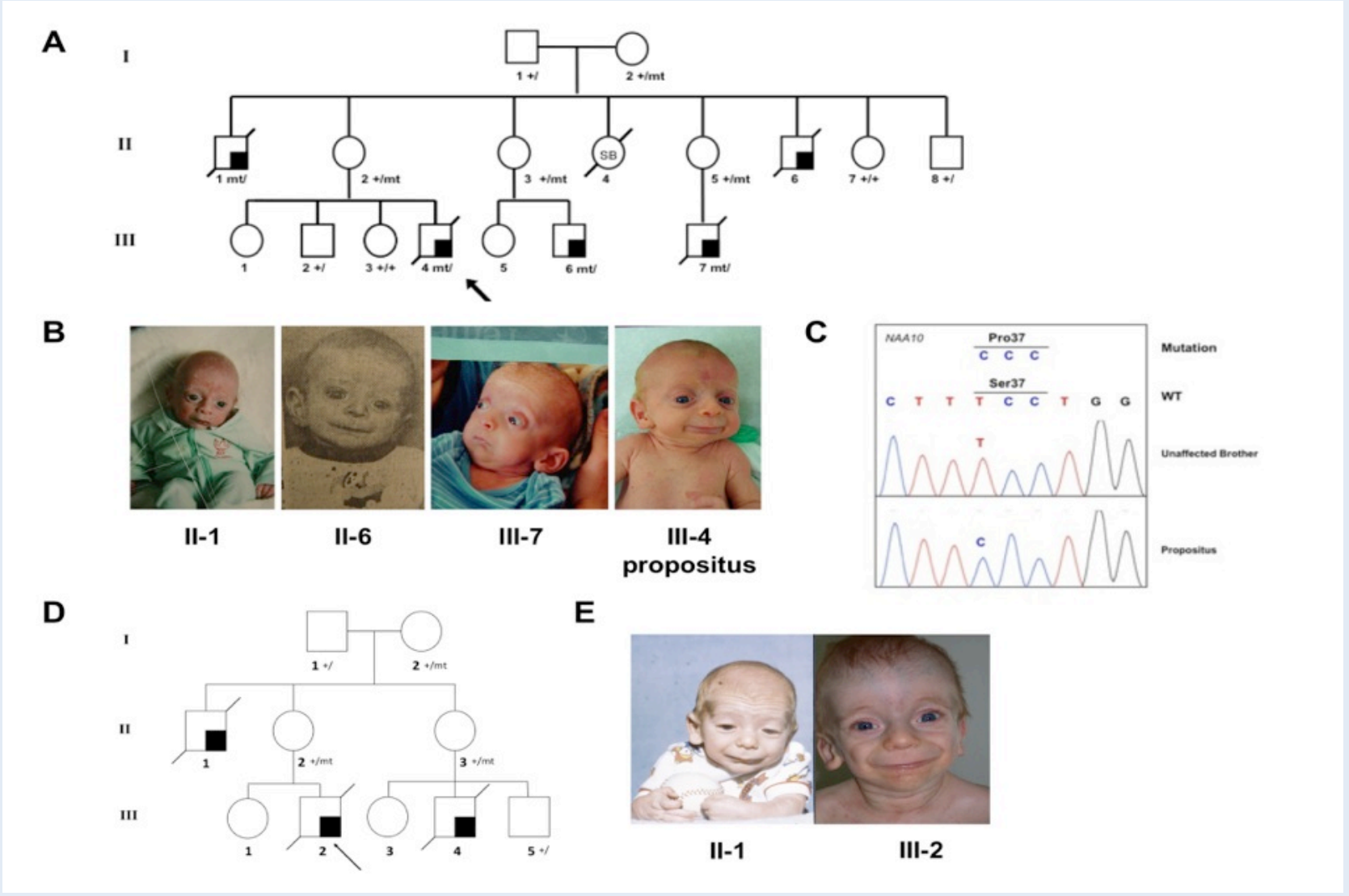
Whole Genome Sequencing Analysis of a severe Idiopathic Intellectual Disability Syndrome.

or

"limitations of theory may not be revealed when the facts are too few" - Knox 1958

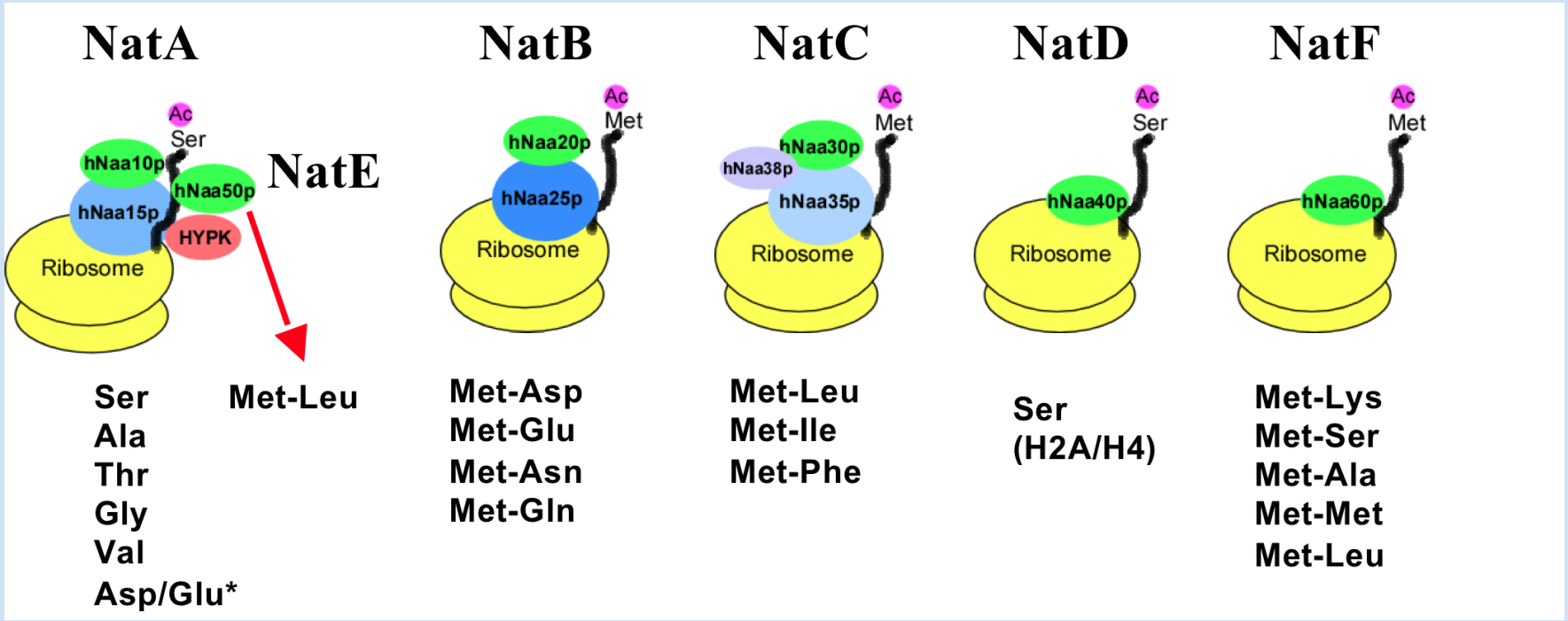
Gholson J. Lyon, M.D. Ph.D.

Ancestry Matters! - Ogden Syndrome



The mutation in NAA10 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

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.	

Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study

Anita Rauch, Dagmar Wieczorek*, Elisabeth Graf*, Thomas Wieland*, Sabine Endele, Thomas Schwarzmayr, Beate Albrecht, Deborah Bartholdi, Jasmin Beygo, Nataliya Di Donato, Andreas Dufke, Kirsten Cremer, Maja Hempel, Denise Horn, Juliane Hoyer, Pascal Joset, Albrecht Röpke, Ute Moog, Angelika Riess, Christian T Thiel, Andreas Tzschach, Antje Wiesener, Eva Wohlleber, Christiane Zweier, Arif B Ekici, Alexander M Zink, Andreas Rump, Christa Meisinger, Harald Grallert, Heinrich Sticht, Annette Schenck, Hartmut Engels, Gudrun Rappold, Evelin Schröck, Peter Wieacker, Olaf Riess, Thomas Meitinger, André Reis†, Tim M Strom†*

Lancet 2012; 380: 1674–82

Published Online
September 27, 2012
http://dx.doi.org/10.1016/S0140-6736(12)61480-9

	Sex	Gene	Online Mendelian Inheritance in Man reference	Type	Genomic change	Protein change	Haploinsufficiency index (%)	PolyPhen2 category (score)
E10-0275	Male	*IQSEC2*	309530	Nonsense	X chromosome: g.53277315G→A	Arg855*1	9·8%	..
BO17/09	Female	*MECP2*	312750	Frameshift	X chromosome: g.153296093_153296115del	Pro401Argfs*8	24·3%	..
ZH58769	Male	*NAA10*	300855	Missense	X chromosome: g.153197564G→A	Arg116Trp	..	Benign (0·233)*
ER52725	Female	*SATB2*	608148	Missense	Chromosome 2: g.200213455A→C	Val381Gly	4·3%	Probably damaging (1)
ER8490	Male	*SCN2A*	613721	Frameshift	Chromosome 2: g.166179821_166179822delCT	Leu611Valfs*35	12·7%	..
MS111684	Male	*SCN2A*	613721	Frameshift	Chromosome 2: g.166172100_166172101insA	Asn503Lysfs*19	12·7%	..
ZH60991	Female	*SCN2A*	613721	Missense	Chromosome 2: g.166201311C→T	Arg937Cys	12·7%	Probably damaging (1)
ER12988	Female	*SCN8A*	614558, 614306	Missense	Chromosome 12: g.52200120G→A	Arg1617Gln	14·5%	Probably damaging (1)
BO22/10	Female	*SETBP1*	269150	Nonsense	Chromosome 18: g.42531079A→T	Lys592*	9·9%	..
PL111540	Male	*SLC2A1*	606777, 612126	Missense	Chromosome 1: g.43396356G→A	Arg153Cys	24·1%	Probably damaging (1)
ES07E0046	Female	*STXBP1*	612164	Missense	Chromosome 9: g.130422363G→C	Ala101Pro	8·7%	Possibly damaging (0·860)†
MR-NET001	Female	*STXBP1*	612164	Splice	Chromosome 9: g.130422308delC	Aberrant splicing predicted	8·7%	..
P4276	Female	*STXBP1*	612164	Missense	Chromosome 9: g.130420659G→A	Glu59Lys	8·7%	Probably damaging (0·994)
BO14/09	Female	*SYNGAP1*	612621	Frameshift	Chromosome 6: g.33410958_33410959insT	Thr878Aspfs*60	23·6%	..
ER53899	Male	*SYNGAP1*	612621	Frameshift	Chromosome 6: g.33405934_33405935delAA	Lys418Argfs*54	23·6%	..
TUBA080997	Female	*TCF4*	610954	Missense	Chromosome 18: g.53070725G→A	Ser110Leu	1·9%	Benign (0·073)

*Molecular modelling suggests that the bulky Trp116 side-chain interferes with coenzyme A binding, thereby affecting enzymatic activity. †The crystal structure indicates that Ala101 is located at the N-terminal region of a sheet structure. The ϕ angle of −158° is not possible for proline, and molecular modelling suggests that this mutation destabilises the structure and probably also hampers ligand binding (appendix).

Table 3: **Missense, nonsense, frameshift, and splice site de-novo variants in genes associated with intellectual disability in each patient–parent trio**

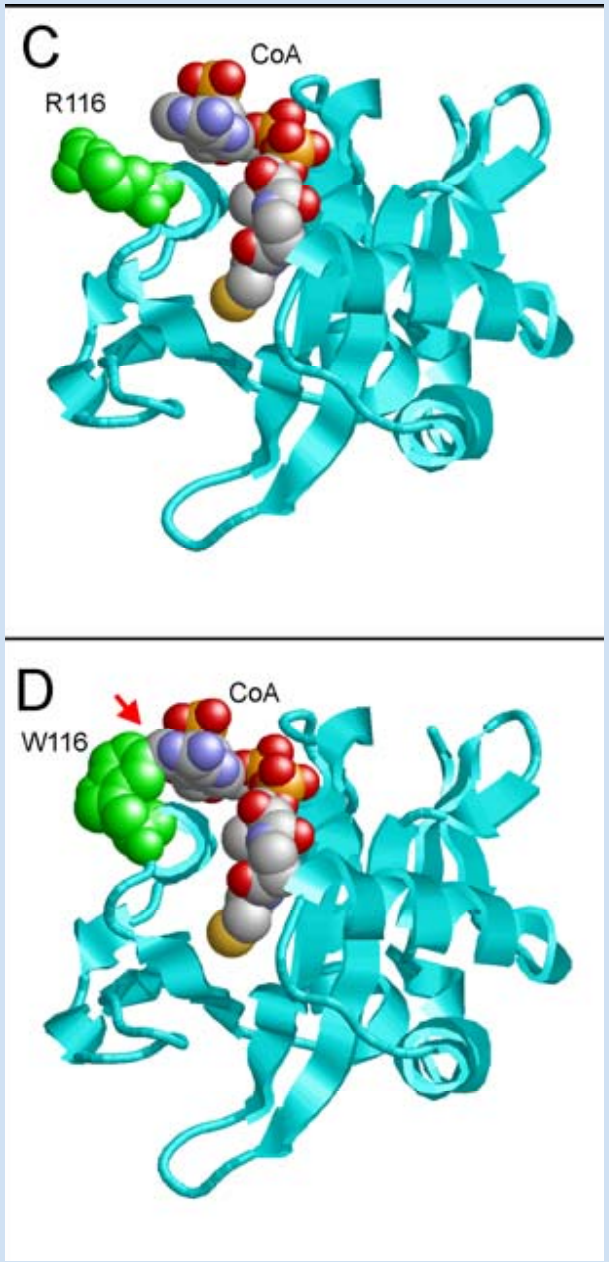
Patient ID	Family history	Gender	Ethnicity	Prenatal anomalies	Weeks of gestation	Birth weight [g/SD]	OFC at birth [cm/SD]	Age at last investi-gation [y]	Height [cm/SD]	OFC [cm/SD]	IQ	Sitting age [mo]	Walking age [mo]	Age at first words [mo]
ZH58769	distantly ID	m	Swiss	-	41	3500/-0·53	n.r.	5 11/12	108/-1·92	50/-1·7	<20	18	/	/

Patient ID	Current speech	Behavioural anomalies	Seizures	Age of onset (seizures) [mo]	MRI (at age [mo])	Minor anomalies	Others	Age of mother at birth [y]	Age of father at birth [y]	Diagnostic tests before exome (besides karyotype, array, FMR1)
ZH58769	-	hyperactivity, hand biting	-	-	enlarged ventricles, reduced periventricular volume, gliotic changes [66]	large ears, small hands/ feet, dias-thema, high palate	truncal hypotonia, hypertonia of extrem.	32	36	AS, mat. X-inact.

“Although two variants were classified as benign, the phenotypic features of the affected patient and predictions based on protein structure suggest that the NAA10 variant has a causal effect (appendix).”

“The children in our study did not have cardiac anomalies associated with Ogden syndrome, the typical facial features of Pitt-Hopkins syndrome, or the ataxia described in children with SCN8A mutations.”

“The R116W mutation of NAA10 is located in the acetyltransferase domain, which was modeled using the structure of the homologous ARD1 acetylase domain from Sulfolobus as template (PDB code: 2X7B).”



C) Location of R116 (green) in the acetyltransferase domain (cyan) of NAA10. The cofactor coenzyme A (CoA) is shown in space-filled presentation and colored according to the atom type. D) In the R116W mutant, W116 preferentially adopts a sidechain orientation that interferes with CoA binding. The steric clashes between W116 and CoA are indicated by a red arrow. These clashes are expected to hamper CoA binding and to reduce the enzymatic activity.

EJHG Open

European Journal of Human Genetics (2013) 00, 1–11
© 2013 Macmillan Publishers Limited All rights reserved 1018-4813/13
www.nature.com/ejhg

ARTICLE

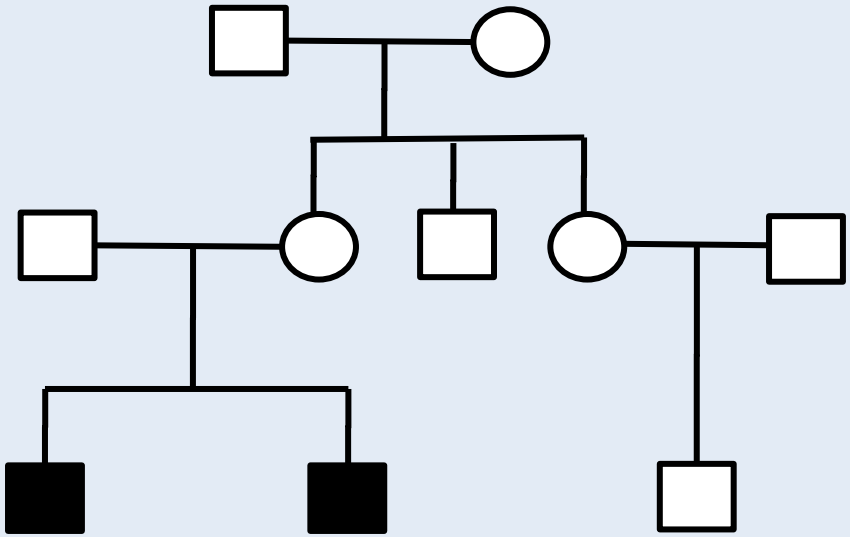
New population-based exome data are questioning the pathogenicity of previously cardiomyopathy-associated genetic variants

Charlotte Andreasen[1,2,5], Jonas B Nielsen[1,2,5], Lena Refsgaard[1,2], Anders G Holst[1,2], Alex H Christensen[1,2], Laura Andreasen[1,2], Ahmad Sajadieh[3], Stig Haunsø[1,2,4], Jesper H Svendsen[1,2,4] and Morten S Olesen*[,1,2]

Cardiomyopathies are a heterogeneous group of diseases with various etiologies. We focused on three genetically determined cardiomyopathies: hypertrophic (HCM), dilated (DCM), and arrhythmogenic right ventricular cardiomyopathy (ARVC). Eighty-four genes have so far been associated with these cardiomyopathies, but the disease-causing effect of reported variants is often dubious. In order to identify possible false-positive variants, we investigated the prevalence of previously reported cardiomyopathy-associated variants in recently published exome data. We searched for reported missense and nonsense variants in the *NHLBI-Go Exome Sequencing Project* (ESP) containing exome data from 6500 individuals. In ESP, we identified 94 variants out of 687 (14%) variants previously associated with HCM, 58 out of 337 (17%) variants associated with DCM, and 38 variants out of 209 (18%) associated with ARVC. These findings correspond to a genotype prevalence of 1:4 for HCM, 1:6 for DCM, and 1:5 for ARVC. PolyPhen-2 predictions were conducted on all previously published cardiomyopathy-associated missense variants. We found significant overrepresentation of variants predicted as being benign among those present in ESP compared with the ones not present. In order to validate our findings, seven variants associated with cardiomyopathy were genotyped in a control population and this revealed frequencies comparable with the ones found in ESP. In conclusion, we identified genotype prevalences up to more than one thousand times higher than expected from the phenotype prevalences in the general population (HCM 1:500, DCM 1:2500, and ARVC 1:5000) and our data suggest that a high number of these variants are not monogenic causes of cardiomyopathy.
European Journal of Human Genetics advance online publication, 9 January 2013; doi:10.1038/ejhg.2012.283

Keywords: cardiomyopathy; exome; next-generation sequencing; HCM; DCM; ARVC

New Syndrome with Dysmorphology, Severe Intellectual Disability, “Autism”, “ADHD”



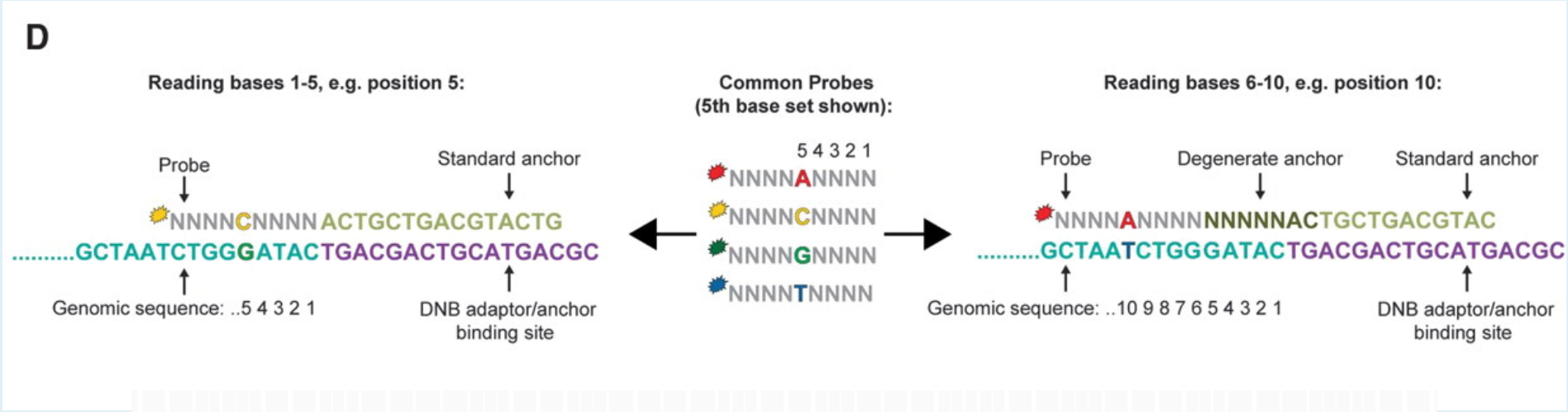
Likely X-linked or Autosomal Recessive, with X-linked being supported by extreme X-skewing in the mother

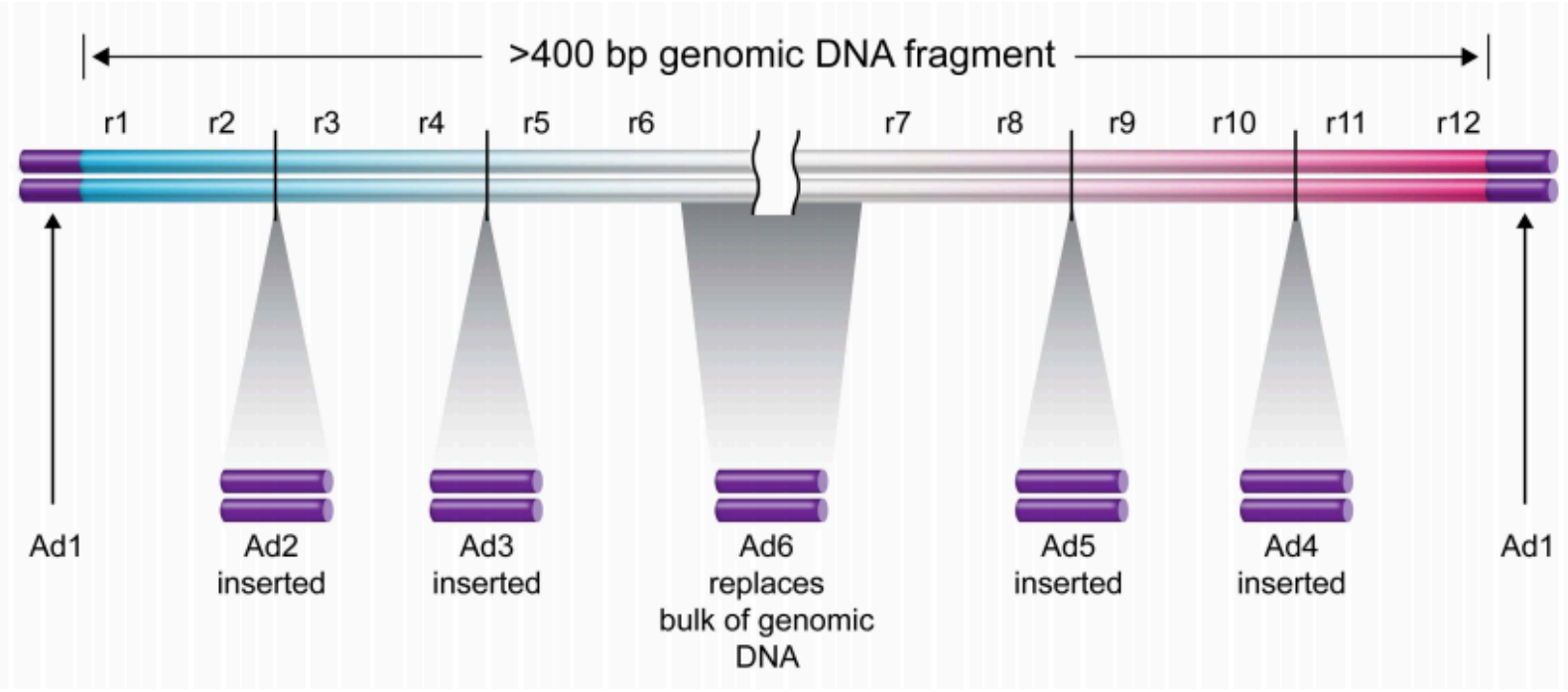
Workup Ongoing for past 10 years

- Numerous genetic tests negative, including negative for Fragile X and many candidate genes.
- No obvious pathogenic CNVs – several microarrays without any definitive result.
- 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

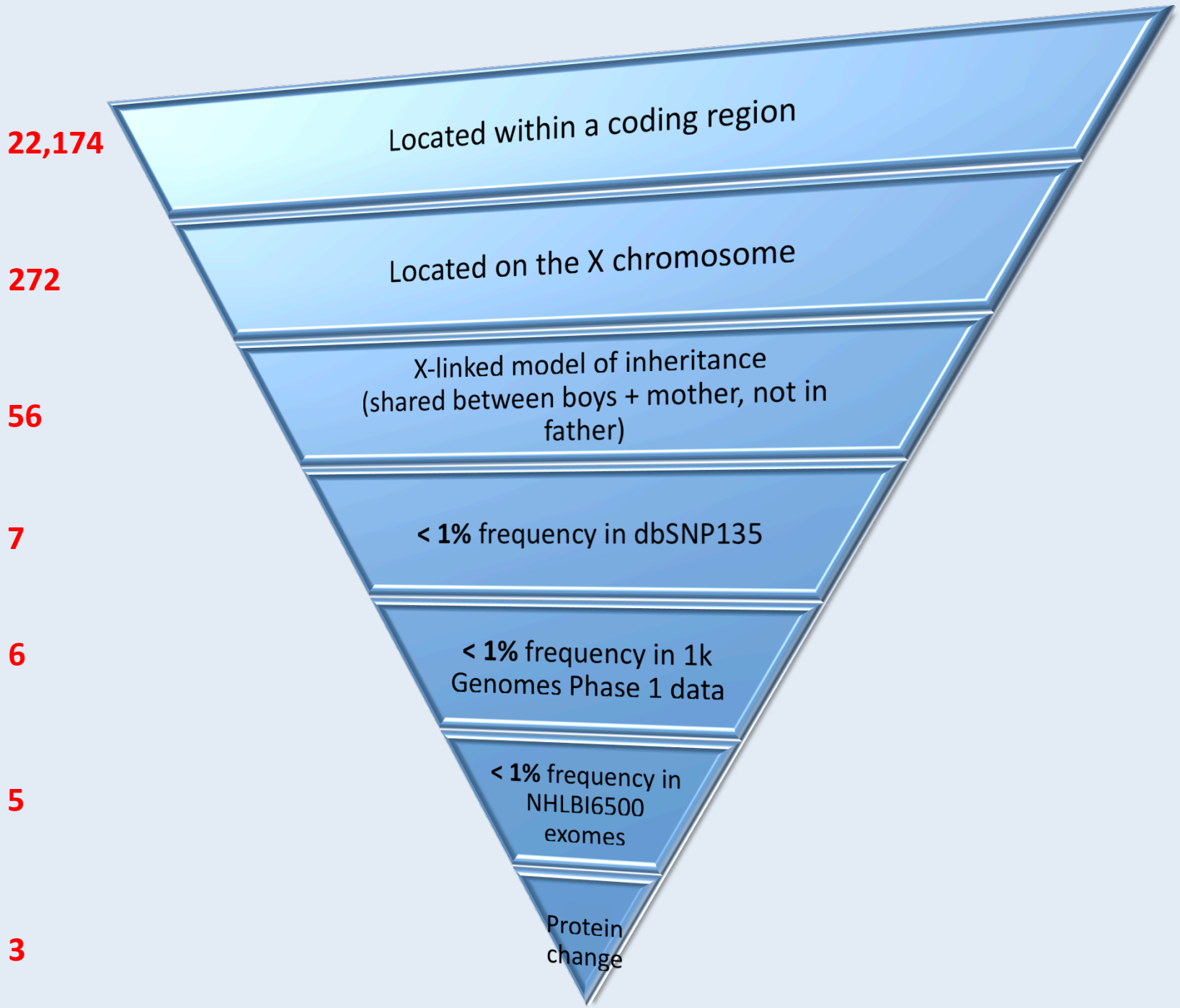
Complete Genomics chemistry - combinatorial probe anchor ligation (cPAL)





Analysis with VAAST, ANNOVAR, and Golden Helix SVS

- No obvious pathogenic CNVs in both brothers.
- No autosomal recessive SNVs or indels in protein-coding regions in both brothers.
- One putative interesting homozygous de novo mutation in both brothers, but validating now with Sanger sequencing. The mechanism of this is uncertain?


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

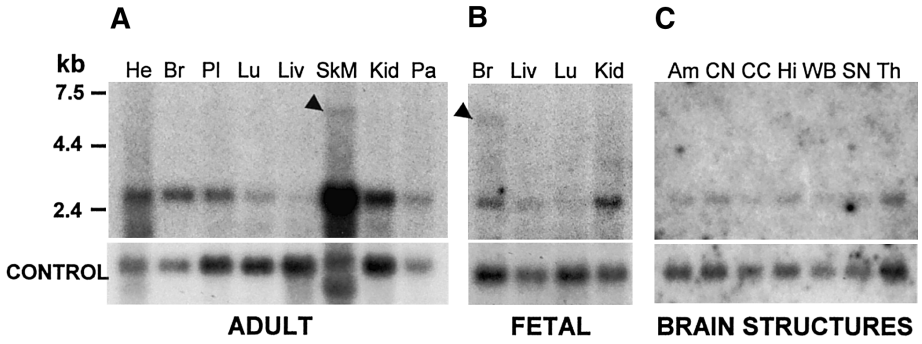
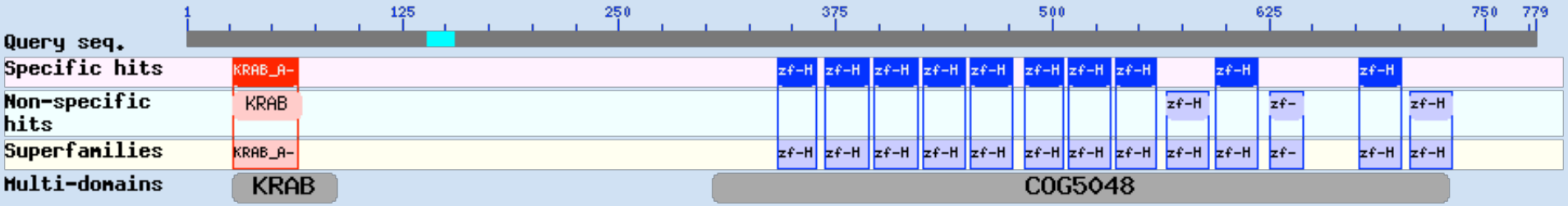
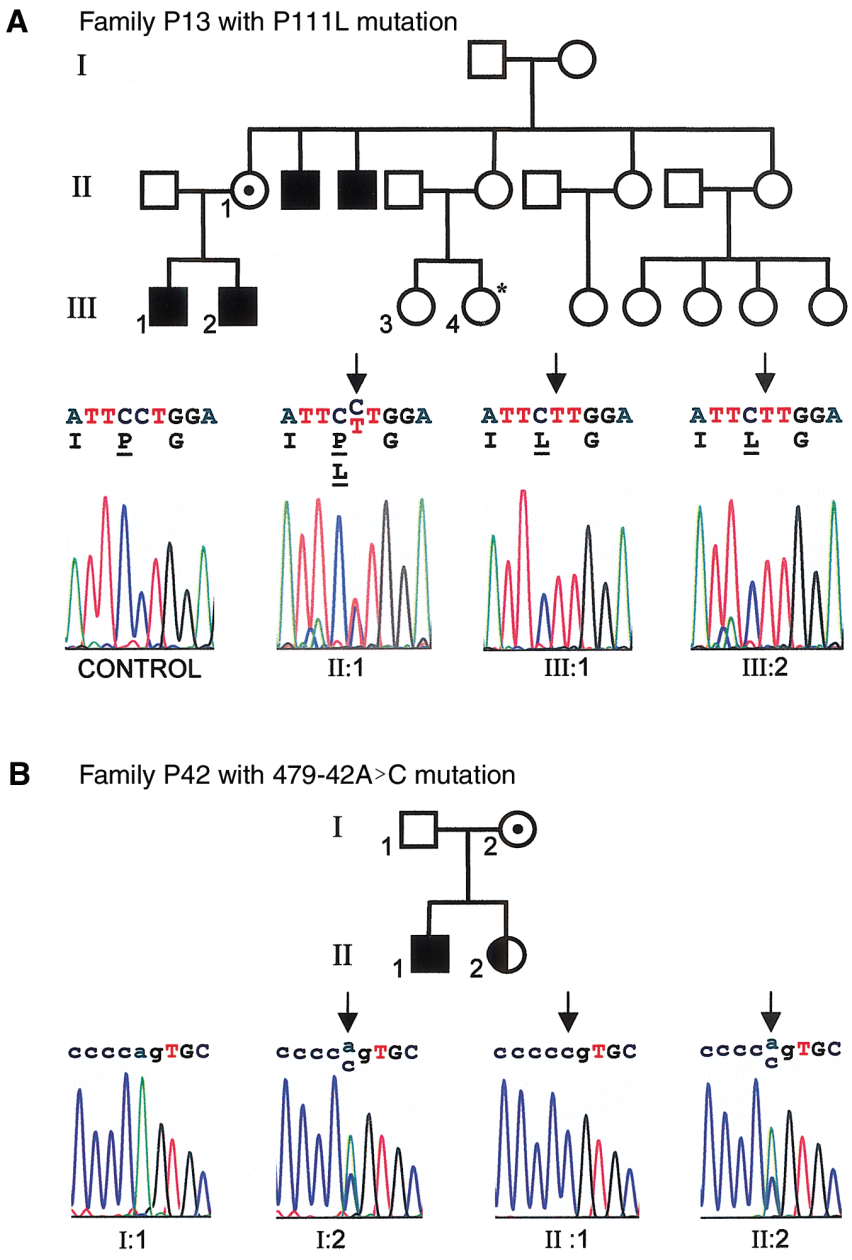


Figure 6 Northern blot hybridization of *ZNF41*, by use of a probe corresponding to nucleotides 621–1099 of *ZNF41* transcript variant 1. *A*, Adult tissues (left to right): heart, brain, placenta, lung, liver, skeletal muscle, kidney, and pancreas. *B*, Fetal tissues (left to right): brain, lung, liver, and kidney. *C*, Adult brain structures (left to right): amygdala, caudate nucleus, corpus callosum, hippocampus, whole brain, substantia nigra, and thalamus. Black arrowheads highlight the presence of a novel 6-kb transcript. *Actin* (*A* and *C*) or *GAPDH* (*B*) served as controls for RNA loading.



- KRAB (Kruppel-associated box) domain -A box.
- The KRAB domain is a transcription repression module, found in a subgroup of the zinc finger proteins (ZFPs) of the C2H2 family, KRAB-ZFPs. KRAB-ZFPs comprise the largest group of transcriptional regulators in mammals, and are only found in tetrapods.
- The KRAB domain is a protein-protein interaction module which represses transcription through recruiting corepressors. The KAP1/ KRAB-AFP complex in turn recruits the heterochromatin protein 1 (HP1) family, and other chromatin modulating proteins, leading to transcriptional repression through heterochromatin formation.



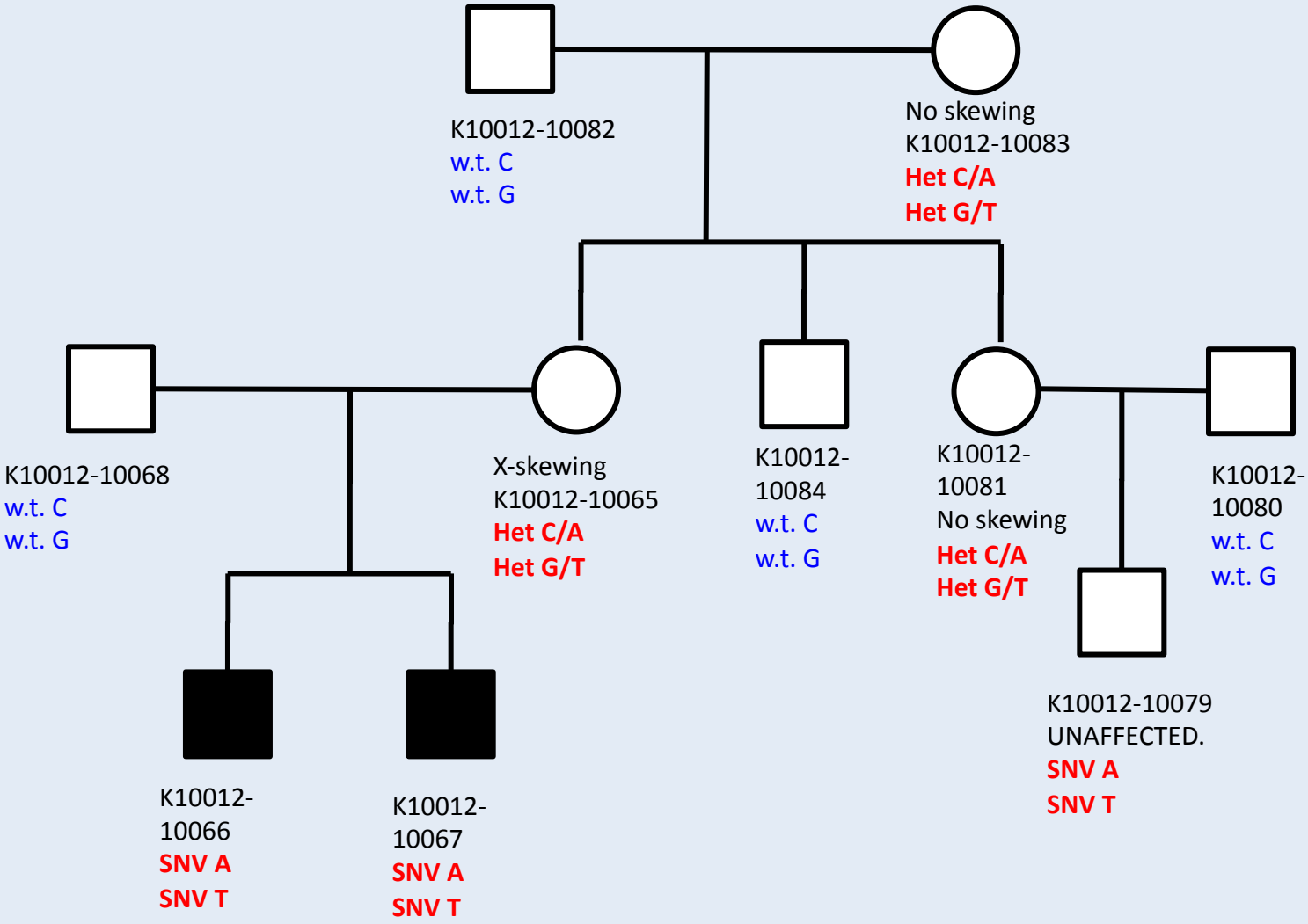
"The P111L change also found in two "male controls" (EVS server, ESP6500).

More recently, we have identified this change in another family with XLID. Co-segregation analysis in this family is in progress.

Furthermore, there are two rare, likely heterozygous ZNF41 frameshift mutations and one heterozygous stop-gained mutation reported in control individuals (ESP6500)."

-Personal communication, Vera Kalscheuer

Sanger validation: ASB12 and ZNF41 mutations



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

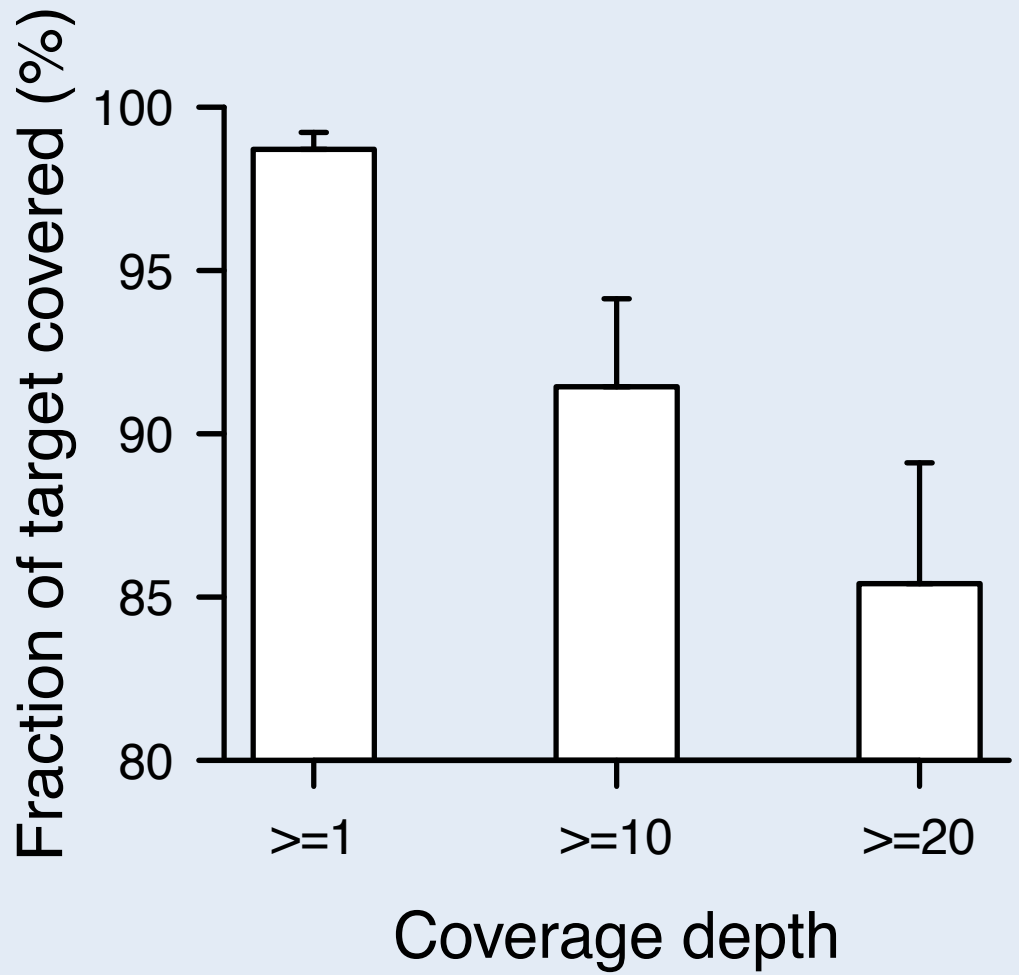
Proving Whether this Mutation is Necessary and Sufficient to result in the phenotype.

- Will need to find a second, unrelated family with same exact mutation and similar phenotype.
- Can also perform in vitro/in vivo studies and structural modeling, and make knock-in mice and/or test in zebrafish, etc… for biological function.
- What about the False Negative Rate in Complete Genomes data? What did we miss?

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

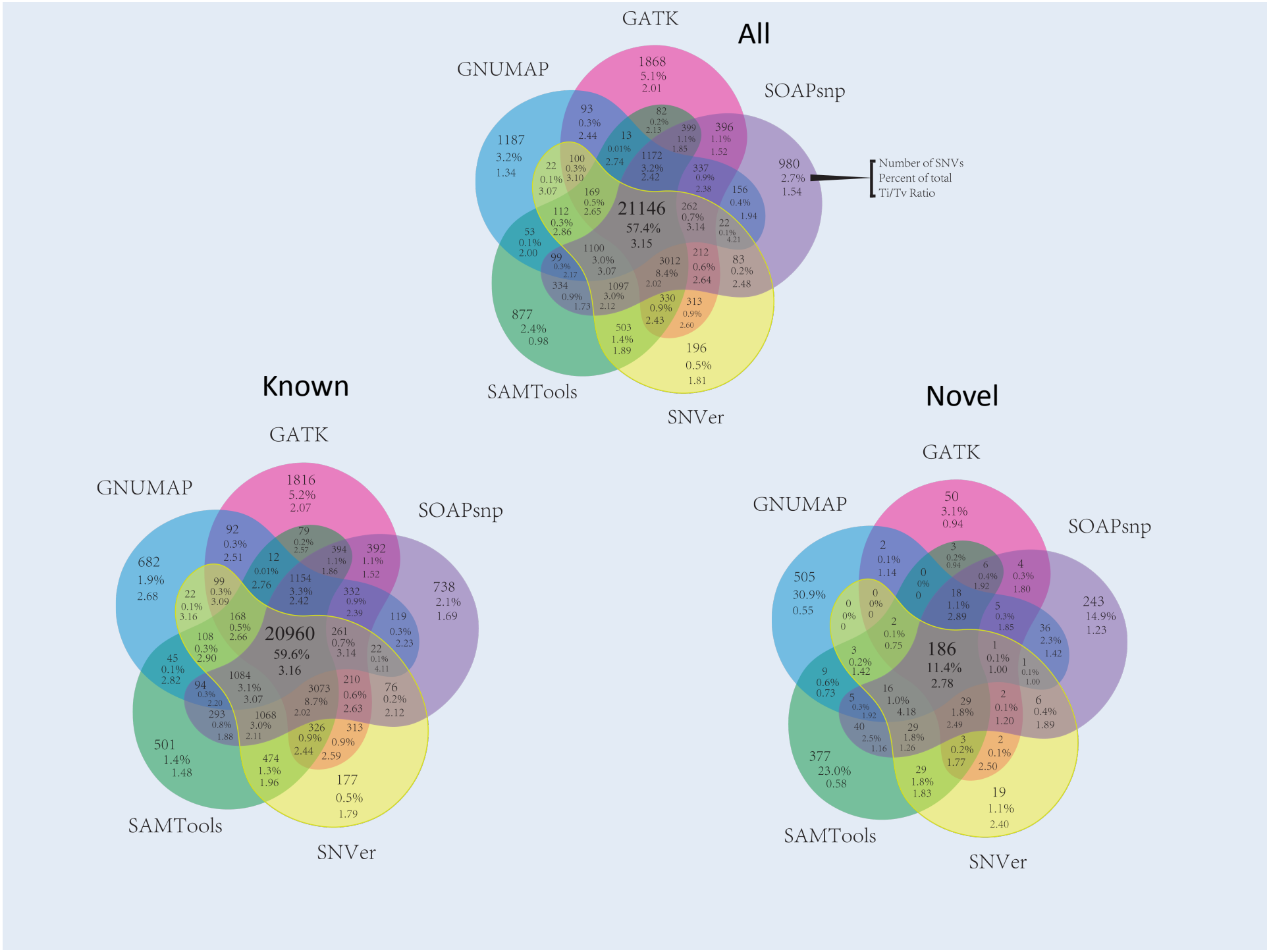
1) BWA - **GATK** (version 1.5) with recommended parameters (GATK IndelRealigner, base quality scores were re-calibrated by GATK Table Recalibration tool. Genotypes called by GATK UnifiedGenotyper. For SNVs and indels.

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.

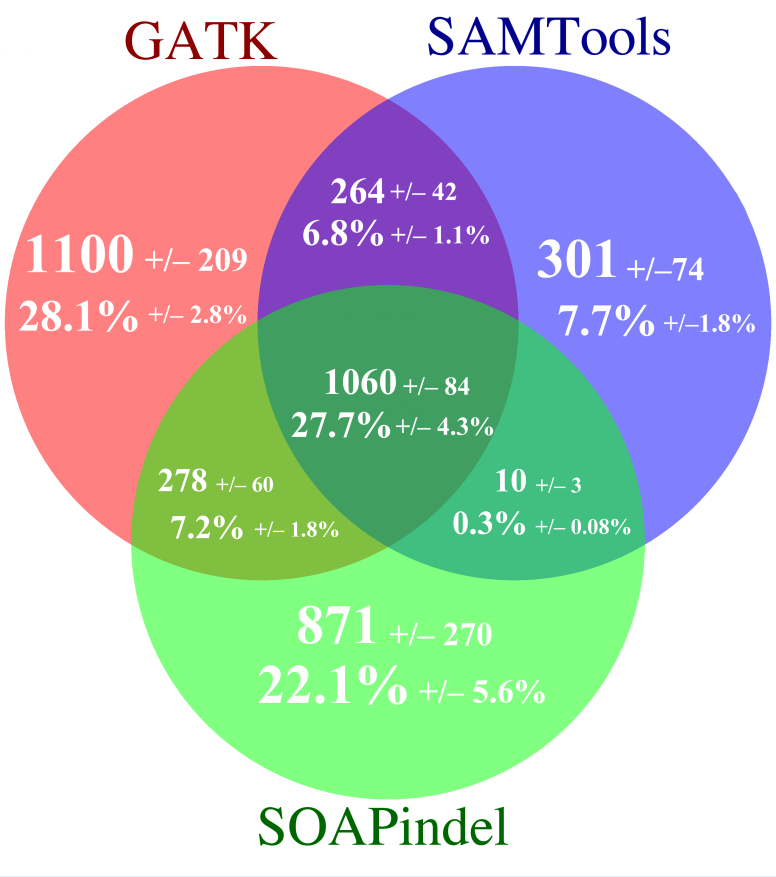
4) GNUMAP-SNP (probabilistic Pair-Hidden Markov which effectively accounts for uncertainty in the read calls as well as read mapping in an unbiased fashion), for SNVs only.

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

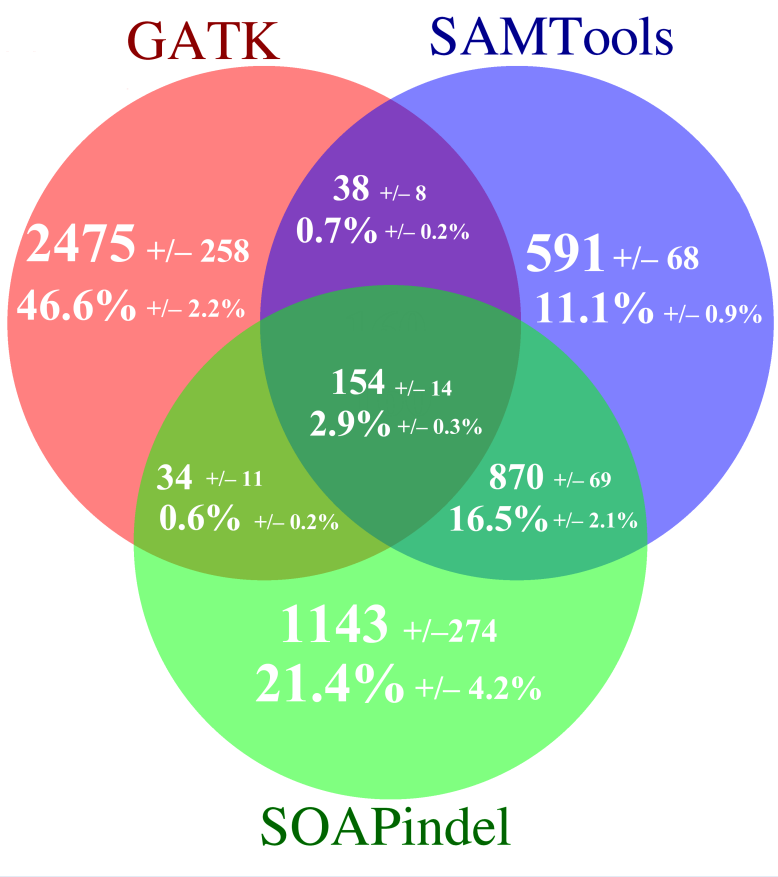

All
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1868 5.1% 2.01
1187 3.2% 1.34
980 2.7% 1.54
93 0.3% 2.44
82 0.2% 2.13
399 1.1% 1.85
396 1.1% 1.52
13 0.01% 2.74
1172 3.2% 2.42
337 0.9% 2.38
156 0.4% 1.94
22 0.1% 3.07
100 0.3% 3.10
169 0.5% 2.65
112 0.3% 2.86
21146 57.4% 3.15
262 0.7% 3.14
22 0.1% 4.21
53 0.1% 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

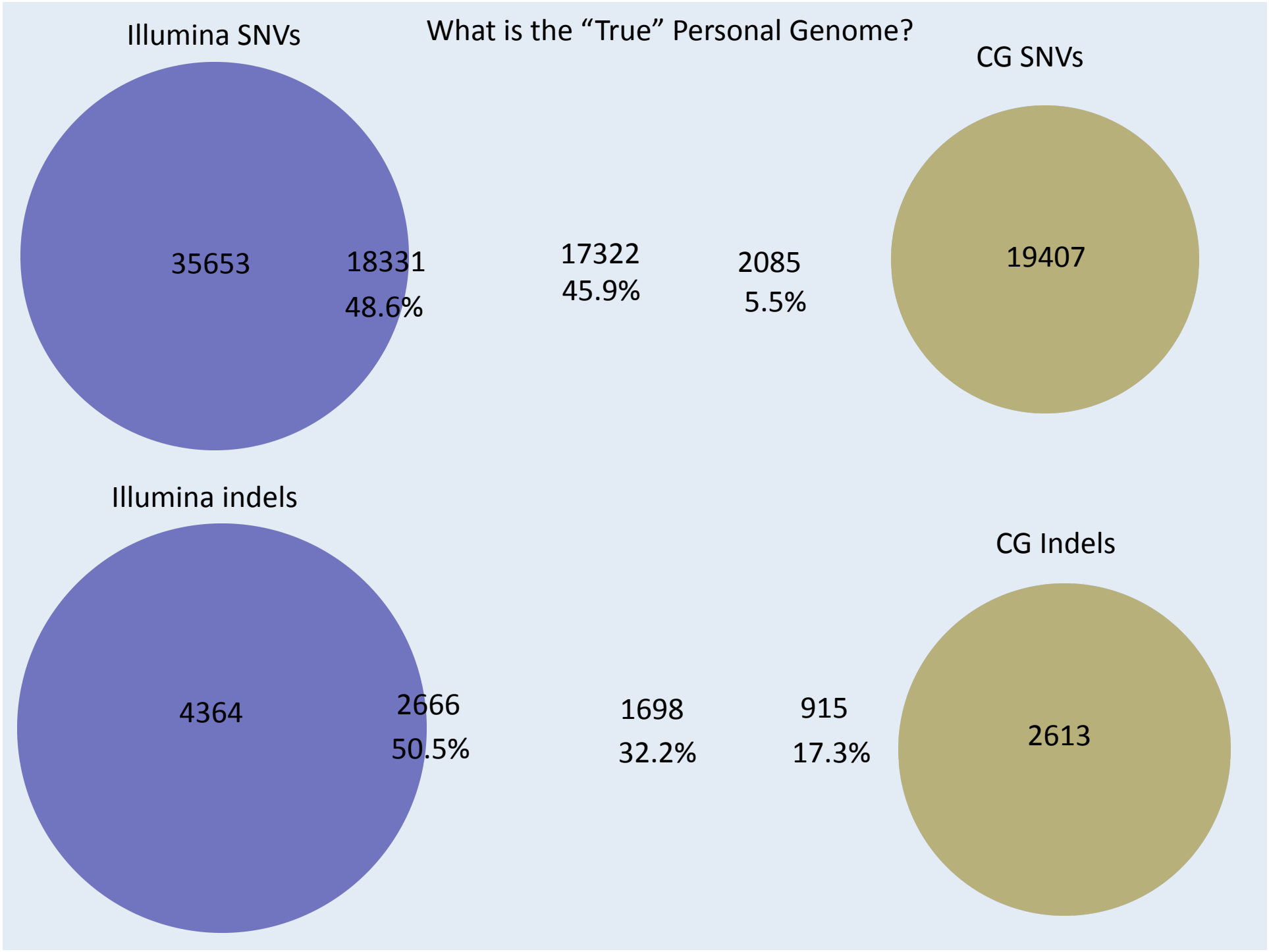


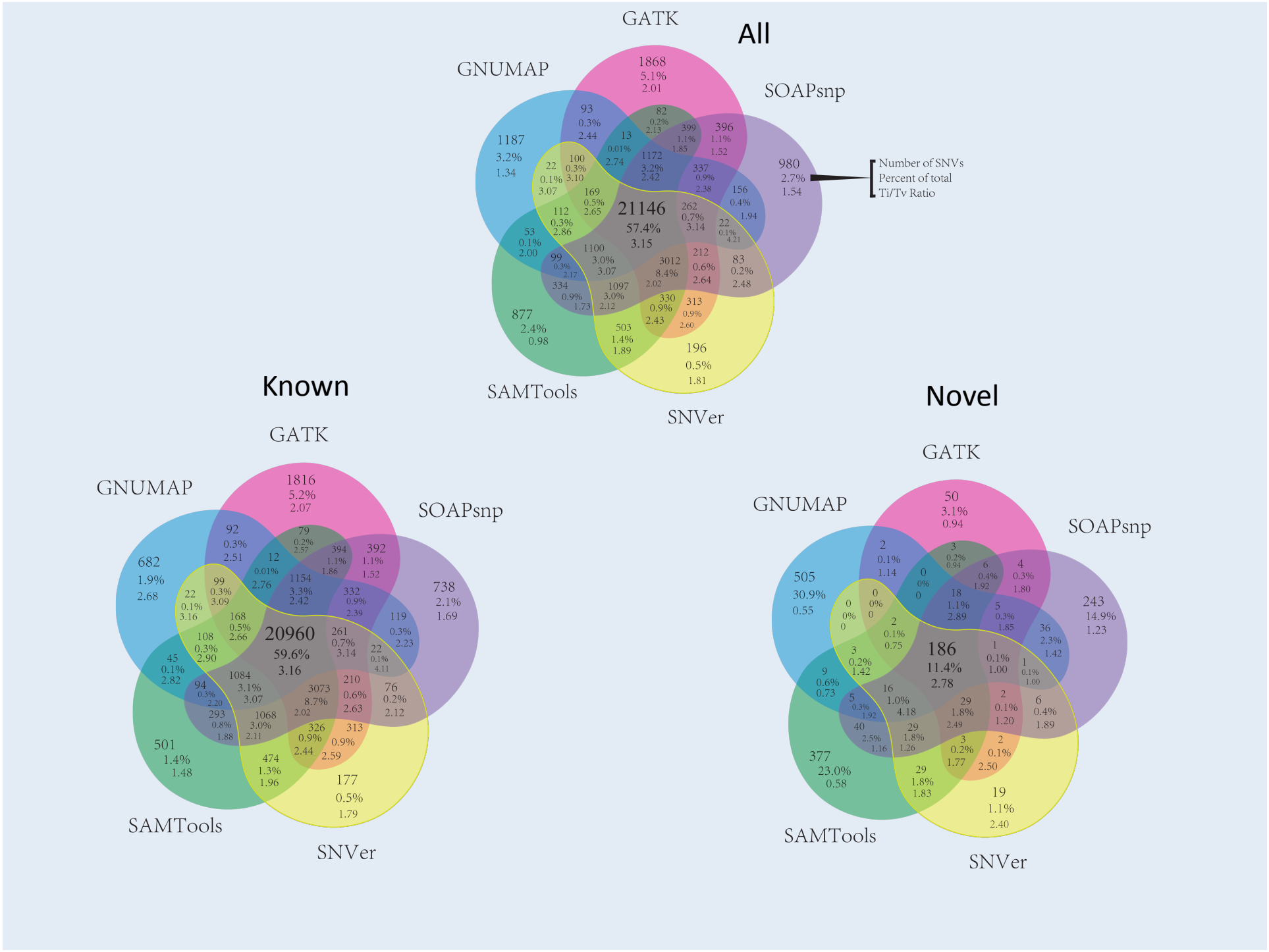
Total mean overlap, plus or minus one standard deviation, observed between three indel calling pipelines: GATK, SOAP-indel, and SAMTools. a) Mean overlap when indel position was the only necessary agreement criterion. **b)** Mean overlap when indel position, base length and base composition were the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

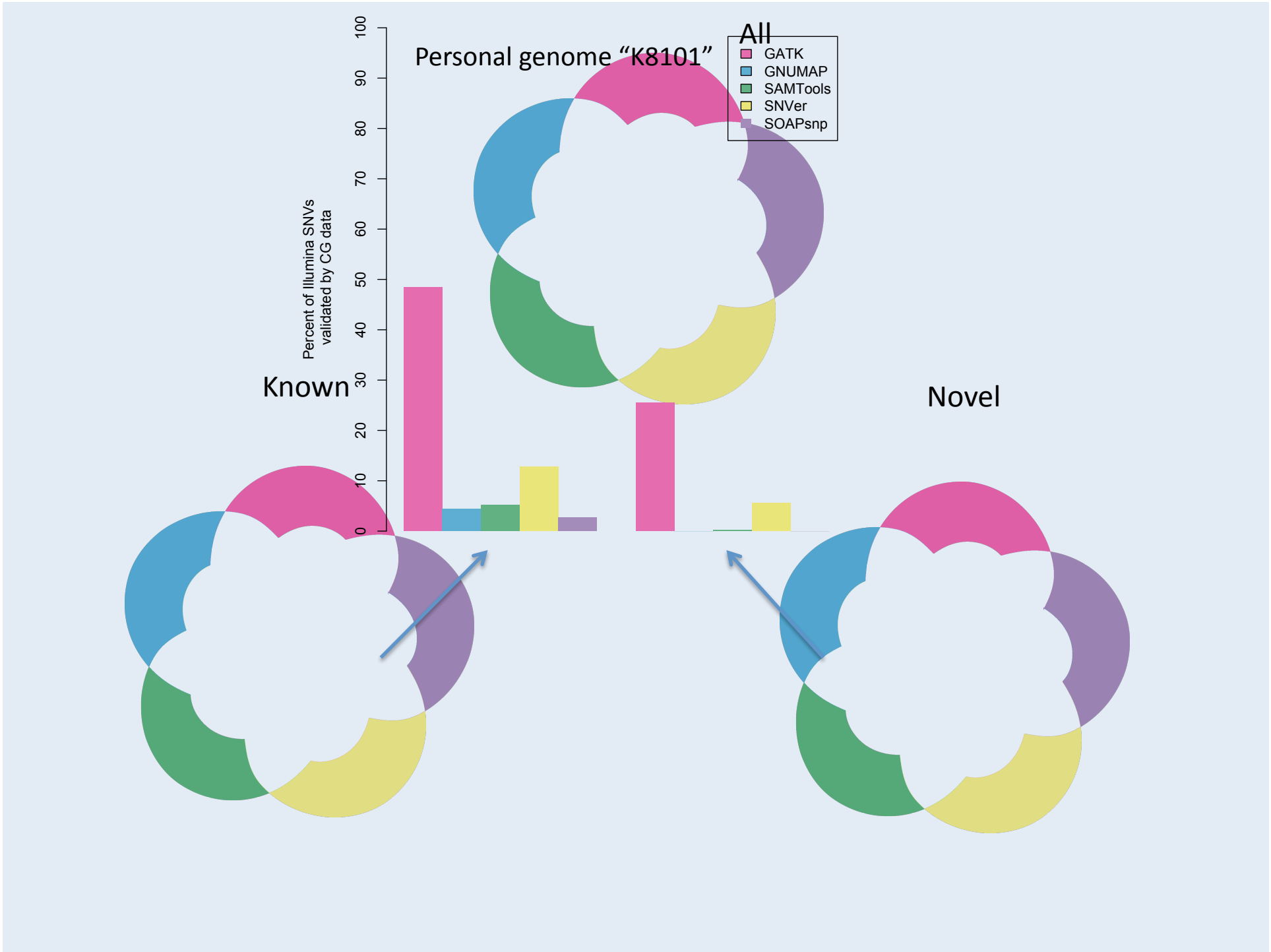
- How reliable are variants that are uniquely called by individual pipelines?

- 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

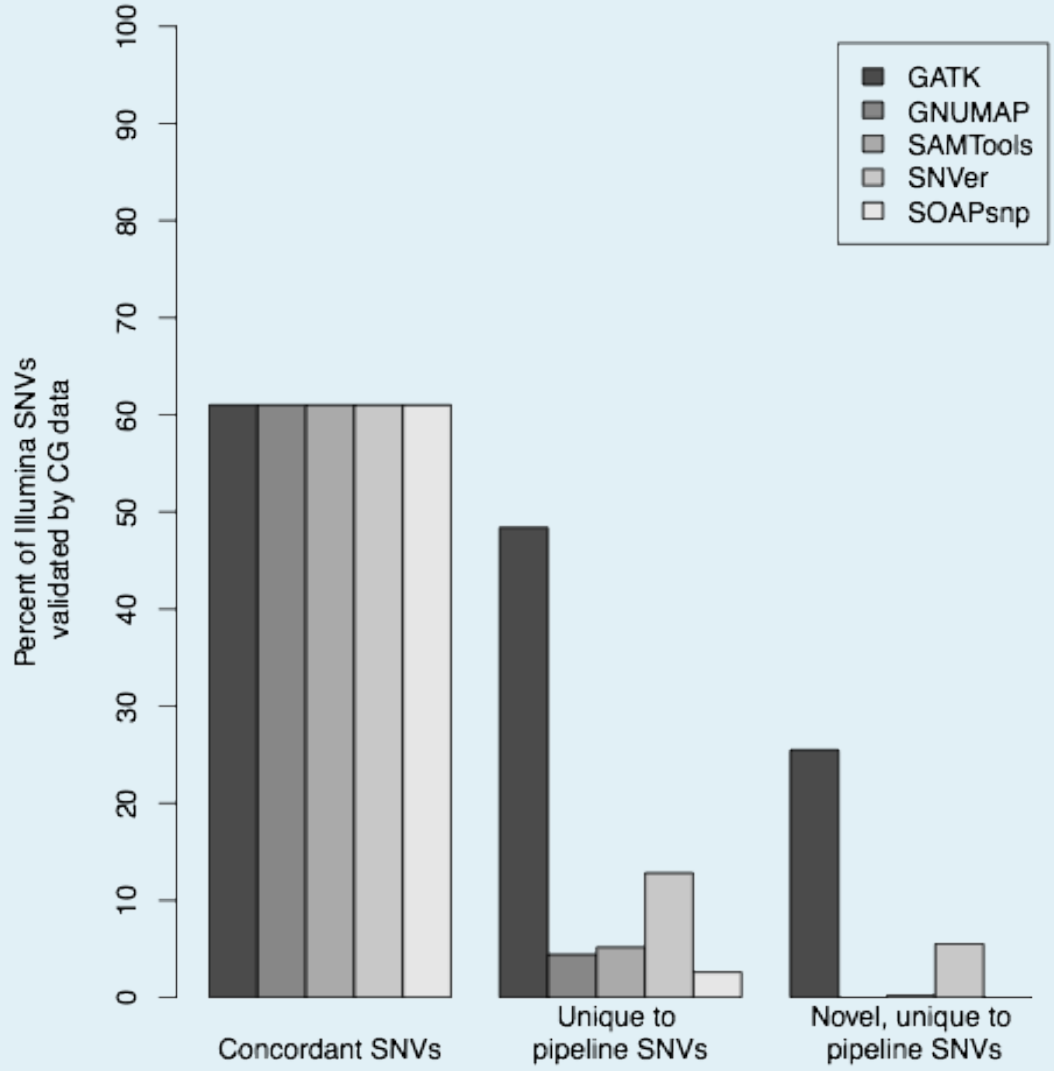

All
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
1868 5.1% 2.01
1187 3.2% 1.34
980 2.7% 1.54
93 0.3% 2.44
82 0.2% 2.13
399 1.1% 1.85
396 1.1% 1.52
13 0.01% 2.74
1172 3.2% 2.42
337 0.9% 2.38
156 0.4% 1.94
22 0.1% 3.07
100 0.3% 3.10
169 0.5% 2.65
112 0.3% 2.86
21146 57.4% 3.15
262 0.7% 3.14
22 0.1% 4.21
53 0.1% 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

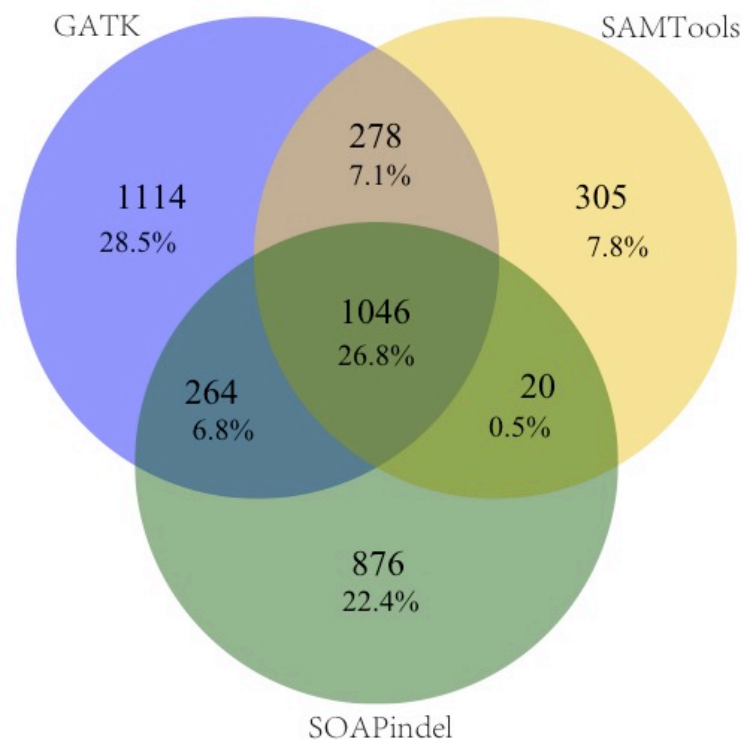
Higher Validation of SNVs with the BWA-GATK pipeline

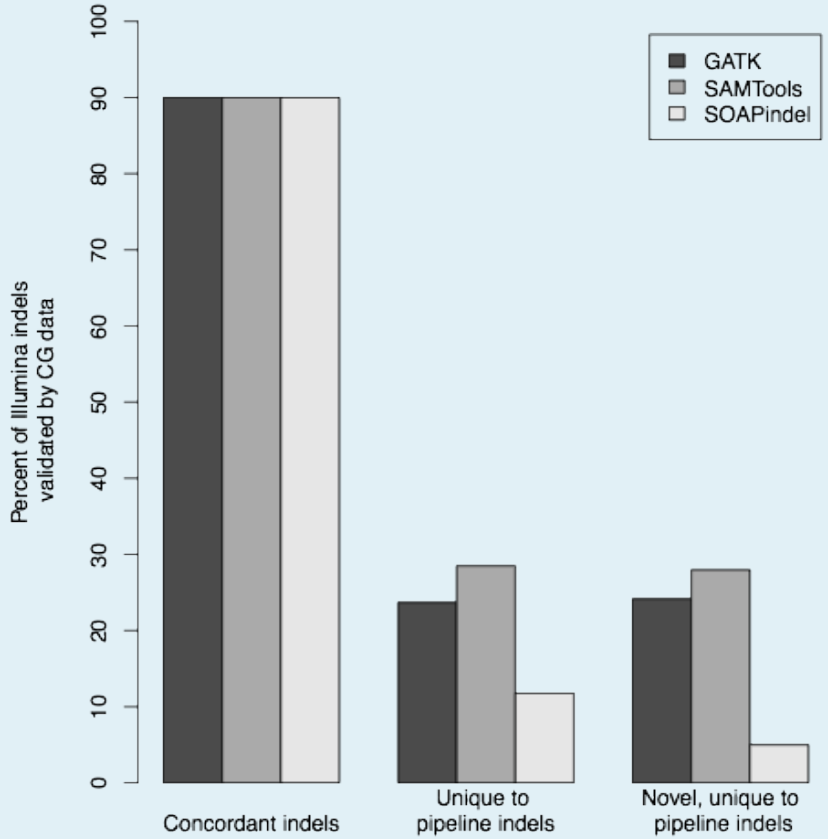
- Reveals higher validation rate of unique-to-pipeline variants, as well as uniquely discovered novel variants, for the variants called by BWA-GATK, in comparison to the other 4 pipelines (including SOAP).

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



Validating Indels with Complete Genomics Data for the 3 pipelines





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]

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

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

Reid Robison

Kai Wang

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

Thomas Arnesen
Rune Evjenth
Johan R. Lillehaug

our study families

Tao Jiang
Guangqing Sun