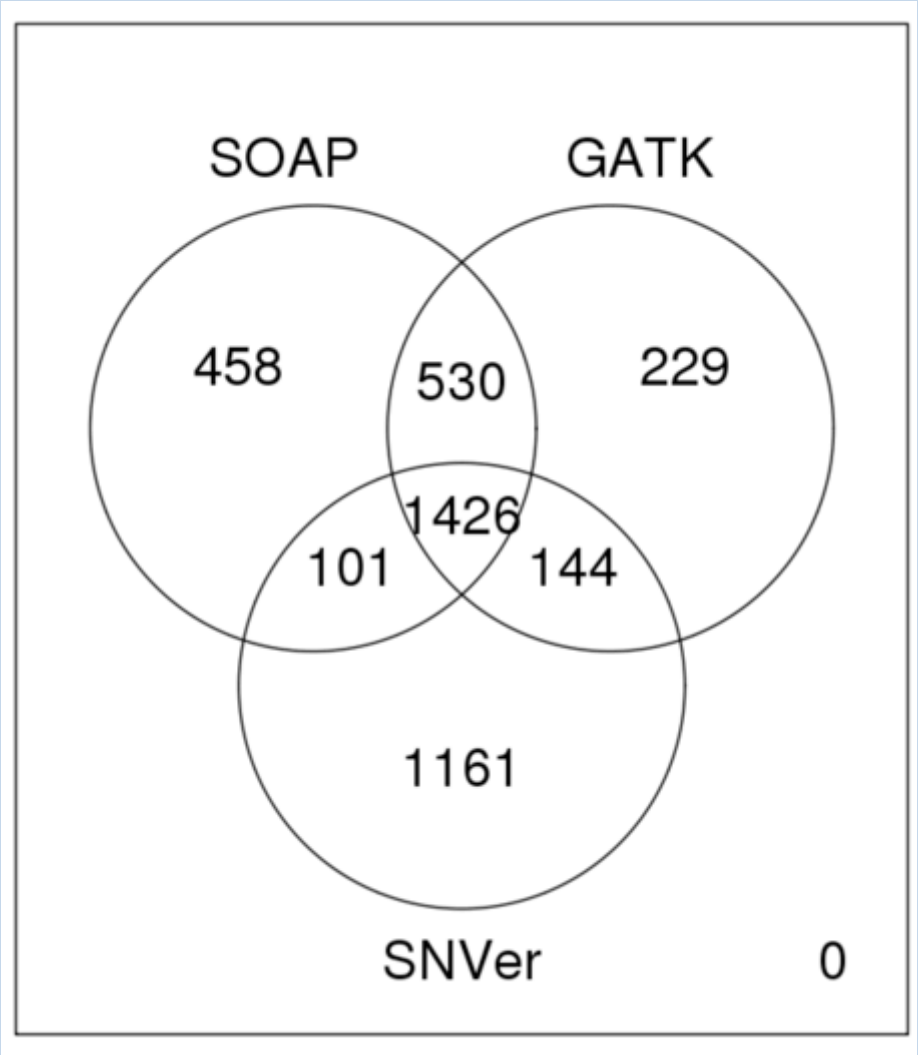
Exome sequencing of one pedigree in a research setting.



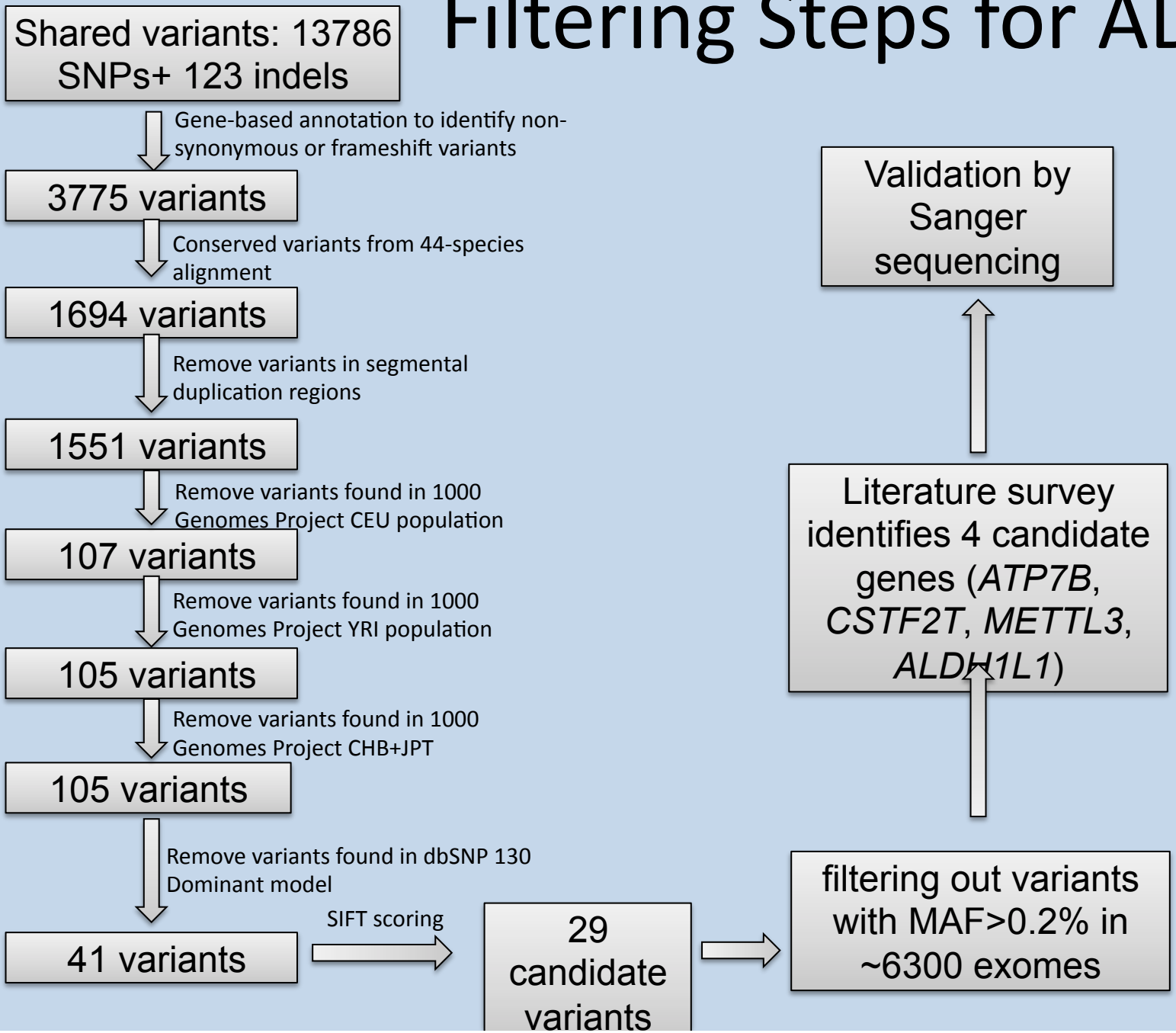
Bioinformatics Analysis for ADHD pedigree

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.

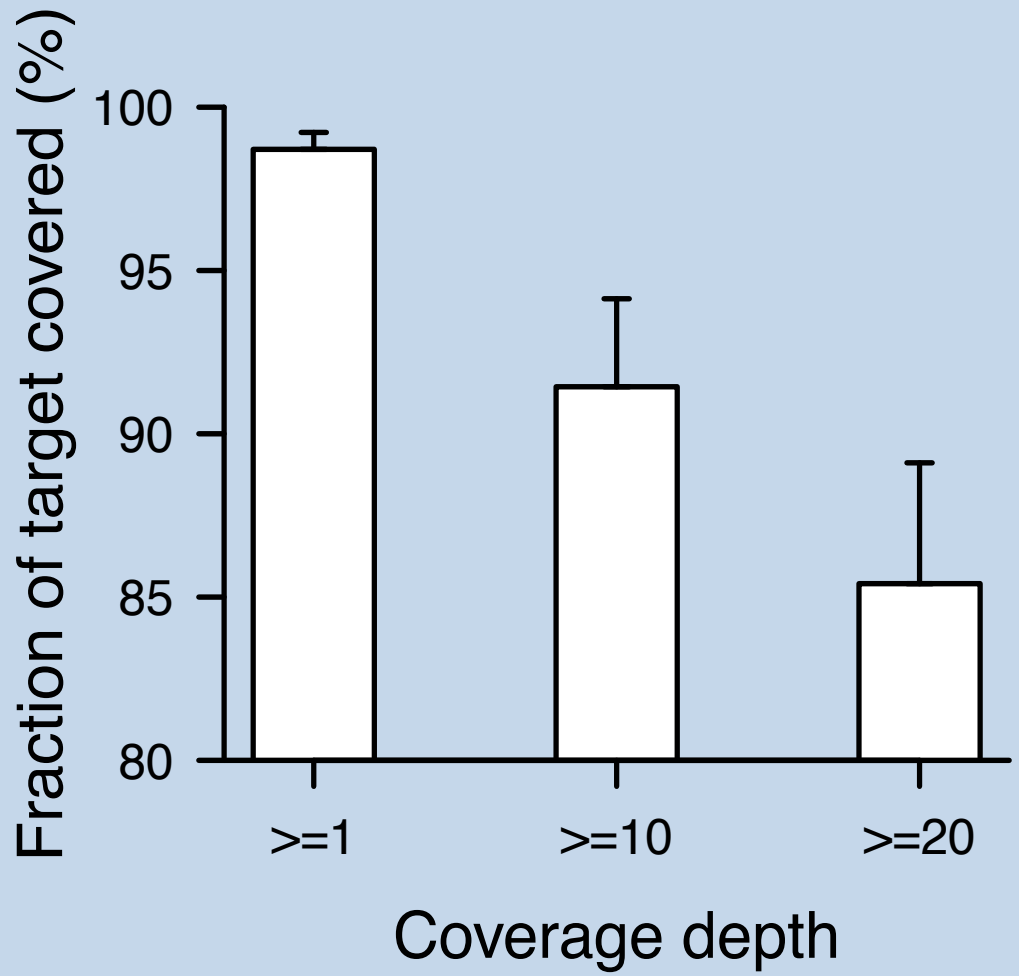
Optimizing Variant Calling in Exomes at BGI in 2011

- 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

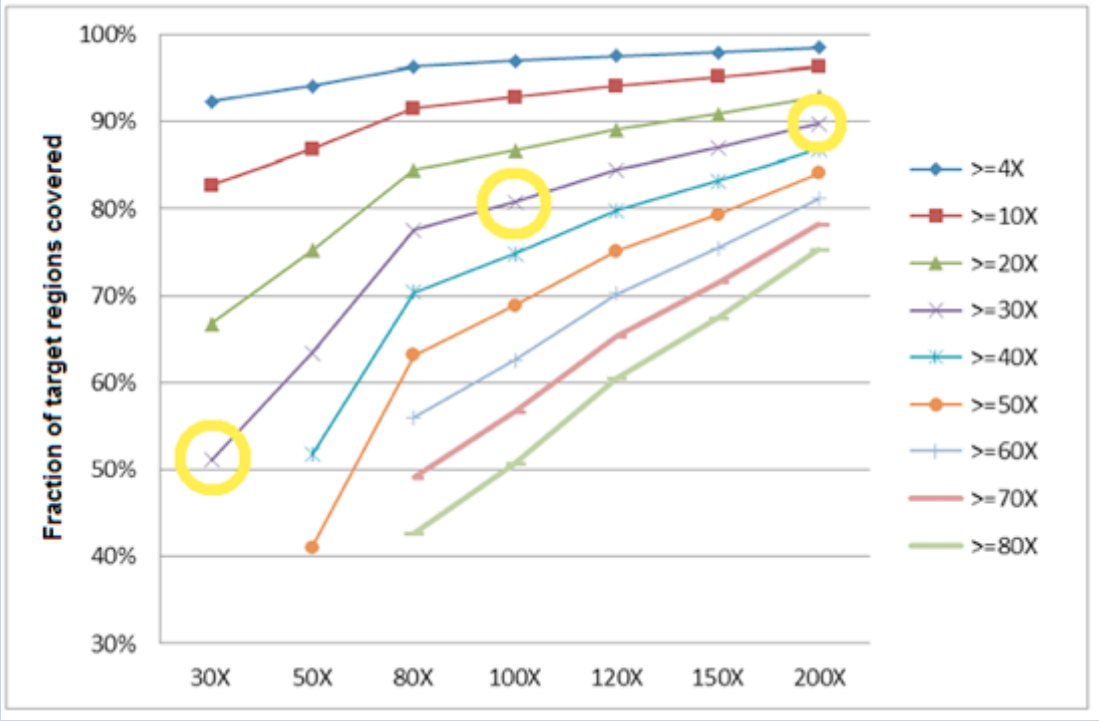


Figure from BGI website: http://bgiamericas.com/news-events/why-deep-exome-sequencing/

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?

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

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

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)

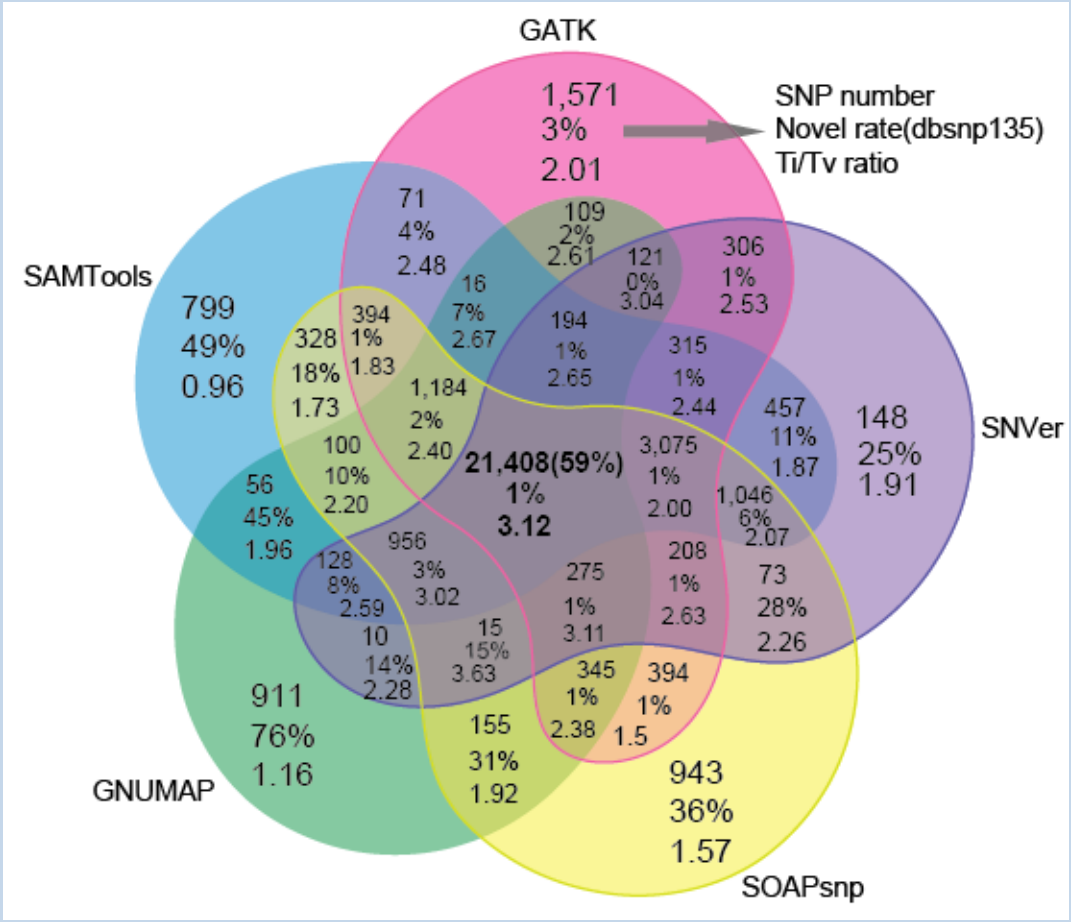
4) GNUMAP-SNP (probabilistic Pair-Hidden Markov which effectively accounts for uncertainty in the read calls as well as read mapping in an unbiased fashion)

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

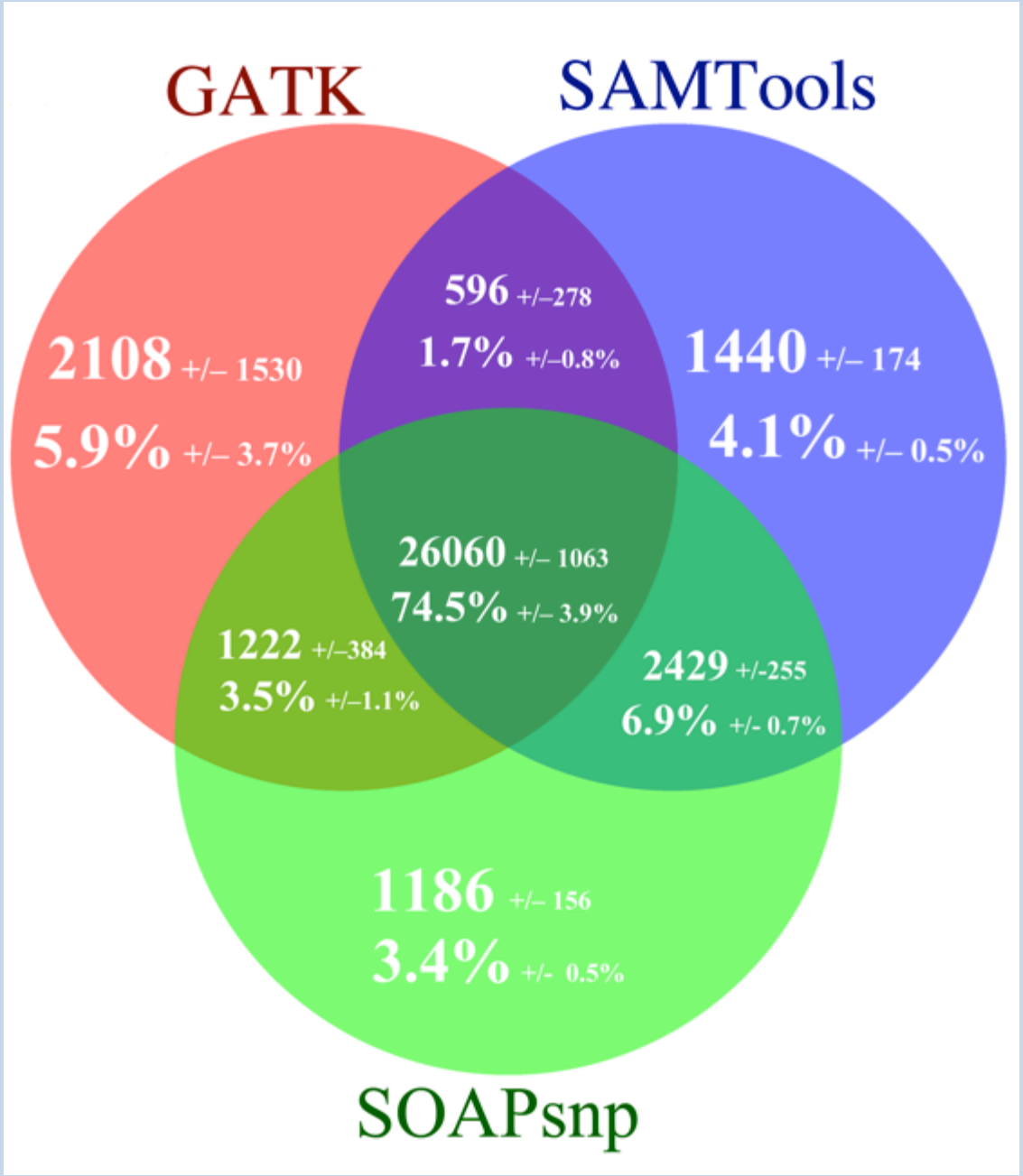
6) BWA - Sam format to Bam format - Picard to remove duplicates - **SCALPEL**

Total SNVs

A)

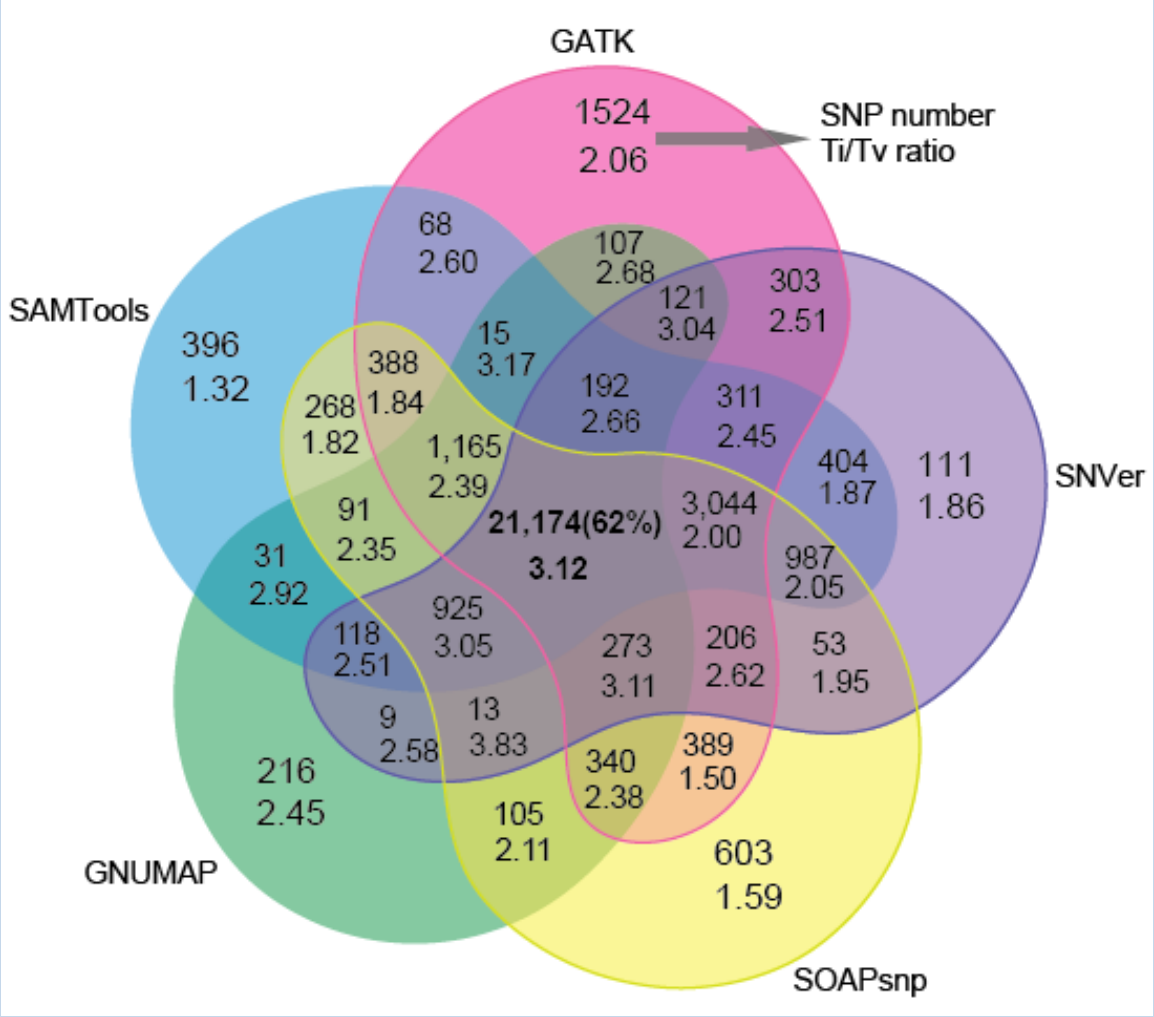


Mean # of total SNVs across 15 exomes, called by 5 pipelines. The percentage in the center of the the Venn diagram(Parenthesis) is the percent of total SNVs called by all five pipelines.


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

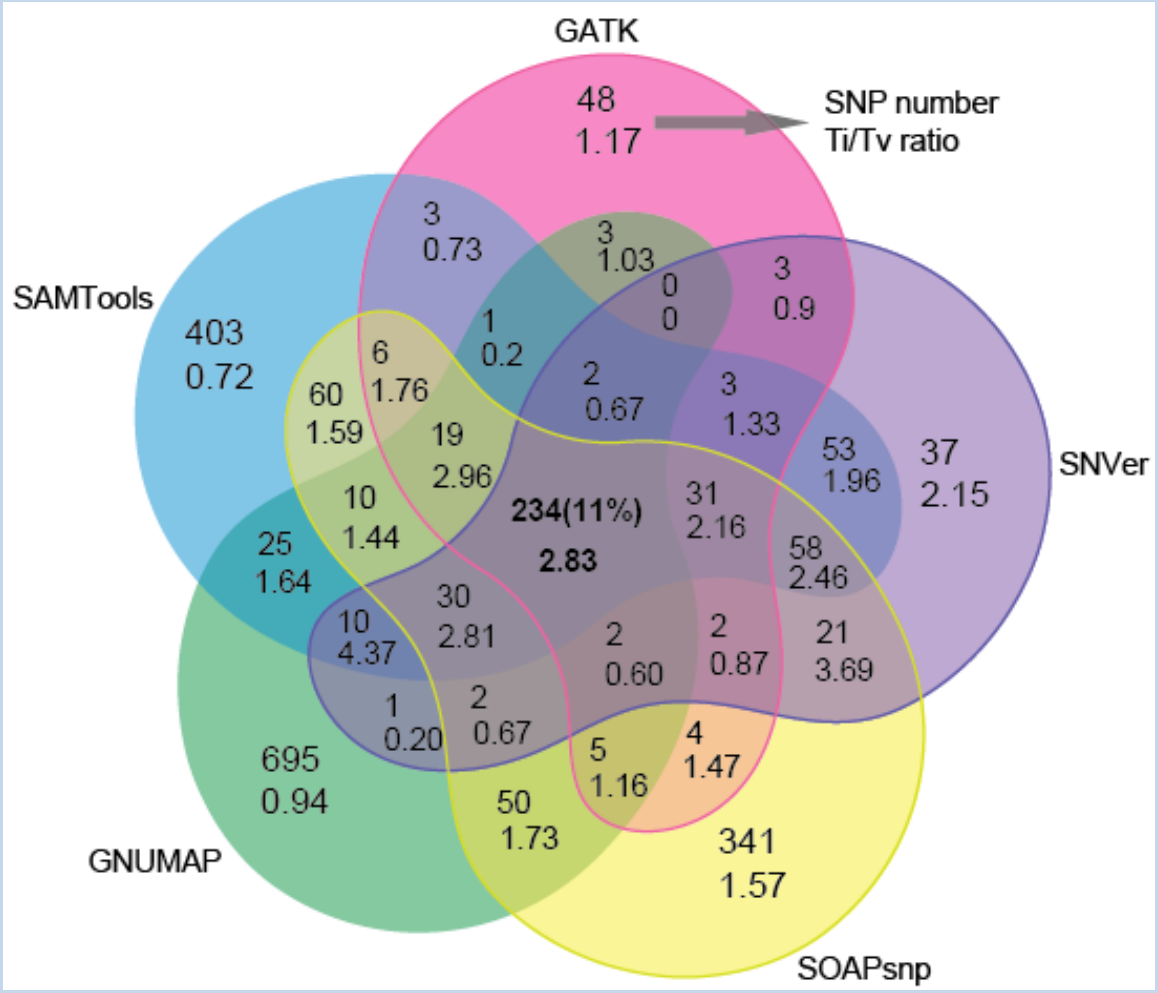
B)



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.

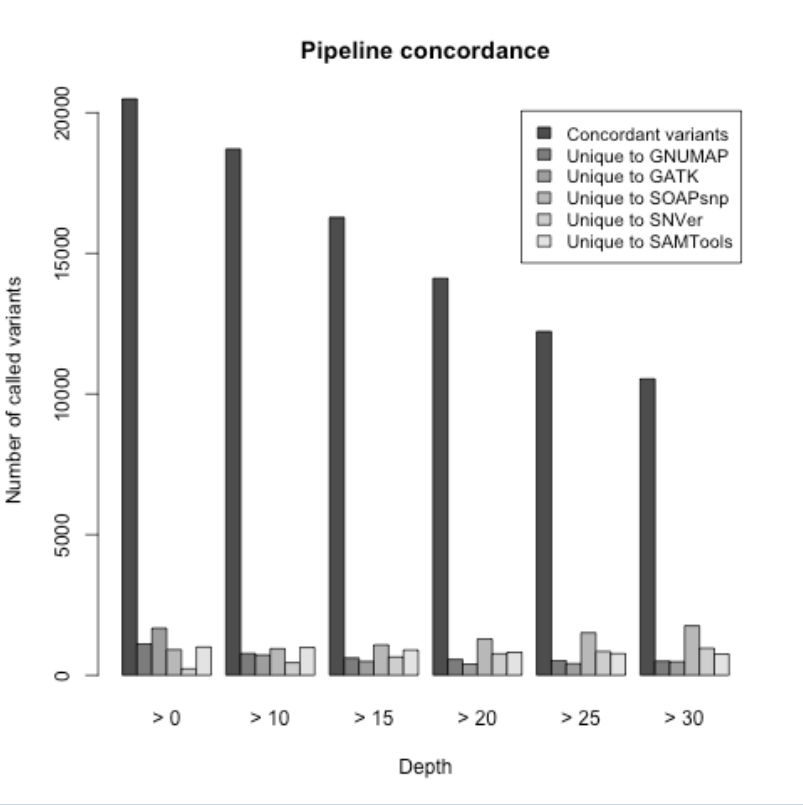
Novel SNVs

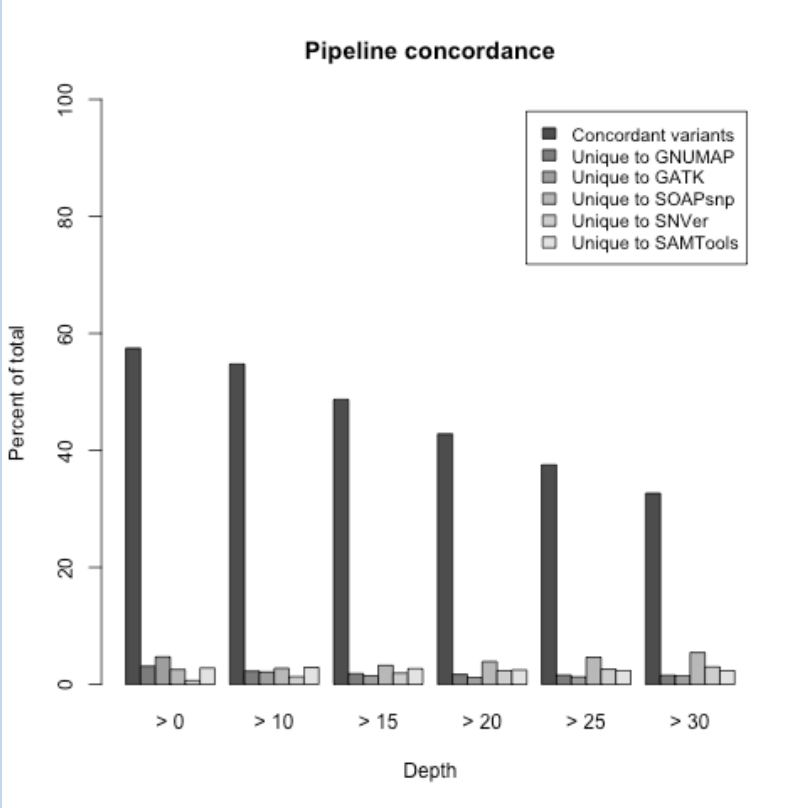
C)



- **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.

Comparing the concordance among the 5 pipelines used to analyze Illumina data, also stratified by read depth from >0 to >30 reads.

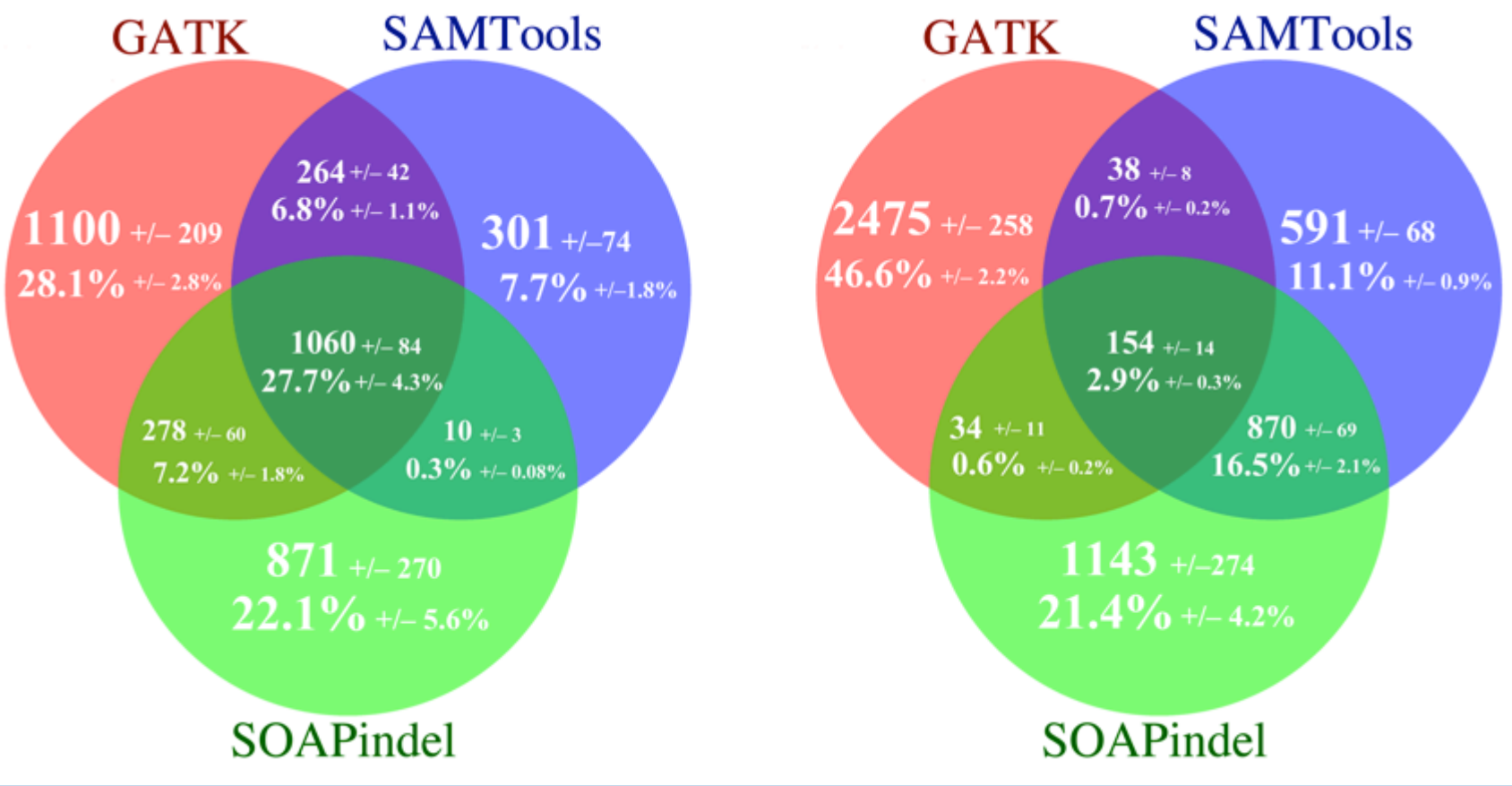




INDELS

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.

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

To conclude, 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.

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