# Genetic and Biochemical Analysis of Childhood-Onset Idiopathic Neuropsychiatric Disorders.

Gholson Lyon, M.D. Ph.D.

CSH

STANLEY INSTITUTE FOR
COGNITIVE GENOMICS
COLD SPRING HARBOR LABORATORY

UFBR
UTAH FOUNDATION FOR
BIOMEDICAL RESEARCH

@GholsonLyon

# Conflicts of Interest

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

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

“Happy families are all alike; every unhappy family is unhappy in its own way.”

**Leo Tolstoy**, *Anna Karenina, Chapter 1, first line*

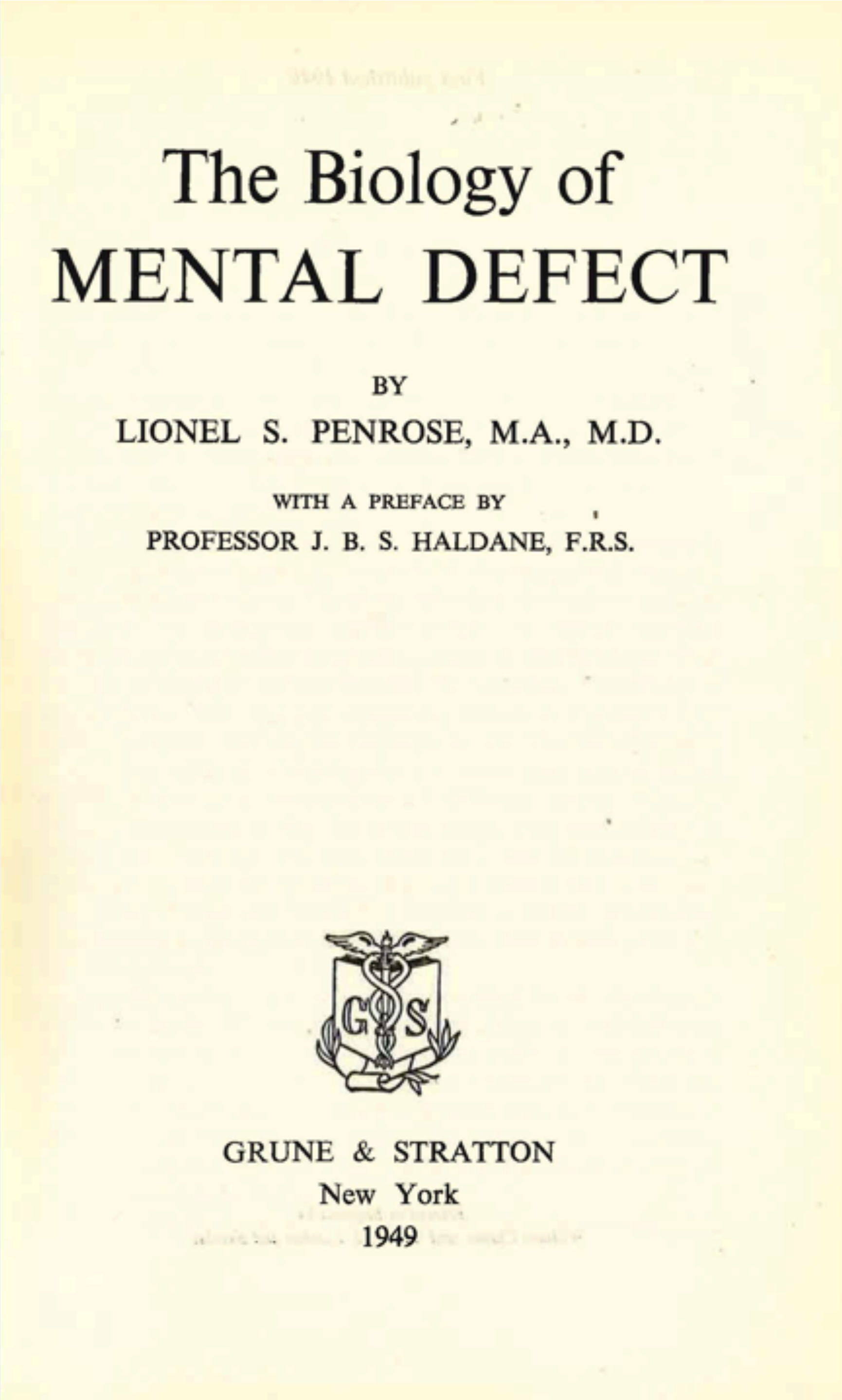
*Russian mystic & novelist (1828 - 1910)*

*First heard quote in human genetics from Mary-Claire King

“We don’t have to look for a model organism anymore, because we *are* the model organisms.”

– Sydney Brenner, Nobel Laureate, quote in 2008

**Hypothesis:** Every human is a unique genetic organism. Therefore, we can find previously unreported idiopathic disorders in humans and identify their genetic basis, thus revealing substantial new biology relevant to medicine.

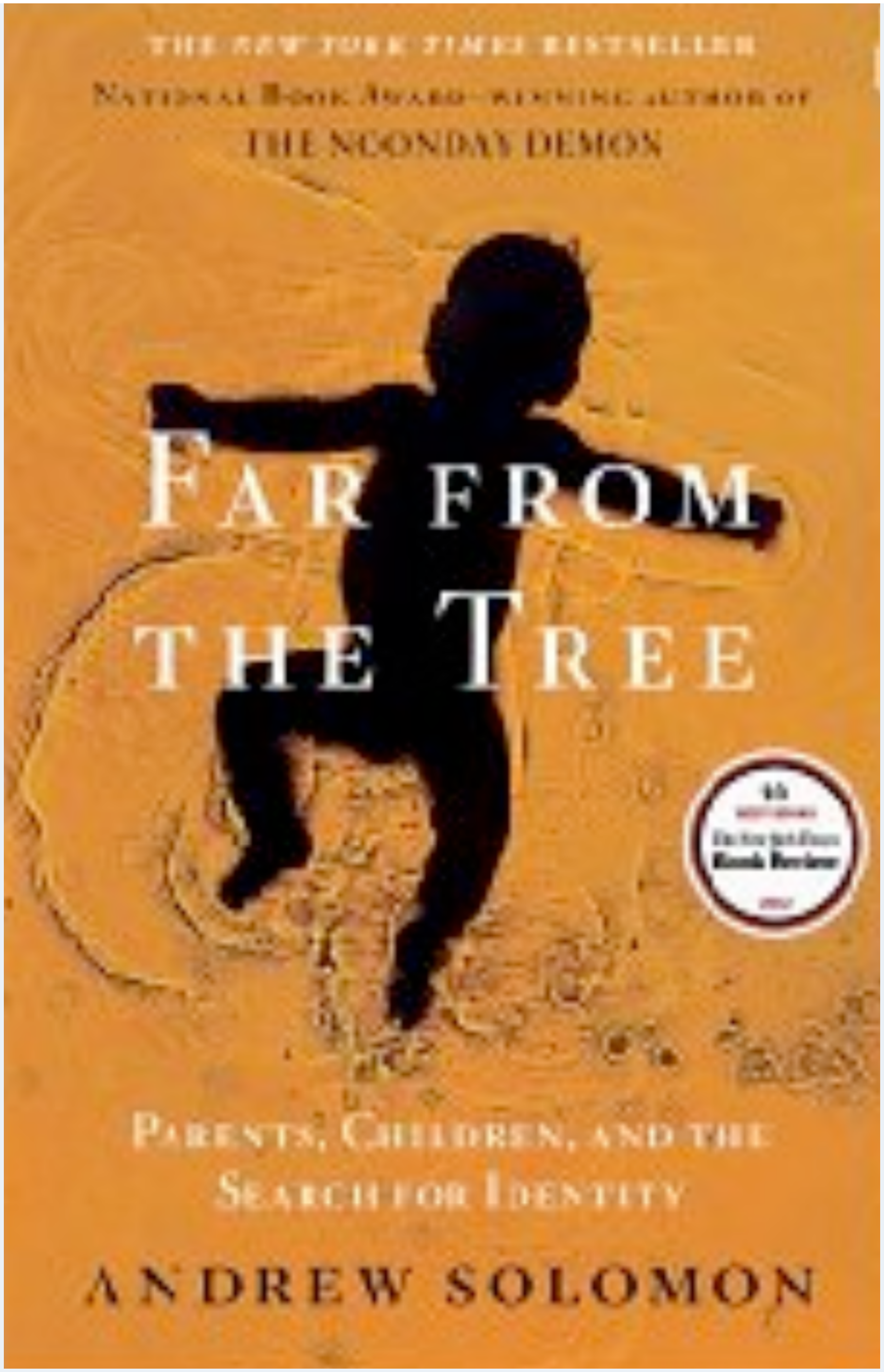


The Biology of
MENTAL DEFECT

BY
LIONEL S. PENROSE, M.A., M.D.

WITH A PREFACE BY
PROFESSOR J. B. S. HALDANE, F.R.S.

GRUNE & STRATTON
New York
1949



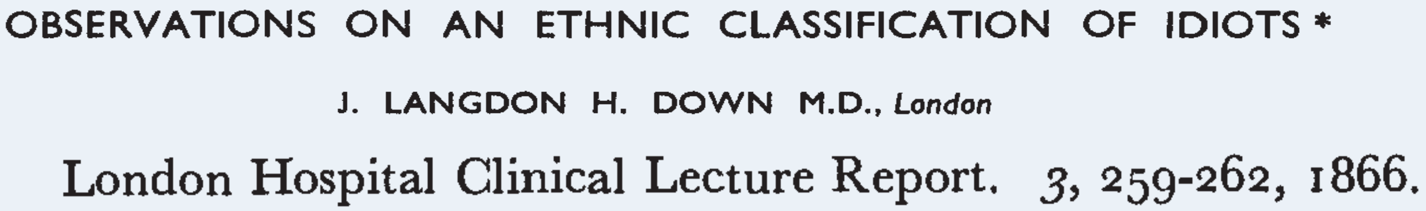
THE NEW YORK TIMES BESTSELLER
NATIONAL BOOK AWARD–WINNING AUTHOR OF
THE NOONDAY DEMON

FAR FROM THE TREE

PARENTS, CHILDREN, AND THE
SEARCH FOR IDENTITY

ANDREW SOLOMON

- Seguin E. 1866, Idiocy and its treatment by the physiological method.
- - “our incomplete studies do not permit actual classification; but it is better to leave things by themselves rather than to force them into classes which have their foundation only on paper”.



OBSERVATIONS ON AN ETHNIC CLASSIFICATION OF IDIOTS *

J. LANGDON H. DOWN M.D., *London*

London Hospital Clinical Lecture Report. 3, 259-262, 1866.

“Those who have given any attention to congenital mental lesions, must have been frequently puzzled how to arrange, in any satisfactory way, the different classes of this defect which may have come under their observation. Nor will the difficulty be lessened by an appeal to what has been written on the subject. The systems of classification are generally so vague and artificial, that, not only do they assist but feebly, in any mental arrangement of the phenomena represented, but they completely fail in exerting any practical influence on the subject.”

# The Big Picture

- Over the course of my entire career, I want to help understand the pathophysiology of severe neuropsychiatric disorders, including such things as developmental delay, mental retardation, autism, psychotic disorders (schizophrenia, bipolar, schizoaffective), Tourette Syndrome and obsessive compulsive disorder.

- I expect this to uncover new biology along the way.

# The toll with Brain Disorders is tremendous.

Most recent analysis in Europe showed that brain disorders cost almost US $1 trillion year per year, more than cancer, cardiovascular disease and diabetes combined.

These brain disorders include:
Mood disorders
**Psychotic Disorders**
Addiction
Anxiety
Dementia
Headache
Other- brain tumor, **child/adolescent developmental disorders (autism, ADHD, tics, etc…),** eating disorders, epilepsy, **mental retardation,** multiple sclerosis, neuromuscular disorders, Parkinson's, personality disorders, sleep disorders, somatoform disorder, stroke and traumatic brain injury.

* Cost of disorders of the brain in Europe 2010.
Gustavsson A, et al.
Eur Neuropsychopharmacol. 2011 Oct;21(10):718-779. Epub 2011 Sep 15.

# UFBR

UTAH FOUNDATION FOR
BIOMEDICAL RESEARCH

## I moved to Utah in July 2009 to find new human genetic syndromes, thus revealing new biology.

- ◆ July 2009-December 2009: Attended weekly genetics case conference in which 10-30 genetic cases are presented weekly, led by Dr. Alan Rope and attended by Drs. John Carey and John Opitz.
- ◆ There are indeed MANY idiopathic disorders not described in the literature, many of which have neuropsychiatric manifestations. I thought about hundreds of such cases, looking for ideal first families to sequence.

## Beyond our Kuhnian inheritance

A recent lecture by Prof Greg Radick questions our scientific inheritance, through textbook histories of genetics and Thomas Kuhn's legacy

http://www.guardian.co.uk/science/the-h-word/2012/aug/28/thomas-kuhn

Vs.

Walter Frank Raphael Weldon

William Bateson

Forthcoming by Greg Radick. Scholarly edition of W. F. R. Weldon's Theory of Inheritance (1904-1905), coedited with Annie Jamieson.

“The fundamental mistake which vitiates all work based upon Mendel’s method is the neglect of ancestry, and the attempt to regard the whole effect upon offspring, produced by a particular parent, as due to the existence in the parent of particular structural characters; while the contradictory results obtained by those who have observed the offspring of parents apparently identical in certain characters show clearly enough that not only the parents themselves, but their race, that is their ancestry, must be taken into account before the result of pairing them can be predicted.”

Weldon, W. F. R. 1902. Mendel’s laws of alternative inheritance in peas. *Biometrika*, 1:228-254.

# “Biological Indeterminacy”

- Bateson became famous as the outspoken Mendelian antagonist of Walter Raphael Weldon, his former teacher, and Karl Pearson who led the biometric school of thinking. This concerned the debate over saltationism versus gradualism (Darwin had been a gradualist, but Bateson was a saltationist). Later, Ronald Fisher and J.B.S. Haldane showed that discrete mutations were compatible with gradual evolution: see the modern evolutionary synthesis.

Biological Indeterminacy. Greenspan RJ. Sci Eng Ethics. 2012 Jul 3

# Walter Frank Raphael Weldon 1860–1906

A Memoir

Karl Pearson

CAMBRIDGE

# 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?

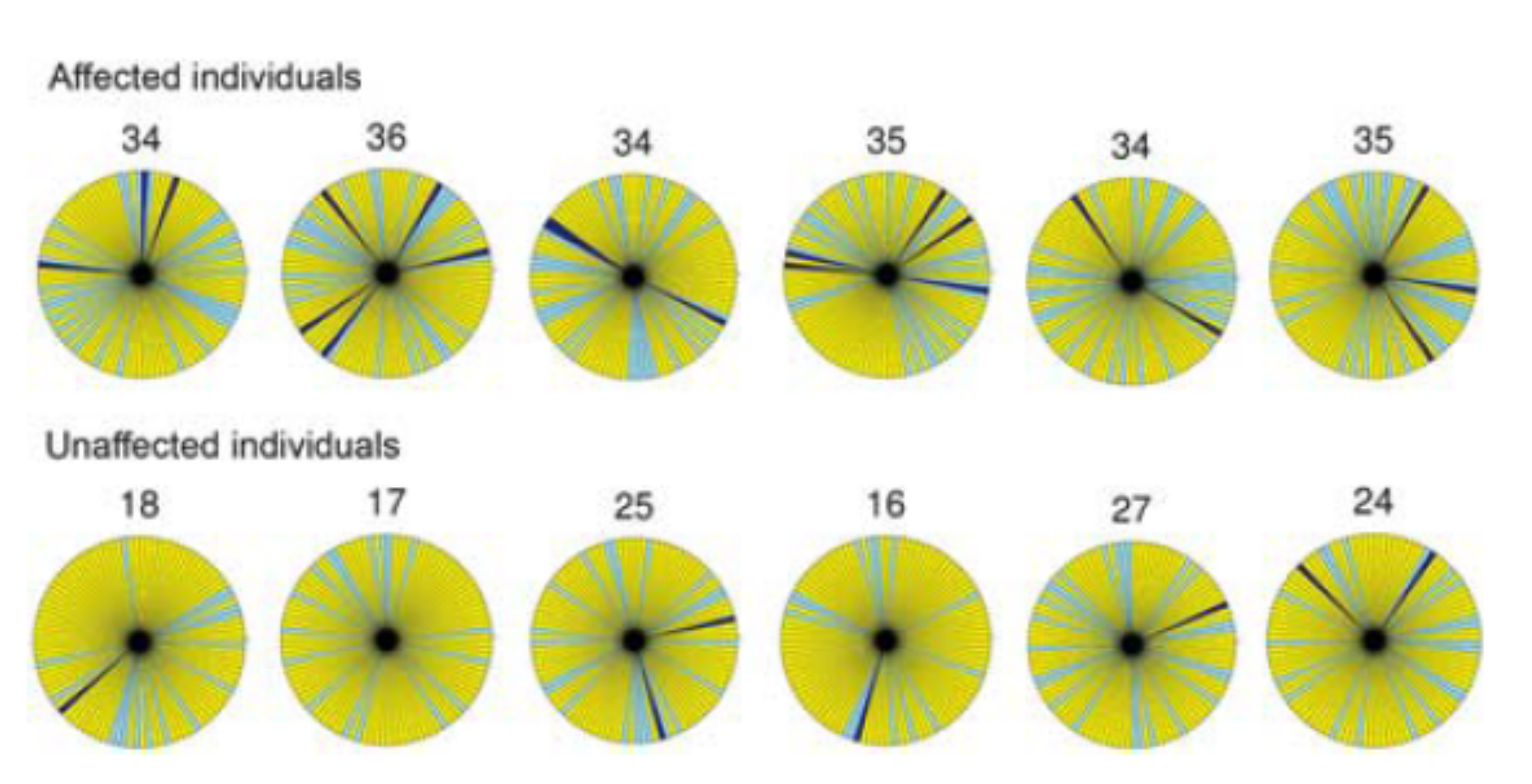
# Some Definitions …

- The words "penetrance" and "expressivity", defined classically as:
- Penetrance: whether someone has ANY symptoms of a disease, i.e. all or none, 0% or 100%. **Nothing in between.**
- Expressivity: how much disease (or how many symptoms) someone with 100% penetrance has.
- This has led to endless confusion!
- Some just use the word "penetrance" to mean the expressivity of disease, i.e. incomplete penetrance, and maybe we should combine the two terms into ONE word with the full expression from 0-100% of phenotypic spectrum.

# **Definitions.** It is unknown what portion of autism will be oligogenic vs. polygenic

- **Oligogenic** – multiple mutations together contributing to aggregate disease, BUT with only 1 mutation of ~ >10% penetrance (or “effect size) in EACH person.
- **Polygenic** – Dozens to hundreds of mutations in different genes in the SAME person, together contributing to the disease in the SAME person, hence **additive** and/or **epistatic** contribution with ~0.01-1% penetrance for each mutation.

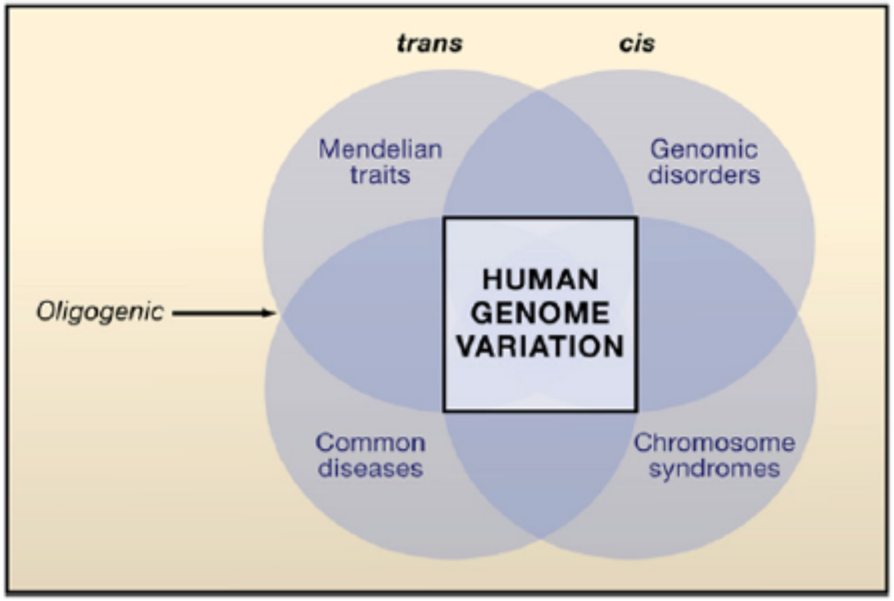
# Example of Polygenic Model

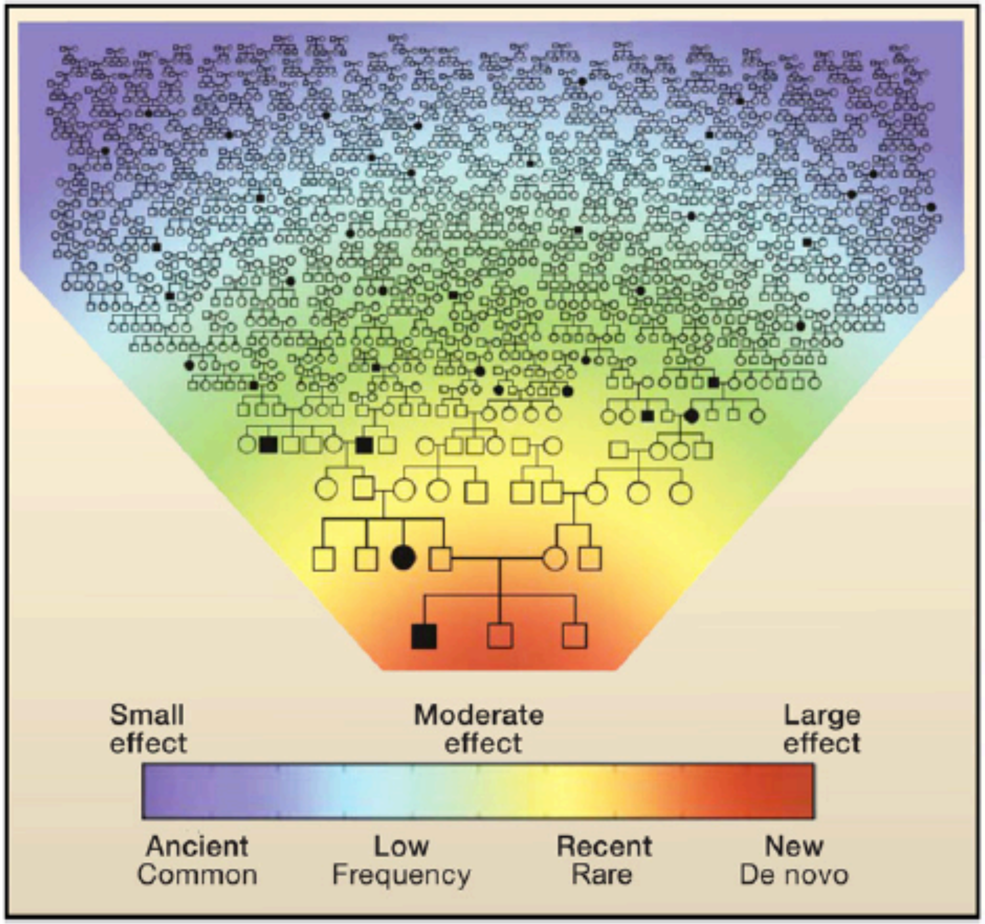


Visscher et al. 2011

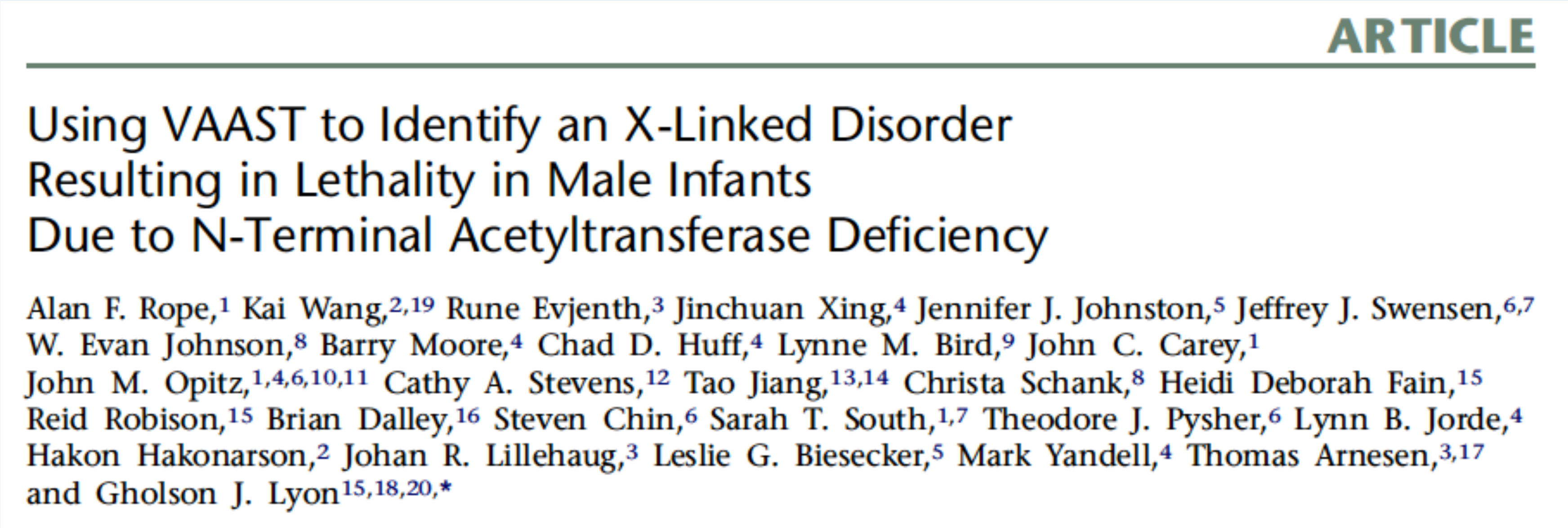
# Clan Genomics and the Complex Architecture of Human Disease

James R. Lupski,[1,2,3,*] John W. Belmont,[1,2] Eric Boerwinkle,[4,5] and Richard A. Gibbs[1,5,*]





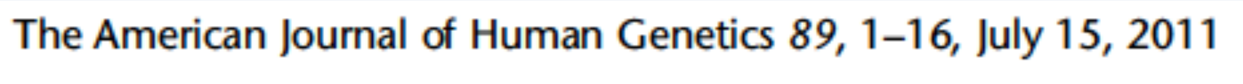
# Discovering a new syndrome and its genetic basis.



ARTICLE

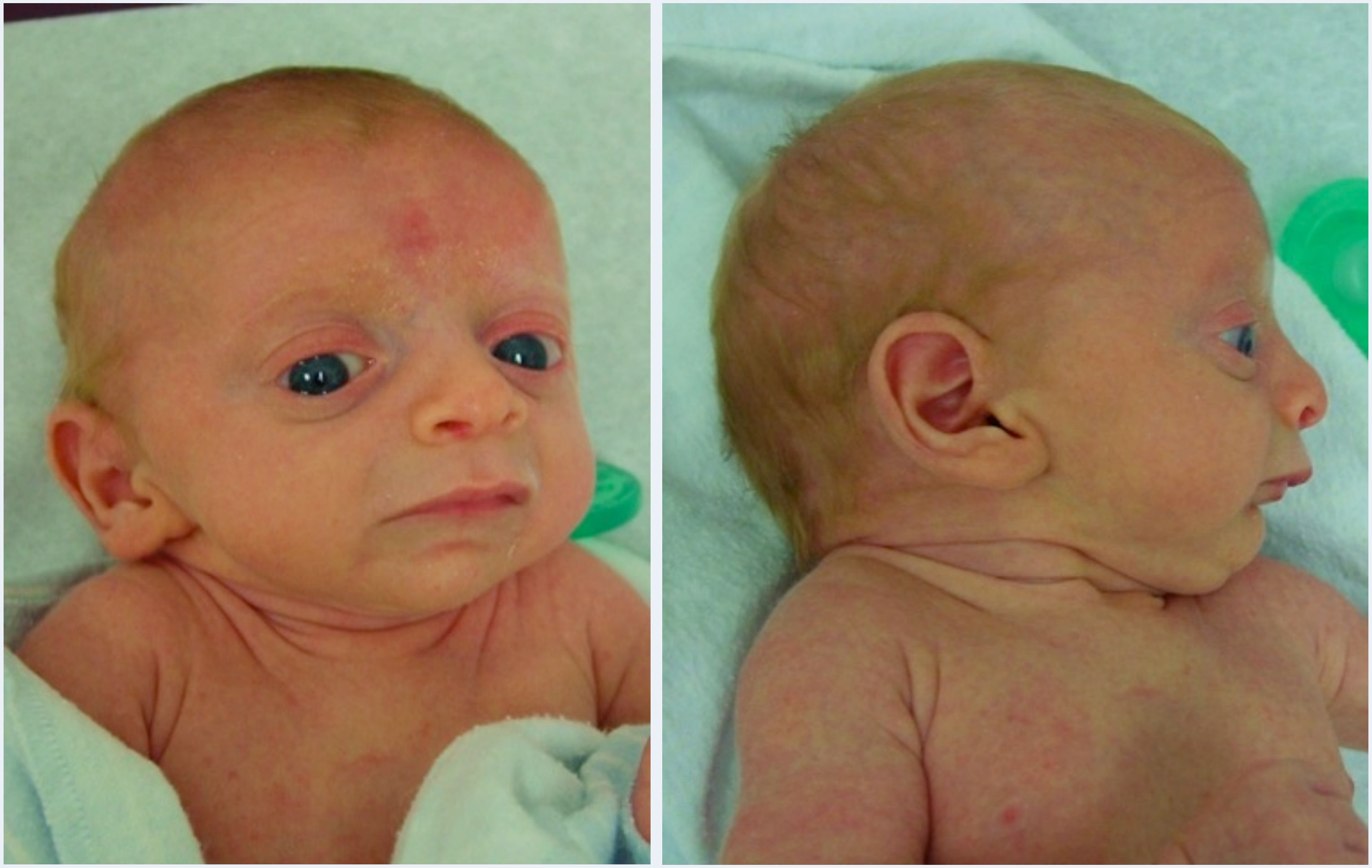
## Using VAAST to Identify an X-Linked Disorder Resulting in Lethality in Male Infants Due to N-Terminal Acetyltransferase Deficiency

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,*]



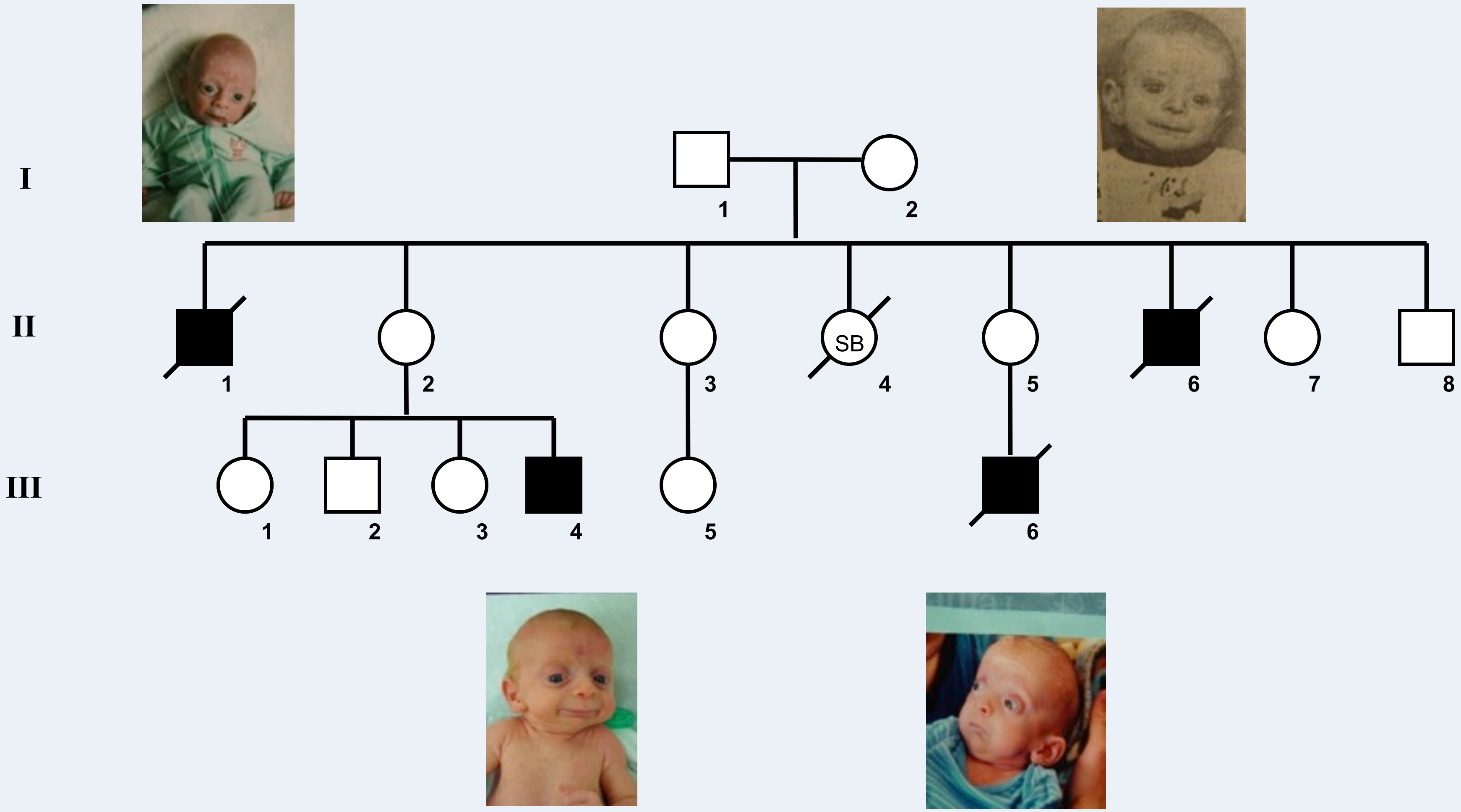
The American Journal of Human Genetics *89*, 1–16, July 15, 2011

# 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

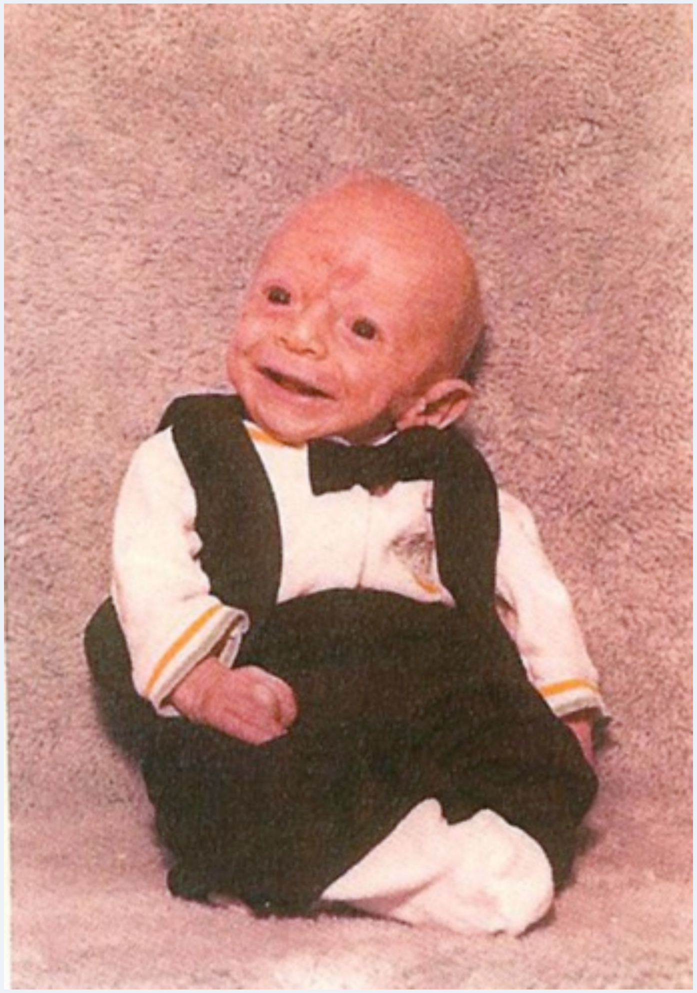
# This is the family in Utah in December 2009.



# I met the entire family on March 29, 2010

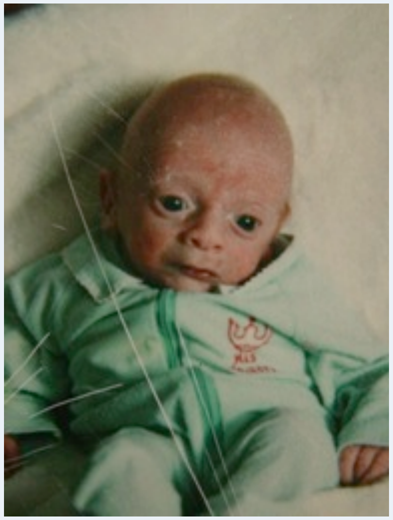
Photo of mother
with son in late 1970’s

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

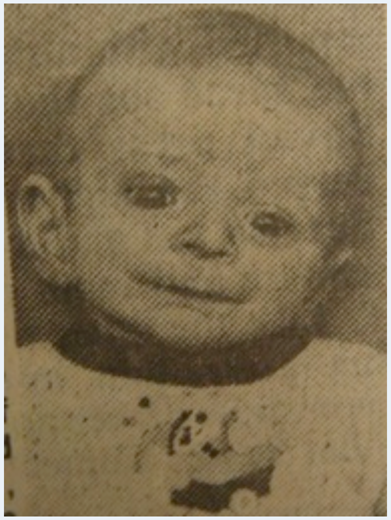


First boy. Called “a little old man” by the family. Died around ~1 year of age, from cardiac arrhythmias.

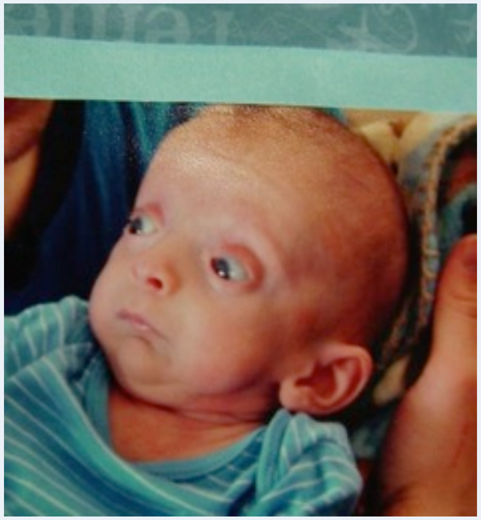
# These are the Affected Boys of Family 1 in 2009.



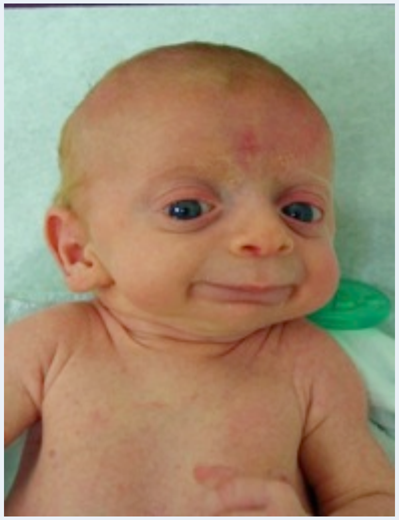
Uncle #1



Uncle #2



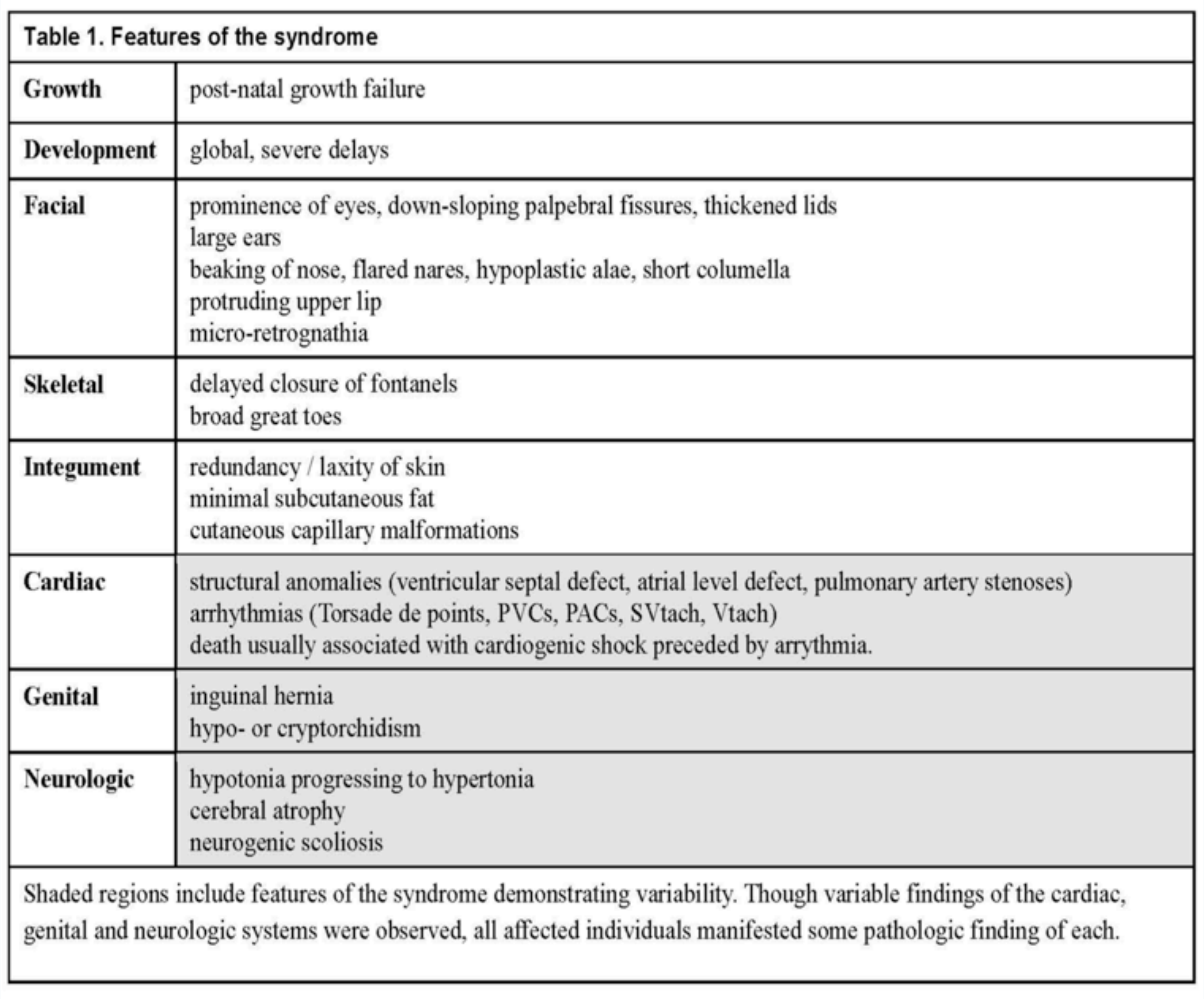
cousin



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.

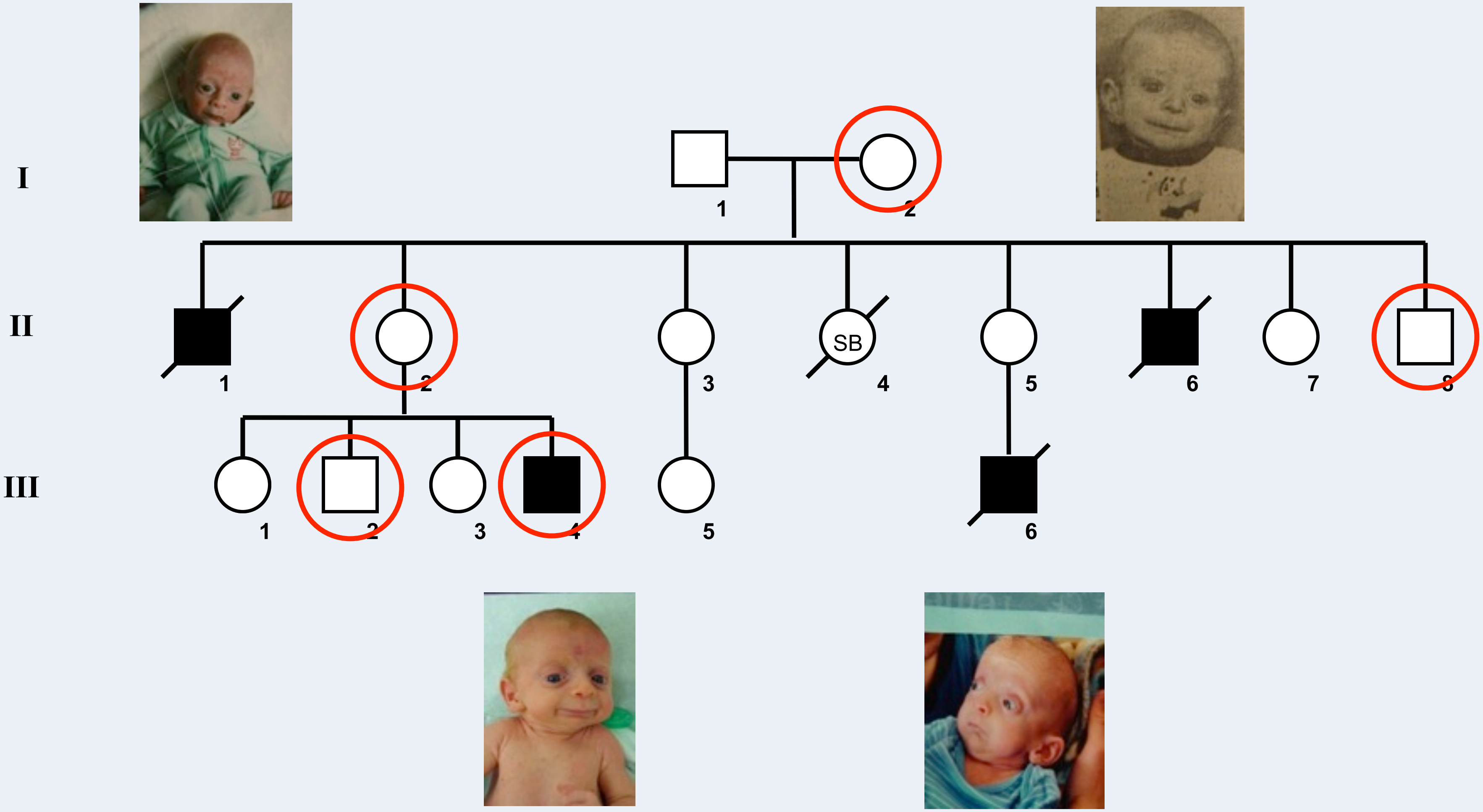
# These are the Major Features of the Syndrome.



| 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<br>large ears<br>beaking of nose, flared nares, hypoplastic alae, short columella<br>protruding upper lip<br>micro-retrognathia |
| **Skeletal** | delayed closure of fontanels<br>broad great toes |
| **Integument** | redundancy / laxity of skin<br>minimal subcutaneous fat<br>cutaneous capillary malformations |
| **Cardiac** | structural anomalies (ventricular septal defect, atrial level defect, pulmonary artery stenoses)<br>arrhythmias (Torsade de points, PVCs, PACs, SVtach, Vtach)<br>death usually associated with cardiogenic shock preceded by arrythmia. |
| **Genital** | inguinal hernia<br>hypo- or cryptorchidism |
| **Neurologic** | hypotonia progressing to hypertonia<br>cerebral atrophy<br>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.

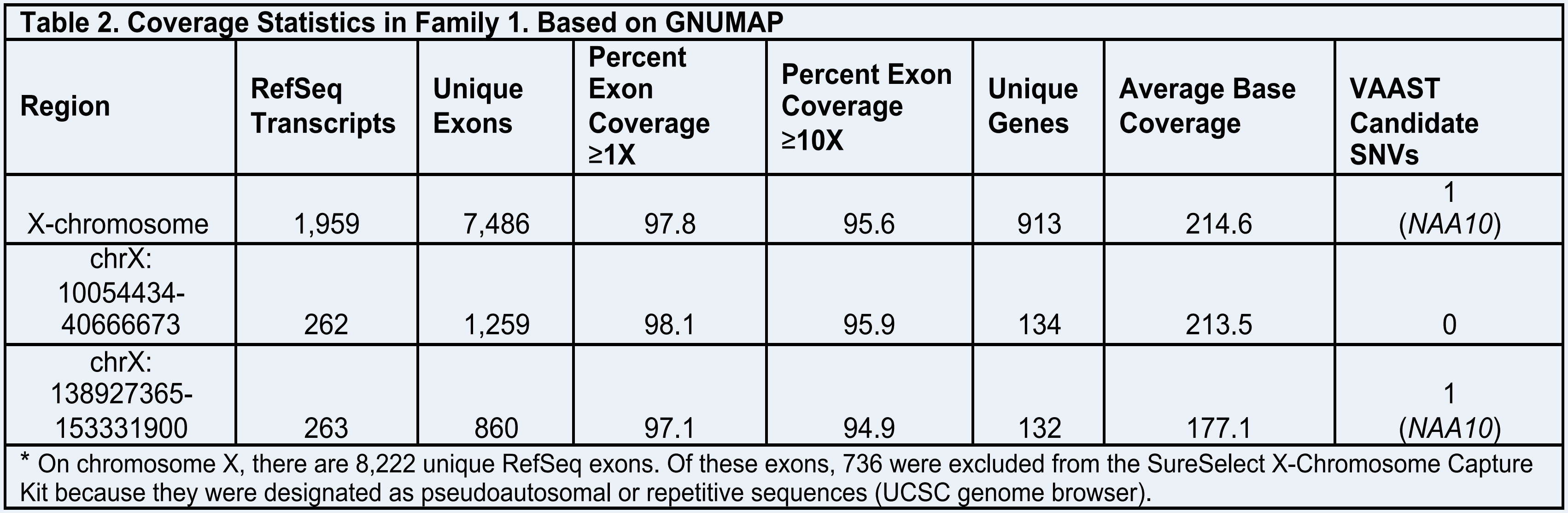
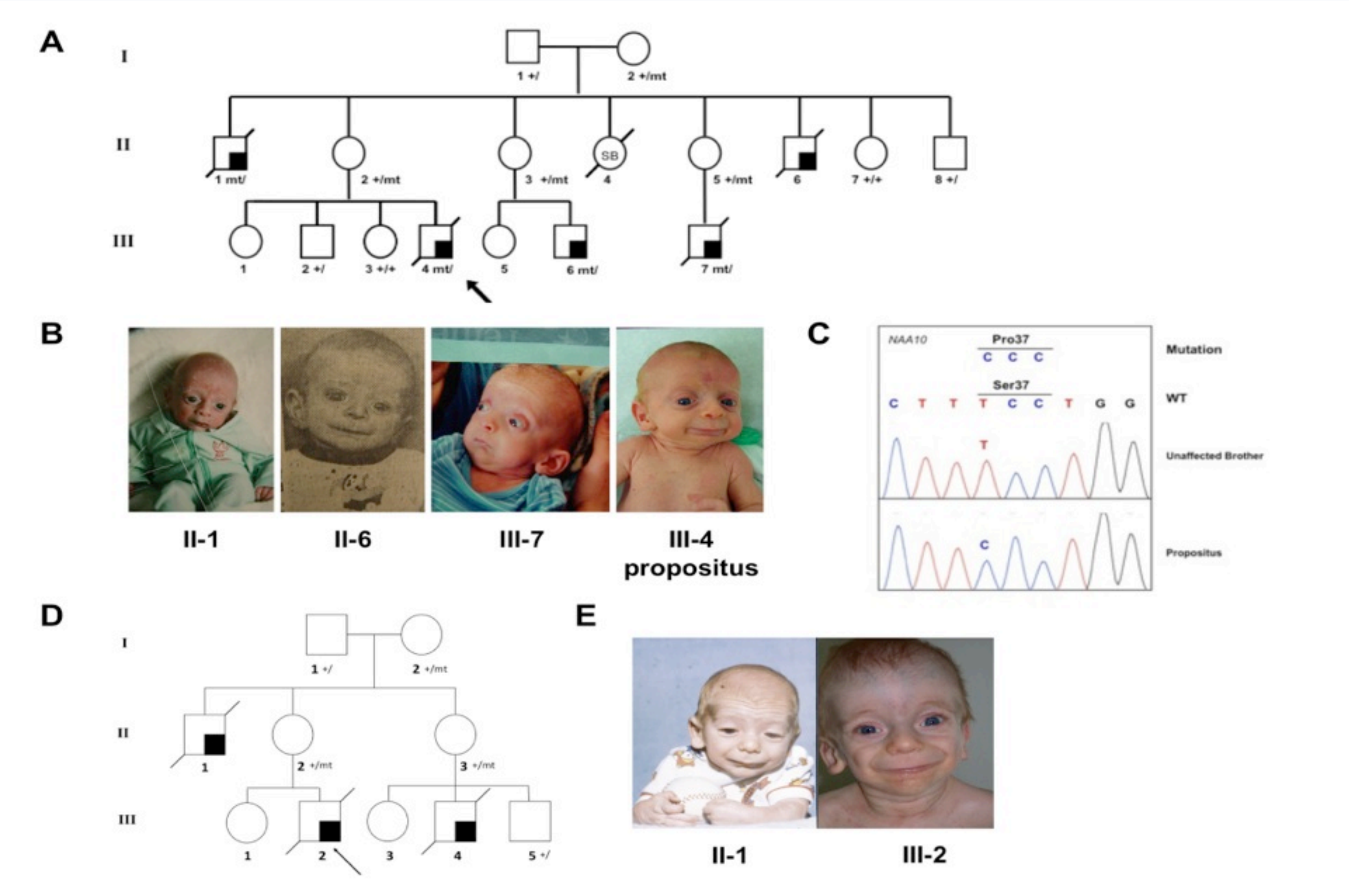


**Table 2. Coverage Statistics in Family 1. Based on GNUMAP**

| 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*) |

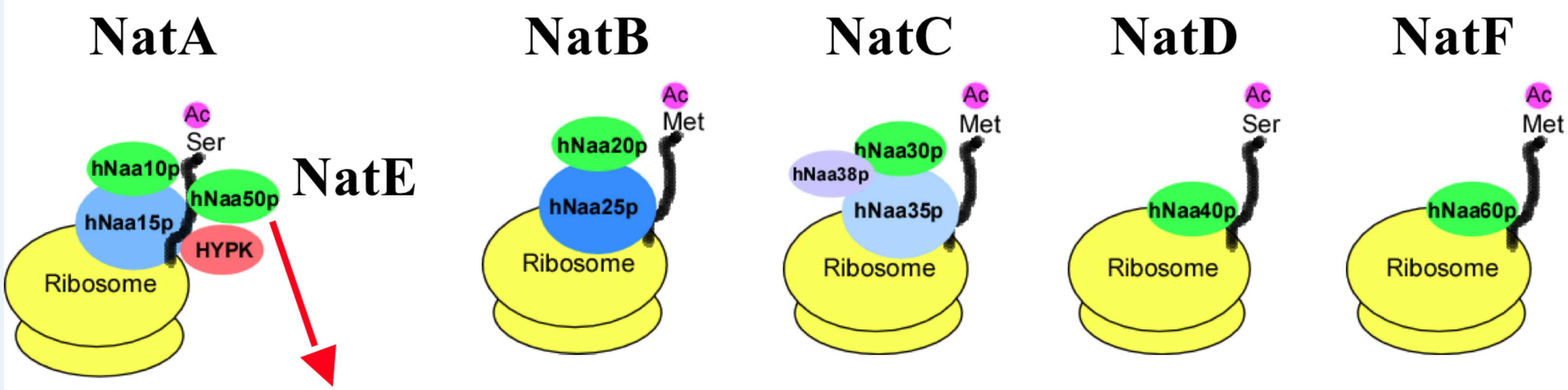
* On chromosome X, there are 8,222 unique RefSeq exons. Of these exons, 736 were excluded from the SureSelect X-Chromosome Capture Kit because they were designated as pseudoautosomal or repetitive sequences (UCSC genome browser).

# Found one mutation that seems to contribute



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.

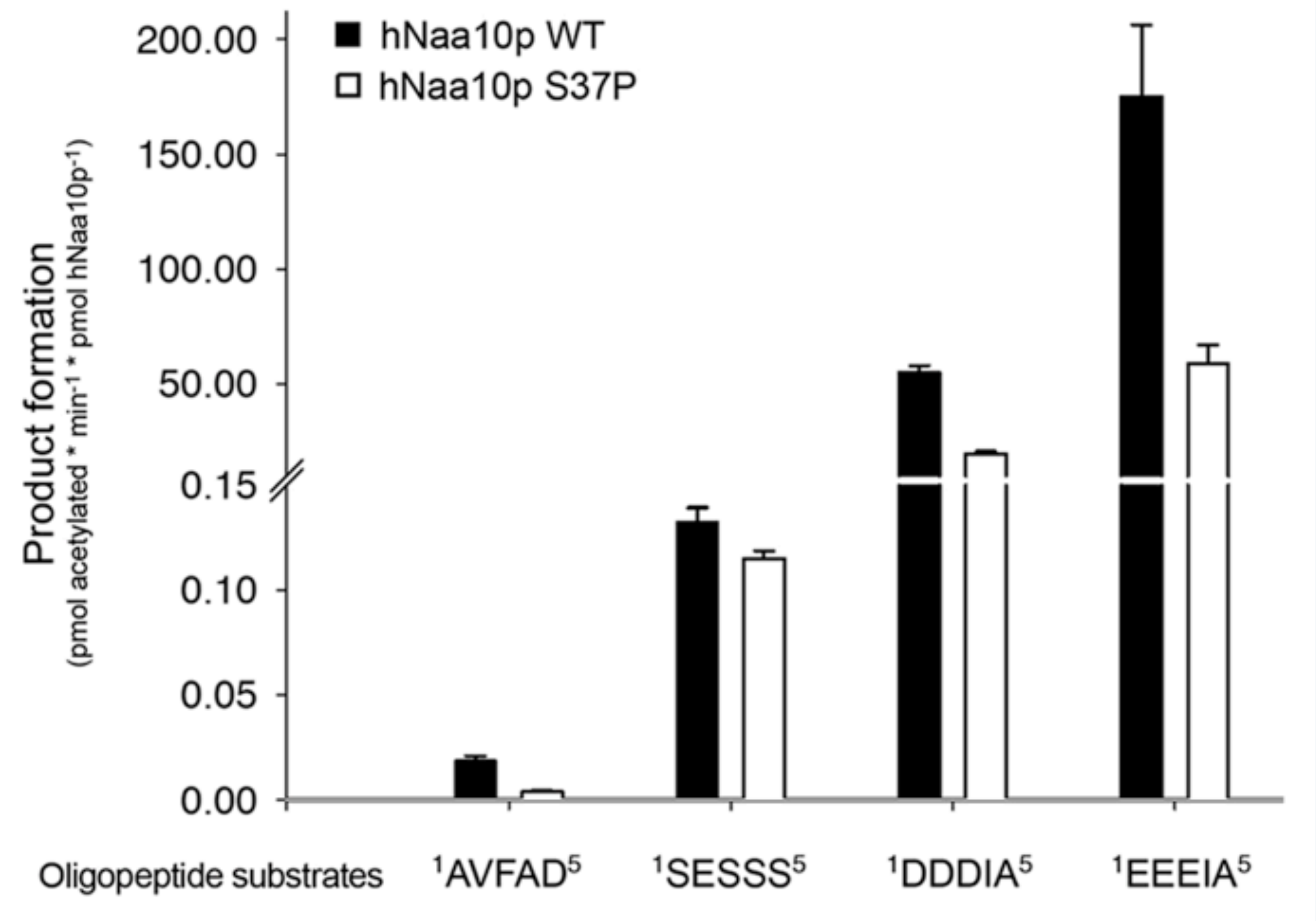
# The mutation disrupts the N-terminal acetylation machinery (NatA) in human cells.



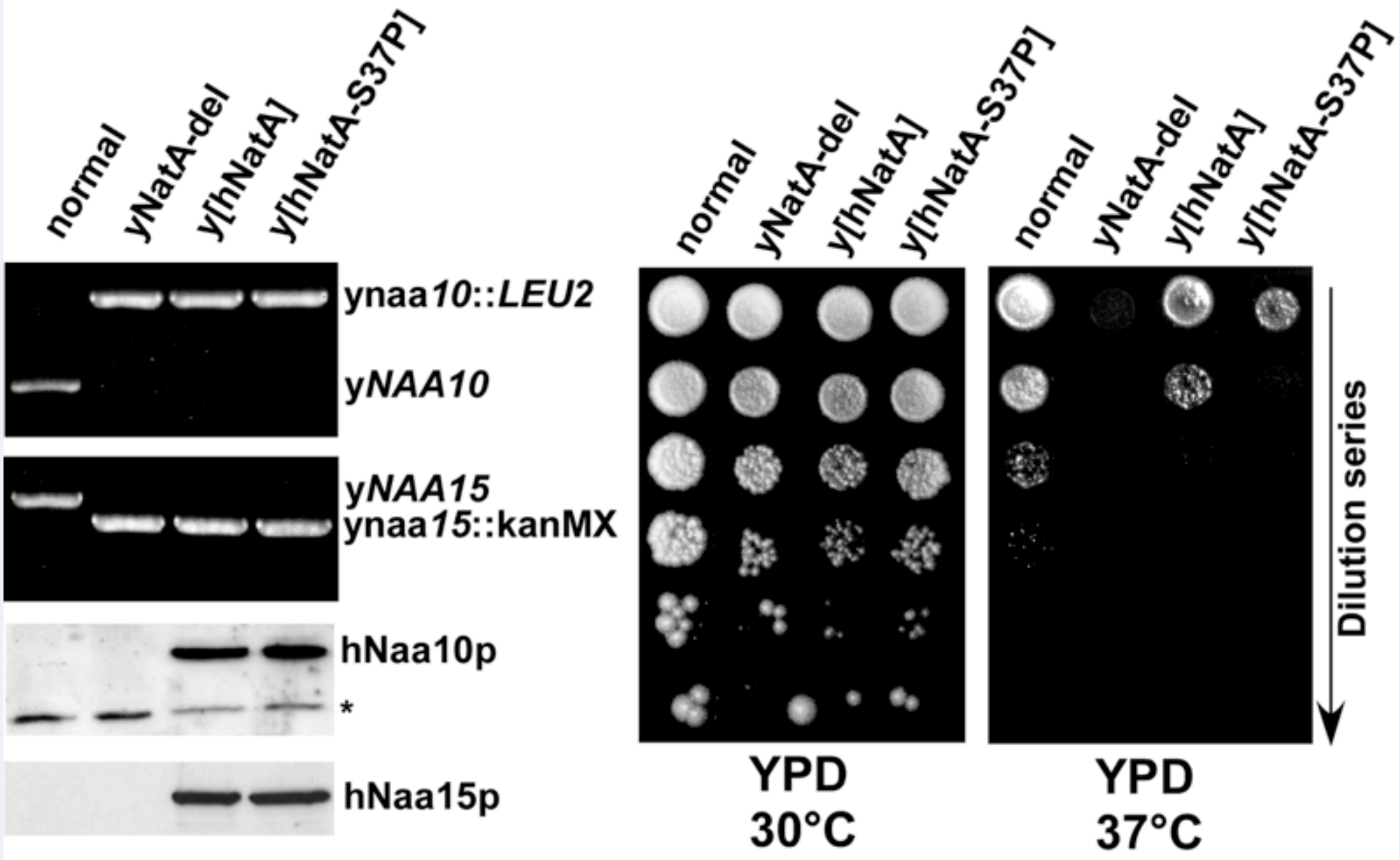
| NatA | NatE | NatB | NatC | NatD | NatF |
|---|---|---|---|---|---|
| Ser<br>Ala<br>Thr<br>Gly<br>Val<br>Asp/Glu* | Met-Leu | Met-Asp<br>Met-Glu<br>Met-Asn<br>Met-Gln | Met-Leu<br>Met-Ile<br>Met-Phe | Ser<br>(H2A/H4) | Met-Lys<br>Met-Ser<br>Met-Ala<br>Met-Met<br>Met-Leu |

Slide courtesy of Thomas Arnesen

# NAT activity of recombinant hNaa10p WT or p.Ser37Pro towards synthetic N-terminal peptides



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

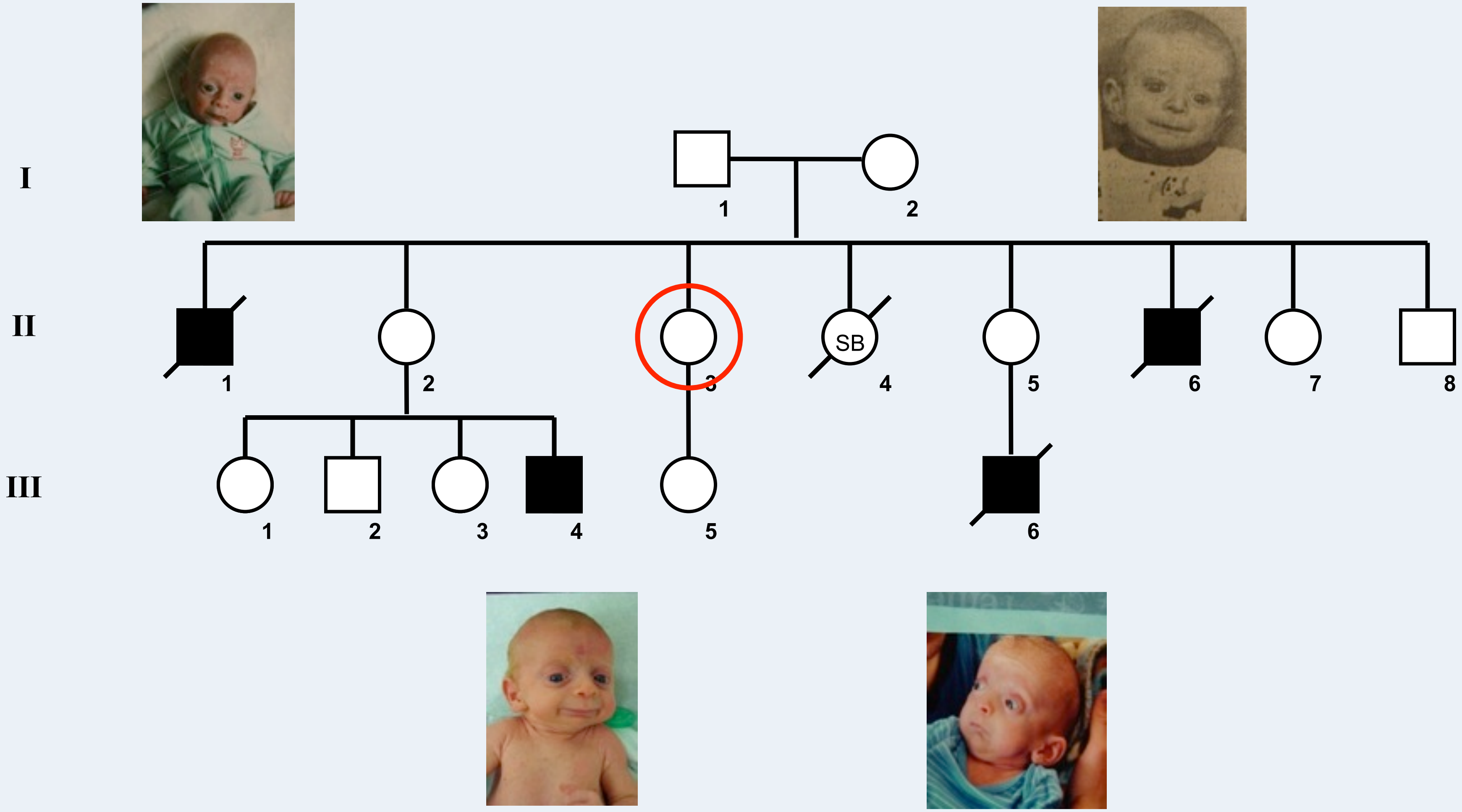


Unpublished data from Thomas Arnesen, do not further distribute.

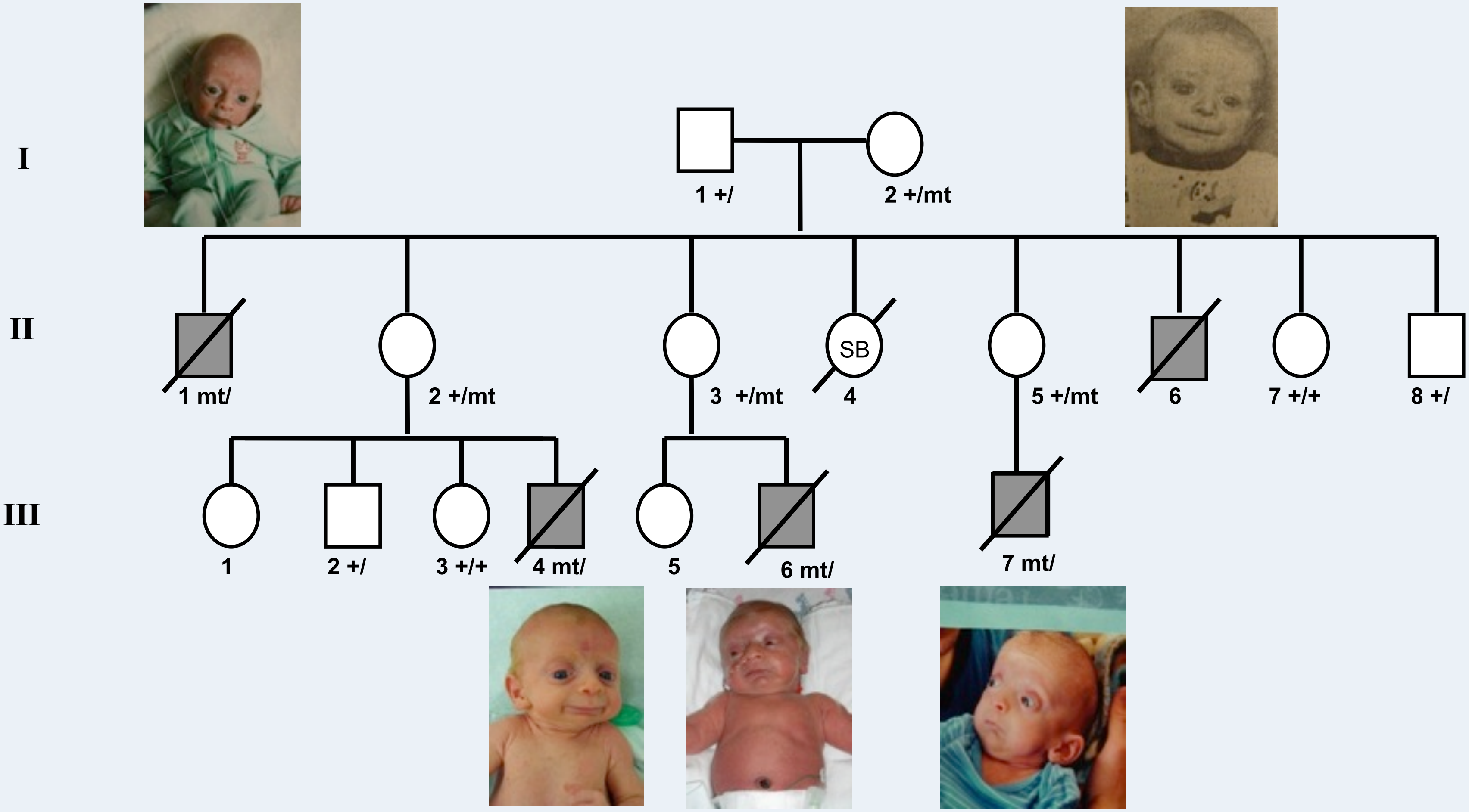
By November 2010, we had good functional data *in vitro* (bacterially expressed proteins) and *in vivo* (yeast, unpublished), leading me to believe we had identified the causative mutation.

*A new mother in the family informs me she is 4 months pregnant, with a boy!*

## The now pregnant mother-to-be is circled in red. Our Sanger Sequencing had shown her to be a carrier of the mutation.



# Family now, with five mutation-positive boys dying from the disease – Ogden Syndrome.



# PRIVACY and PROGRESS in Whole Genome Sequencing

Presidential Commission *for the* Study of Bioethical Issues

October 2012

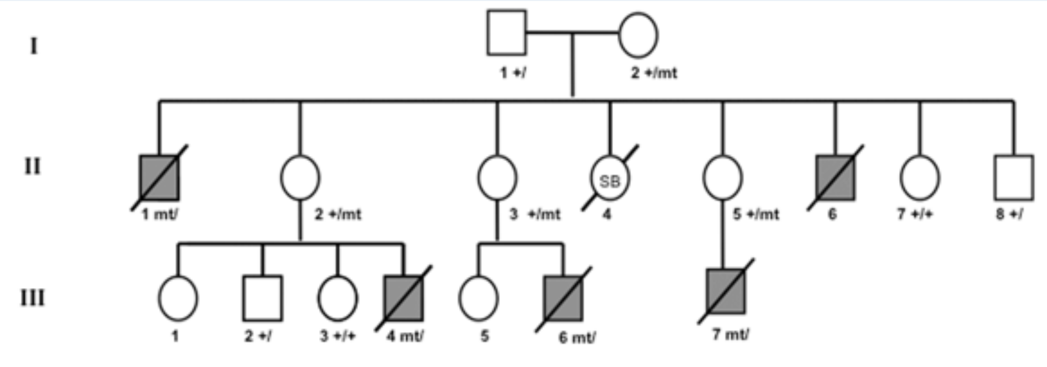
### Recommendation 4.1

**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.**

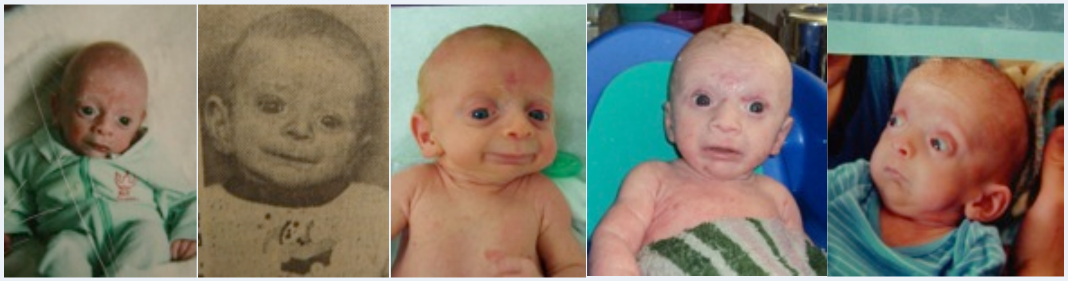
Performing all whole genome sequencing in CLIA-approved laboratories would remove one of the barriers to data sharing. It would help ensure that whole genome sequencing generates high-quality data that clinicians and researchers can use to draw clinically relevant conclusions. It would also ensure that 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]

# Ancestry Matters! - Ogden Syndrome

A

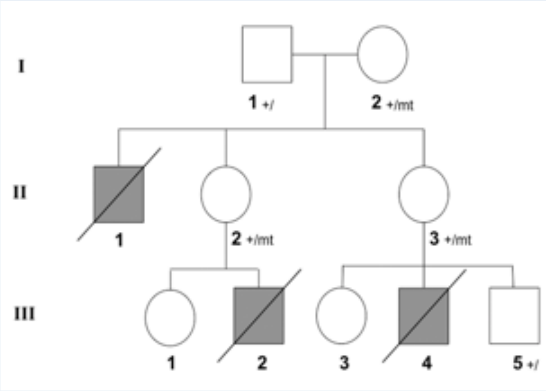


B

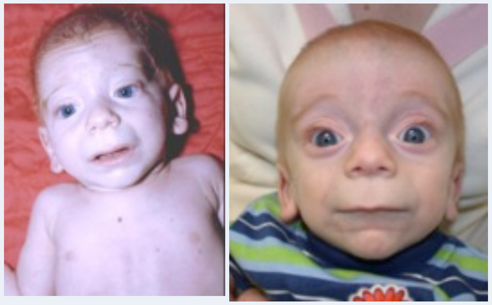


II-1 II-6 III-4 III-6 III-7

C



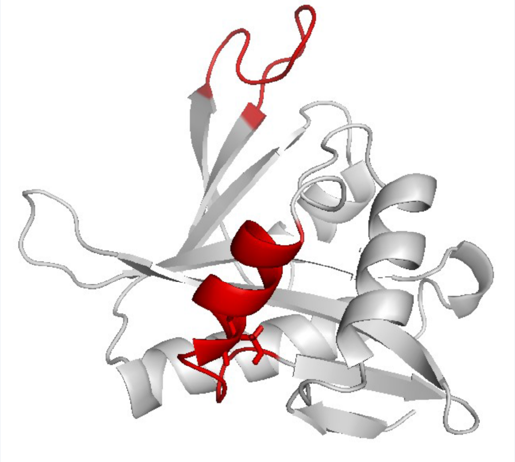
D



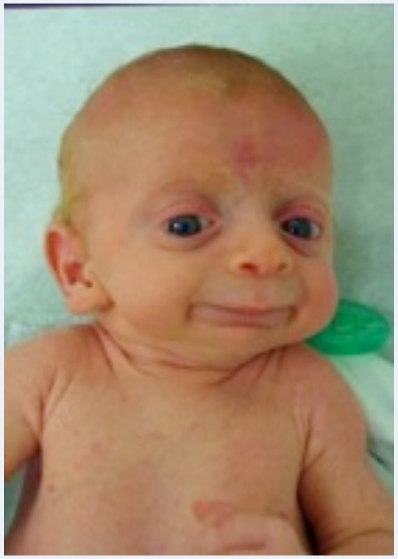
II-1 III-2

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.

# Big Question:



?

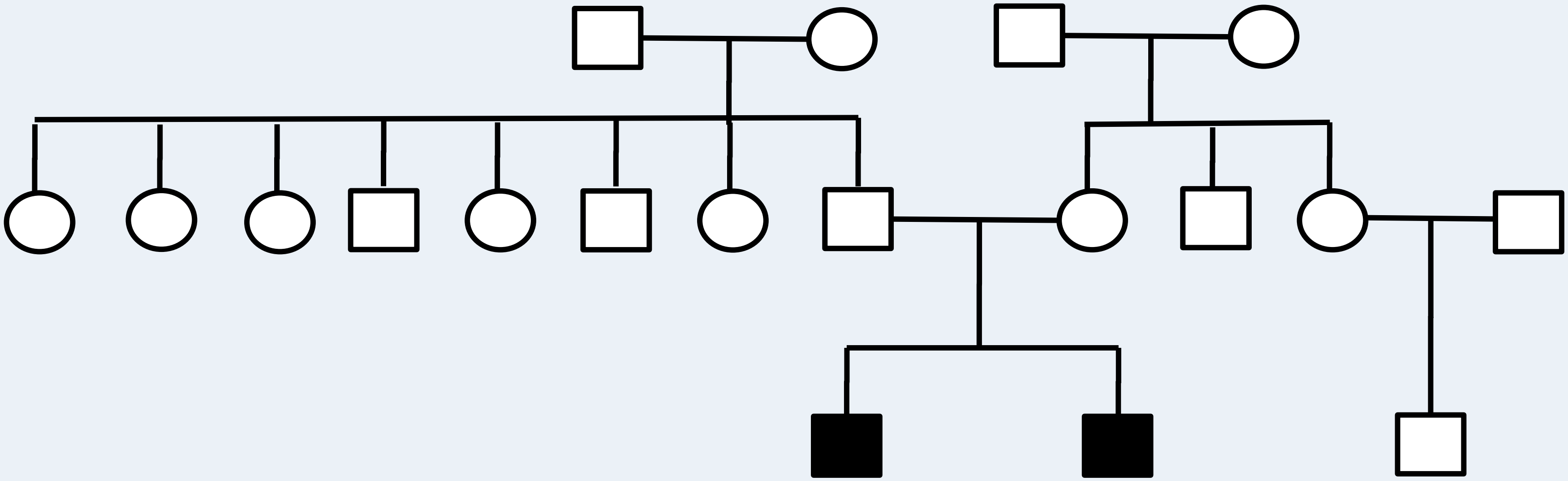


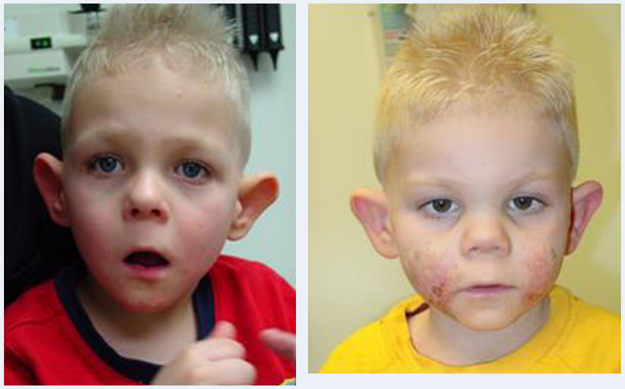
Simulated structure of S37P mutant

Max Doerfel

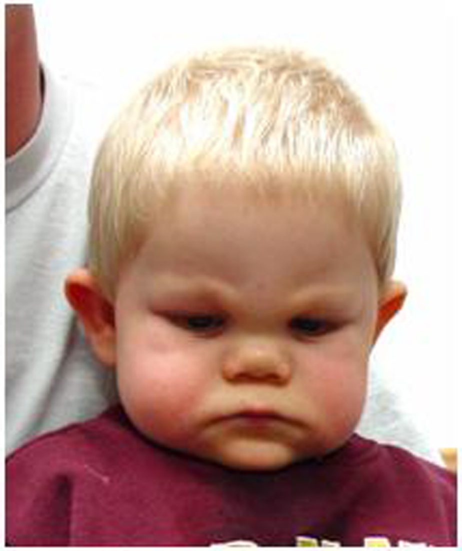
Yiyang Wu

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

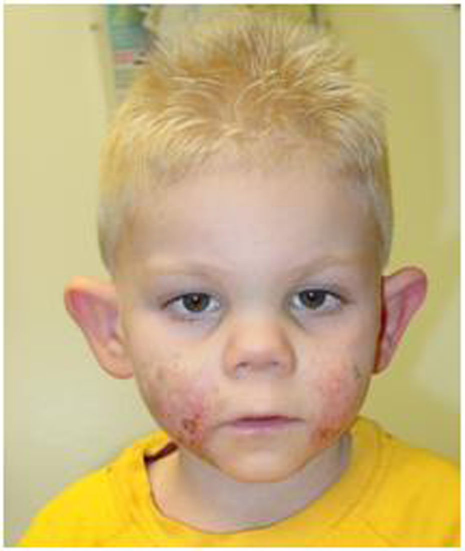




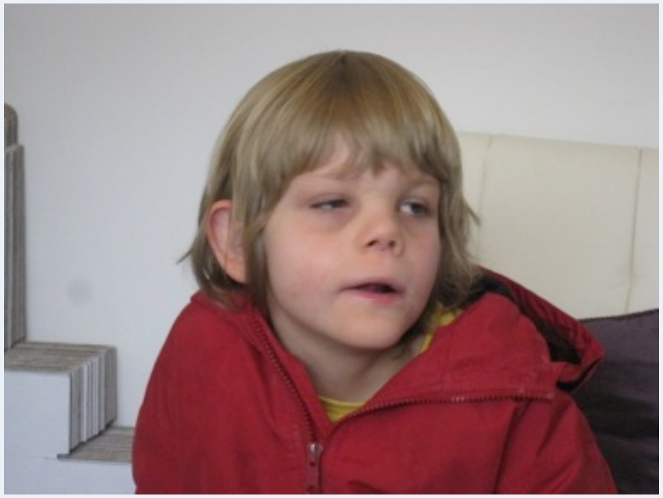
Could be X-linked, Autosomal Recessive, multi-allelic or polygenic threshold effect?



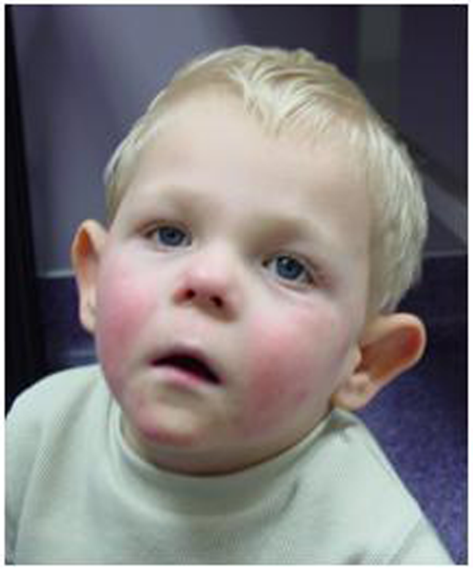
1.5 years old



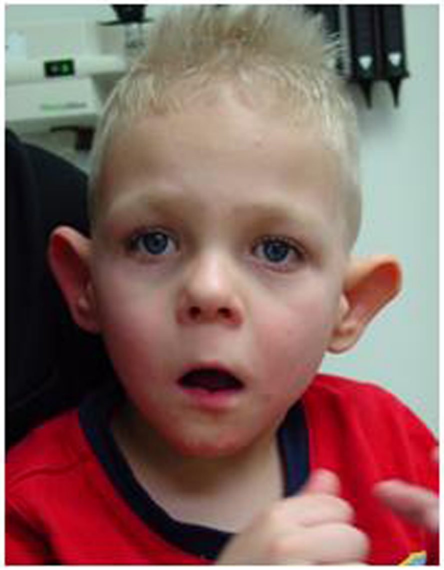
3.5 years old



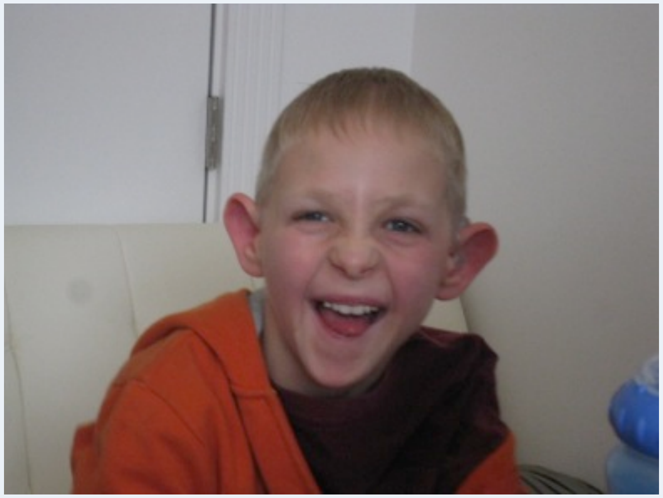
7 years old



3 years old



5 years old



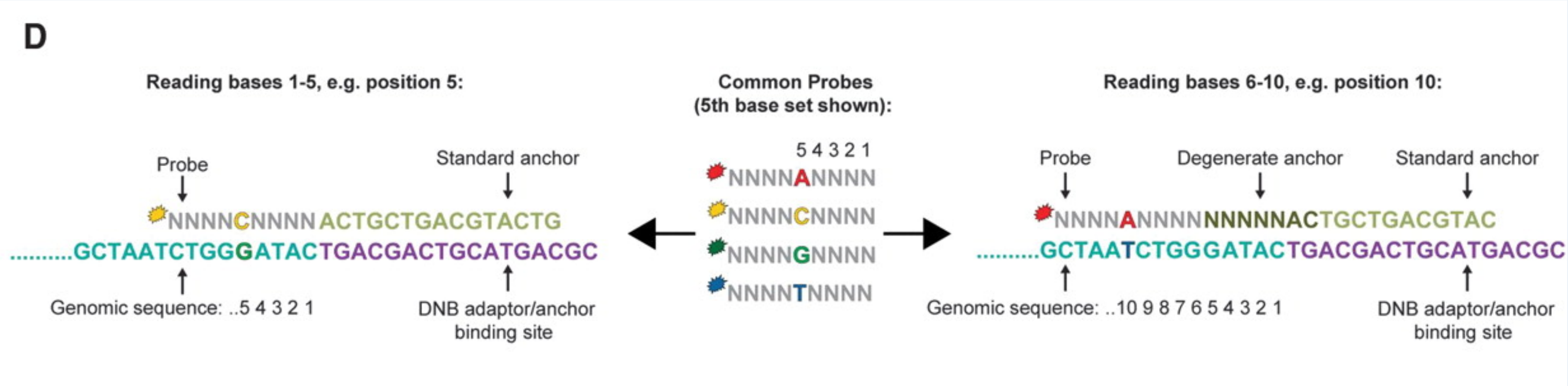
9 years old

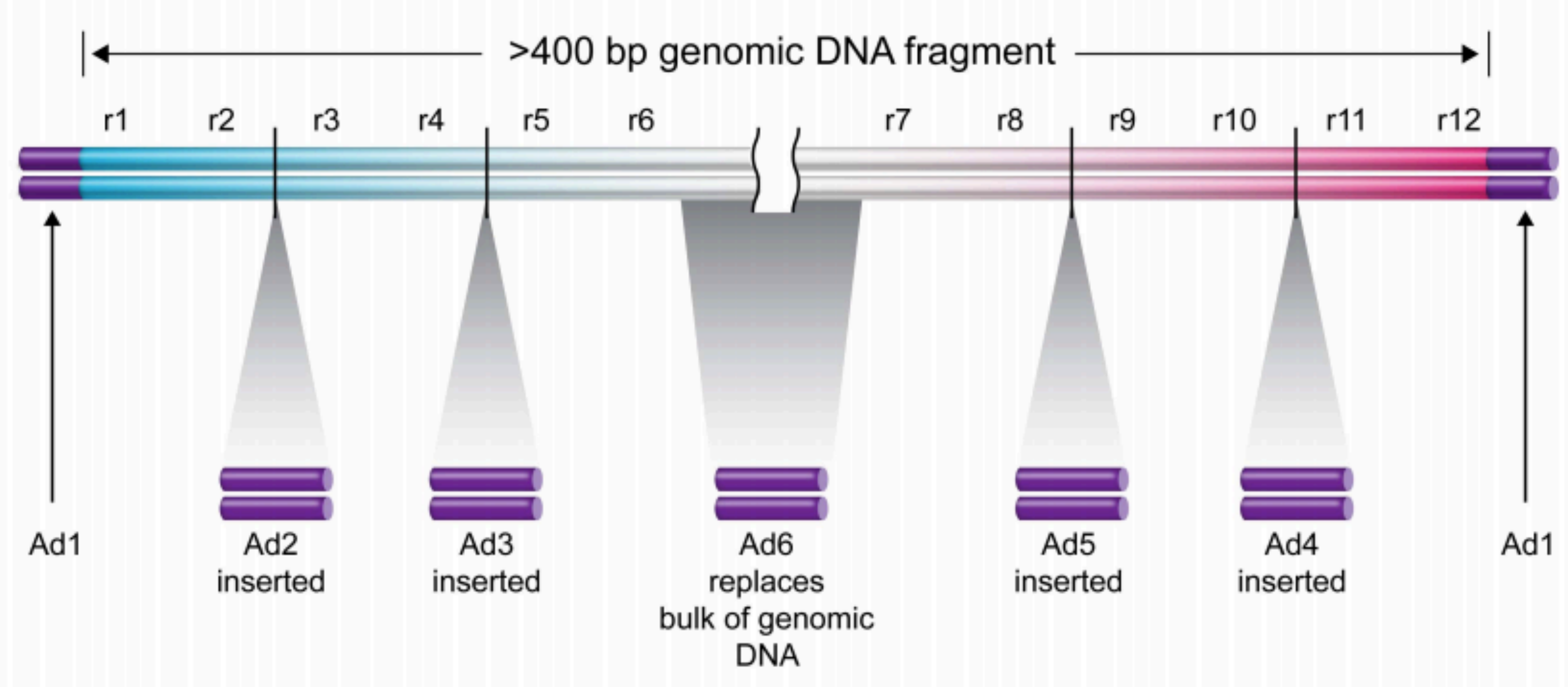
# Workup Ongoing for past 10 years

- Numerous genetic tests negative, including negative for Fragile X and many candidate genes.
- No obvious pathogenic CNVs – 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

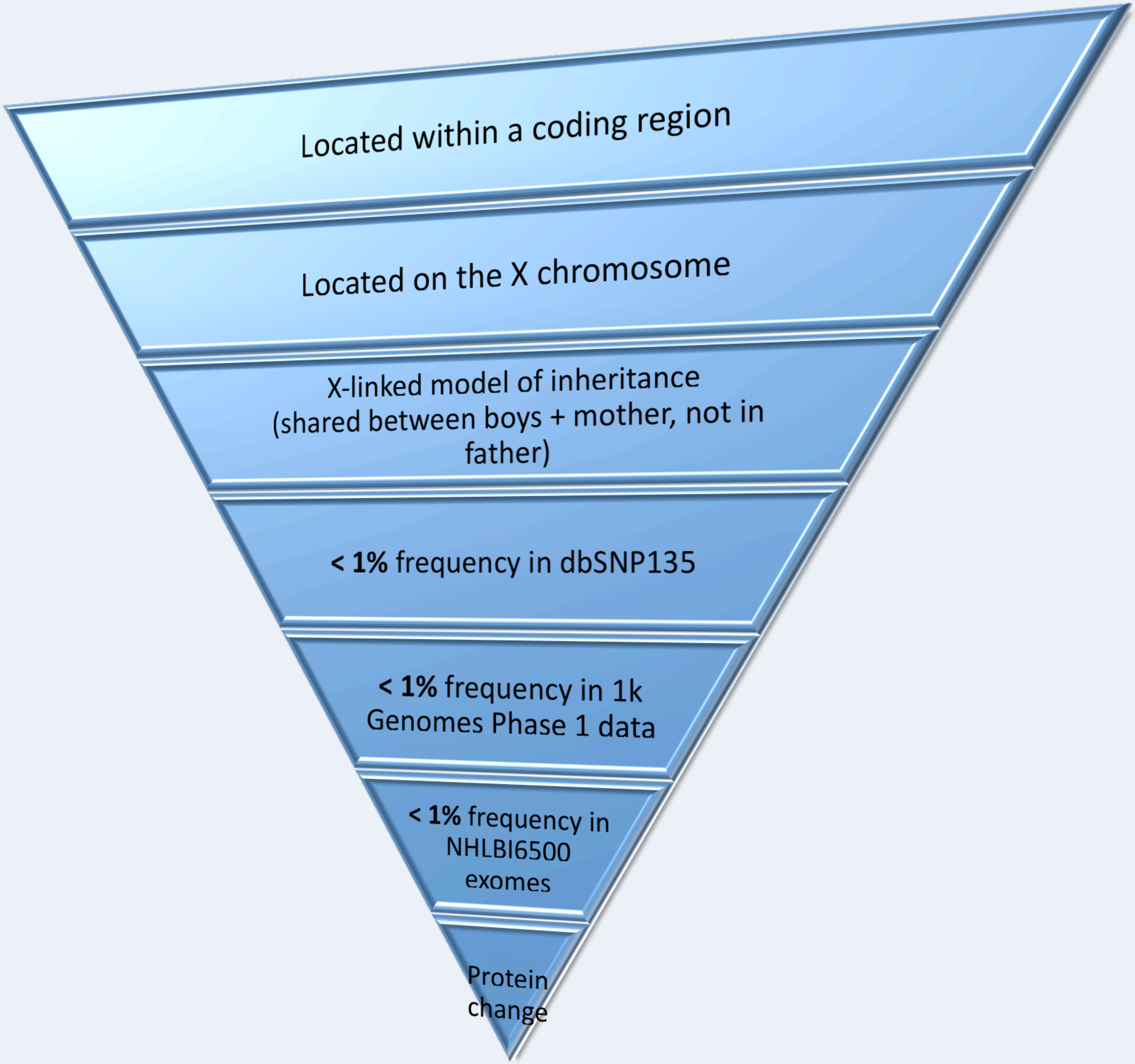
# Complete Genomics chemistry - combinatorial probe anchor ligation (cPAL)


D
Reading bases 1-5, e.g. position 5:
Probe
Standard anchor
NNNNCNNNN ACTGCTGACGTACTG
..........GCTAATCTGGGATACTGACGACTGCATGACGC
Genomic sequence: ..5 4 3 2 1
DNB adaptor/anchor binding site
Common Probes (5th base set shown):
5 4 3 2 1
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



## Analysis with VAAST, ANNOVAR, and Golden Helix SVS

- No obvious pathogenic CNVs in both brothers.
- One putative set of interesting compound heterozygote mutations in both brothers, but validating now with Sanger sequencing. The mechanism of this is uncertain?
- No other obvious disease-contributory autosomal recessive SNVs or indels in protein-coding regions in both brothers.



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*

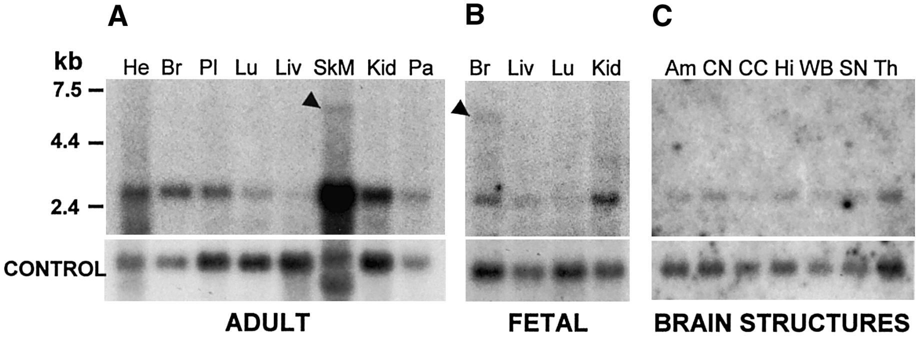
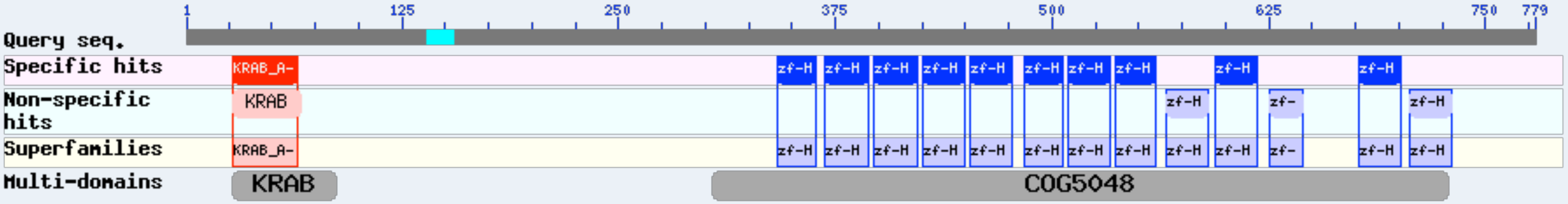
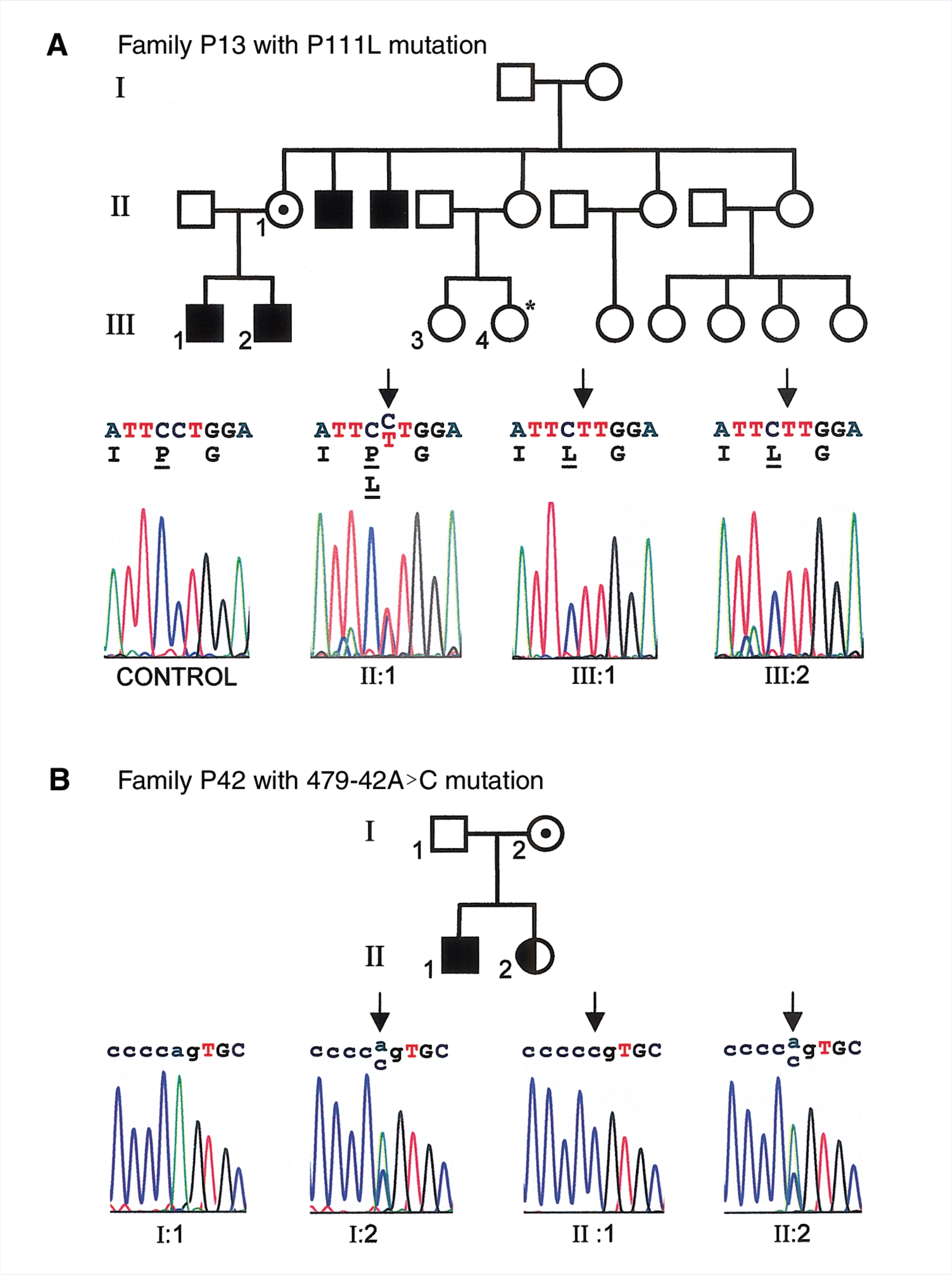


**Figure 6** Northern blot hybridization of *ZNF41*, by use of a probe corresponding to nucleotides 621–1099 of *ZNF41* transcript variant 1. *A*, Adult tissues (left to right): heart, brain, placenta, lung, liver, skeletal muscle, kidney, and pancreas. *B*, Fetal tissues (left to right): brain, lung, 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.



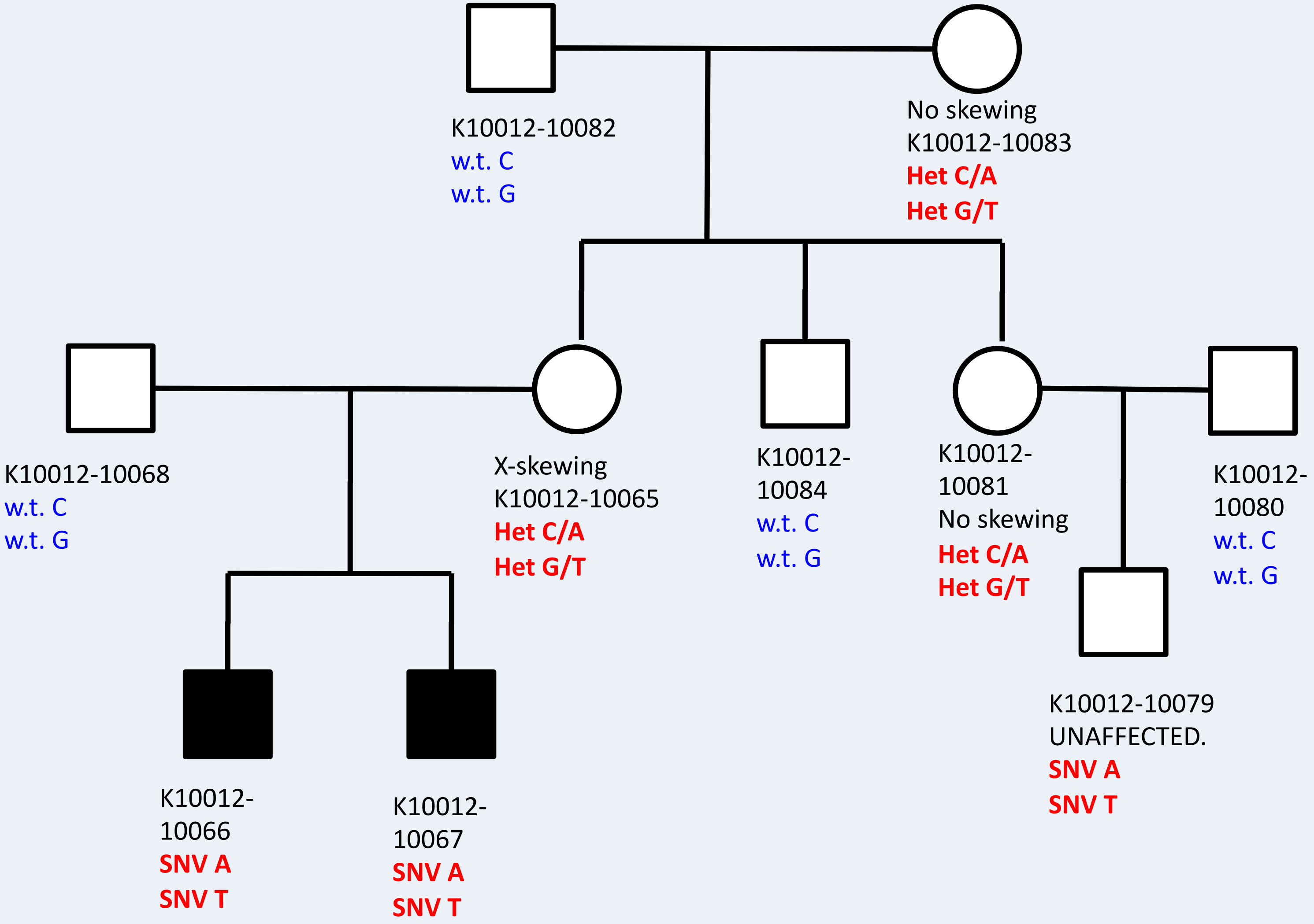
"The P111L change also found in two "male controls" (EVS server, ESP6500).

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



K10012-10082
w.t. C
w.t. G

No skewing
K10012-10083
**Het C/A**
**Het G/T**

K10012-10068
w.t. C
w.t. G

X-skewing
K10012-10065
**Het C/A**
**Het G/T**

K10012-10084
w.t. C
w.t. G

K10012-10081
No skewing
**Het C/A**
**Het G/T**

K10012-10080
w.t. C
w.t. G

K10012-10066
**SNV A**
**SNV T**

K10012-10067
**SNV A**
**SNV T**

K10012-10079
UNAFFECTED.
**SNV A**
**SNV T**

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

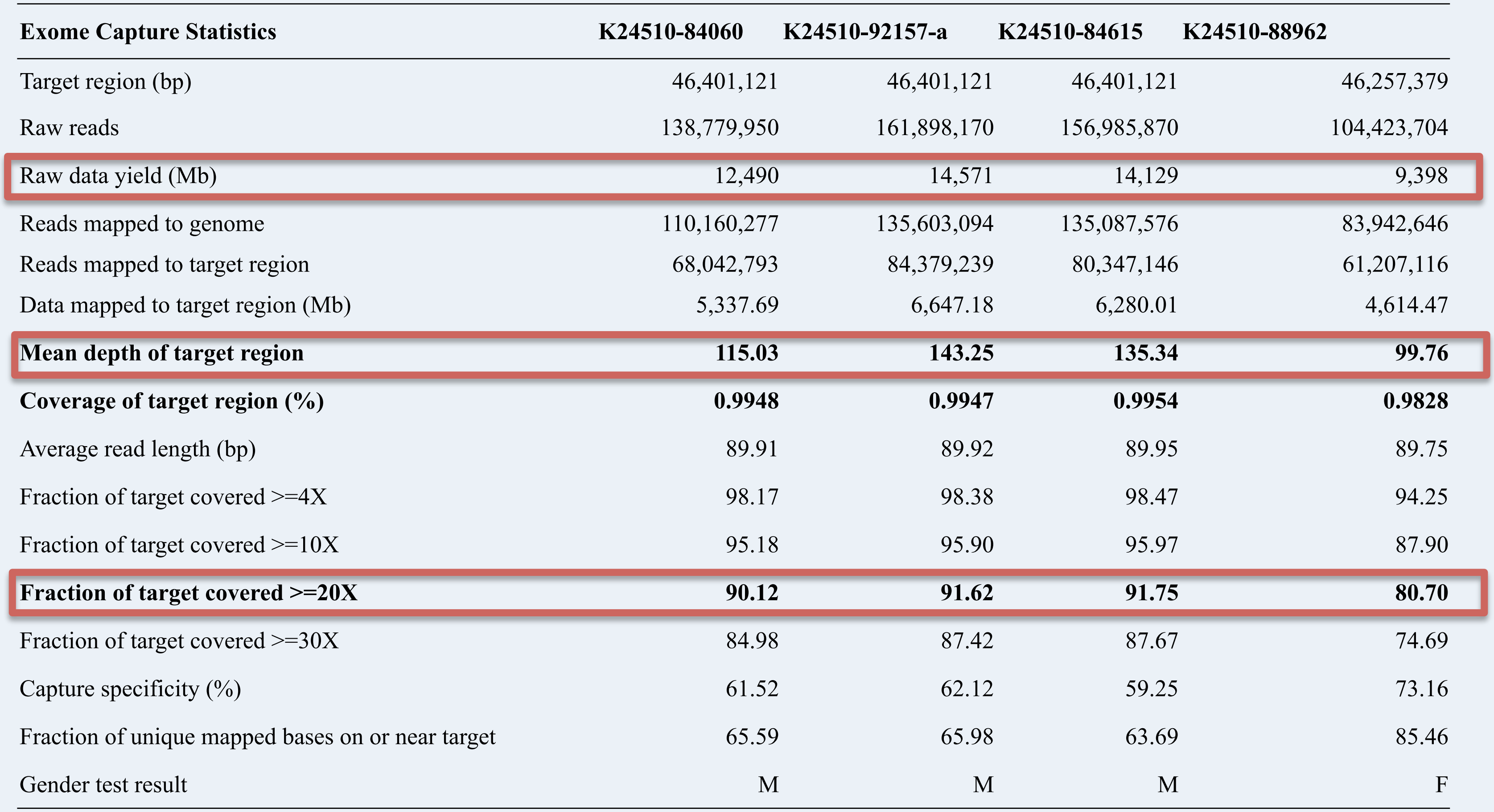
**Proving Whether this Mutation is Necessary and Sufficient to result in the phenotype.**

- 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.

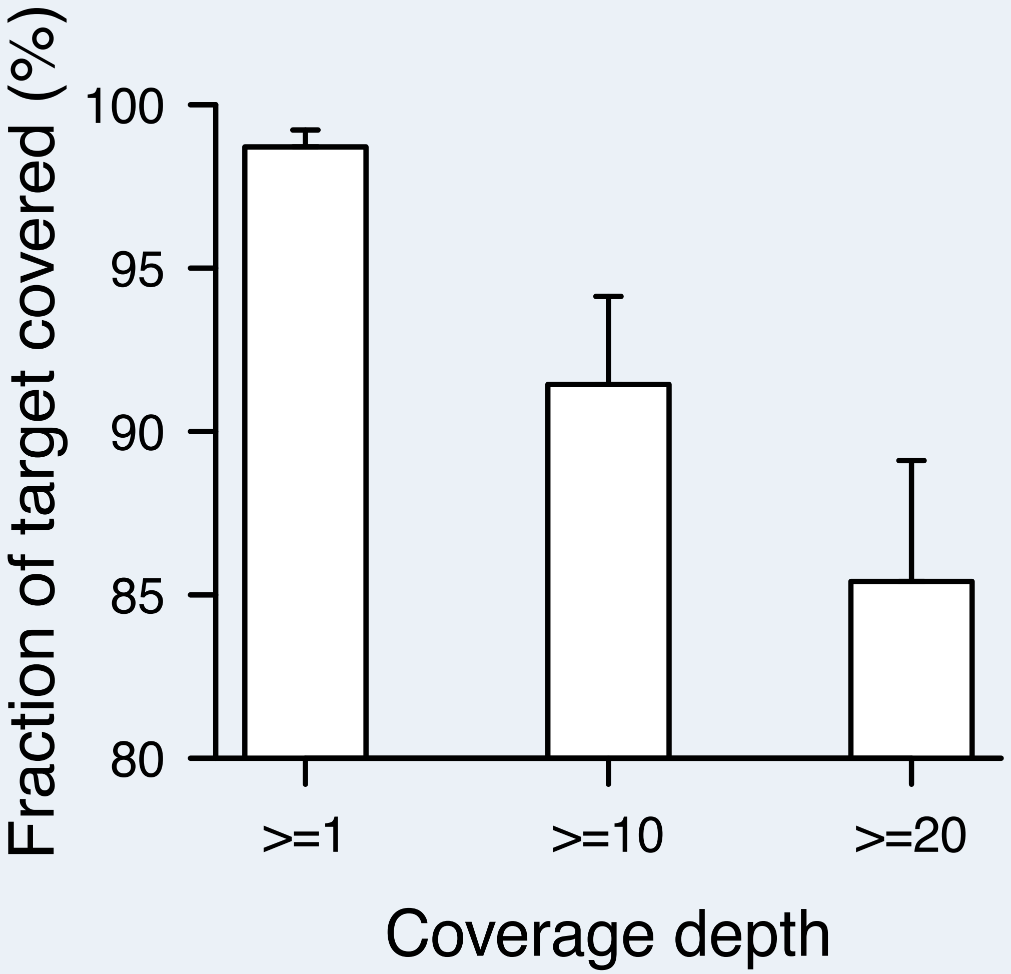
- But what about the False Negative Rate in Complete Genomes data? What did we miss?

# 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

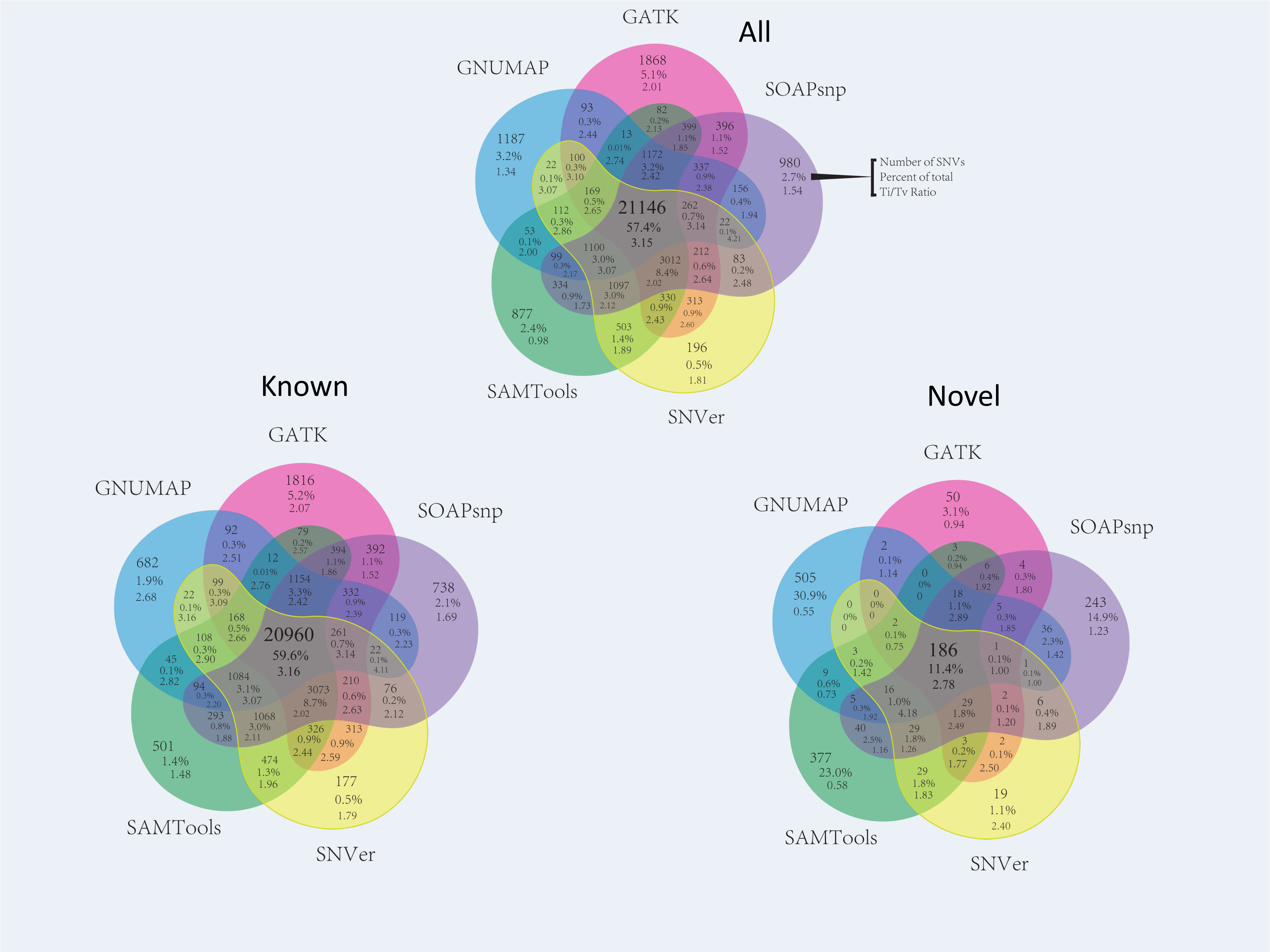
1) BWA - **GATK** (version 1.5) with recommended parameters (GATK IndelRealigner, base quality scores were re-calibrated by GATK Table Recalibration tool. Genotypes called by GATK UnifiedGenotyper. For SNVs and indels.

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

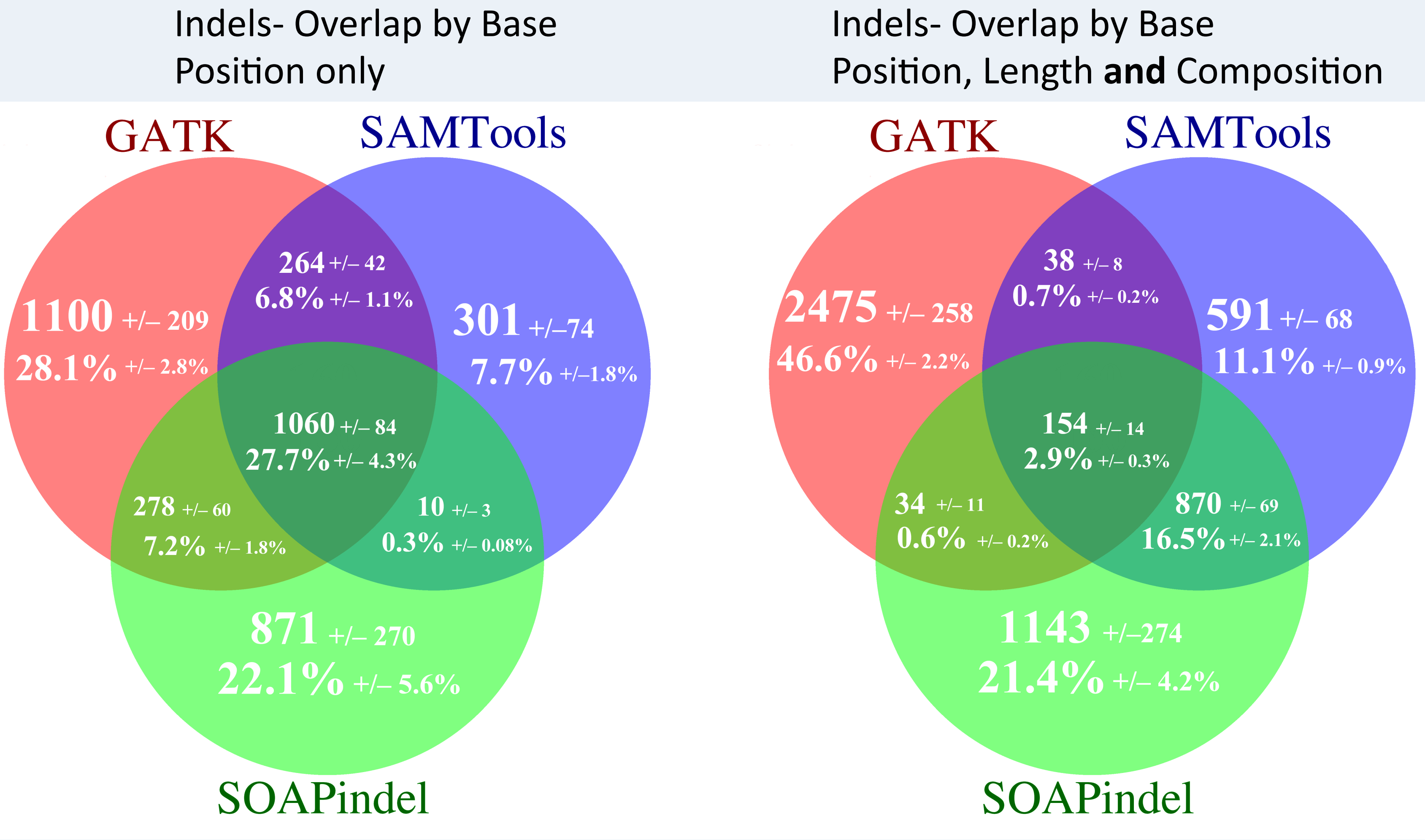
**3)** **SOAP**-Align – SOAPsnp for SNVs– and BWA-SOAPindel (adopts local assembly based on an extended de Bruijn graph) for indels.

**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
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
22
0.1%
3.07
100
0.3%
3.10
169
0.5%
2.65
21146
57.4%
3.15
262
0.7%
3.14
156
0.4%
1.94
112
0.3%
2.86
22
0.1%
4.21
53
0.1%
2.00
99
0.3%
2.17
1100
3.0%
3.07
3012
8.4%
2.02
212
0.6%
2.64
83
0.2%
2.48
334
0.9%
1.73
1097
3.0%
2.12
330
0.9%
2.43
313
0.9%
2.60
877
2.4%
0.98
503
1.4%
1.89
196
0.5%
1.81
Number of SNVs
Percent of total
Ti/Tv Ratio
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
99
0.3%
3.09
22
0.1%
3.16
168
0.5%
2.66
20960
59.6%
3.16
261
0.7%
3.14
119
0.3%
2.23
108
0.3%
2.90
22
0.1%
4.11
45
0.1%
2.82
94
0.3%
2.20
1084
3.1%
3.07
3073
8.7%
2.02
210
0.6%
2.63
76
0.2%
2.12
293
0.8%
1.88
1068
3.0%
2.11
326
0.9%
2.44
313
0.9%
2.59
501
1.4%
1.48
474
1.3%
1.96
177
0.5%
1.79
Novel
GATK
GNUMAP
SOAPsnp
SAMTools
SNVer
50
3.1%
0.94
2
0.1%
1.14
3
0.2%
0.94
6
0.4%
1.92
4
0.3%
1.80
505
30.9%
0.55
0
0%
0
18
1.1%
2.89
243
14.9%
1.23
0
0%
0
0
0%
0
2
0.1%
0.75
5
0.3%
1.85
36
2.3%
1.42
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

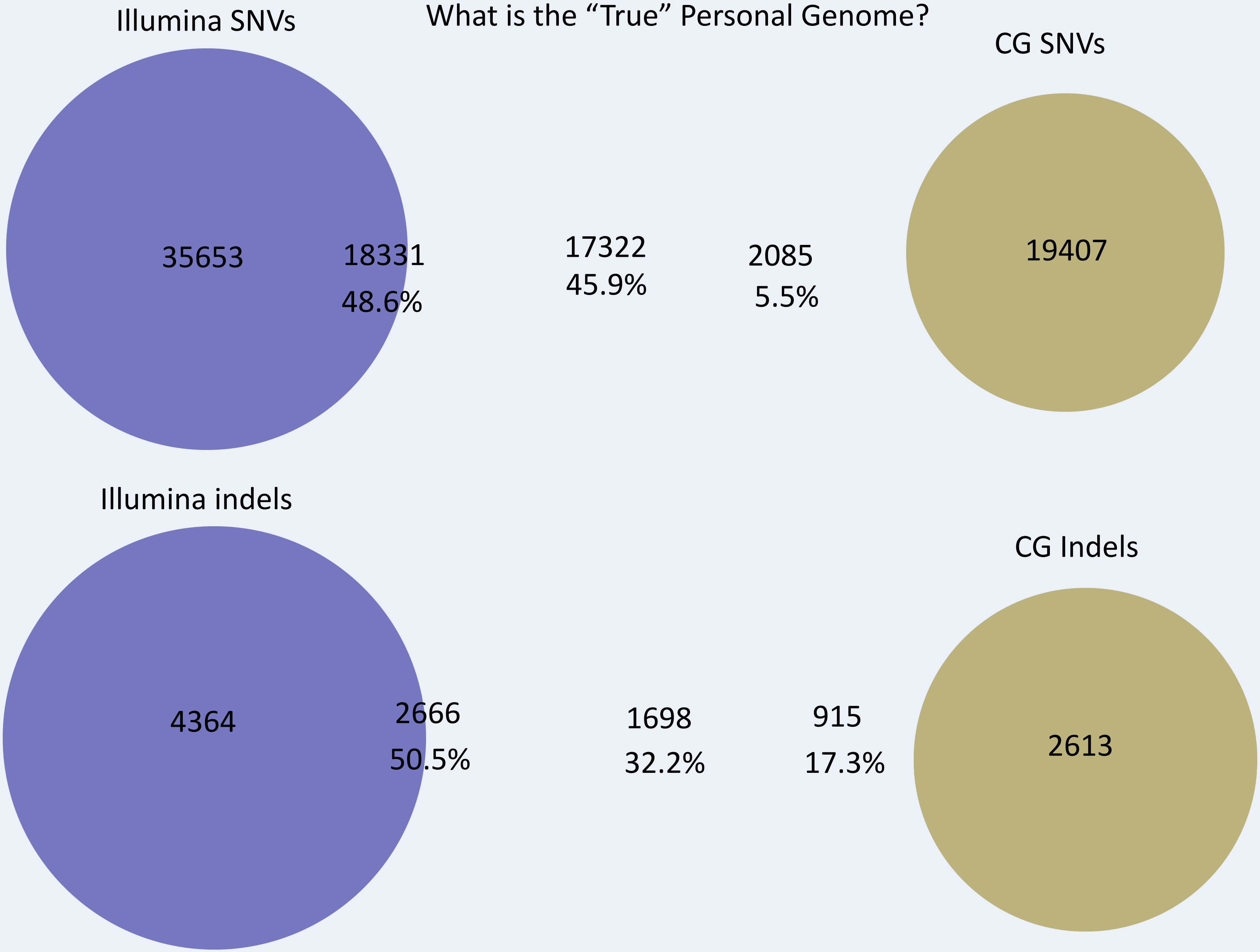
**Total mean overlap, plus or minus one standard deviation, observed between three indel calling pipelines: GATK, SOAP-indel, and SAMTools. a)** Mean overlap when indel position was the only necessary agreement criterion. **b)** Mean overlap when indel position, base length and base composition were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

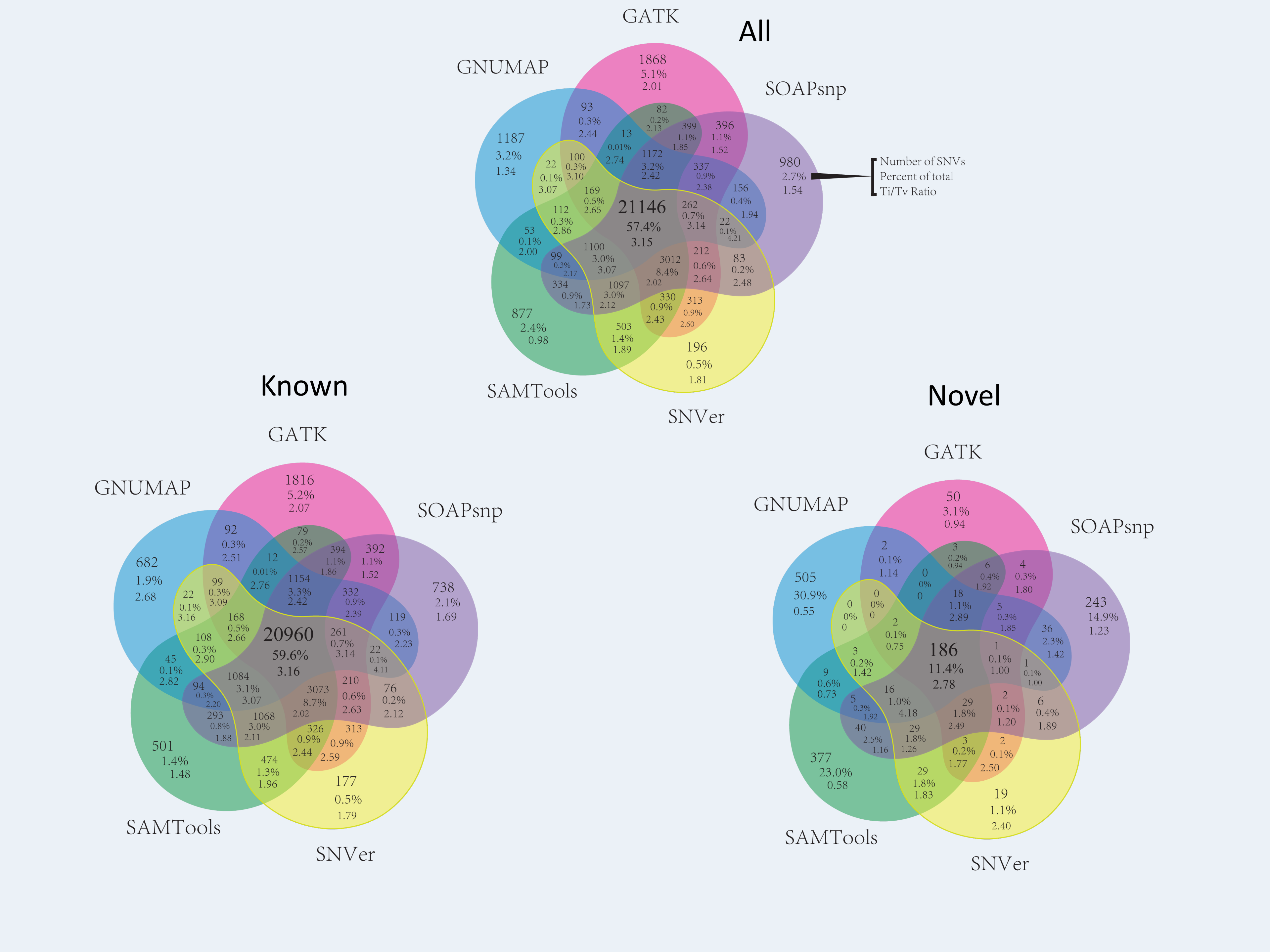
- How reliable are variants that are uniquely called by individual pipelines?

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

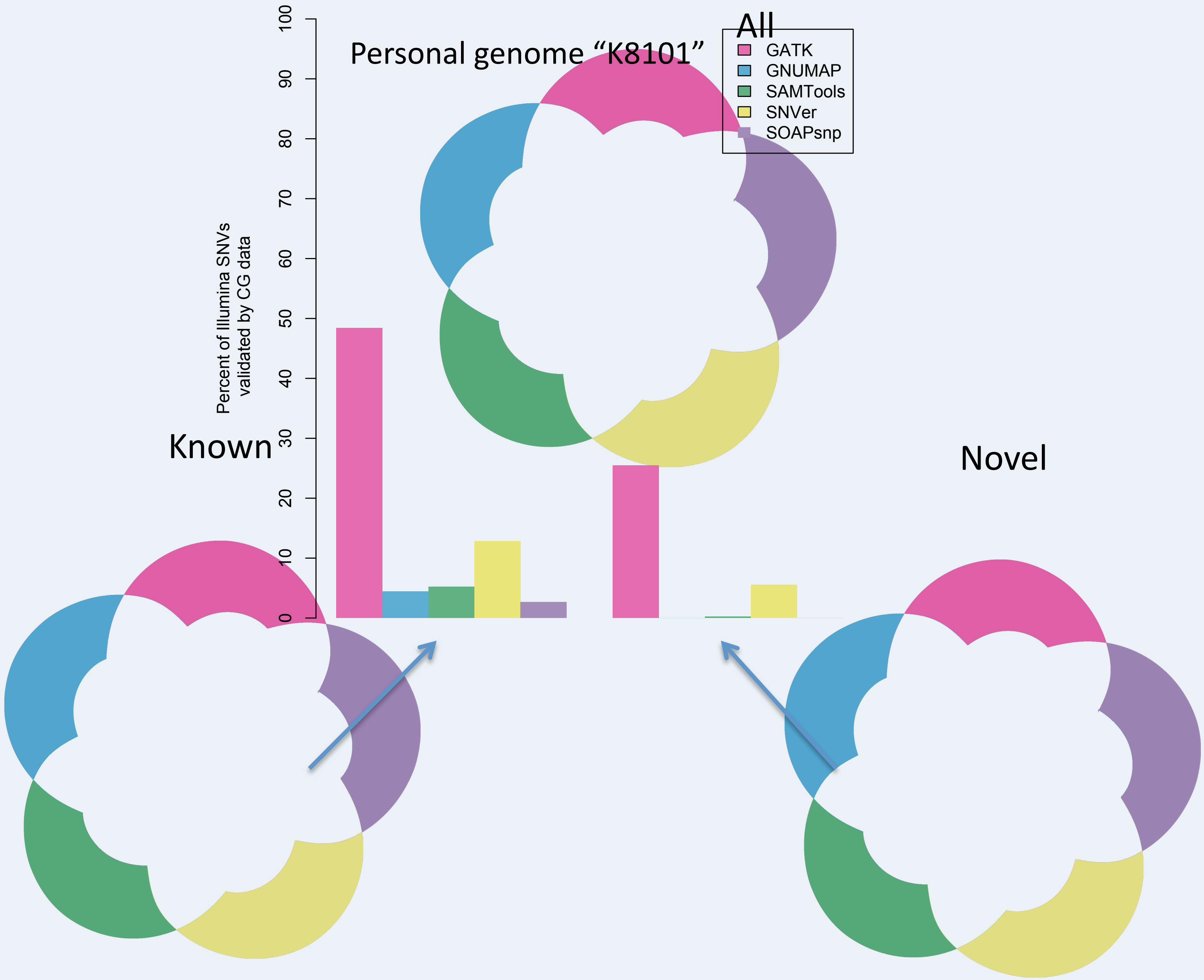
# Cross validation using orthogonal sequencing technology (Complete Genomics)

# What is the “True” Personal Genome?




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
100 0.3% 3.10
22 0.1% 3.07
1172 3.2% 2.42
337 0.9% 2.38
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
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
99 0.3% 3.09
22 0.1% 3.16
1154 3.3% 2.42
332 0.9% 2.39
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
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
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

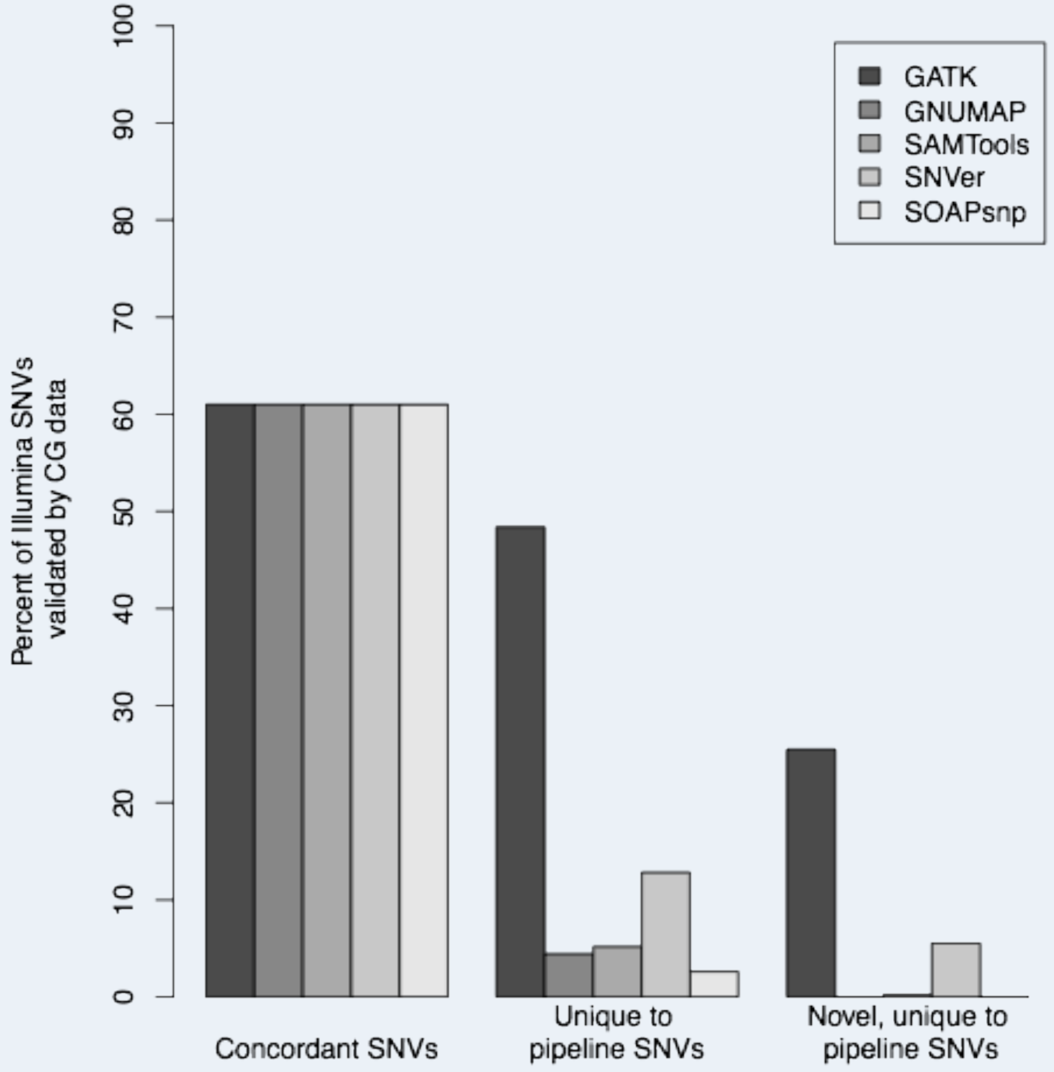
# Personal genome “K8101”



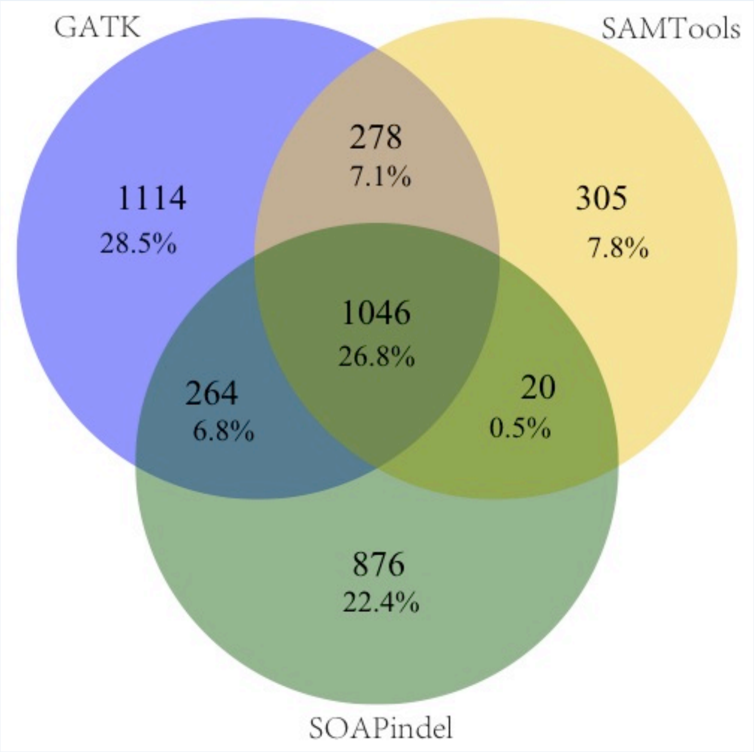
# Higher Validation of SNVs with the BWA-GATK pipeline

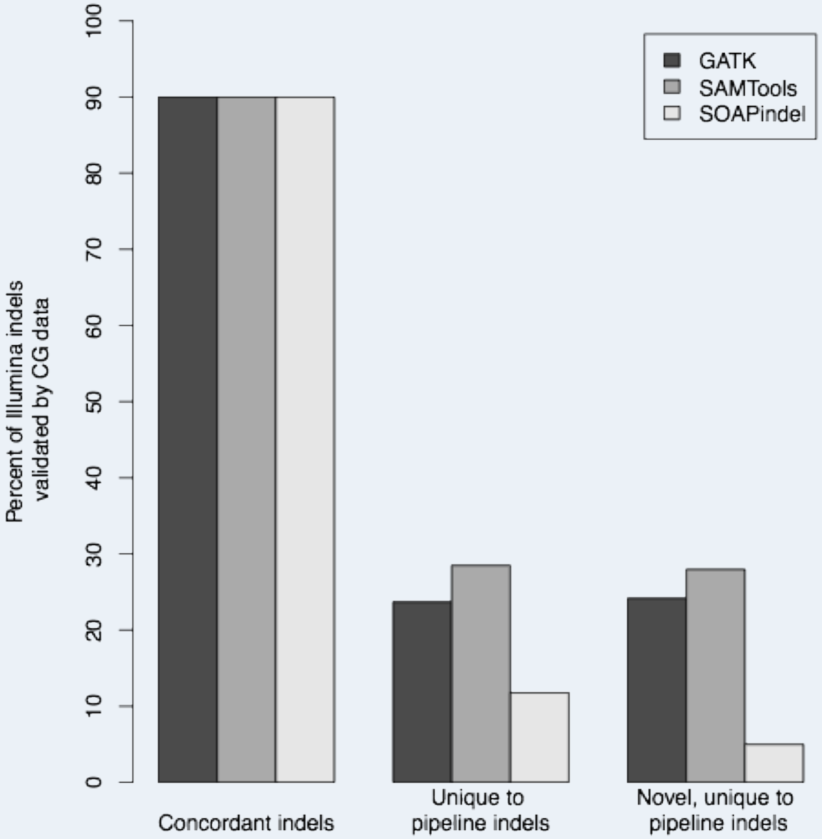
- Reveals higher validation rate of unique-to-pipeline variants, as well as uniquely discovered novel variants, for the variants called by BWA-GATK, in comparison to the other 4 pipelines (including SOAP).

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

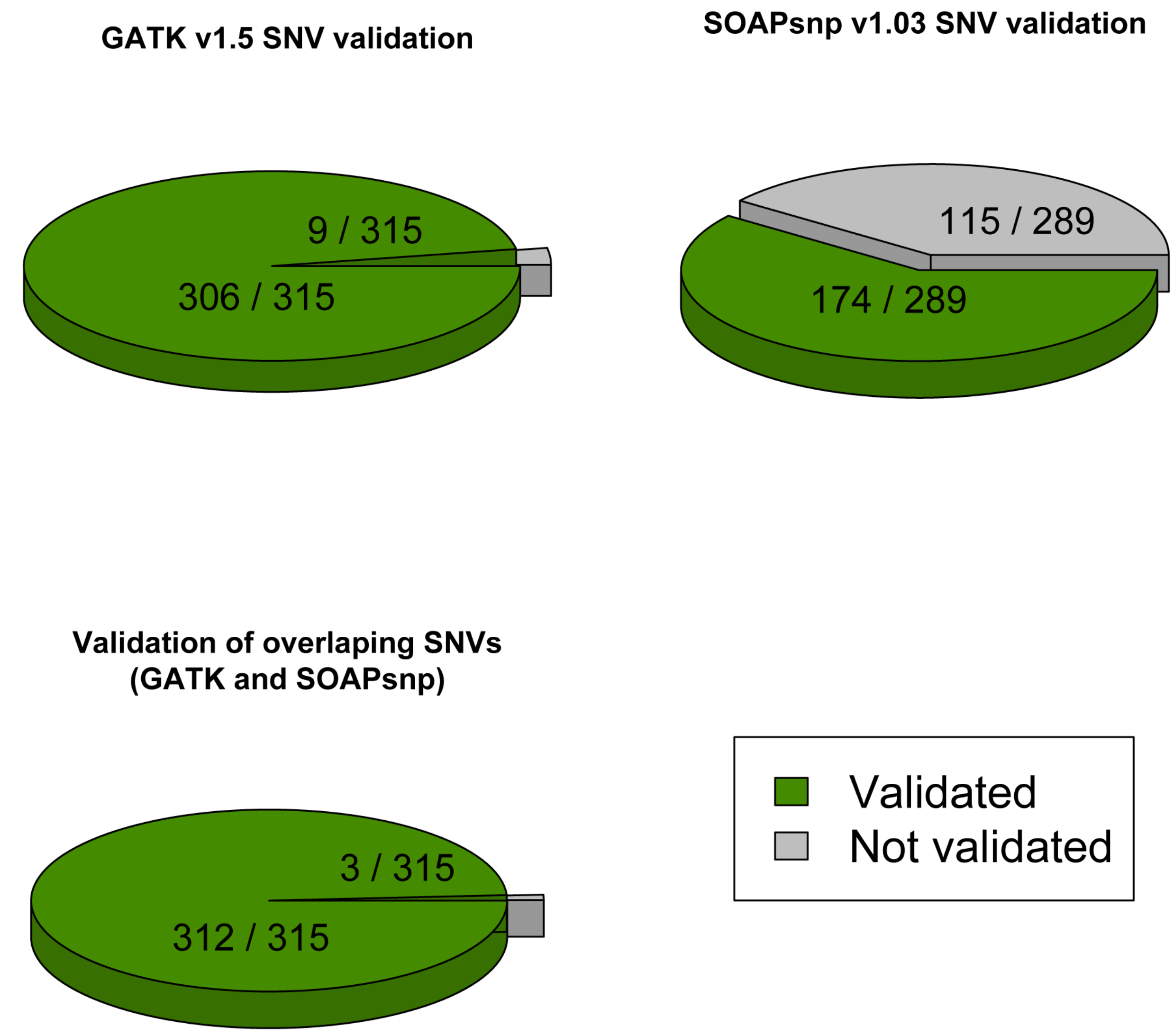


# Validating Indels with Complete Genomics Data for the 3 pipelines





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

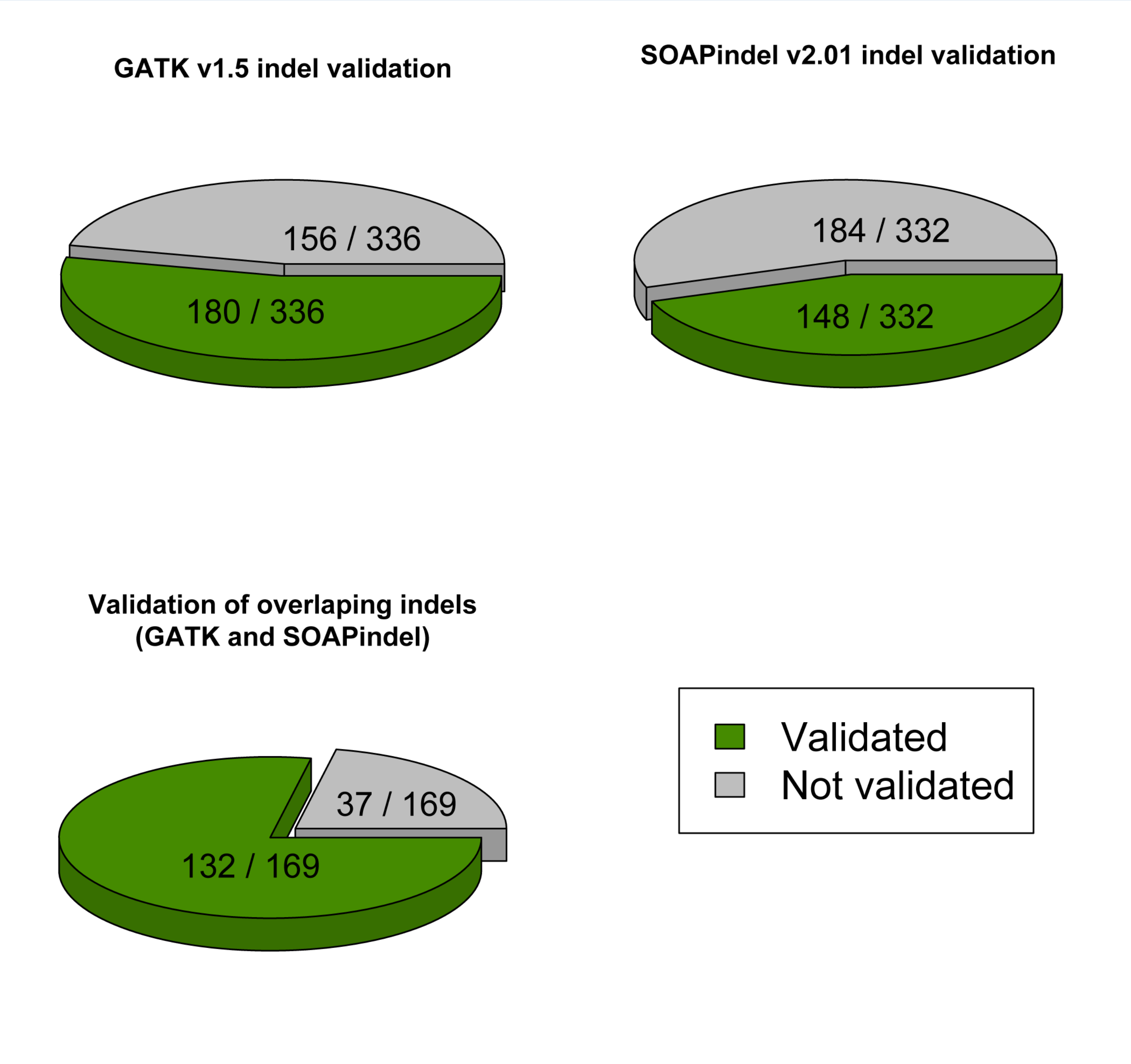


GATK v1.5 SNV validation

SOAPsnp v1.03 SNV validation

Validation of overlaping SNVs
(GATK and SOAPsnp)

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



GATK v1.5 indel validation

SOAPindel v2.01 indel validation

Validation of overlaping indels (GATK and SOAPindel)

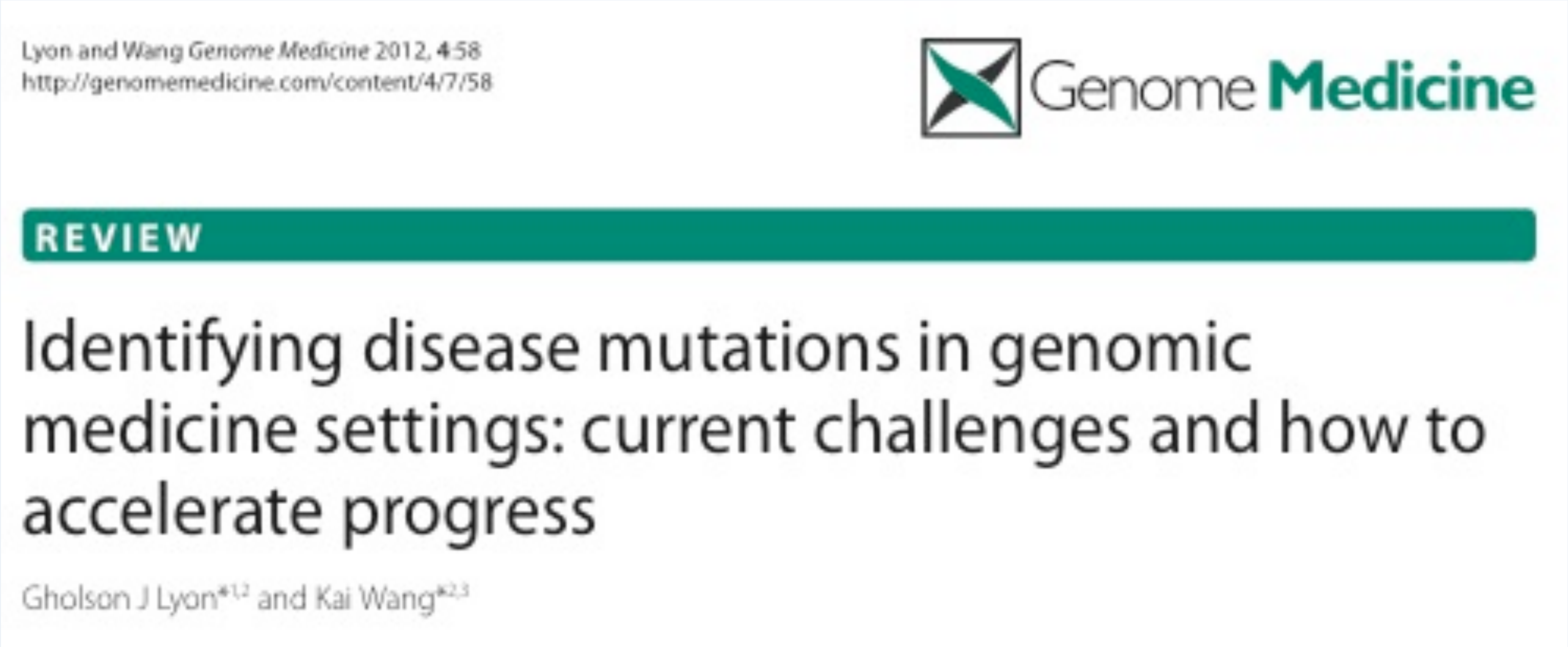
# Clinical Validity?

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 for thousands of volunteer families.

ancestry.com

PatientsLikeMe

23andMe



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

Genome **Medicine**

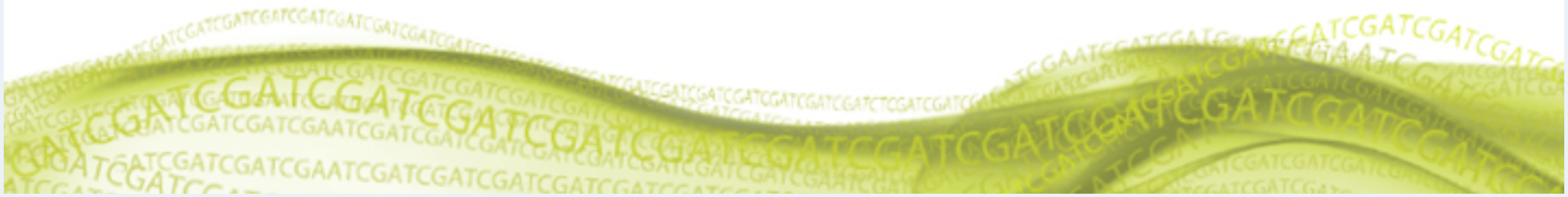
REVIEW

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

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

# 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

# Clinical Validity with Worldwide Human Genetic Variation "database"?

PatientsLikeMe

# Conclusions

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

# 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

THE END

Extra Slides to follow this last slide.

# “To Cause”, or “Causative”

- **Koch’s postulates for the 21st century (from Wikipedia)**
- Koch’s postulates have played an important role in microbiology, yet they have major limitations. For example, Koch was well aware that in the case of cholera, the causal agent, *Vibrio cholerae*, could be found in both sick and healthy people, invalidating his first postulate. Furthermore, viral diseases were not yet discovered when Koch formulated his postulates, and there are many viruses that do not cause illness in all infected individuals, a requirement of the first postulate.
- More recently, modern nucleic acid-based microbial detection methods have made Koch’s original postulates even less relevant. These nucleic acid-based methods make it possible to identify microbes that are associated with a disease, but in many cases the microbes are uncultivable. Also, nucleic acid-based detection methods are very sensitive, and they can often detect the very low levels of viruses that are found in healthy people without disease.
- The use of these new methods has led to revised versions of Koch’s postulates: Fredricks and Relman[11] have suggested the following set of Koch’s postulates for the 21st century:
- A nucleic acid sequence belonging to a putative pathogen should be present in most cases of an infectious disease. Microbial nucleic acids should be found preferentially in those organs or gross anatomic sites known to be diseased, and not in those organs that lack pathology.
- Fewer, or no, copies of pathogen-associated nucleic acid sequences should occur in hosts or tissues without disease.
- With resolution of disease, the copy number of pathogen-associated nucleic acid sequences should decrease or become undetectable. With clinical relapse, the opposite should occur.
- When sequence detection predates disease, or sequence copy number correlates with severity of disease or pathology, the sequence-disease association is more likely to be a causal relationship.
- The nature of the microorganism inferred from the available sequence should be consistent with the known biological characteristics of that group of organisms.
- Tissue-sequence correlates should be sought at the cellular level: efforts should be made to demonstrate specific in situ hybridization of microbial sequence to areas of tissue pathology and to visible microorganisms or to areas where microorganisms are presumed to be located.
- These sequence-based forms of evidence for microbial causation should be reproducible.