Challenges of Clinical Implementation of Genomic Medicine

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

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

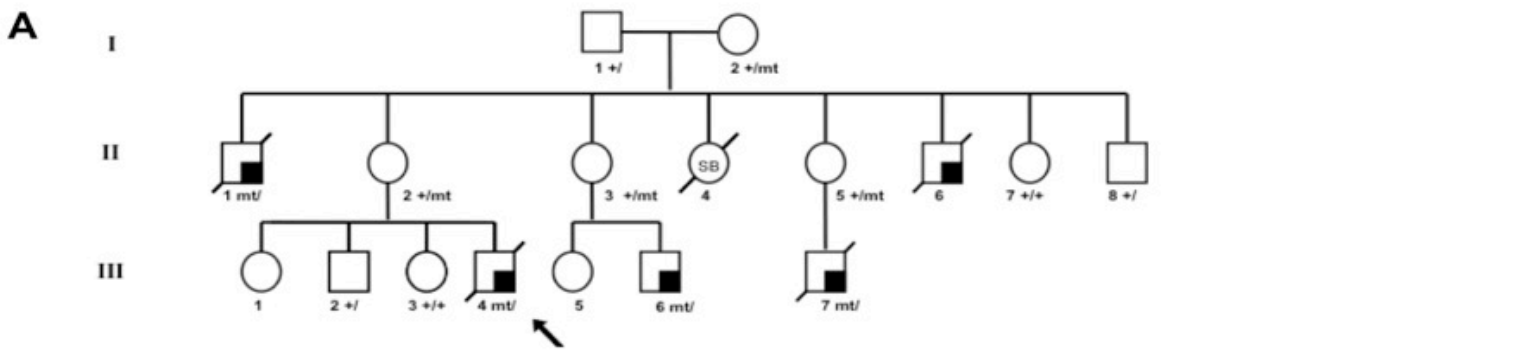
Signature

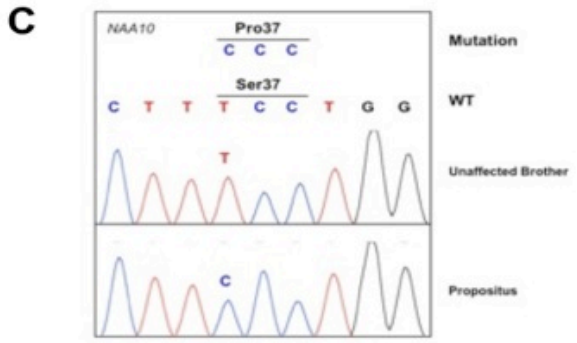
14-Apr-2011
Date

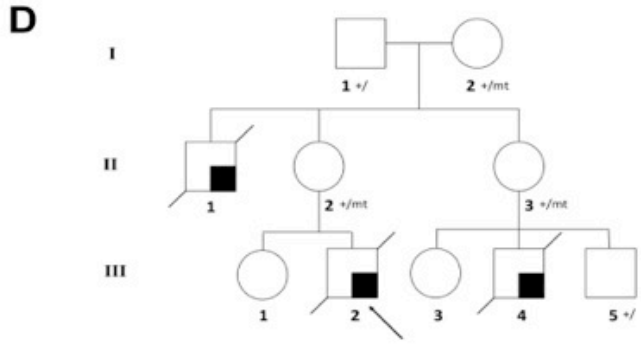
Penetrance and Expressivity

- 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?

Ancestry Matters! - Ogden Syndrome

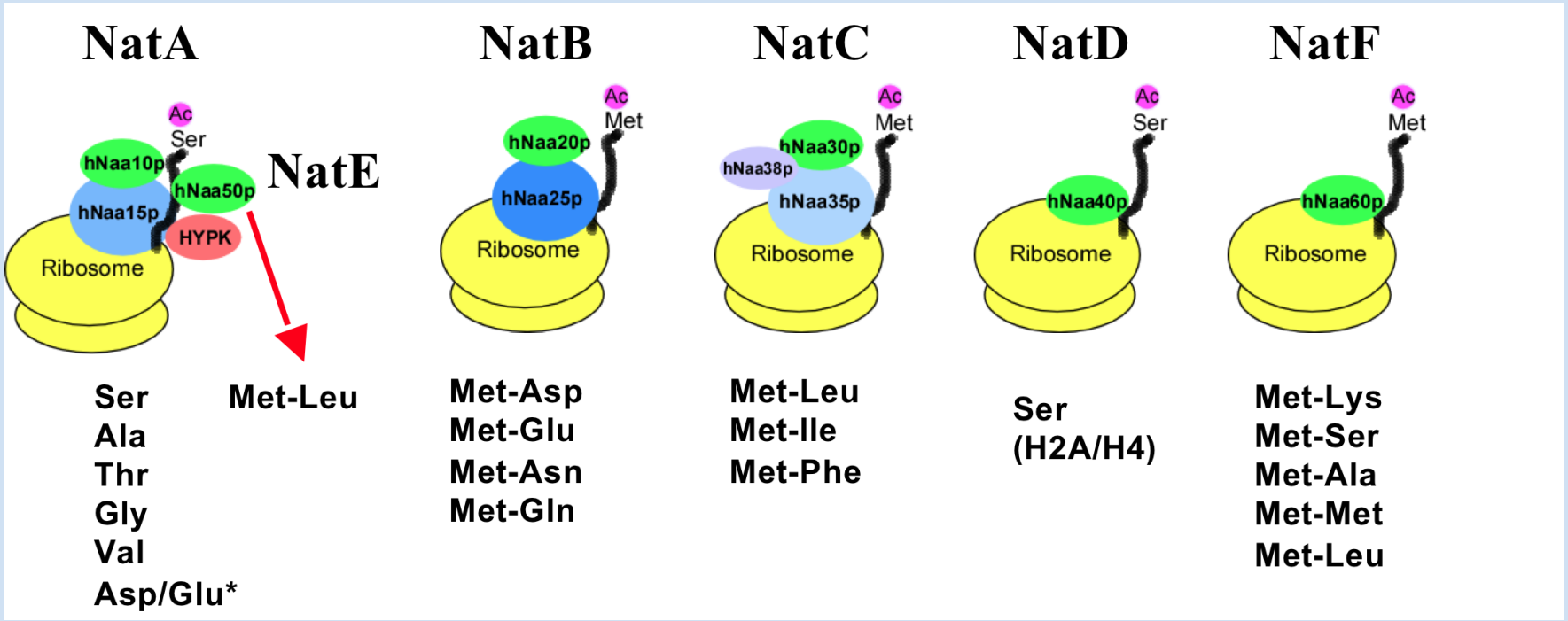






The mutation is **necessary**, but we do not know if it is **sufficient** to cause this phenotype in ANY genetic background. It simply “contributes to” the phenotype.

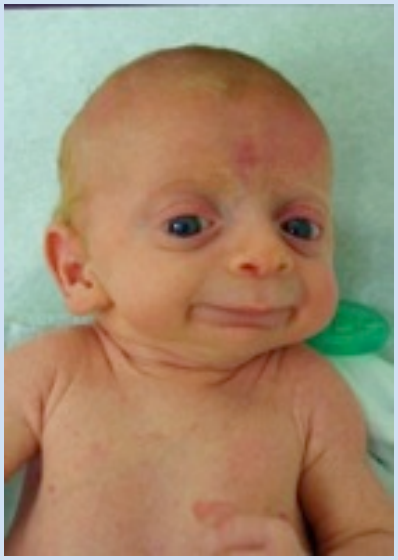
The mutation disrupts the N-terminal acetylation machinery (NatA) in human cells.



Slide courtesy of Thomas Arnesen

Big Question:

?

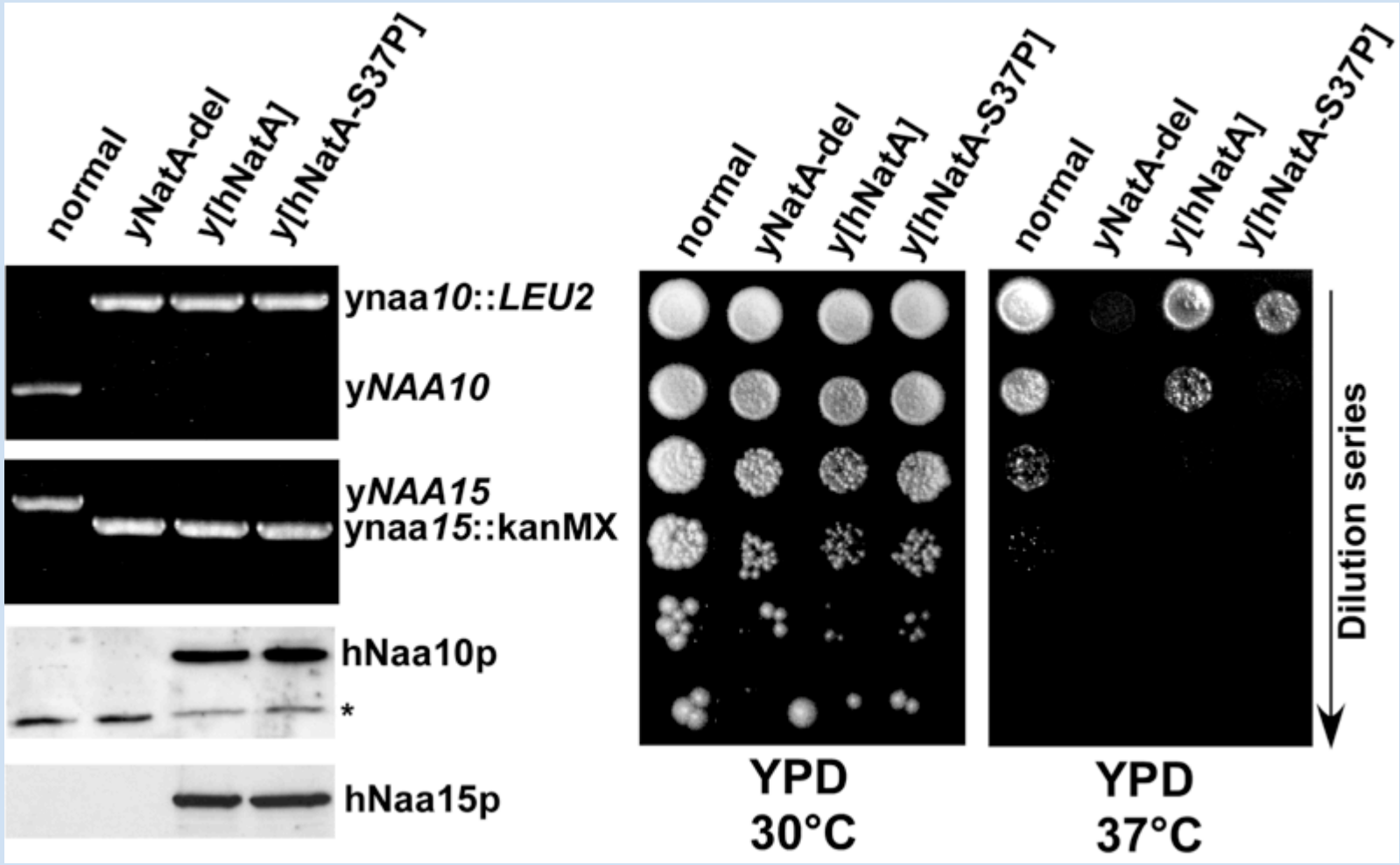


Simulated structure of S37P mutant

Max Doerfel

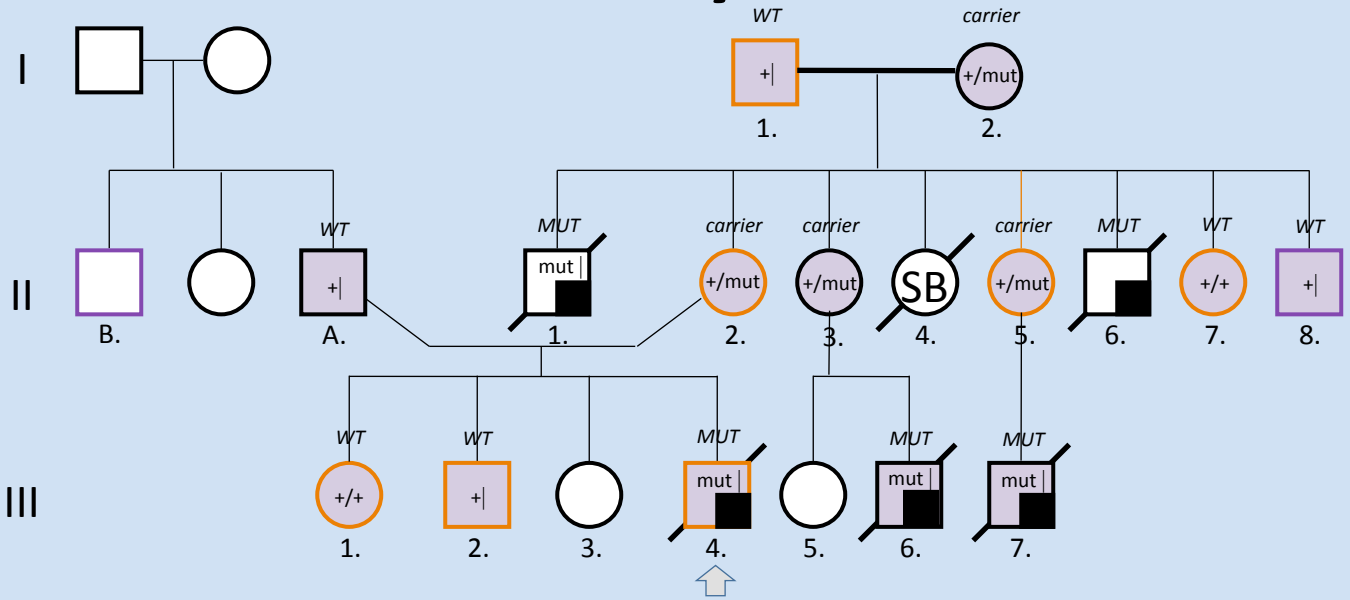
Yiyang Wu

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



Unpublished data from Thomas Arnesen, do not further distribute.

Proteomics Analysis of EBV-transformed cell lines from family members



= male ♂

= female ♀

= FFPE DNA (for patient III.7.) or DNA from blood available (and for some of them: EBV transformed cell lines available + skin fibroblast of patient III.6.)

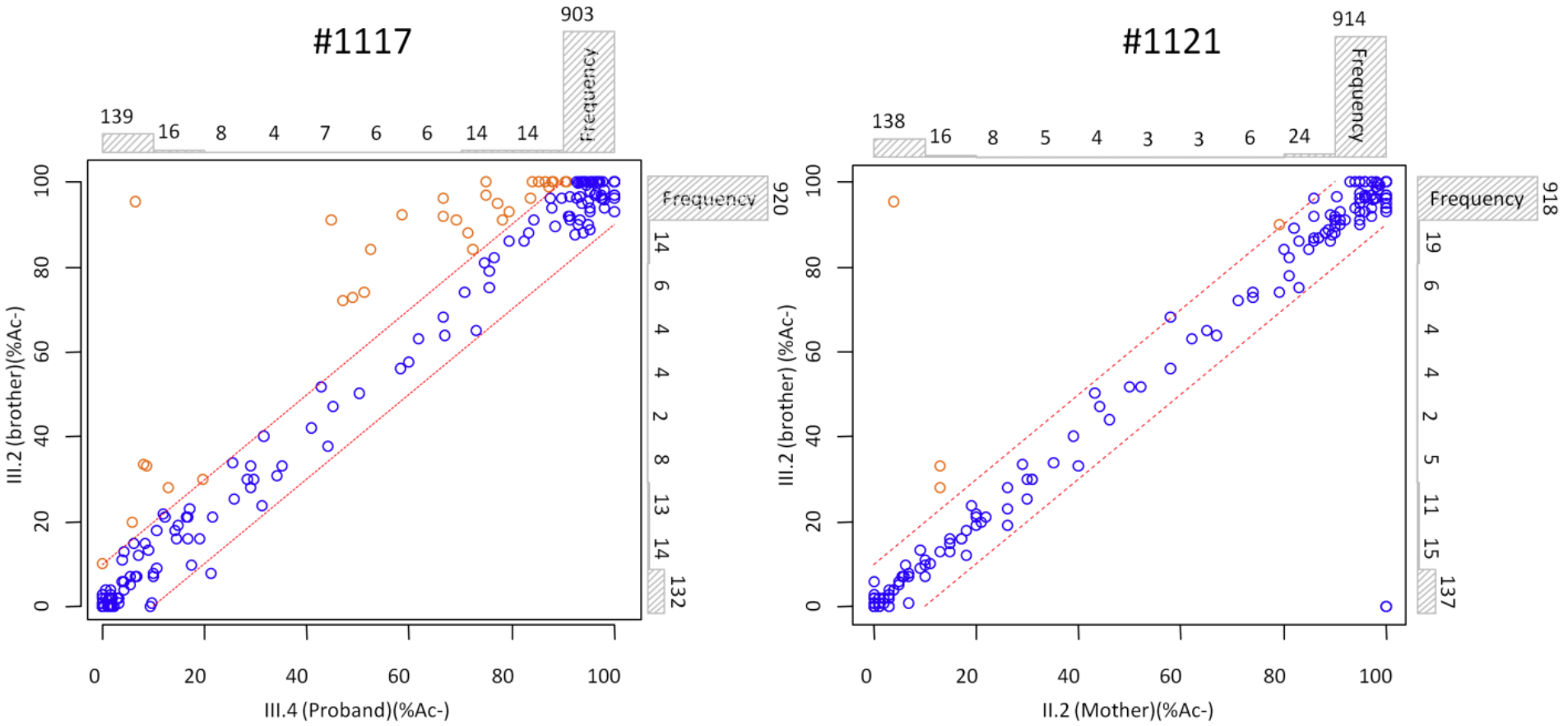
SB = stillborn

= proband

= patient samples analyzed by N-terminal COFRADIC analyses **(#1 to #5)**

= patient samples prepared for N-terminal COFRADIC analyses (but still to be analyzed) **(#8 and #9)**

III.4. proband hemizygous, mutant (89323) **(#1a)** (#1b)
II.2. mother of proband, carrier (89324) **(#2)**
II.A. married-in father of proband, WT(89325)
III.2. brother of proband, WT(90526) **(#3)**
III.1. sister of proband, WT (90527) **(#4)**
I.2. grandmother of proband, carrier (90528)
I.1. married-in grandfather of proband, WT(90529) **(#5)**
II.7. aunt of proband, WT (90530) **(#6)**
II.3. aunt of proband, carrier (90531)
II.B. married-in uncle of proband, WT(90532) (#8)
II.8. uncle of proband, WT(90688) (#9)
II.5. aunt of proband, carrier with deceased boy (90797) **(#7)**

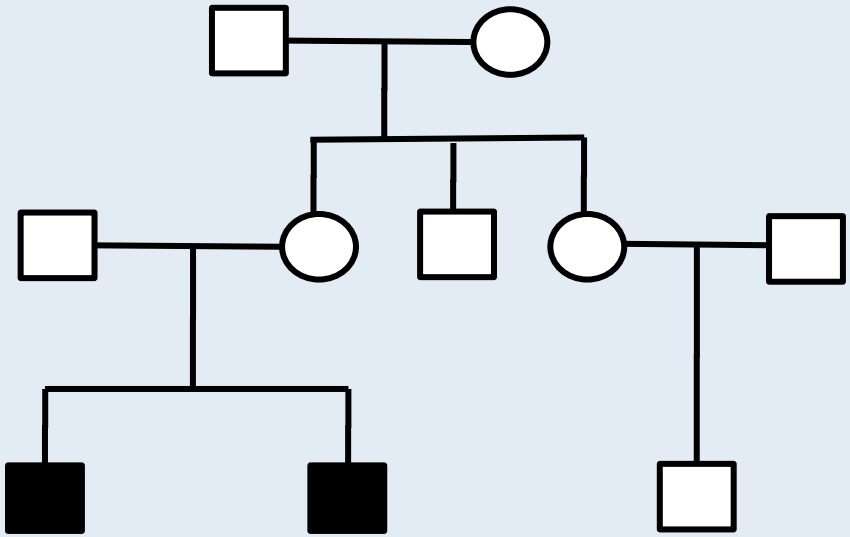


Scatterplots displaying the correlation of the degrees of Nα-acetylation when comparing a control (brother WT)(in this case Y-axis) and the proband or mother(carrier) (Y-axis) N-terminome datasets. The N-termini displaying a significant variation in the degree of Nα-acetylation (see above) are highlighted in orange.

Results from EBV-transformed lymphocytes

AAAEEEDGGPEGPNR	66	86	92	91	87	84	91	Q99942	RNF5_HUMAN	E3 ubiquitin-protein ligase RNF5	Membrane; Multi-pass membrane protein. Mitochondrion membrane. Endoplasmic reticulum membrane.
AADTQVSETLKR	52	80	84	84	80	81	85	Q92616	GCN1L_HUMAN	Translational activator GCN1	
AAESALQVVEKLQAR	58	89	92	92	87	87	92	Q14241	ELOA1_HUMAN	Transcription elongation factor B polypeptide 3	Nucleus.
AVFADLDLR	66	95	96	96	96	95	100	P78346	RPP30_HUMAN	Ribonuclease P protein subunit p30	Nucleus;nucleolus.
MVEKEEAGGGISEEEAAQYDR	69	90	91	94	91	95	96	Q9UBE0	SAE1_HUMAN	SUMO-activating enzyme subunit 1	Nucleus.
MLGAPDESSVR	51	79	74	79	70	72	80	Q7Z4S6	KI21A_HUMAN	Kinesin-like protein KIF21A	Cytoplasm;cytoskeleton.
MLSPEAER	74	97	97	97	96	97	97	Q9NUG6	PDRG1_HUMAN	p53 and DNA damage-regulated protein 1	Cytoplasm.
AAGGGGGSSKASSSSASSAGALESSLDR	72	85	84	85	82	84	87	Q5VT52	RPRD2_HUMAN	Regulation of nuclear pre-mRNA domain-containing protein 2	
GEEANDDKKPTTKFELER	79	91	93	93	94	93	92	Q92989	CLP1_HUMAN	Polyribonucleotide 5'-hydroxyl-kinase Clp1	Nucleus.

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

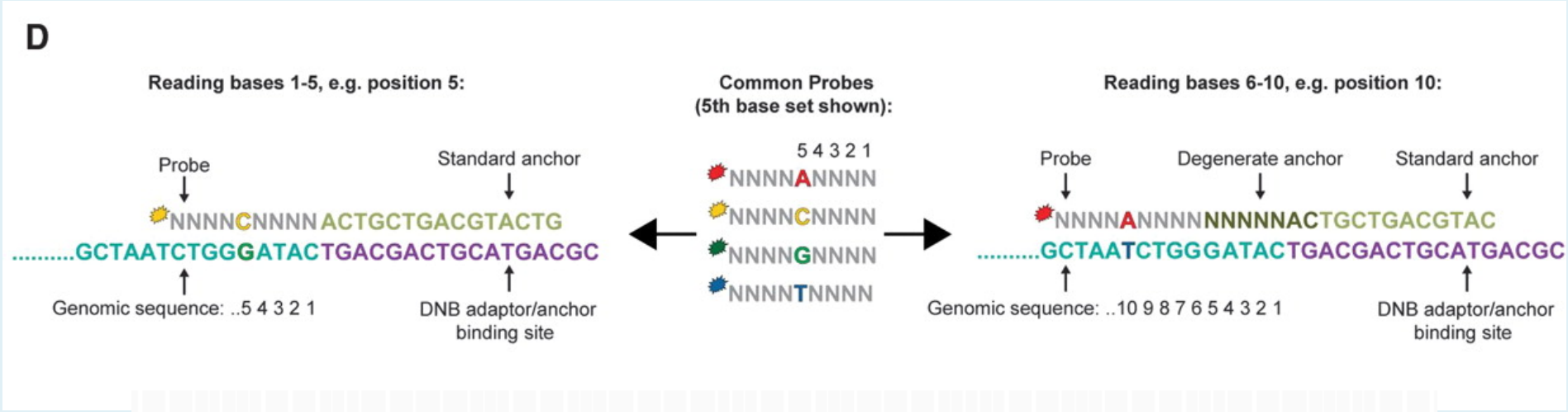
5 years old

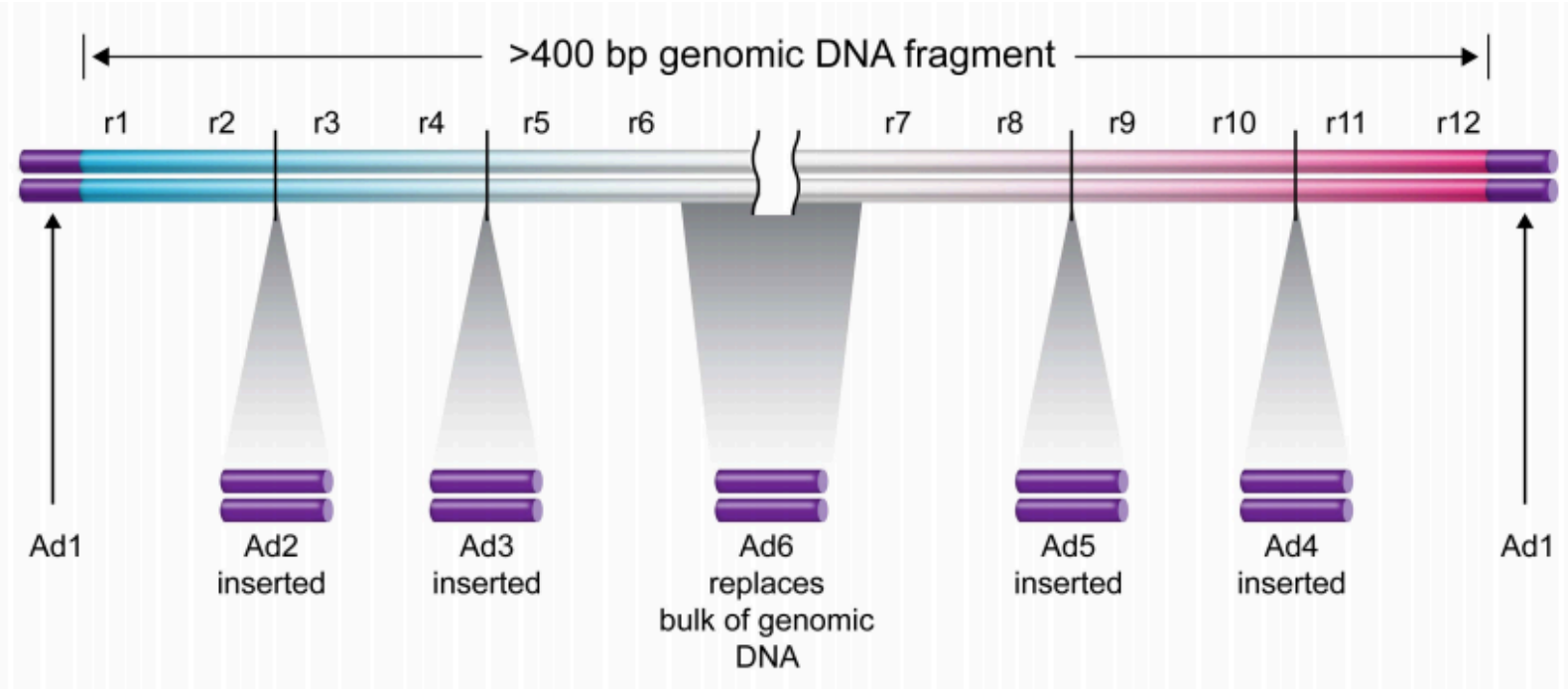
9 years old

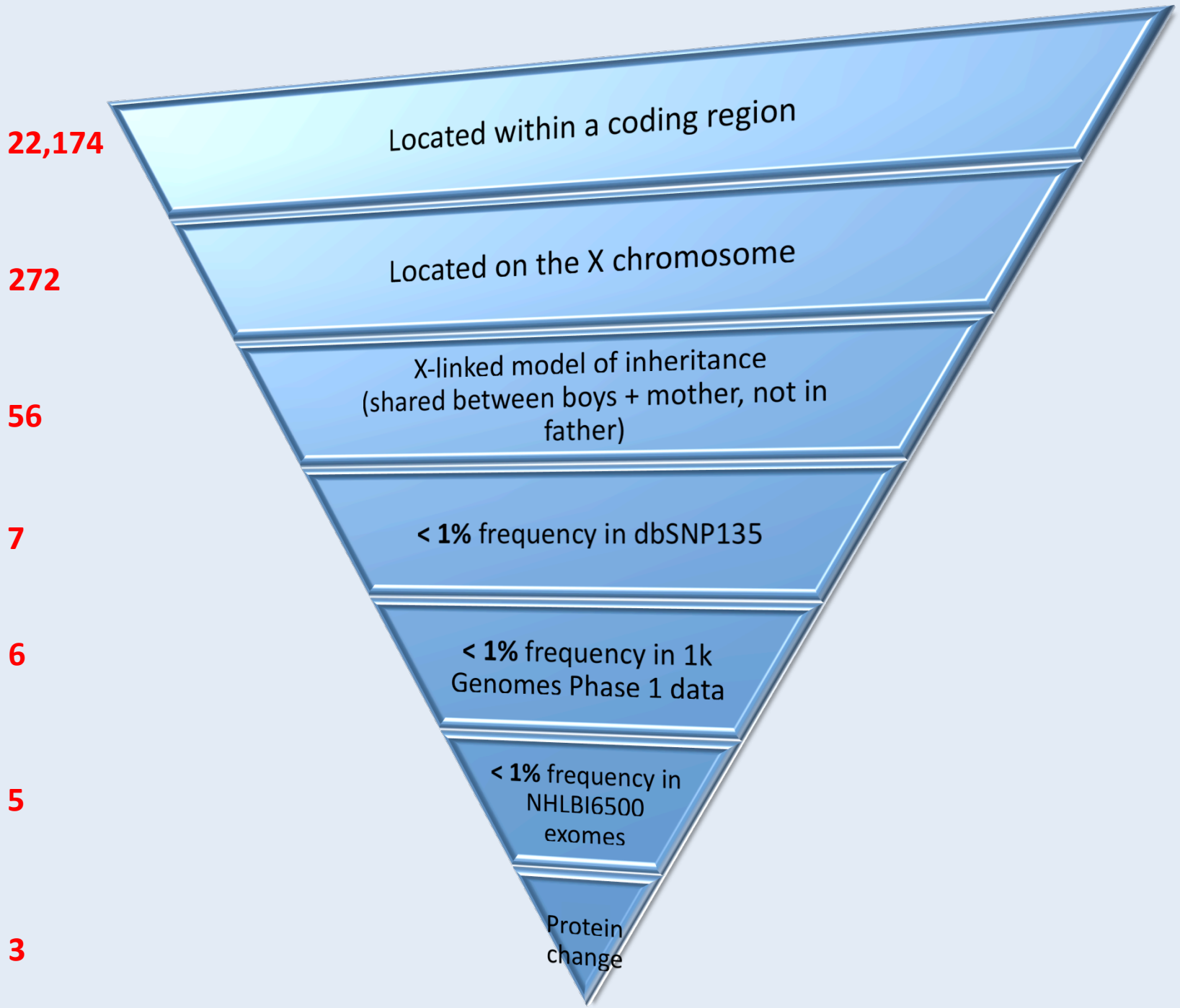
Workup Ongoing for past 10 years

- 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.

Complete Genomics chemistry - combinatorial probe anchor ligation (cPAL)






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

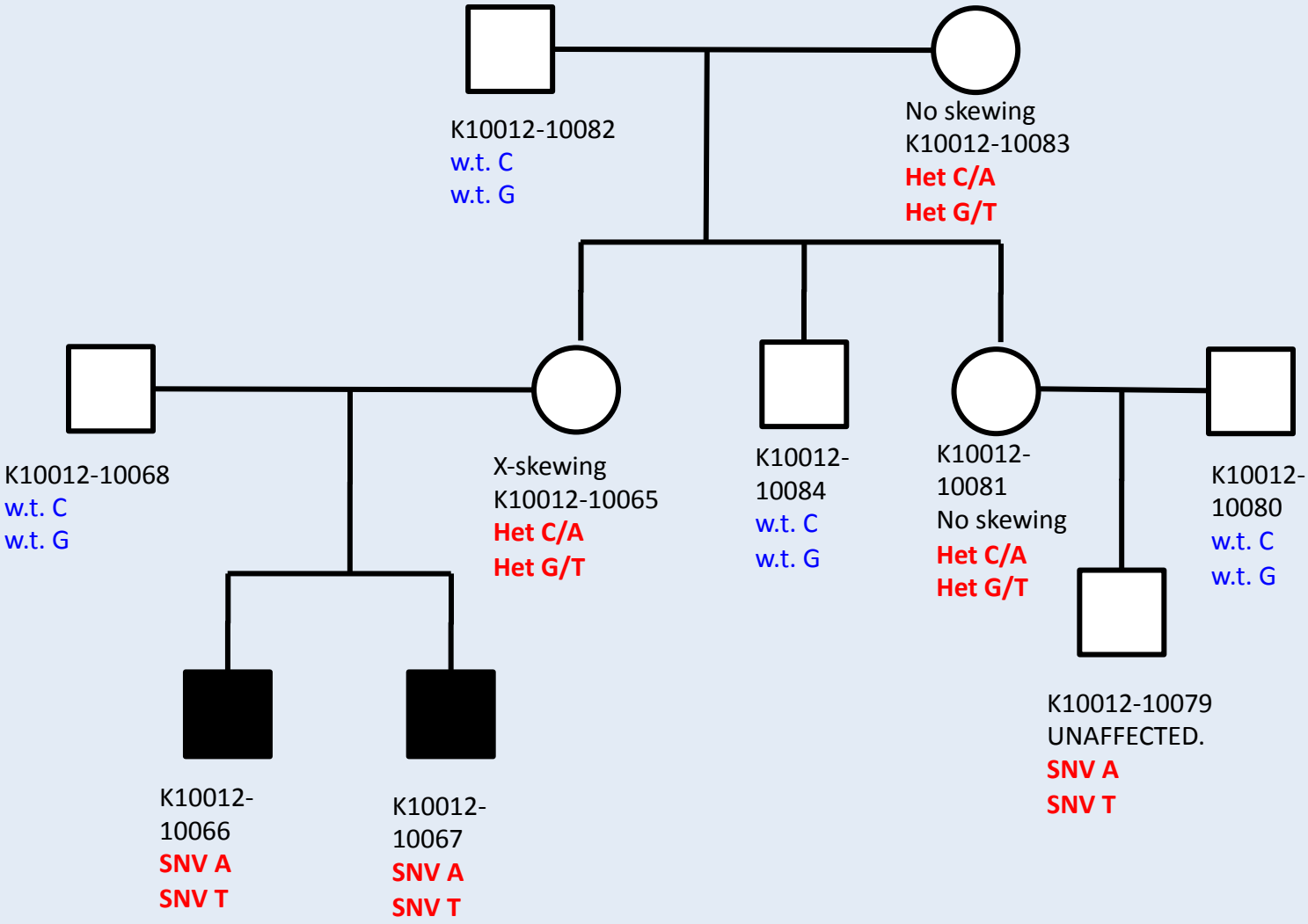
Mutations in the *ZNF41* Gene Are Associated with Cognitive Deficits: Identification of a New Candidate for X-Linked Mental Retardation

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

Toward more comprehensive "personal genomes"

- 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?

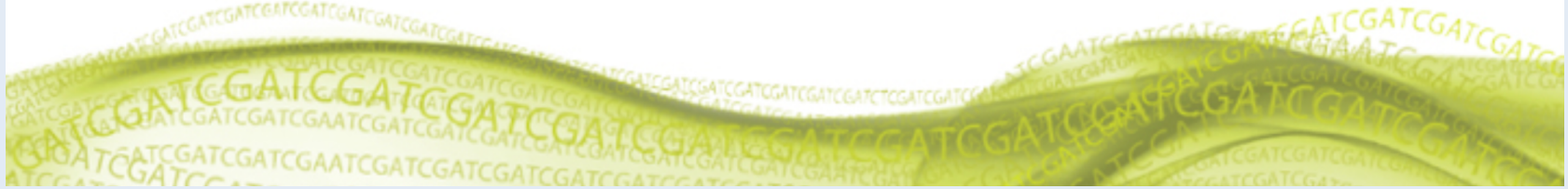
Moving Exome and WGS into a Clinical Setting requires both Analytic and Clinical Validity

- 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

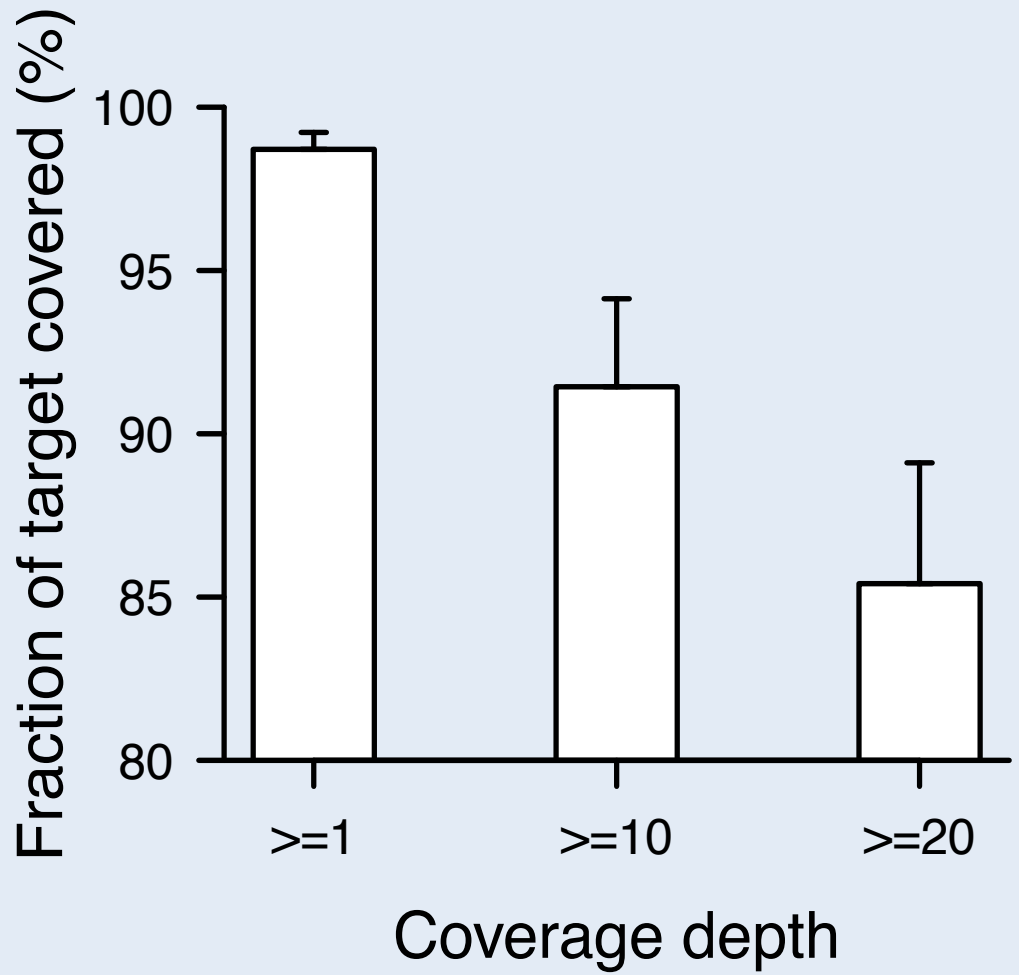
www.everygenome.com
CLIA#: 05D1092911

- ~$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.

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

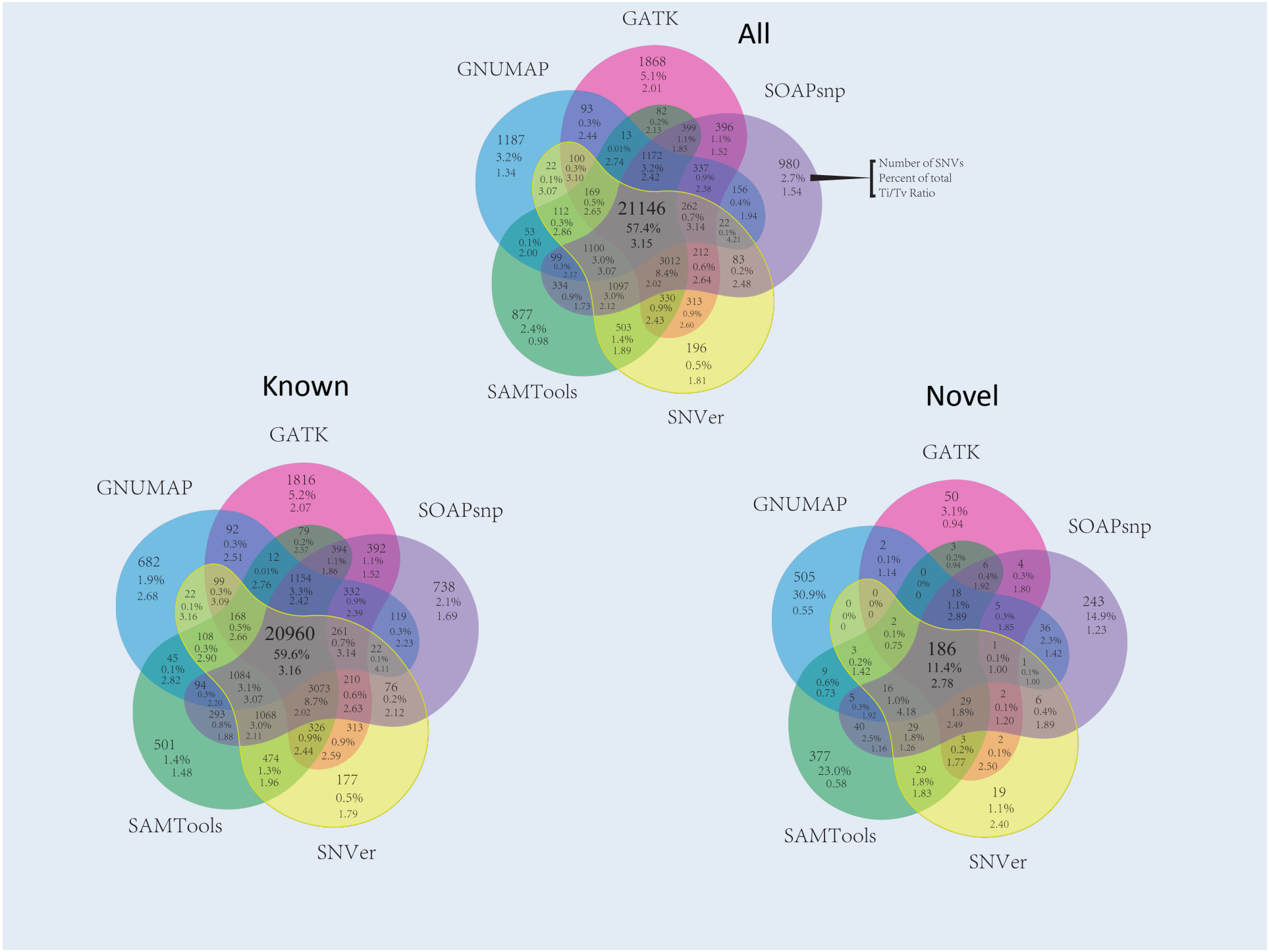
1) BWA - **GATK** (version 1.5) with recommended parameters (GATK IndelRealigner, base quality scores were re-calibrated by GATK Table Recalibration tool. Genotypes called by GATK UnifiedGenotyper. For SNVs and indels.

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.

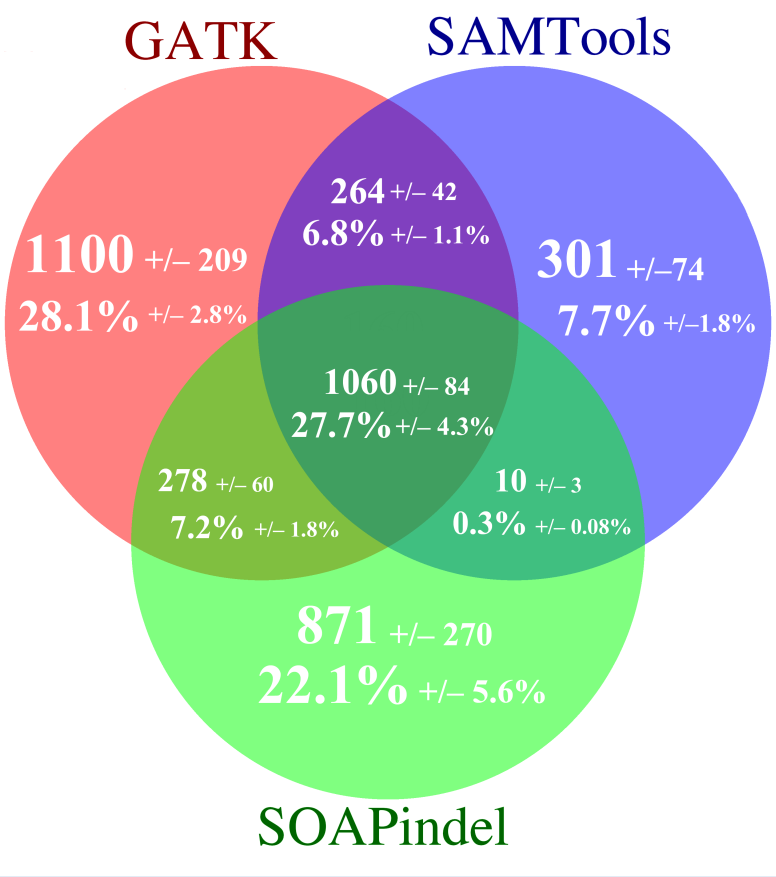
4) GNUMAP-SNP (probabilistic Pair-Hidden Markov which effectively accounts for uncertainty in the read calls as well as read mapping in an unbiased fashion), for SNVs only.

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

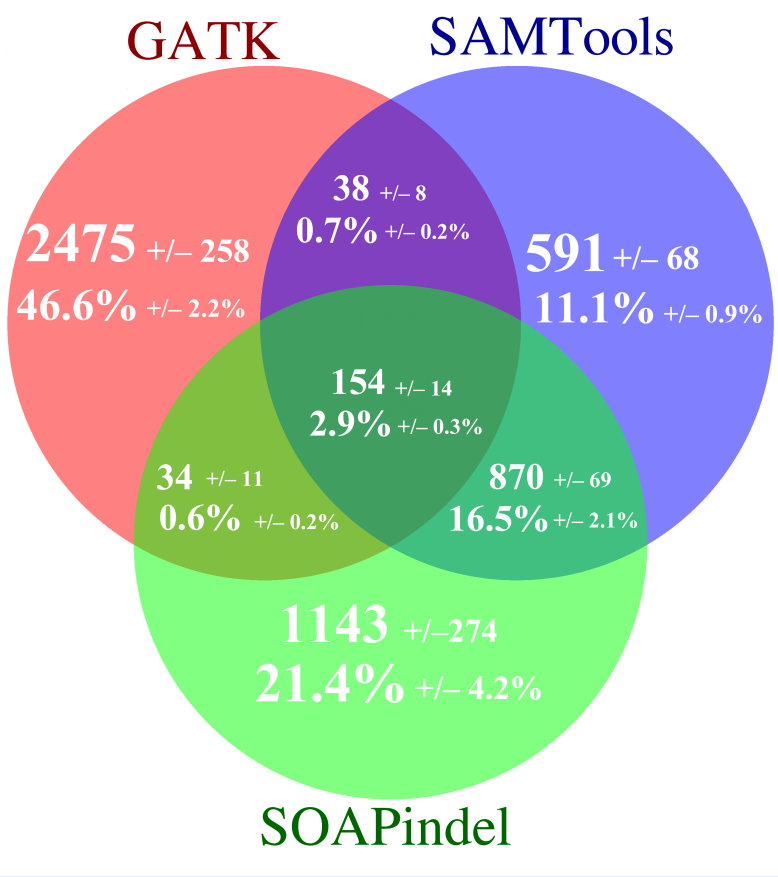

All
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1868 5.1% 2.01
1187 3.2% 1.34
980 2.7% 1.54
93 0.3% 2.44
82 0.2% 2.13
399 1.1% 1.85
396 1.1% 1.52
13 0.01% 2.74
1172 3.2% 2.42
337 0.9% 2.38
156 0.4% 1.94
22 0.1% 3.07
100 0.3% 3.10
169 0.5% 2.65
112 0.3% 2.86
21146 57.4% 3.15
262 0.7% 3.14
22 0.1% 4.21
53 0.1% 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
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
92 0.3% 2.51
79 0.2% 2.57
394 1.1% 1.86
392 1.1% 1.52
12 0.01% 2.76
1154 3.3% 2.42
332 0.9% 2.39
99 0.3% 3.09
22 0.1% 3.16
168 0.5% 2.66
119 0.3% 2.23
20960 59.6% 3.16
261 0.7% 3.14
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
505 30.9% 0.55
243 14.9% 1.23
2 0.1% 1.14
3 0.2% 0.94
6 0.4% 1.92
4 0.3% 1.80
0 0% 0
18 1.1% 2.89
5 0.3% 1.85
36 2.3% 1.42
0 0% 0
0 0% 0
2 0.1% 0.75
186 11.4% 2.78
1 0.1% 1.00
1 0.1% 1.00
3 0.2% 1.42
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

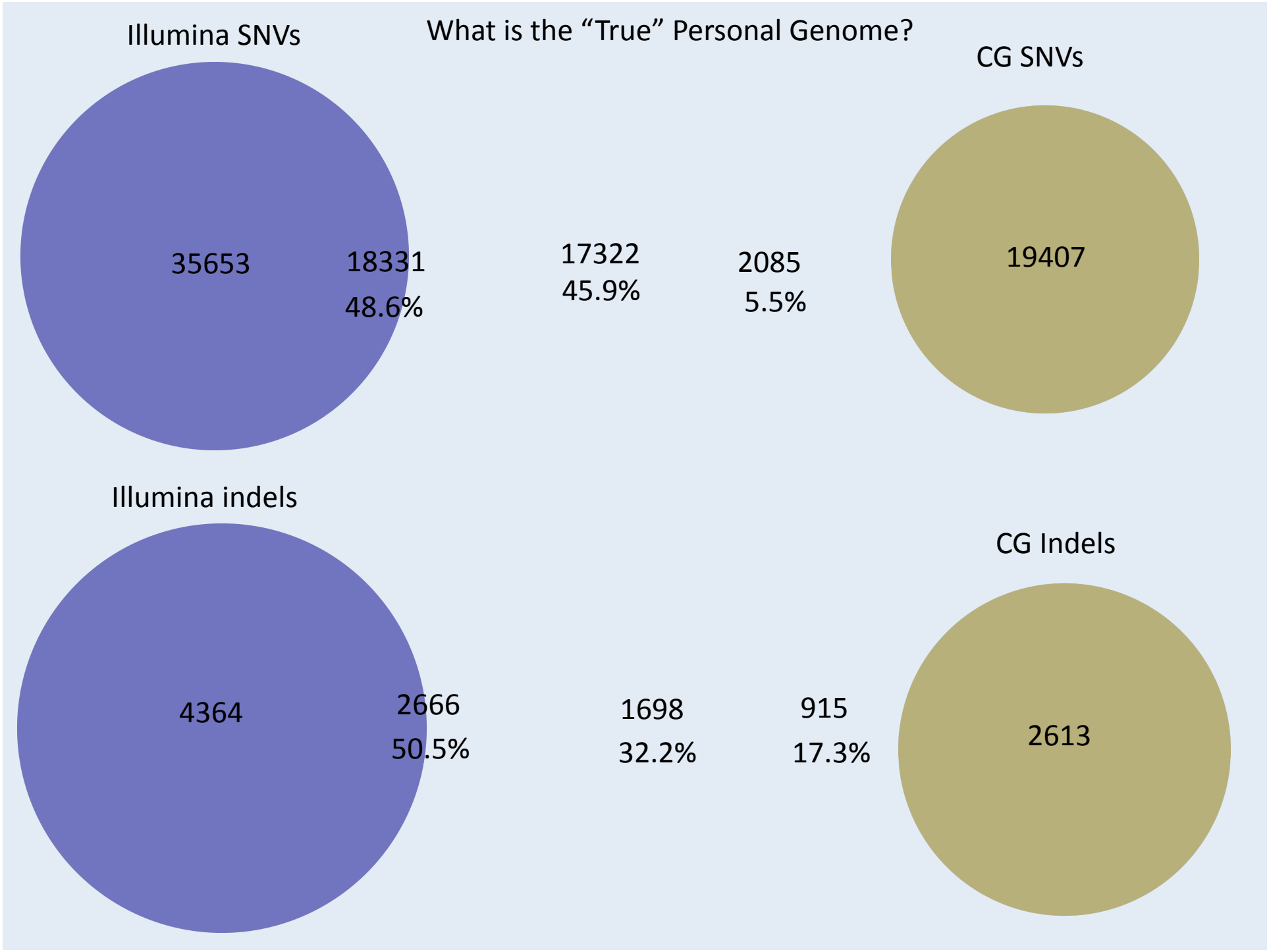


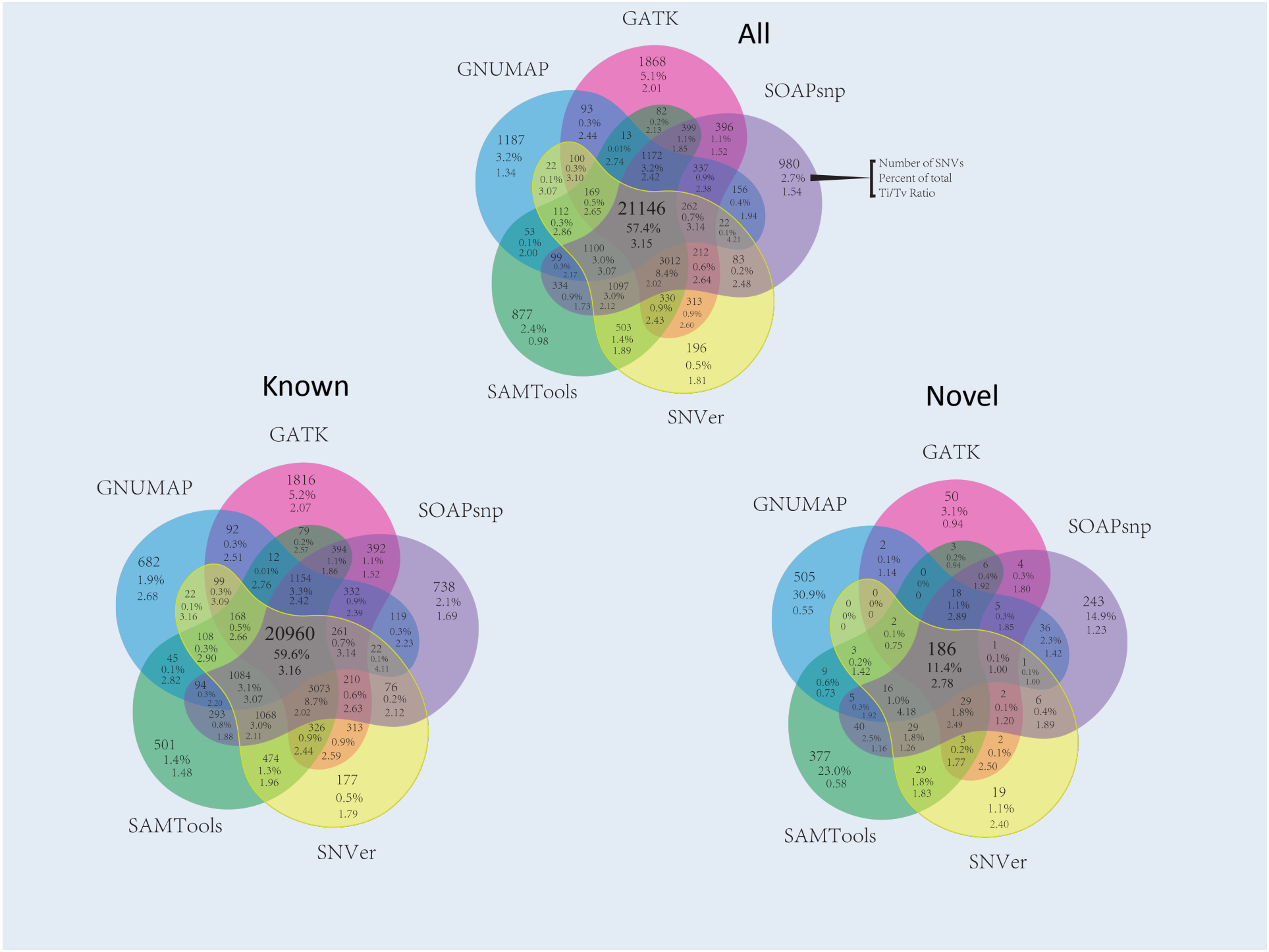
Total mean overlap, plus or minus one standard deviation, observed between three indel calling pipelines: GATK, SOAP-indel, and SAMTools. a) Mean overlap when indel position was the only necessary agreement criterion. **b)** Mean overlap when indel position, base length and base composition were the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

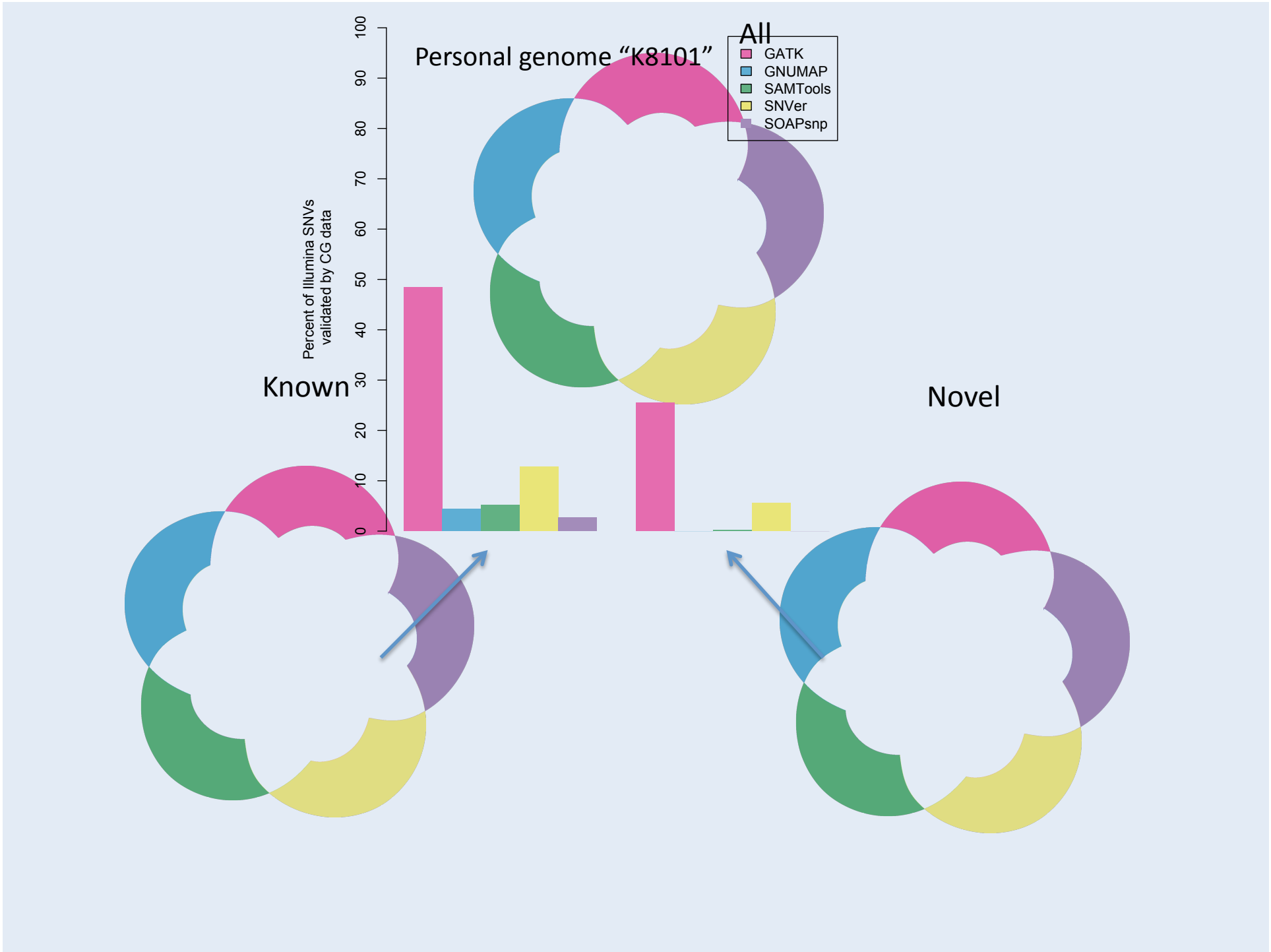
- How reliable are variants that are uniquely called by individual pipelines?

- 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?
Illumina SNVs
CG SNVs
35653
18331
48.6%
17322
45.9%
2085
5.5%
19407
Illumina indels
CG Indels
4364
2666
50.5%
1698
32.2%
915
17.3%
2613

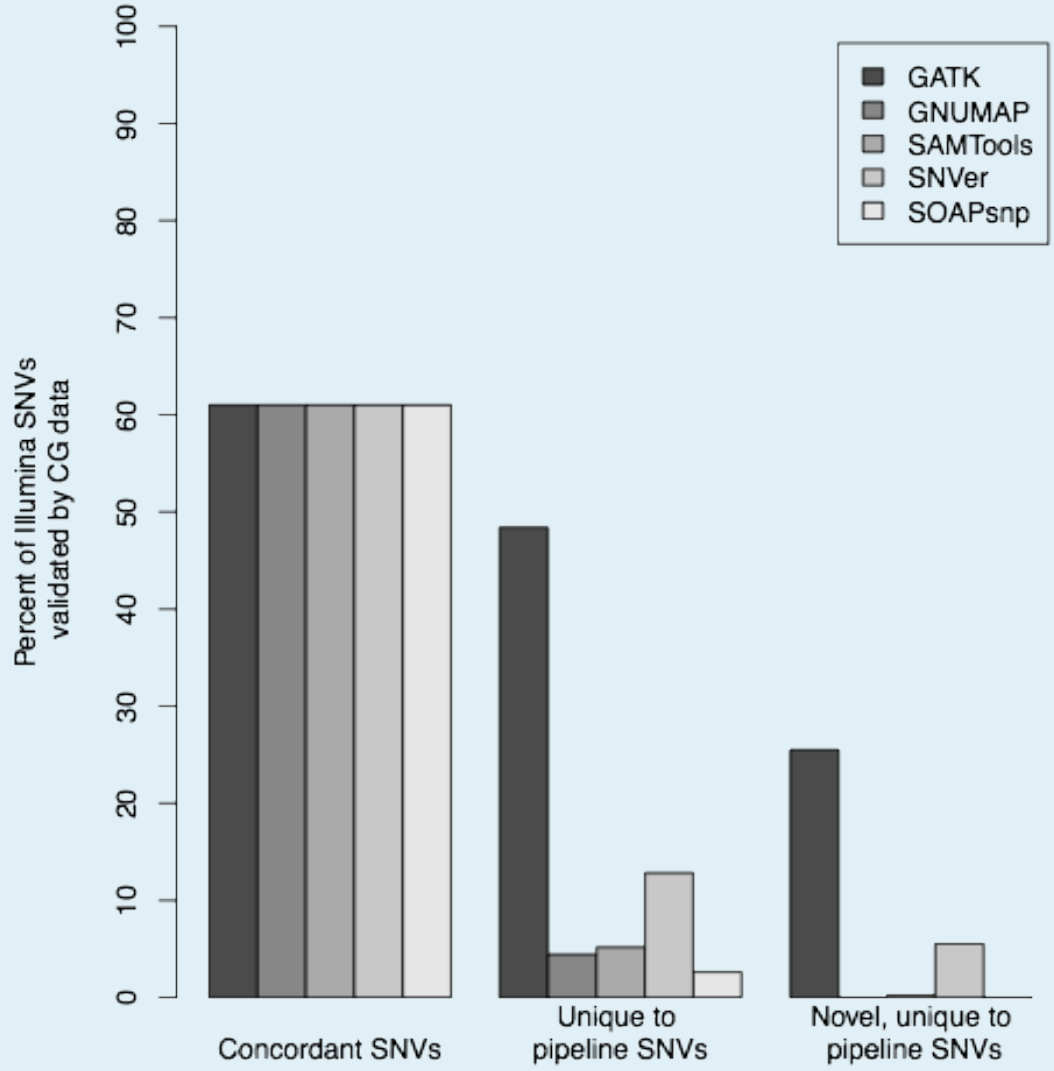

All
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1868 5.1% 2.01
1187 3.2% 1.34
980 2.7% 1.54
93 0.3% 2.44
82 0.2% 2.13
399 1.1% 1.85
396 1.1% 1.52
13 0.01% 2.74
1172 3.2% 2.42
337 0.9% 2.38
156 0.4% 1.94
22 0.1% 3.07
100 0.3% 3.10
169 0.5% 2.65
112 0.3% 2.86
21146 57.4% 3.15
262 0.7% 3.14
22 0.1% 4.21
53 0.1% 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
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
92 0.3% 2.51
79 0.2% 2.57
394 1.1% 1.86
392 1.1% 1.52
12 0.01% 2.76
1154 3.3% 2.42
332 0.9% 2.39
99 0.3% 3.09
22 0.1% 3.16
168 0.5% 2.66
119 0.3% 2.23
20960 59.6% 3.16
261 0.7% 3.14
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
505 30.9% 0.55
243 14.9% 1.23
2 0.1% 1.14
3 0.2% 0.94
6 0.4% 1.92
4 0.3% 1.80
0 0% 0
18 1.1% 2.89
5 0.3% 1.85
36 2.3% 1.42
0 0% 0
0 0% 0
2 0.1% 0.75
186 11.4% 2.78
1 0.1% 1.00
1 0.1% 1.00
3 0.2% 1.42
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


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

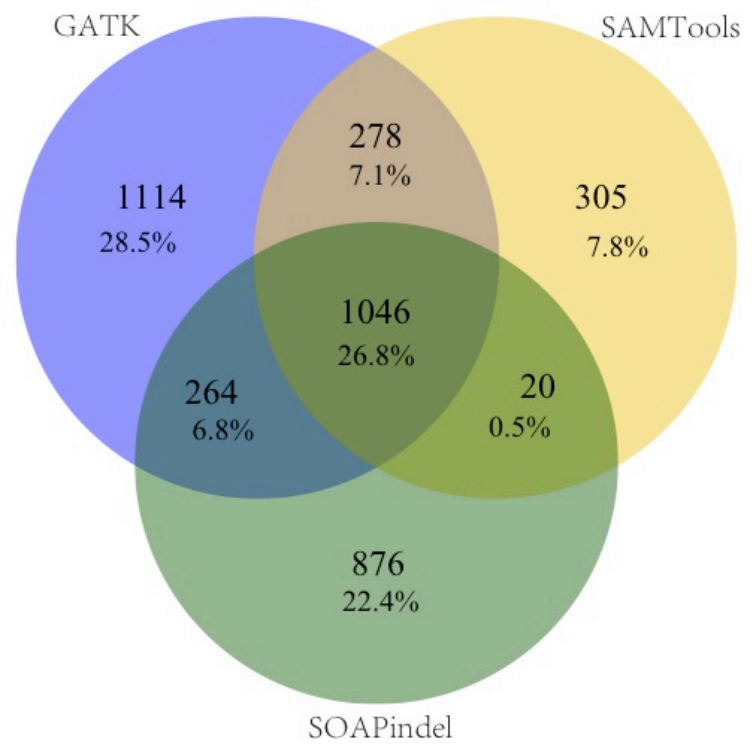
Higher Validation of SNVs with the BWA-GATK pipeline

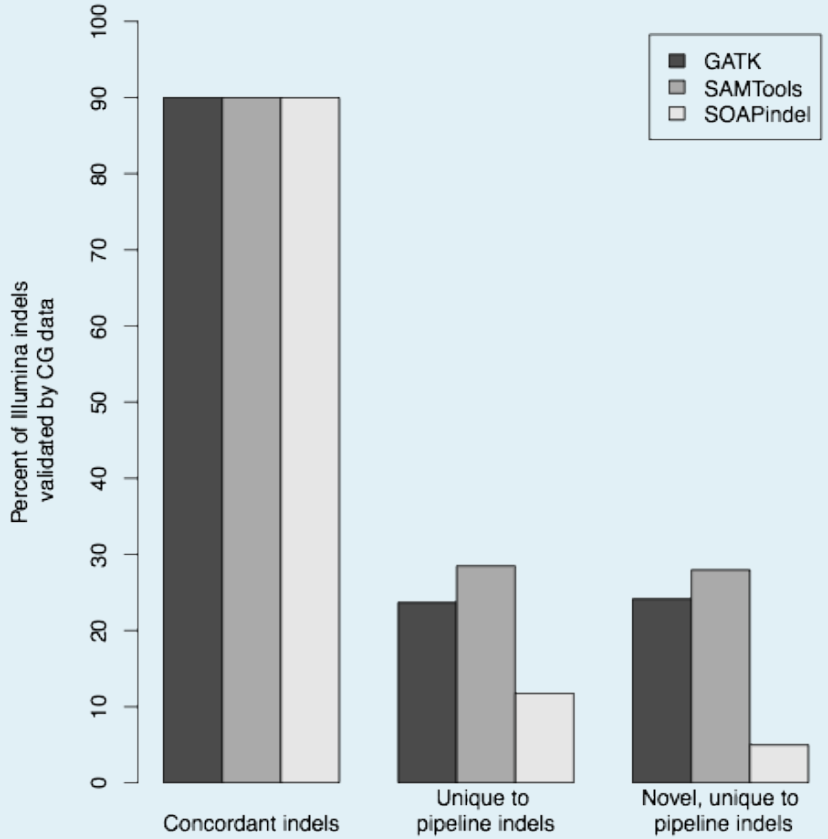
- Reveals higher validation rate of unique-to-pipeline variants, as well as uniquely discovered novel variants, for the variants called by BWA-GATK, in comparison to the other 4 pipelines (including SOAP).

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



Validating Indels with Complete Genomics Data for the 3 pipelines





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

Dick McCombie
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