Challenges of Clinical Implementation of Genomic Medicine

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

CSH STANLEY INSTITUTE FOR COGNITIVE GENOMICS
COLD SPRING HARBOR LABORATORY

UTAH FOUNDATION FOR BIOMEDICAL RESEARCH

Punch Line

We need much more baseline whole genome sequencing in large pedigrees or clans to at least begin to understand genotype-phenotype correlations.

Ancestry Matters!

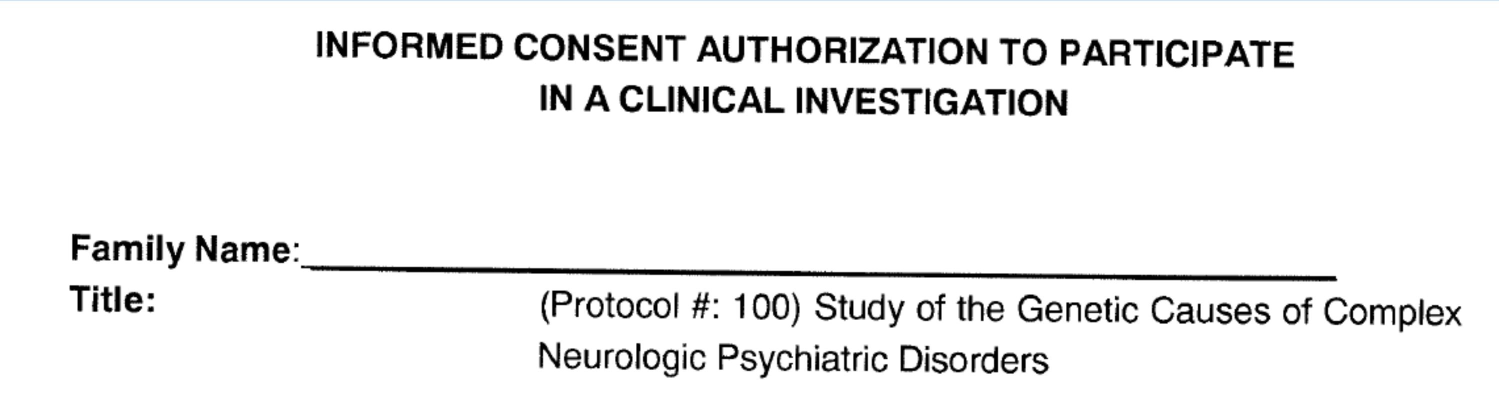
Penetrance and Expressivity

- We do not really know the penetrance or expressivity of pretty much ALL mutations in **humans**, as we have not systematically sequenced or karyotyped any genetic alteration in **MILLIONS** of well-phenotyped people.
- Do single mutations drive outcome predominately, or are the results modified substantially by other mutations and/or environment? Is there really such a thing as genetic determinism for **MANY** mutations?

THE UNIVERSITY OF UTAH

UTAH FOUNDATION FOR BIOMEDICAL RESEARCH

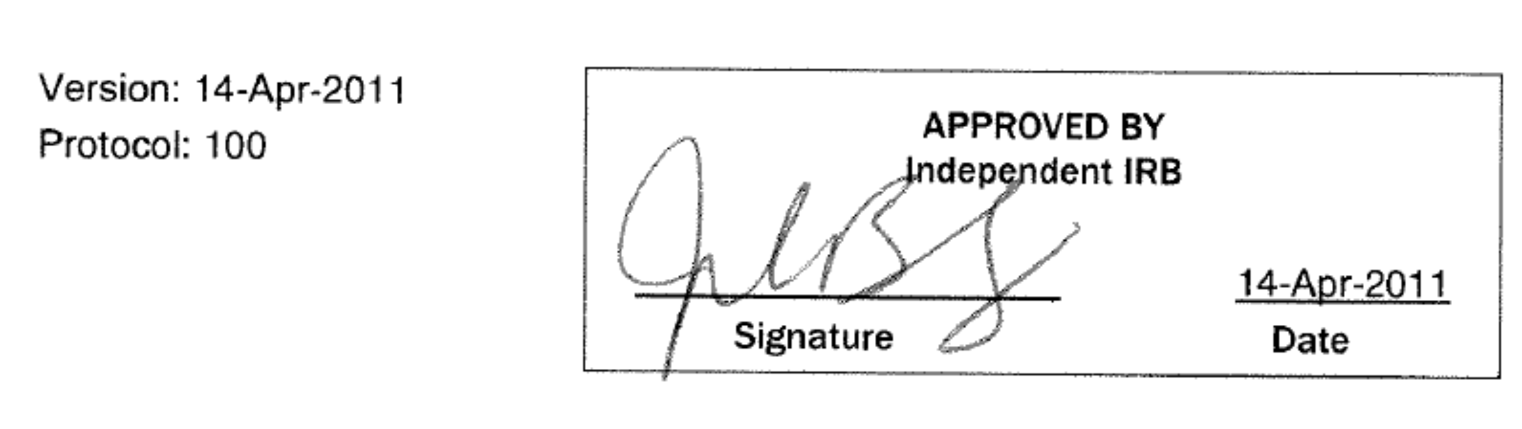
Utah Foundation for Biomedical Research



INFORMED CONSENT AUTHORIZATION TO PARTICIPATE
IN A CLINICAL INVESTIGATION

Family Name:____

Title: (Protocol #: 100) Study of the Genetic Causes of Complex Neurologic Psychiatric Disorders



Version: 14-Apr-2011
Protocol: 100

APPROVED BY
Independent IRB

____
Signature

14-Apr-2011
Date

Long-range Plans: ~750 DNA samples from many pedigrees with 455 of these genotyped thus far on Illumina 610K/2.5M arrays and 15 with high-depth exome, and 8 with CG whole genomes.

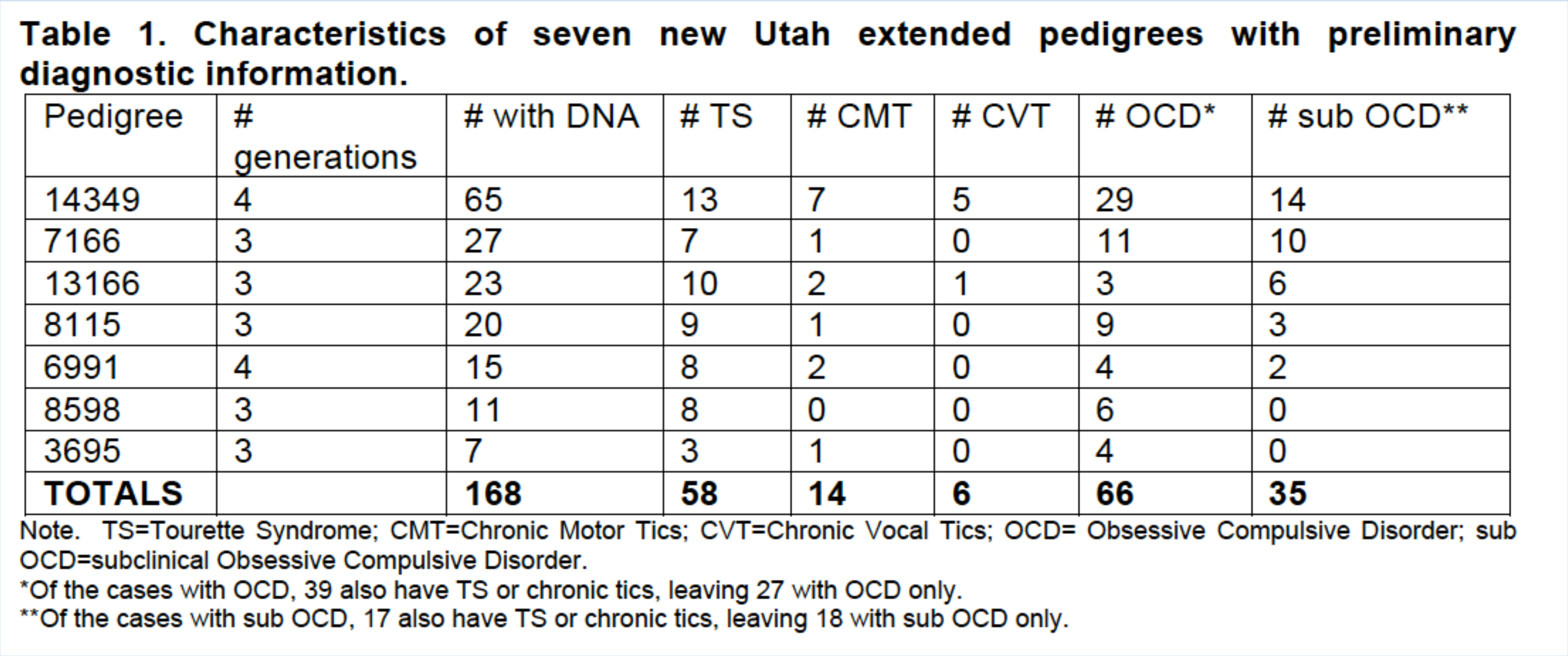


Table 1. Characteristics of seven new Utah extended pedigrees with preliminary diagnostic information.

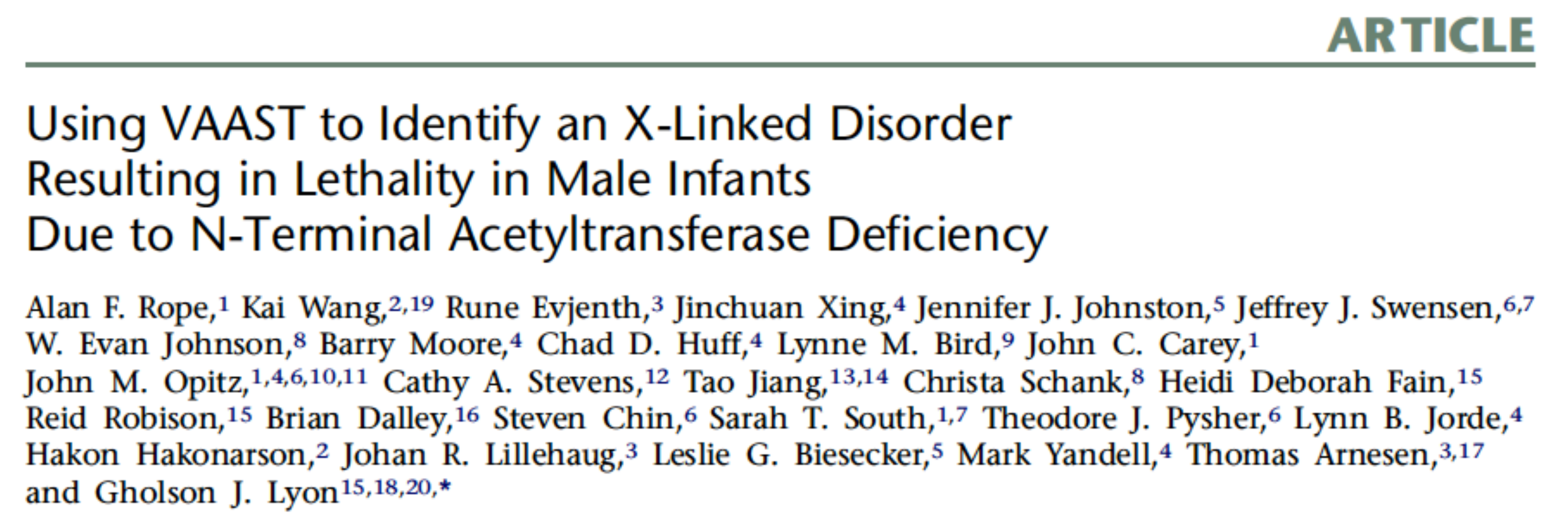
Pedigree	# generations	# with DNA	# TS	# CMT	# CVT	# OCD*	# sub OCD**
14349	4	65	13	7	5	29	14
7166	3	27	7	1	0	11	10
13166	3	23	10	2	1	3	6
8115	3	20	9	1	0	9	3
6991	4	15	8	2	0	4	2
8598	3	11	8	0	0	6	0
3695	3	7	3	1	0	4	0
TOTALS		**168**	**58**	**14**	**6**	**66**	**35**

Note. TS=Tourette Syndrome; CMT=Chronic Motor Tics; CVT=Chronic Vocal Tics; OCD= Obsessive Compulsive Disorder; sub OCD=subclinical Obsessive Compulsive Disorder.

*Of the cases with OCD, 39 also have TS or chronic tics, leaving 27 with OCD only.

**Of the cases with sub OCD, 17 also have TS or chronic tics, leaving 18 with sub OCD only.

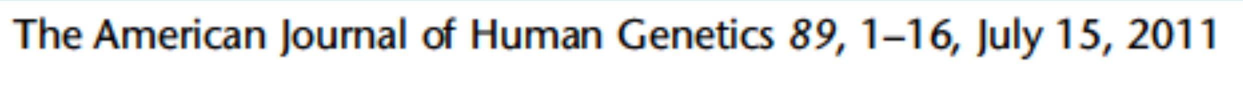
A new syndrome and its genetic basis.



ARTICLE

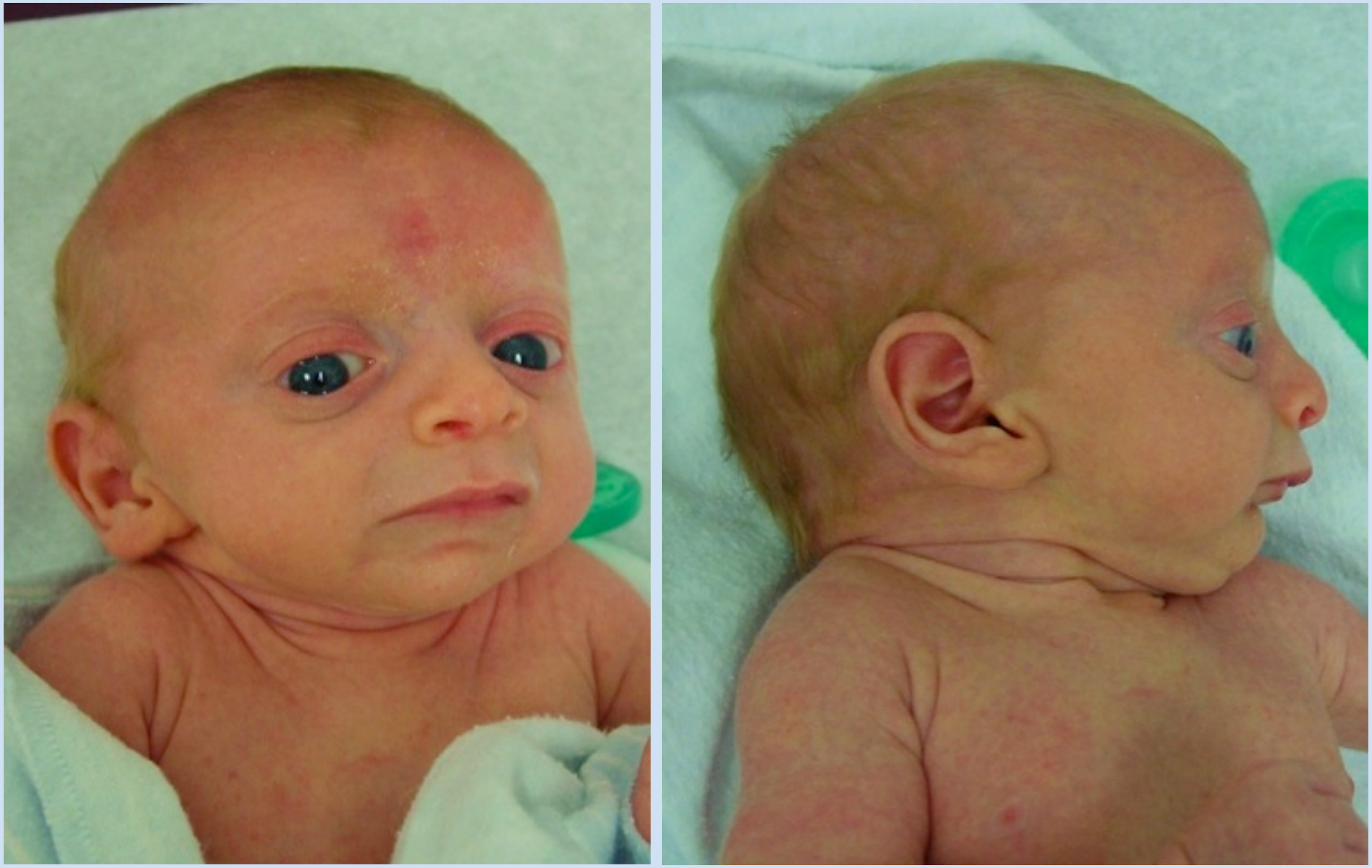
Using VAAST to Identify an X-Linked Disorder Resulting in Lethality in Male Infants Due to N-Terminal Acetyltransferase Deficiency

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,*]



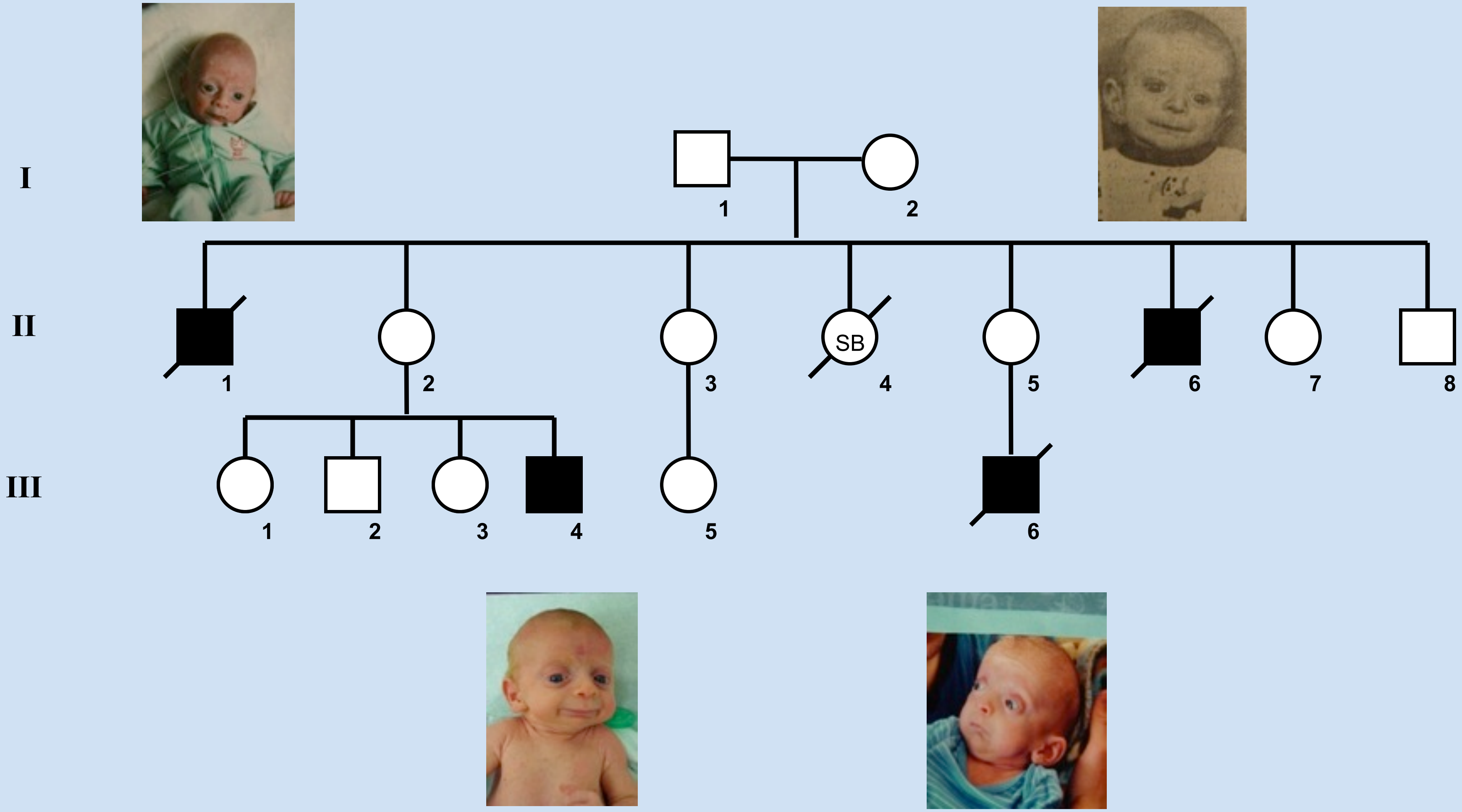
The American Journal of Human Genetics *89*, 1–16, July 15, 2011

This is the “Proband” photograph presented at Case Conference.



prominence of eyes, down-sloping palpebral fissures, thickened eyelids, large ears, beaking of nose, flared nares, hypoplastic nasal alae, short columella, protruding upper lip, micro-retrognathia

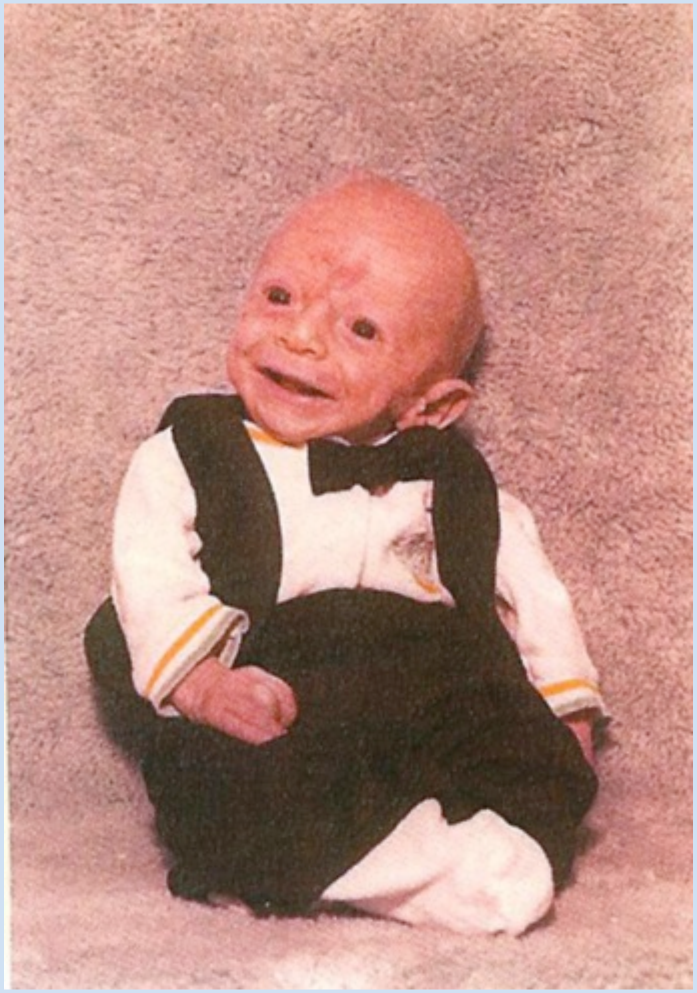
This is the family in Utah in December 2009.



I met the entire family on March 29, 2010

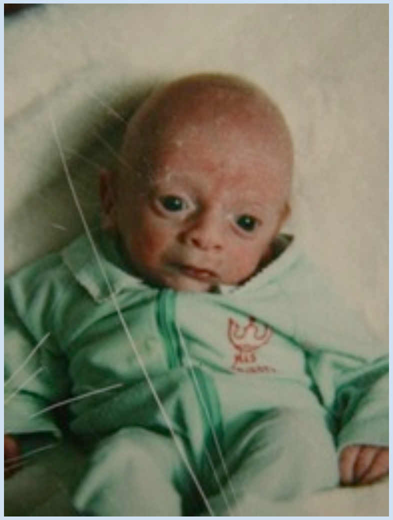
Photo of mother
with son in late 1970's

This is the first boy in the late 1970’s.

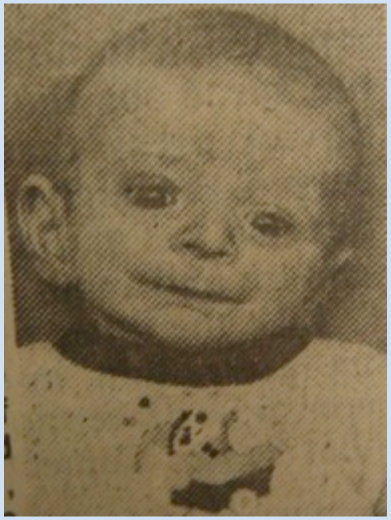


First boy. Called “a little old man” by the family. Died around ~1 year of age, from cardiac arrhythmias.

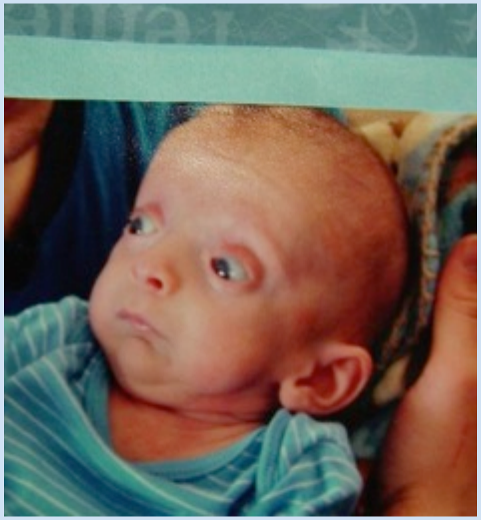
These are the Affected Boys of Family 1 in 2009.



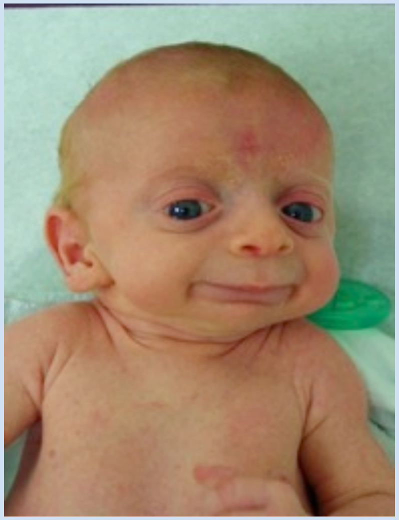
Uncle #1



Uncle #2



cousin



Proband- Sutter

Affected males had the consistent presentation of an aged appearance, a distinct and recognizable combination of craniofacial anomalies, post-natal growth failure, hypotonia, global developmental delays, cryptorchidism, arrhythmia, and eventual death from cardiac failure.

These are the Major Features of the Syndrome.

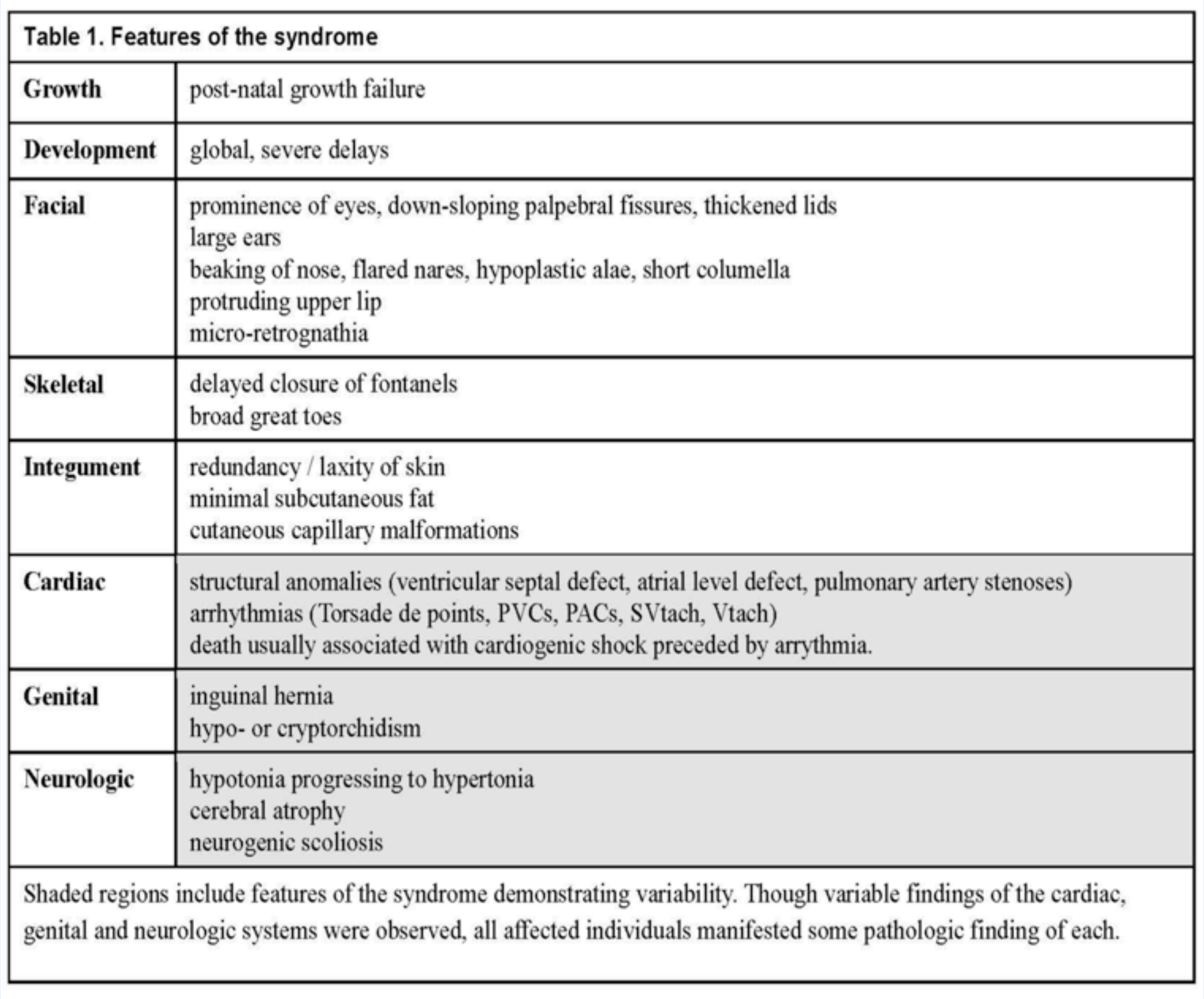
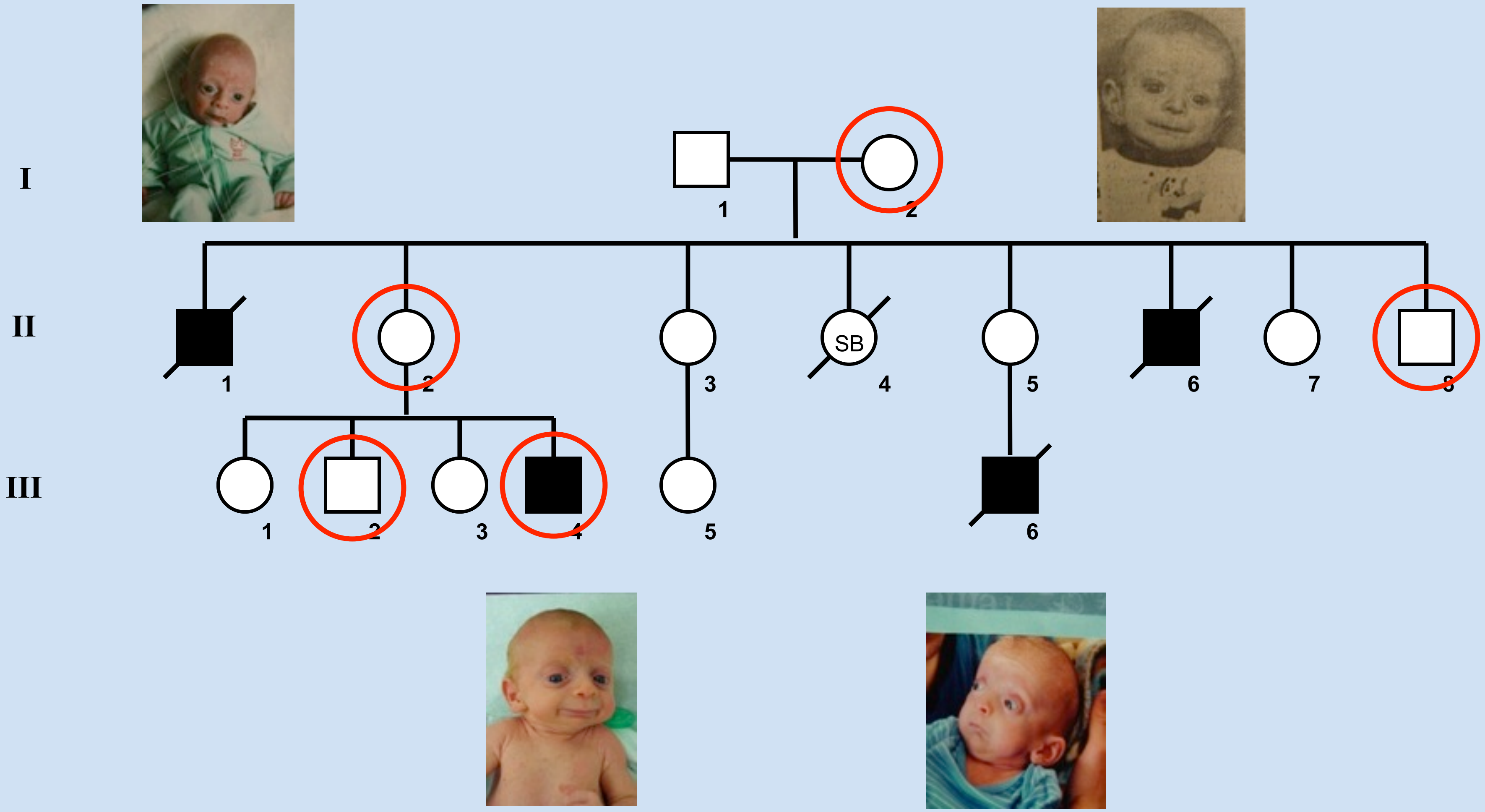


Table 1. Features of the syndrome

Growth	post-natal growth failure
Development	global, severe delays
Facial	prominence of eyes, down-sloping palpebral fissures, thickened lids large ears beaking of nose, flared nares, hypoplastic alae, short columella protruding upper lip micro-retrognathia
Skeletal	delayed closure of fontanels broad great toes
Integument	redundancy / laxity of skin minimal subcutaneous fat cutaneous capillary malformations
Cardiac	structural anomalies (ventricular septal defect, atrial level defect, pulmonary artery stenoses) arrhythmias (Torsade de points, PVCs, PACs, SVtach, Vtach) death usually associated with cardiogenic shock preceded by arrythmia.
Genital	inguinal hernia hypo- or cryptorchidism
Neurologic	hypotonia progressing to hypertonia cerebral atrophy neurogenic scoliosis

Shaded regions include features of the syndrome demonstrating variability. Though variable findings of the cardiac, genital and neurologic systems were observed, all affected individuals manifested some pathologic finding of each.

Experimental Design for Sequencing is Critical.



- We performed X-chromosome exon capture with Agilent, followed by Next Gen Sequencing with Illumina.

- We analyzed the data with ANNOVAR and VAAST (Variant Annotation, Analysis and Search Tool). New computational tools for identifying disease-causing mutations by individual genome sequencing.

Yandell, M. *et al.* 2011. “A probabilistic disease-gene finder for personal genomes.” *Genome Res*. 21 (2011). doi:10.1101/gr.123158.111.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164.

The Exon Capture and Coverage was high depth.

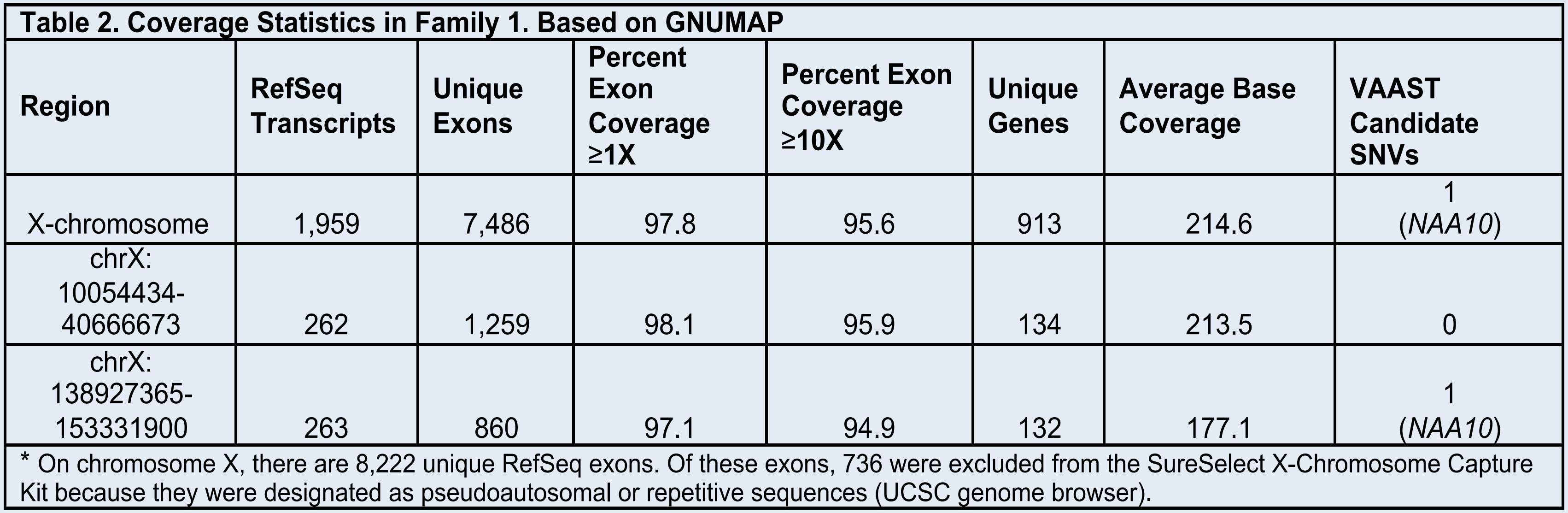
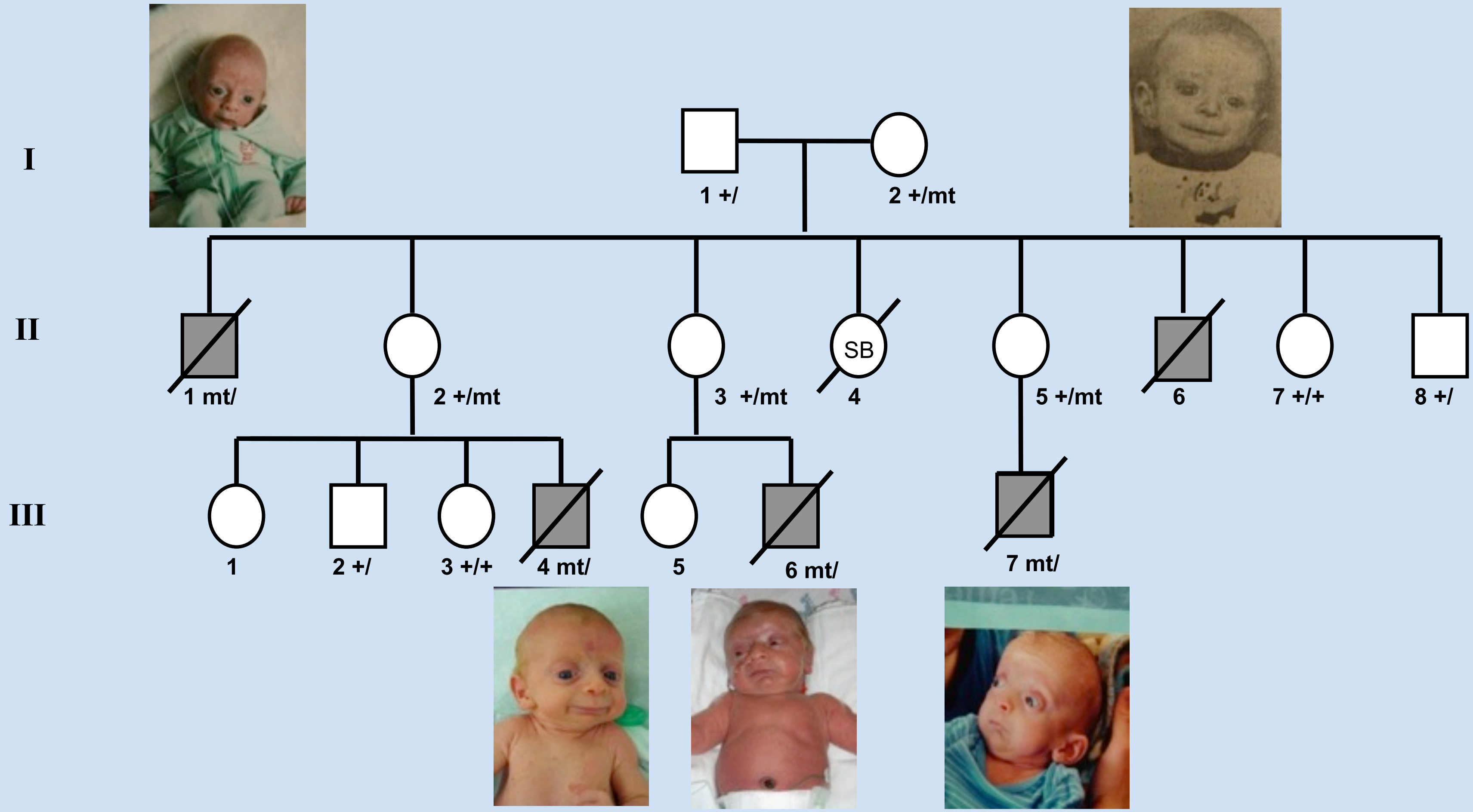


Table 2. Coverage Statistics in Family 1. Based on GNUMAP

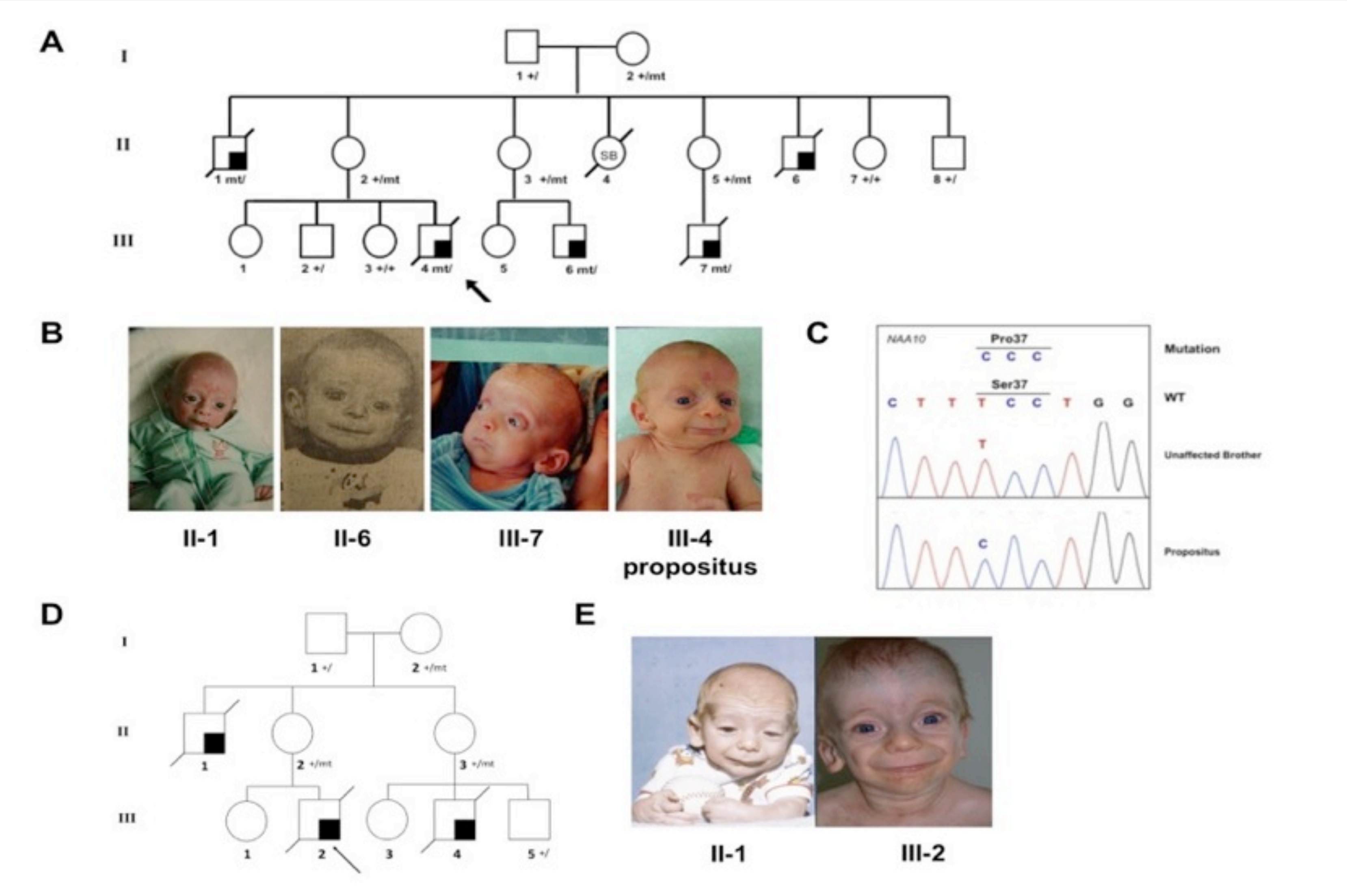
Region	**RefSeq Transcripts**	**Unique Exons**	**Percent Exon Coverage ≥1X**	**Percent Exon Coverage ≥10X**	**Unique Genes**	**Average Base Coverage**	**VAAST Candidate SNVs**
X-chromosome	1,959	7,486	97.8	95.6	913	214.6	1 (*NAA10*)
chrX: 10054434-40666673	262	1,259	98.1	95.9	134	213.5	0
chrX: 138927365-153331900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

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

Family now, with five mutation-positive boys dying from the disease.

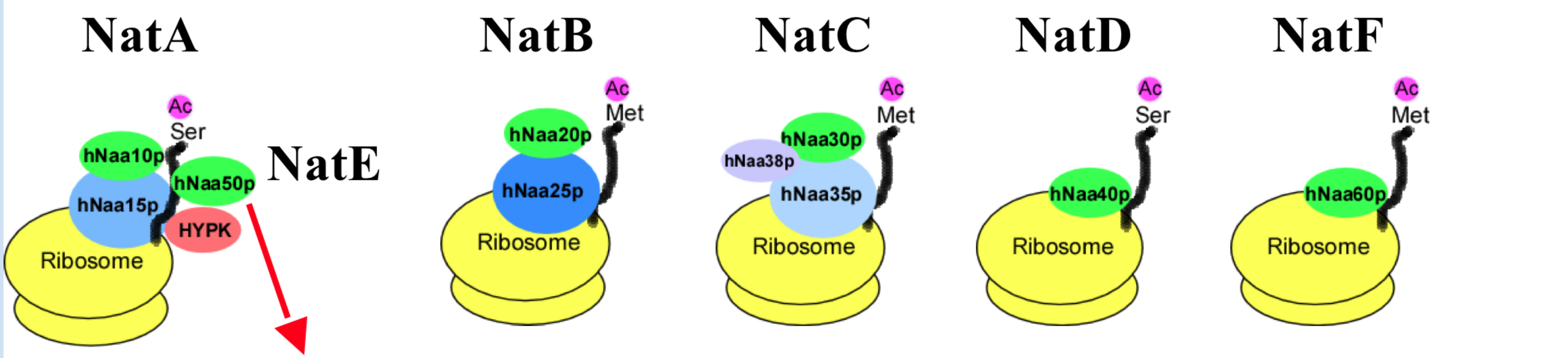


Ancestry Matters! - Ogden Syndrome



The mutation is **necessary**, but we do not know if it is **sufficient** to cause this phenotype in ANY genetic background. It simply "contributes to" the phenotype.

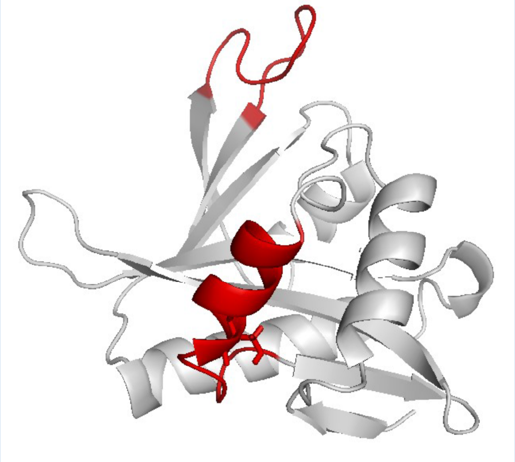
The mutation disrupts the N-terminal acetylation machinery (NatA) in human cells.



NatA	NatE	NatB	NatC	NatD	NatF
Ser	Met-Leu	Met-Asp	Met-Leu	Ser	Met-Lys
Ala		Met-Glu	Met-Ile	(H2A/H4)	Met-Ser
Thr		Met-Asn	Met-Phe		Met-Ala
Gly		Met-Gln			Met-Met
Val					Met-Leu
Asp/Glu*					

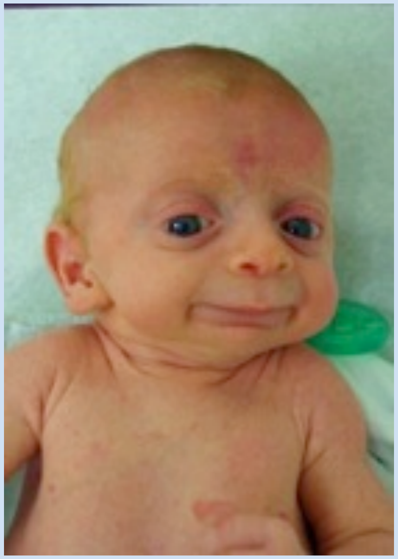
Slide courtesy of Thomas Arnesen

Big Question:

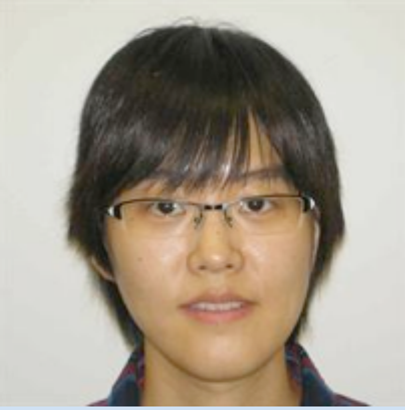


Simulated structure of S37P mutant

?

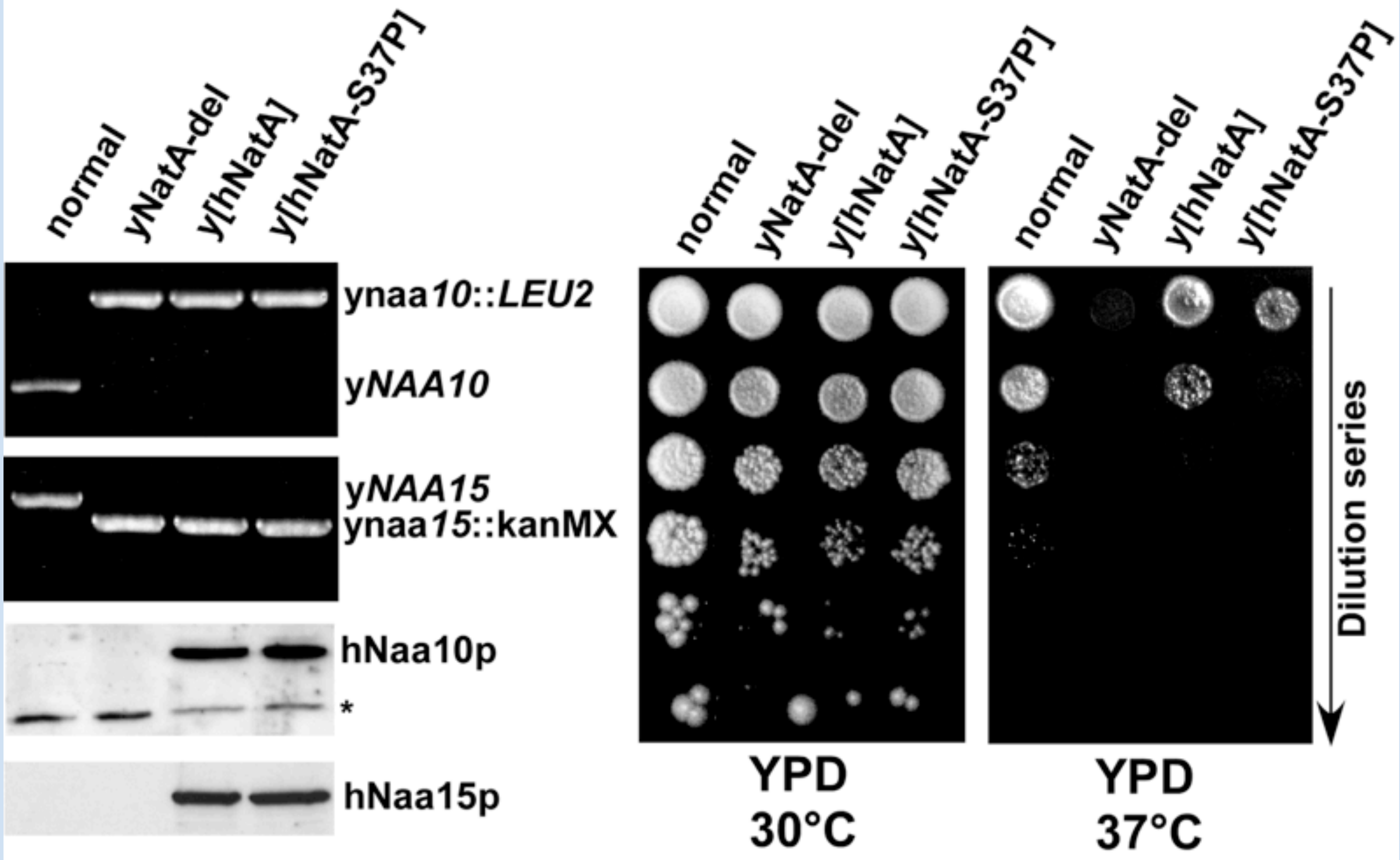


Max Doerfel



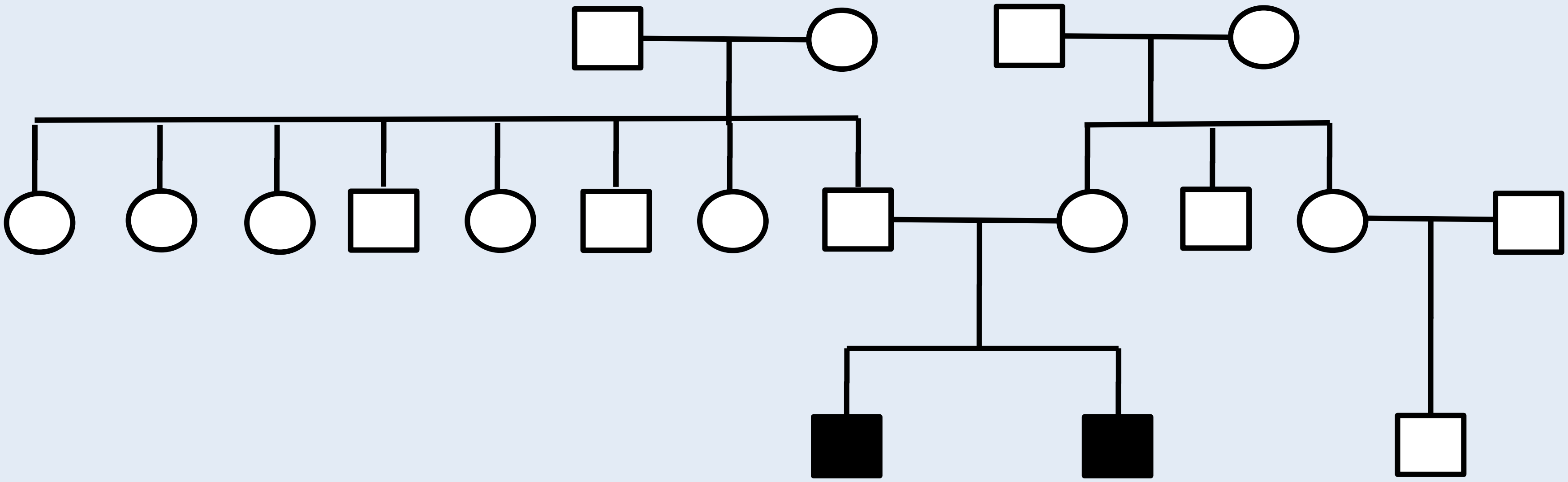
Yiyang Wu

hNaa10p-S37P is functionally impaired *in vivo* using a yeast model.



Unpublished data from Thomas Arnesen, do not further distribute.

New Syndrome with Dysmorphology, Mental Retardation, “Autism”, “ADHD”



Likely X-linked or Autosomal Recessive, with X-linked being supported by extreme X-skewing in the mother

1.5 years old

3.5 years old

7 years old

3 years old

5 years old

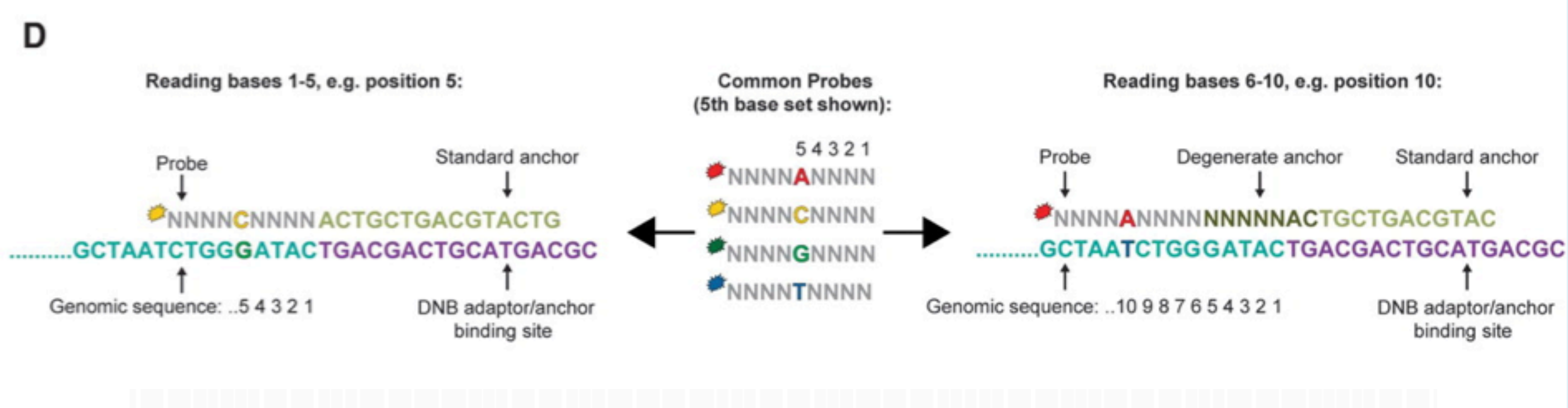
9 years old

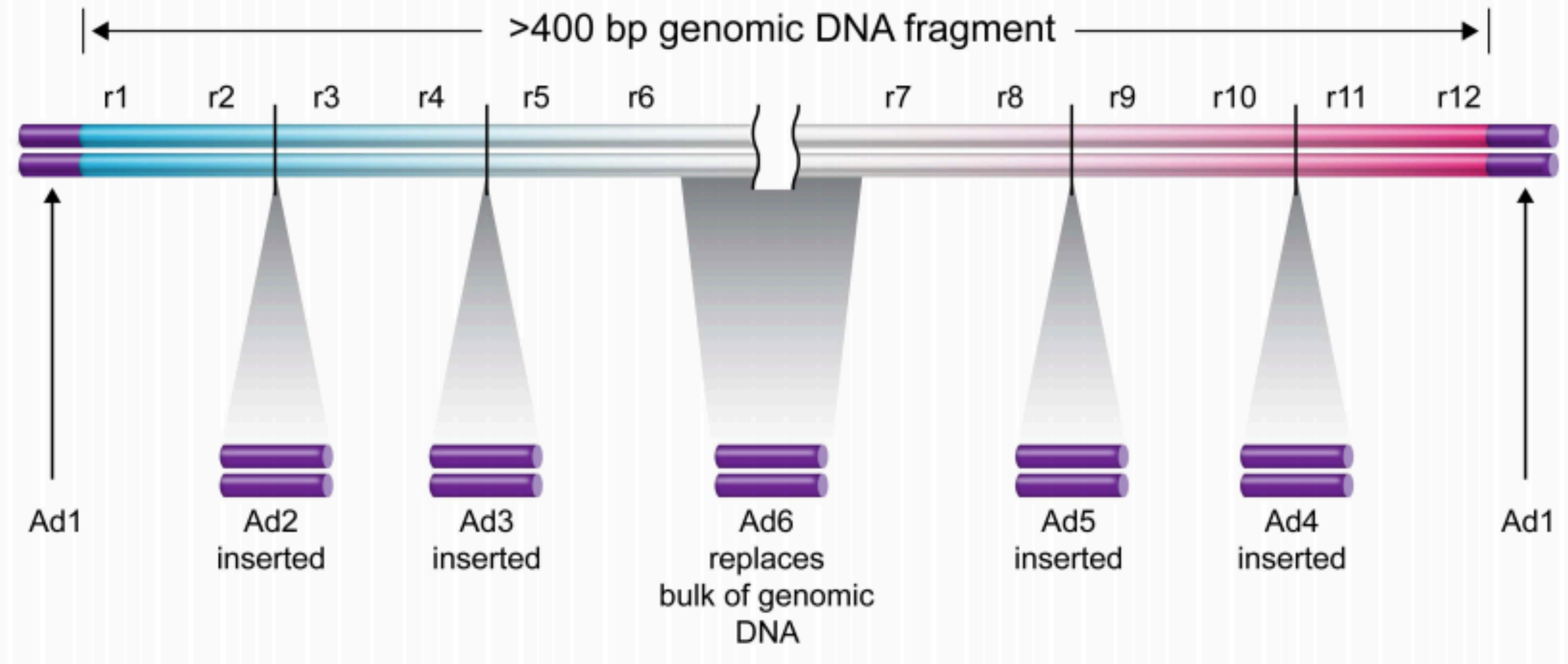
Workup Ongoing for past 10 years

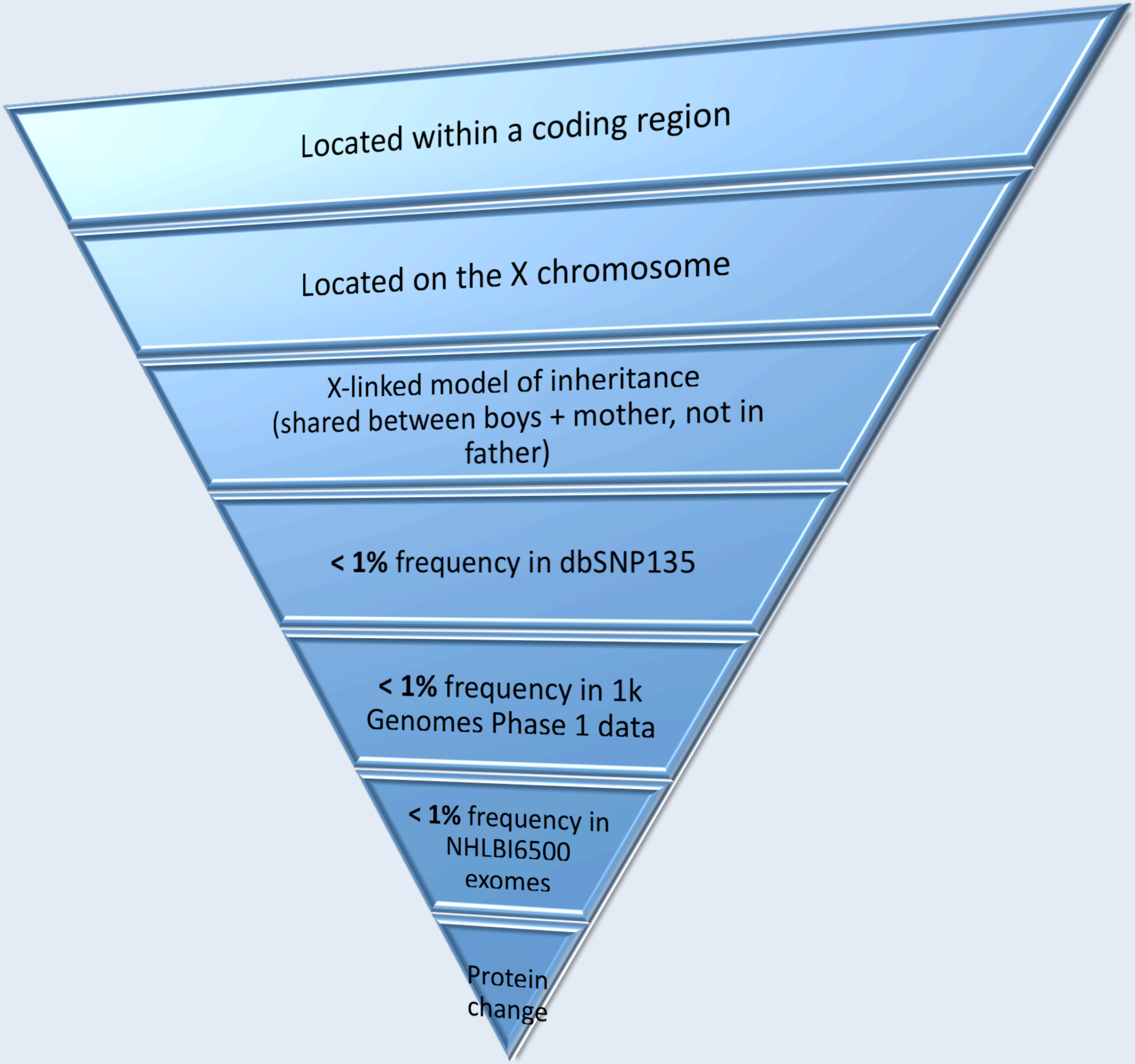
- Numerous genetic tests negative, including negative for Fragile X and many candidate genes.
- No obvious pathogenic CNVs – microarrays normal.
- Sequenced whole genomes of Mother, Father and Two Boys, using Complete Genomics, obtained data in June of this year, i.e. version 2.0 CG pipeline.

Jason O'Rawe,
analyst

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





22,174	Located within a coding region
272	Located on the X chromosome
56	X-linked model of inheritance (shared between boys + mother, not in father)
7	< 1% frequency in dbSNP135
6	< 1% frequency in 1k Genomes Phase 1 data
5	< 1% frequency in NHLBI6500 exomes
3	Protein change

Variant classification

Variant	Reference	Alternate	Classification	Gene 1	Transcript 1	Exon 1	HGVS Coding 1	HGVS Protein 1
X:47307978-SNV	G	T	Nonsyn SNV	ZNF41	NM_007130	5	c.1191C>A	p.Asp397Glu
X:63444792-SNV	C	A	Nonsyn SNV	ASB12	NM_130388	2	c.739G>T	p.Gly247Cys
X:70621541-SNV	T	C	Nonsyn SNV	TAF1	NM_004606	25	c.4010T>C	p.Ile1337Thr

SIFT classification

Chromosome	Position	Reference	Coding?	SIFT Score	Score <= 0.05	Ref/Alt Alleles
X	47307978	G	YES	0.649999976	0	G/T
X	63444792	C	YES	0	1	C/A
X	70621541	T	YES	0.009999999776	1	T/C

VAAST score

RANK	Gene	p-value	p-value-ci	Score	Variants
1	ASB12	1.56E-11	1.55557809307134e-11,0.000290464582480396	38.63056297	chrX:63444792;38.63;C->A;G->C;0,3
2	TAF1	1.56E-11	1.55557809307134e-11,0.000290464582480396	34.51696816	chrX:70621541;34.52;T->C;I->T;0,3
3	ZNF41	1.56E-11	1.55557809307134e-11,0.000290464582480396	32.83011803	chrX:47307978;32.83;G->T;D->E;0,3

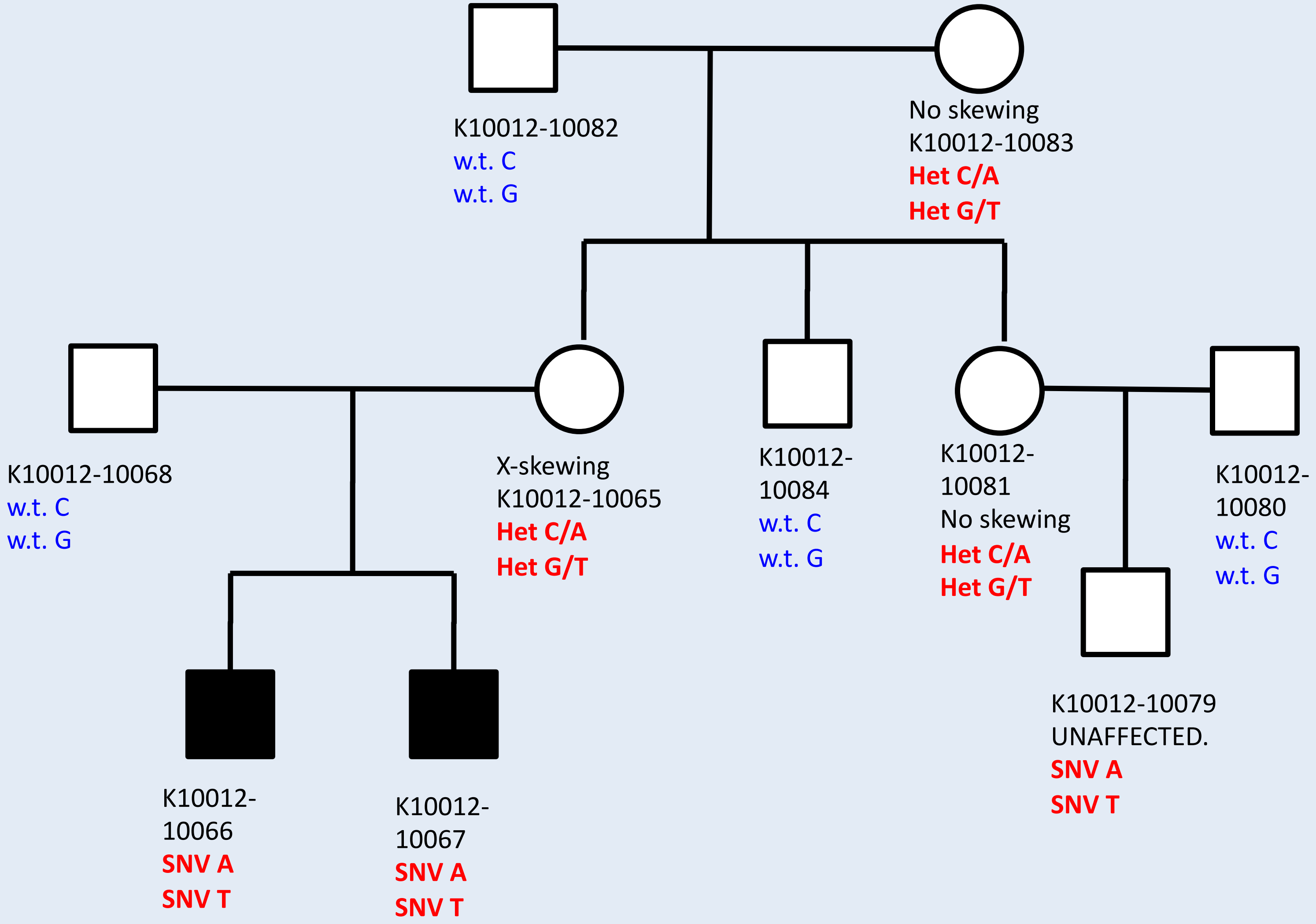
Mutations in the *ZNF41* Gene Are Associated with Cognitive Deficits: Identification of a New Candidate for X-Linked Mental Retardation

Sarah A. Shoichet,[1] Kirsten Hoffmann,[1] Corinna Menzel,[1] Udo Trautmann,[2] Bettina Moser,[1] Maria Hoeltzenbein,[1] Bernard Echenne,[3] Michael Partington,[4] Hans van Bokhoven,[5] Claude Moraine,[6] Jean-Pierre Fryns,[7] Jamel Chelly,[8] Hans-Dieter Rott,[2] Hans-Hilger Ropers,[1] and Vera M. Kalscheuer[1]

[1]Max-Planck-Institute for Molecular Genetics, Berlin; [2]Institute of Human Genetics, University of Erlangen-Nuremberg, Erlangen-Nuremberg; [3]Centre Hospitalier Universitaire de Montpellier, Hôpital Saint-Eloi, Montpellier, France, [4]Hunter Genetics and University of Newcastle, Waratah, Australia; [5]Department of Human Genetics, University Medical Centre, Nijmegen, The Netherlands; [6]Services de Génétique–INSERM U316, CHU Bretonneau, Tours, France; [7]Center for Human Genetics, Clinical Genetics Unit, Leuven, Belgium; and [8]Institut Cochin de Génétique Moleculaire, Centre National de la Recherche Scientifique/INSERM, CHU Cochin, Paris

Am. J. Hum. Genet. 73:1341–1354, 2003

Sanger validation: ASB12 and ZNF41 mutations



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

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

Ancestry Matters!

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

- Tool Building for Human Genetics

- Can we reliably detect a comprehensive, and accurate, set of variants using more than one pipeline, or even more than one sequencing platform?

- How much data is enough, and how reliable and reproducible are variant calls?

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

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

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

Understand Your Genome Symposium

During this two-day educational event, industry experts will discuss the clinical implementation of whole-genome next-generation sequencing (NGS) technology.

illumina®

Ordering Physician:
Gholson Lyon, MD
Steinmann Institute
10 West Broadway, Suite #820
Salt Lake City, UT 84101

Individual Genome Sequence Results

Clinical Report

www.everygenome.com
CLIA#: 05D1092911

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

PRIVACY and PROGRESS in Whole Genome Sequencing

Presidential Commission
for the Study of Bioethical Issues

October 2012

Recommendation 4.1

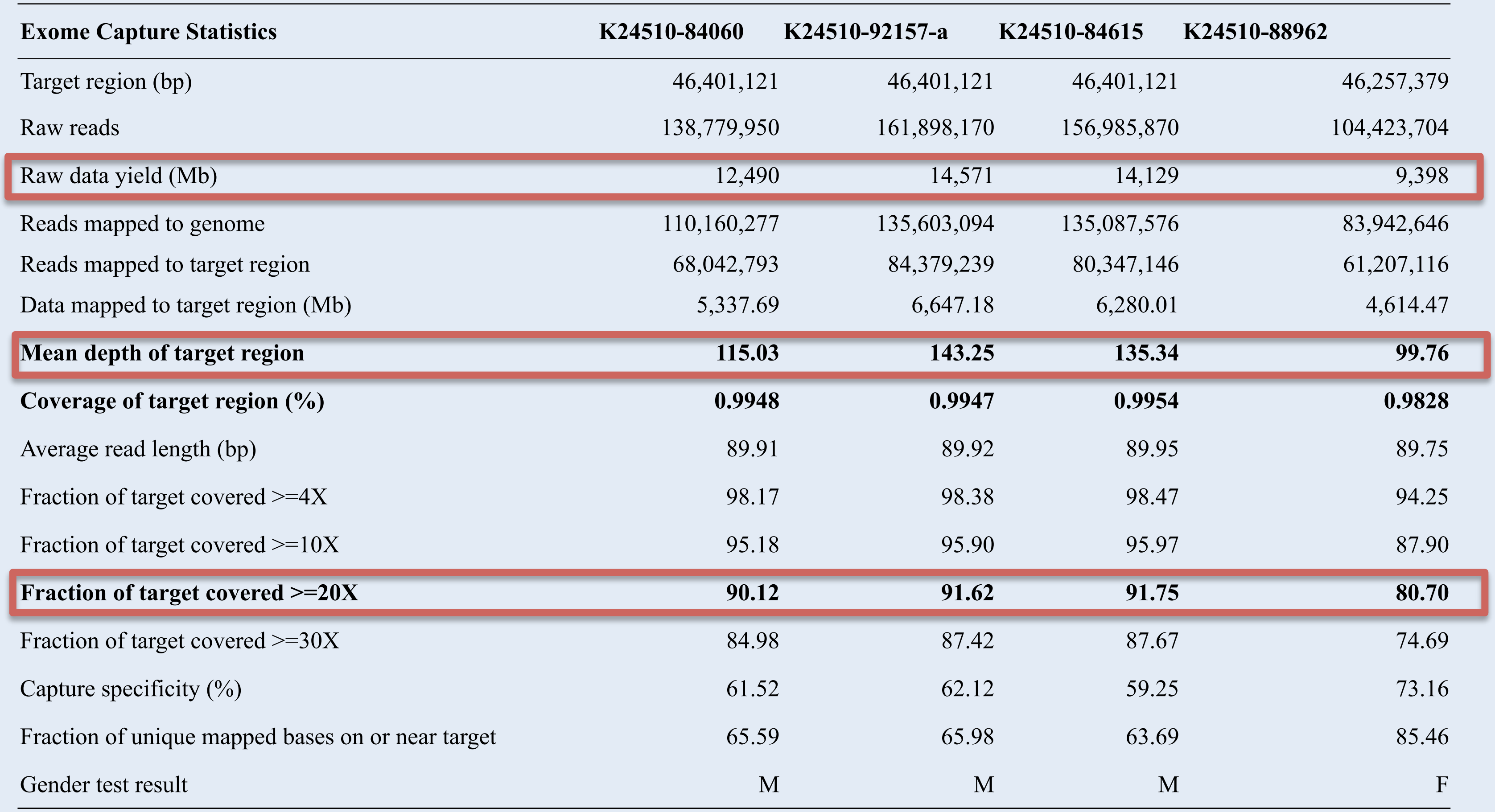
Funders of whole genome sequencing research, relevant clinical entities, and the commercial sector should facilitate explicit exchange of information between genomic researchers and clinicians, while maintaining robust data protection safeguards, so that whole genome sequence and health data can be shared to advance genomic medicine.

Performing all whole genome sequencing in CLIA-approved laboratories would remove one of the barriers to data sharing. It would help ensure that whole genome sequencing generates high-quality data that clinicians and researchers can use to draw clinically relevant conclusions. It would also ensure that individuals who obtain their whole genome sequence data could share them more confidently in patient-driven research initiatives, producing more meaningful data. That said, current sequencing technologies and those in development are diverse and evolving, and standardization is a substantial challenge. Ongoing efforts, such as those by the Standardization of Clinical Testing working group are critical to achieving standards for ensuring the reliability of whole genome sequencing results, and facilitating the exchange and use of these data.[216]

Optimizing Variant Calling in Exomes at BGI in 2011

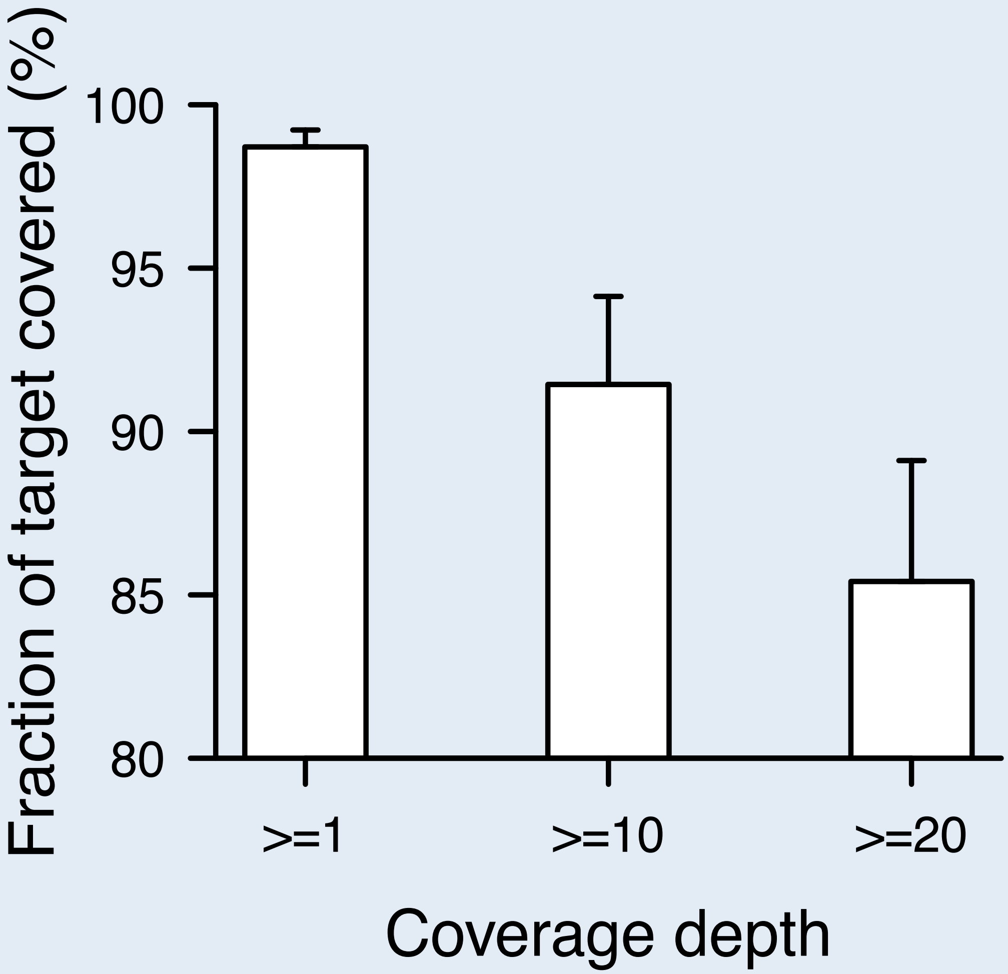
- Agilent v2 44 MB exome kit
- Illumina Hi-Seq for sequencing.
- Average coverage ~100-150x.
- Depth of sequencing of >80% of the target region with >20 reads or more per base pair.
- Comparing various pipelines for alignment and variant-calling.

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



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

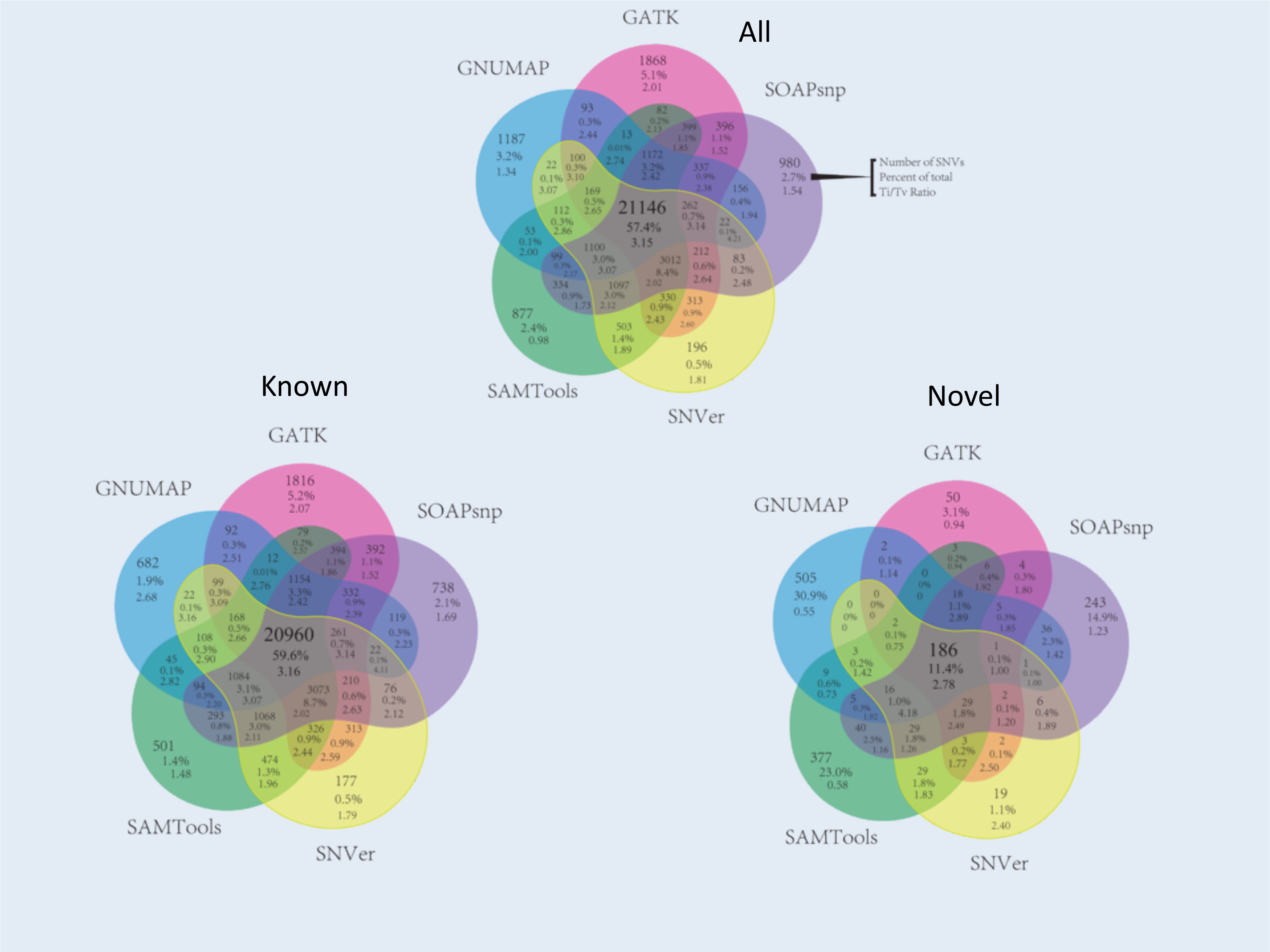
1) BWA - **GATK** (version 1.5) with recommended parameters (GATK IndelRealigner, base quality scores were re-calibrated by GATK Table Recalibration tool. Genotypes called by GATK UnifiedGenotyper. For SNVs and indels.

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.

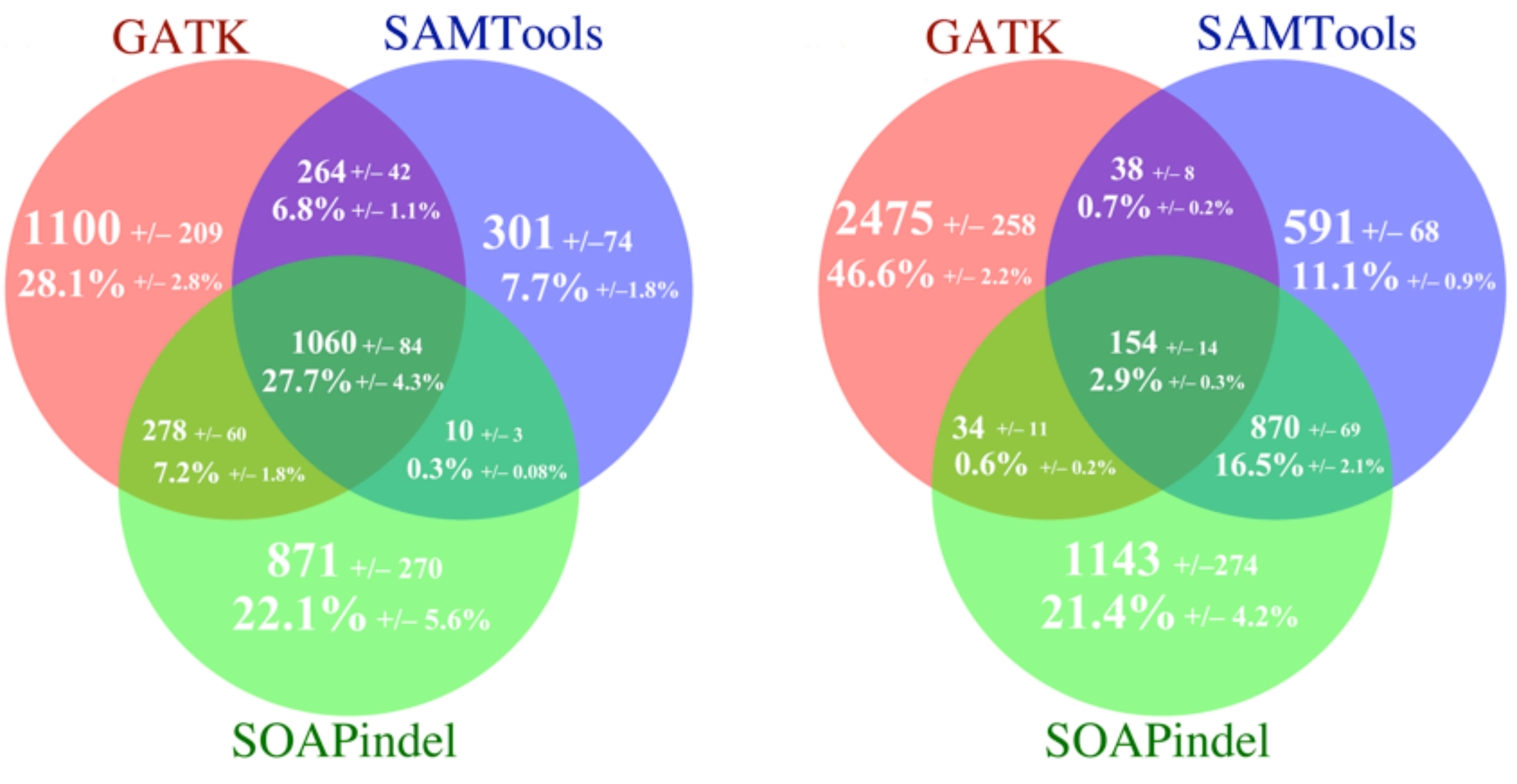
5) BWA - Sam format to Bam format - Picard to remove duplicates – **SNVer** , for SNVs only


All
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
1100
3.0%
3.07
3012
8.4%
2.02
212
0.6%
2.64
83
0.2%
2.48
99
0.3%
2.17
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
1084
3.1%
3.07
3073
8.7%
2.02
210
0.6%
2.63
76
0.2%
2.12
94
0.3%
2.20
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
16
1.0%
4.18
29
1.8%
2.49
2
0.1%
1.20
6
0.4%
1.89
5
0.3%
1.92
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

INDELS

Indels- Overlap by Base Position only

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



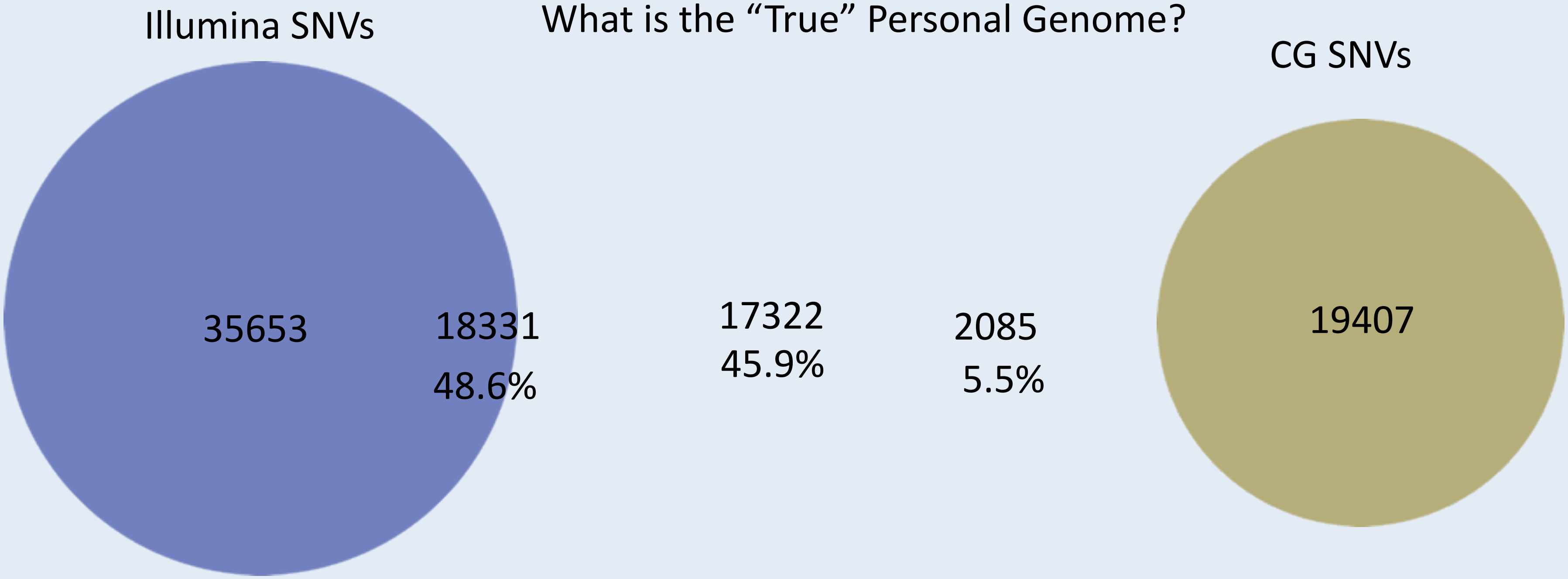
Total mean overlap, plus or minus one standard deviation, observed between three indel calling pipelines: GATK, SOAP-indel, and SAMTools. a) Mean overlap when indel position was the only necessary agreement criterion. **b)** Mean overlap when indel position, base length and base composition were the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

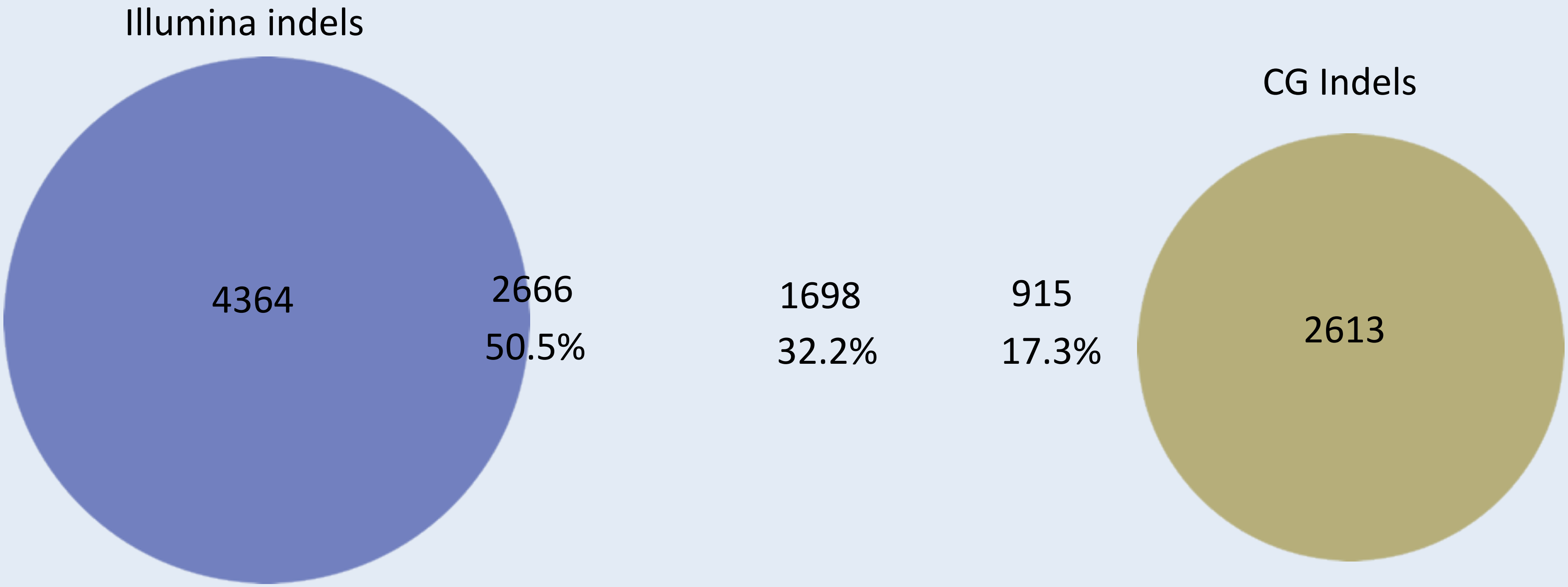
- How reliable are variants that are uniquely called by individual pipelines?

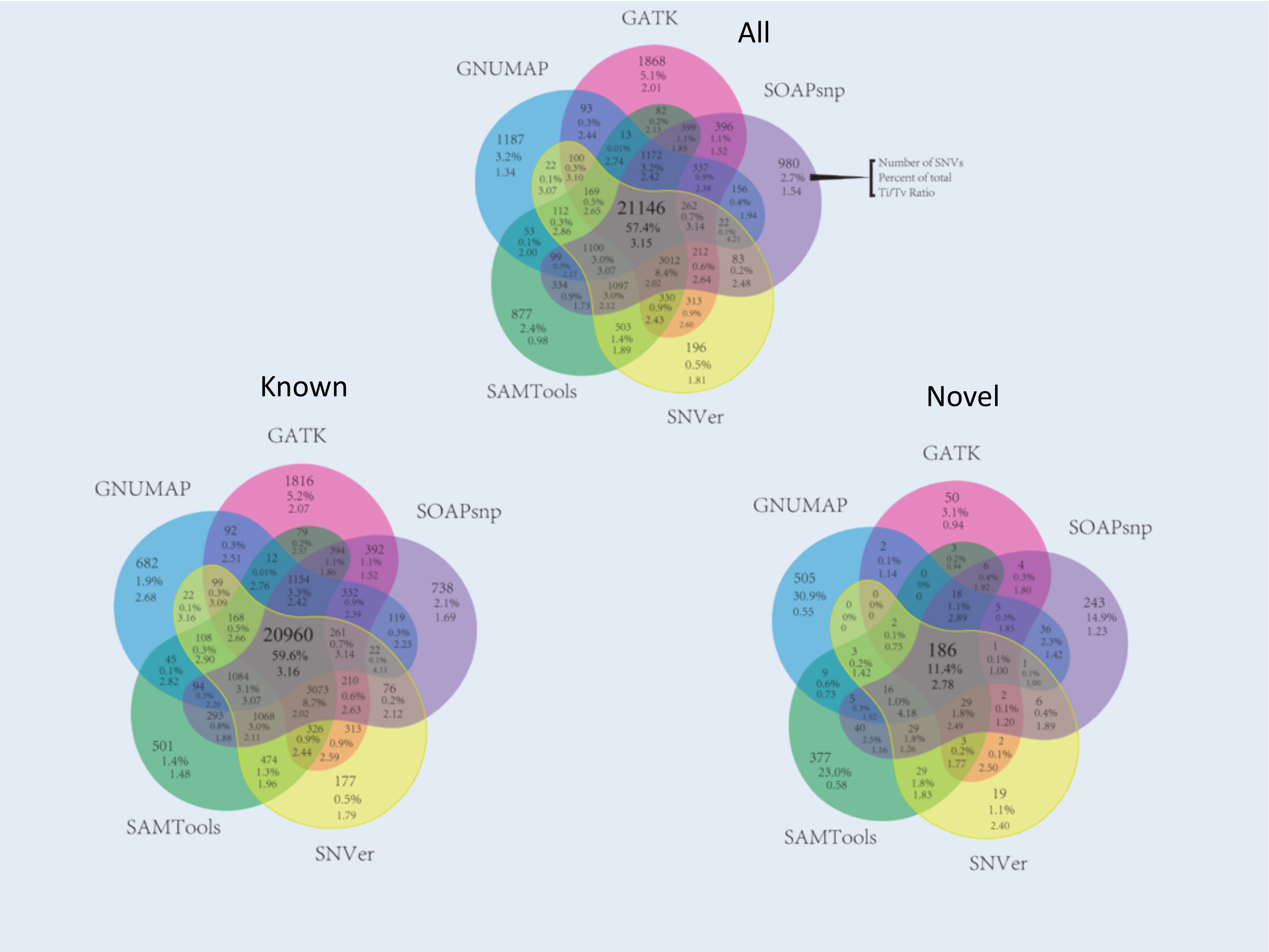
- Are some pipelines better at detecting rare, or novel variants than others?

Cross validation using orthogonal sequencing technology (Complete Genomics)

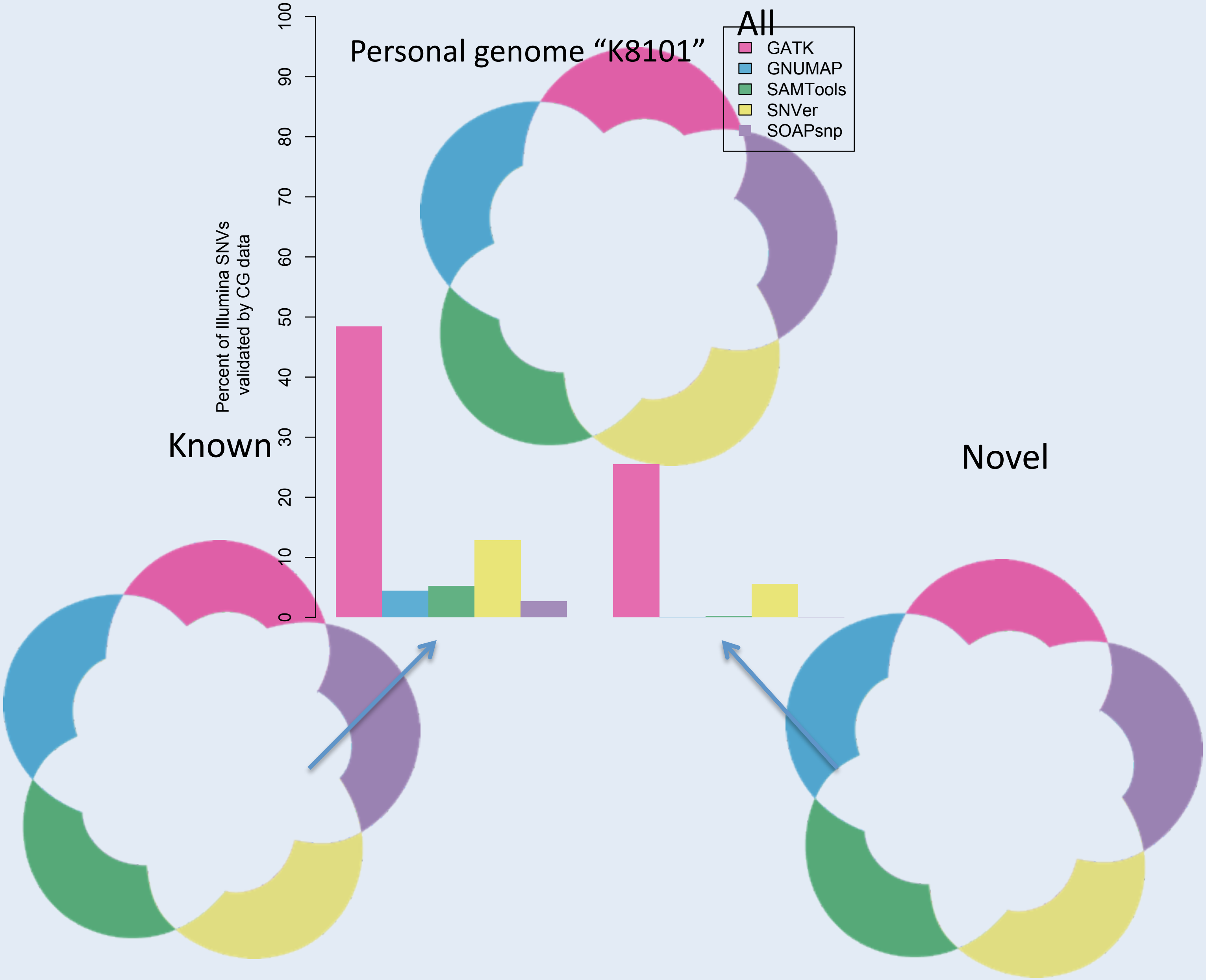
What is the “True” Personal Genome?





All
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1868 5.1% 2.01
93 0.3% 2.44
82 0.2% 2.13
1187 3.2% 1.34
13 0.01% 2.74
399 1.1% 1.85
396 1.1% 1.52
980 2.7% 1.54
Number of SNVs
Percent of total
Ti/Tv Ratio
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
156 0.4% 1.94
21146 57.4% 3.15
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
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
682 1.9% 2.68
12 0.01% 2.76
394 1.1% 1.86
392 1.1% 1.52
738 2.1% 1.69
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
119 0.3% 2.23
20960 59.6% 3.16
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
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
505 30.9% 0.55
0 0% 0
6 0.4% 1.92
4 0.3% 1.80
243 14.9% 1.23
0 0% 0
0 0% 0
18 1.1% 2.89
5 0.3% 1.85
2 0.1% 0.75
36 2.3% 1.42
186 11.4% 2.78
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
377 23.0% 0.58
29 1.8% 1.83
19 1.1% 2.40

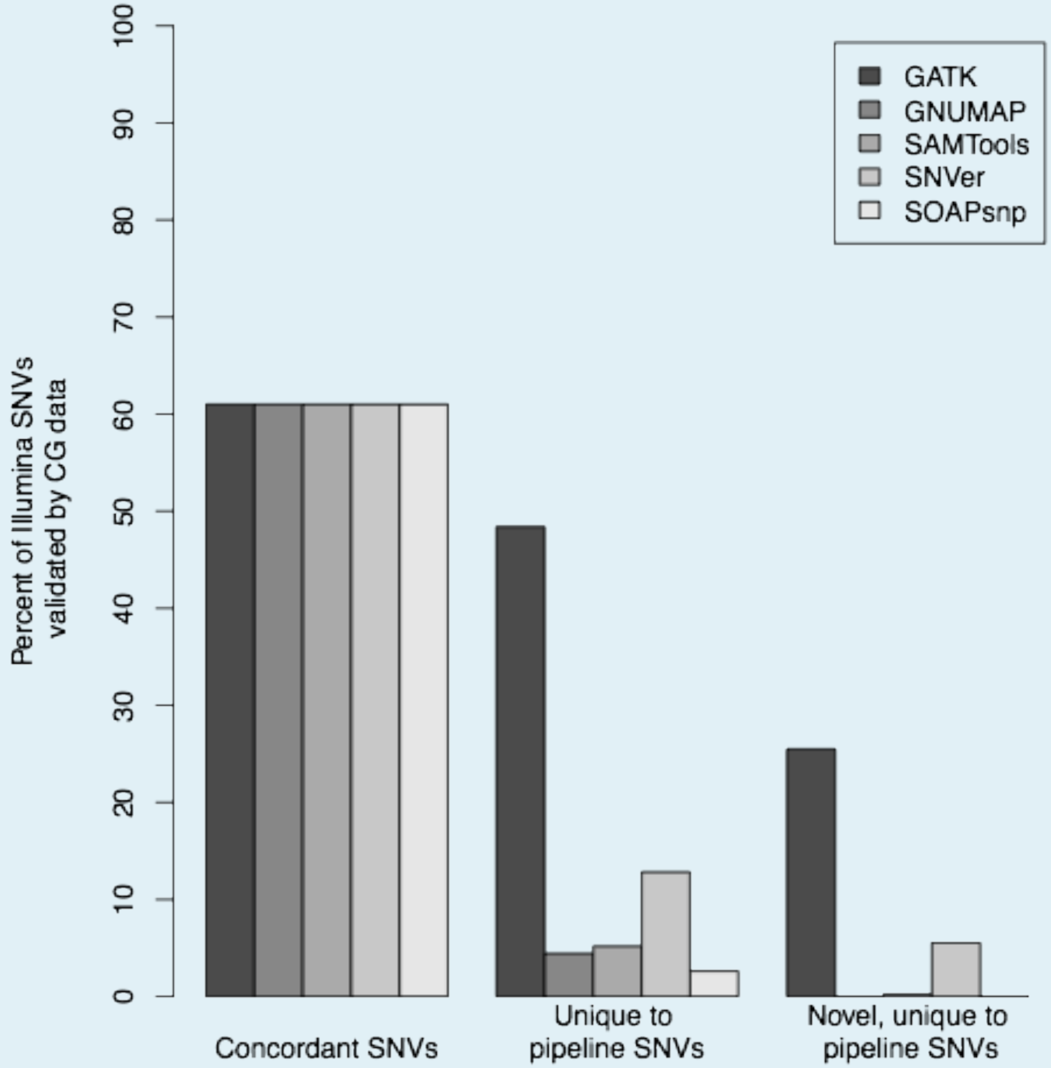
Personal genome “K8101”



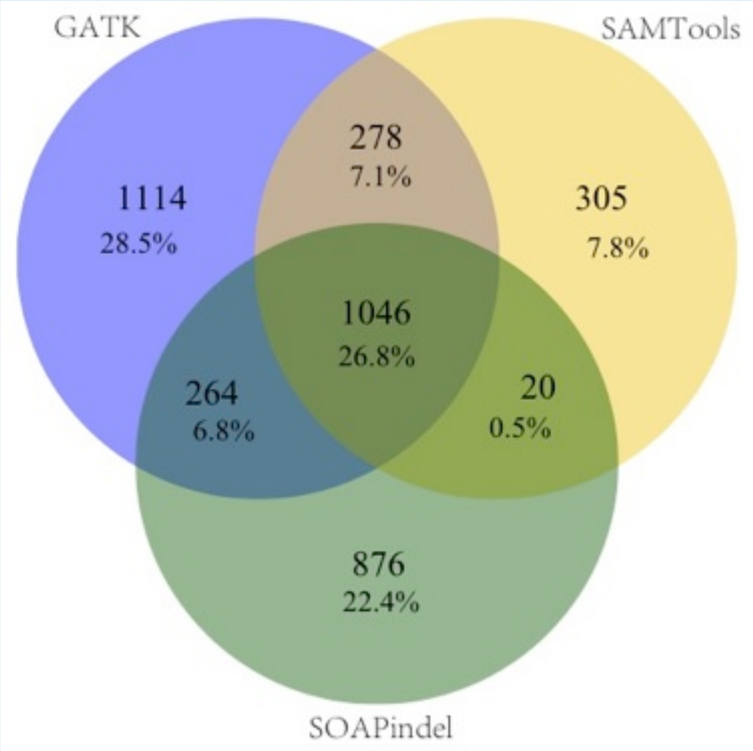
Higher Validation of SNVs with the BWA-GATK pipeline

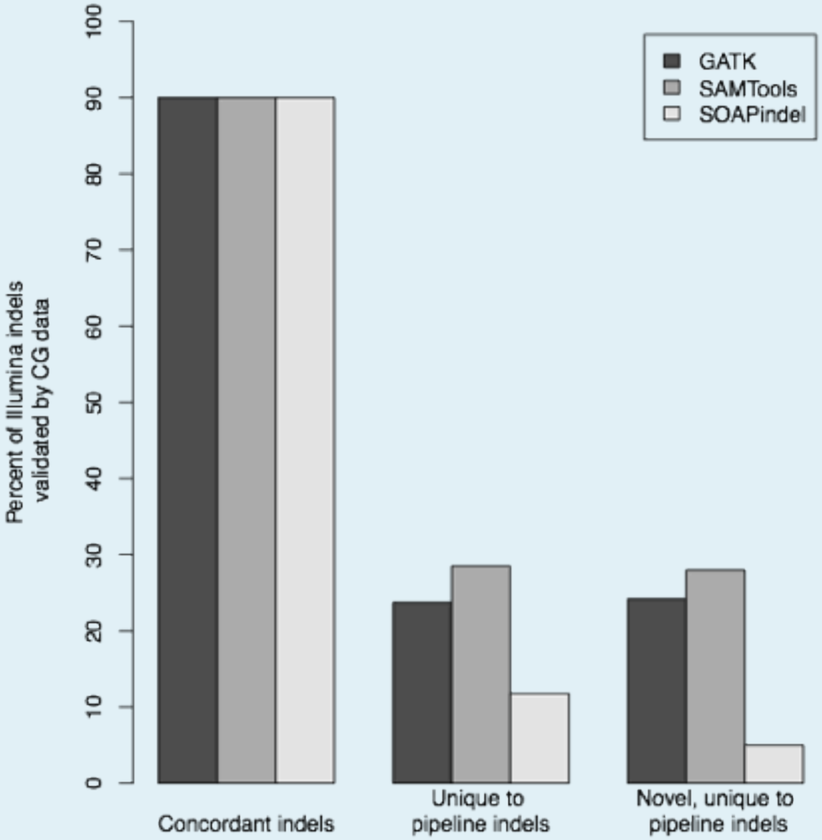
- Reveals higher validation rate of unique-to-pipeline variants, as well as uniquely discovered novel variants, for the variants called by BWA-GATK, in comparison to the other 4 pipelines (including SOAP).

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

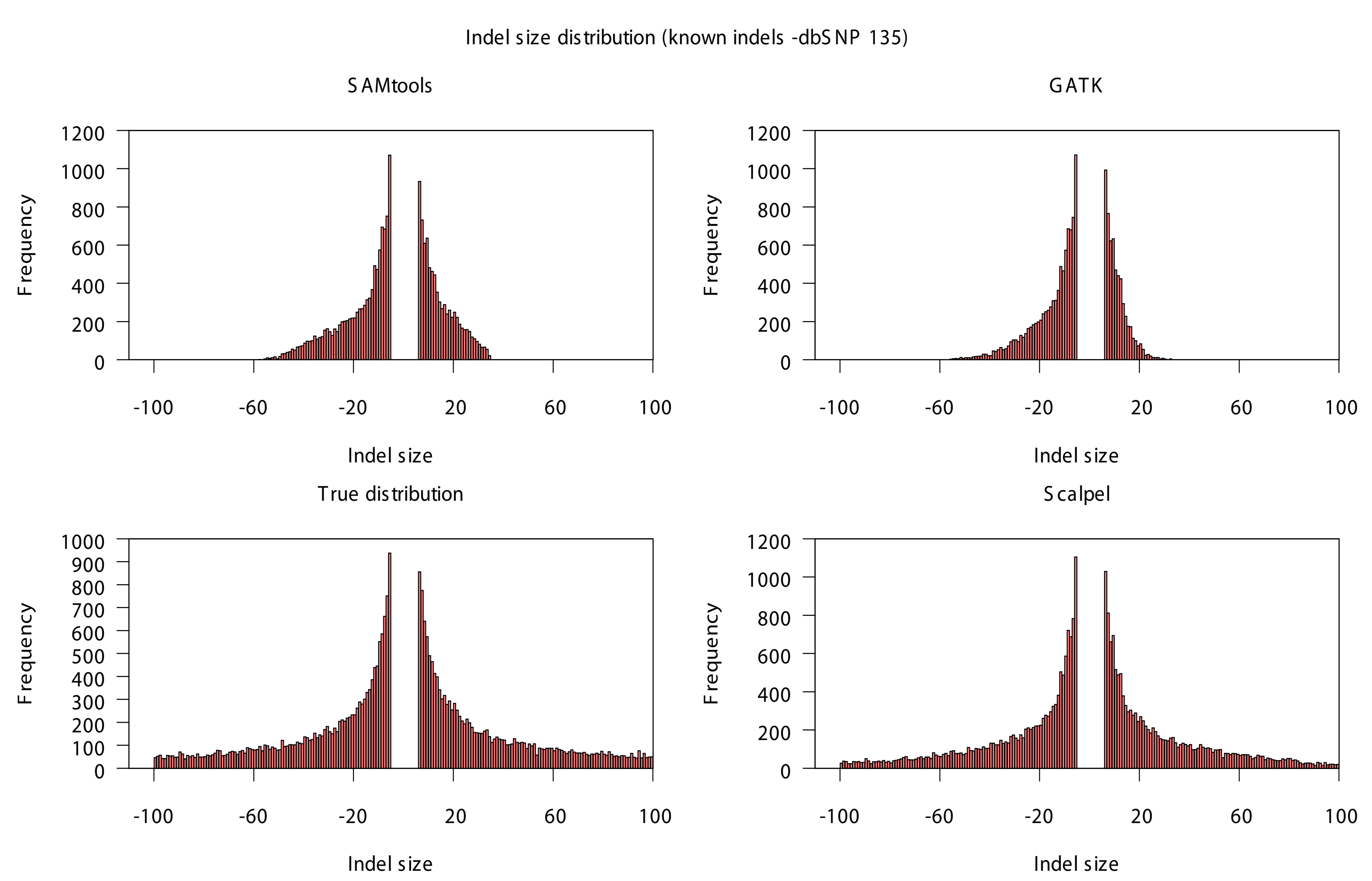


Validating Indels with Complete Genomics Data for the 3 pipelines

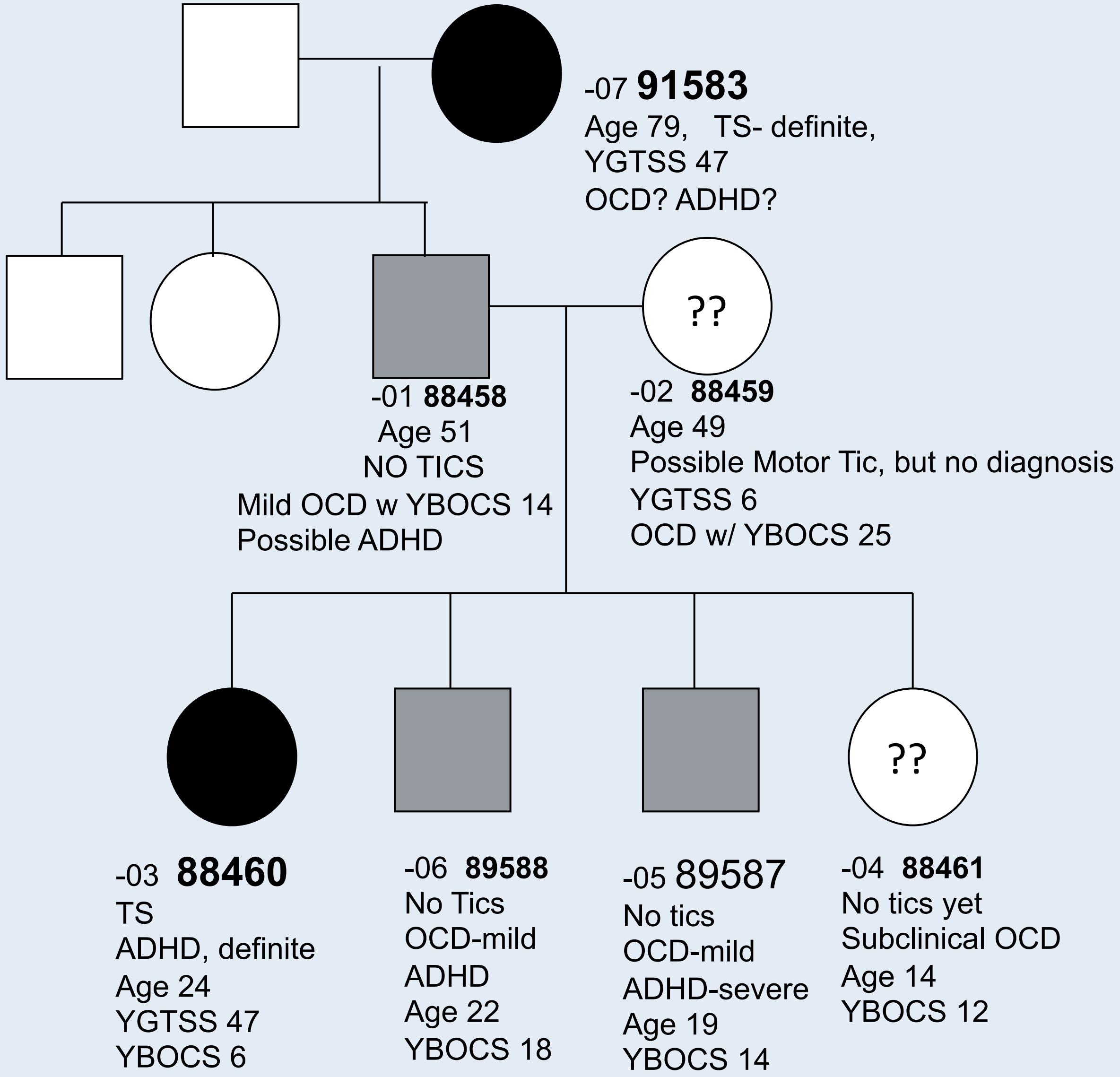




Simulated Data vs. 3 pipelines



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

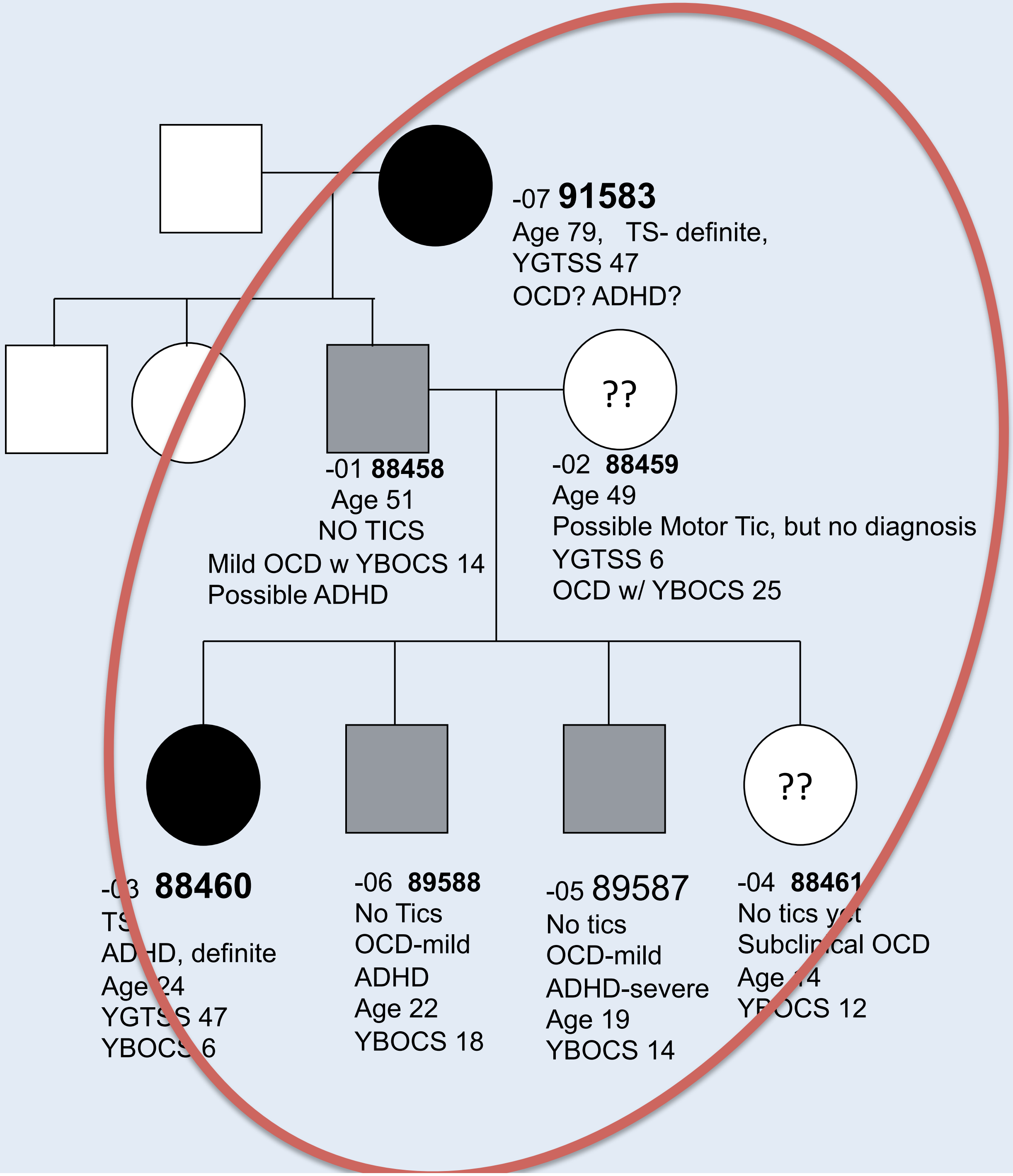
-03 **88460**
TS
ADHD, definite
Age 24
YGTSS 47
YBOCS 6

-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

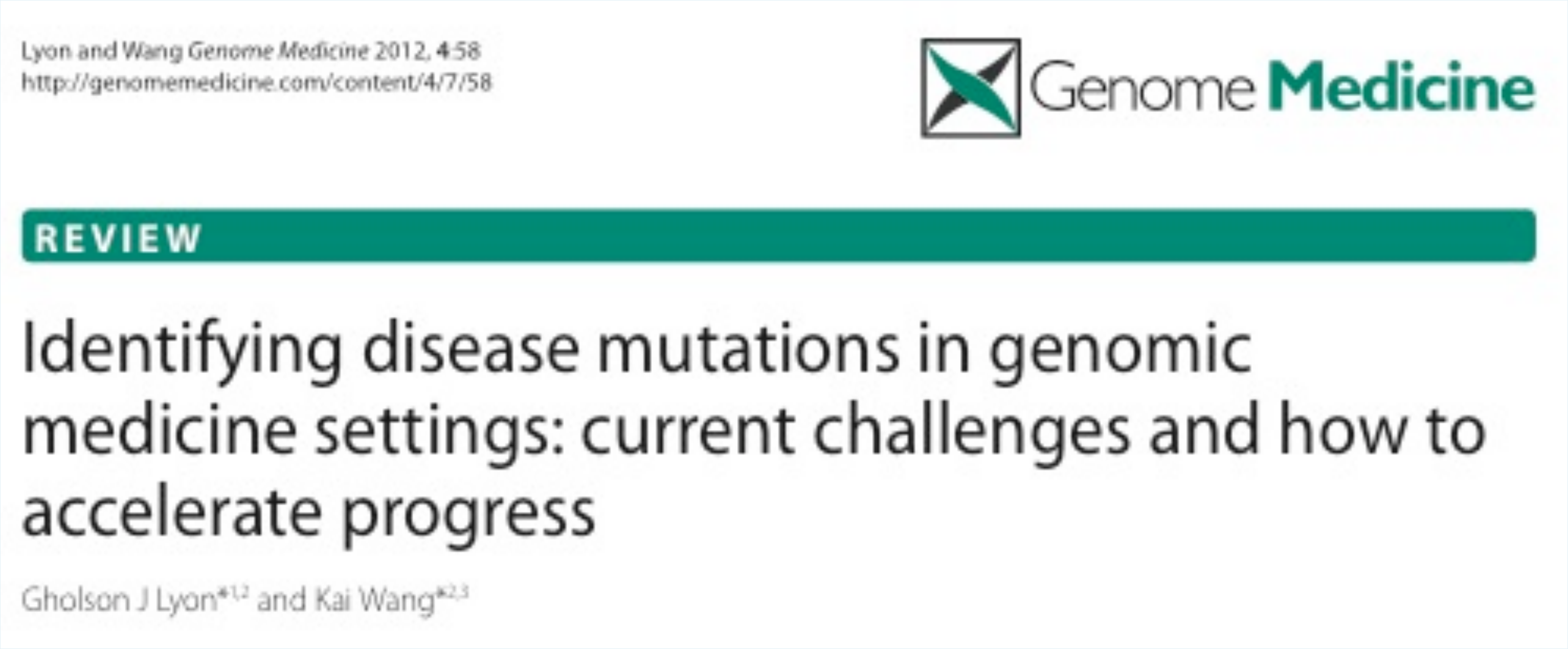
Optimizing pipeline based on literature value of ~1 true de novo protein-altering mutation per exome

Family 1	Number of putative "de-novo" coding nonsynonymous or nonsense SNVs detected without using grandparent as a filter	Number of putative "de-novo" coding nonsynonymous or nonsense SNVs detected when also using one grandparent as a filter
Child A	241	1
Child B	211	0
Child C	102	6
Child D	242	3
Family 2		
Child A	49	N/A- No Grandparent available
Child B	41	N/A - No Grandparent available

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.

Clinical Validity?

This is SO complex that the only solid way forward is with a “networking of science” model, i.e. online database with genotype and phenotype longitudinally tracked.



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

Genome **Medicine**

REVIEW

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

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

Clinical Validity with Worldwide Human Genotype-Phenotype“database”?

PatientsLikeMe

Conclusions

- Ancestry, i.e. genetic background, matters!
- We need to sequence whole genomes of large pedigrees, and then construct super-family structures, starting in Utah.
- Collectively, we need to improve the accuracy of “whole” genomes, and also enable the sharing of genotype and phenotype data broadly, among researchers, the research participants and consumers.

Acknowledgments

THE UNIVERSITY OF UTAH

Alan Rope
John C. Carey
Steven Chin
Brian Dalley
Heidi Deborah Fain
Chad D. Huff
W. Evan Johnson
Lynn B. Jorde
Barry Moore
John M. Opitz
Theodore J. Pysher
Christa Schank
Sarah T. South
Jeffrey J Swensen
Jinchuan Xing
Mark Yandell

UTAH FOUNDATION FOR BIOMEDICAL RESEARCH

Reid Robison
Edwin Nyambi

USC
Kai Wang

CSH STANLEY INSTITUTE FOR COGNITIVE GENOMICS
COLD SPRING HARBOR LABORATORY

Jason O'Rawe
Yiyang Wu
Max Doerfel
Michael Schatz
Giuseppe Narzisi
Jennifer Parla
Shane McCarthy
Jesse Gillis

UNIVERSITAS BERGENSIS

Thomas Arnesen
Rune Evjenth
Johan R. Lillehaug

our study families

华大基因
BGI

Tao Jiang
Guangqing Sun
Jun Wang