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

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

CSH Cold Spring Harbor Laboratory
Stanley Institute for Cognitive Genomics

Acknowledgements

CSH Cold Spring Harbor Laboratory
STANLEY INSTITUTE FOR COGNITIVE GENOMICS

Jason O'Rawe

Yiyang Wu

USC TROJANS

Kai Wang

Michael Schatz
Giuseppe Narzisi
Eric Antoniou
Dick McCombie
Sequencing core facility

华大基因
BGI

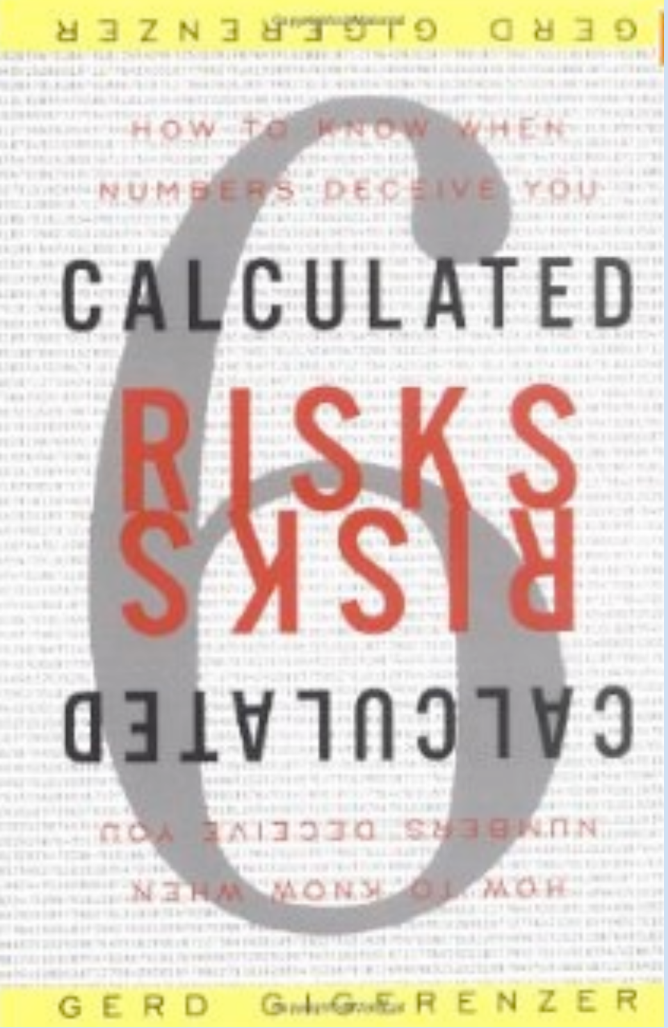
Tao Jiang
Guangqing Sun

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?

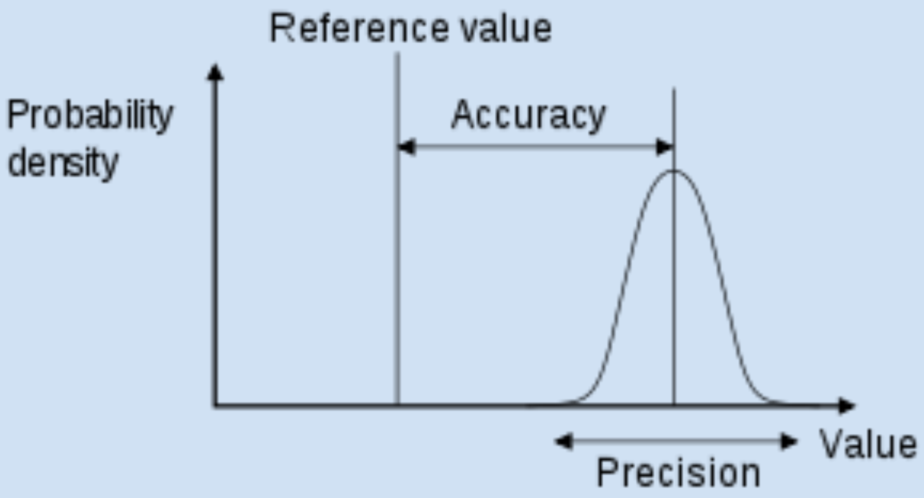
Illusions of Certainty. Everything is Probabilistic.

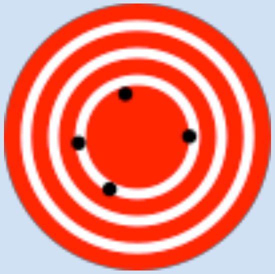


HOW TO KNOW WHEN NUMBERS DECEIVE YOU

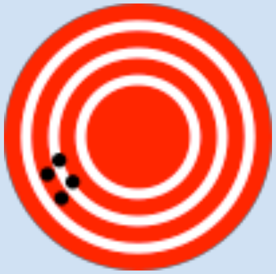
CALCULATED RISKS

GERD GIGERENZER





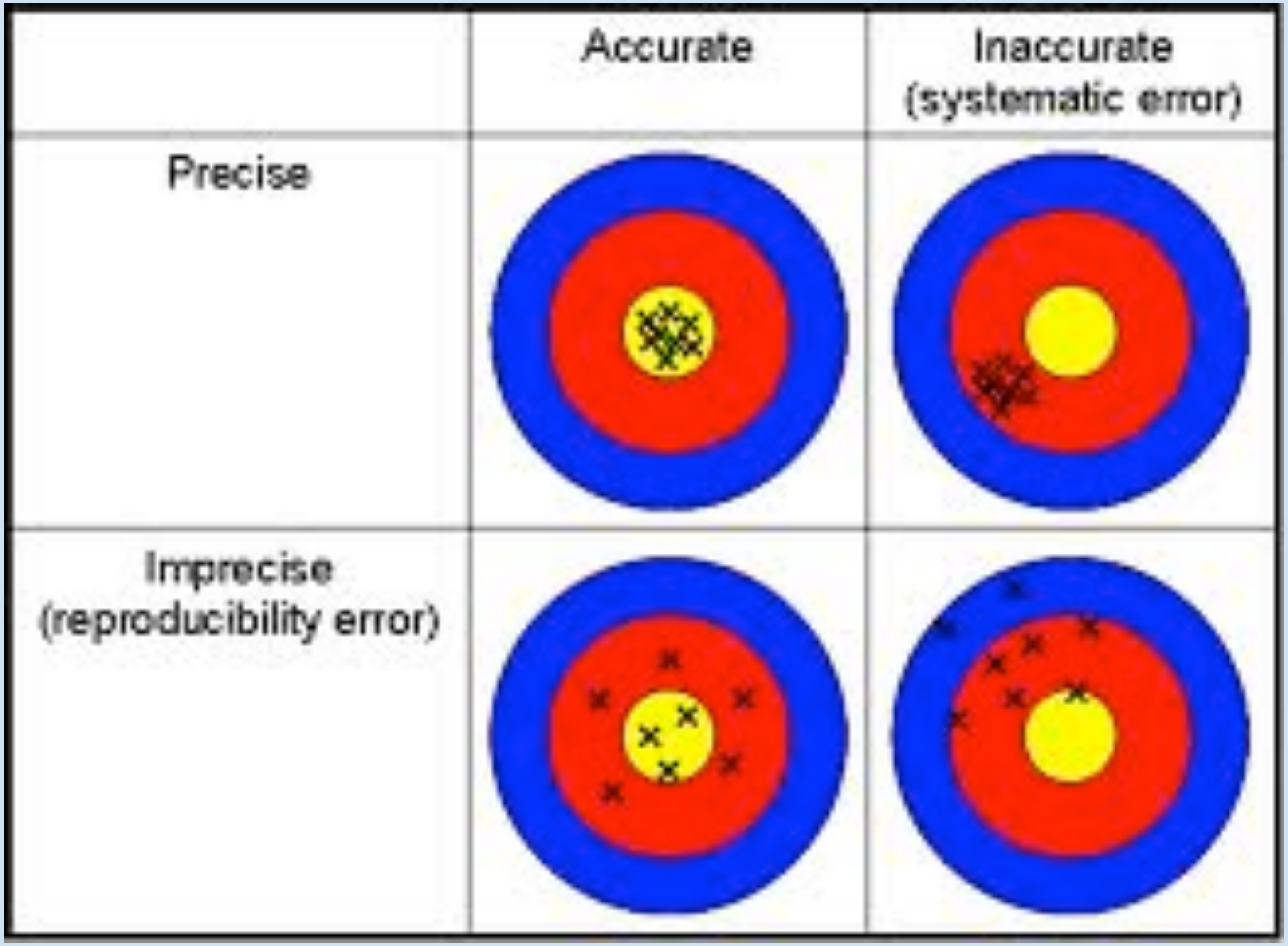
High accuracy, but low precision



High precision, but low accuracy

In the fields of science, engineering, industry, and statistics, the **accuracy** of a measurement system is the degree of closeness of measurements of a quantity to that quantity's actual (true) value. The **precision** of a measurement system, also called reproducibility or repeatability, is the degree to which repeated measurements under unchanged conditions show the same results.

http://en.wikipedia.org/wiki/Accuracy_and_precision

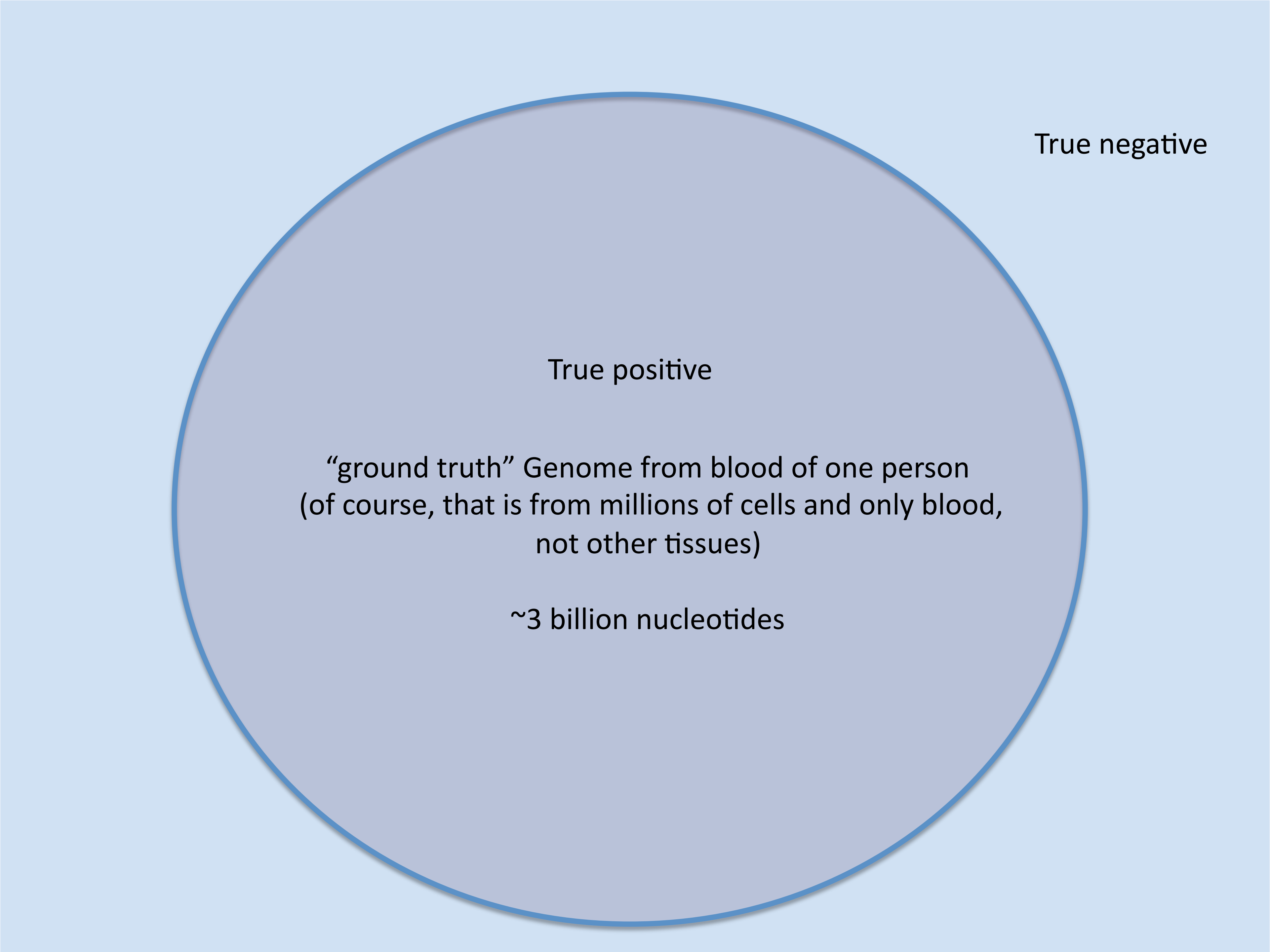


	Accurate	Inaccurate (systematic error)
Precise		
Imprecise (reproducibility error)		

Accuracy

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

An accuracy of 100% means that the measured values are exactly the same as the given values.

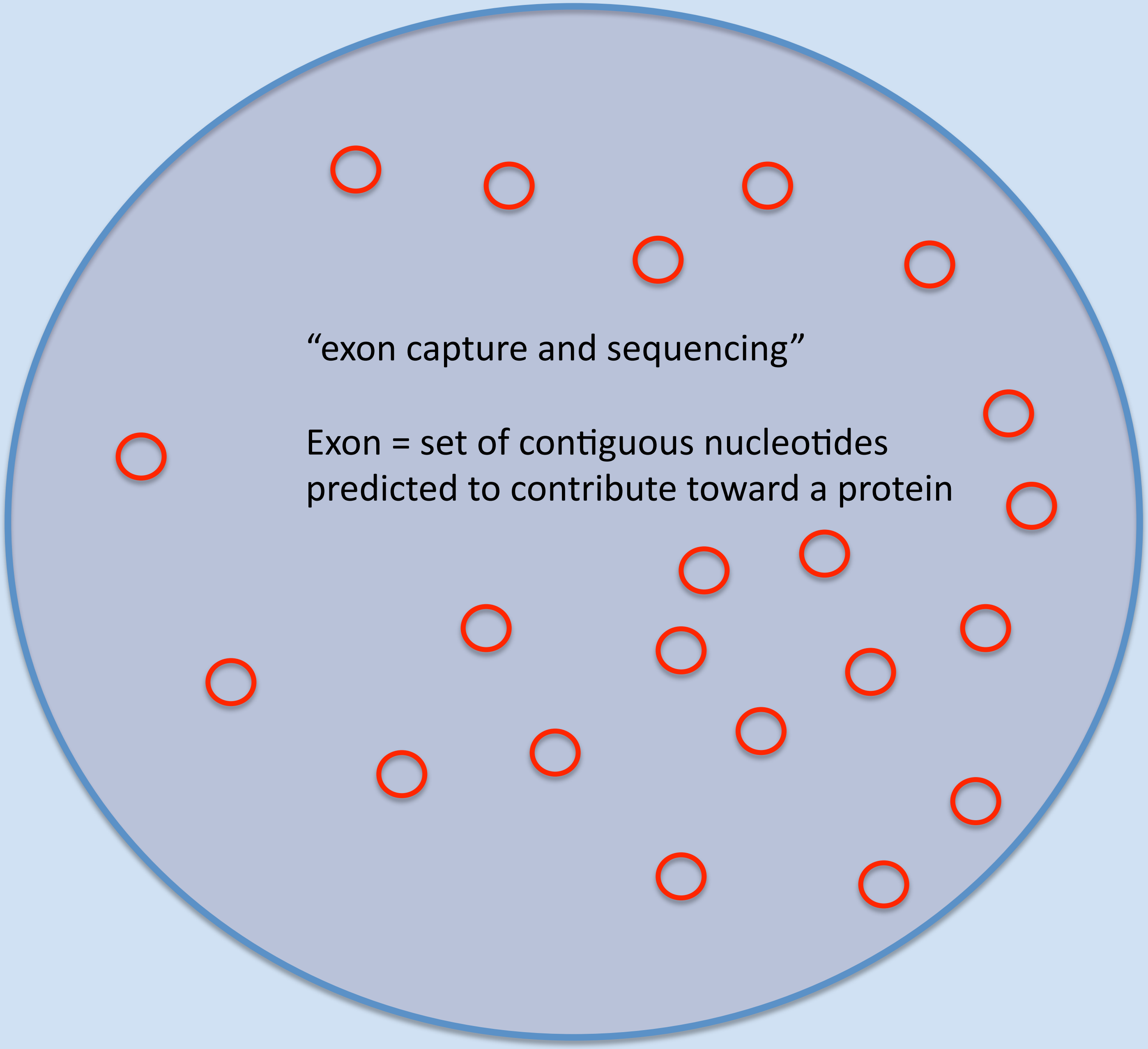


True negative

True positive

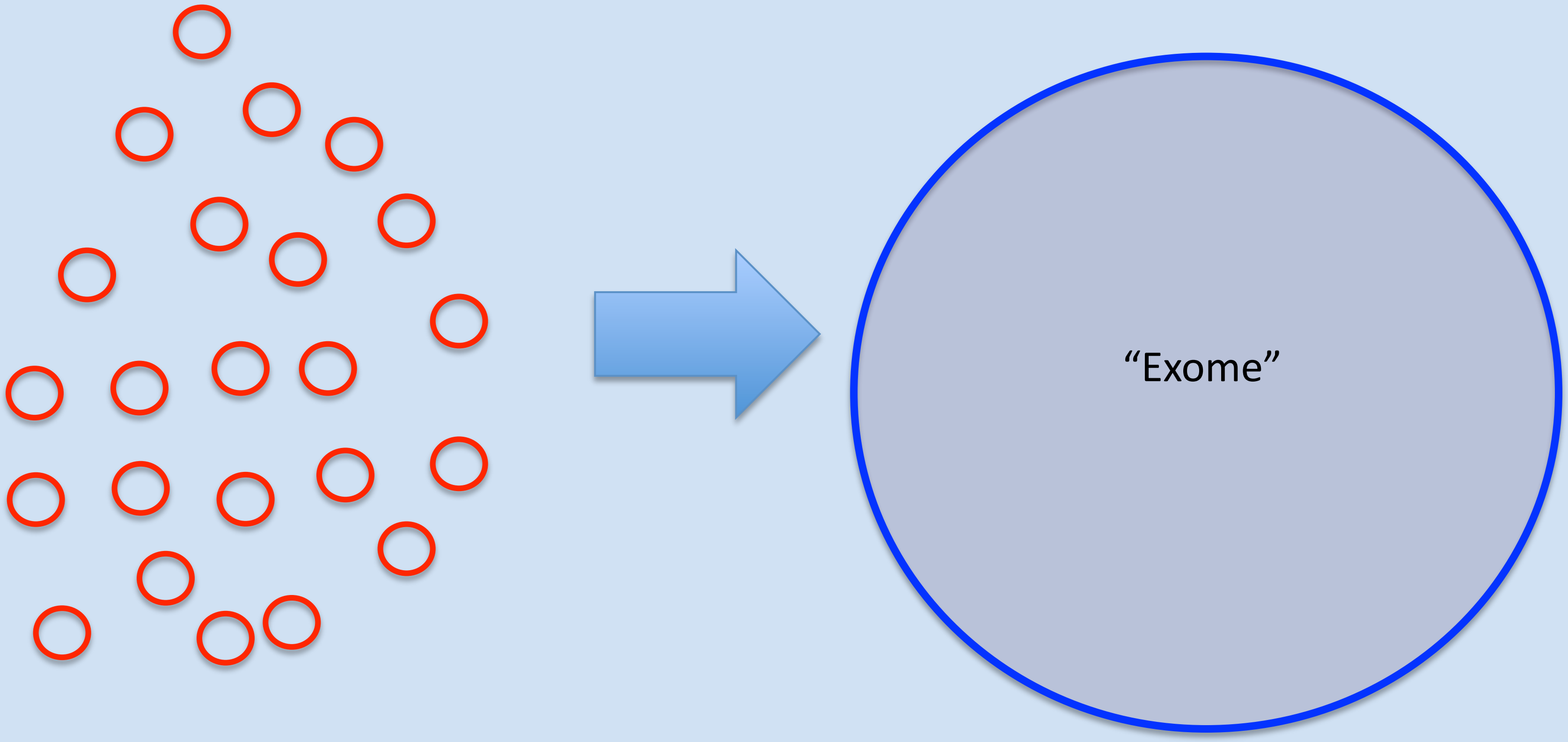
“ground truth” Genome from blood of one person
(of course, that is from millions of cells and only blood,
not other tissues)

~3 billion nucleotides



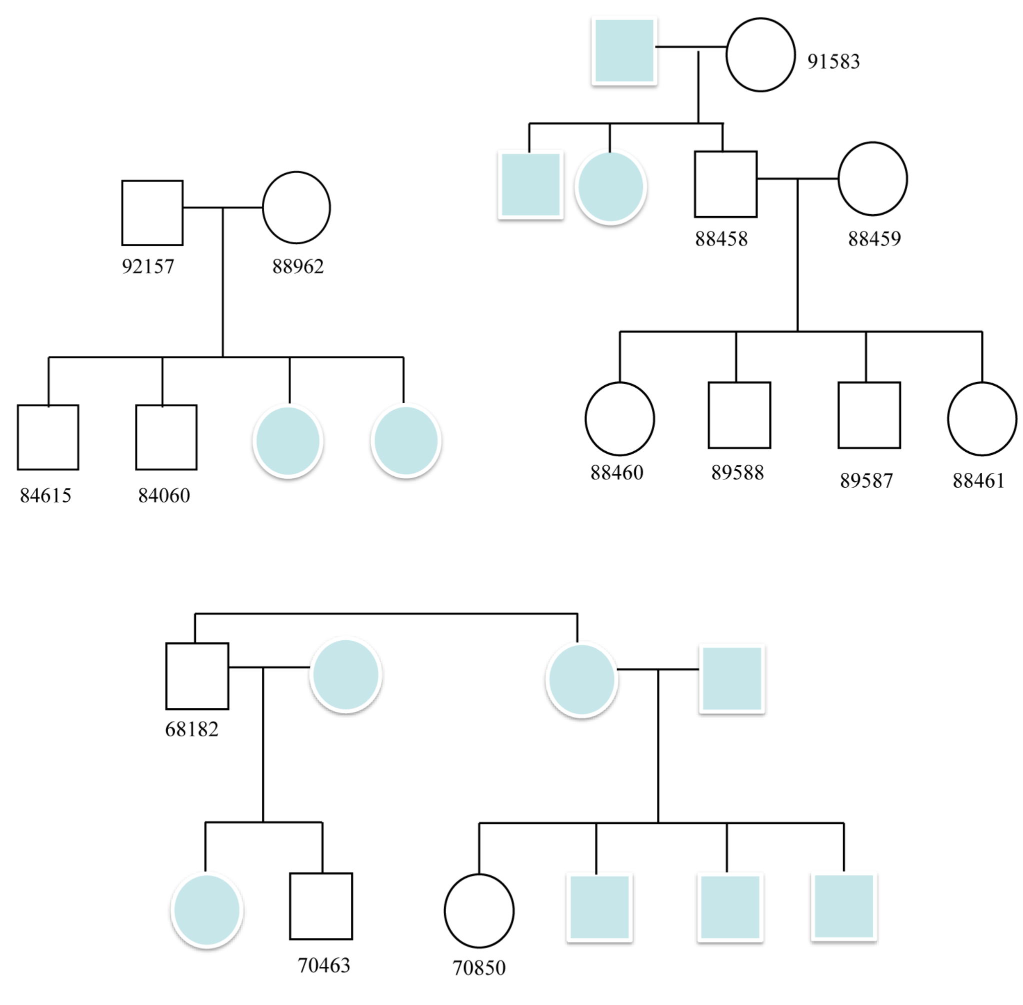
“exon capture and sequencing”

Exon = set of contiguous nucleotides predicted to contribute toward a protein

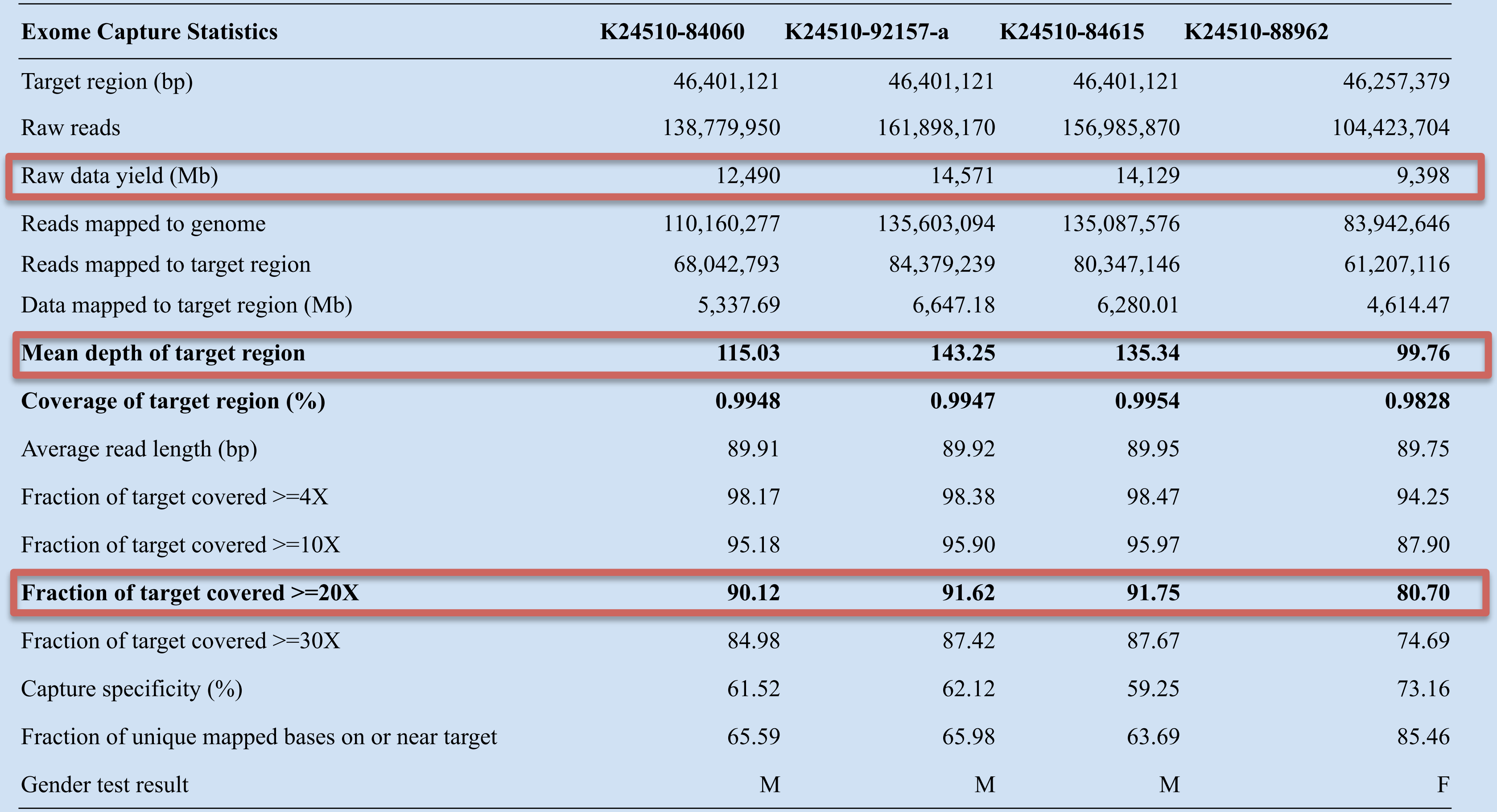


“Exome”

Chose to sequence 15 “exomes”

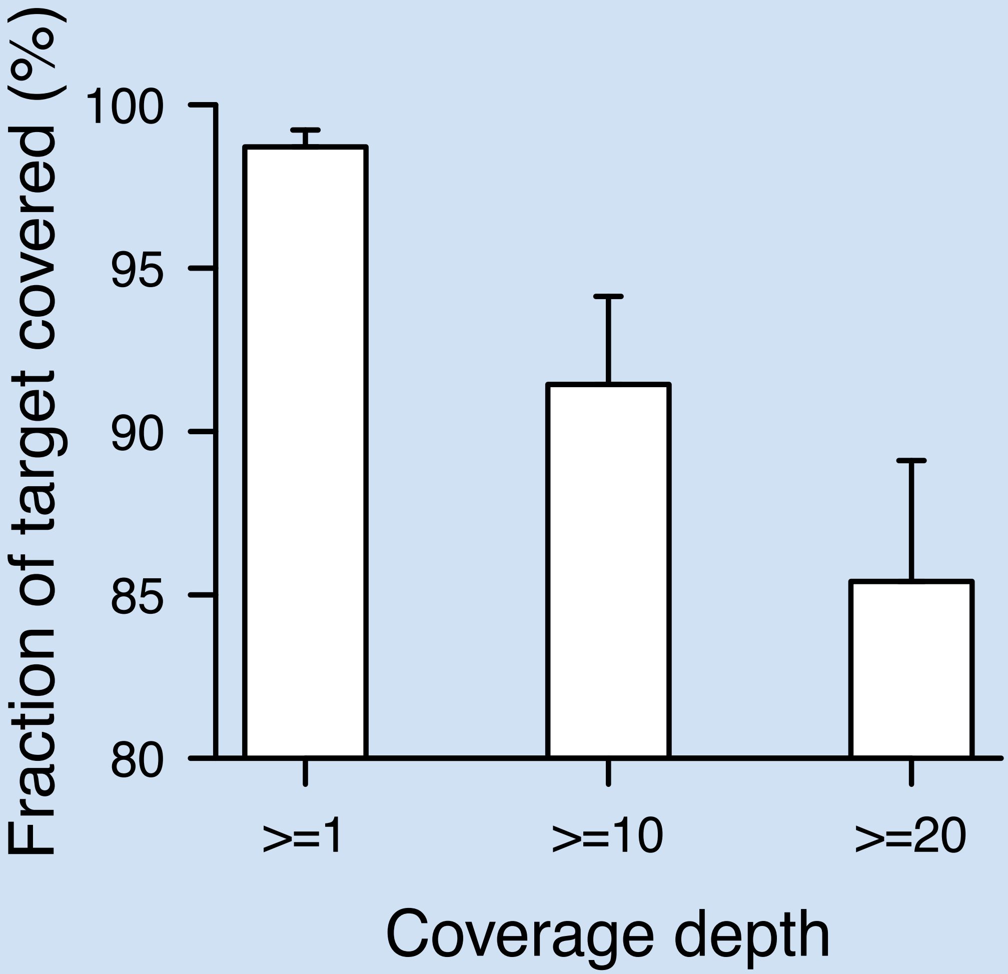


2-3 rounds of sequencing at BGI to attain goal of >80% of target region at >20 reads per base pair



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



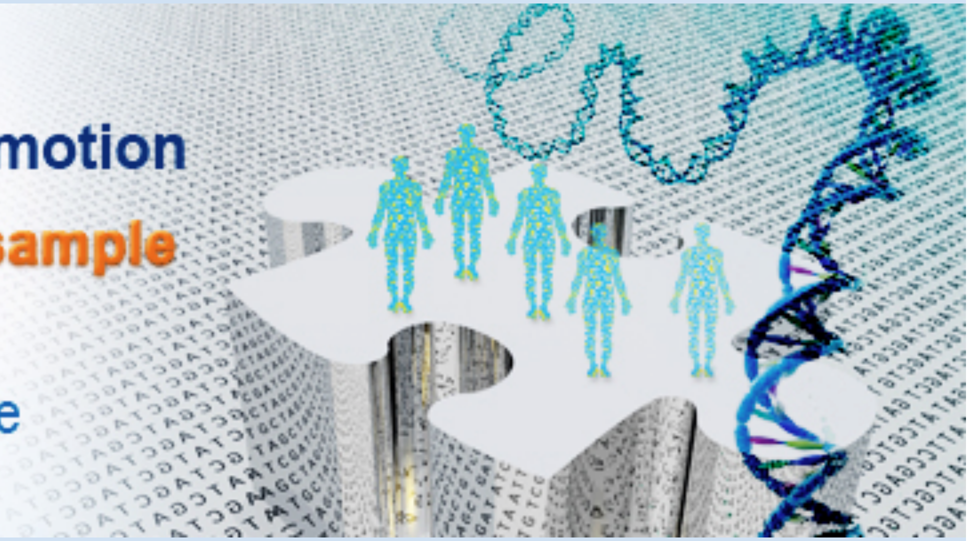
Experimental Design

- Evaluate robustness of variant calling implemented by different bioinformatics analysts.
- Looking at False Positives and False Negatives.
- How reliable are variants that are uniquely called by individual pipelines?
- Are some pipelines better at detecting rare, or novel variants than others?

Human Exome Sequencing Promotion

50X: $899/sample 100X: $1299/sample (SNP & Indel Included)

38,000 Exomes Sequenced by BGI to Date



Promotion Details (valid for Americas and Europe customers NOW through MAY 31)

A. The 899 USD/sample package – 50X human exome sequencing

Agilent SureSelect 50/51M Capture kit
100 bp paired-end sequencing on HiSeq 2000
5 Gb high quality* sequencing data
50X average coverage for target regions guaranteed
SNP & Indel calling and annotation included

B. The 1299 USD/sample package – 100X human exome sequencing

Agilent SureSelect 50/51M Capture kit
100 bp paired-end sequencing on HiSeq 2000
10 Gb high quality* sequencing data
100X average coverage for target regions guaranteed
SNP & Indel calling and annotation included

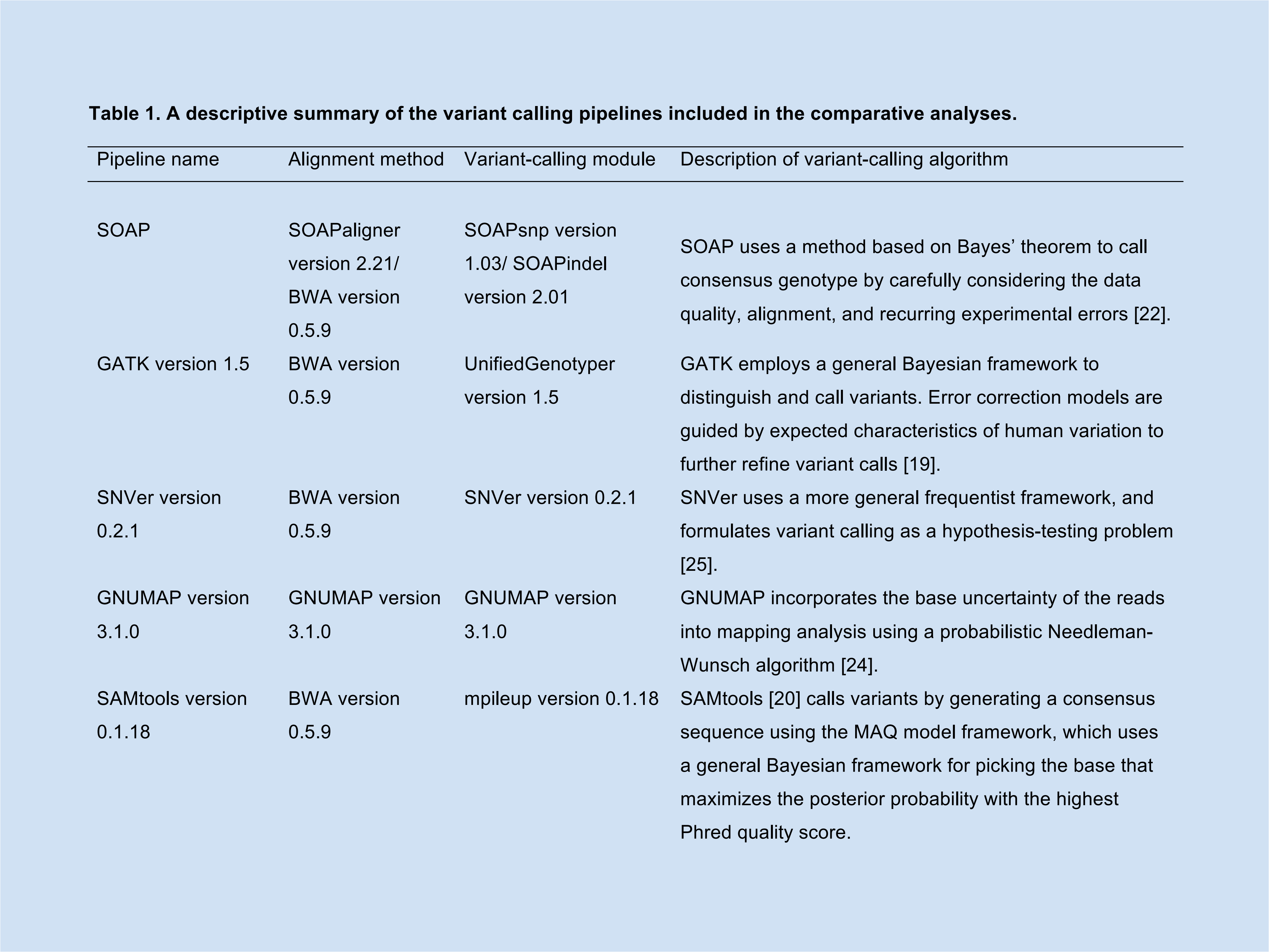
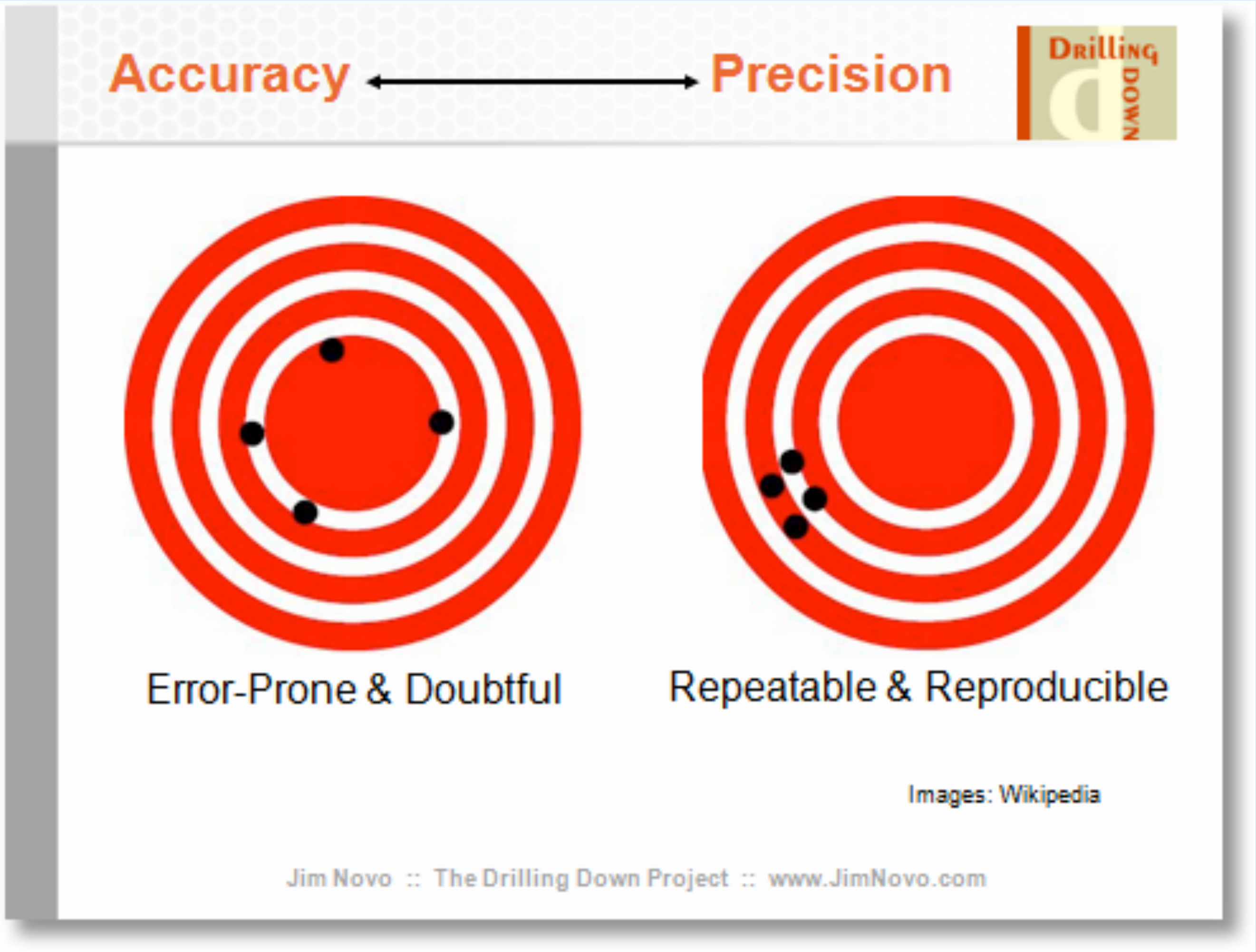


Table 1. A descriptive summary of the variant calling pipelines included in the comparative analyses.

Pipeline name	Alignment method	Variant-calling module	Description of variant-calling algorithm
SOAP	SOAPaligner version 2.21/ BWA version 0.5.9	SOAPsnp version 1.03/ SOAPindel version 2.01	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 version 1.5	BWA version 0.5.9	UnifiedGenotyper version 1.5	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 version 0.2.1	BWA version 0.5.9	SNVer version 0.2.1	SNVer uses a more general frequentist framework, and formulates variant calling as a hypothesis-testing problem [25].
GNUMAP version 3.1.0	GNUMAP version 3.1.0	GNUMAP version 3.1.0	GNUMAP incorporates the base uncertainty of the reads into mapping analysis using a probabilistic Needleman-Wunsch algorithm [24].
SAMtools version 0.1.18	BWA version 0.5.9	mpileup version 0.1.18	SAMtools [20] calls variants by generating a consensus sequence using the MAQ model framework, which uses a general Bayesian framework for picking the base that maximizes the posterior probability with the highest Phred quality score.



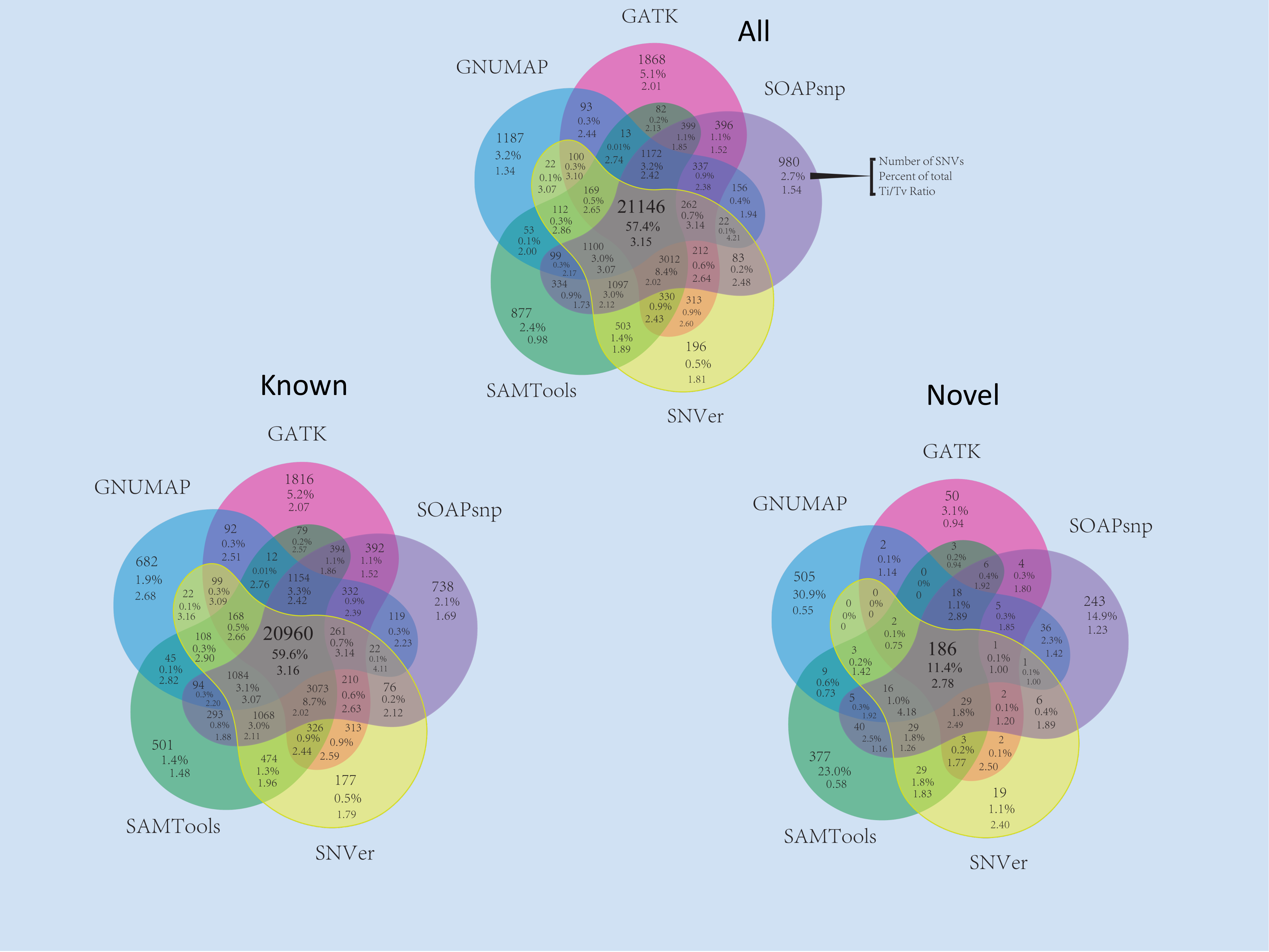
Accuracy ←→ Precision

Error-Prone & Doubtful

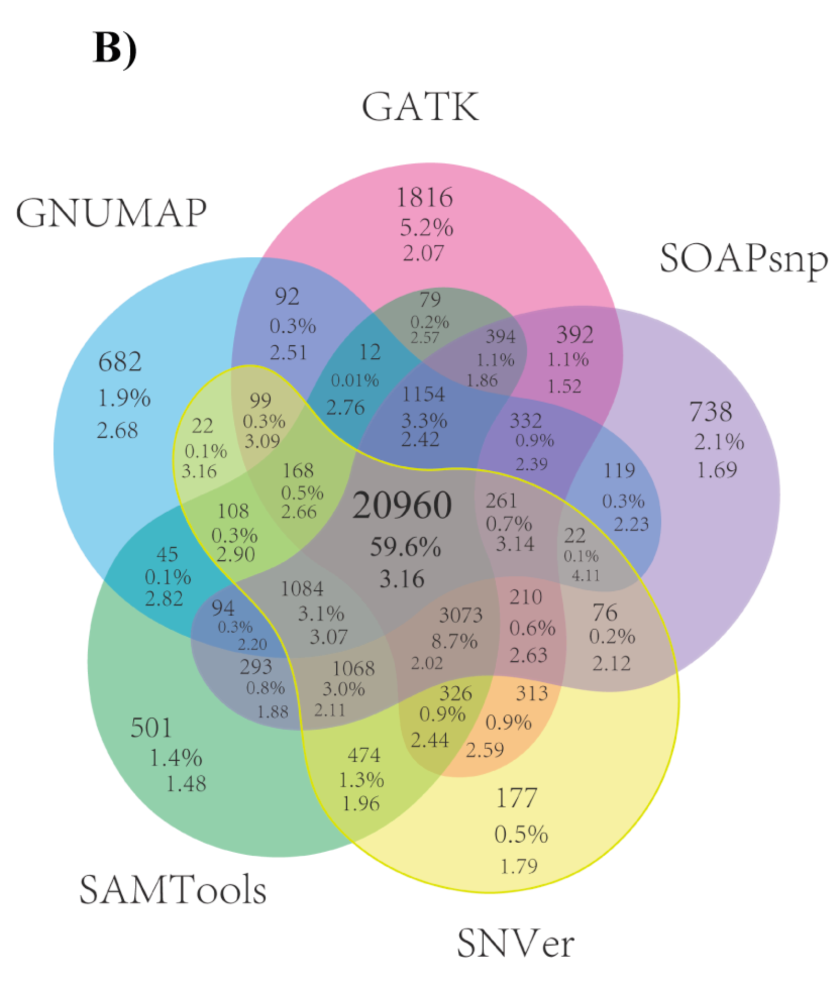
Repeatable & Reproducible

Images: Wikipedia

Jim Novo :: The Drilling Down Project :: www.JimNovo.com

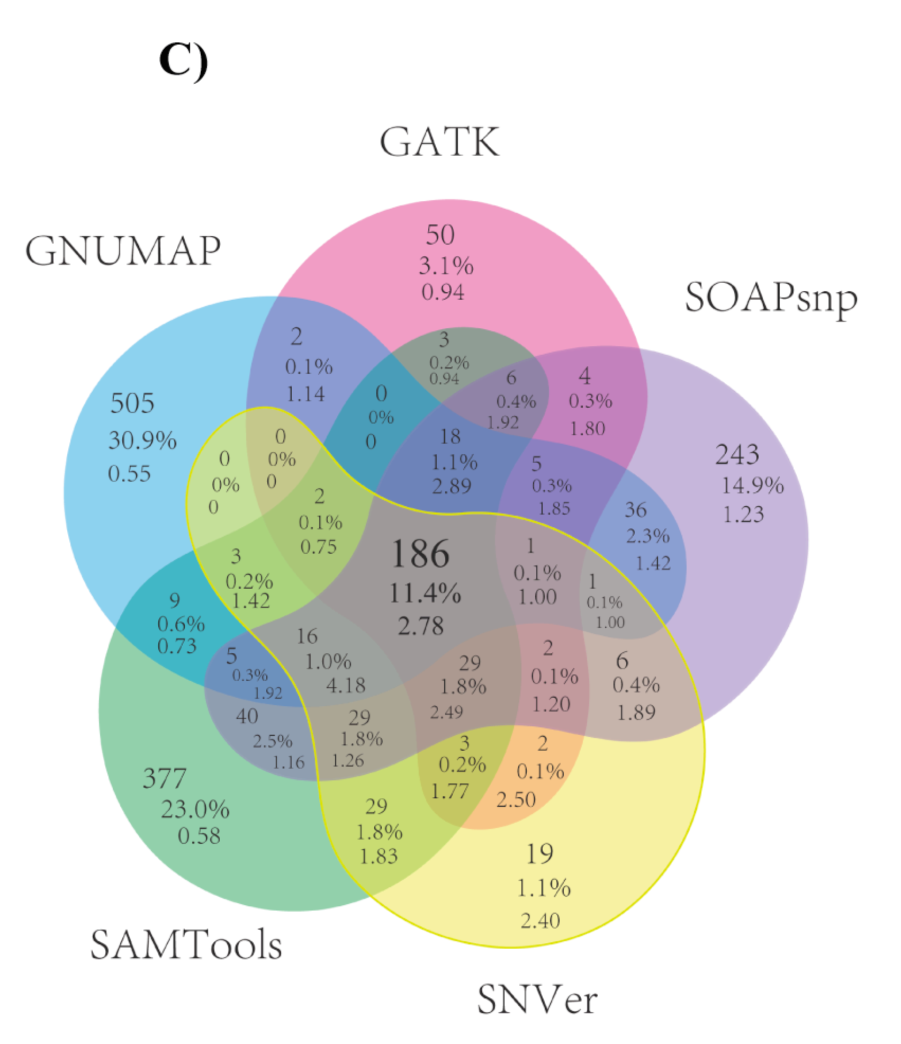

All
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1868 5.1% 2.01
1187 3.2% 1.34
980 2.7% 1.54
877 2.4% 0.98
196 0.5% 1.81
93 0.3% 2.44
82 0.2% 2.13
13 0.01% 2.74
399 1.1% 1.85
396 1.1% 1.52
22 0.1% 3.07
100 0.3% 3.10
1172 3.2% 2.42
337 0.9% 2.38
169 0.5% 2.65
21146 57.4% 3.15
156 0.4% 1.94
112 0.3% 2.86
262 0.7% 3.14
22 0.1% 4.21
53 0.1% 2.00
1100 3.0% 3.07
99 0.3% 2.17
3012 8.4% 2.02
212 0.6% 2.64
83 0.2% 2.48
334 0.9% 1.73
1097 3.0% 2.12
330 0.9% 2.43
313 0.9% 2.60
503 1.4% 1.89
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
92 0.3% 2.51
79 0.2% 2.57
12 0.01% 2.76
394 1.1% 1.86
392 1.1% 1.52
22 0.1% 3.16
99 0.3% 3.09
1154 3.3% 2.42
332 0.9% 2.39
168 0.5% 2.66
20960 59.6% 3.16
119 0.3% 2.23
108 0.3% 2.90
261 0.7% 3.14
22 0.1% 4.11
45 0.1% 2.82
1084 3.1% 3.07
94 0.3% 2.20
3073 8.7% 2.02
210 0.6% 2.63
76 0.2% 2.12
293 0.8% 1.88
1068 3.0% 2.11
326 0.9% 2.44
313 0.9% 2.59
474 1.3% 1.96
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
2 0.1% 1.14
3 0.2% 0.94
0 0% 0
6 0.4% 1.92
4 0.3% 1.80
0 0% 0
0 0% 0
18 1.1% 2.89
5 0.3% 1.85
2 0.1% 0.75
186 11.4% 2.78
36 2.3% 1.42
3 0.2% 1.42
1 0.1% 1.00
1 0.1% 1.00
9 0.6% 0.73
16 1.0% 4.18
5 0.3% 1.92
29 1.8% 2.49
2 0.1% 1.20
6 0.4% 1.89
40 2.5% 1.16
29 1.8% 1.26
3 0.2% 1.77
2 0.1% 2.50
29 1.8% 1.83

Known SNVs



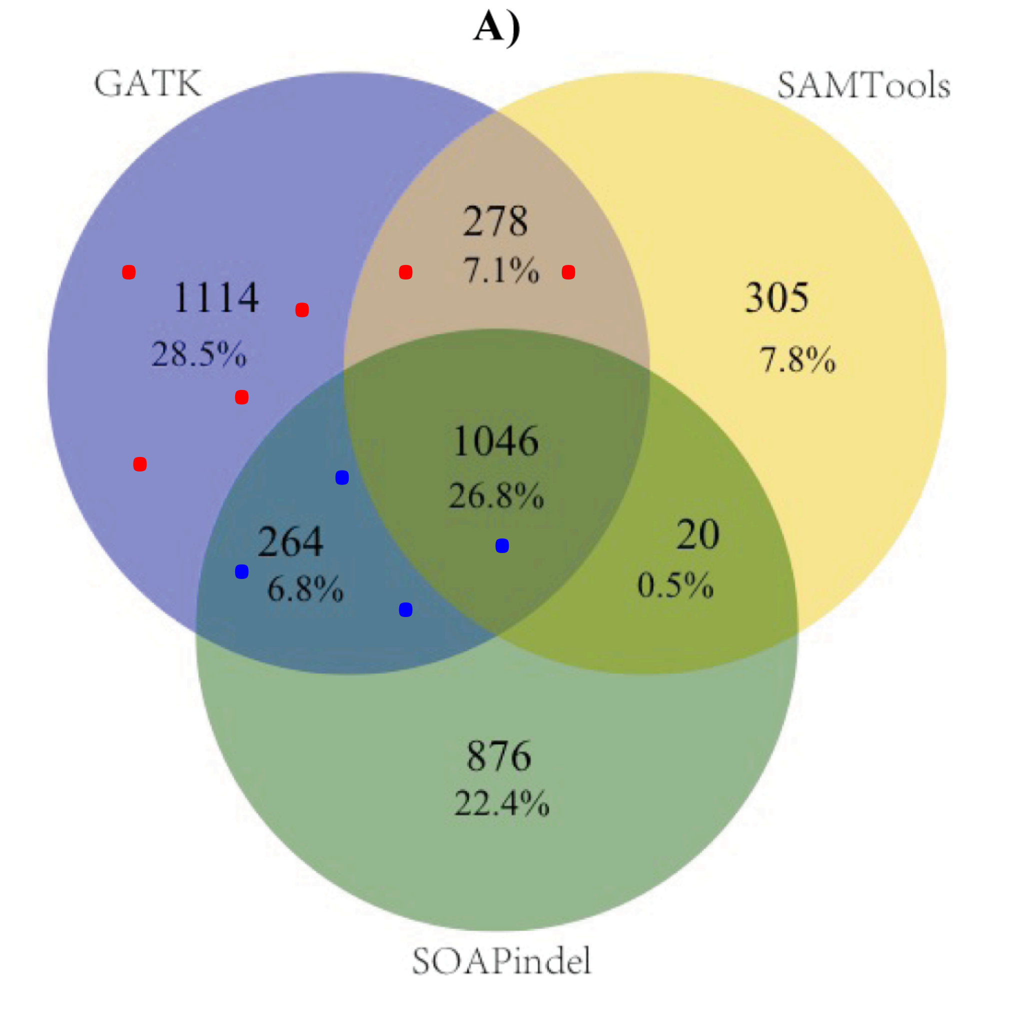
B) Mean # of known SNVs (present in dbSNP135) found by 5 pipelines across 15 exomes. The percentage in the center of the the Venn diagram is the percent of known SNVs called by all five pipelines.

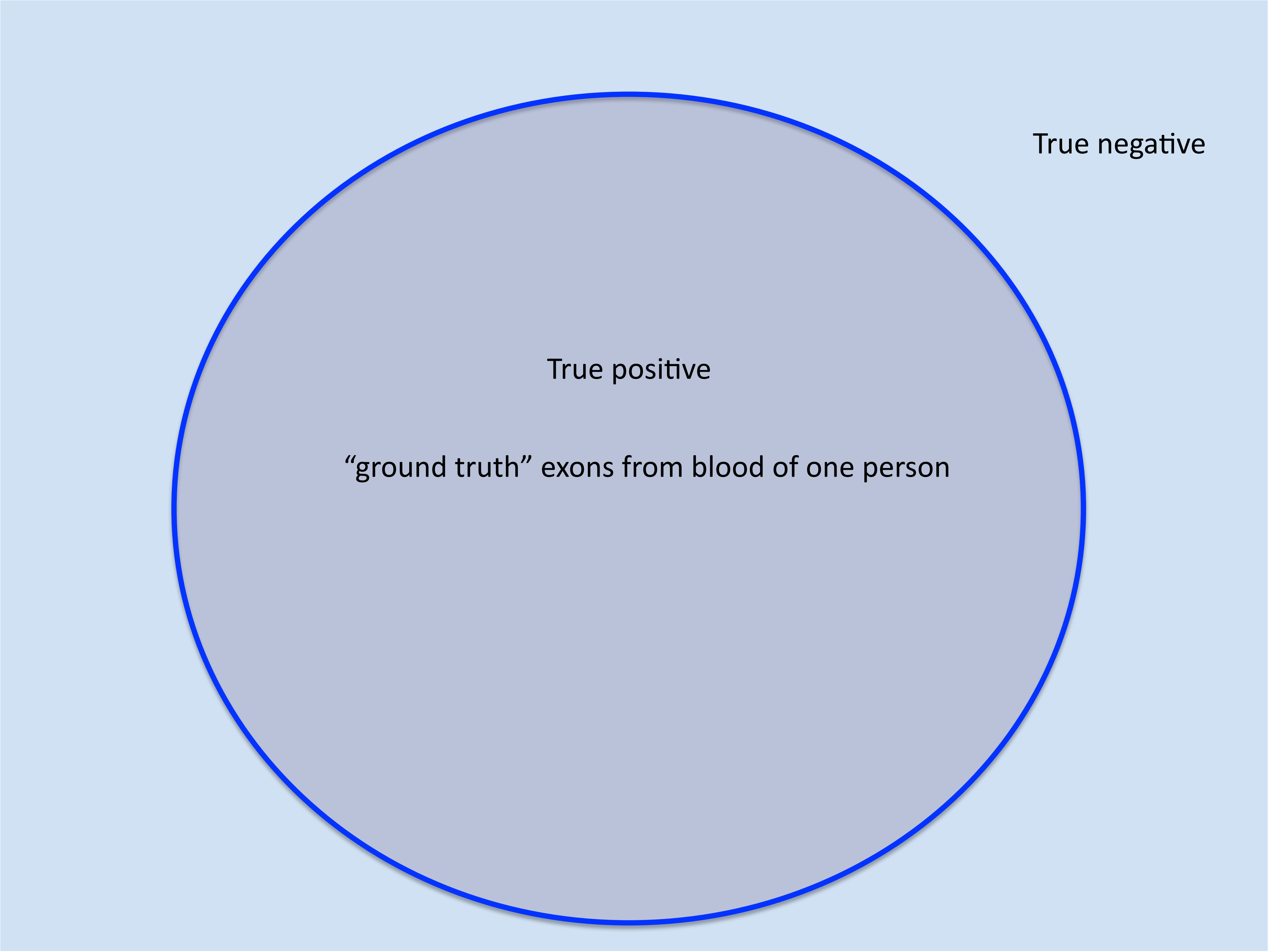
Novel SNVs



C) 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 called by GATK, SOAP and SAMtools

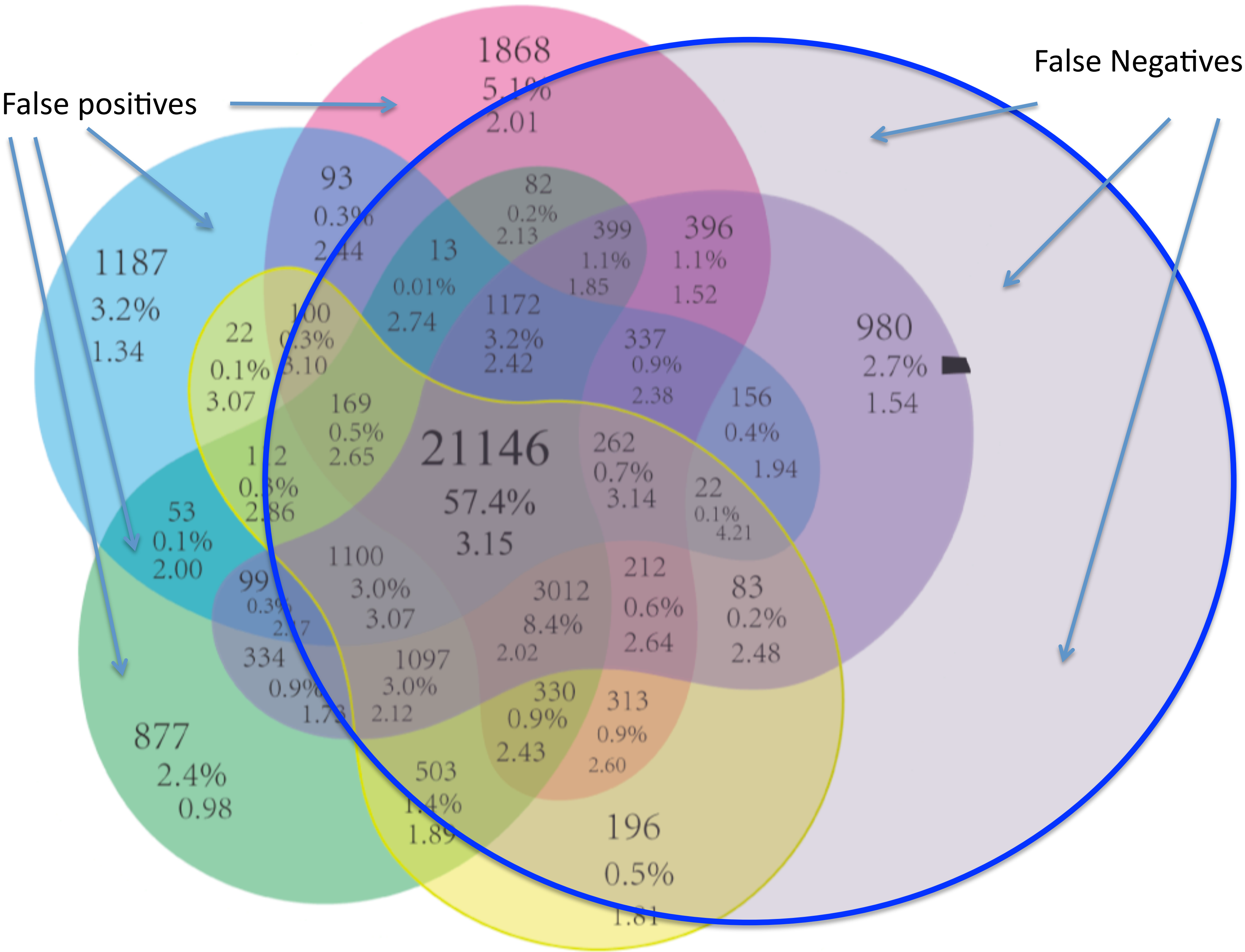




True negative

True positive

“ground truth” exons from blood of one person

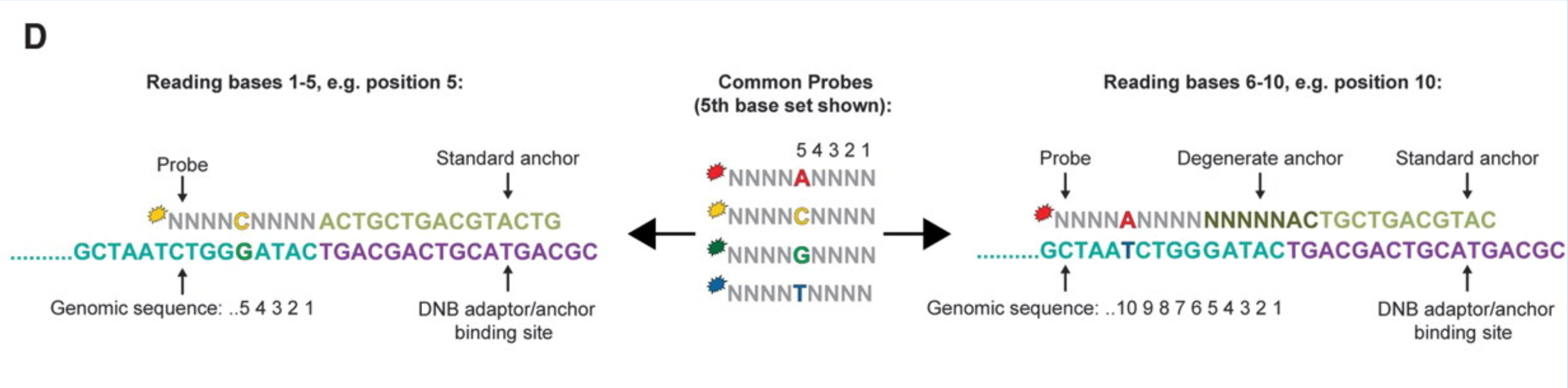


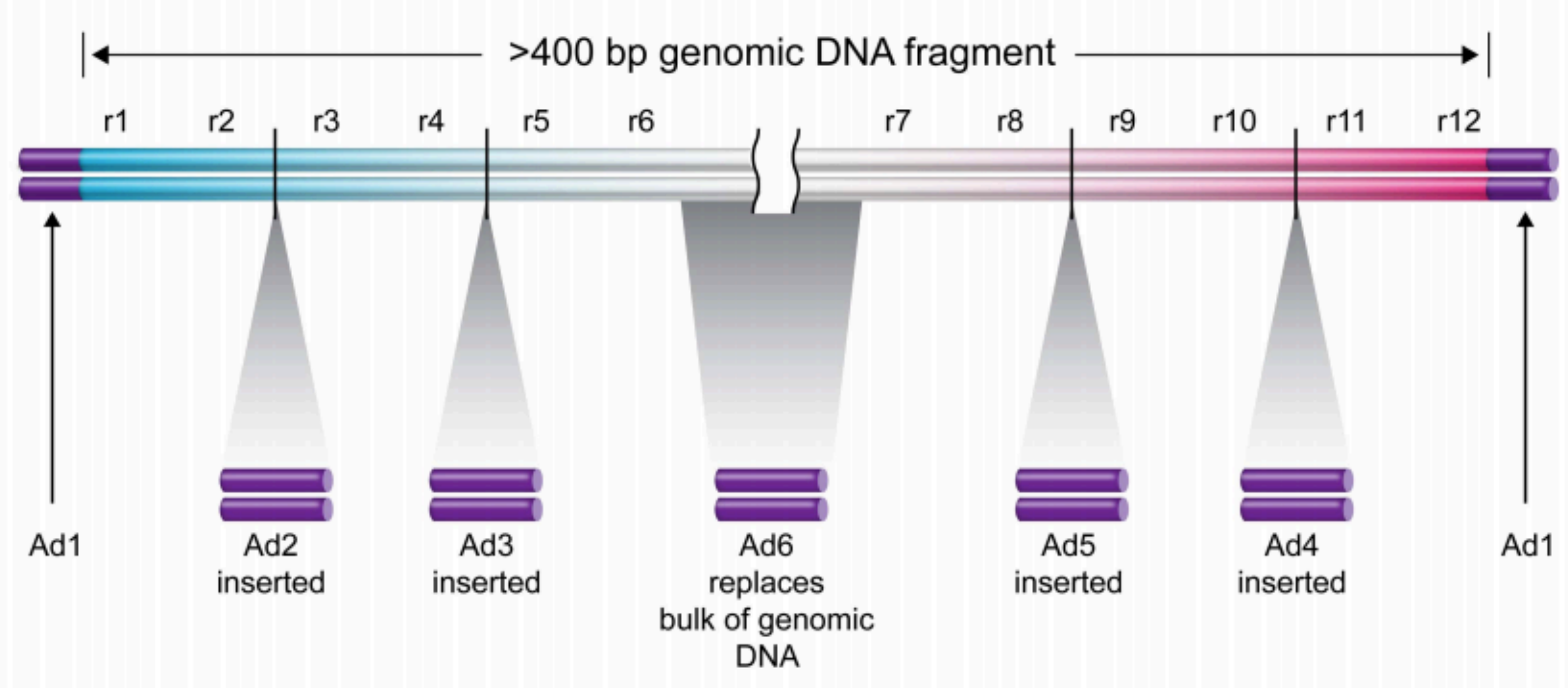
False positives

False Negatives

Cross validation using orthogonal sequencing technology (Complete Genomics)

Complete Genomics chemistry - combinatorial probe anchor ligation (cPAL)


D
Reading bases 1-5, e.g. position 5:
Probe
Standard anchor
NNNNCNNNNN ACTGCTGACGTACTG
..........GCTAATCTGGGATACTGACGACTGCATGACGC
Genomic sequence: ..5 4 3 2 1
DNB adaptor/anchor binding site
Common Probes (5th base set shown):
5 4 3 2 1
NNNNANNNNN
NNNNCNNNNN
NNNNGNNNNN
NNNNTNNNNN
Reading bases 6-10, e.g. position 10:
Probe
Degenerate anchor
Standard anchor
NNNNANNNNNNNNNNACTGCTGACGTAC
..........GCTAATCTGGGATACTGACGACTGCATGACGC
Genomic sequence: ..10 9 8 7 6 5 4 3 2 1
DNB adaptor/anchor binding site



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.*

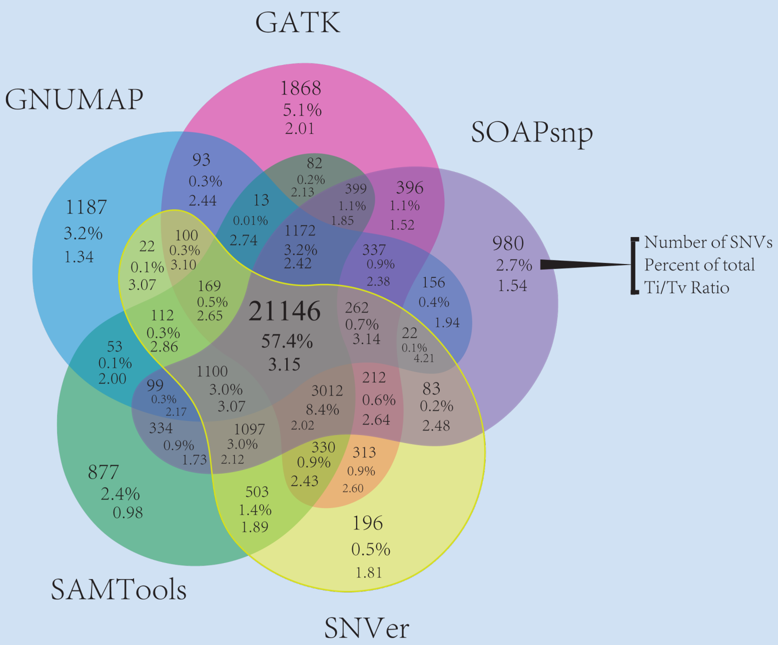
COMPLETE GENOMICS CALL	OTHER PLATFORM	PLATFORM-SPECIFIC SNVs	VALIDATION RATE	EST FPs	FPR
Het or Hom SNV	No SNV Reported	99K	17/18 = 94.4%	5,577	0.16%
No-call or Hom-Ref	SNV Reported	345K	2/15 = 13.3%	299,115	8.2%

***Table 3.** False Positive Rate.*

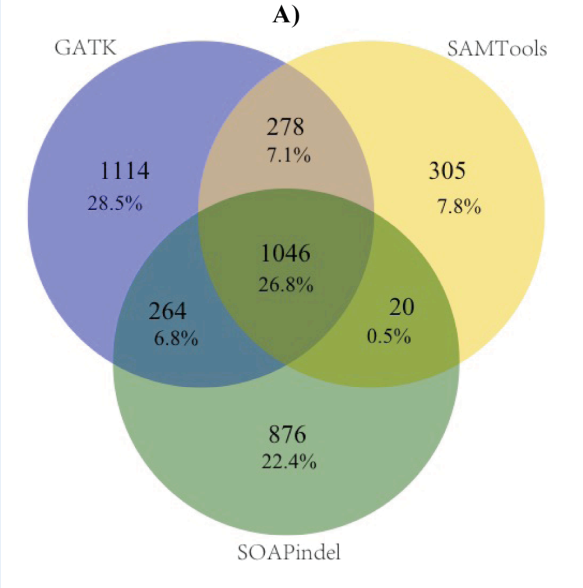
Performance comparison of whole-genome sequencing platforms

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]

VOLUME 30 NUMBER 1 JANUARY 2012 NATURE BIOTECHNOLOGY

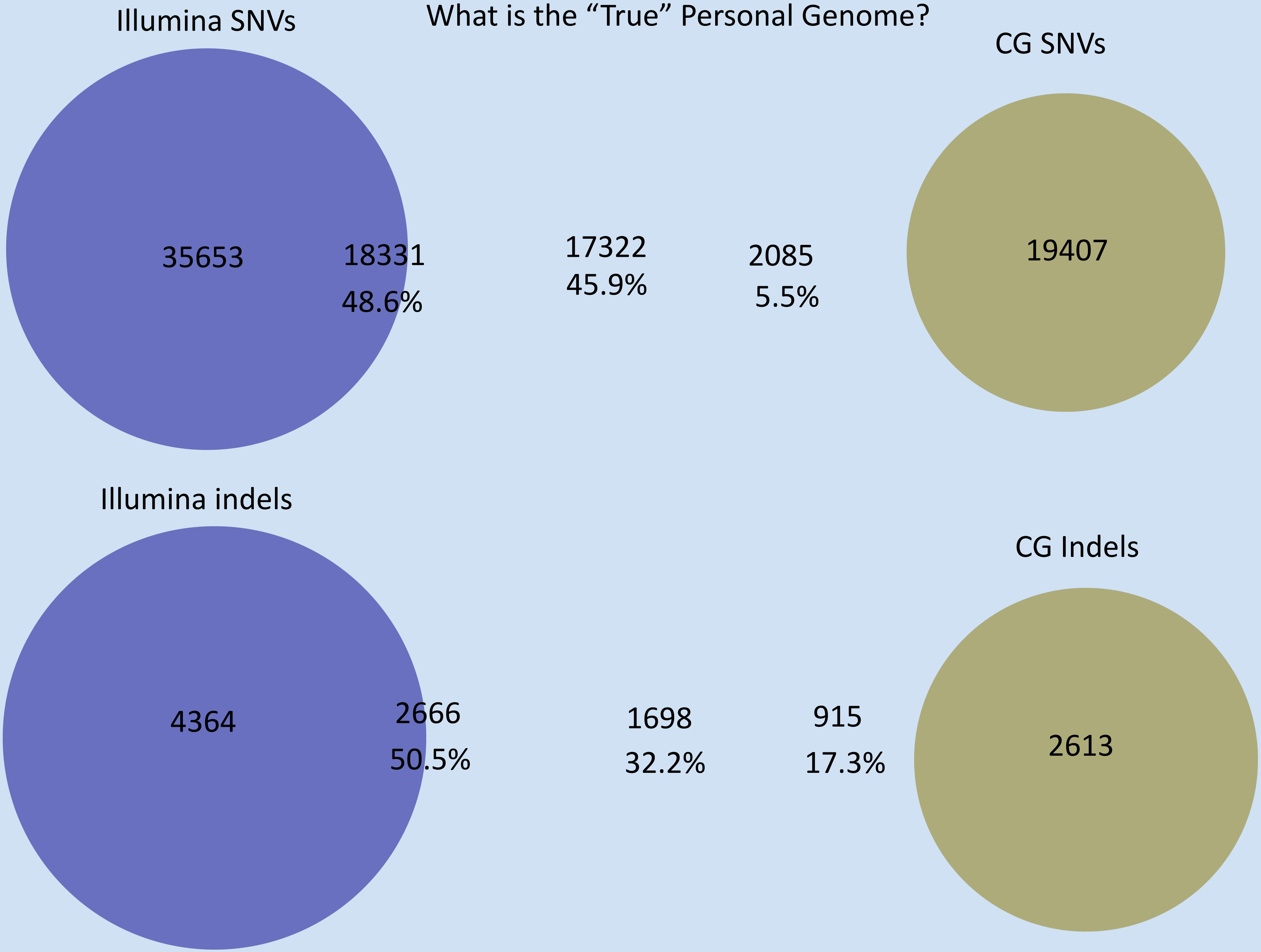


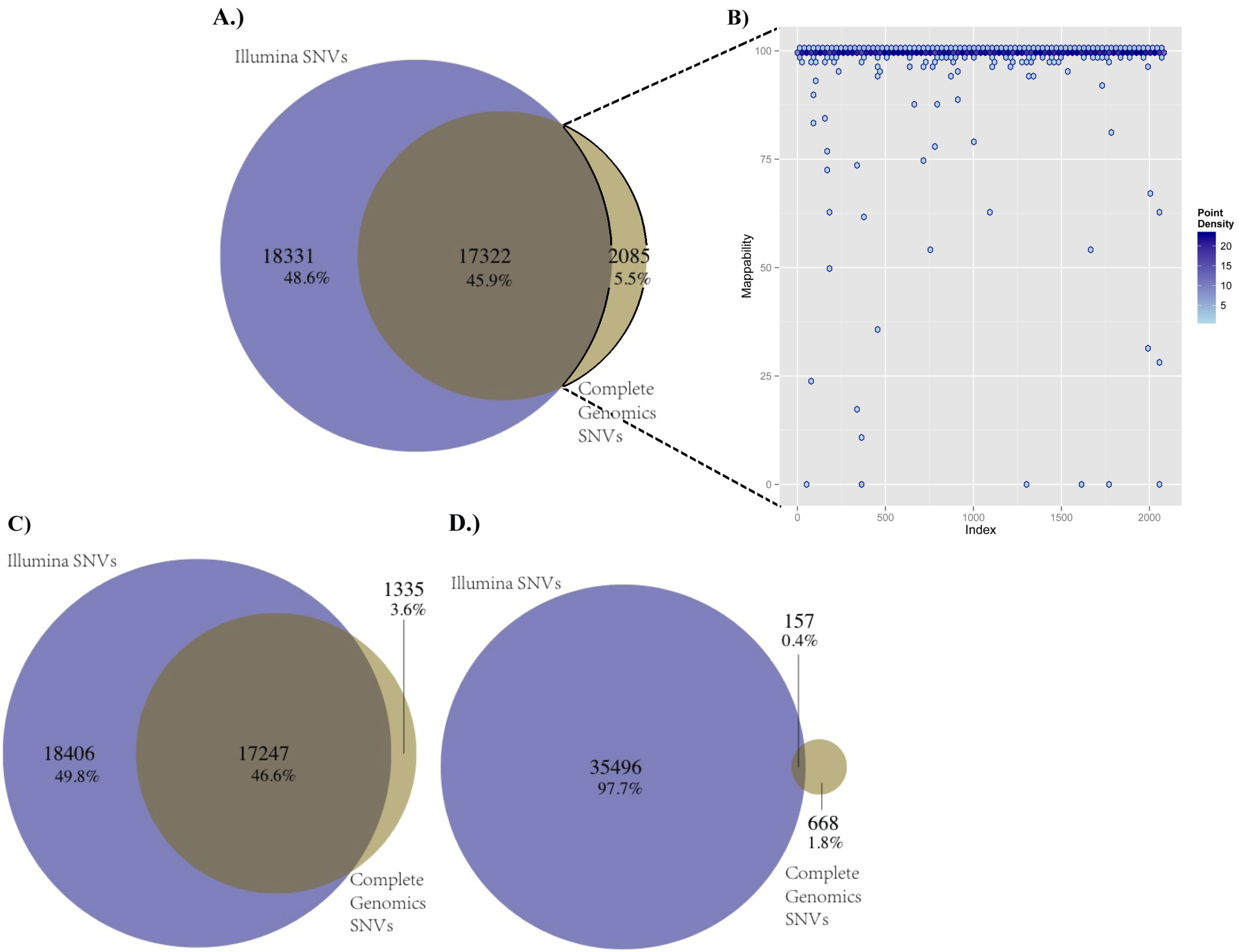
Union of all SNVs from 5 pipelines

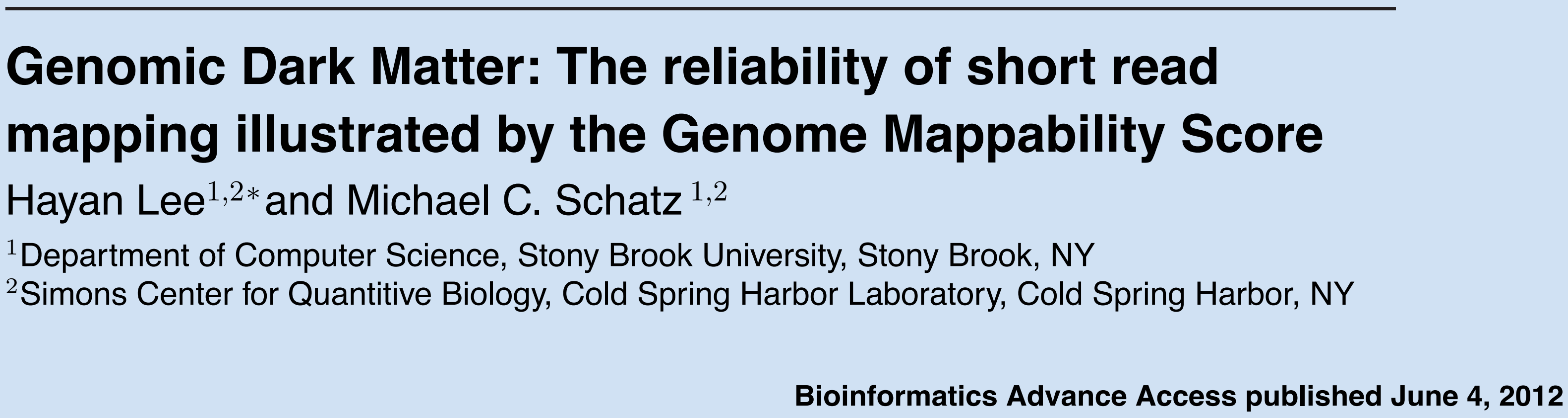


Union of all indels from 3 pipelines

What is the “True” Personal Genome?




A.)
Illumina SNVs
18331
48.6%
17322
45.9%
2085
5.5%
Complete
Genomics
SNVs
B)
Mappability
100
75
50
25
0
0
500
1000
1500
2000
Index
Point
Density
20
15
10
5
C)
Illumina SNVs
18406
49.8%
17247
46.6%
1335
3.6%
Complete
Genomics
SNVs
D.)
Illumina SNVs
35496
97.7%
157
0.4%
668
1.8%
Complete
Genomics
SNVs



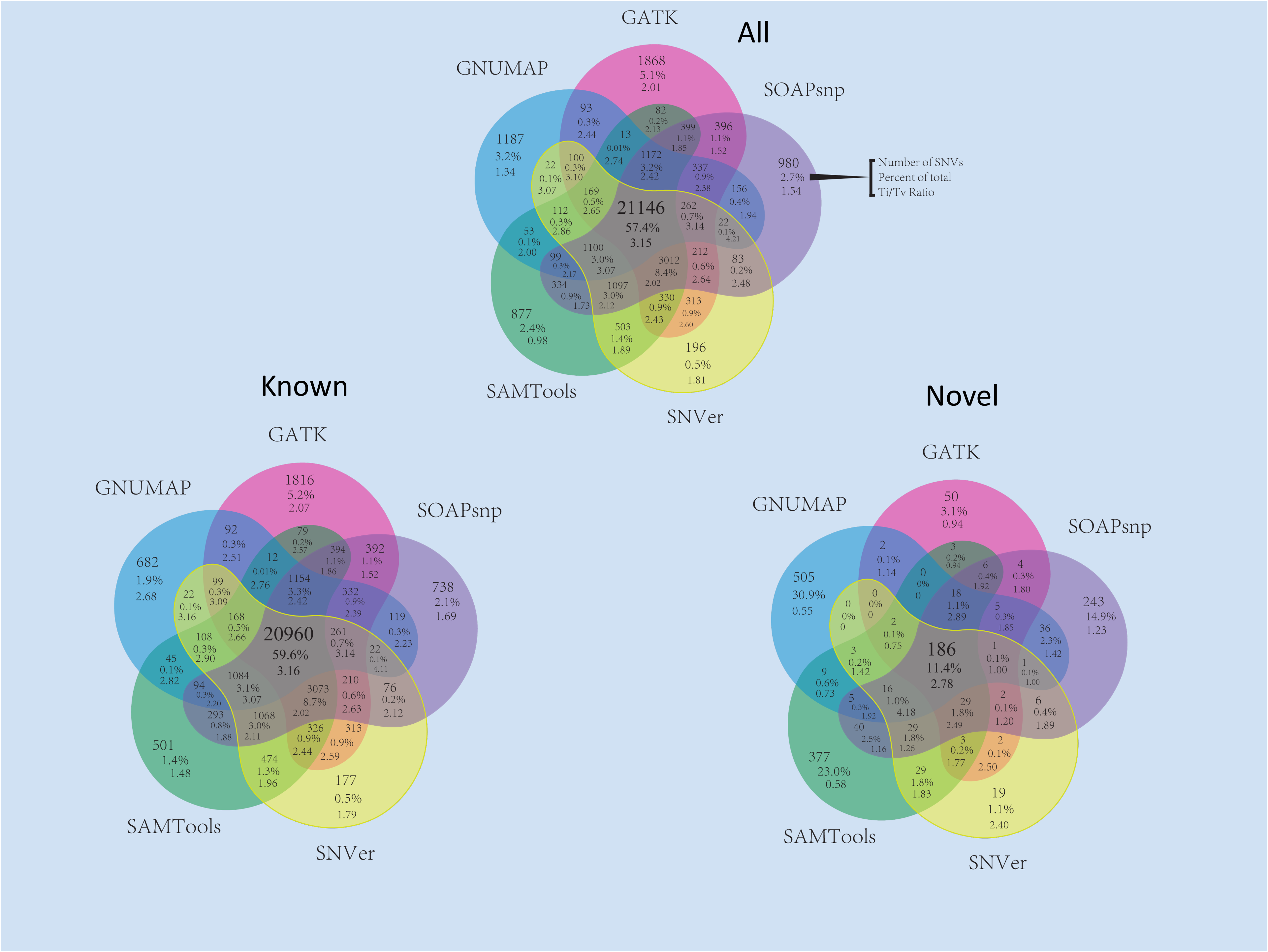
Genomic Dark Matter: The reliability of short read mapping illustrated by the Genome Mappability Score

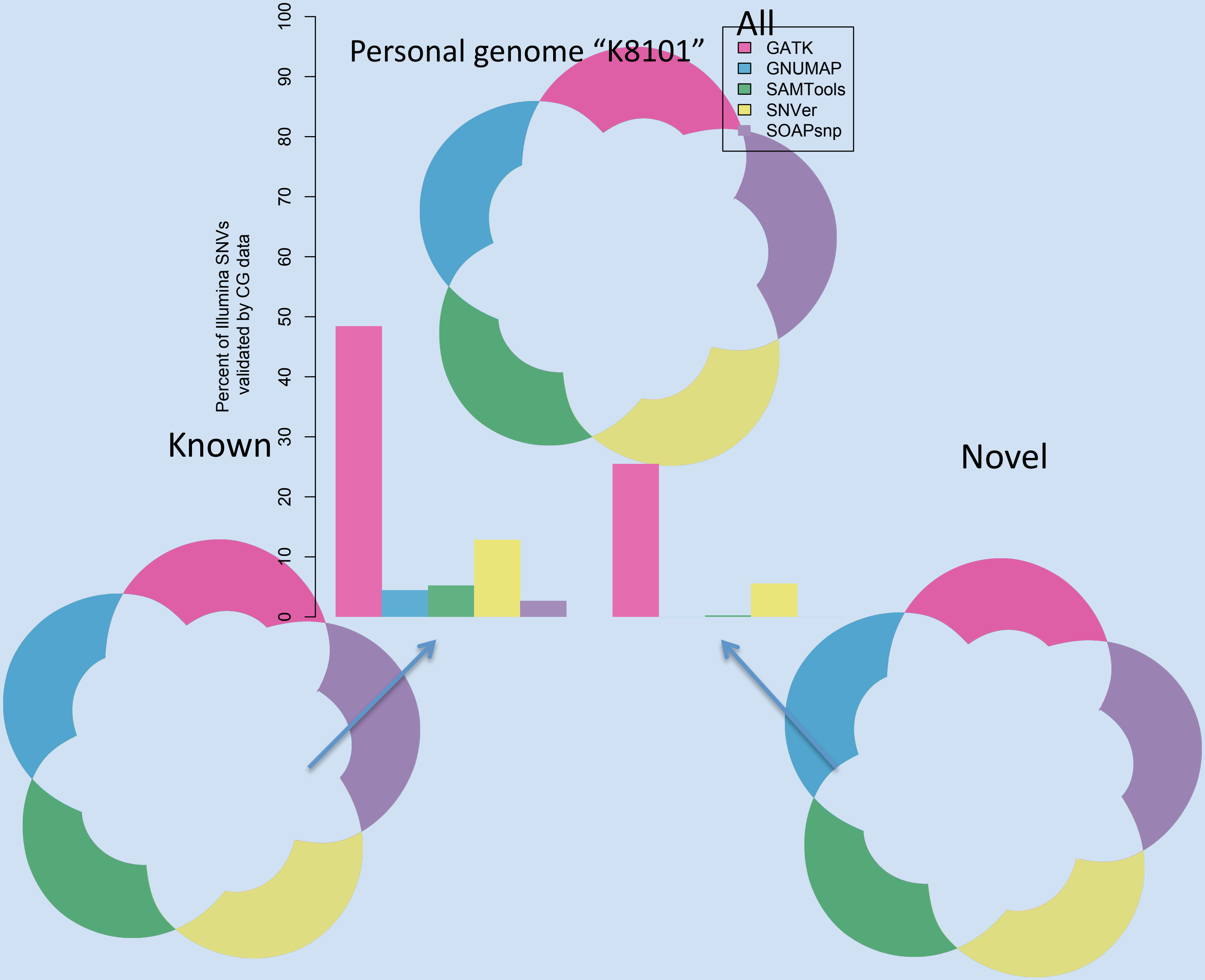
Hayan Lee[1,2*] and Michael C. Schatz [1,2]

[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.


All
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1868 5.1% 2.01
93 0.3% 2.44
82 0.2% 2.13
399 1.1% 1.85
396 1.1% 1.52
1187 3.2% 1.34
13 0.01% 2.74
1172 3.2% 2.42
337 0.9% 2.38
980 2.7% 1.54
Number of SNVs
Percent of total
Ti/Tv Ratio
22 0.1% 3.07
100 0.3% 3.10
169 0.5% 2.65
21146 57.4% 3.15
262 0.7% 3.14
156 0.4% 1.94
112 0.3% 2.86
22 0.1% 4.21
53 0.1% 2.00
99 0.3% 2.17
1100 3.0% 3.07
3012 8.4% 2.02
212 0.6% 2.64
83 0.2% 2.48
334 0.9% 1.73
1097 3.0% 2.12
330 0.9% 2.43
313 0.9% 2.60
877 2.4% 0.98
503 1.4% 1.89
196 0.5% 1.81
Known
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1816 5.2% 2.07
92 0.3% 2.51
79 0.2% 2.57
394 1.1% 1.86
392 1.1% 1.52
682 1.9% 2.68
12 0.01% 2.76
1154 3.3% 2.42
332 0.9% 2.39
738 2.1% 1.69
22 0.1% 3.16
99 0.3% 3.09
168 0.5% 2.66
20960 59.6% 3.16
261 0.7% 3.14
119 0.3% 2.23
108 0.3% 2.90
22 0.1% 4.11
45 0.1% 2.82
94 0.3% 2.20
1084 3.1% 3.07
3073 8.7% 2.02
210 0.6% 2.63
76 0.2% 2.12
293 0.8% 1.88
1068 3.0% 2.11
326 0.9% 2.44
313 0.9% 2.59
501 1.4% 1.48
474 1.3% 1.96
177 0.5% 1.79
Novel
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
50 3.1% 0.94
2 0.1% 1.14
3 0.2% 0.94
6 0.4% 1.92
4 0.3% 1.80
505 30.9% 0.55
0 0% 0
18 1.1% 2.89
5 0.3% 1.85
243 14.9% 1.23
0 0% 0
0 0% 0
2 0.1% 0.75
186 11.4% 2.78
1 0.1% 1.00
36 2.3% 1.42
3 0.2% 1.42
1 0.1% 1.00
9 0.6% 0.73
5 0.3% 1.92
16 1.0% 4.18
29 1.8% 2.49
2 0.1% 1.20
6 0.4% 1.89
40 2.5% 1.16
29 1.8% 1.26
3 0.2% 1.77
2 0.1% 2.50
377 23.0% 0.58
29 1.8% 1.83
19 1.1% 2.40

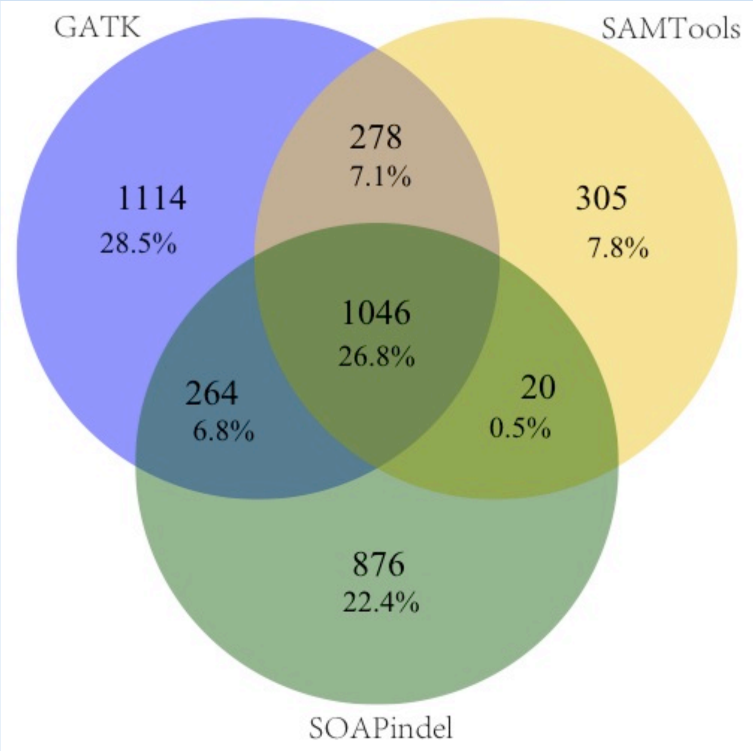


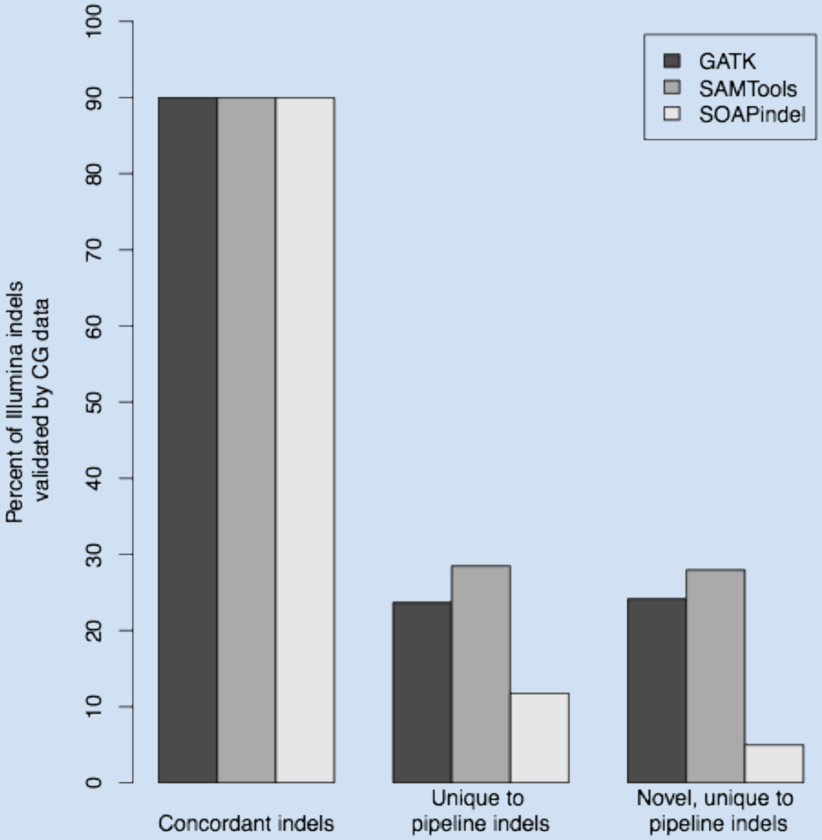
Personal genome “K8101”

Higher Validation by CG of SNVs with the BWA-GATK(v1.5) pipeline

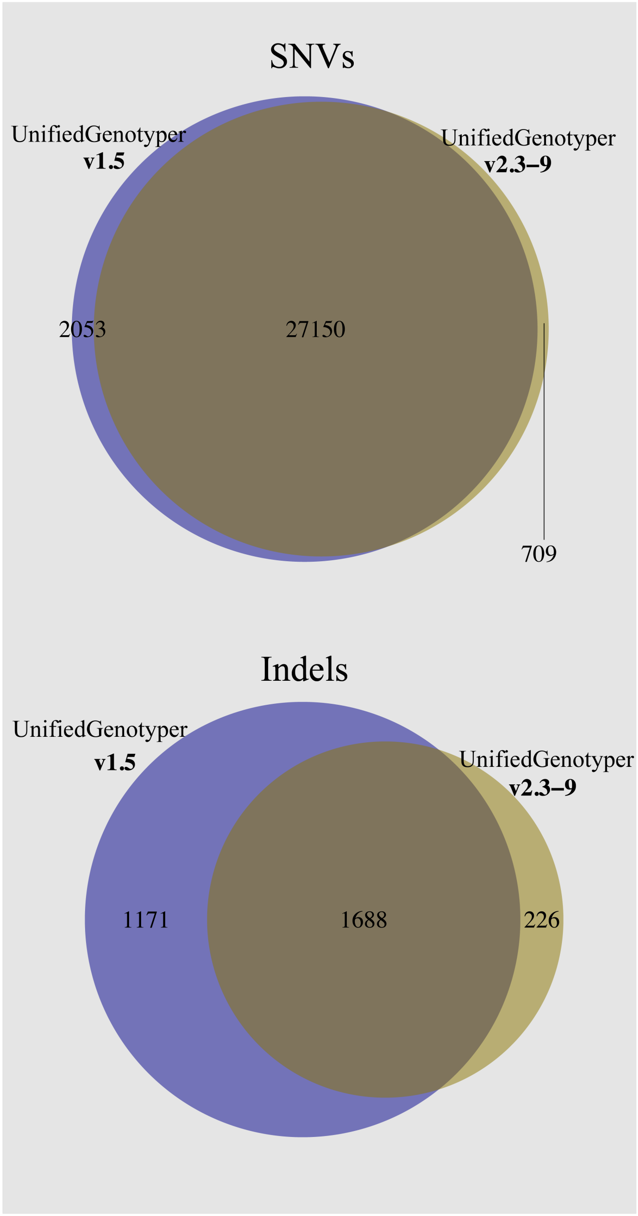
- Reveals higher validation rate of unique-to-pipeline variants, as well as uniquely discovered novel variants, for the variants called by BWA-GATK(v1.5), in comparison to the other 4 pipelines (including SOAP).

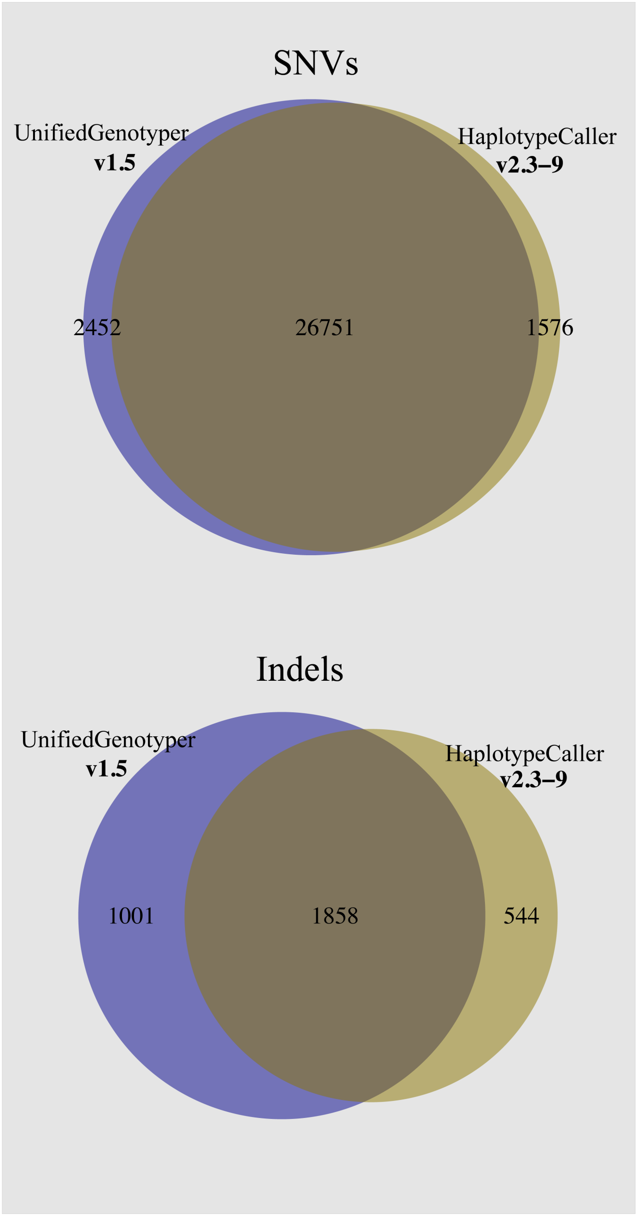
Validating Indels with Complete Genomics Data for the 3 pipelines



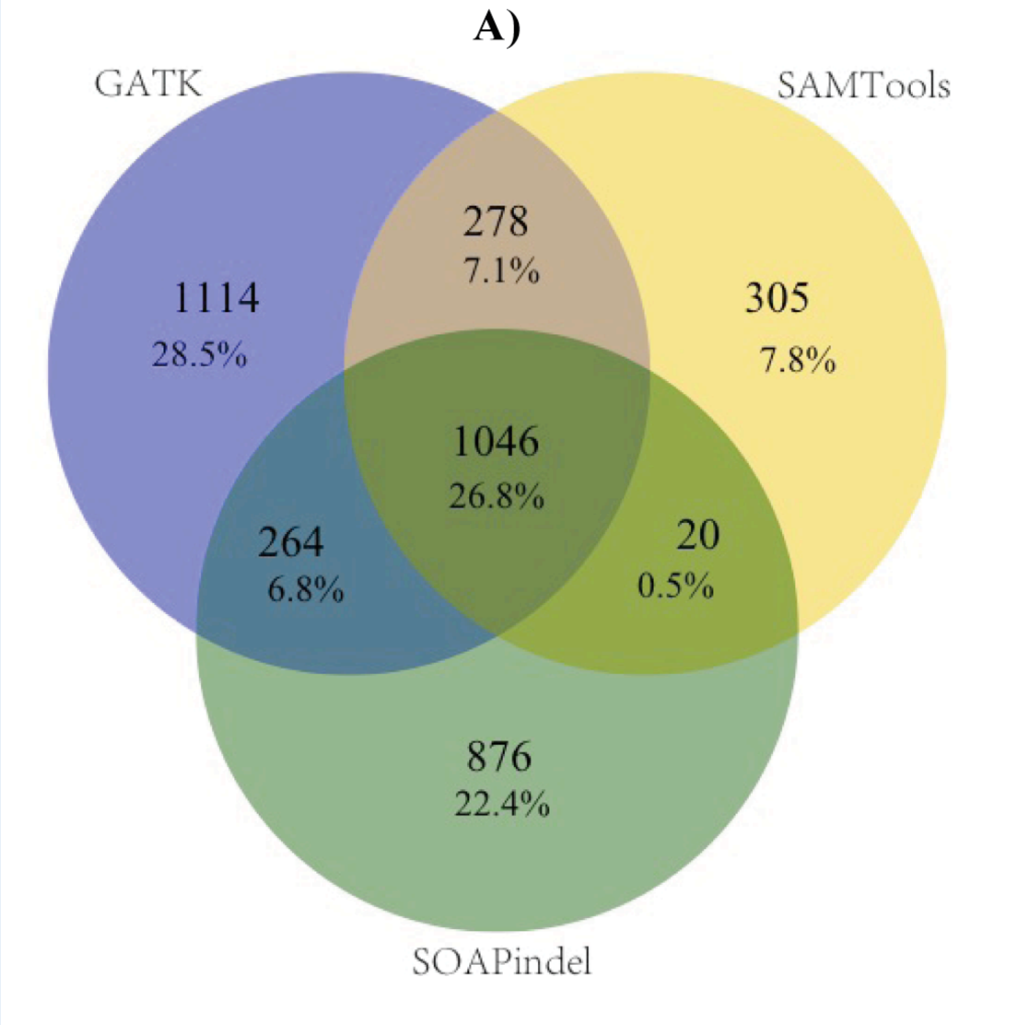


Comparing to New Versions of GATK





Validation of SNVs and Indels called by GATK, SOAP and both, with another platform

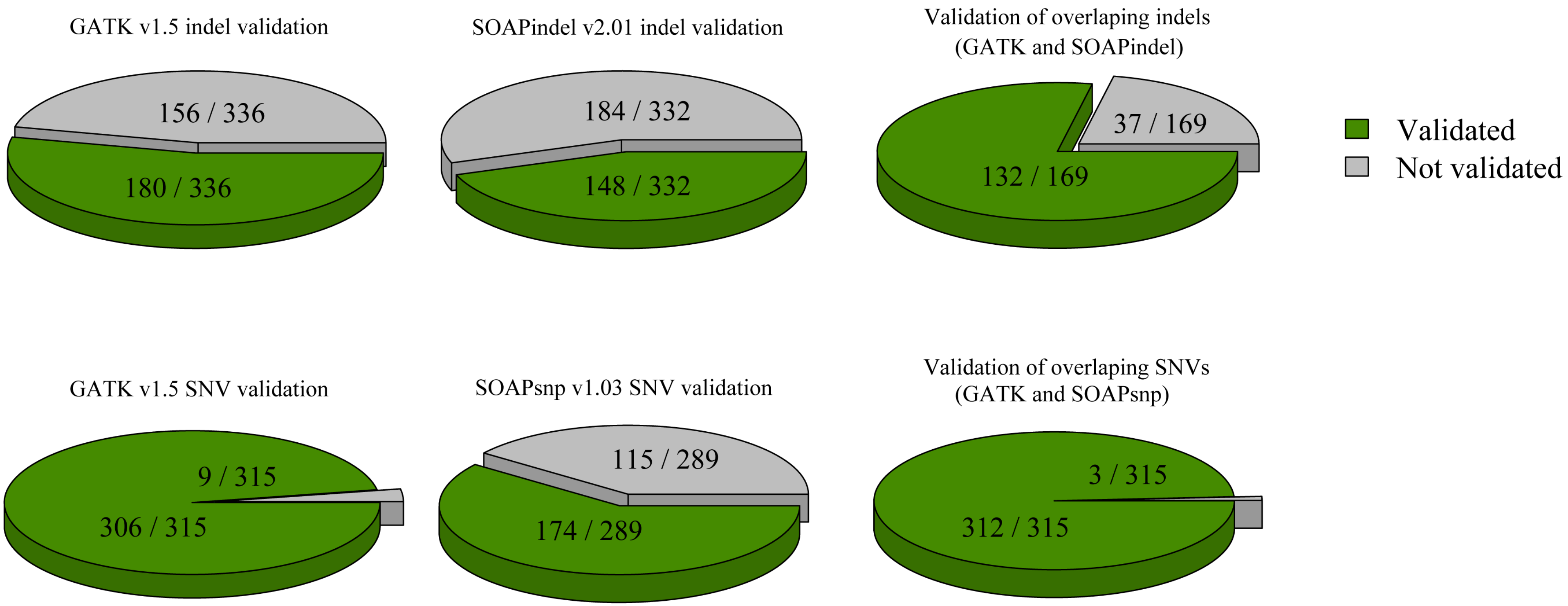


Indels

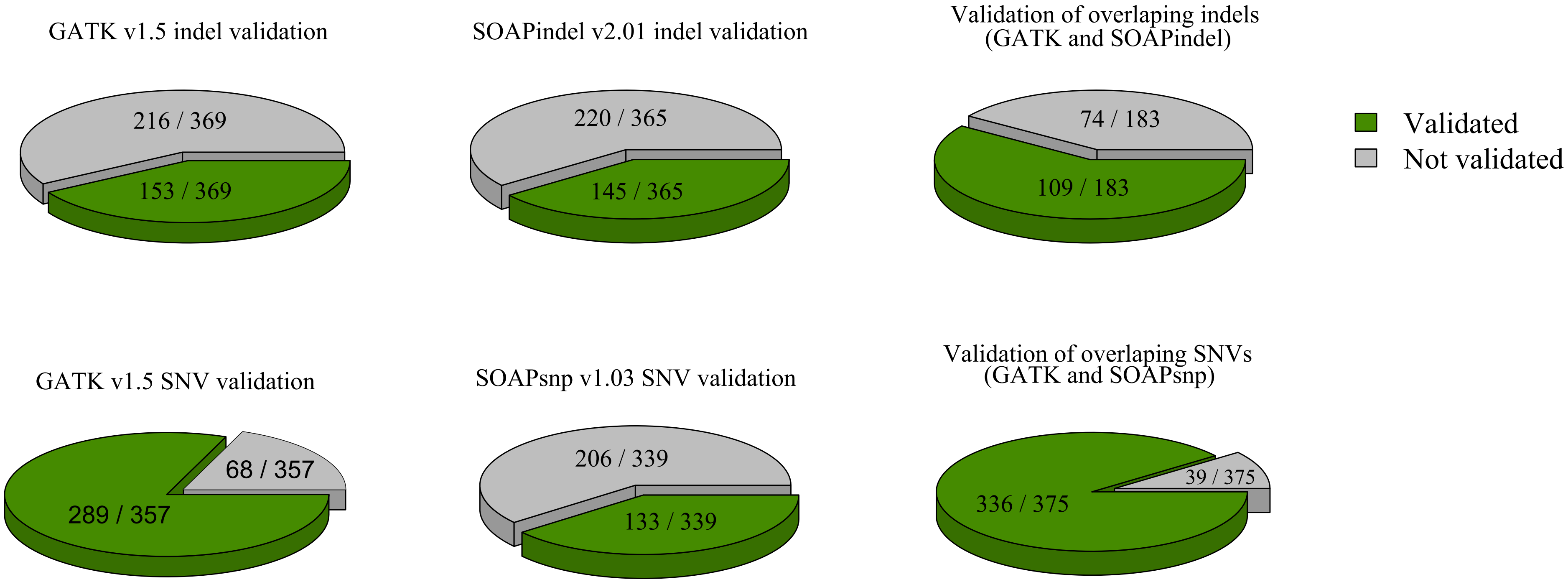
Validation with PCR amplicons and MiSeq 150 bp reads at ~5000x coverage

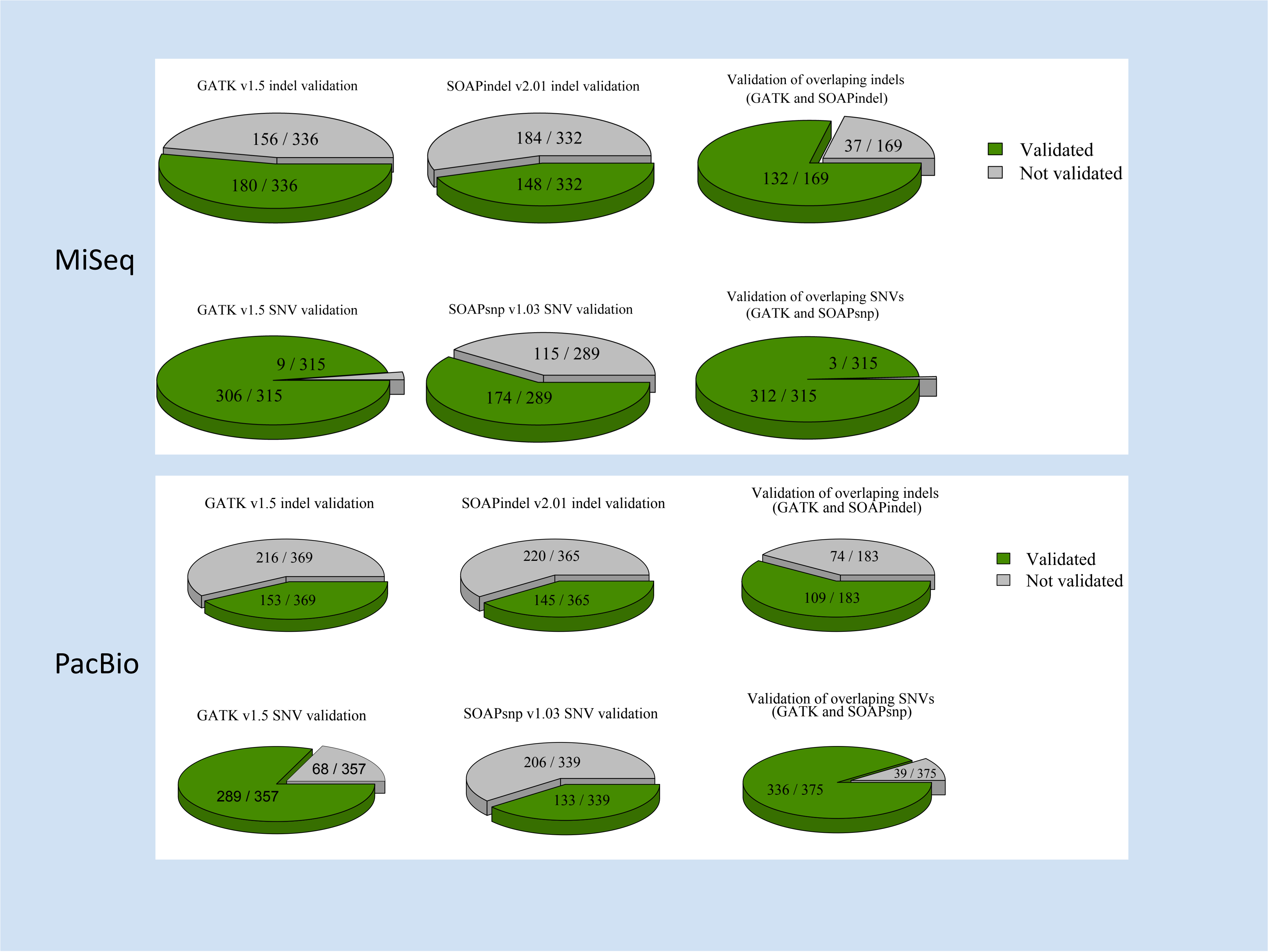
1,140 SNVs, with random sampling of 380 from the set of unique-to-GATK SNVs, 380 from the set of unique-to-SOAPsnp SNVs, and 380 from the set that were overlapping between these two pipelines.

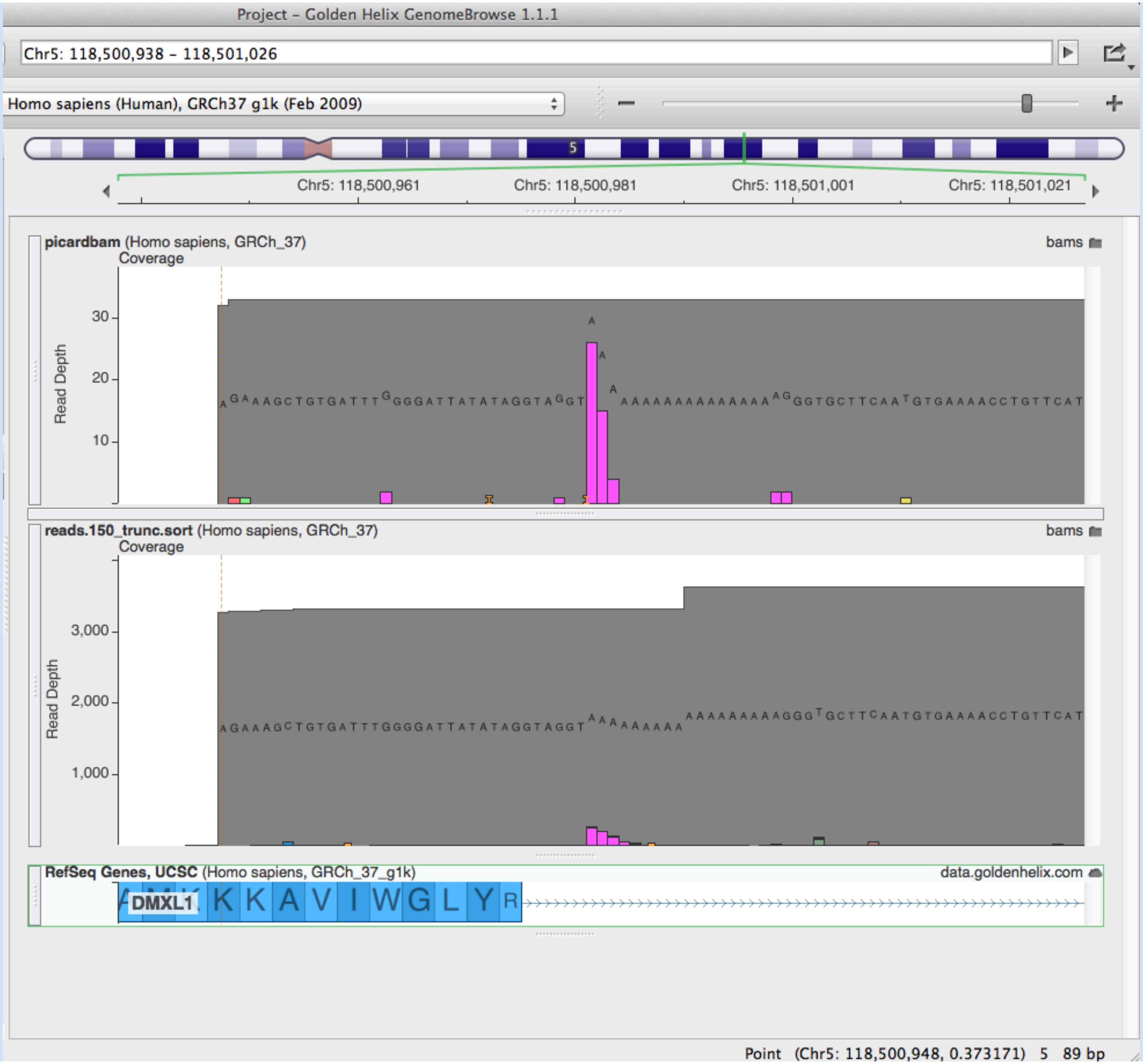
960 indels, with random sampling of 386 from the unique-to-GATK indel set, 387 from the unique-to-SOAPindel set, and 187 from the set of indels overlapping between the two (SOAPindel and GATK).



Validation of ~2000 PCR amplicons with PacBio reads from two SMRT cells (~50,000 useable reads per cell)




MiSeq
GATK v1.5 indel validation
156 / 336
180 / 336
SOAPindel v2.01 indel validation
184 / 332
148 / 332
Validation of overlaping indels
(GATK and SOAPindel)
37 / 169
132 / 169
🟩 Validated
□ Not validated
GATK v1.5 SNV validation
9 / 315
306 / 315
SOAPsnp v1.03 SNV validation
115 / 289
174 / 289
Validation of overlaping SNVs
(GATK and SOAPsnp)
3 / 315
312 / 315
PacBio
GATK v1.5 indel validation
216 / 369
153 / 369
SOAPindel v2.01 indel validation
220 / 365
145 / 365
Validation of overlaping indels
(GATK and SOAPindel)
74 / 183
109 / 183
🟩 Validated
□ Not validated
GATK v1.5 SNV validation
68 / 357
289 / 357
SOAPsnp v1.03 SNV validation
206 / 339
133 / 339
Validation of overlaping SNVs
(GATK and SOAPsnp)
39 / 375
336 / 375



Project – Golden Helix GenomeBrowse 1.1.1

Chr5: 118,500,938 – 118,501,026

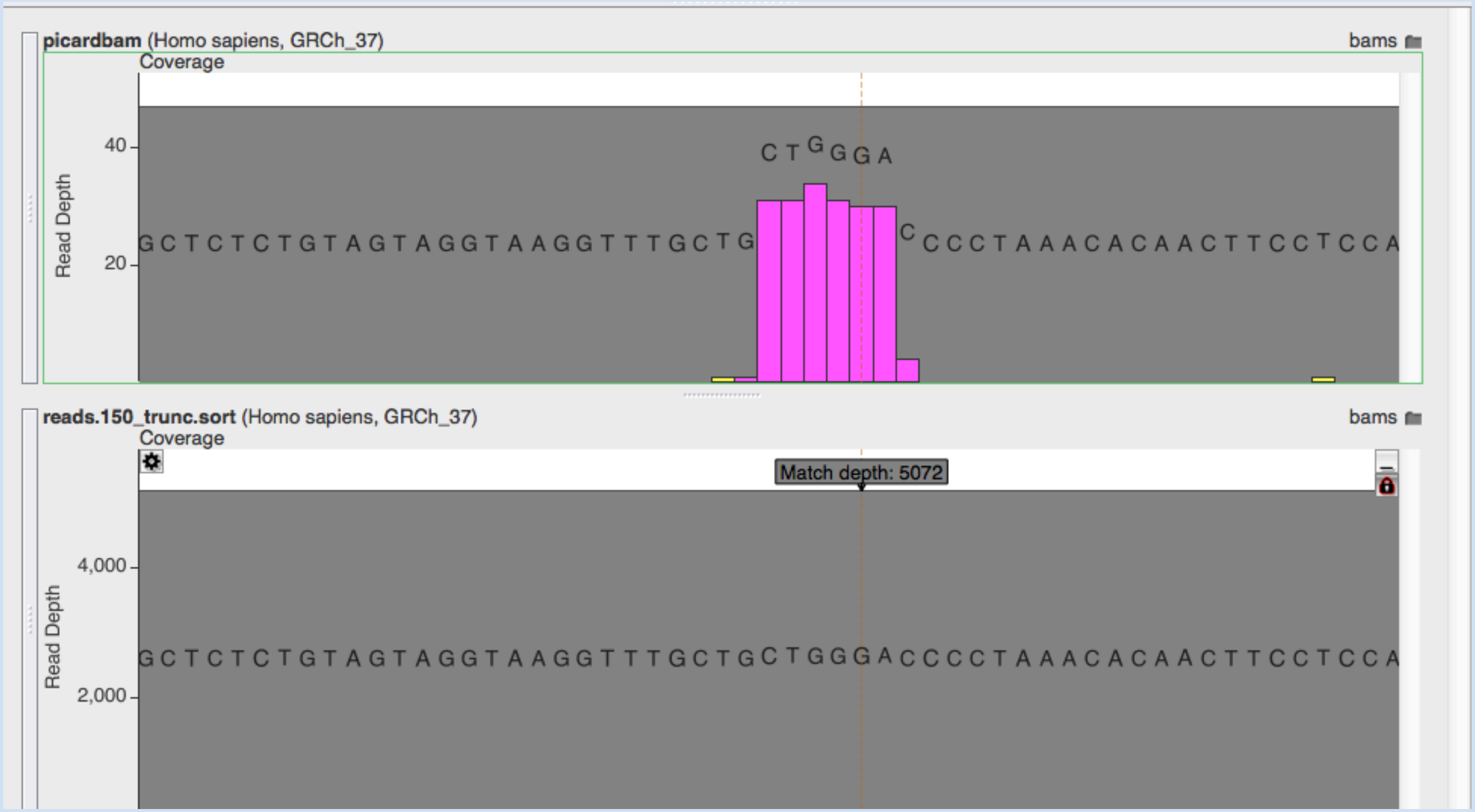
Homo sapiens (Human), GRCh37 g1k (Feb 2009)

picardbam (Homo sapiens, GRCh_37)

reads.150_trunc.sort (Homo sapiens, GRCh_37)

RefSeq Genes, UCSC (Homo sapiens, GRCh_37_g1k)

Point (Chr5: 118,500,948, 0.373171) 5 89 bp



picardbam (Homo sapiens, GRCh_37)

bams

Coverage

Read Depth
40
20
C T G G G A
G C T C T C T G T A G T A G G T A A G G T T T G C T G
C
C C C T A A A C A C A A C T T C C T C C A

reads.150_trunc.sort (Homo sapiens, GRCh_37)

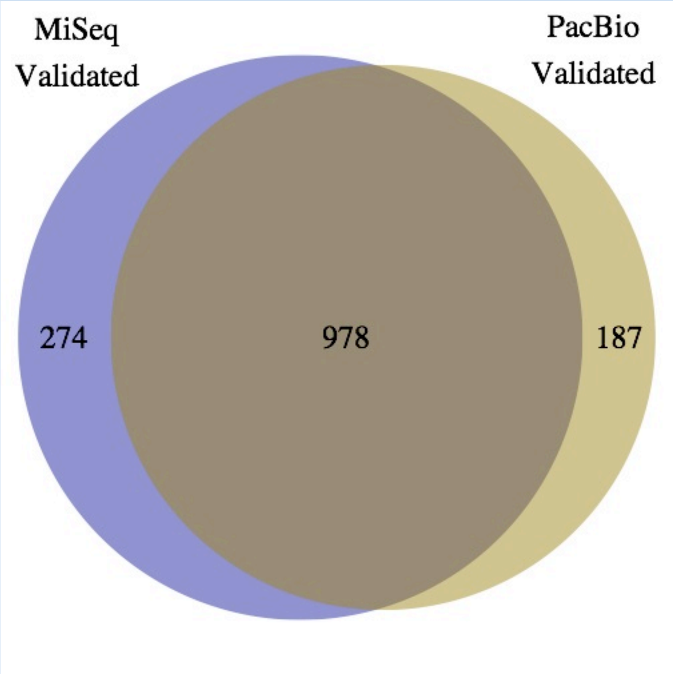
bams

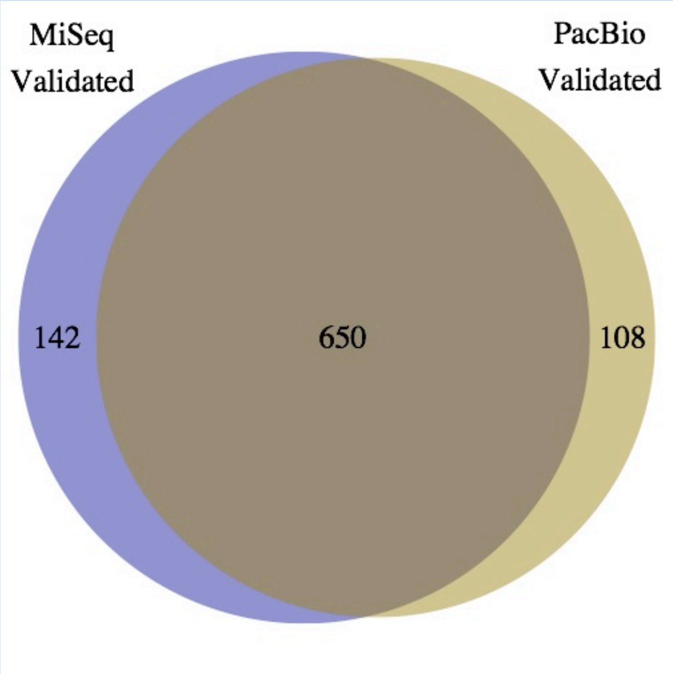
Coverage

Match depth: 5072

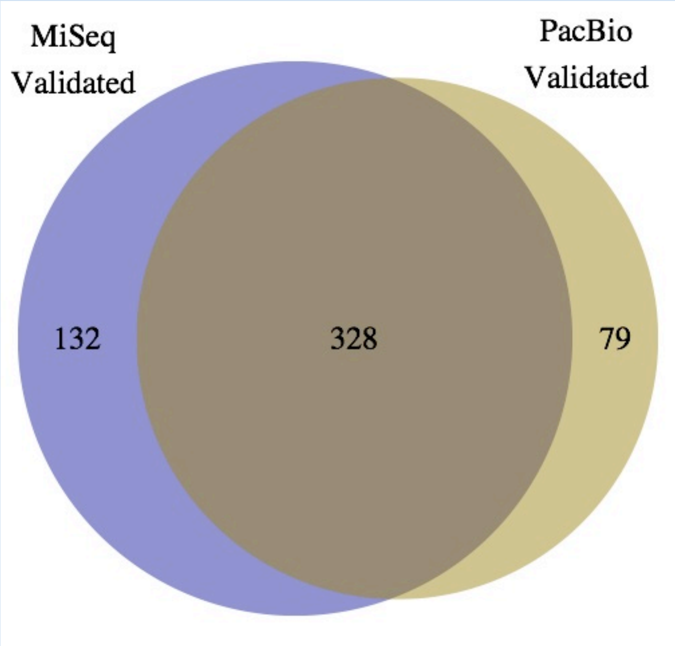
Read Depth
4,000
2,000
G C T C T C T G T A G T A G G T A A G G T T T G C T G C T G G G A C C C C T A A A C A C A A C T T C C T C C A

SNVs
+ Indels



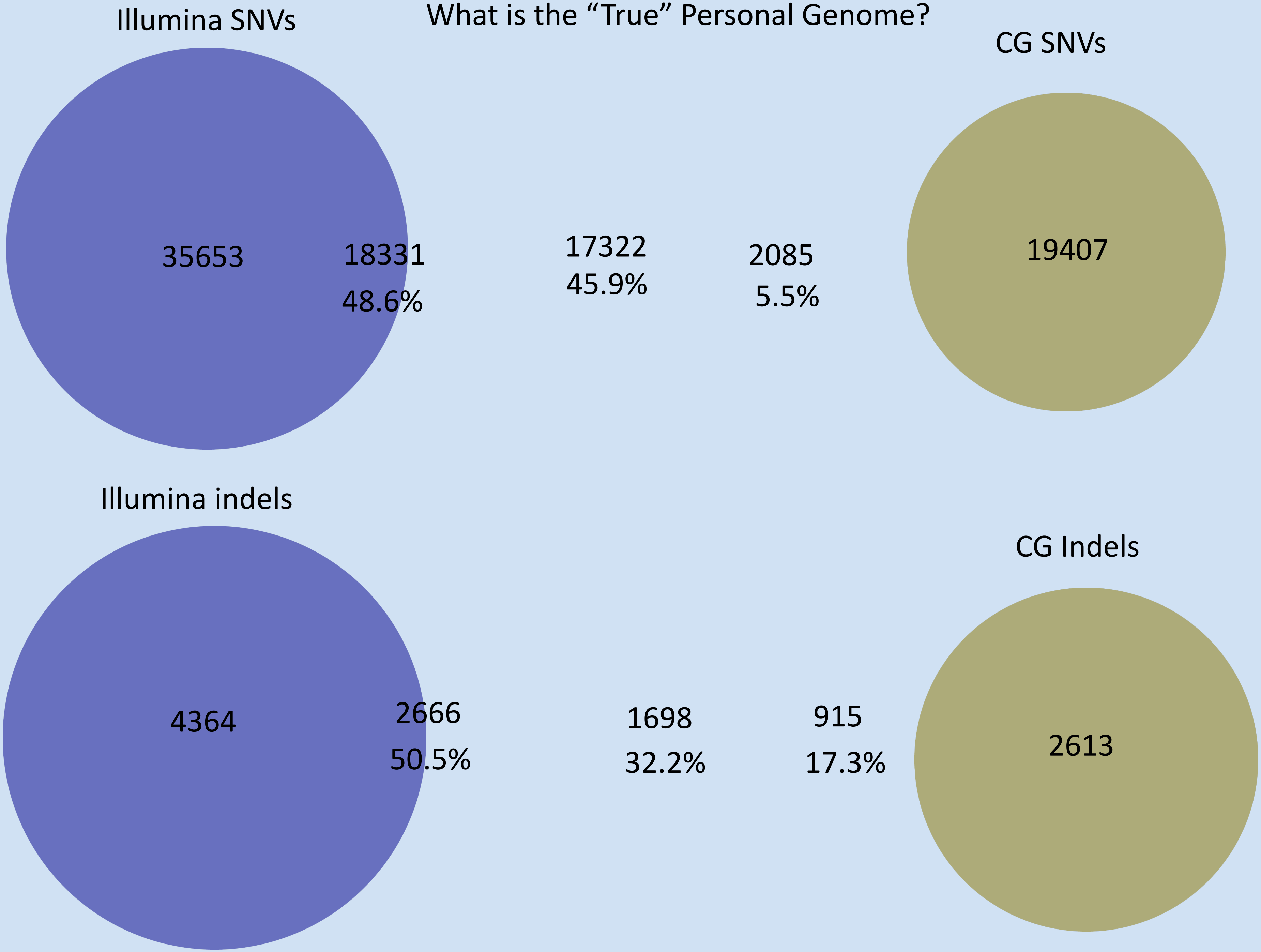


SNVs



Indels

What is the “True” Personal Genome?

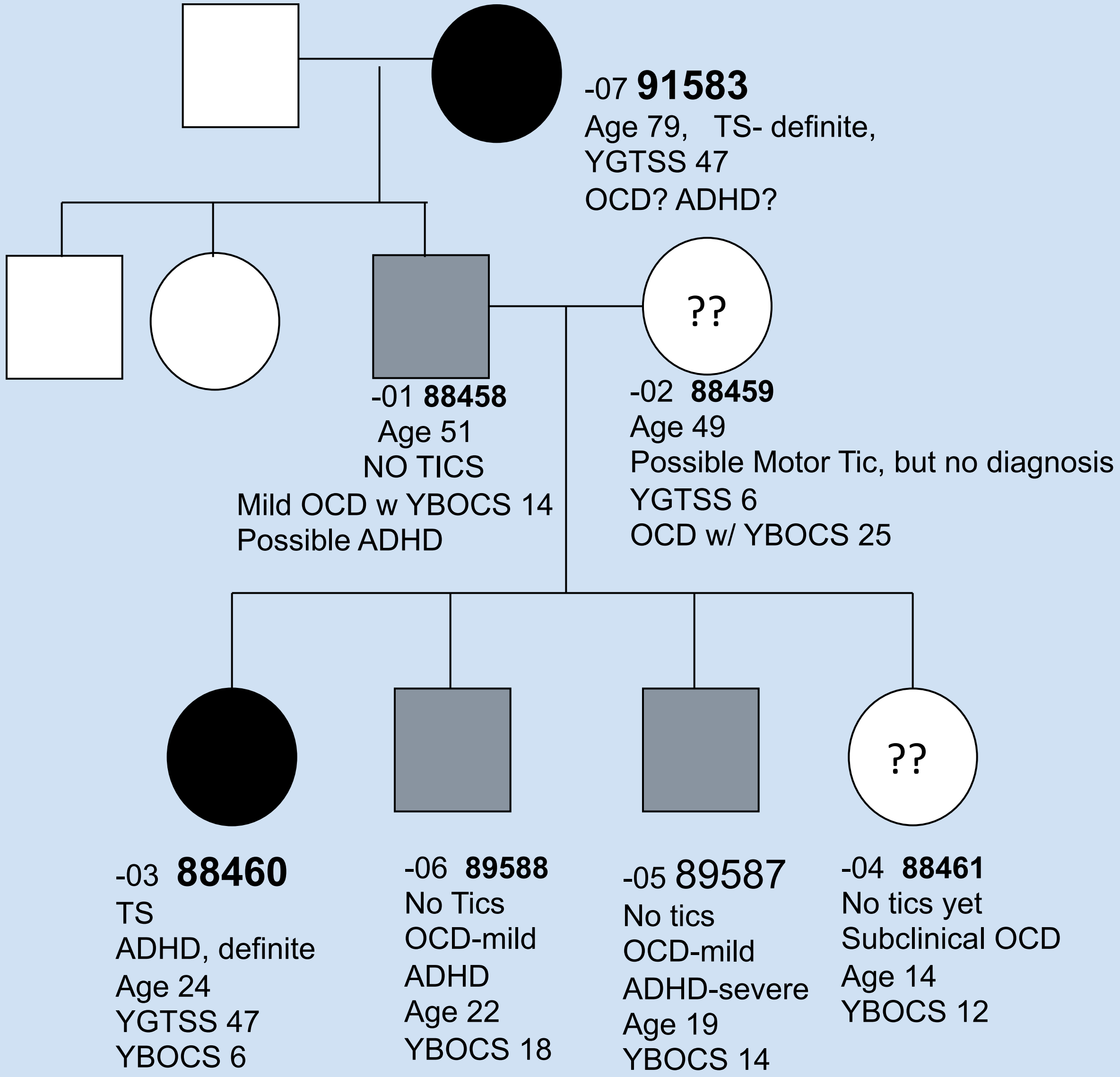


Optimizing the Variant Calling Pipeline Using Family Relationships

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.

TDT- 09 -1018
K26679



-07 **91583**
Age 79,   TS- definite,
YGTSS 47
OCD? ADHD?

-01 **88458**
Age 51
NO TICS
Mild OCD w YBOCS 14
Possible ADHD

-02 **88459**
Age 49
Possible Motor Tic, but no diagnosis
YGTSS 6
OCD w/ YBOCS 25

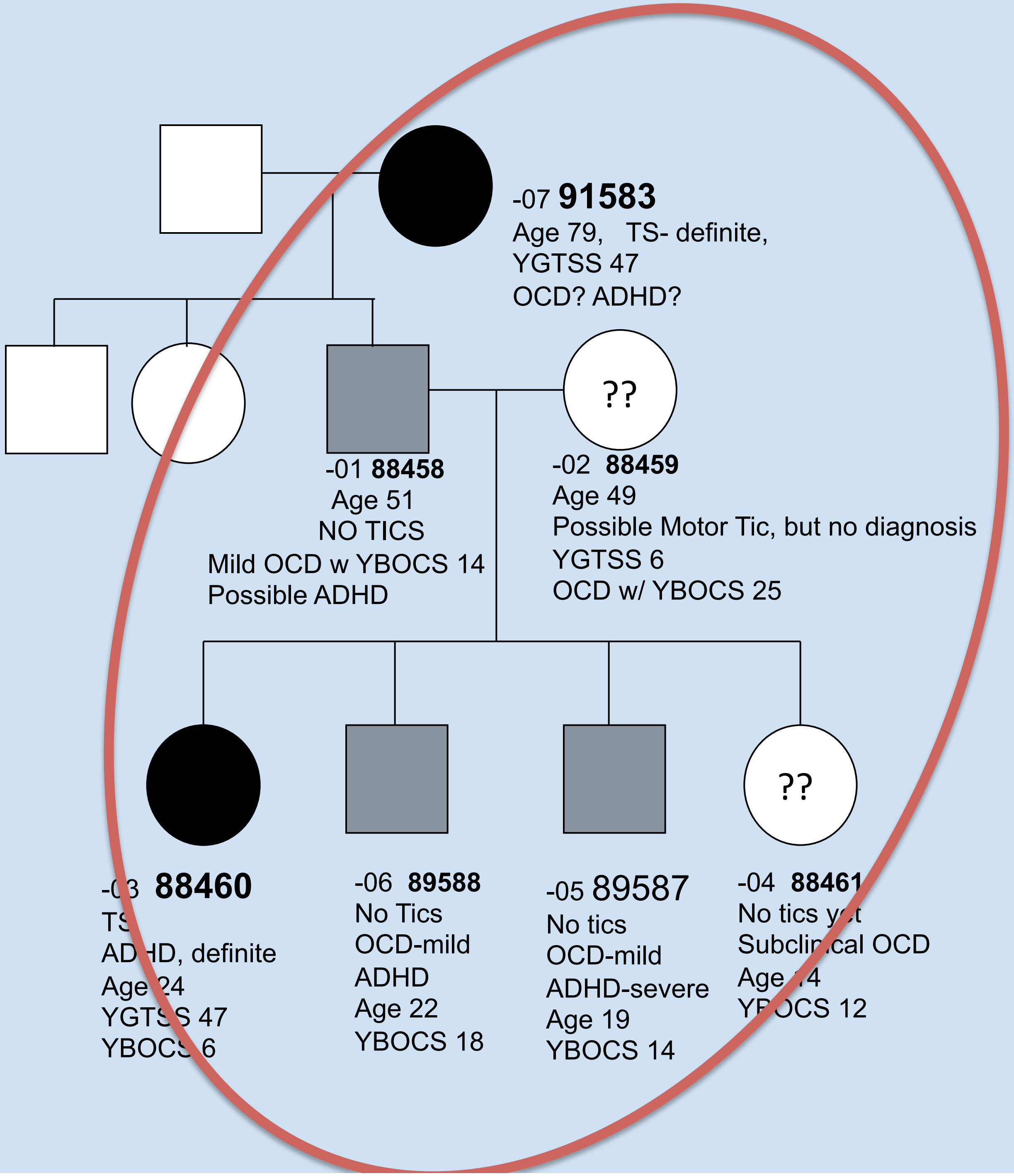
-03 **88460**
TS
ADHD, definite
Age 24
YGTSS 47
YBOCS 6

-06 **89588**
No Tics
OCD-mild
ADHD
Age 22
YBOCS 18

-05 89587
No tics
OCD-mild
ADHD-severe
Age 19
YBOCS 14

-04 **88461**
No tics yet
Subclinical OCD
Age 14
YBOCS 12

TDT- 09 -1018
K26679



-07 **91583**
Age 79, TS- definite,
YGTSS 47
OCD? ADHD?

-01 **88458**
Age 51
NO TICS
Mild OCD w YBOCS 14
Possible ADHD

??
-02 **88459**
Age 49
Possible Motor Tic, but no diagnosis
YGTSS 6
OCD w/ YBOCS 25

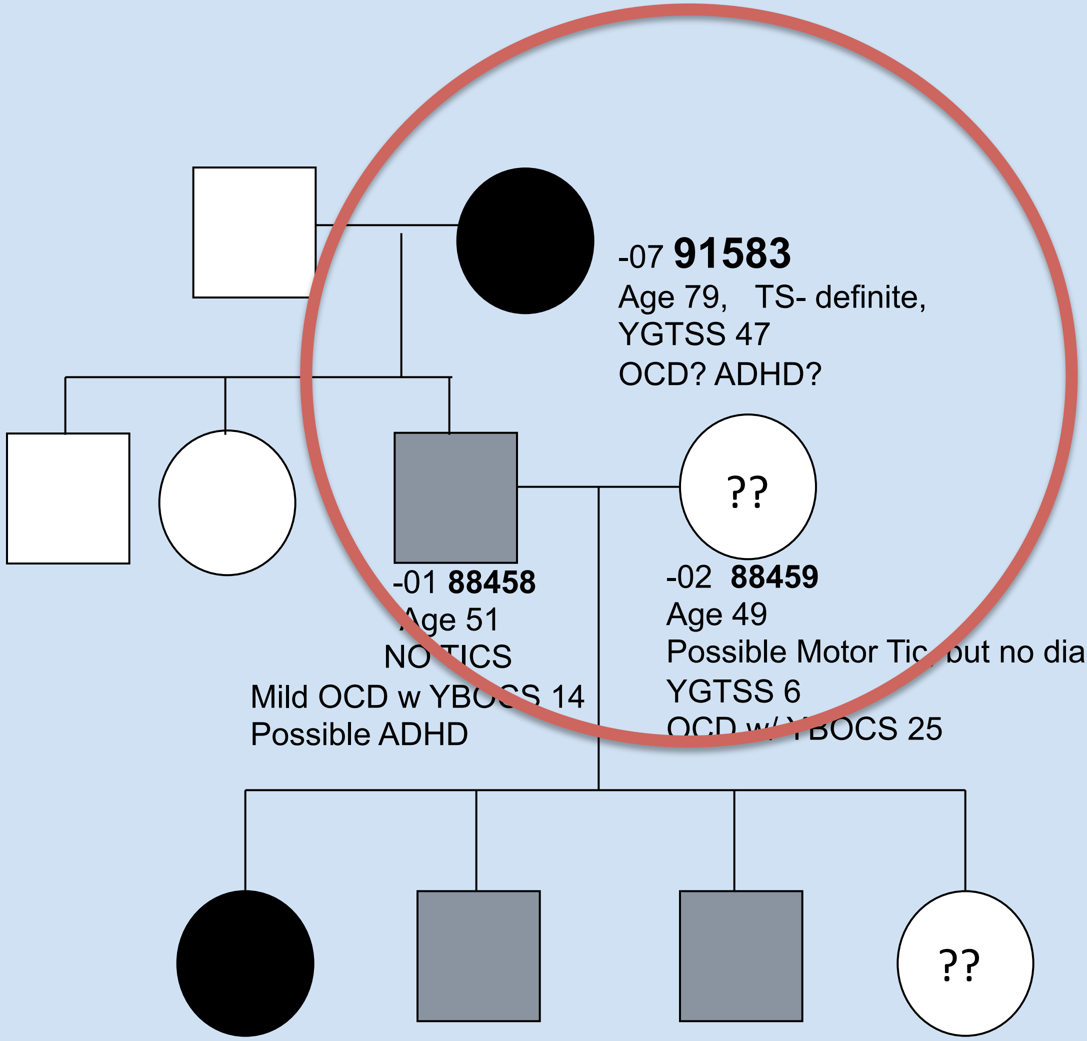
-03 **88460**
TS
ADHD, definite
Age 24
YGTSS 47
YBOCS 6

-06 **89588**
No Tics
OCD-mild
ADHD
Age 22
YBOCS 18

-05 89587
No tics
OCD-mild
ADHD-severe
Age 19
YBOCS 14

??
-04 **88461**
No tics yet
Subclinical OCD
Age 14
YBOCS 12

TDT- 09 -1018
K26679



-07 **91583**
Age 79, TS- definite,
YGTSS 47
OCD? ADHD?

“Parents”

-01 **88458**
Age 51
NO TICS
Mild OCD w YBOCS 14
Possible ADHD

-02 **88459**
Age 49
Possible Motor Tics but no diagnosis
YGTSS 6
OCD w/ YBOCS 25

-03 **88460**
TS
ADHD, definite
Age 24
YGTSS 47
YBOCS 6

-06 **89588**
No Tics
OCD-mild
ADHD
Age 22
YBOCS 18

-05 89587
No tics
OCD-mild
ADHD-severe
Age 19
YBOCS 14

-04 **88461**
No tics yet
Subclinical OCD
Age 14
YBOCS 12

Analysis based on various pipelines

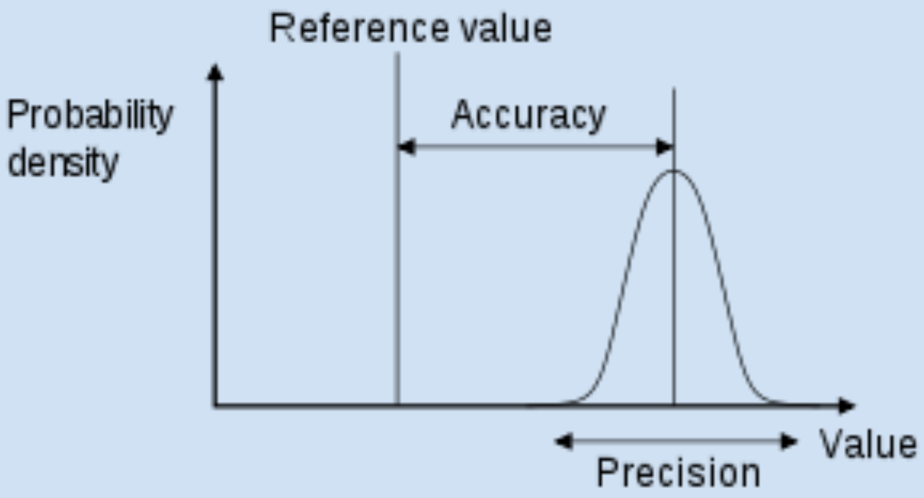
- “Parents” in this case means the mother, father AND grandmother.
- Taking the **Union** of SNVs from all 5 pipelines from “Parents”, and subtract that from the **Union** of all SNVs in each child.
- Or Subtract the **Union** of these “Parents” from the SNVs in the child **concordant** between 5 pipelines.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

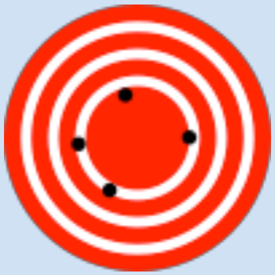
Table 3. *De novo* single-nucleotide variants (SNVs) were detected in two families contained within the 15 study exomes.

Family 1	Number of putative *de novo* coding non-synonymous or nonsense SNVs detected	
	Without using the grandparents as a filter	Using the grandparents as a filter
Child A	241	1
Child B	211	0
Child C	102	6
Child D	242	3
Family 2		
Child A	49	NA[a]
Child B	41	NA[a]

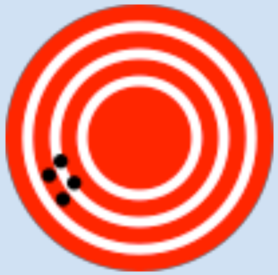
[a]N/A, no grandparent available.

Family 1 had a grandparent available for filtering purposes, whereas family 2 did not. To minimize false positives in the pool of SNVs associated with each child, only highly concordant SNVs were used (SNVs detected by all five pipelines). To construct a comprehensive set of SNVs for each parent, and hence increase filtering accuracy, false negatives for parent SNVs were reduced by taking the union of all SNV calls from all five pipelines.





High accuracy, but low precision



High precision, but low accuracy

In the fields of science, engineering, industry, and statistics, the **accuracy** of a measurement system is the degree of closeness of measurements of a quantity to that quantity's actual (true) value. The **precision** of a measurement system, also called reproducibility or repeatability, is the degree to which repeated measurements under unchanged conditions show the same results.

http://en.wikipedia.org/wiki/Accuracy_and_precision

Conclusions

- Sequencing a grandparent seems to help eliminate errors derived from the current depth of sequencing coverage in the mother and father.
- For now, we advocate using more than one pipeline on one set of sequencing data, but we expect the field to move toward >2 sequencing platforms per sample.
- Still need substantial work on indel-calling and validation.

Acknowledgements

CSH Cold Spring Harbor Laboratory
STANLEY INSTITUTE FOR COGNITIVE GENOMICS

Jason O'Rawe

Yiyang Wu

USC TROJANS

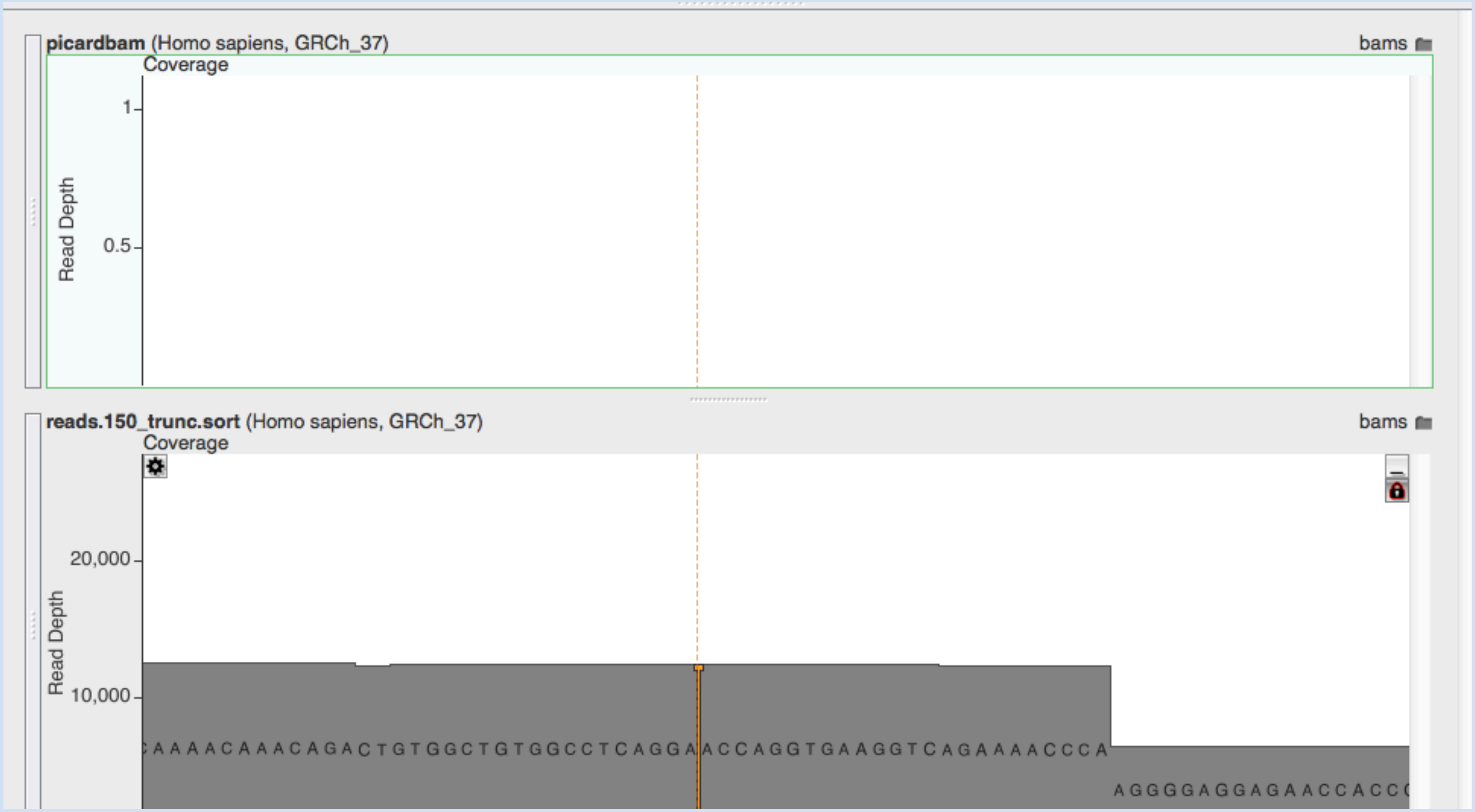
Kai Wang

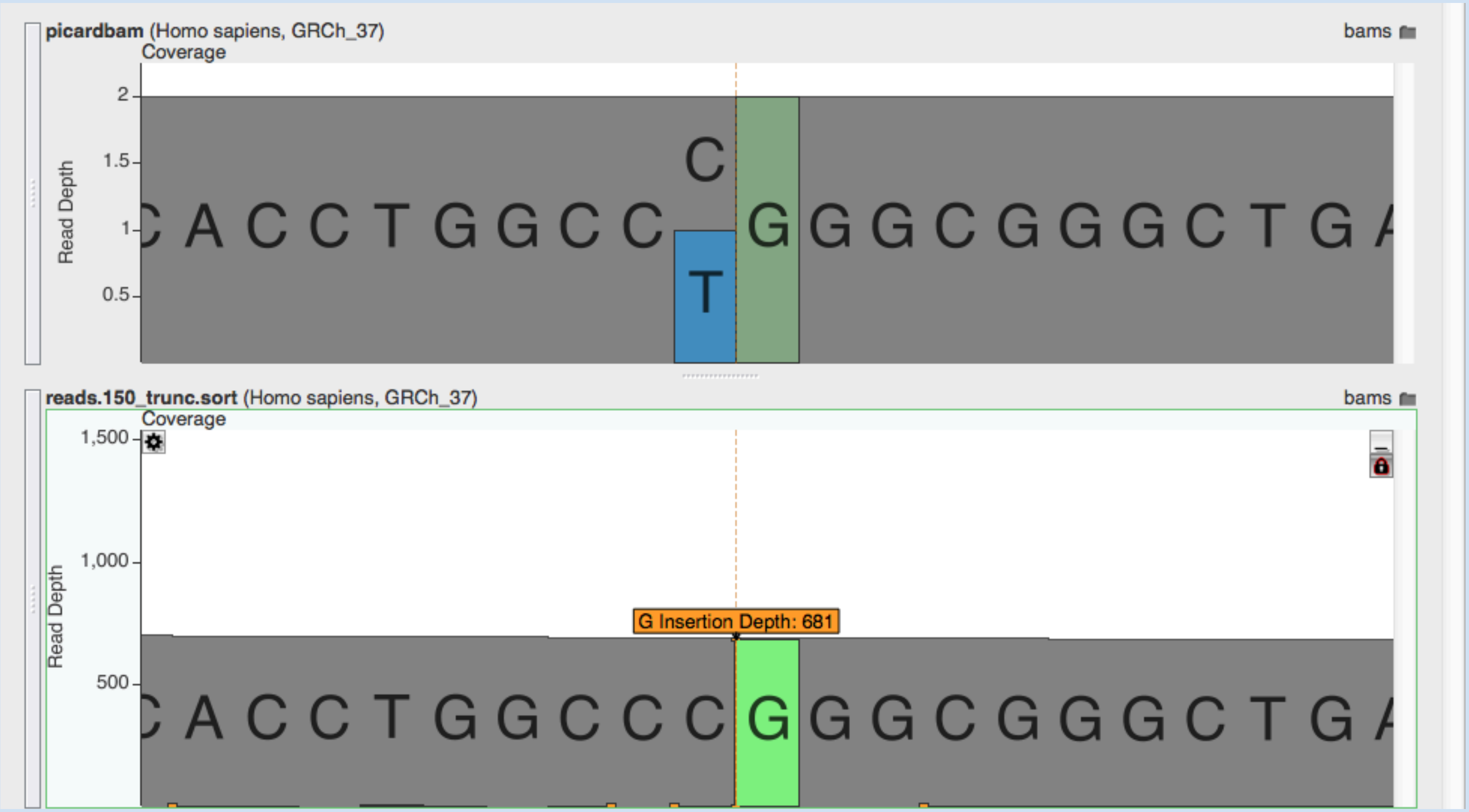
Michael Schatz
Giuseppe Narzisi
Eric Antoniou
Dick McCombie
Sequencing core facility

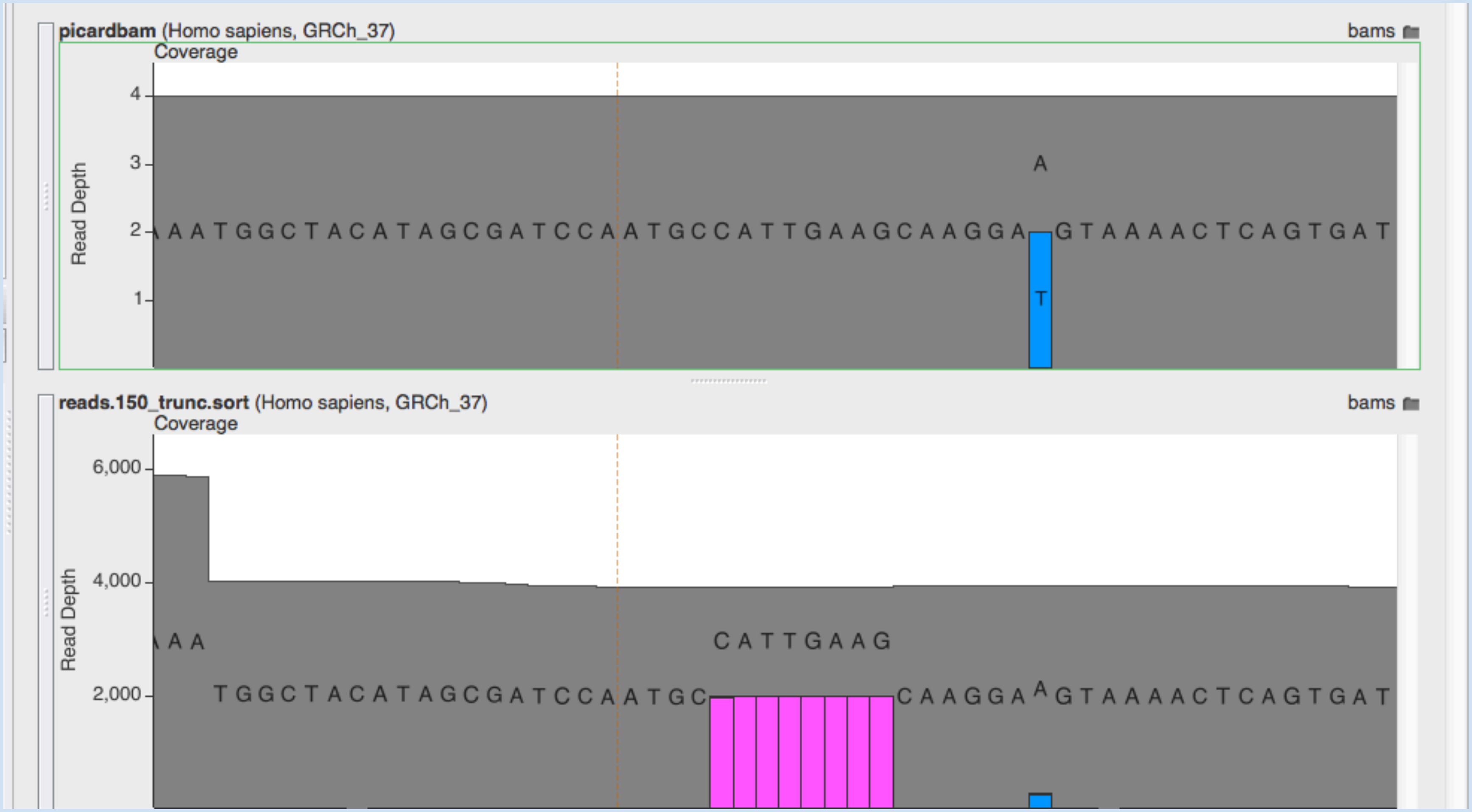
华大基因
BGI

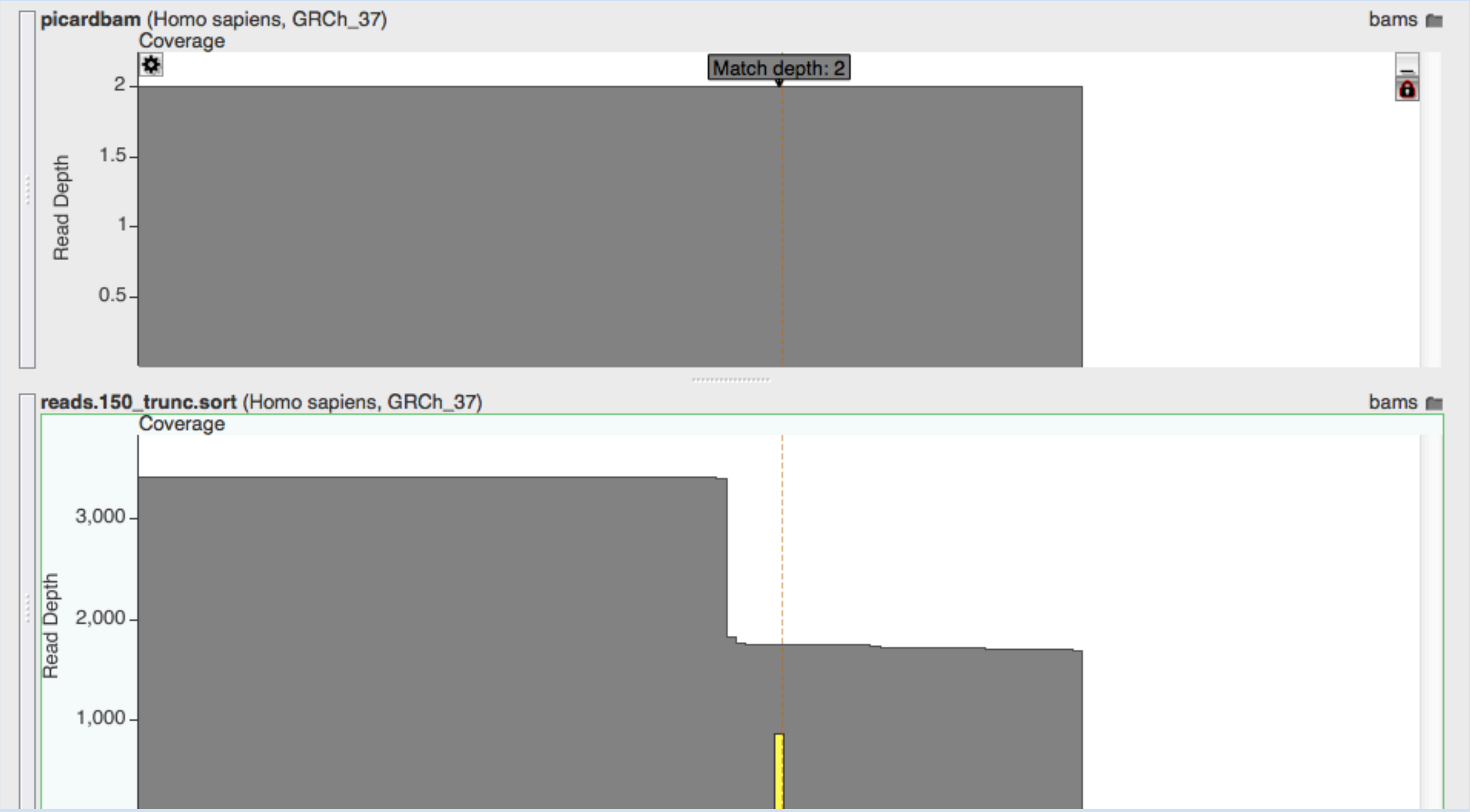
Tao Jiang
Guangqing Sun

EXTRA SLIDES – Not Shown


picardbam (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
1
0.5
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
20,000
10,000
CAAAACAAACAGACTGTGGCTGTGGCCTCAGGAACCAGGTGAAGGTCAGAAAACCCA
AGGGGAGGAGAACCACC


picardbam (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
2
1.5
1
0.5
C
C A C C T G G C C
T
G G G C G G G C T G A
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
1,500
1,000
500
G Insertion Depth: 681
C A C C T G G C C C G G G C G G G C T G A


picardbam (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
4
3
2
1
A
AAATGGCTACATAGCGATCCAATGCCATTGAAGCAAGGA GTAAAACTCAGTGAT
T
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
6,000
4,000
2,000
AAA
CATTGAAG
TGGCTACATAGCGATCCAATGC CAAGGA A GTAAAACTCAGTGAT

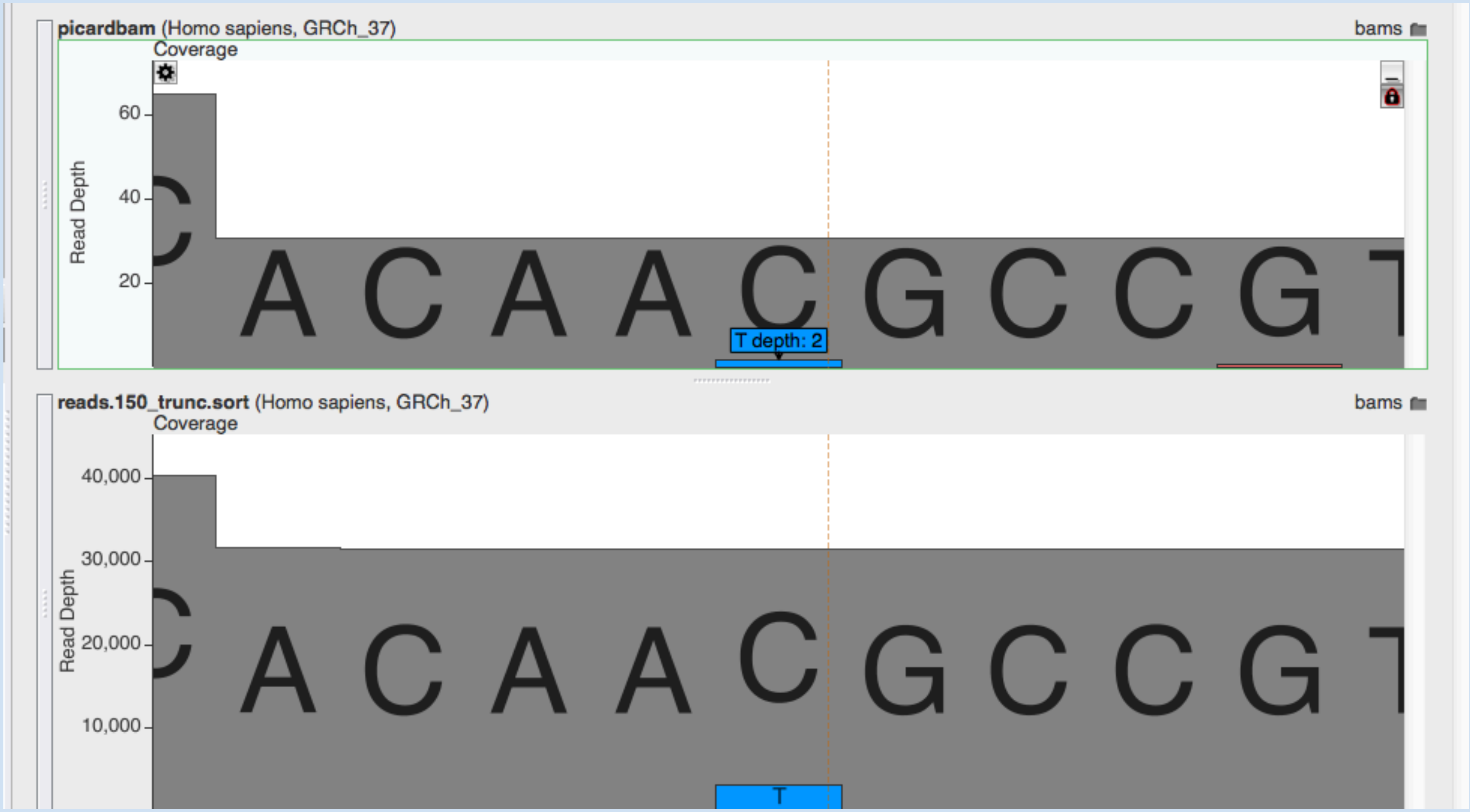


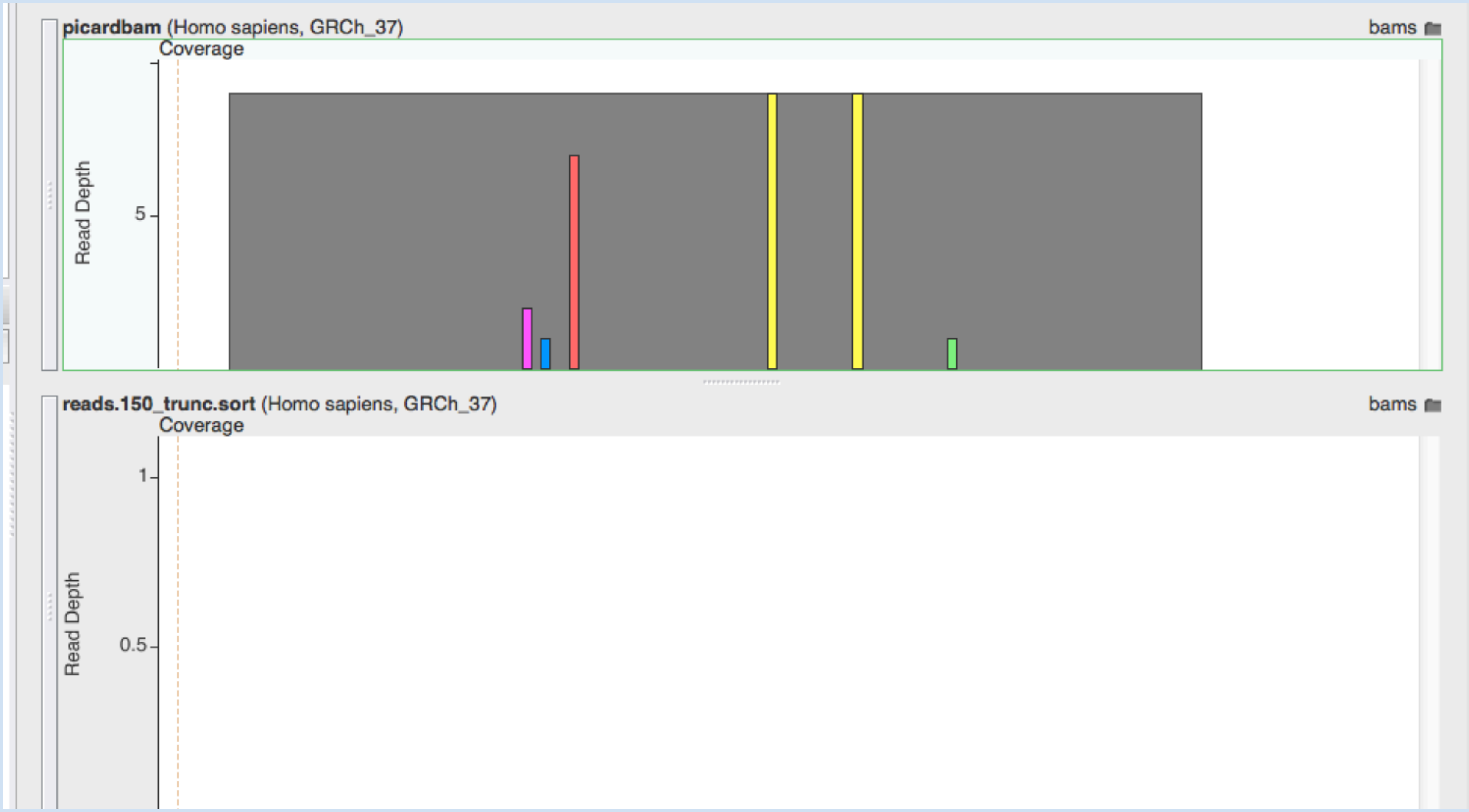
picardbam (Homo sapiens, GRCh_37)

Coverage

reads.150_trunc.sort (Homo sapiens, GRCh_37)

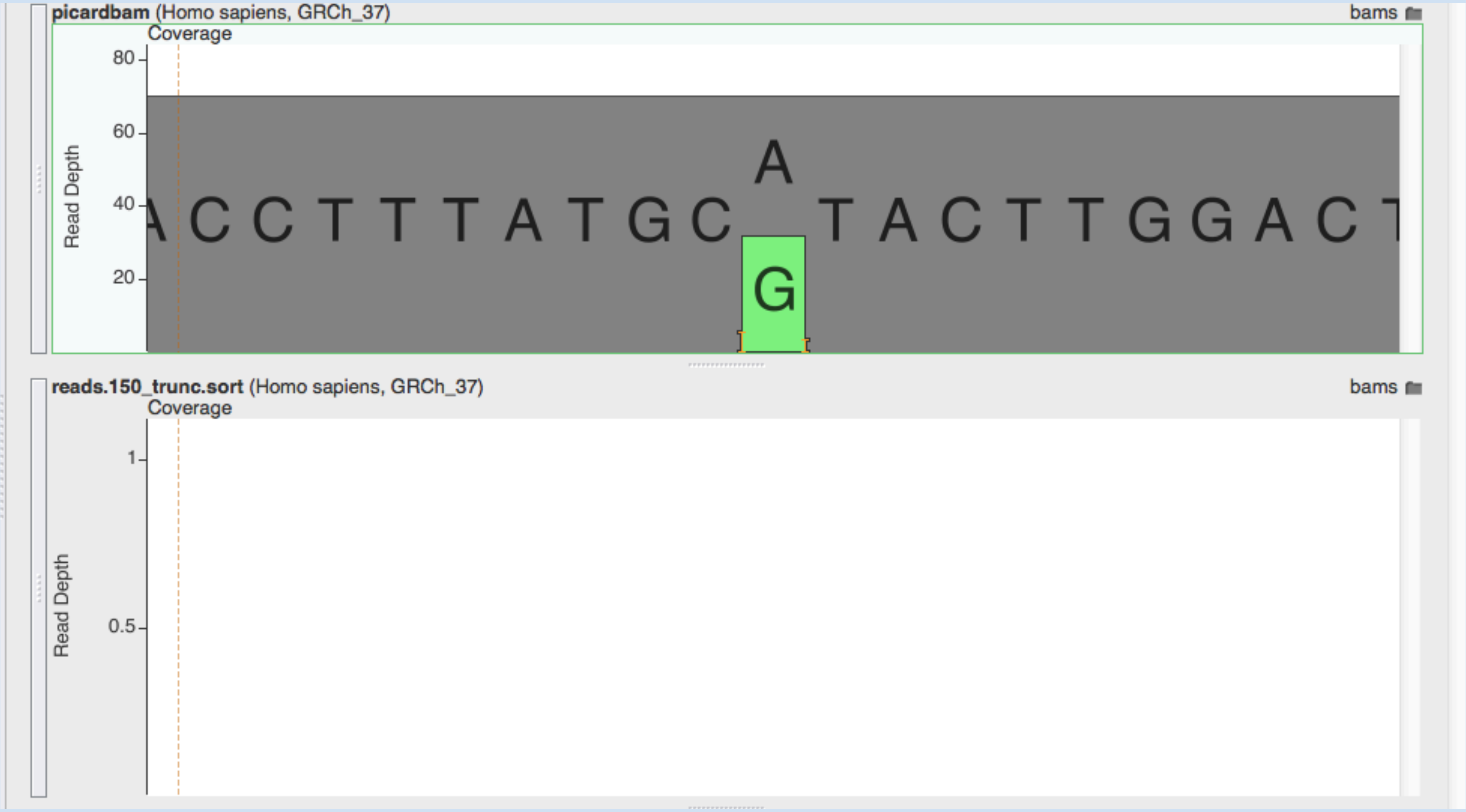
Coverage

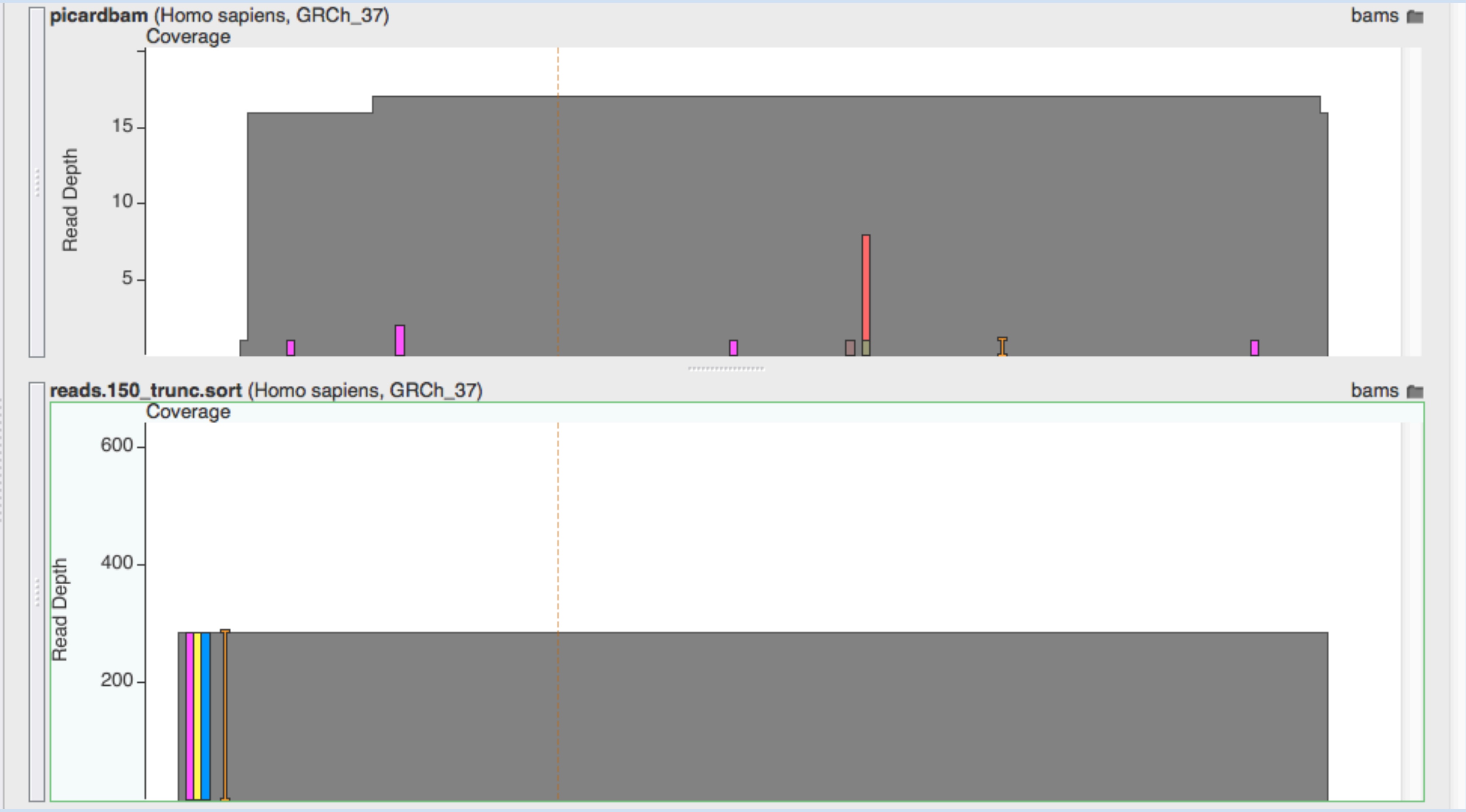

picardbam (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
60
40
20
C A C A A C G C C G T
T depth: 2
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
40,000
30,000
20,000
10,000
C A C A A C G C C G T
T



picardbam (Homo sapiens, GRCh_37)

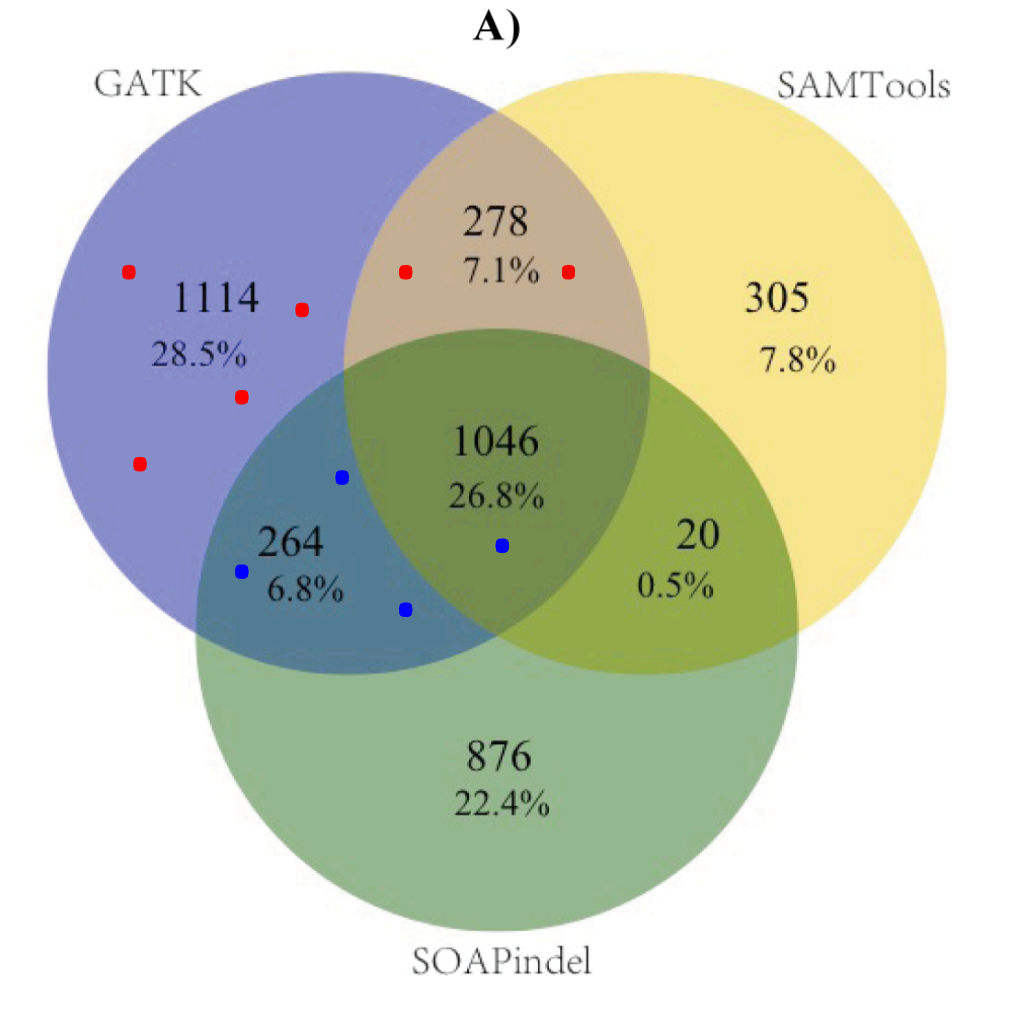
reads.150_trunc.sort (Homo sapiens, GRCh_37)


picardbam (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
80
60
40
20
A
A C C T T T A T G C T A C T T G G A C T
G
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
1
0.5

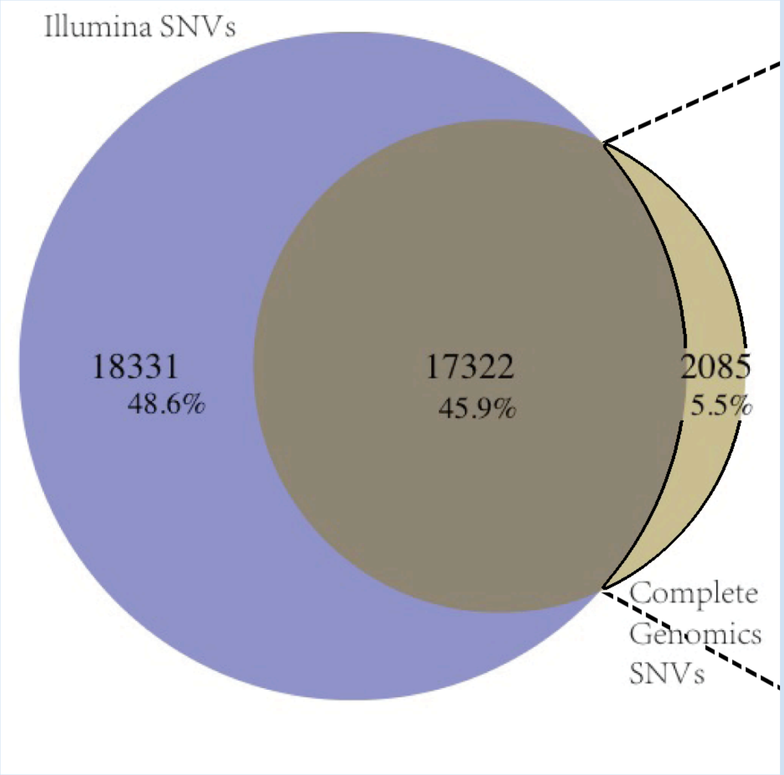


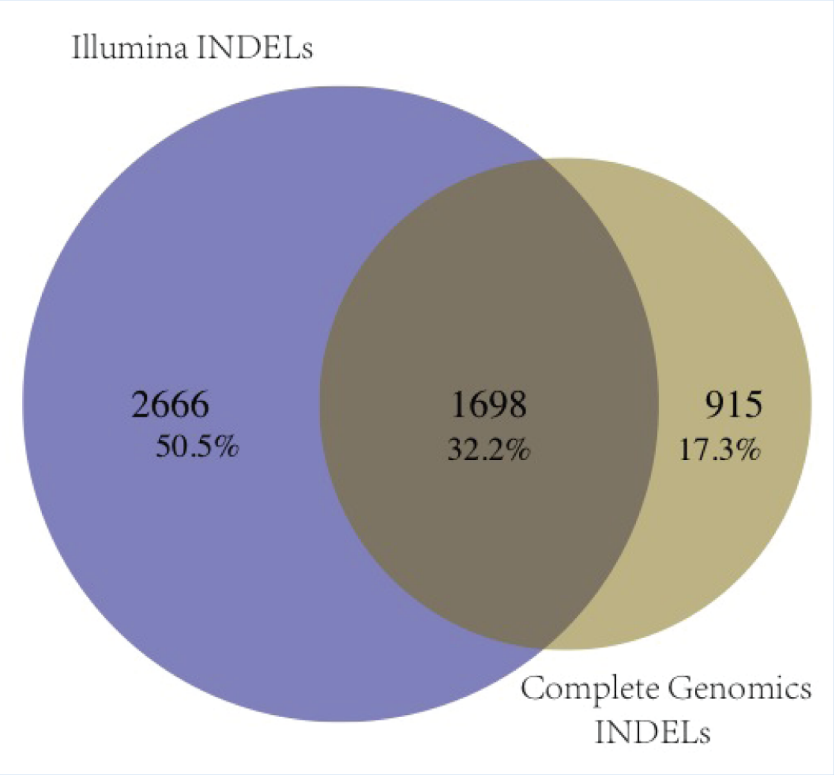
picardbam (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
15
10
5
reads.150_trunc.sort (Homo sapiens, GRCh_37)
bams
Coverage
Read Depth
600
400
200

Indels called by GATK, SOAP and SAMtools

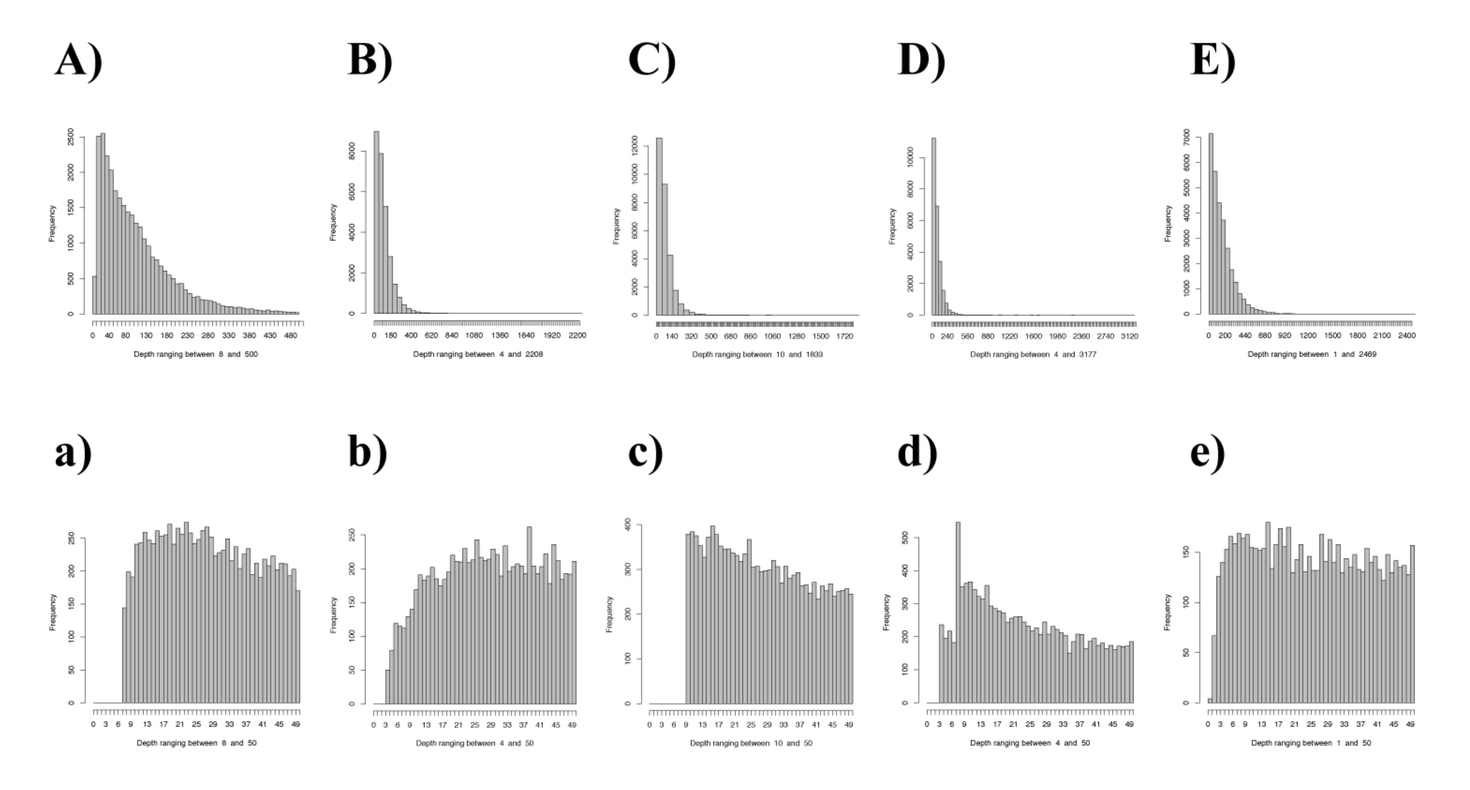


Validation of SNVs and Indels with an additional platform

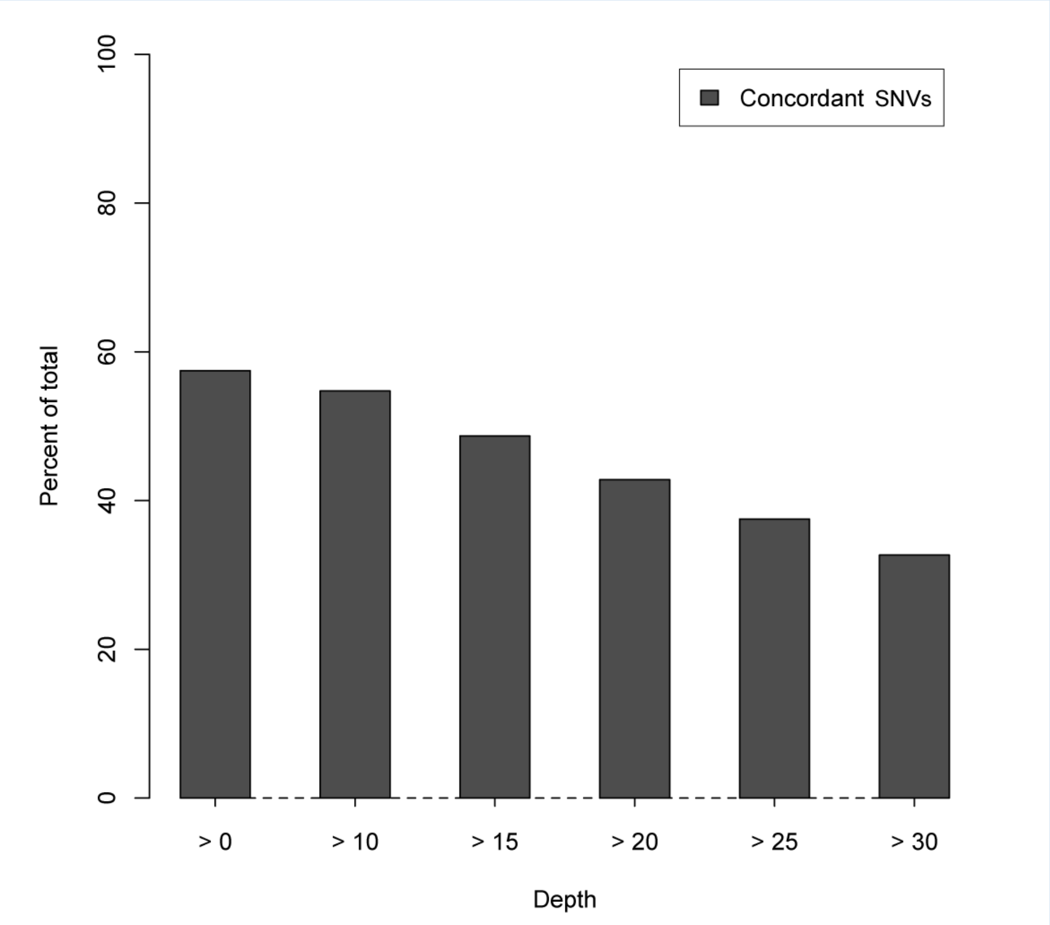




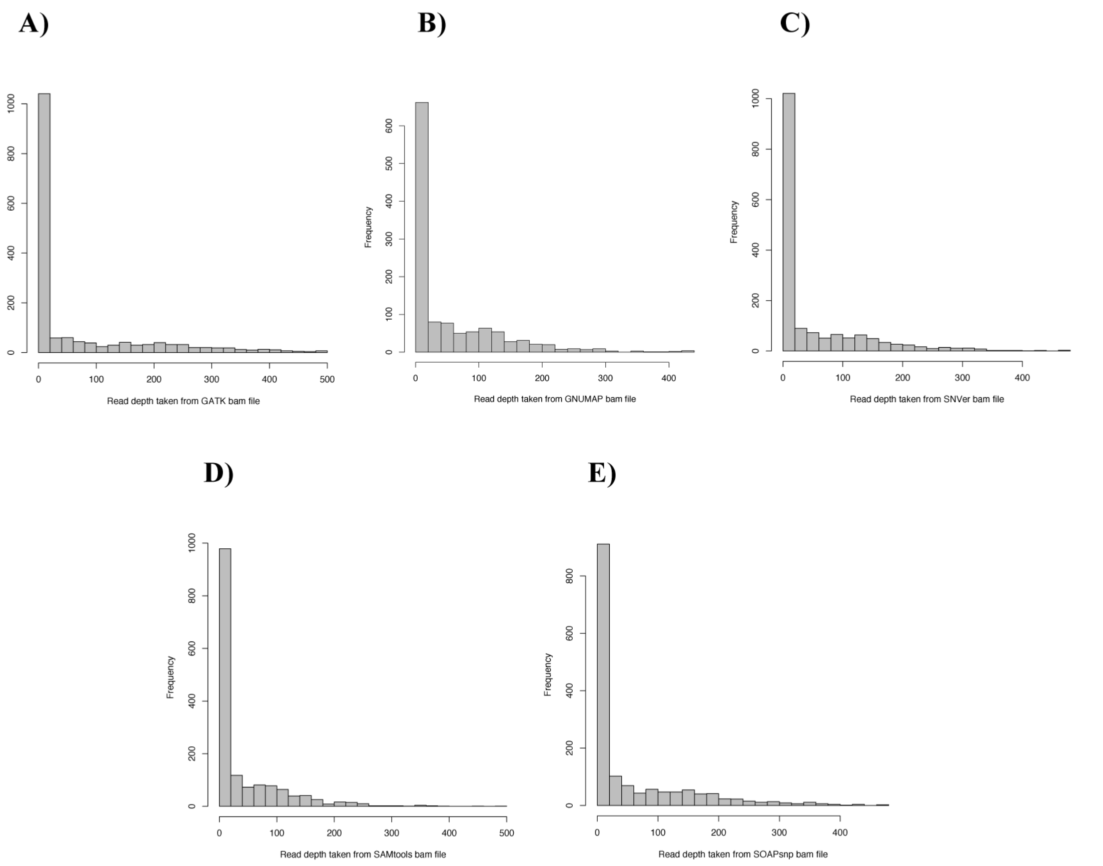
Read Depth



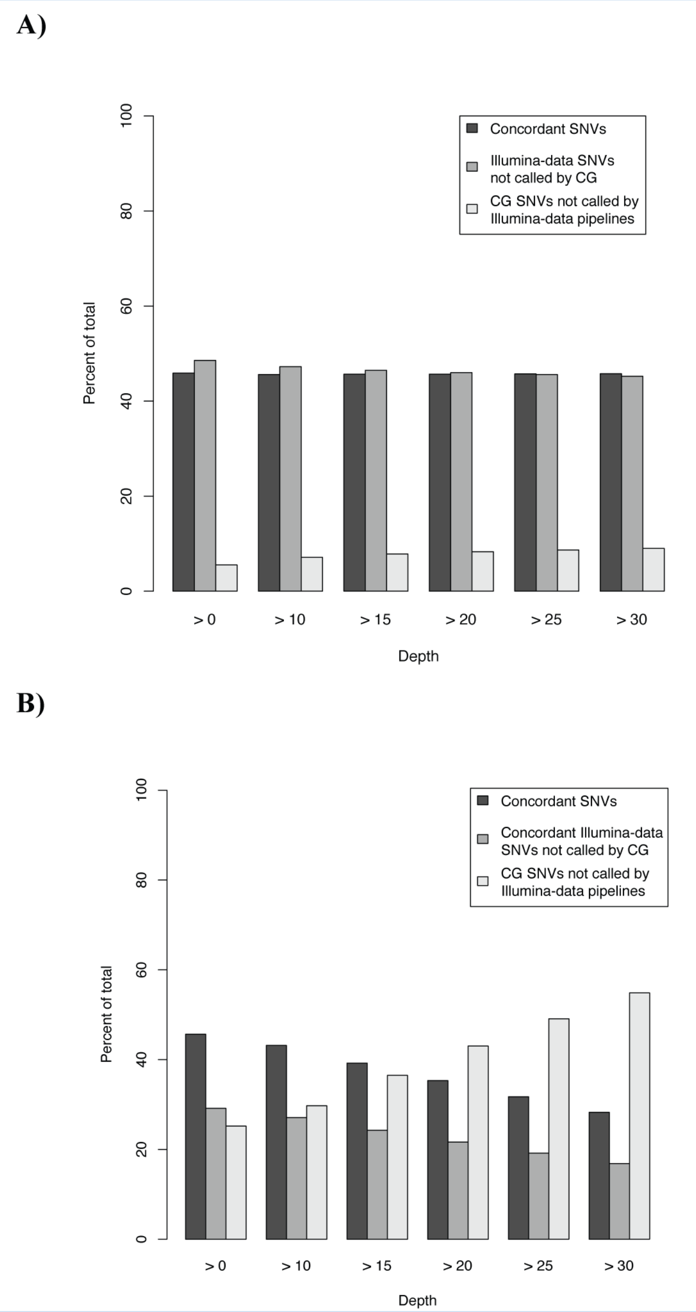
Additional file 2, Figure S3. Histograms of Illumina read depth taken from each pipeline's independently aligned BAM file at genomic coordinates of SNVs called by each of the 5 alignment and variant calling pipelines. **A)** SOAPsnp, **B)** SNVer, **C)** SAMTools, **D)** GNUMAP and **E)** GATK, respectively. Frequency of read depths for all SNVs (**A**, **B**, **C**, **D**, and **E**) as well as for SNVs having depths between 0 and 50 (**a**, **b**, **c**, **d**, and **e**) were plotted.

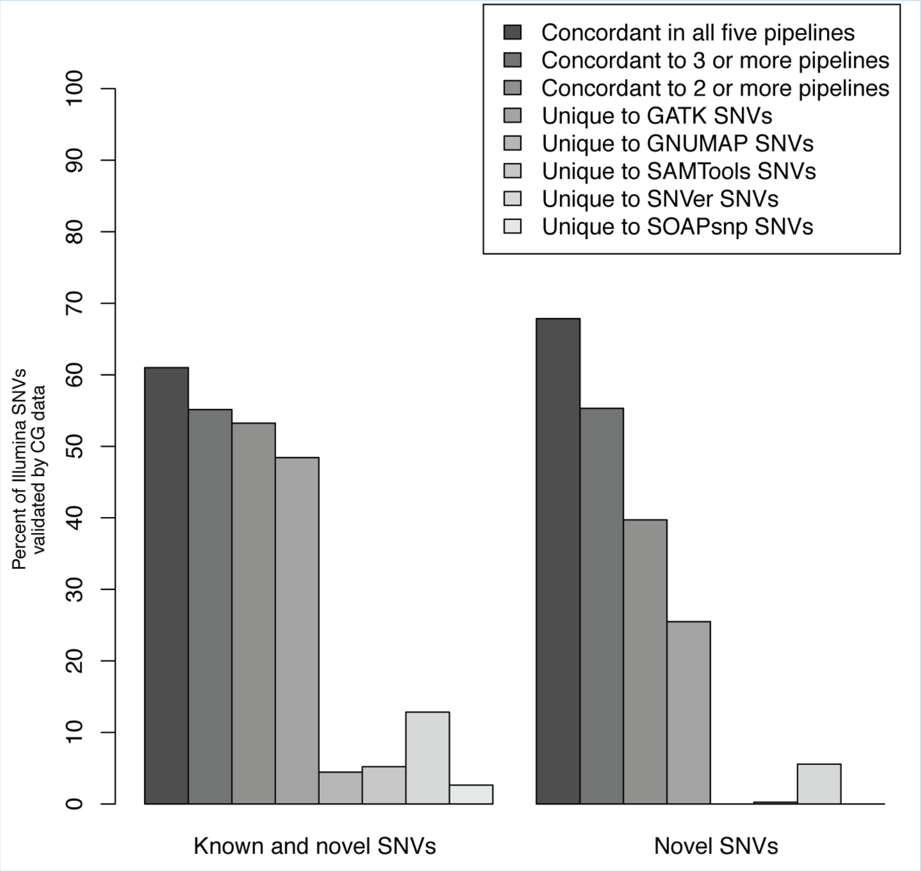


Additional file 2, Figure S4. SNV concordance for a single exome, "k8101-49685", between five alignment and variant detection pipelines: GATK, SOAPsnp, SNVer, SAMTools, and GNUMAP. Concordance between each pipeline was determined by matching the genomic coordinate as well as the base pair change and zygosity for each detected SNV. Concordance was measured at varying Illumina read depth threshold values in each independently aligned BAM file, ranging from >0 (no threshold) to >30 reads.

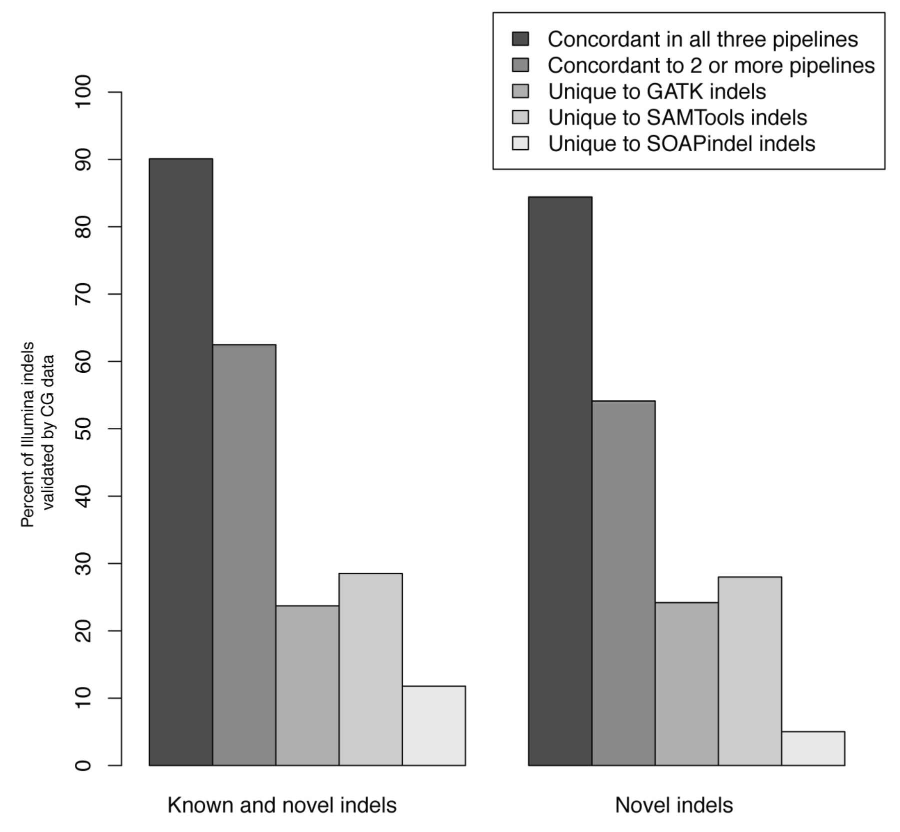


Additional file 2, Figure S5. Histograms of read depth taken from each of the five Illumina pipeline's independently aligned BAM file at genomic coordinates of SNVs that were found by Complete Genomics but not by any of the 5 Illumina pipelines: GATK, GNUMAP, SNVer, SAMTools and SOAPsnp, **A**, **B**, **C**, **D** and **E** respectively. All coordinates fell within the range of the Agilent SureSelect v.2 exons.


A)
100
80
60
40
20
0
Percent of total
Concordant SNVs
Illumina-data SNVs not called by CG
CG SNVs not called by Illumina-data pipelines
> 0
> 10
> 15
> 20
> 25
> 30
Depth
B)
100
80
60
40
20
0
Percent of total
Concordant SNVs
Concordant Illumina-data SNVs not called by CG
CG SNVs not called by Illumina-data pipelines
> 0
> 10
> 15
> 20
> 25
> 30
Depth

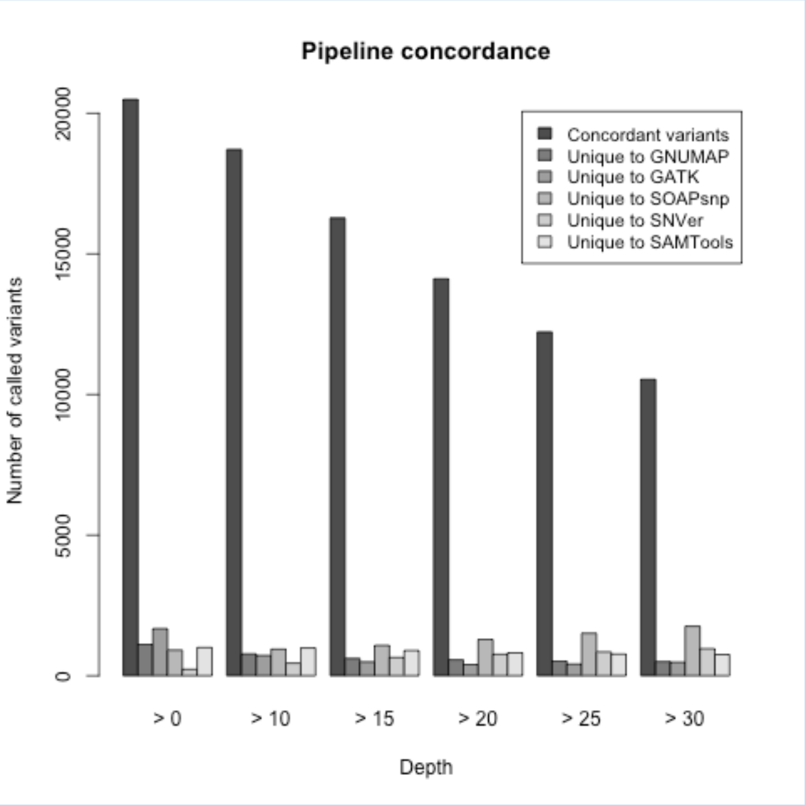


Additional file 2, Figure S8. SNVs called by each Illumina-data pipeline were cross-validated using SNVs called by Complete Genomics, an orthogonal sequencing technology, in sample “k8101-49685”. The percentage of Illumina SNVs that were validated by CG sequencing was measured for variants having varying degrees of Illumina-data pipeline concordance. The same analysis was performed for variants that were considered novel (absent in dbSNP135).



Additional file 2, Figure S9. Indels called by each Illumina-data pipeline were cross-validated using indels called by Complete Genomics for sample “k8101-49685”. The percentage of Illumina indels that were validated by CG sequencing was measured across varying degrees of Illumina pipeline concordance. The same analysis was done for novel indels (indels not found in dbSNP 135).

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



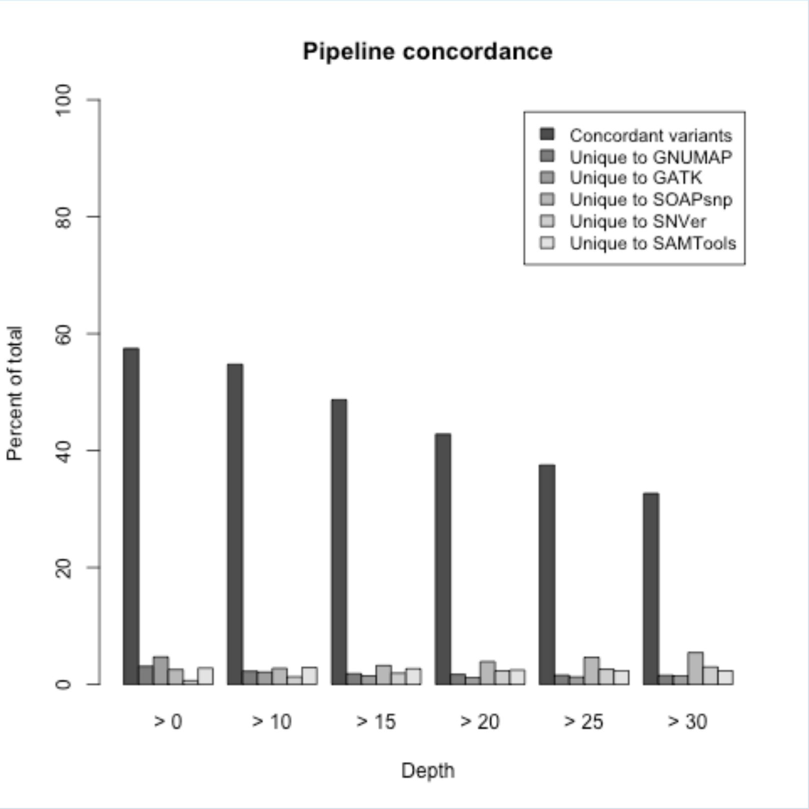


Table 2. Quality evaluation of variant detection using different variant-calling pipelines.

	Sensitivity		Specificity	
	Mean*	SD	Mean*	SD
SOAPsnp	94.68	2.26	99.79	0.03
GATK1.5	95.34	1.16	99.72	0.08
SNVer	92.33	4.40	99.78	0.04
GNUMAP	86.60	3.23	99.64	0.06
SAMtools	94.47	4.22	99.59	0.16
Any pipeline	97.67	1.20	99.62	0.11
≥2 pipelines*	96.64	2.28	99.69	0.07
≥3 pipelines*	95.62	3.13	99.73	0.05
≥4 pipelines*	92.60	3.40	99.82	0.04
5 pipelines*	80.58	5.26	99.87	0.01

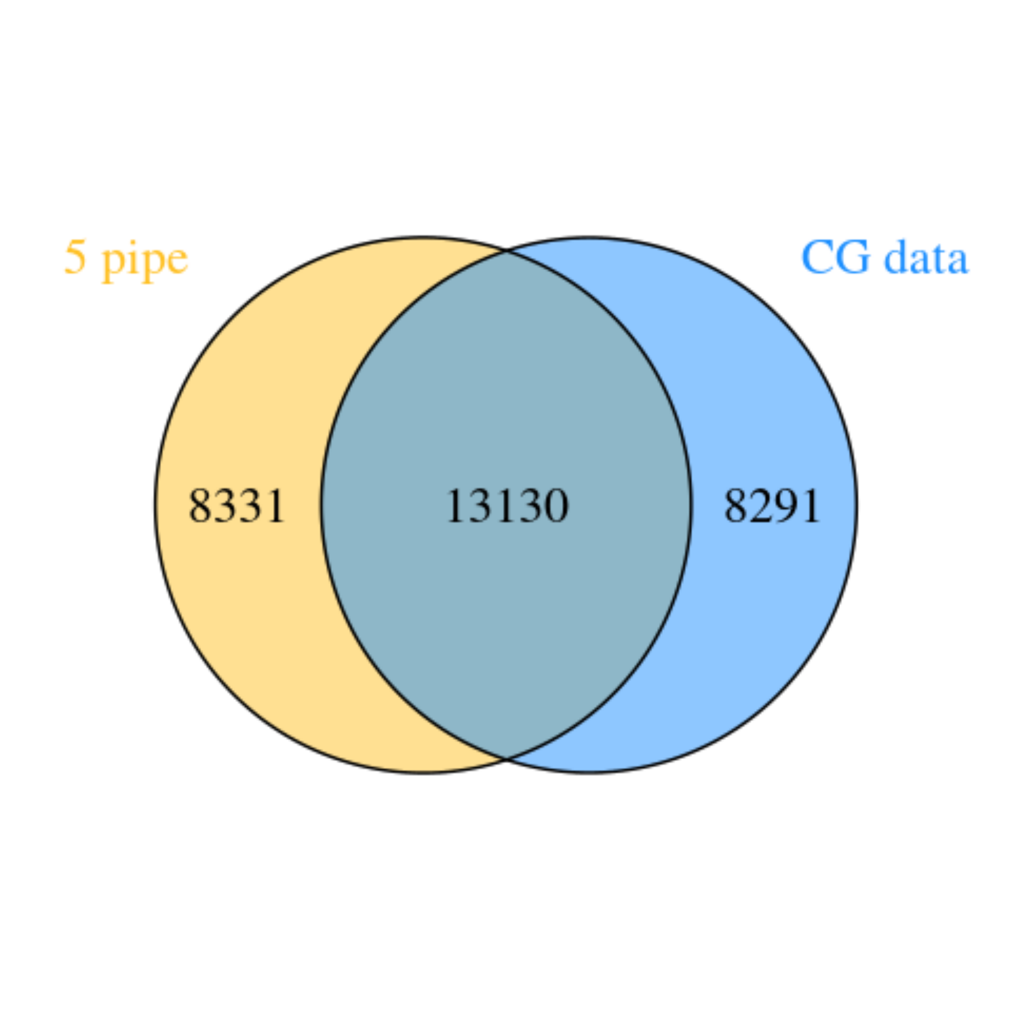
*Intersection of variants contained in the number of pipelines specified. Sensitivity and specificity was calculated for each pipeline by comparing Illumina Human610-Quad version 1 SNP arrays with exome-capture sequencing results, based on the four samples whose genotyping data was available.

Table S1. **Concordance rates with common SNPs genotyped on Illumina 610K genotyping chips.**

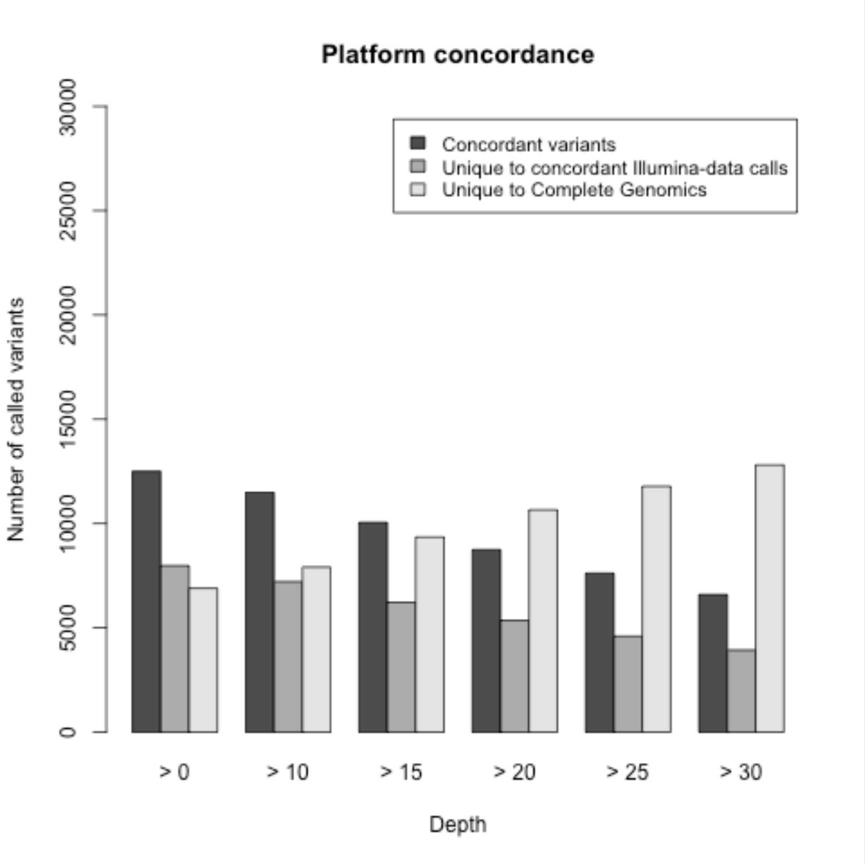
Sample	Software	Compared Sites	Concordance Sites	Concordance rate
Mother-1	SOAPsnp	6088	6074	99.77%
	GATK 1.5	6249	6224	99.60%
	SNVer	5723	5708	99.74%
	GNUMAP	5458	5434	99.56%
	SAMTools	5885	5848	99.37%
Son-1	SOAPsnp	6366	6353	99.80%
	GATK 1.5	6341	6323	99.72%
	SNVer	6255	6239	99.74%
	GNUMAP	5850	5828	99.62%
	SAMTools	6383	6362	99.67%
Son-2	SOAPsnp	6412	6401	99.83%
	GATK 1.5	6426	6413	99.80%
	SNVer	6336	6325	99.83%
	GNUMAP	5906	5889	99.71%
	SAMTools	6477	6450	99.58%
Father-1	SOAPsnp	6247	6238	99.86%
	GATK 1.5	6304	6288	99.75%
	SNVer	6205	6192	99.79%
	GNUMAP	5805	5786	99.67%
	SAMTools	6344	6327	99.73%

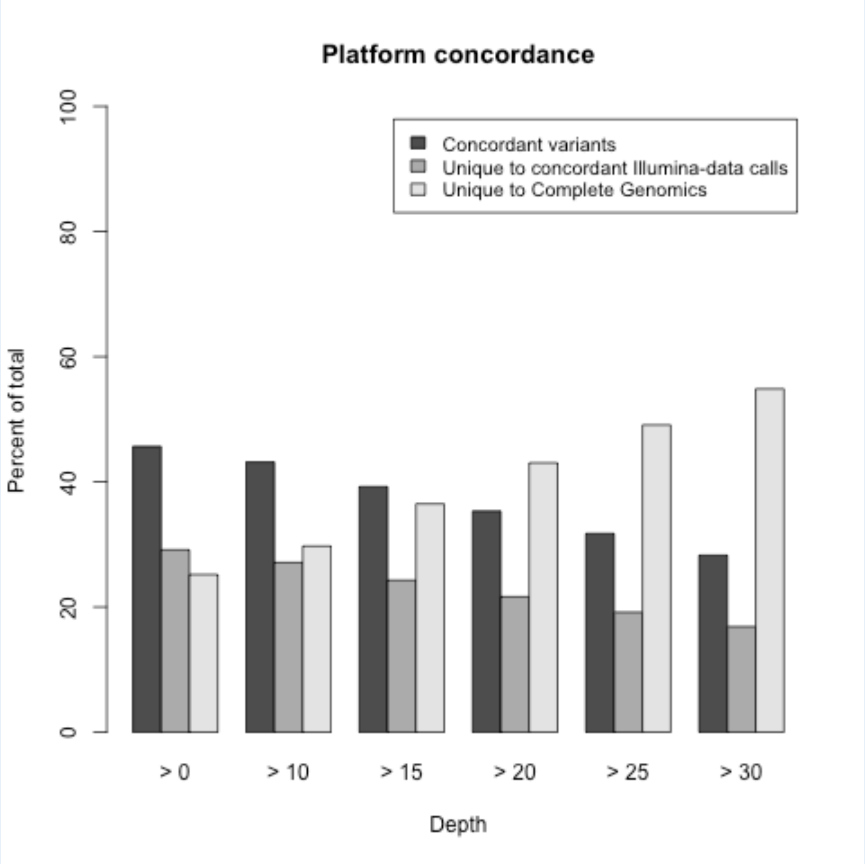
All pipelines are very good with identifying already known, common SNPs.

Taking SNVs concordant in 5 Illumina pipelines, and comparing to SNVs in Complete Genomics Data from same sample

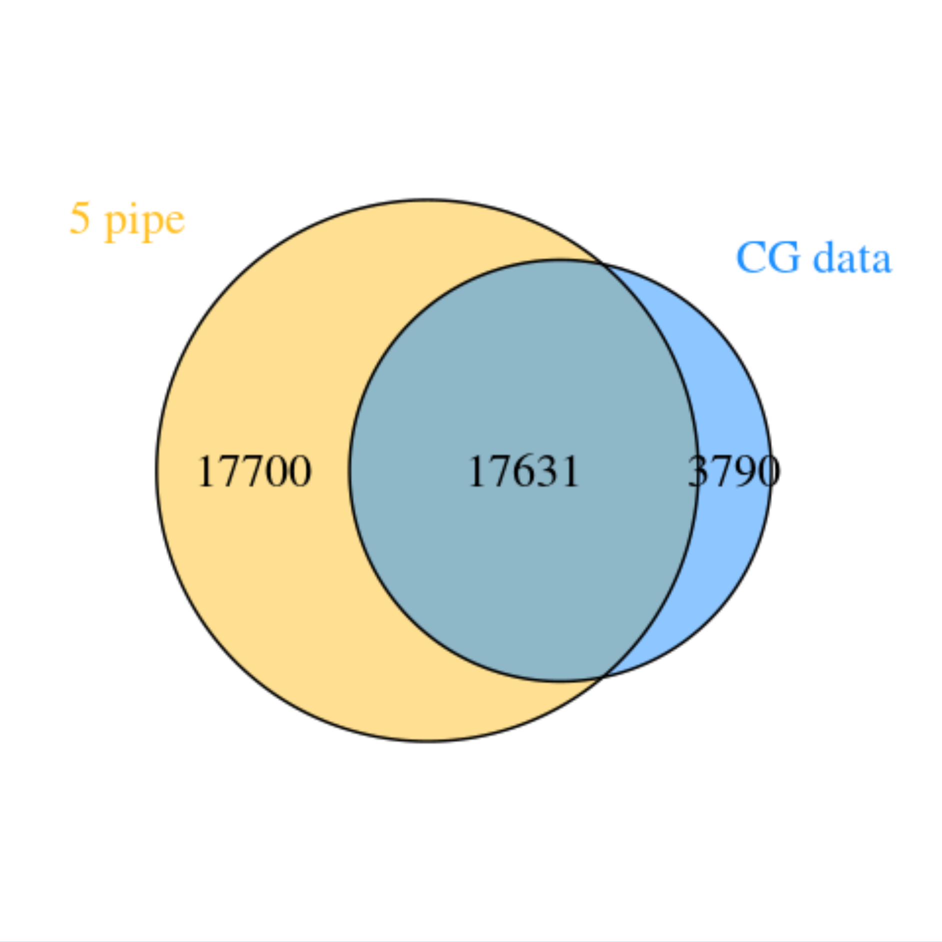


Taking SNVs concordant in 5 Illumina pipelines as per READ DEPTH, and comparing to SNVs in Complete Genomics Data from same sample

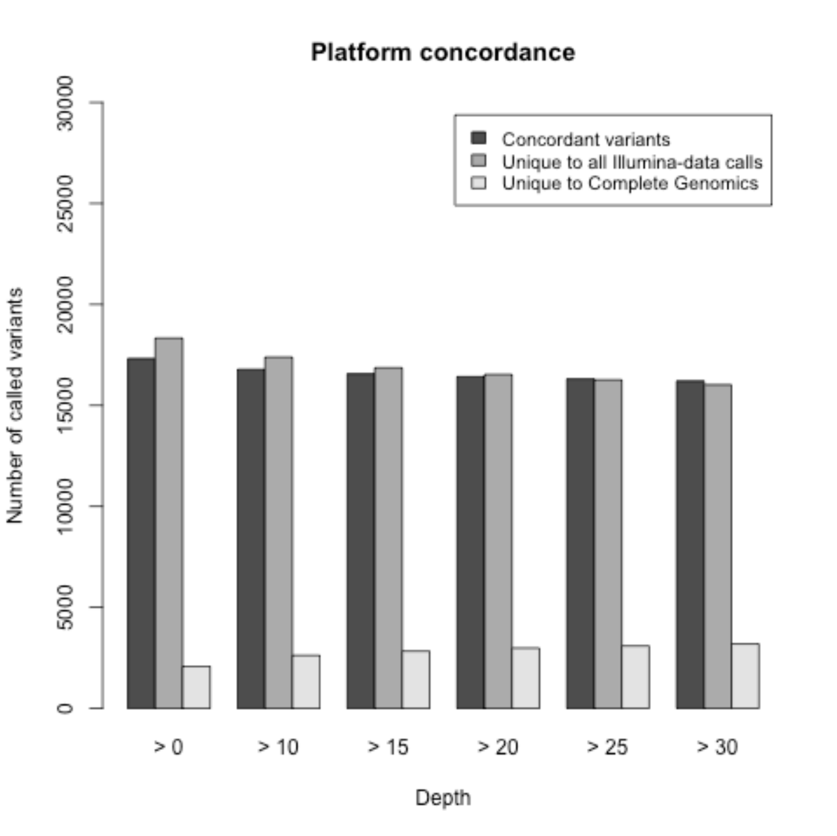


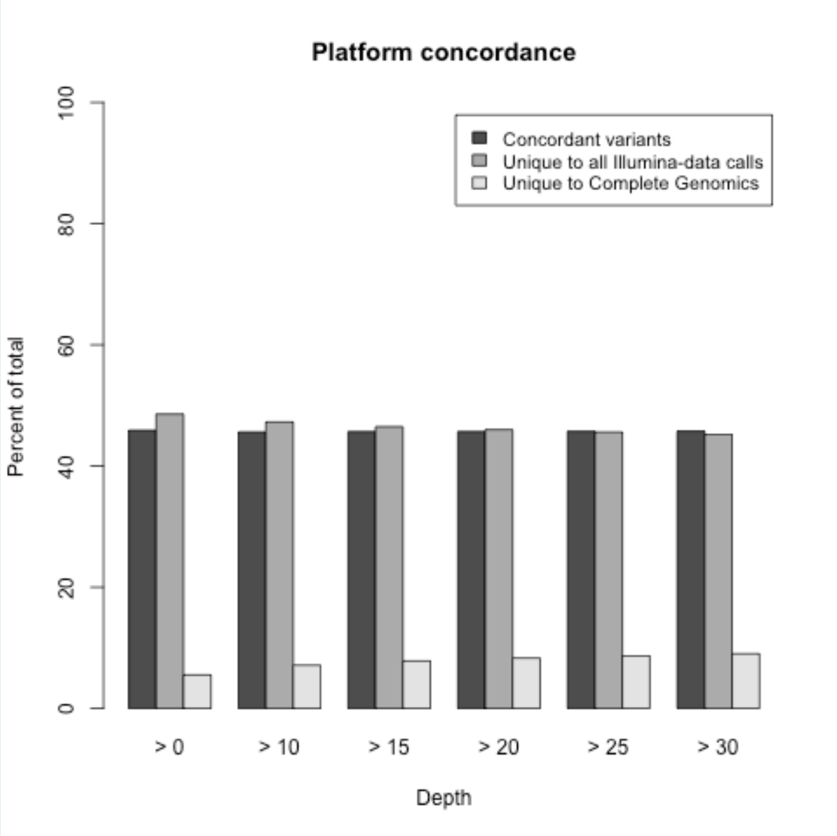


Taking SNVs found by ALL 5 Illumina pipelines (Union), and comparing to SNVs in Complete Genomics Data from same sample

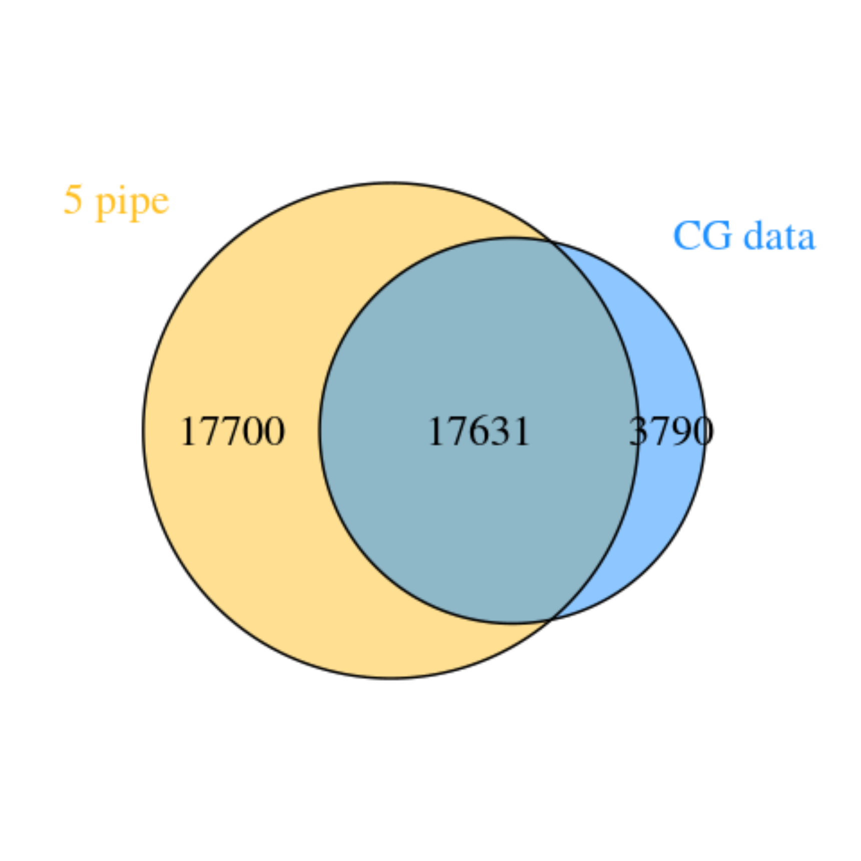


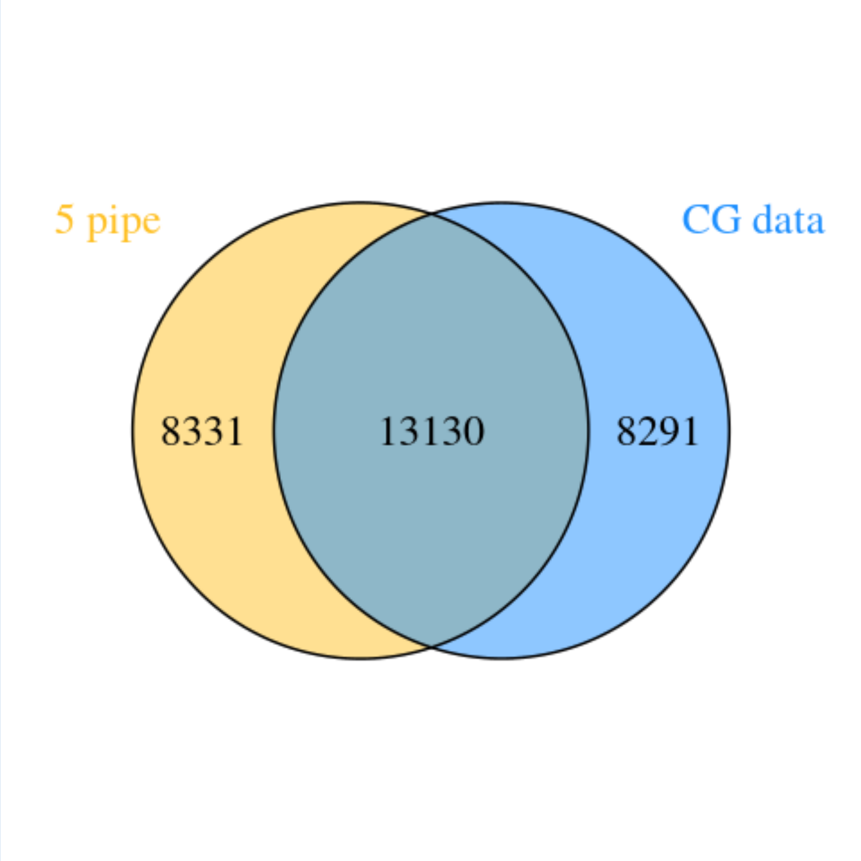
Taking the UNION of all SNVs called by Illumina pipelines, as per READ DEPTH, and comparing to SNVs in Complete Genomics Data from same sample





Comparing the UNION versus the CONCORDANCE of 5 pipelines to the Complete Genomics Data

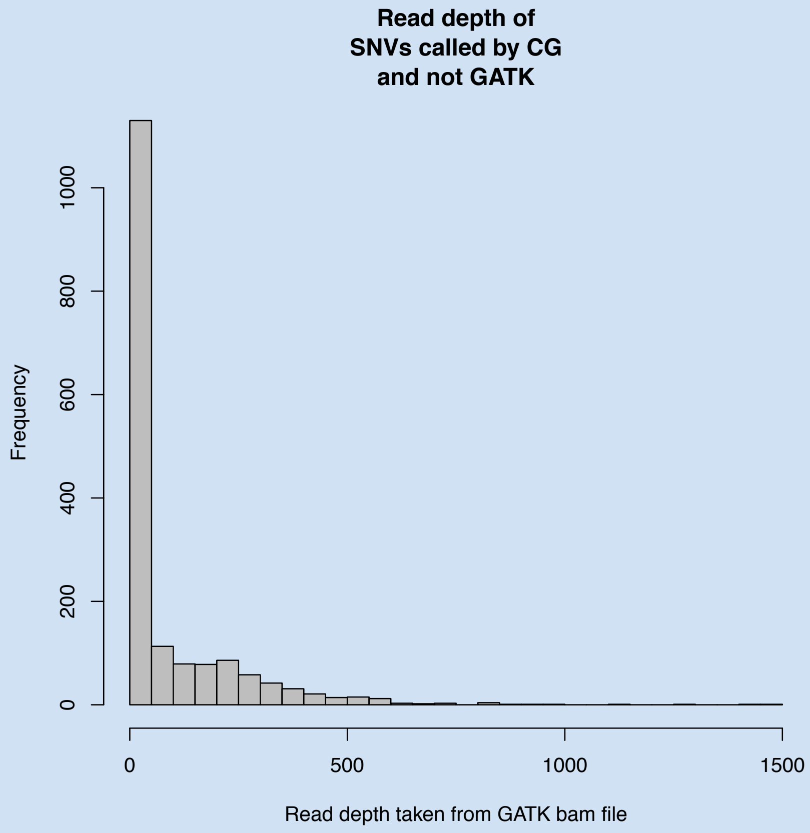


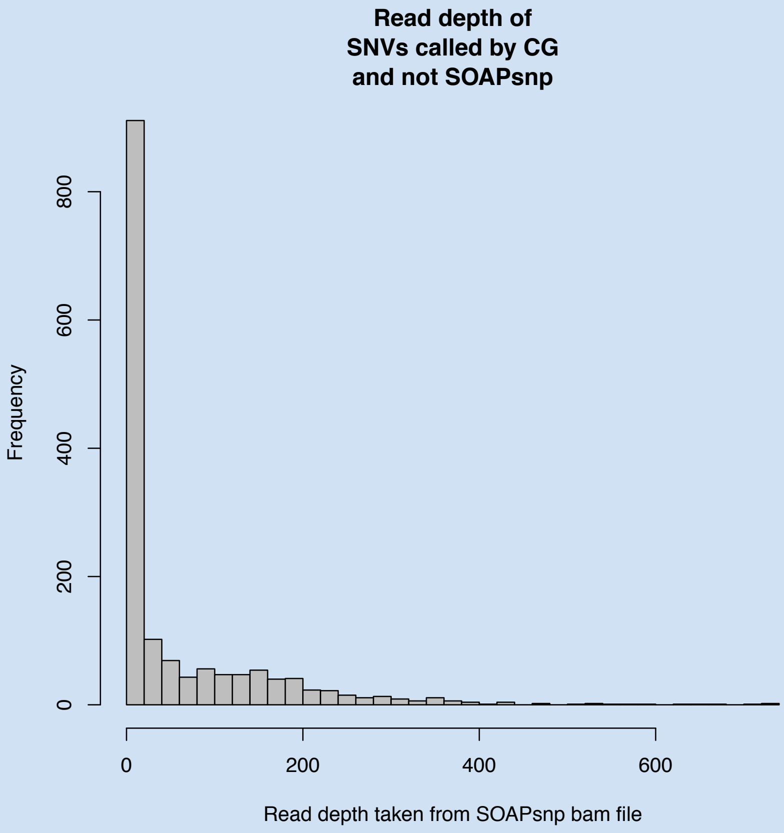


Union of Illumina variants

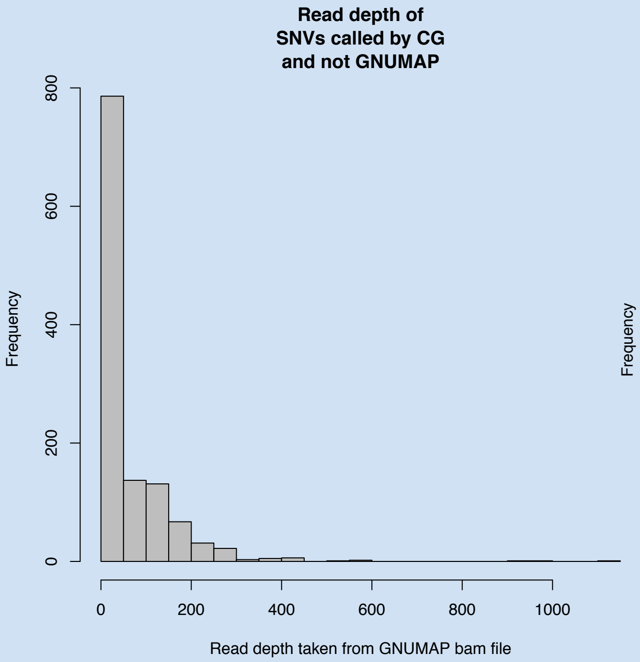
Concordant Illumina variants

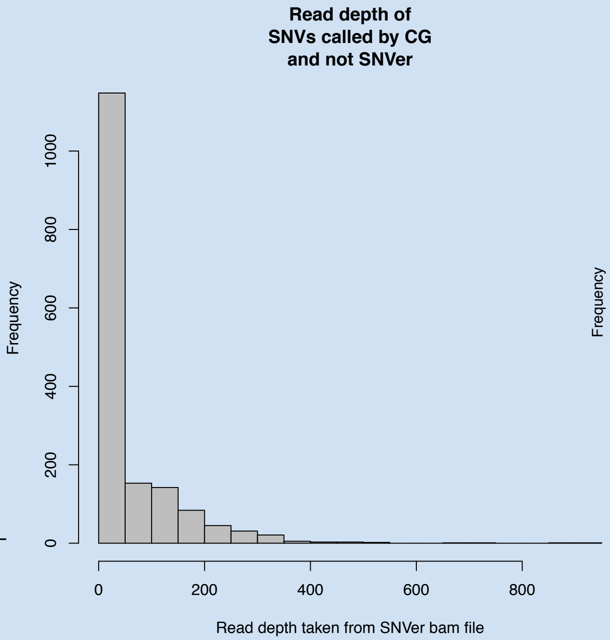
Read Depth of Illumina Reads for variants called by Complete Genomics but NOT by GATK or SOAP pipelines

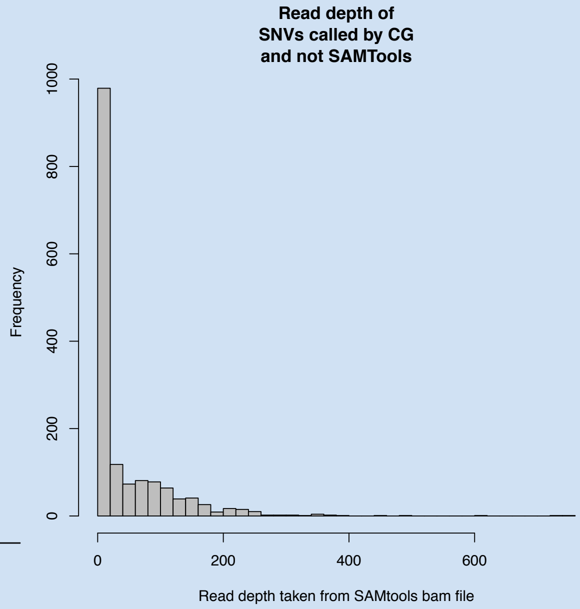




Read Depth of Illumina Reads for variants called by Complete Genomics but NOT by GNUMAP, SNVer or SamTools pipelines







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.

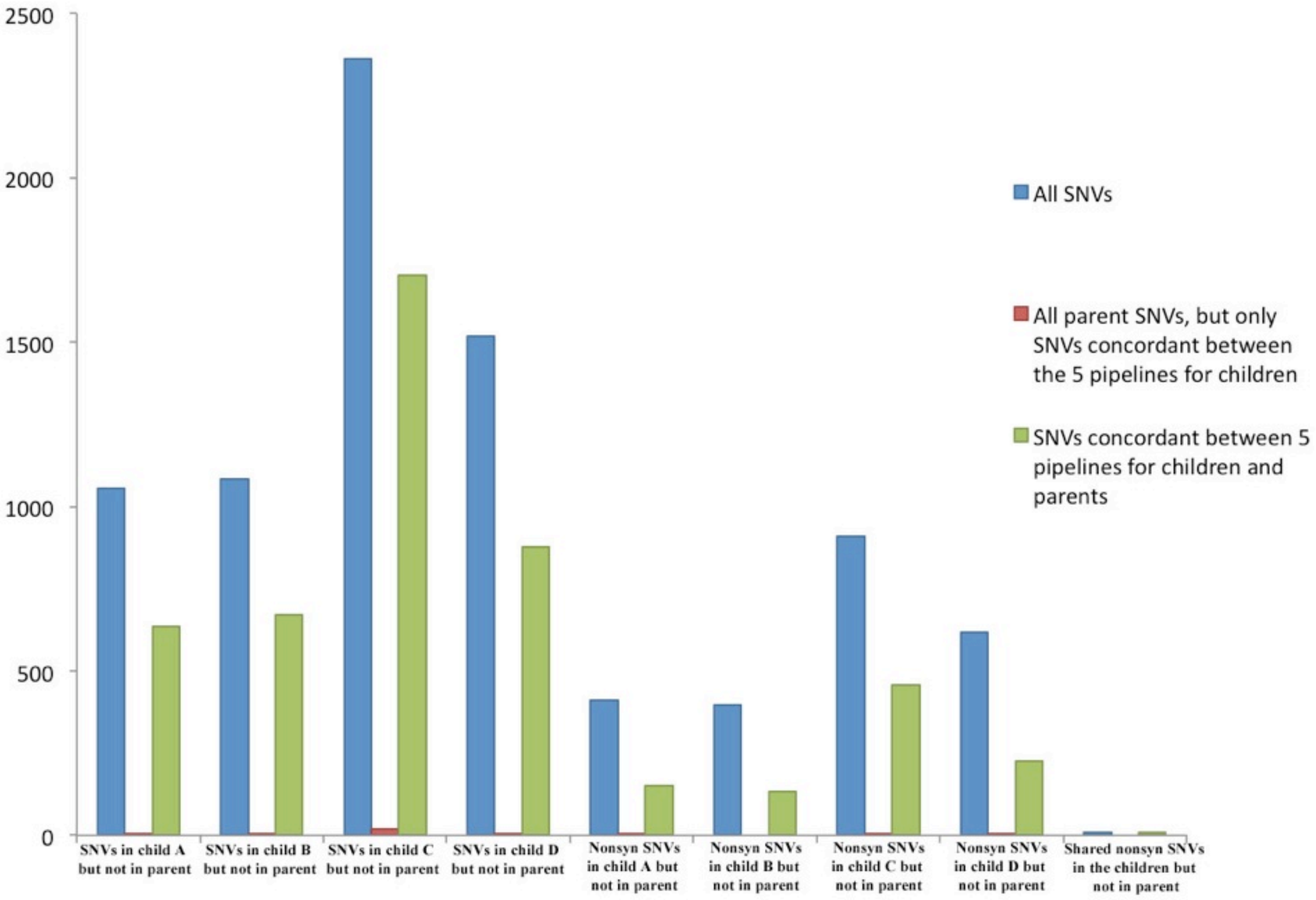

2500
2000
1500
1000
500
0
🟦 All SNVs
🟥 All parent SNVs, but only SNVs concordant between the 5 pipelines for children
🟩 SNVs concordant between 5 pipelines for children and parents
SNVs in child A but not in parent
SNVs in child B but not in parent
SNVs in child C but not in parent
SNVs in child D but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

Figure 1. Mean single-nucleotide variants (SNV) concordance over 15 exomes between five alignment and variant-calling pipelines. The alignment method used, followed by the SNV variant calling algorithm is annotated here in shorthand: BWA-GATK, SOAP-Align-SOAPsnp, BWA-SNVer, BWA-SAMtools, and GNUMAP-GNUMAP. **(A)** Mean SNV concordance between each pipeline was determined by matching the genomic coordinate as well as the base-pair change and zygosity for each detected SNV. **(B)** The same analysis as in (A) but filtered to include only SNVs already found in dbSNP135. **(C)** The same analysis as in (A), but filtered to include novel SNVs (that is, SNVs not found in dbSNP135).

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.

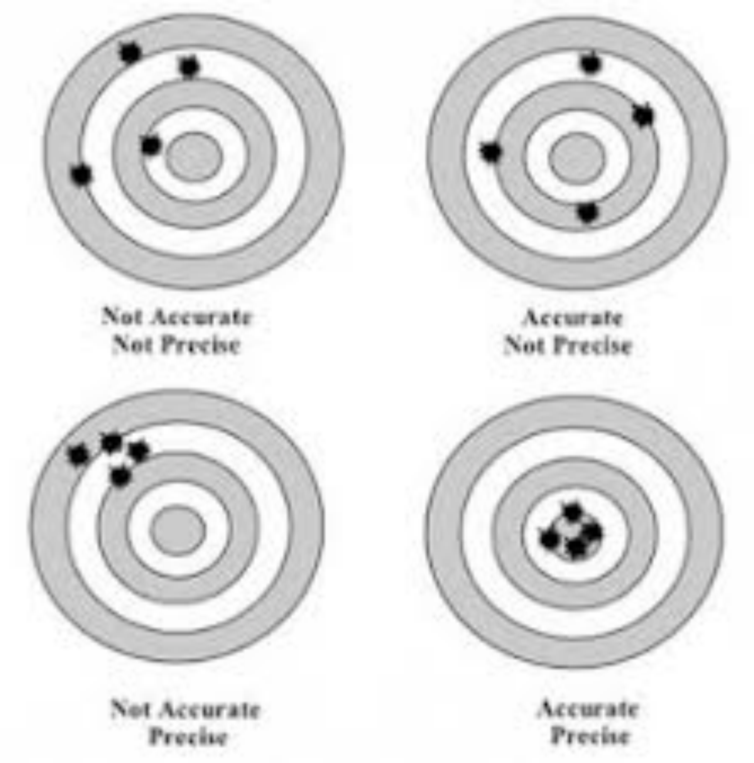
2) BWA - Sam format to Bam format-Picard to remove duplicates - **SamTools** version 0.1.18 to generate genotype calls -- The “mpileup” command in SamTools were used for identify SNPs and indels.

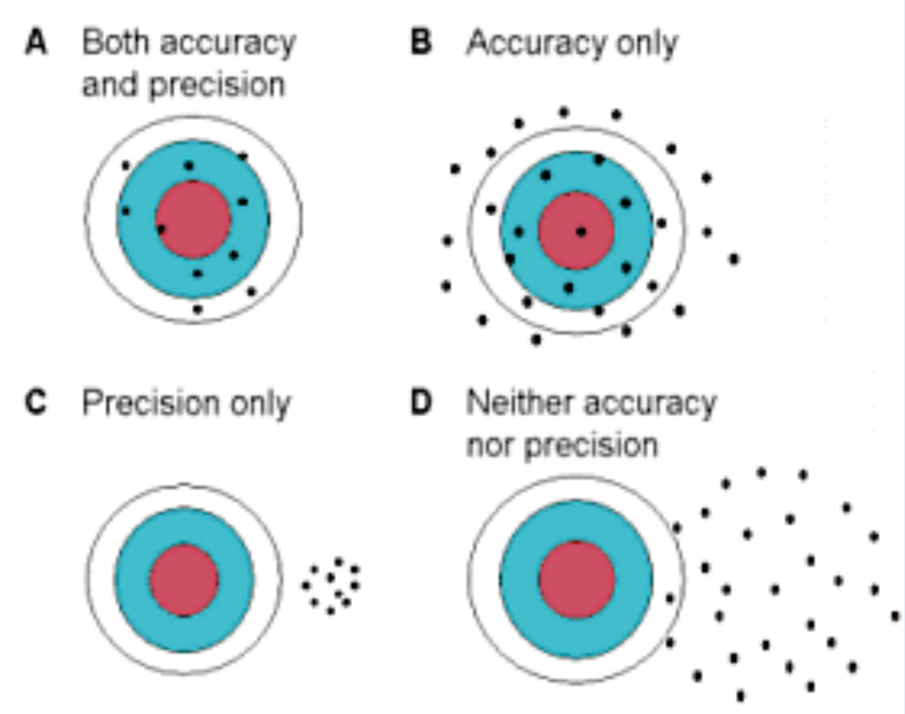
3) **SOAP**-Align – SOAPsnp – then BWA-SOAPindel (adopts local assembly based on an extended de Bruijn graph)

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**


Not Accurate
Not Precise
Accurate
Not Precise
Not Accurate
Precise
Accurate
Precise


A Both accuracy and precision
B Accuracy only
C Precision only
D Neither accuracy nor precision

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

1) BWA - **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. For SNVs and indels.

2) BWA - **SamTools** version 0.1.18 to generate genotype calls -- The "mpileup" command in SamTools was used for identify SNVs and indels.

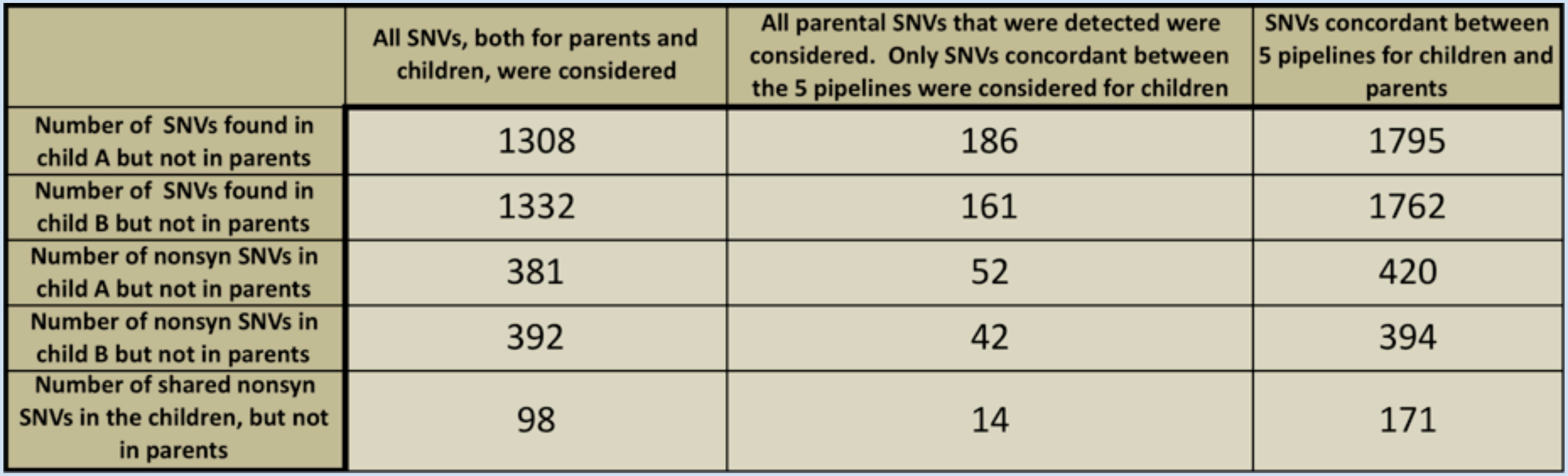
3) **SOAP**-Align – SOAPsnp for SNVs– and BWA-SOAPindel (adopts local assembly based on an extended de Bruijn graph) for indels.

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), for SNVs only.

5) BWA - Sam format to Bam format - Picard to remove duplicates – **SNVer** , for SNVs only

	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

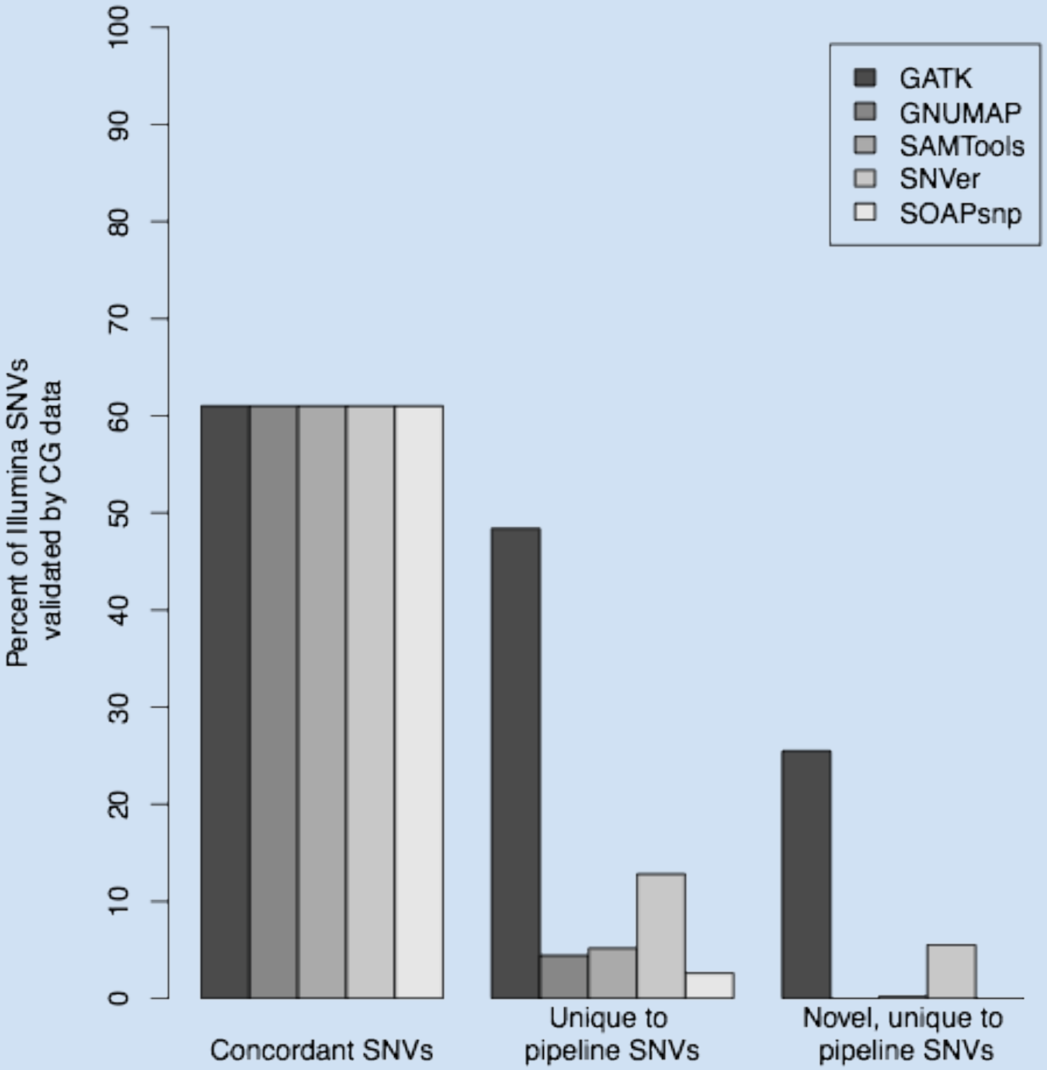
Optimizing pipeline based on literature value of ~1 true de novo protein-altering mutation per exome



	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.

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



CSH Cold Spring Harbor Laboratory
STANLEY INSTITUTE FOR COGNITIVE GENOMICS

THE
STANLEY MEDICAL
RESEARCH INSTITUTE

NIMH
National Institute
of Mental Health