Implementation of Variant Calling Algorithms in Clinical Genome Sequencing

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

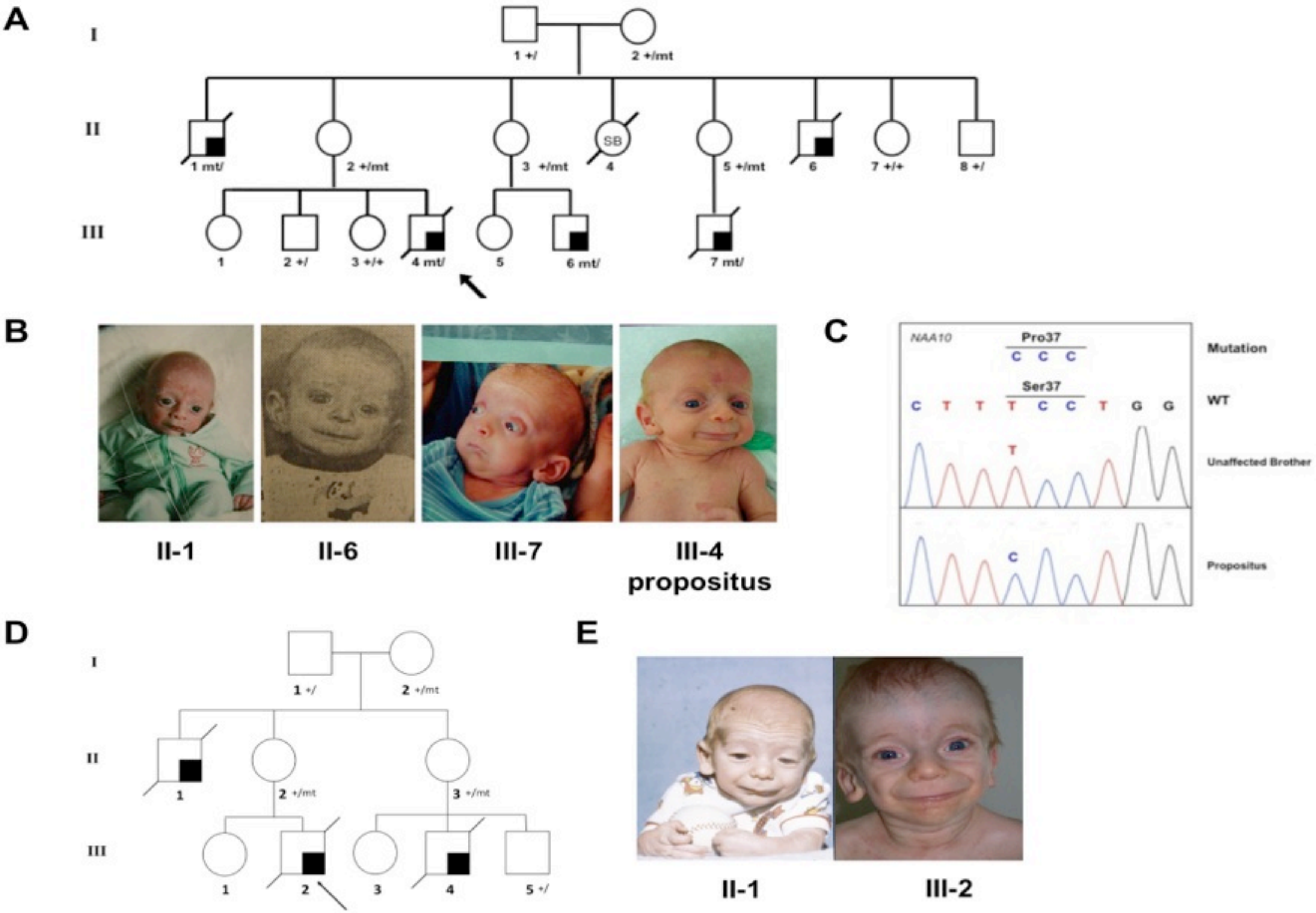
- I do not accept salary from anyone other than my current employer, CSHL.

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

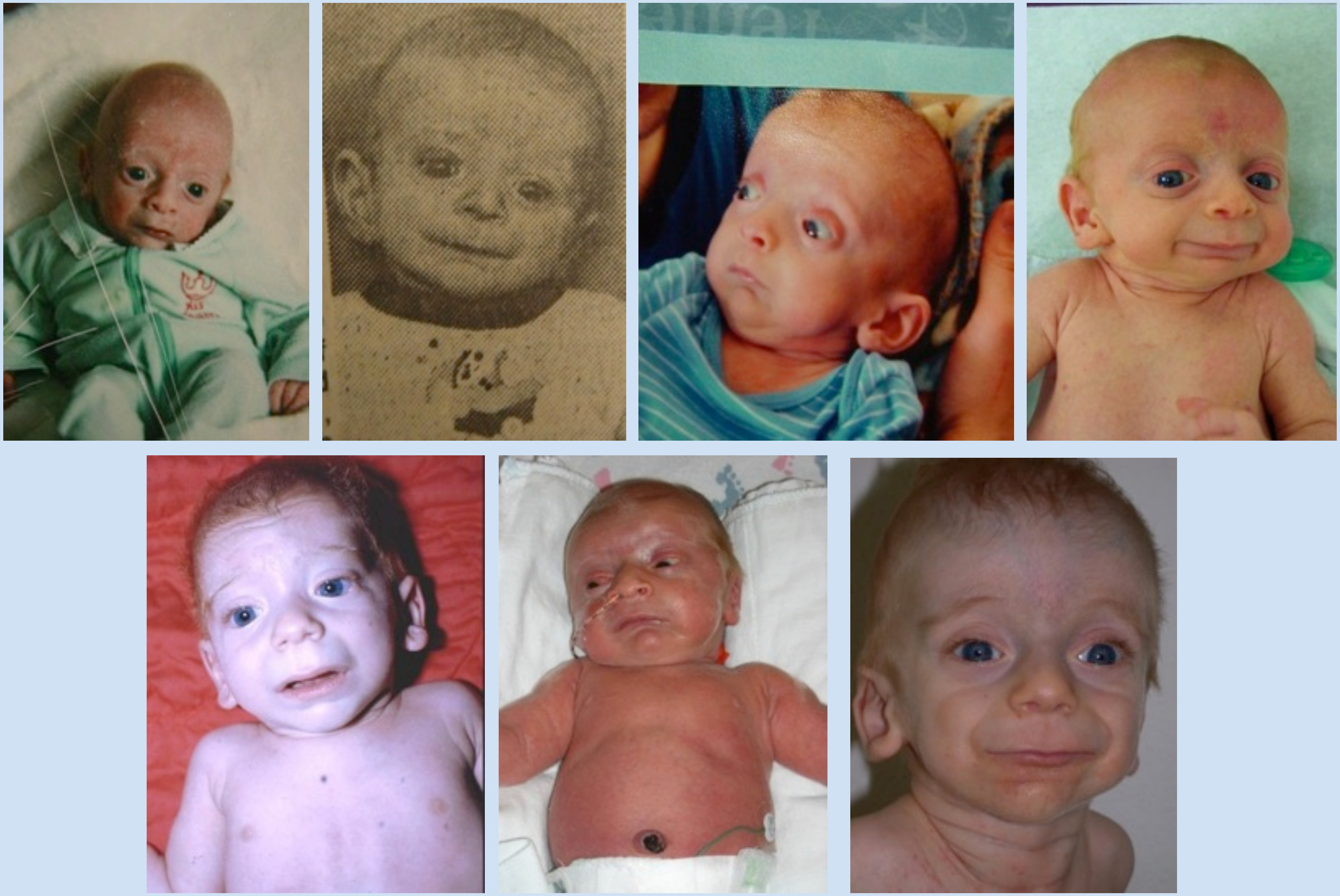
- I worked on the Clarity Challenge as an unpaid medical consultant to: Omicia
  Unlocking Individualized Medicine

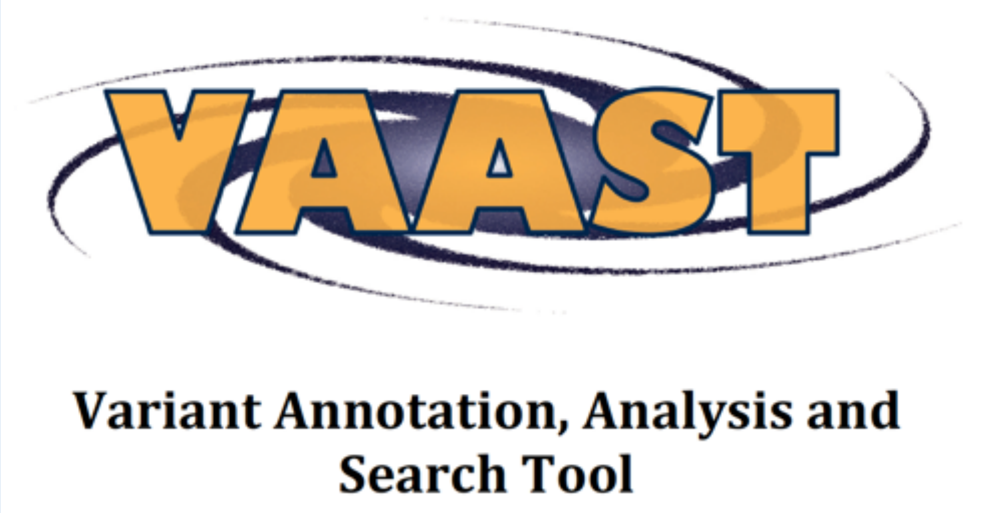
THE UNIVERSITY OF UTAH

UTAH foundation for BIOMEDICAL research


A
I
1 +/
2 +/mt
II
1 mt/
2 +/mt
3 +/mt
SB
4
5 +/mt
6
7 +/+
8 +/
III
1
2 +/
3 +/+
4 mt/
5
6 mt/
7 mt/
B
II-1
II-6
III-7
III-4
propositus
C
NAA10
Pro37
C C C
Mutation
Ser37
C T T T C C T G G
WT
T
Unaffected Brother
C
Propositus
D
I
1 +/
2 +/mt
II
1
2 +/mt
3 +/mt
III
1
2
3
4
5 +/
E
II-1
III-2

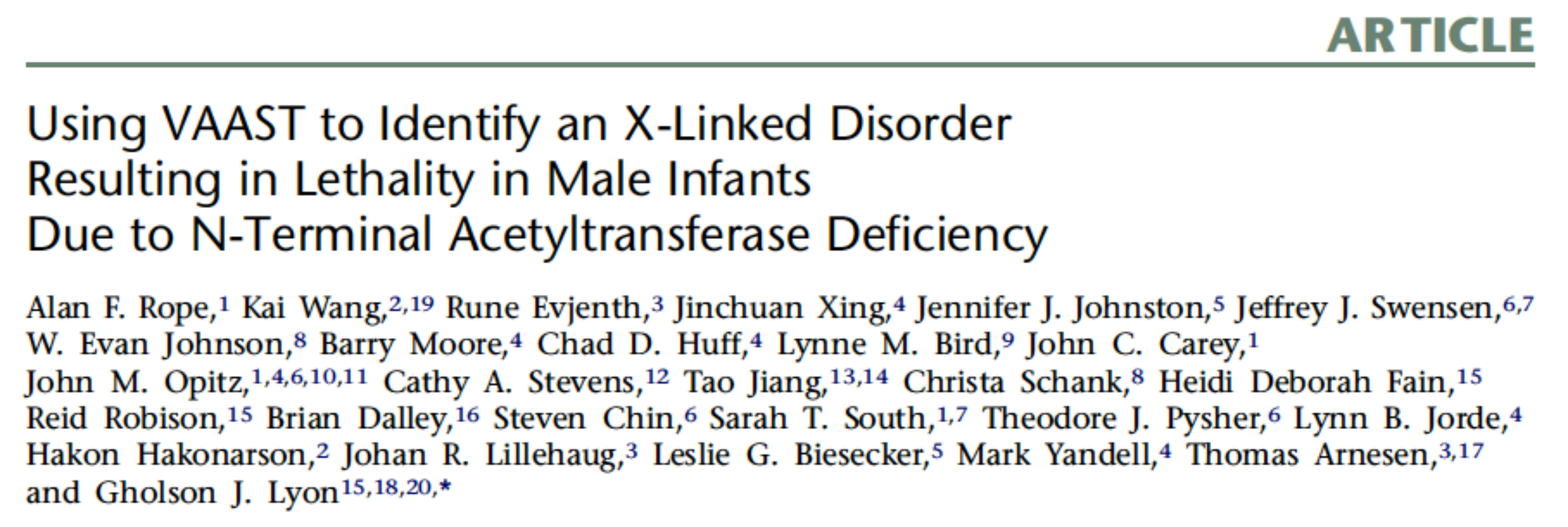
Ogden Syndrome, in honor of where the first family lives, in Ogden, Utah





Variant Annotation, Analysis and Search Tool

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

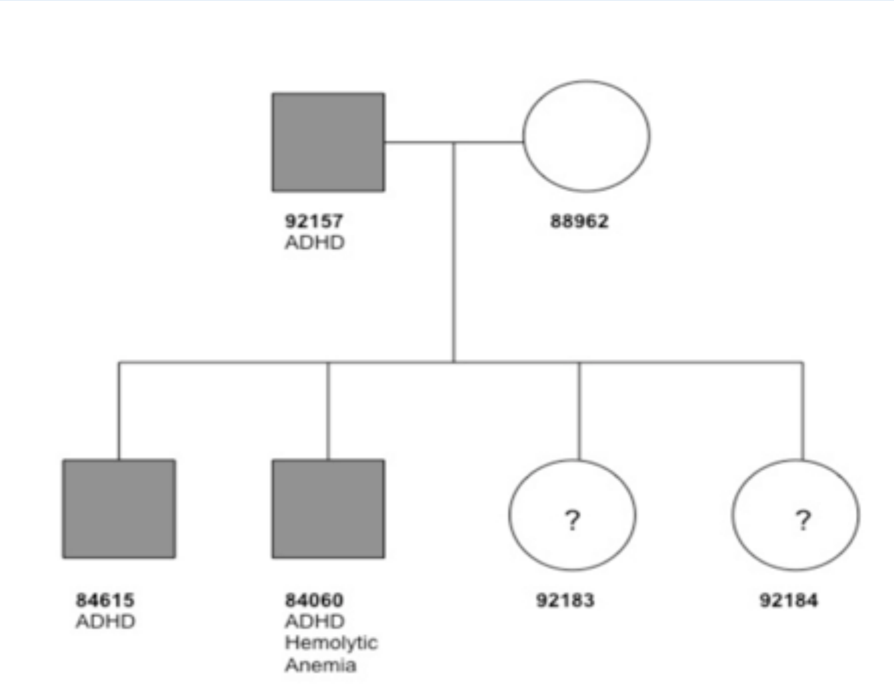
Using VAAST to Identify an X-Linked Disorder Resulting in Lethality in Male Infants Due to N-Terminal Acetyltransferase Deficiency

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,*]

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.



Moving Exome and WGS into a Clinical Setting requires both Analytic and Clinical Validity

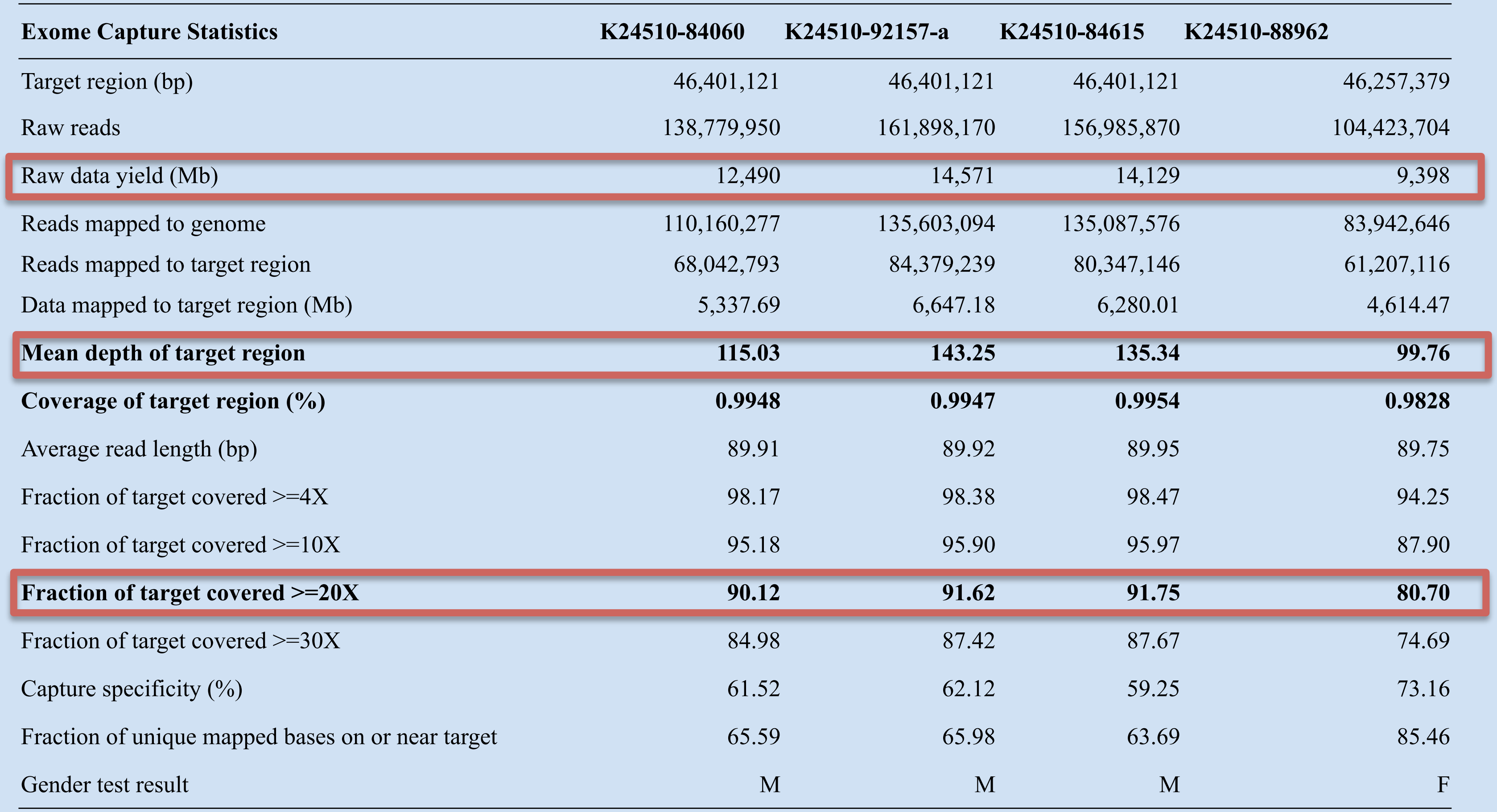
- Analytical Validity: the test is accurate with high sensitivity and specificity.

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

Optimizing Variant Calling in Exomes at BGI in 2011

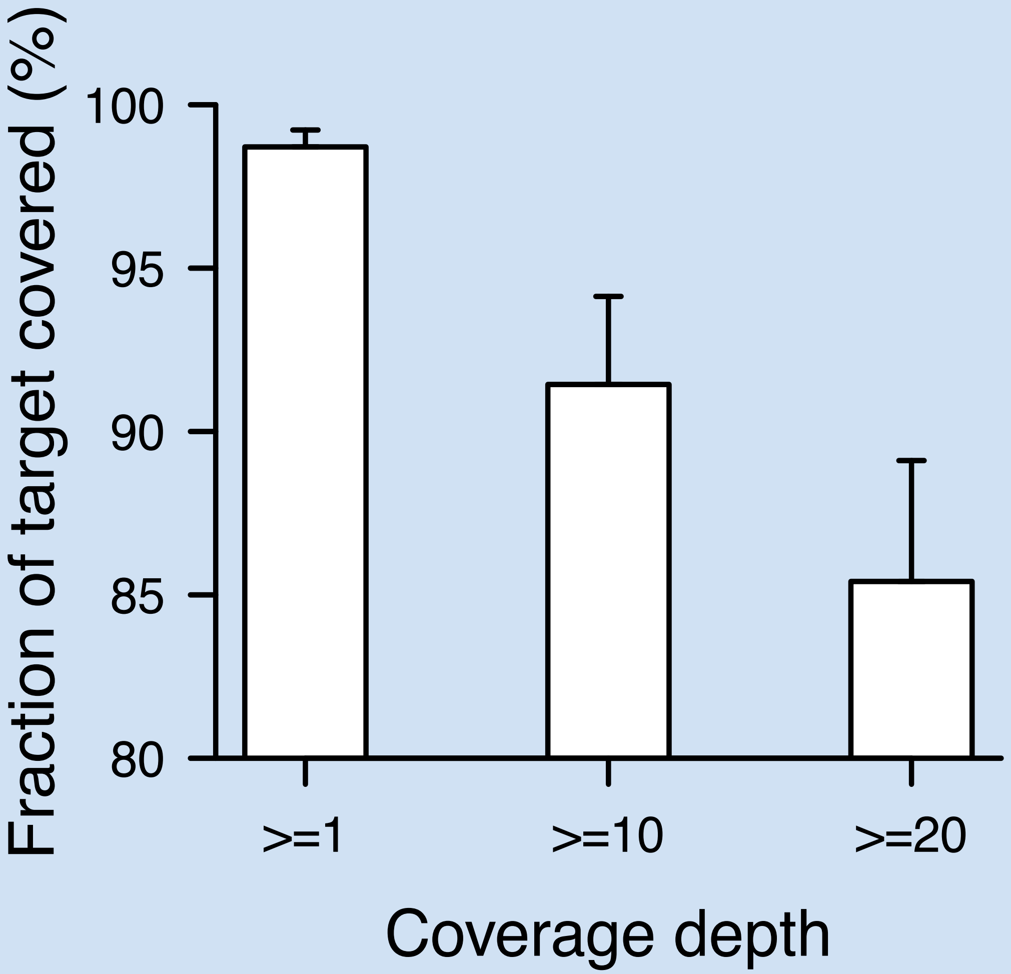
- Agilent v2 44 MB exome kit
- Illumina Hi-Seq for sequencing.
- Average coverage ~100-150x.
- Depth of sequencing of >80% of the target region with >20 reads or more per base pair.
- Comparing various pipelines for alignment and variant-calling.

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

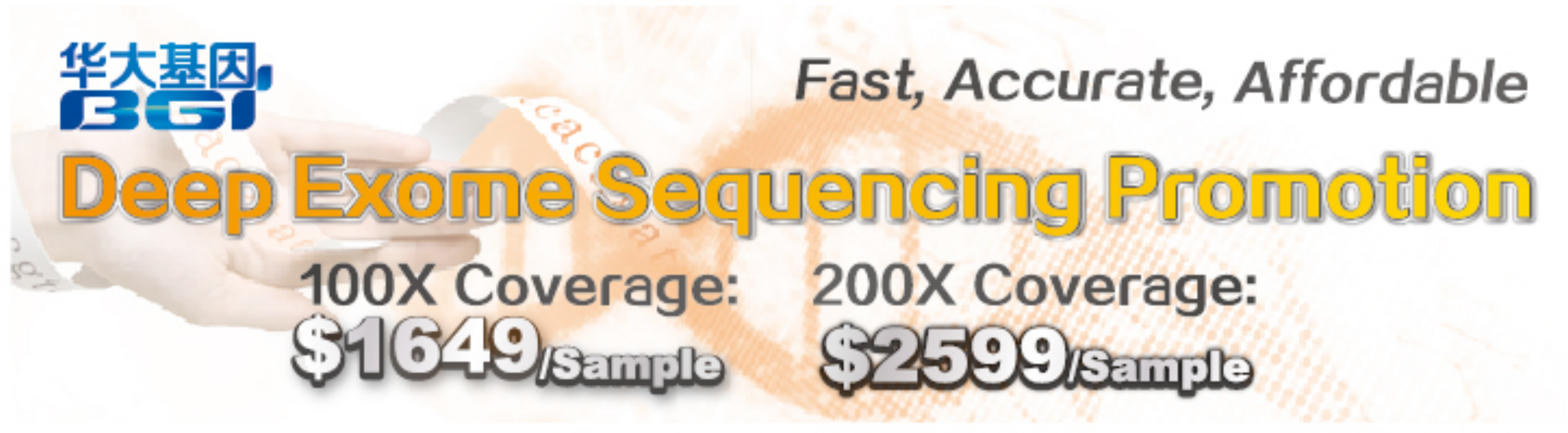


Exome Capture Statistics	**K24510-84060**	**K24510-92157-a**	**K24510-84615**	**K24510-88962**
Target region (bp)	46,401,121	46,401,121	46,401,121	46,257,379
Raw reads	138,779,950	161,898,170	156,985,870	104,423,704
Raw data yield (Mb)	12,490	14,571	14,129	9,398
Reads mapped to genome	110,160,277	135,603,094	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



BGI appears to have followed the lead of the other major genome sequencing centers (Broad, WashU and Baylor) and embraced “Deep Exomes” at this point.



华大基因 BGI

Fast, Accurate, Affordable

Deep Exome Sequencing Promotion

100X Coverage: $1649/Sample

200X Coverage: $2599/Sample

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.

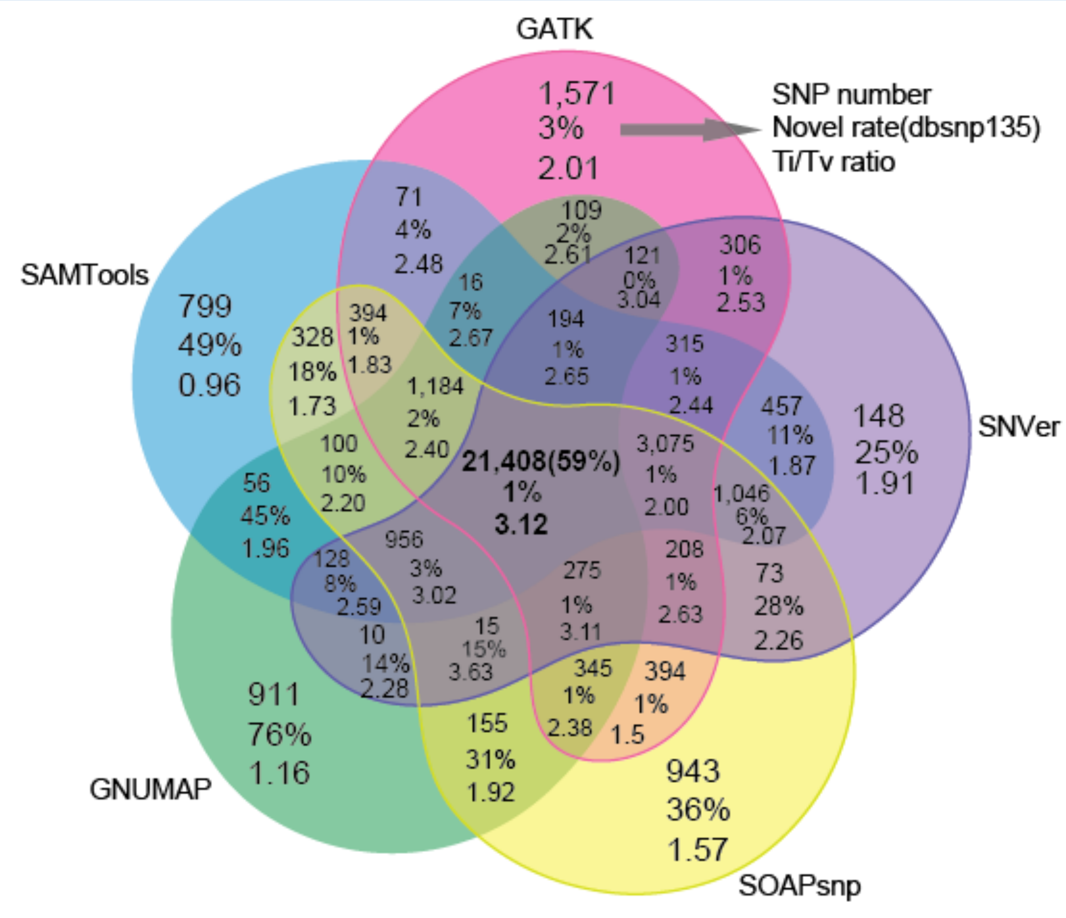
3) **SOAP**-Align – SOAPsnp – and BWA-SOAPindel (adopts local assembly based on an extended de Bruijn graph)

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

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

Total SNVs

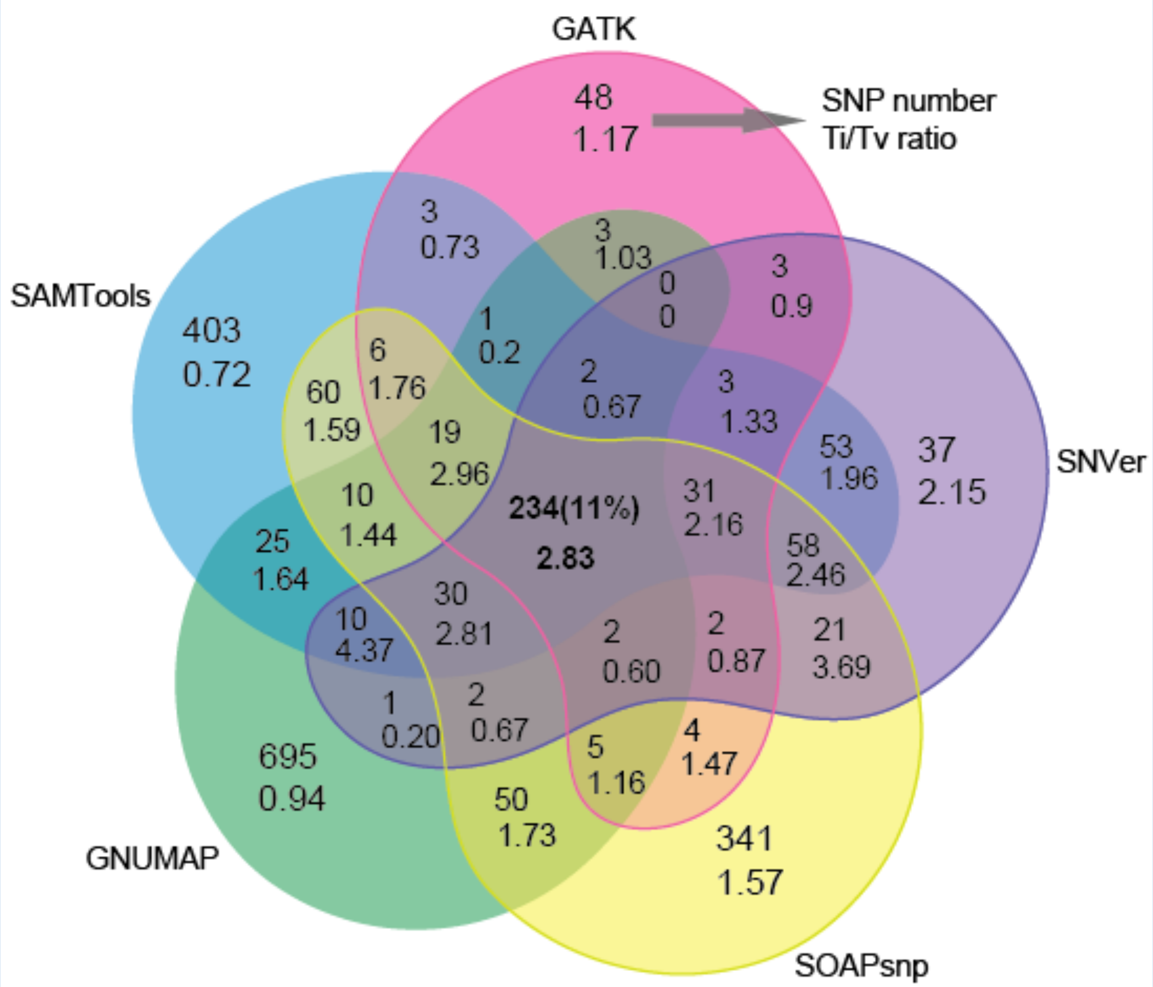
A)



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

Novel SNVs

C)

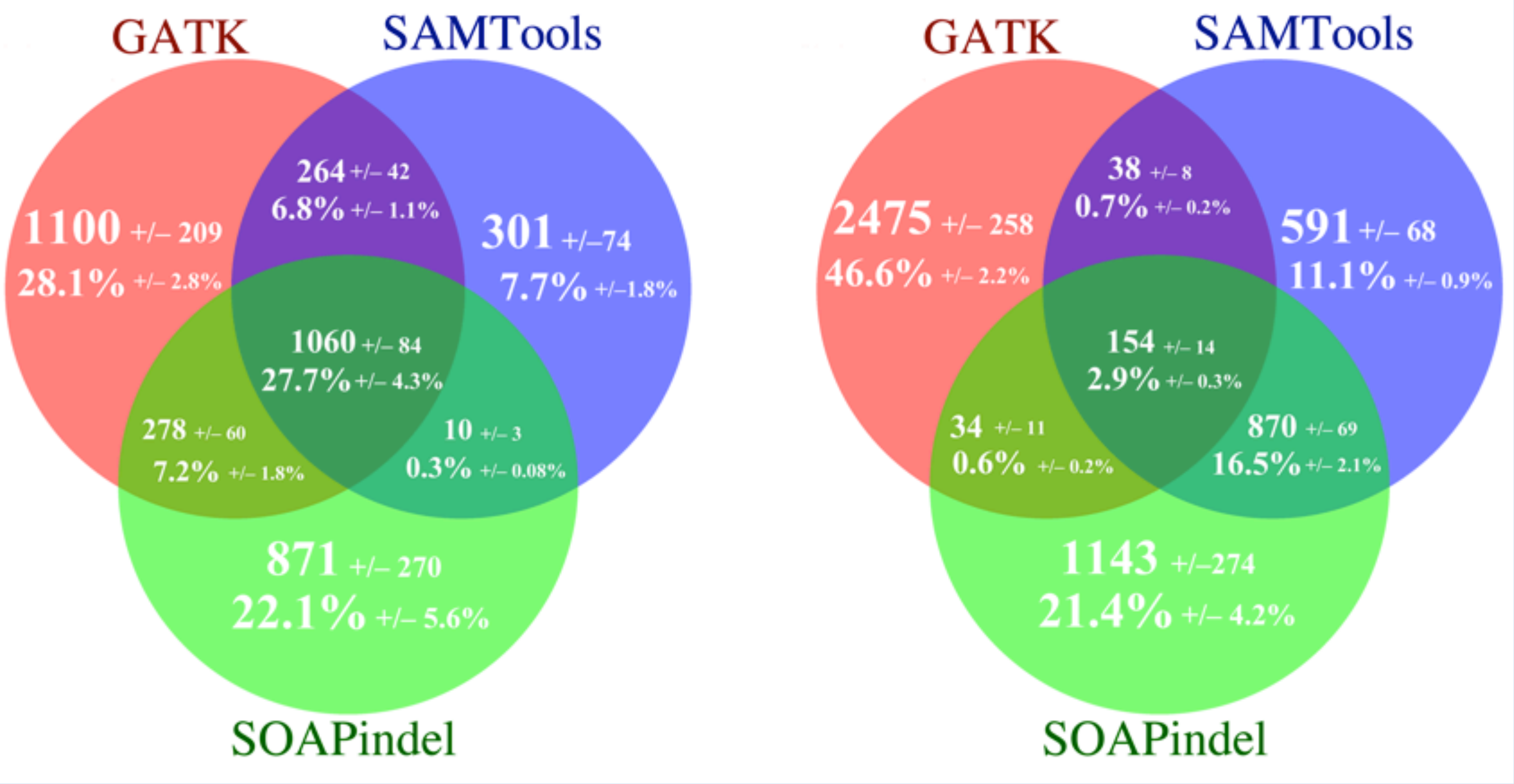


- **C)** Mean # of novel SNVs (not present in dbSNP135) found by 5 pipelines across 15 exomes. The percentage in the center of the Venn diagram is the percent of novel SNVs called by all five pipelines.

INDELS

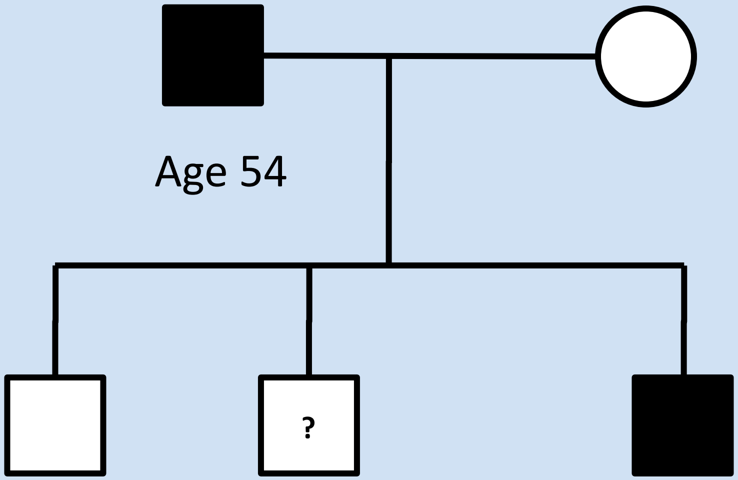
Indels- Overlap by Base Position only

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



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

Another Pedigree –K8101



Age 54

Age 25

Age 24

Age 19
prodromal, likely bipolar

Collected 35 DNA samples from the extended family, due to very large excess of major depression, bipolar, Tourette and OCD.

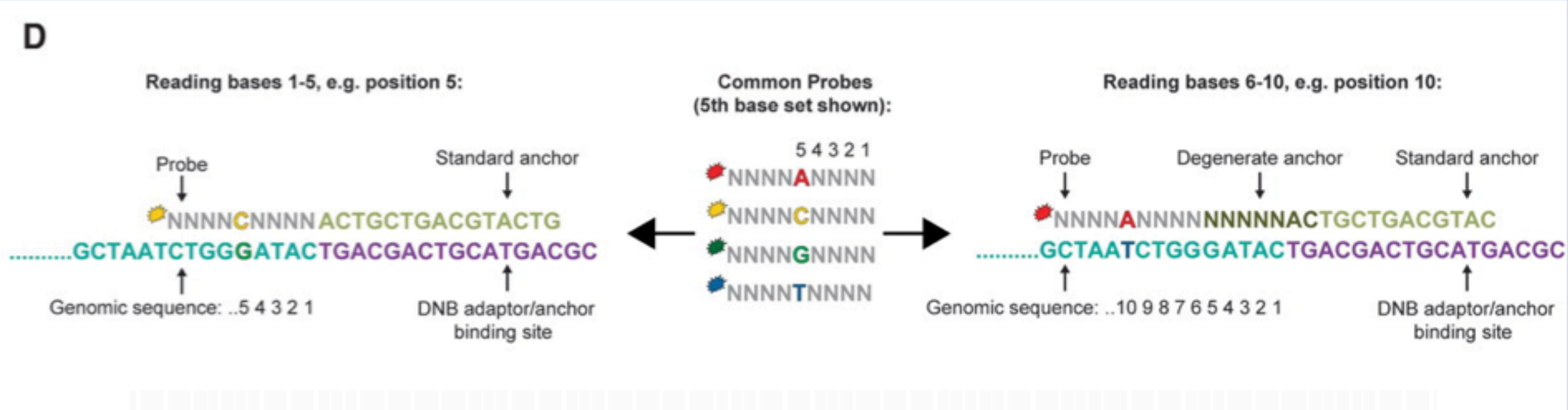
Case Presentation

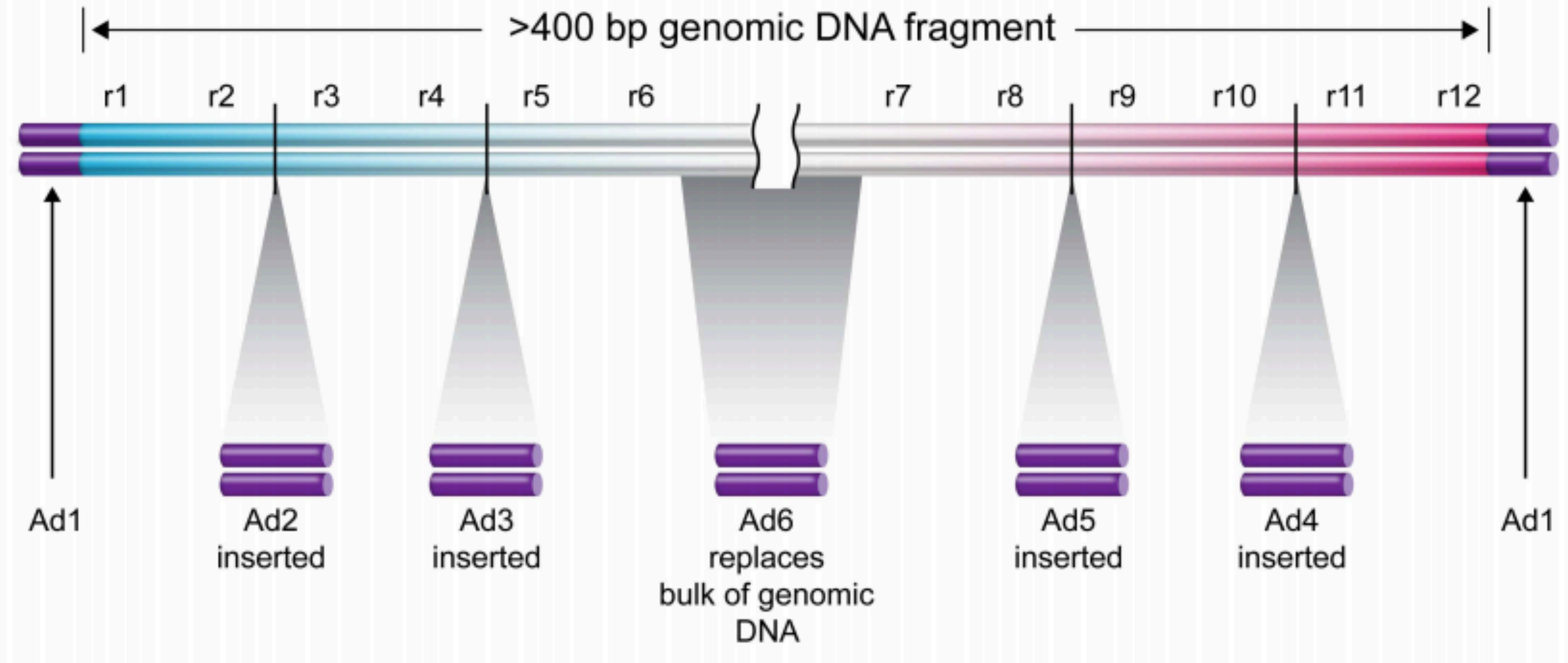
- Male, age 55 currently.
- Psychotic break at age 20 with bipolar features.
- Evolution into schizoaffective disorder over next 25 years.
- Also with severe obsessive compulsive disorder and severe Tourette Syndrome
- At least two very severe suicide attempts at age 22, including throwing self under a truck one time and then driving head-on into another car (with death of two passengers in other car, found not guilty by reason of insanity).

- Extensive medication trials over many years, along with anterior capsulotomy with very little effect for the OCD.

- Current meds:
  - Klonopin
  - Nicotinamide
  - Lunesta
  - Ativan
  - Lithium
  - Seroquel
  - Lamictal
  - Luvox

Complete Genomics chemistry - combinatorial probe anchor ligation (cPAL)


D
Reading bases 1-5, e.g. position 5:
Probe
Standard anchor
NNNNCNNNNACTGCTGACGTACTG
..........GCTAATCTGGGATACTGACGACTGCATGACGC
Genomic sequence: ..5 4 3 2 1
DNB adaptor/anchor binding site
Common Probes (5th base set shown):
5 4 3 2 1
NNNNANNNN
NNNNCNNNN
NNNNGNNNN
NNNNTNNNN
Reading bases 6-10, e.g. position 10:
Probe
Degenerate anchor
Standard anchor
NNNNANNNNNNNNNACTGCTGACGTAC
.........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

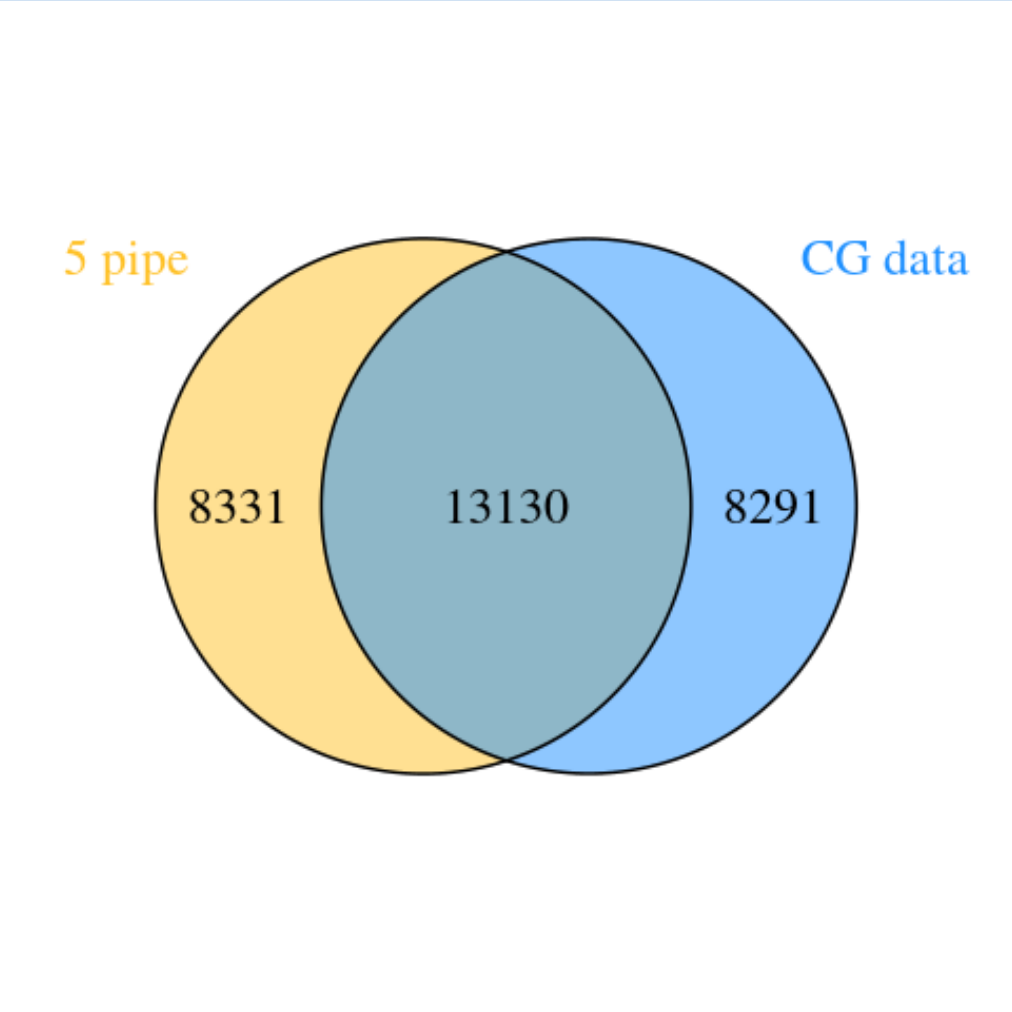
	FALSE POSITIVES	EST FPs	FALSE NEGATIVES	TOTAL DISCORDANCES	CONCORDANCE
Discordant SNVs per called MB	1.56 x 10-6	4,450	1.67 x 10-6	3.23 x 10-6	99.9997% of bases

Table 2. *Concordance of Technical Replicates.*

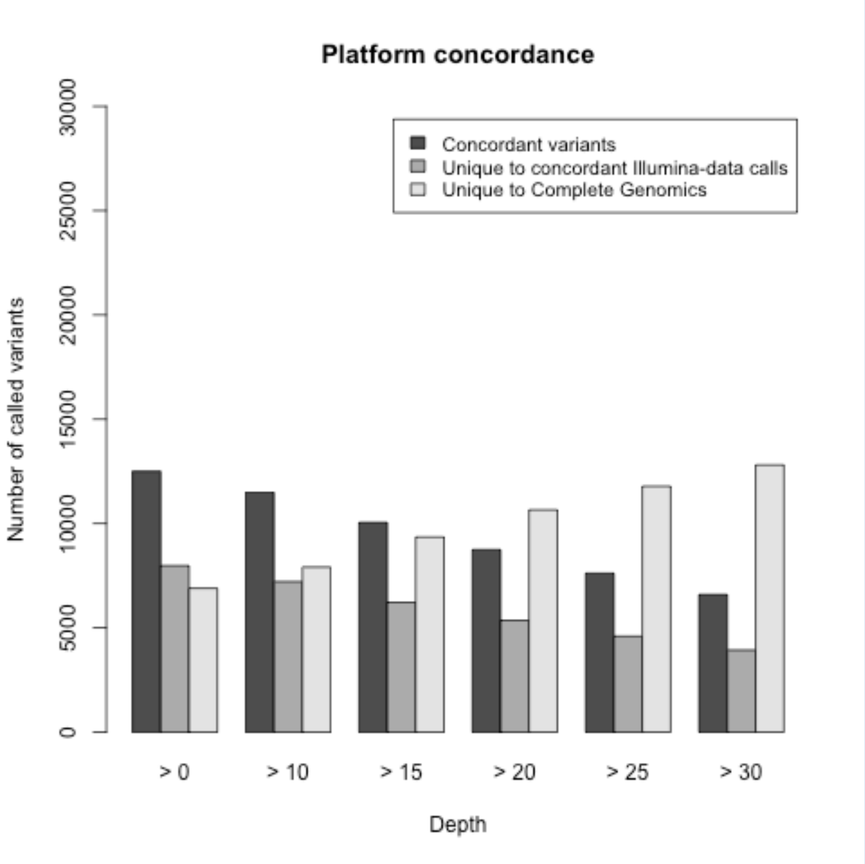
COMPLETE GENOMICS CALL	OTHER PLATFORM	PLATFORM-SPECIFIC SNVs	VALIDATION RATE	EST FPs	FPR
Het or Hom SNV	No SNV Reported	99K	17/18 = 94.4%	5,577	0.16%
No-call or Hom-Ref	SNV Reported	345K	2/15 = 13.3%	299,115	8.2%

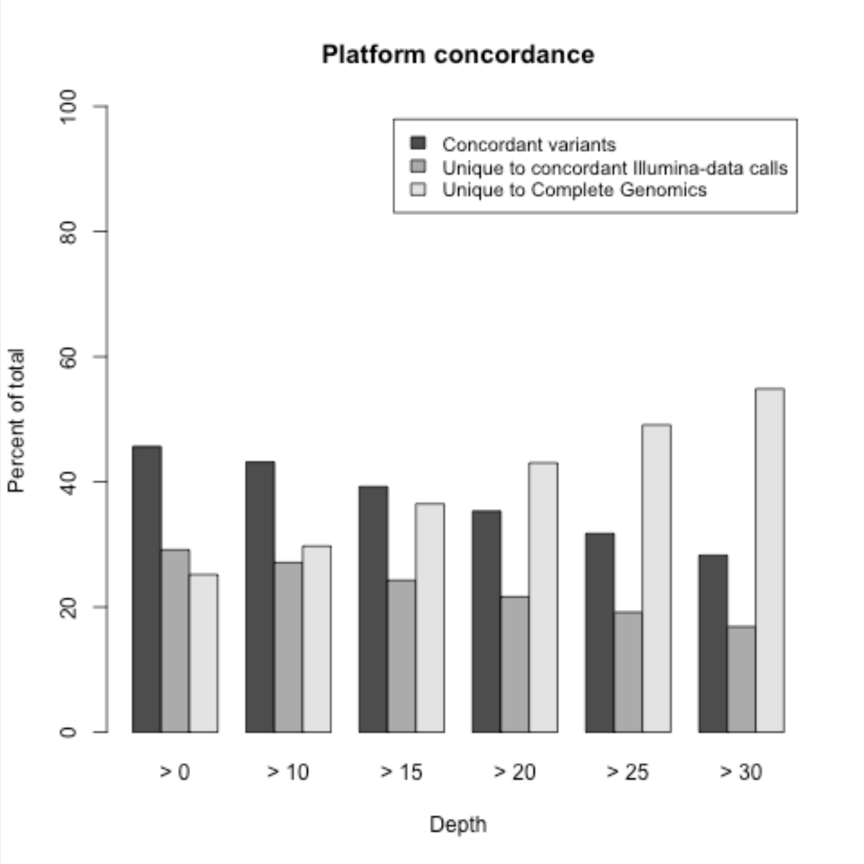
Table 3. *False Positive Rate.*

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

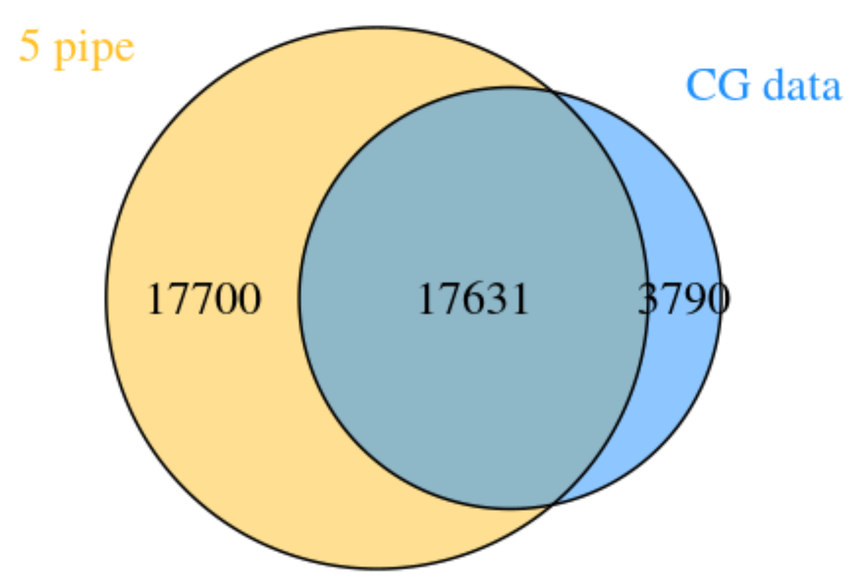


Taking SNVs concordant in 5 Illumina pipelines as per READ DEPTH, and comparing to SNVs in Complete Genomics Data from same sample

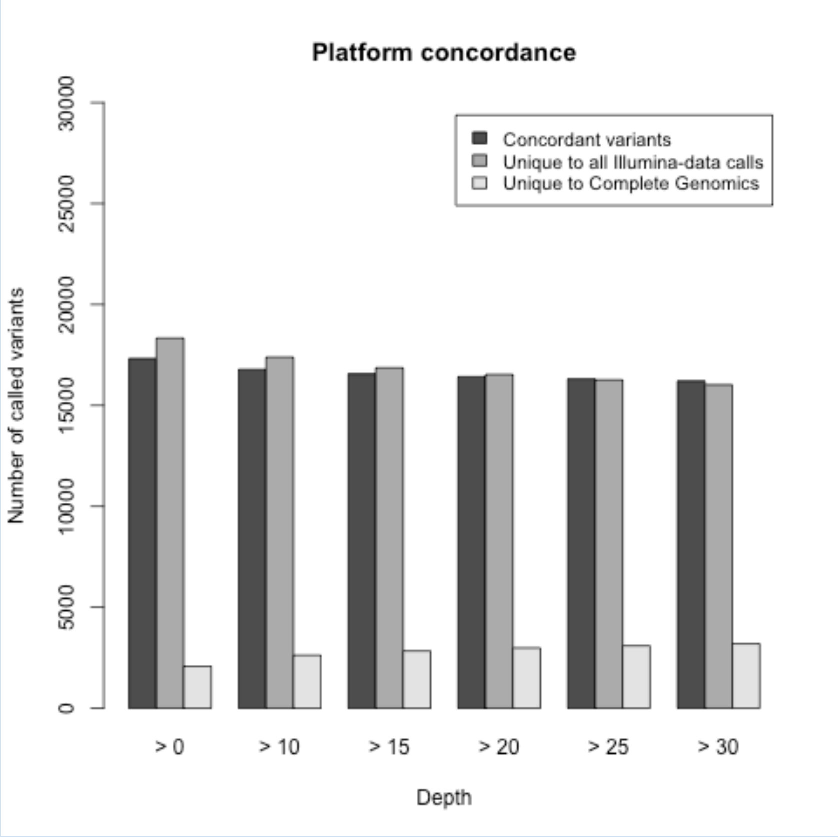


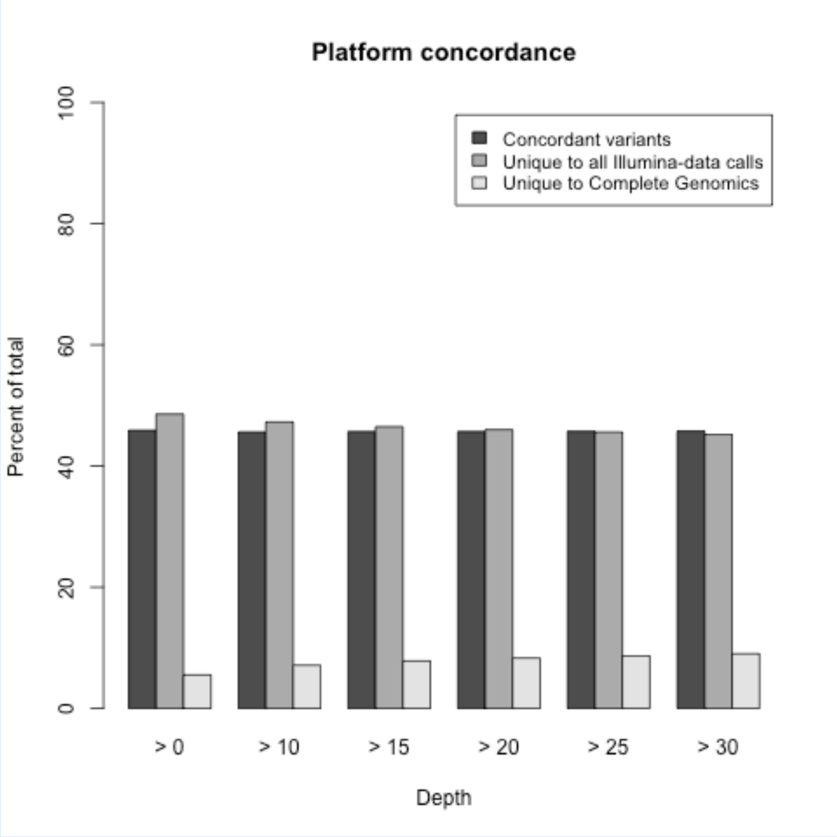


Taking SNVs found by ALL 5 Illumina pipelines (Union), and comparing to SNVs in Complete Genomics Data from same sample

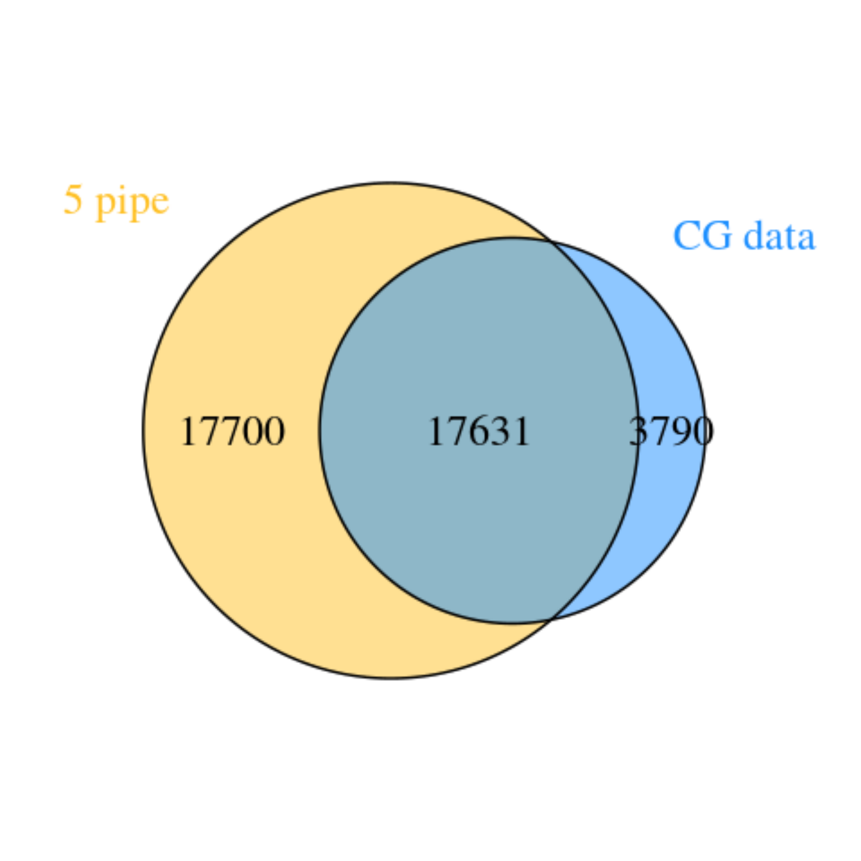


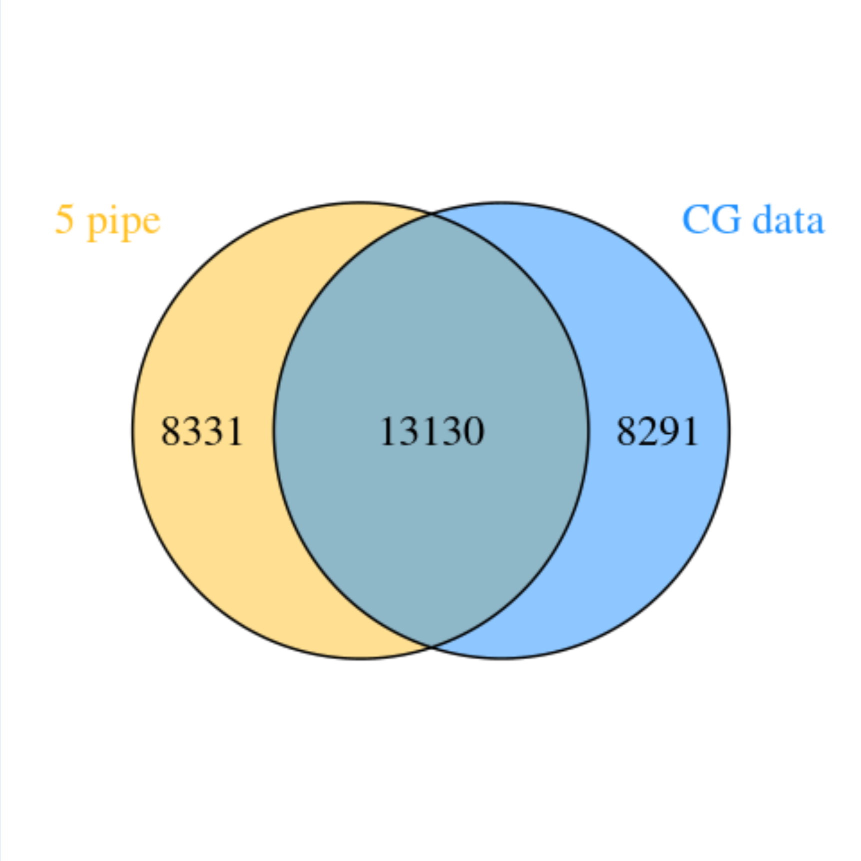
Taking the UNION of all SNVs called by Illumina pipelines, as per READ DEPTH, and comparing to SNVs in Complete Genomics Data from same sample





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

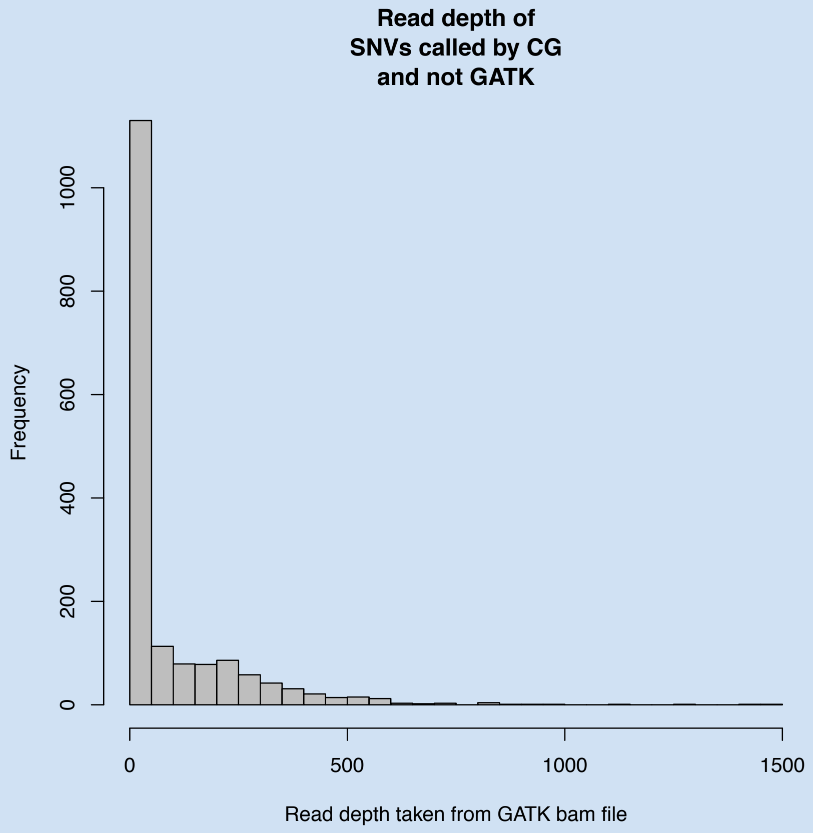


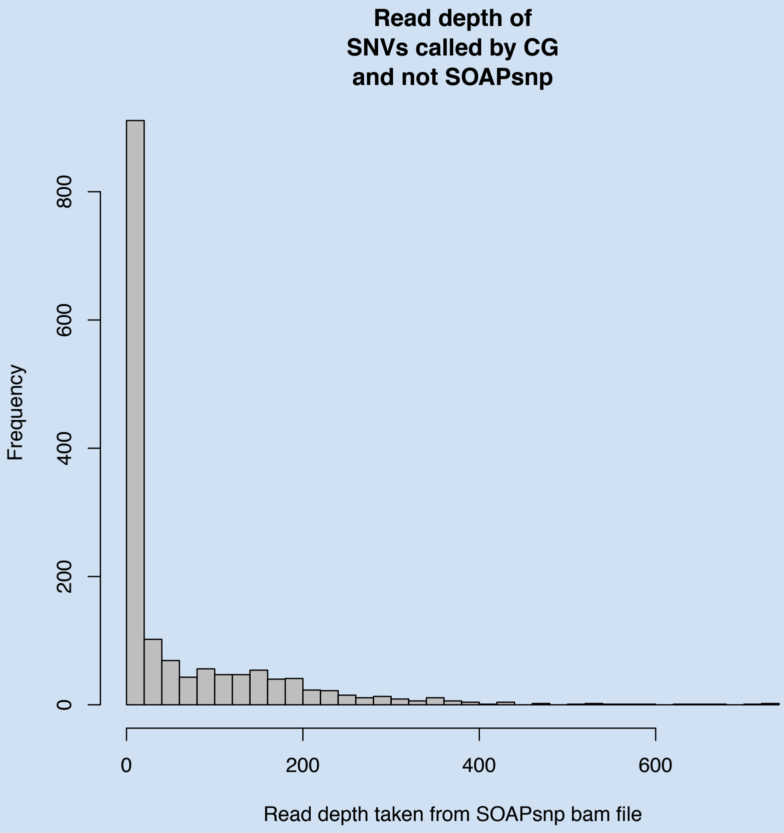


Union of Illumina variants

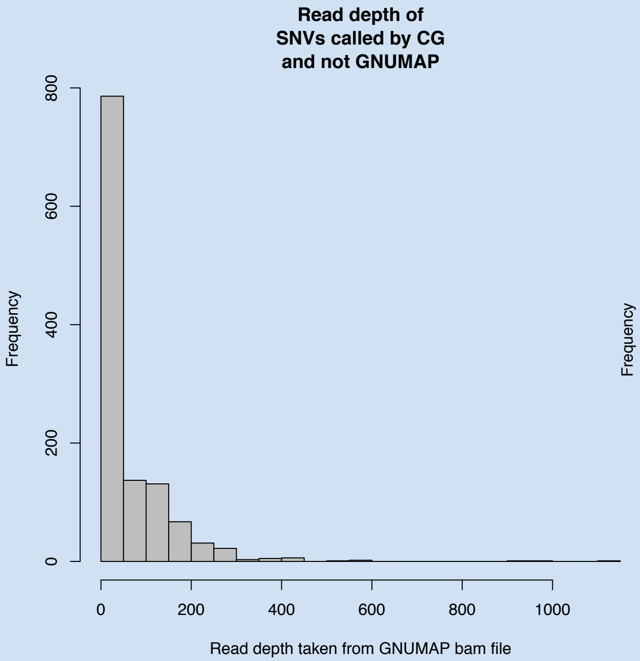
Concordant Illumina variants

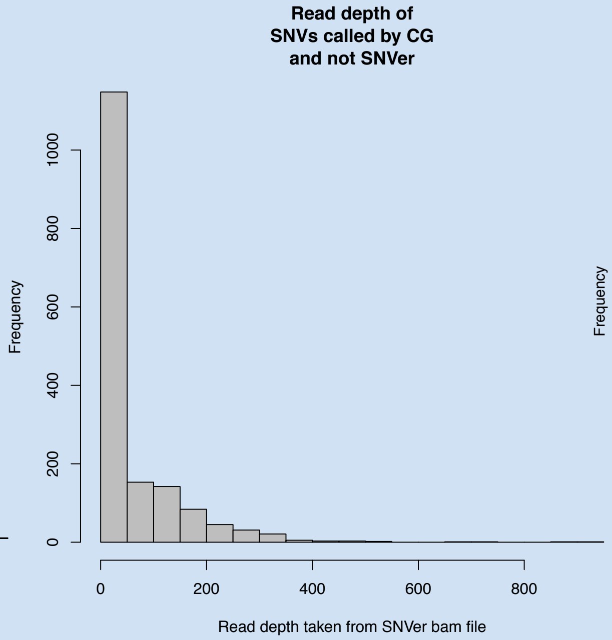
Read Depth of Illumina Reads for variants called by Complete Genomics but NOT by GATK or SOAP pipelines

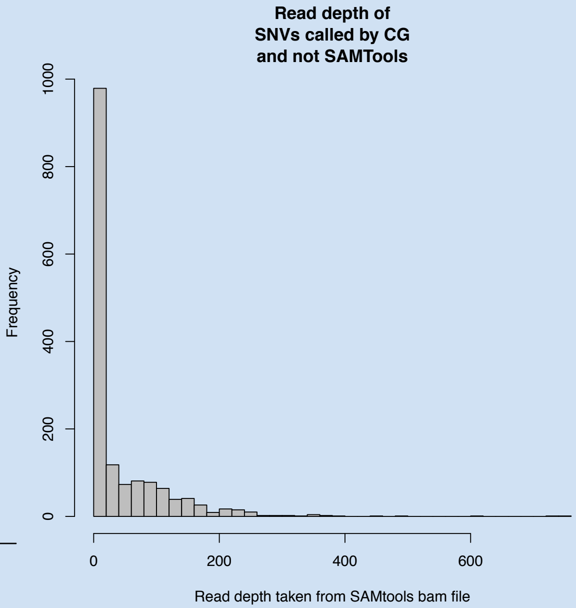


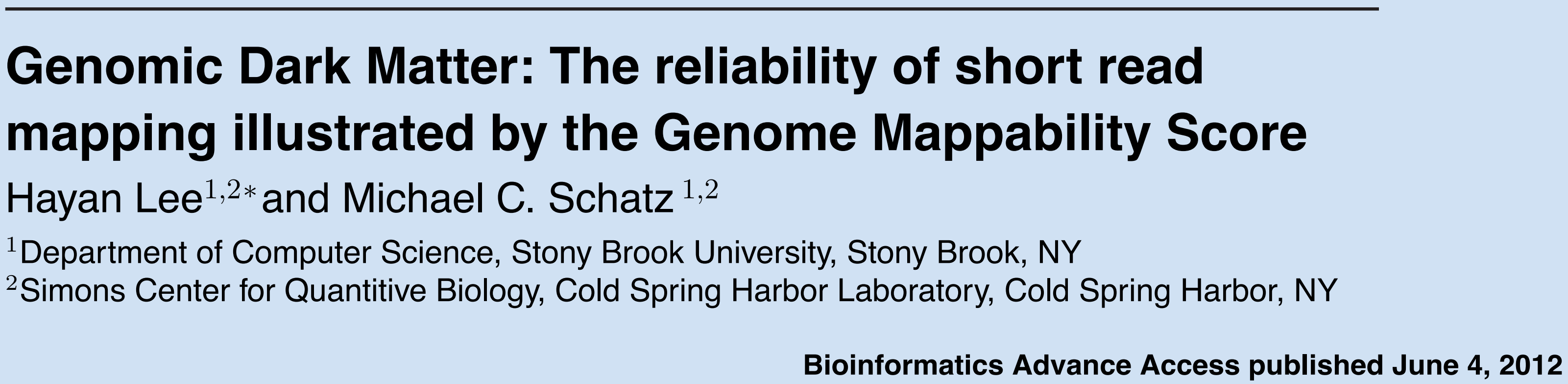


Read Depth of Illumina Reads for variants called by Complete Genomics but NOT by GNUMAP, SNVer or SamTools pipelines









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

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

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

Bioinformatics Advance Access published June 4, 2012

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

Genomic Dark Matter, cont….

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

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.

Acknowledgments

THE UNIVERSITY OF UTAH

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

UTAH FOUNDATION FOR BIOMEDICAL RESEARCH

Reid Robison
Edwin Nyambi

USC TROJANS

Kai Wang

CH

Zhi Wei
Lifeng Tian
Hakon Hakonarson

our study families

UNIVERSITAS BERGENSIS

Thomas Arnesen
Rune Evjenth
Johan R. Lillehaug

CSH STANLEY INSTITUTE FOR COGNITIVE GENOMICS
COLD SPRING HARBOR LABORATORY

Jason O'Rawe
Michael Schatz
Giuseppe Narzisi

华大基因
BGI

Tao Jiang
Guangqing Sun
Jun Wang

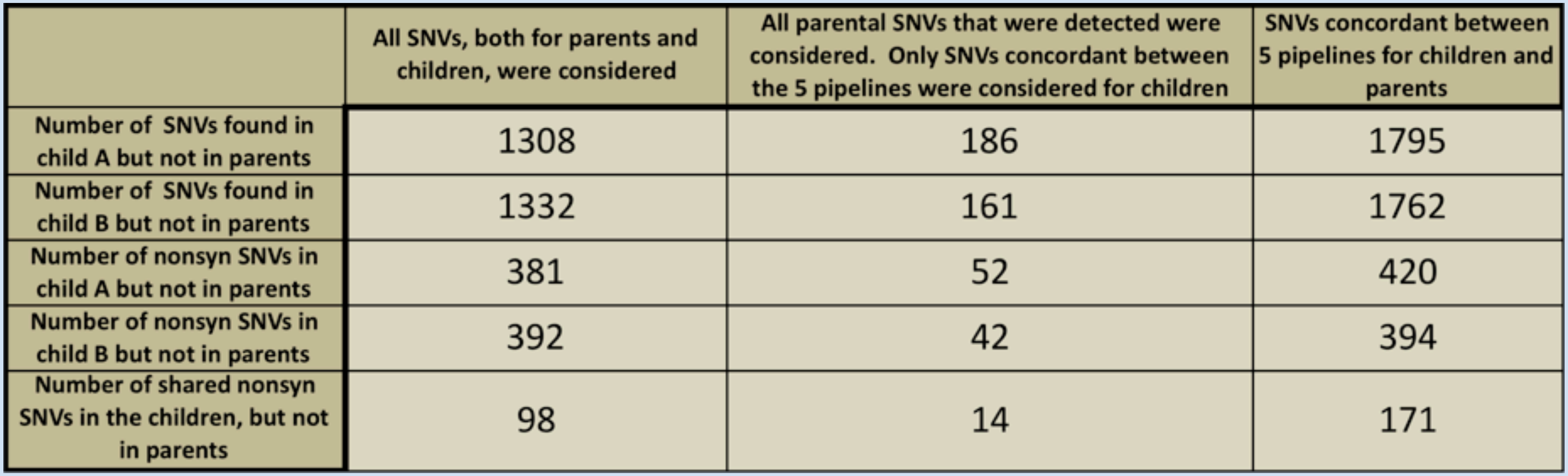
Extra Slides Not Covered in Talk

Optimizing the Variant Calling Pipeline Using Family Relationships

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

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

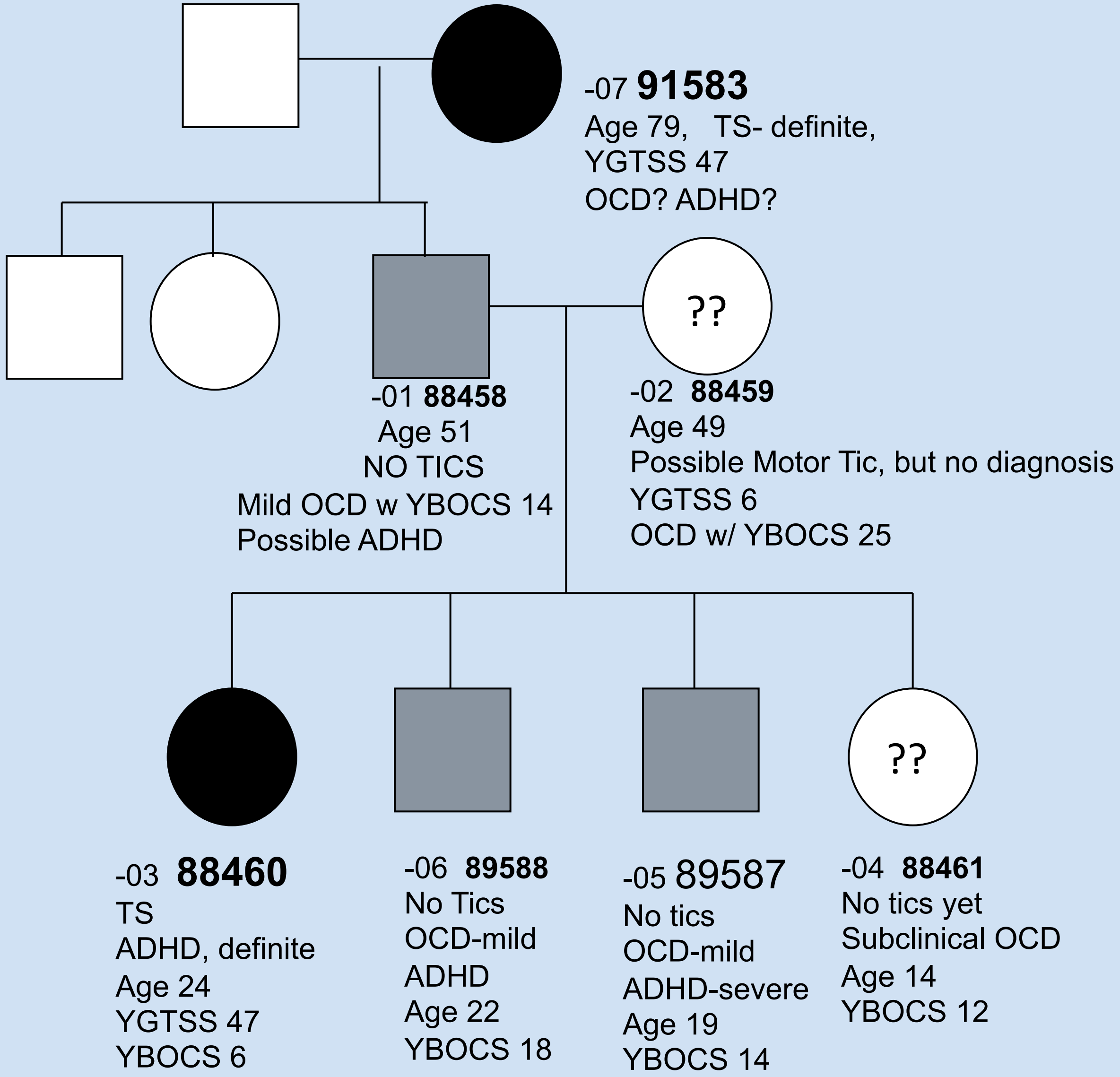
Optimizing pipeline based on literature value of ~1 true de novo protein-altering mutation per exome



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

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

TDT- 09 -1018
K26679



-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

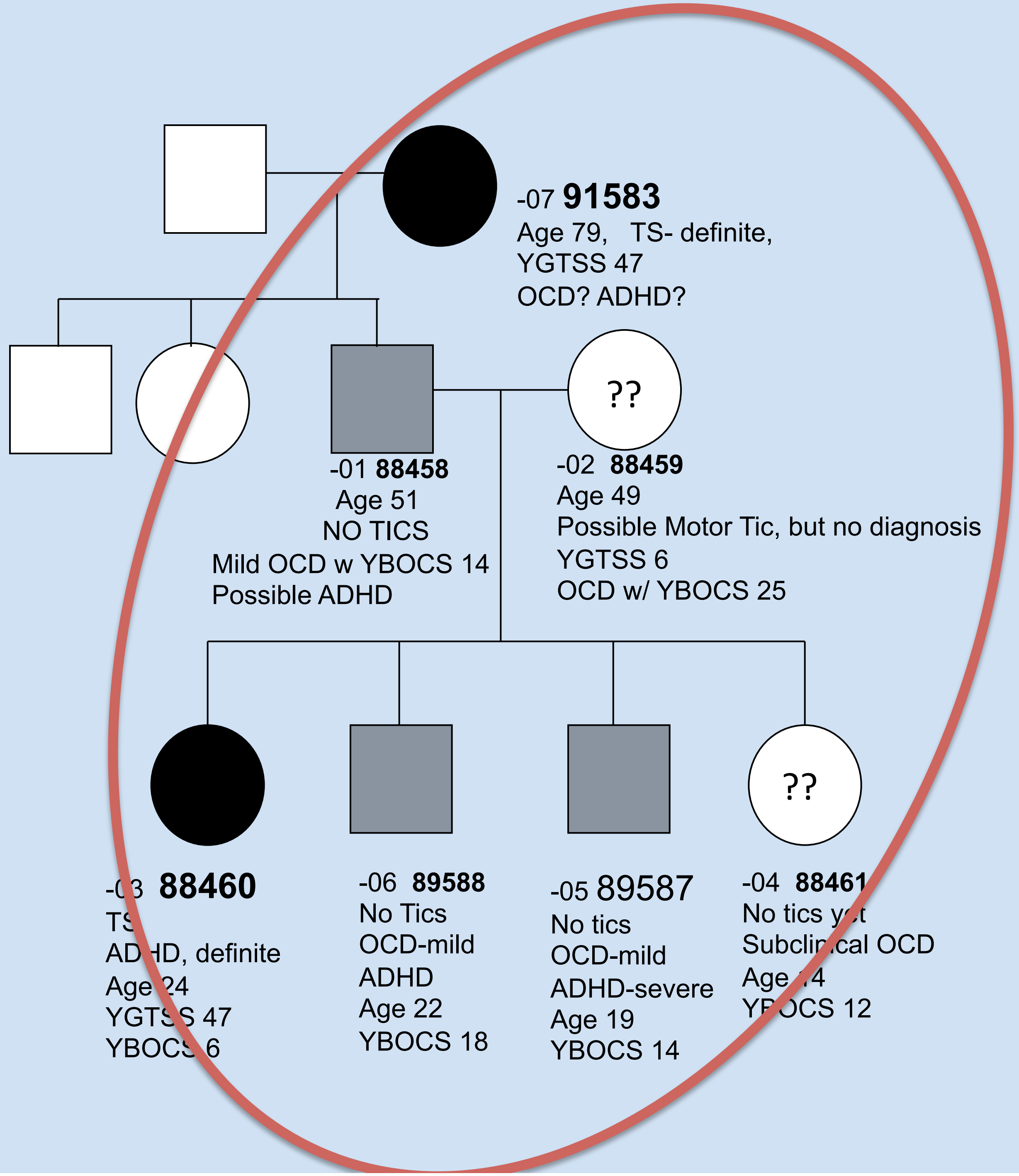
-03 **88460**
TS
ADHD, definite
Age 24
YGTSS 47
YBOCS 6

-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

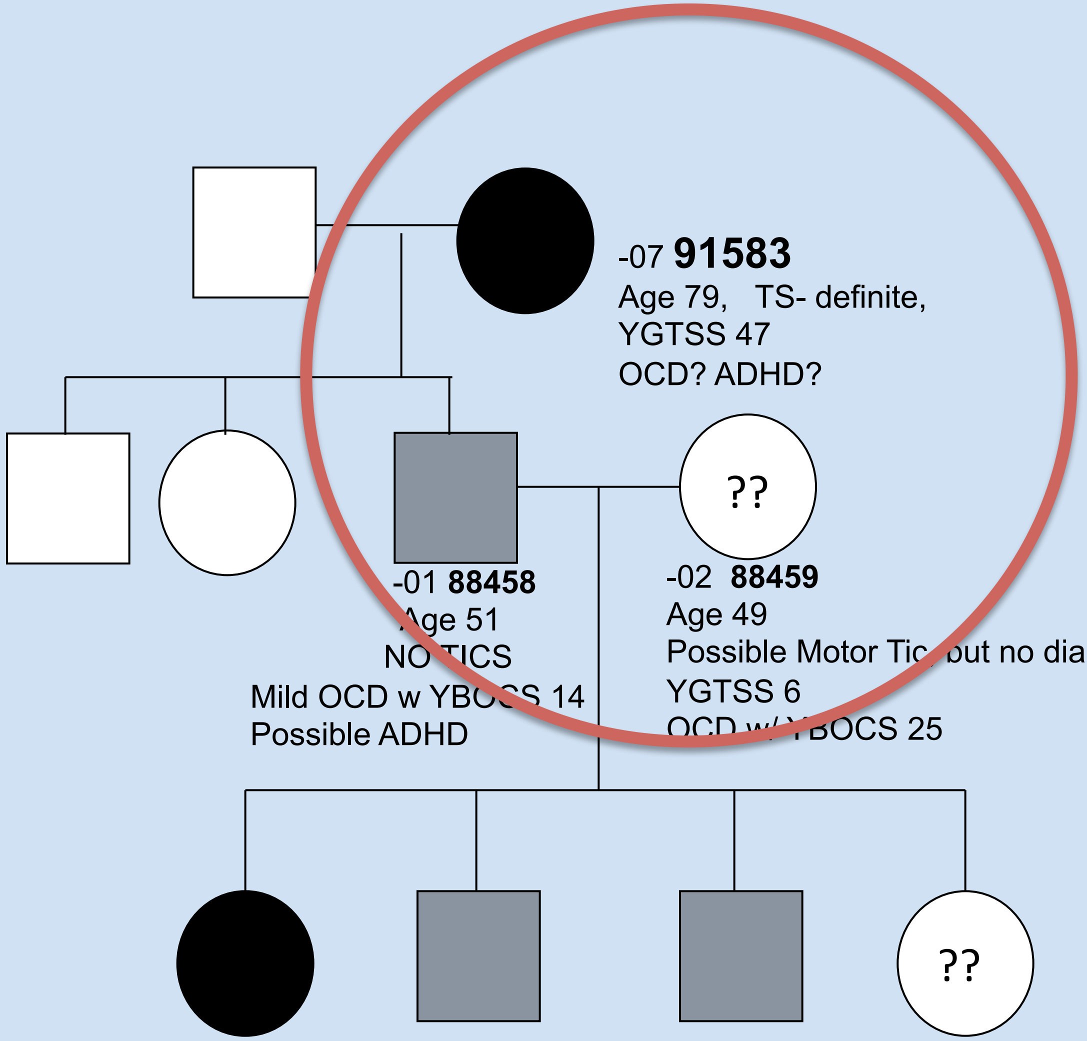
-03 **88460**
TS
ADHD, definite
Age 24
YGTSS 47
YBOCS 6

-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



"Parents"

-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

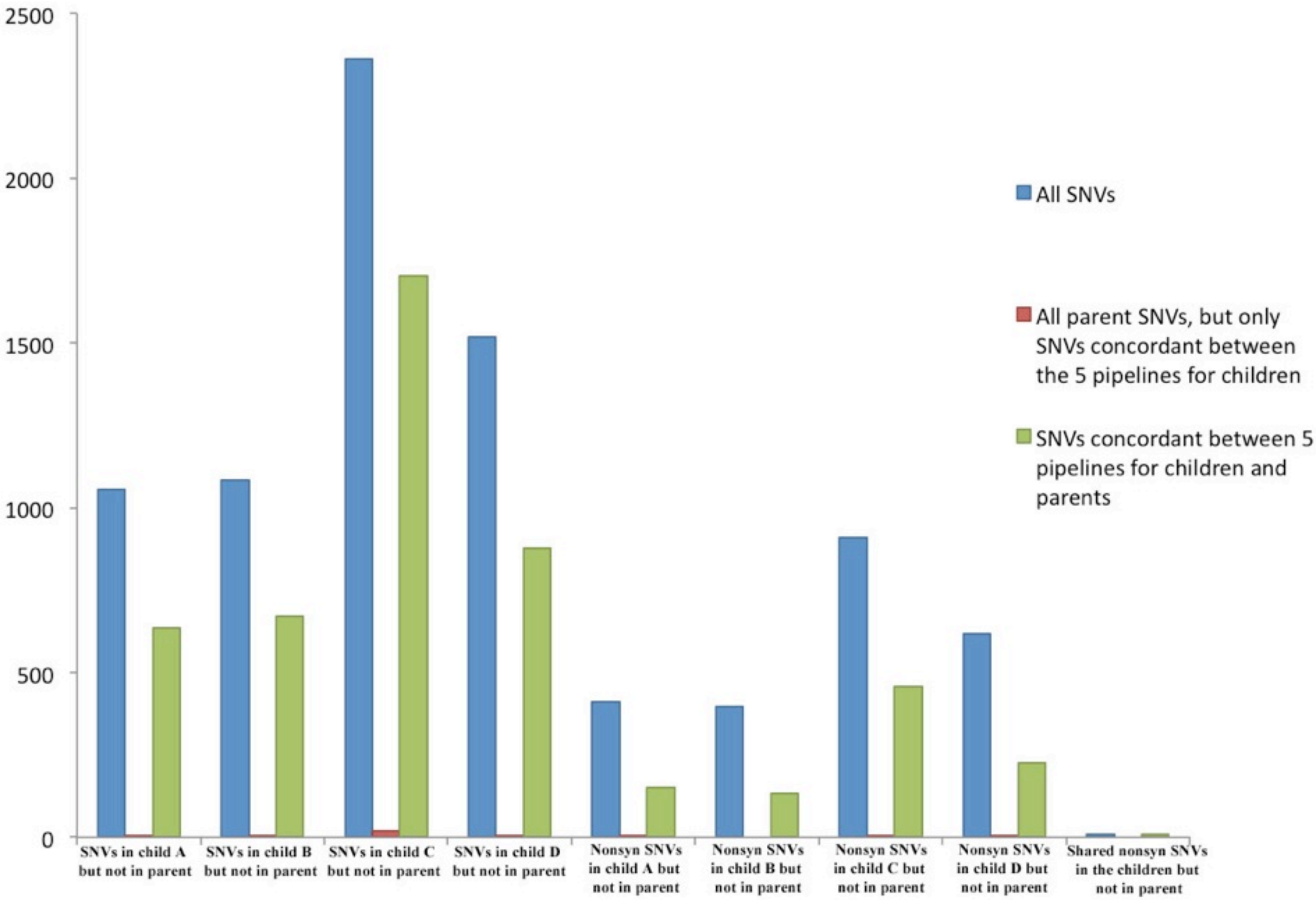
-05 89587
No tics
OCD-mild
ADHD-severe
Age 19
YBOCS 14

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

Analysis based on various pipelines

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

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


2500
2000
1500
1000
500
0
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
All SNVs
All parent SNVs, but only SNVs concordant between the 5 pipelines for children
SNVs concordant between 5 pipelines for children and parents

Preliminary Conclusions

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

VAAST shows that probabilistic ranking will be very useful going forward

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

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

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

Preliminary Conclusions

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

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

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