Returning Research Results from Next-Generation Sequencing and Analysis to Patients with Idiopathic Disorders?

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

"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)

*Quote introduced to genetics community by 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 UNIVERSITY OF UTAH

UTAH foundation for BIOMEDICAL research

I moved to Utah July 2009 to find new rare diseases AND to study disease in large pedigrees, controlling for environment and population effects.

- **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 "Mendelian" disorders not described in the literature and NOT in OMIM, many of which have neuropsychiatric manifestations.**

Story #1

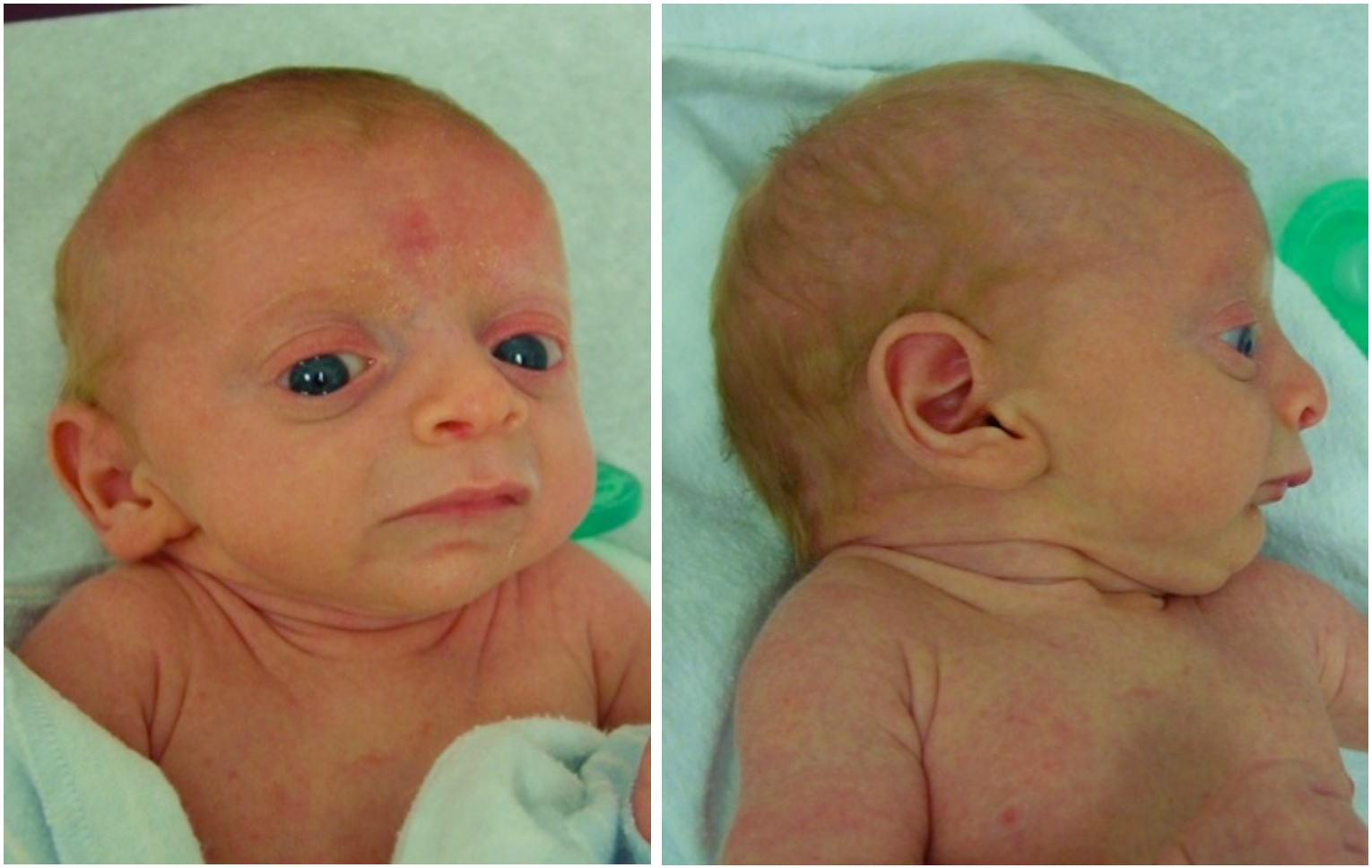
ARTICLE

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,*]

American Journal of Human Genetics **89**, 28-43 (2011).

Finding a new disease and figuring out its genetic basis



“little old man”, 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

I met the entire family on March 29, 2010, collecting all blood for DNA and cell lines.

Photo of mother, now a grandmother,
with son in late 1970's

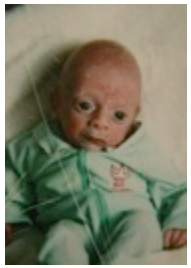
*Picture NOT for public distribution.

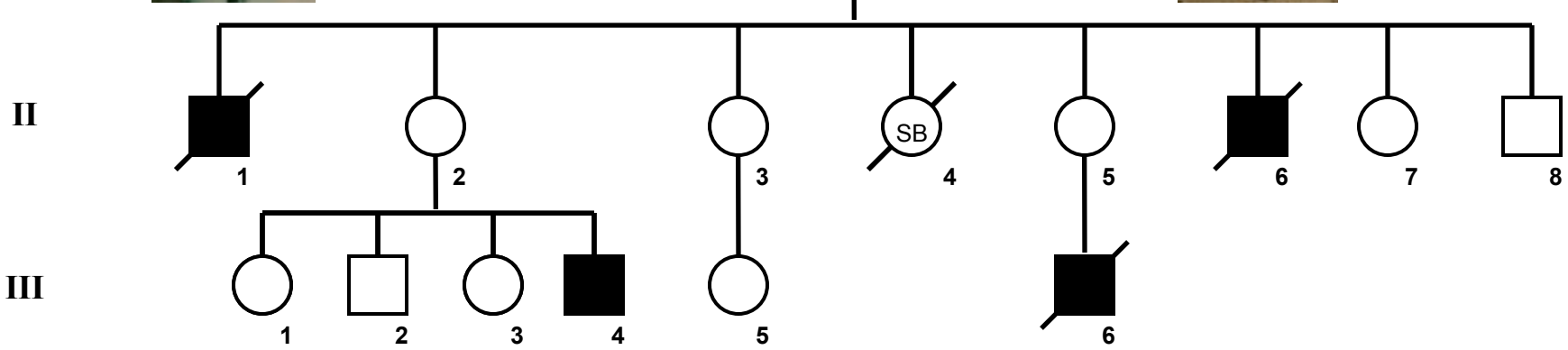
This is the first boy in the late 1970's.

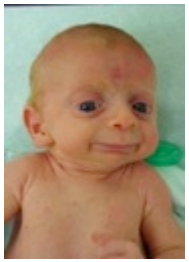
First boy. Called "a little old man" by the family. Died around ~1 year of age, from cardiac arrhythmias.

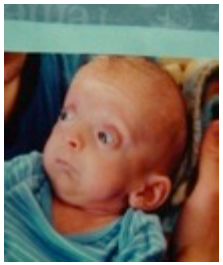
*Picture NOT for public distribution.

This is the family in Utah in March 2010.









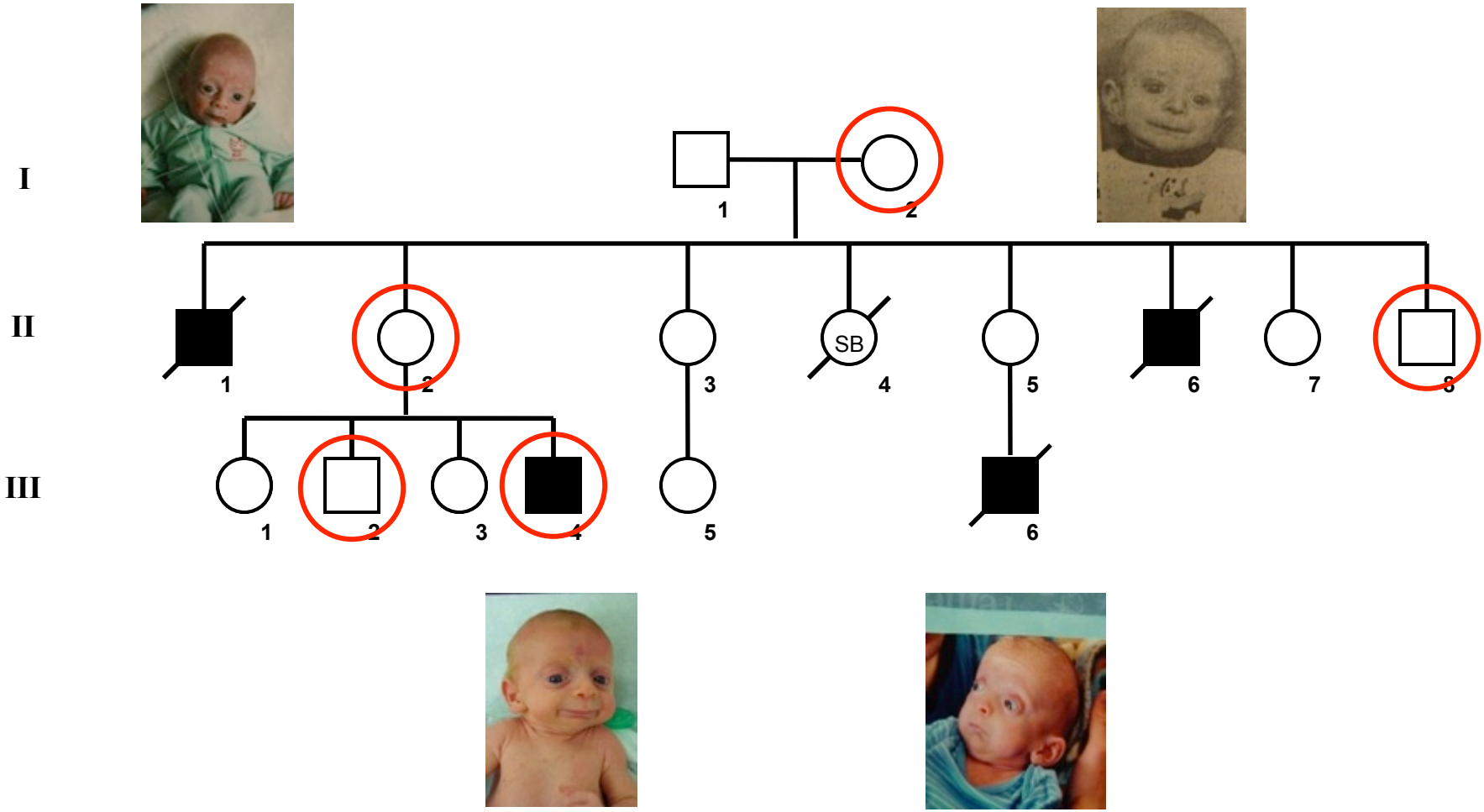
These are the Major Features of the Syndrome.

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.	

Proband died in early 2010, and I attended his autopsy.

Remarkably, for the most part, gross inspection at autopsy and histopathology were unremarkable for clues explaining the pathogenesis of this condition, other than cardiac failure.

Experimental Design for Sequencing is Critical.



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

- **Analysis with ANNOVAR and VAAST (Variant Annotation, Analysis and Search Tool).**

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

Rope, A.F., Wang, K., Evjenth, R., Xing, J., Johnston, J.J., Swensen, J.J., Johnson, W.E., Moore, B., Huff, C.D., Bird, L.M., et al. (2011). Using VAAST to Identify an X-Linked Disorder Resulting in Lethality in Male Infants Due to N-Terminal Acetyltransferase Deficiency. American Journal of Human Genetics.

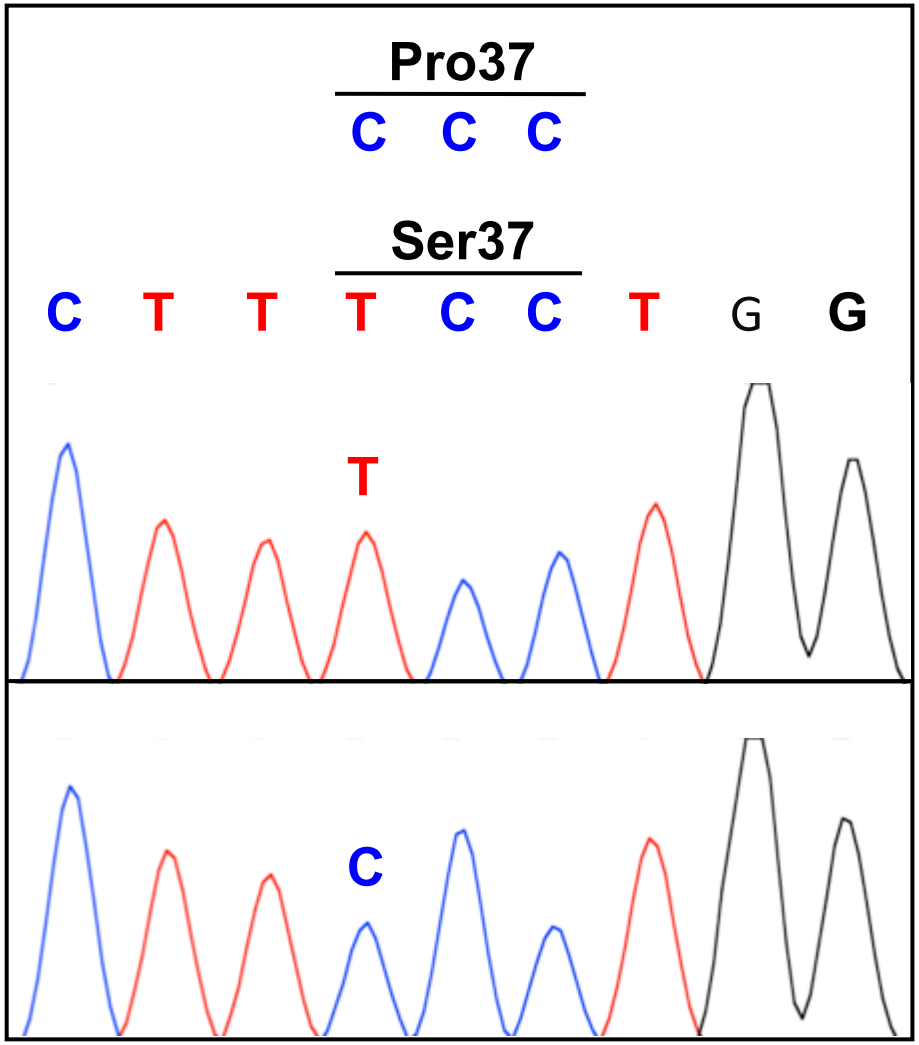
Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164.

Analysis with VAAST readily identified a few likely candidates.

Table 4. Summary of the Filtering Procedure and Candidate Genes with VAAST

SNV-Calling Pipeline	GATK	Samtools	GNUMAP
III-4 (total SNVs)	1546	1499	2168
III-4 (nsSNVs)	146	114	155
VAAST candidate genes (NAA10 ranking)	4 (3)	3 (2)	5 (2)
Present in III-4 and mother II-2 (nsSNVs)	122	107	116
VAAST candidate genes (NAA10 ranking)	3 (2)	2 (1)	2 (2)
Present in III-4, mother II-2, and grandmother I-2 (nsSNVs)	115	95	104
VAAST candidate genes (NAA10 ranking)	2 (1)	2 (1)	1 (1)
Present in III-4, II-2, and I-2, absent in brother III-2 and uncle II-8 (nsSNVs)	8	6	8
VAAST candidate genes (NAA10 ranking)	1 (1)	1 (1)	2 (1)

This is the causative mutation. There are ~6 billion nucleotide in a diploid human genome, but this one REALLY matters!



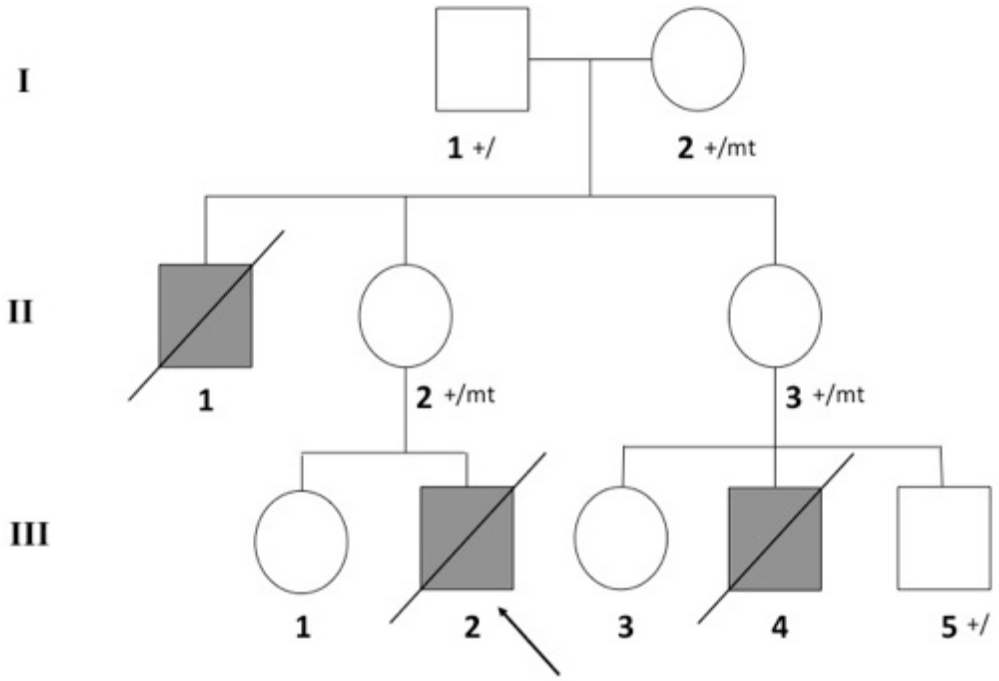
- **The discovery (or “Eureka moment”) occurred when blinded Sanger sequencing showed perfect segregation of one mutation with the disease.**

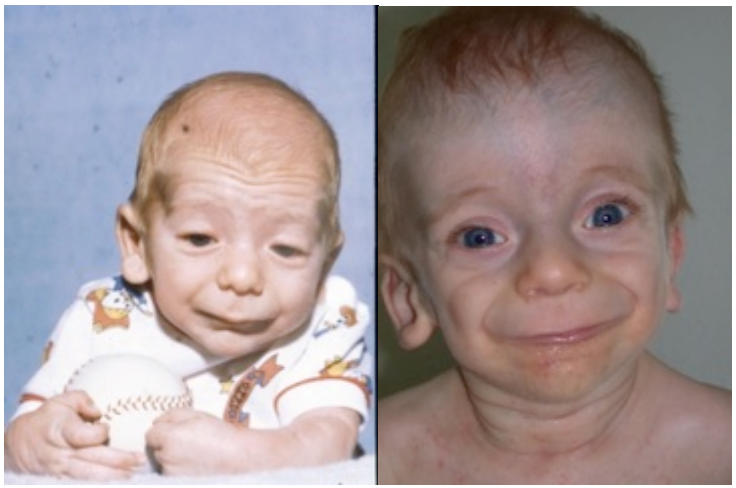
- **Mutation present in Proband, Carrier Mother, Carrier Grandmother and other carrier mothers.**

- **Absent in unaffected brother and unaffected uncle.**

- **Also present in DNA from formalin-fixed paraffin-embedded tissue from two other deceased boys, found in pathology department, saved in one case for 30 years.**

A Second Family was Identified with Same Mutation, Despite the Fact that the Two Families do NOT share a common founder.

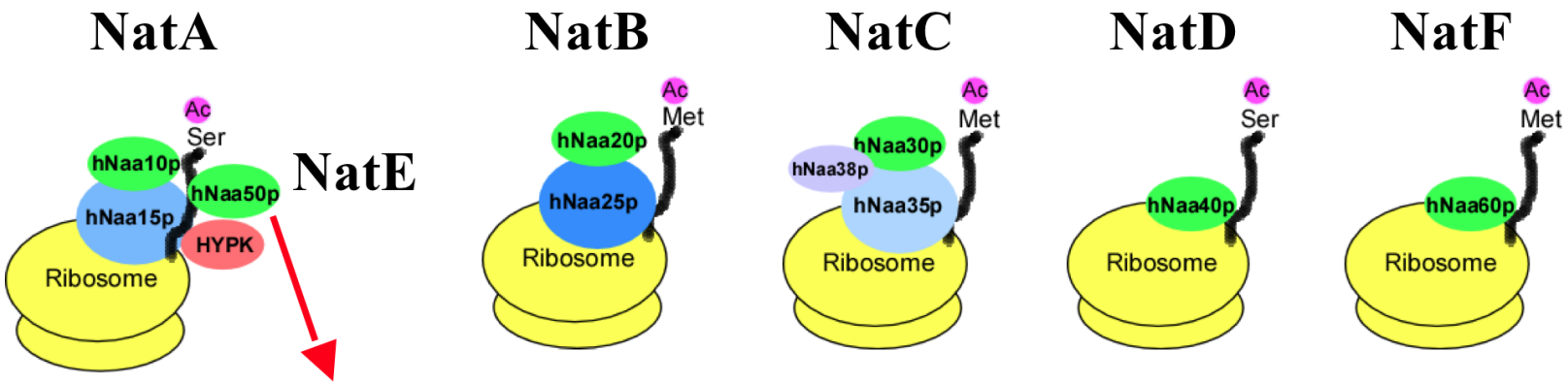




II-1 III-2

Contributed by Les Biesecker and colleagues at NIH

The N-terminal acetylation machinery in human cells.



Ser
Ala
Thr
Gly
Val
Asp/Glu*

Met-Leu

Met-Asp
Met-Glu
Met-Asn
Met-Gln

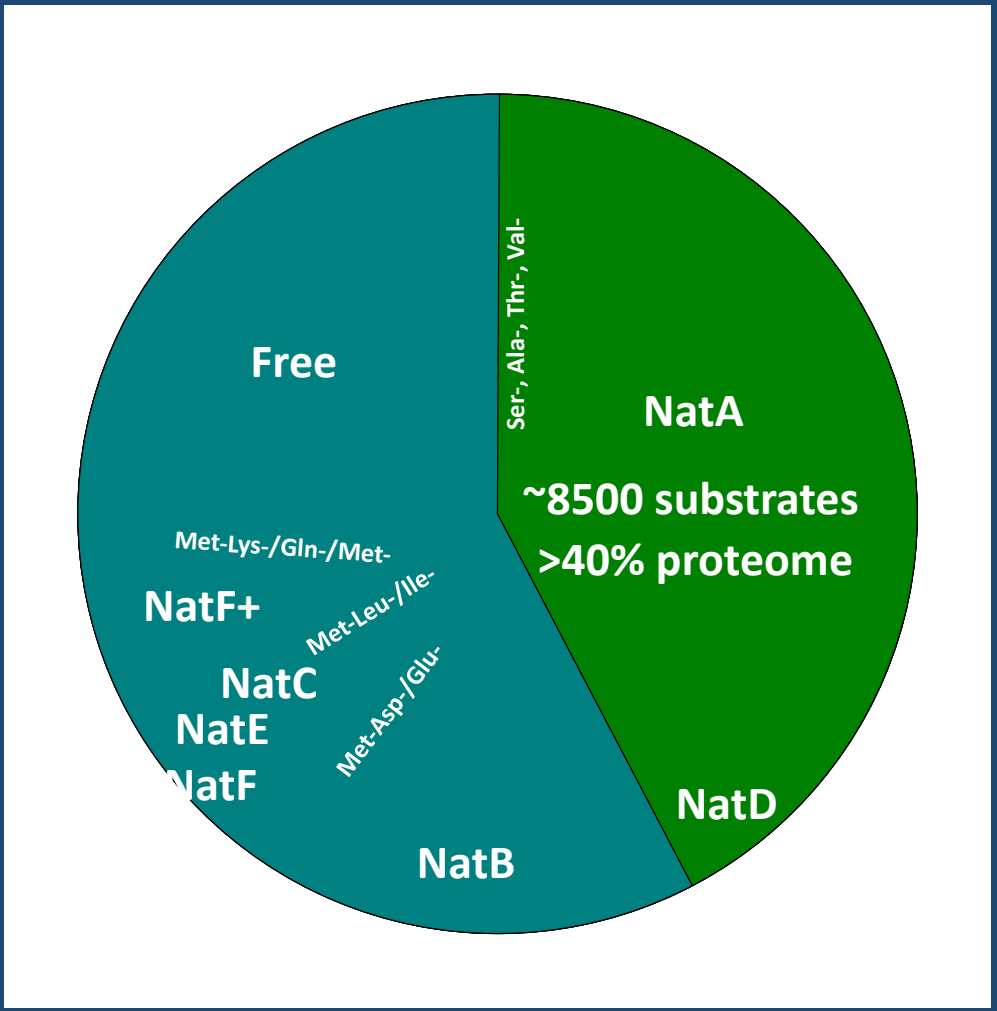
Met-Leu
Met-Ile
Met-Phe

Ser
(H2A/H4)

Met-Lys
Met-Ser
Met-Ala
Met-Met
Met-Leu

Slide courtesy of Thomas Arnesen

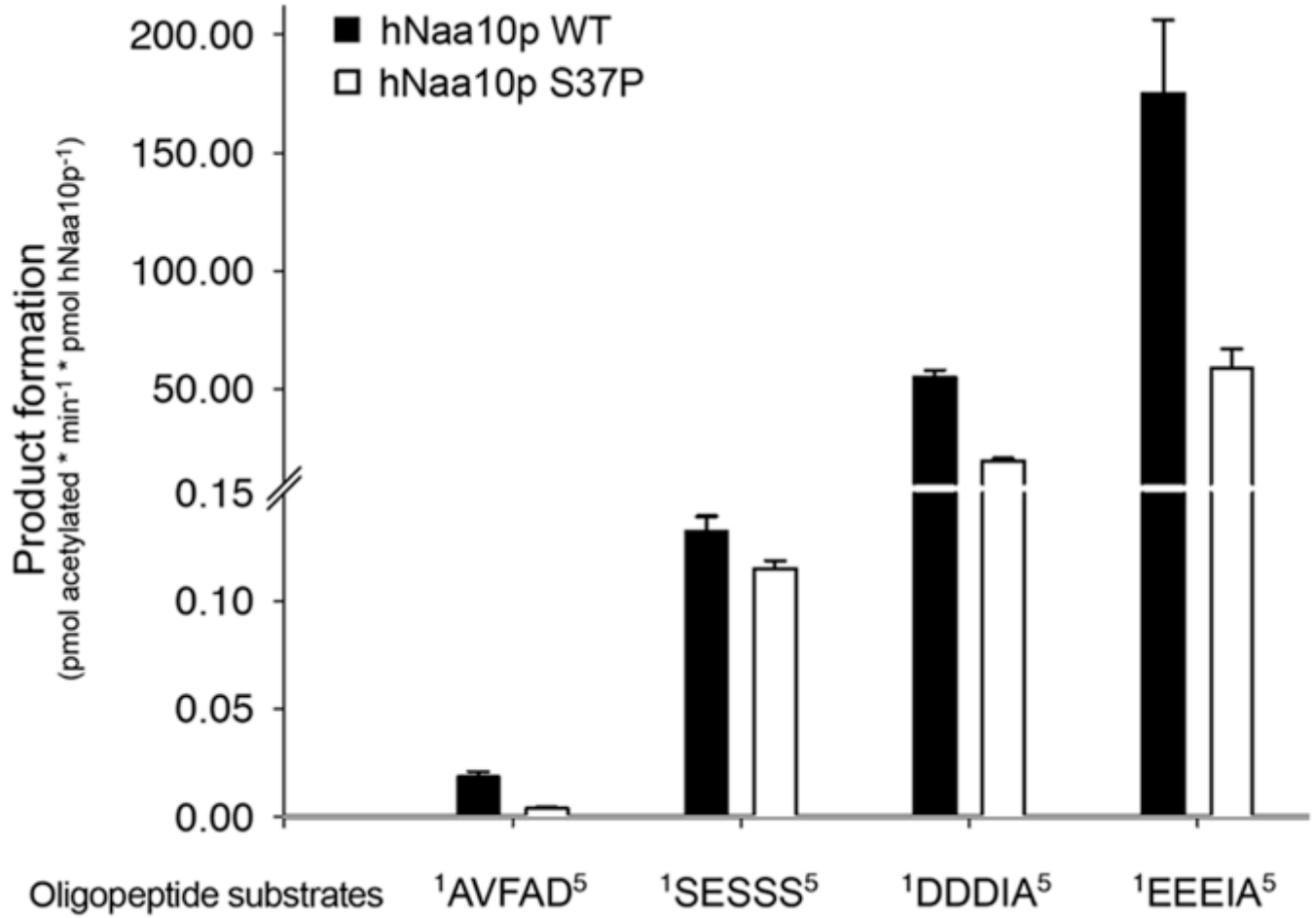
The human N-Acetylome



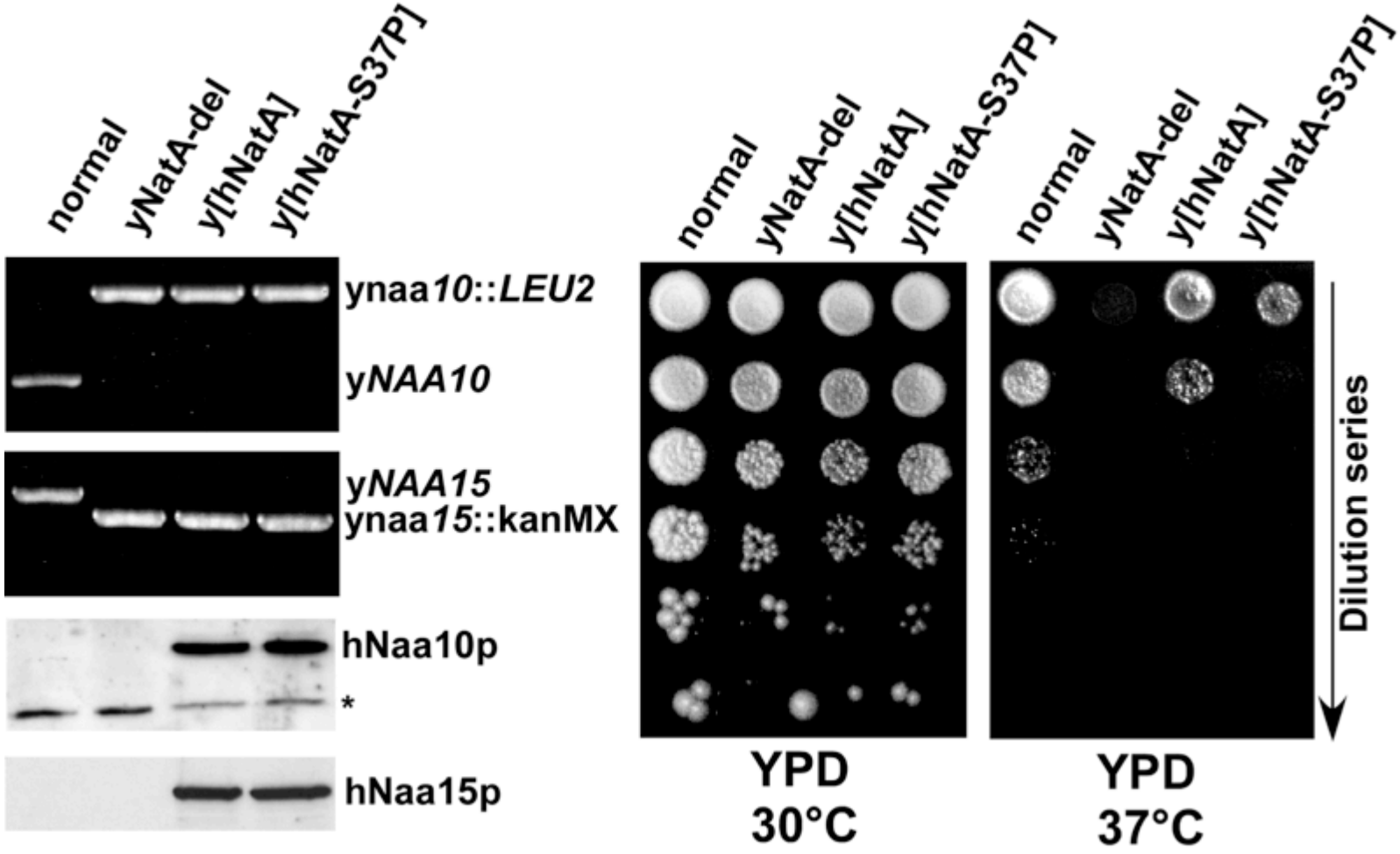
- A majority of soluble human proteins are N-terminally acetylated
- NatA is a major protein modifying enzyme of the human proteome

Arnesen T *et al.*, *Proc Natl Acad Sci USA*, 2009; Van Damme P *et al.*, *PLoS Genet*, 2011

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.

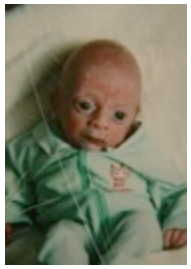


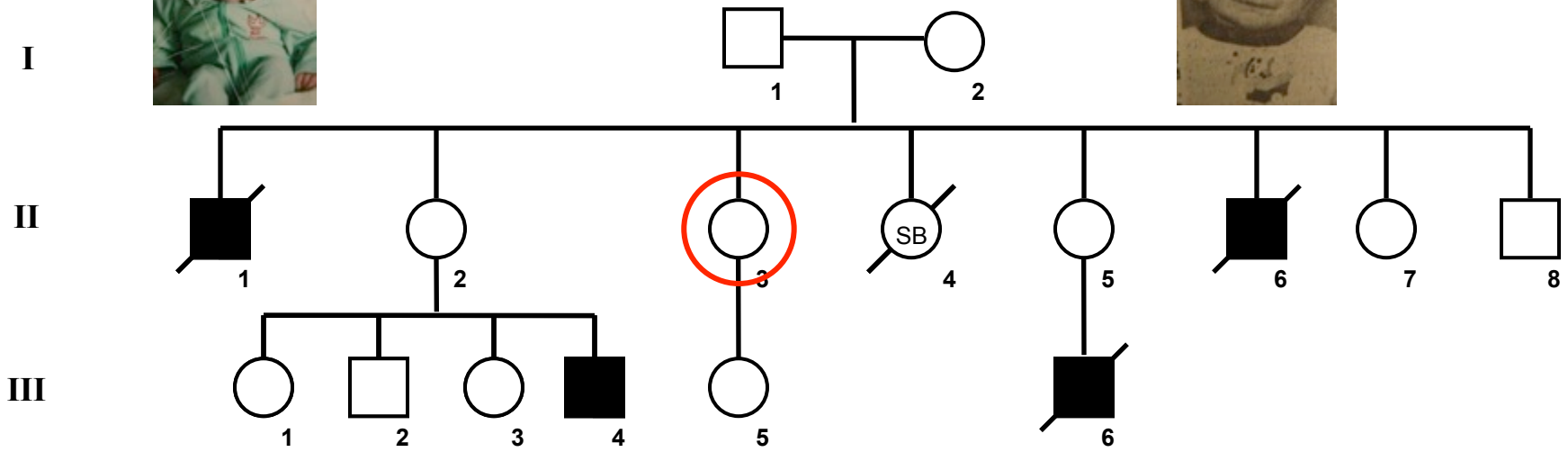
Unpublished, figure courtesy of Thomas Arnesen

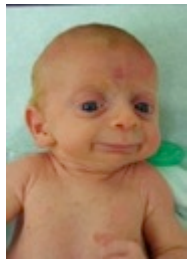
By November 2010, we had good functional data *in vitro* (bacterially expressed proteins) and *in vivo* (yeast, unpublished), leading me to believe we had identified the causative mutation.

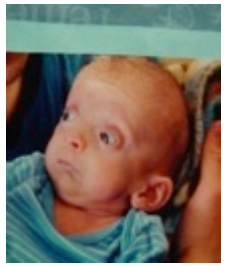
A new mother in the family informs me she is 4 months pregnant, with a boy!

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







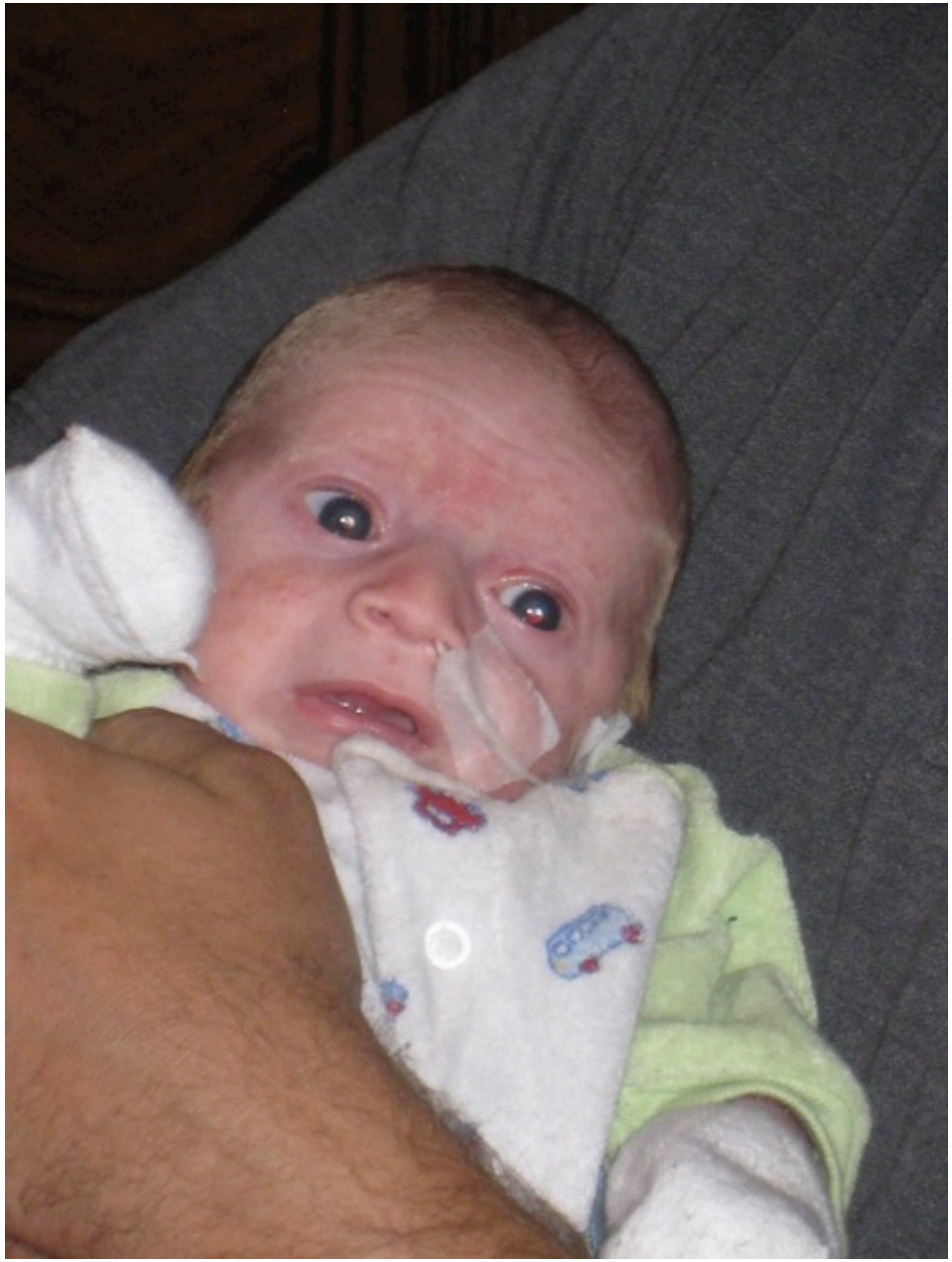


BUT, how do we give such research results back to patients?

Hippocratic Oath: FIRST DO NO HARM.

MAJOR ISSUES

- I am a physician but not HER physician, therefore I had NOT entered into a “physician-patient contract” with her.
- This was not a “diagnostic test”. This was research.
- The IRB protocol dictated that research results cannot be returned to these research subjects.
- There are, of course, MANY reasons for why all of this IS the way that it IS…. Discussed at length by prior speakers at this meeting.
- So, I did NOT return the results to the mother to be.



Mother four months Pregnant Nov 2010

Baby born March 2011.

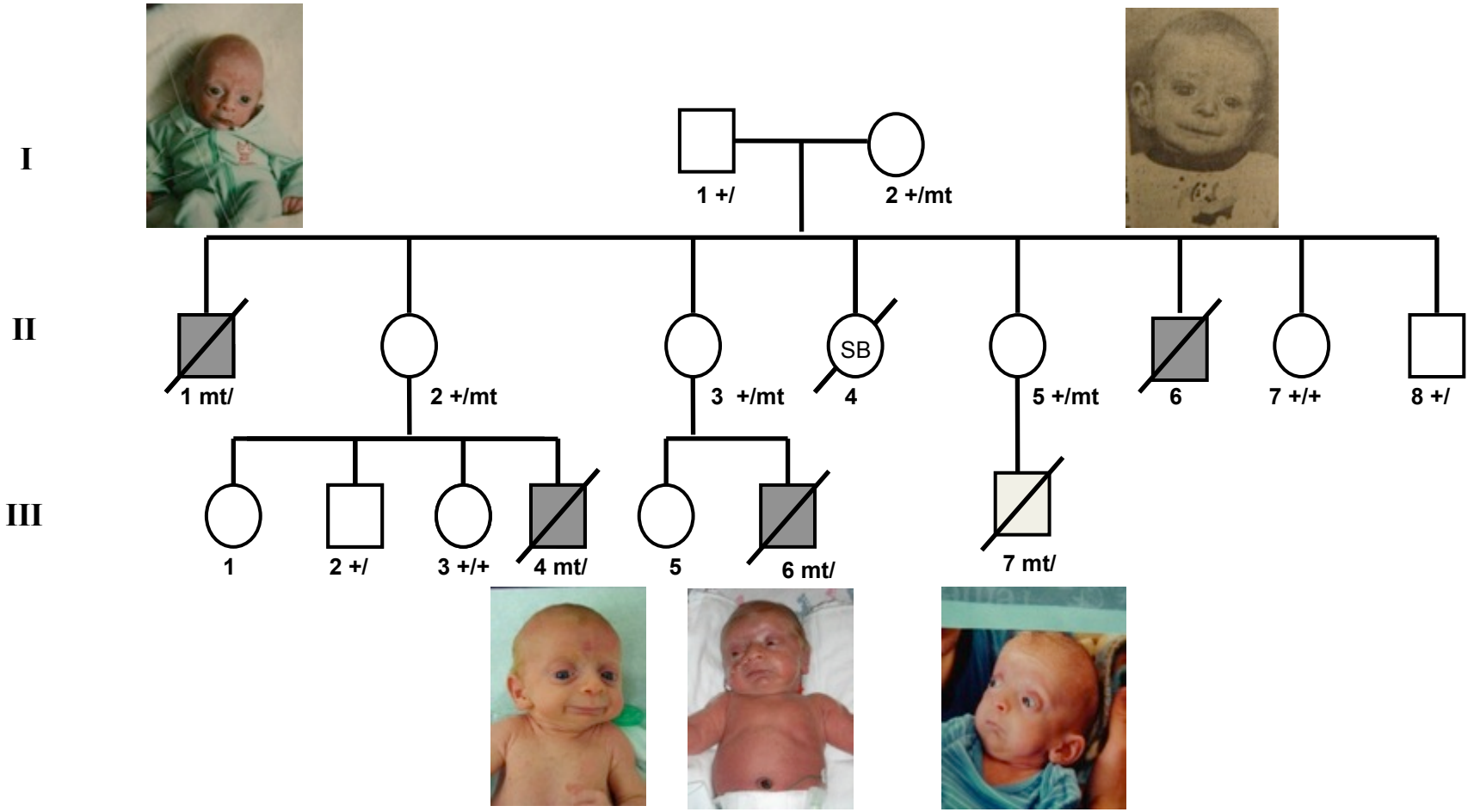
Affected with Disease.

He died June 2011, same week as publication of our paper in AJHG.

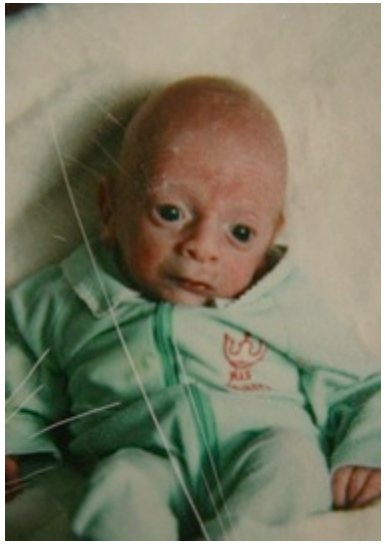
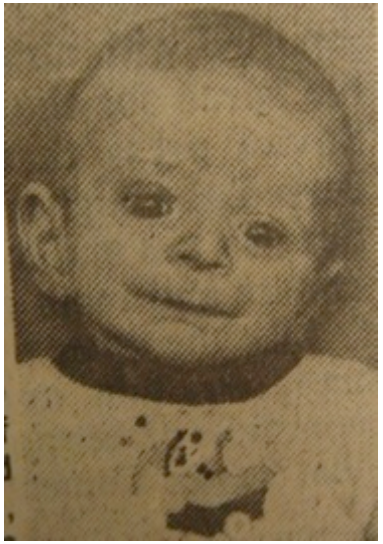
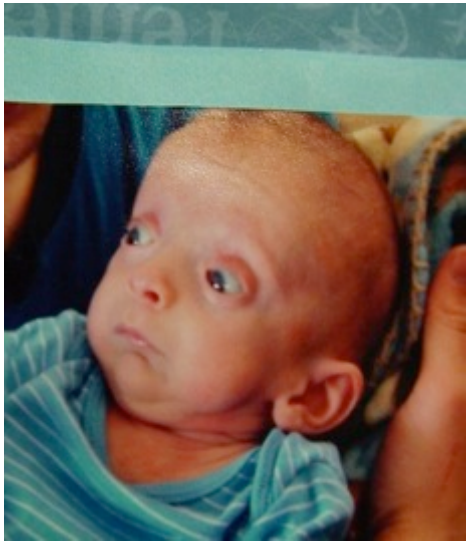
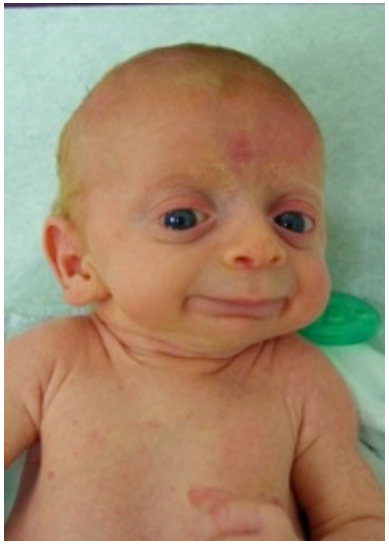
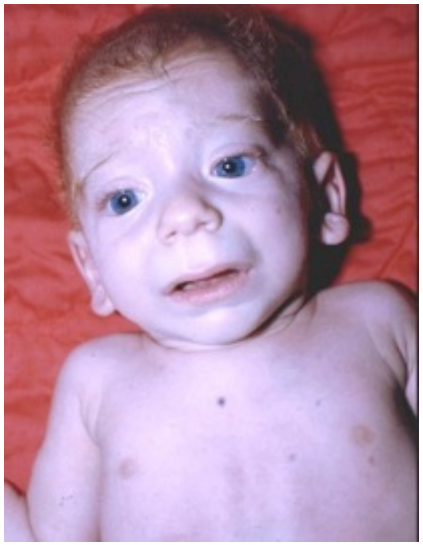
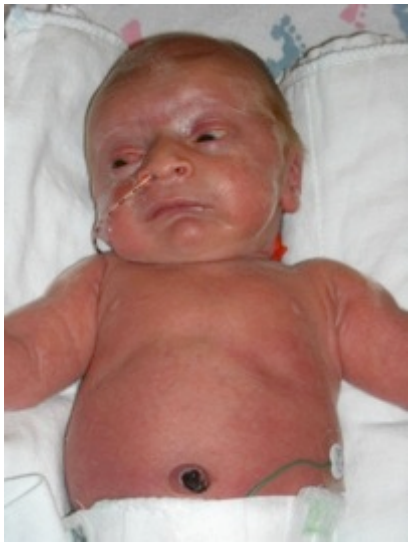
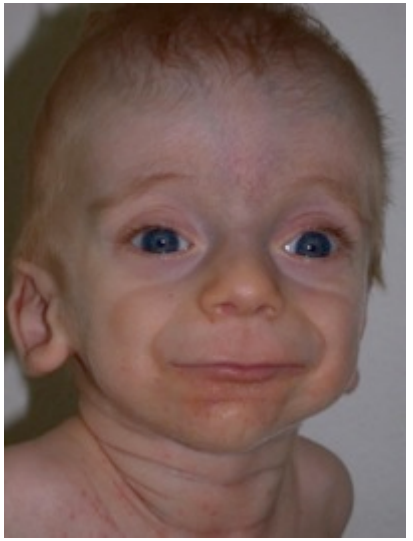
We did develop a CLIA-certified test for the future.

*Picture NOT for public distribution.

Family now in September 2011



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

Story #2 – Psychiatric Genetics is MUCH harder!

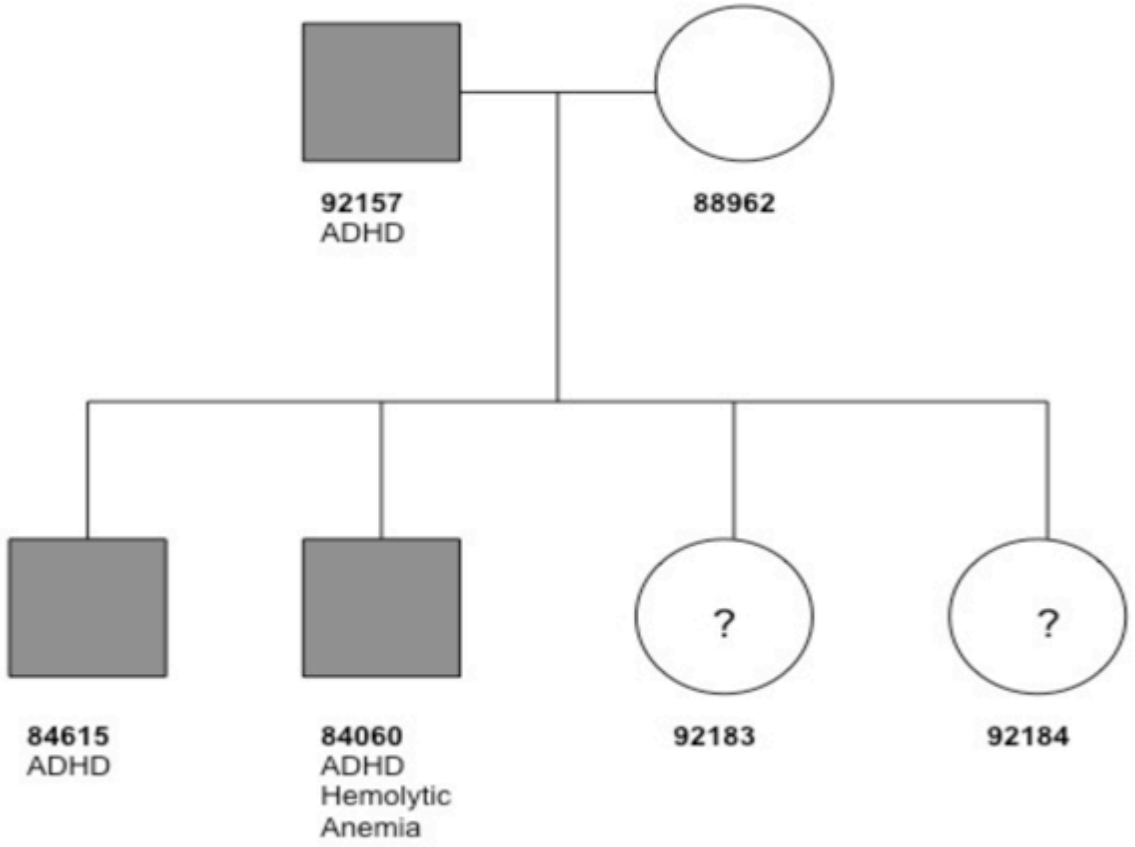
Exome Sequencing and Unrelated Findings in the Context of Complex Disease Research: Ethical and Clinical Implications

Gholson J. Lyon, Tao Jiang, Richard Van Wijk, Wei Wang, Paul Mark Bodily, Jinchuan Xing, Lifeng Tian, Reid J. Robison, Mark Clement, Lin Yang, Peng Zhang, Ying Liu, Barry Moore, Joseph T. Glessner, Josephine Elia, Fred Reimherr, Wouter W. van Solinge, Mark Yandell, Hakon Hakonarson, Jun Wang, William Evan Johnson, Zhi Wei, and Kai Wang

Discovery Medicine®
www.discoverymedicine.com
ISSN: 1539-6509; eISSN: 1944-7930

I am very supportive of open-access publishing!

This is the ADHD pedigree in January 2010, admittedly small but very well phenotyped for four members.



Phenotyping is Critically Important in Neuropsychiatric Disorders!

Supplementary Table 1. ADHD measures during a clinical trial of methylphenidate transdermal system.

		92157	84060	84615
Baseline				
	WRAADDS	16	22	16
	ODD	1	11	7
	CAARS	40	55	38
	CGI-S	4	4	4
Active Medication				
	WRAADDS	0	4	3
	ODD	0	1	3
	CAARS	10	0	13
	CGI-I	1	1	1
	CGI-S	1	3	2
Placebo				
	WRAADDS	15	24	20
	ODD	6	8	7
	CAARS	33	51	42
	CGI-I	4	4	N/A
	CGI-S	4	5	N/A

WRAADDS: Total score on the Wender Reimherr Adult ADD Scale
ODD: Oppositional Defiant Disorder scaore on the WRAADDS ODD subscale
CAARS: Total score Connor's Adult ADHD Rating Scale
CGI-S: Clinical Global Impression, Severity score.

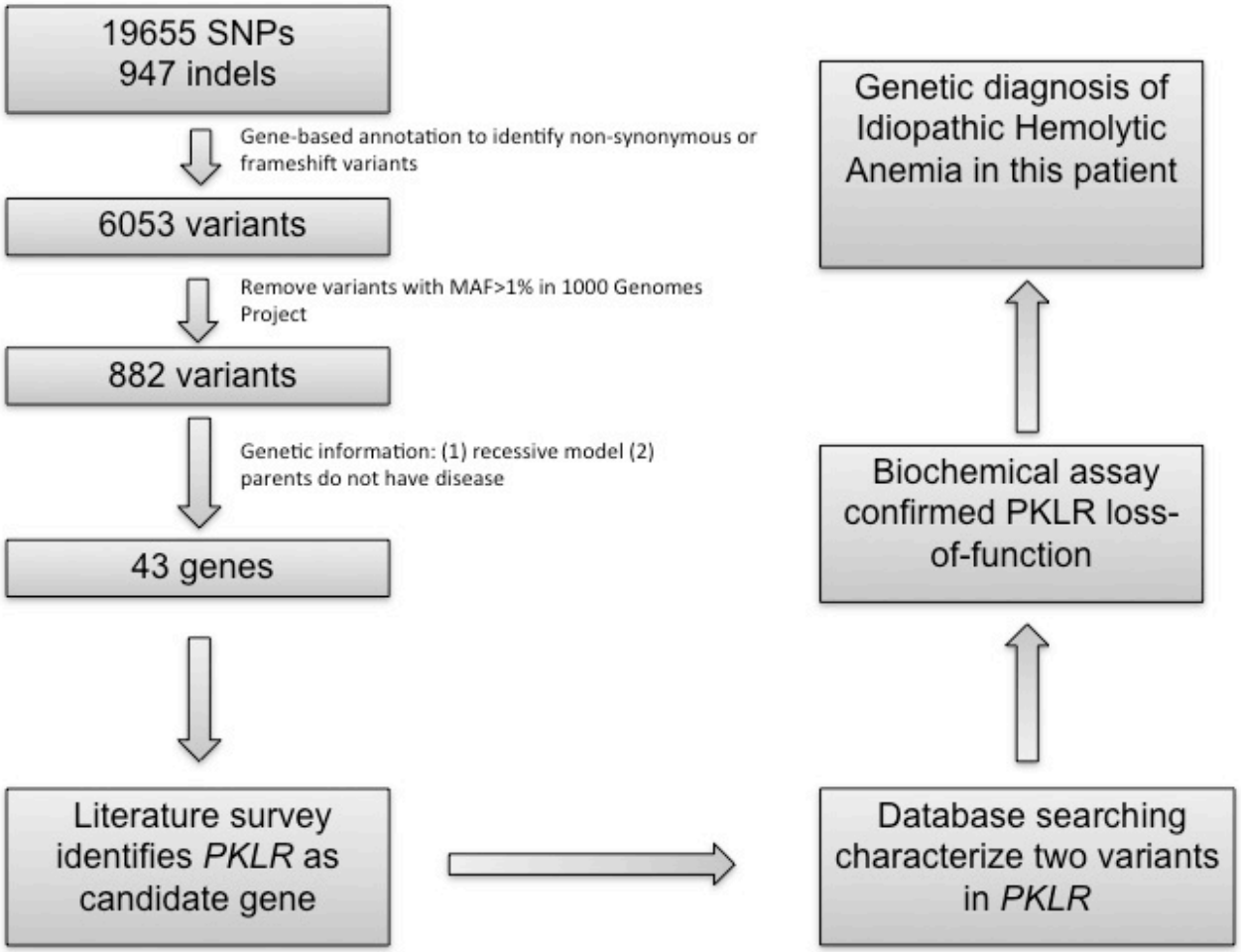
Exome Sequencing performed early 2010

While analyzing the exome data, research subject (age ~24) informs me that he recently had his spleen removed!

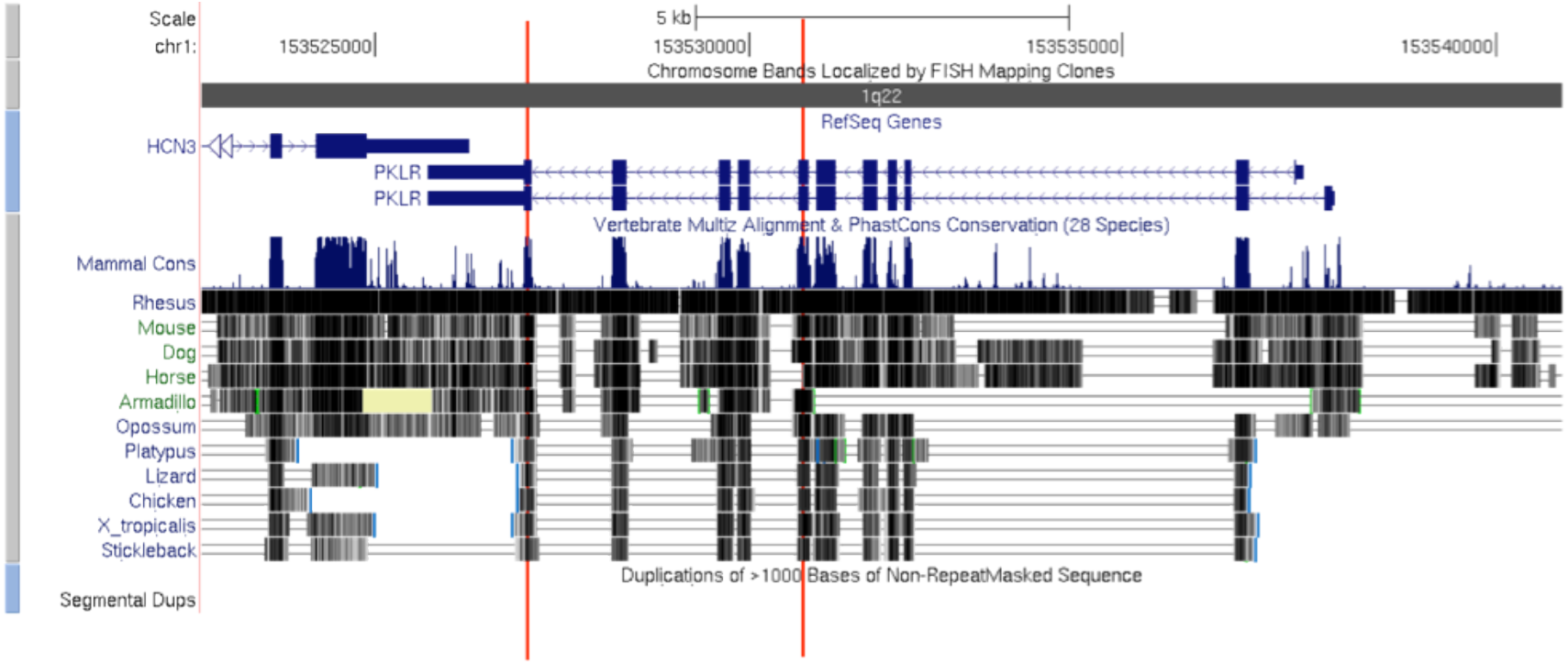
He has idiopathic hemolytic anemia, since childhood….

Although I am not his physician, I still feel an ethical and moral obligation to try to figure out what is going on….

This was our filtering pipeline.



Compound Heterozygote in *PKLR*, with each mutation inherited from one parent.



Some Additional Data to support the causation of these variants for idiopathic hemolytic anemia.

Table 2: Biochemical assays of enzyme activities in the patient affected with idiopathic hemolytic anemia confirmed *PKLR* deficiency. PK, pyruvate kinase; HK, hexokinase; G6PD, glucose-6-phosphate dehydrogenase.

	Patient 84060	Control	Reference values
PK (U/gHb)	3.3 L	8.6	6.1 – 12.3
HK (U/gHb)	3.2 H	1.1	0.8 – 1.5
G6PD (U/gHb)	15.8 H	9.2	6.4 – 10.5

Table 3: Bioinformatics prediction on the functional impact of two PKLR mutations. A mutation is regarded as deleterious if the SIFT<0.05, or PolyPhen>0.85, or PhyloP>0.95, or MutationTaster/LRT prediction as "D" (deleterious).

Mutation	SIFT	PolyPhen 2	PhyloP	LRT	MutationTaster
R569Q	0.03	0.84	0.97	D	D
G341A	0	0.889	1	D	D

Structural Modeling is also consistent with deleterious effects of these mutations.

- **In our case, we communicated the research results back to the hematologist, asking the doctor to follow up with appropriate genetic tests and counseling.**

- **There is NOTHING "Incidental" about Unrelated Findings.**

- **Sequencing a bunch of exomes and finding random rare variants MIGHT be "incidental", but actually proving that these variants CAUSE the disease is NOT simple or "incidental" or "accidental".**

- **I would suggest calling these "unrelated findings", rather than "incidental".**

A Word on the Identification of Rare Variants

- We found many very rare variants segregating with ADHD in father and two sons,
- BUT we cannot prove with certainty that any of these variants cause the disease, either by themselves or in aggregate or with epistatic effects.

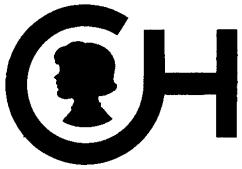
- Biology experiments must be performed to further bolster causality.

- **Proving CAUSALITY for variants is critically important, as it is not enough just to present anecdotes or associations.**

Conclusions

- We have identified a new, previously idiopathic, human disorder likely resulting from a defect in amino-terminal acetylation of proteins, opening the door to new biology with this major protein modification.
- We have used next generation sequencing to figure out the genetic basis of a case of idiopathic hemolytic anemia, and grappled with how to give back “unrelated findings”.
- The time is now to find and study idiopathic diseases in humans, using next generation sequencing. There are likely MANY previously undescribed diseases.
- I would suggest that researchers perform CLIA-certified sequencing UP FRONT, either with exomes or whole genomes, so that we can return results to research subjects, when and if causality is proven beyond reasonable doubt.

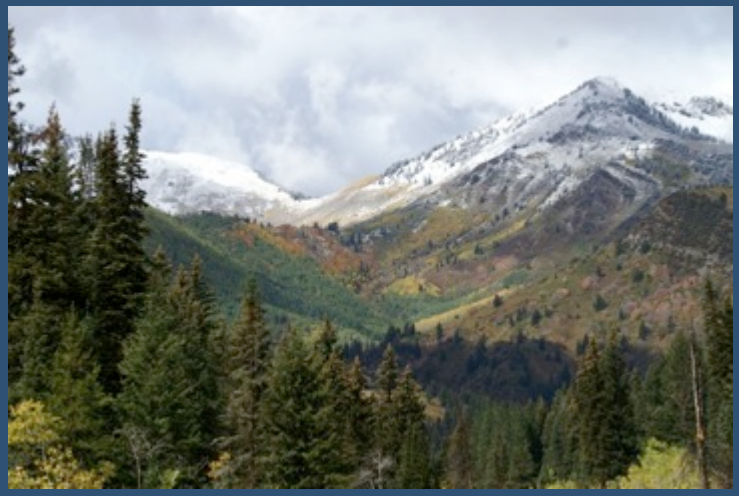
Acknowledgements



The VAAST DEVELOPMENT GROUP

www.yandell-lab.org

- Mark Yandell
- Martin Reese[!]
- Hao Hu[+]
- Barry Moore[+]
- Steve Chervitz[+!]
- Chad Huff[x]
- Jinchuan Xing[x+]
- Marc Singleton[+]
- Edward Kiruluta[!]
- Archie Russell[!]
- Fidel Salas[!]
- Ginger Guozhen Fan[+]



[+]Yandell lab, [!]Omicia, [x]Jorde Lab

Omicia

The University of Utah

Acknowledgments

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
Reid Robison
Christa Schank
Sarah T. South
Jeffrey J Swensen
Jinchuan Xing
Mark Yandell

Reid Robison
Edwin Nyambi
Matt Adams
Josh Hansen
Leone Lobendahn

Gholson J. Lyon
Kai Wang
Zhi Wei
Lifeng Tian

our study families

Thomas Arnesen
Rune Evjenth
Johan R. Lillehaug

Leslie G. Biesecker
Jenifer J Johnston
Lynne M. Bird
Cathy A. Stevens

Tao Jiang
Jun Wang

THE END

The Exon Capture and Coverage was high depth with Average Base Coverage of 214x, but…..

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).							

Analysis with VAAST readily identified a few likely candidates.

Table 3. Summary of the filtering procedure and candidate genes using VAAST			
SNV calling pipeline	GATK	Samtools	GNUMAP
III-4 (total SNVs)	1546	1499	2168
III-4 (nsSNVs)	146	114	155
VAAST candidate genes (NAA10 ranking)	4 (3)	3 (2)	5 (2)
Present in III-4 and mother II-2 (nsSNVs)	122	107	116
VAAST candidate genes (NAA10 ranking)	3 (2)	2 (1)	2 (2)
Present in III-4, mother II-2, and grandmother I-2 (nsSNVs)	115	95	104
VAAST candidate genes (NAA10 ranking)	2 (1)	2 (1)	1 (1)
Present in III-4, II-2, and I-2, absent in brother III-2 and uncle II-8 (nsSNVs)	8	6	8
VAAST candidate genes (NAA10 ranking)	1 (1)	1 (1)	2 (1)

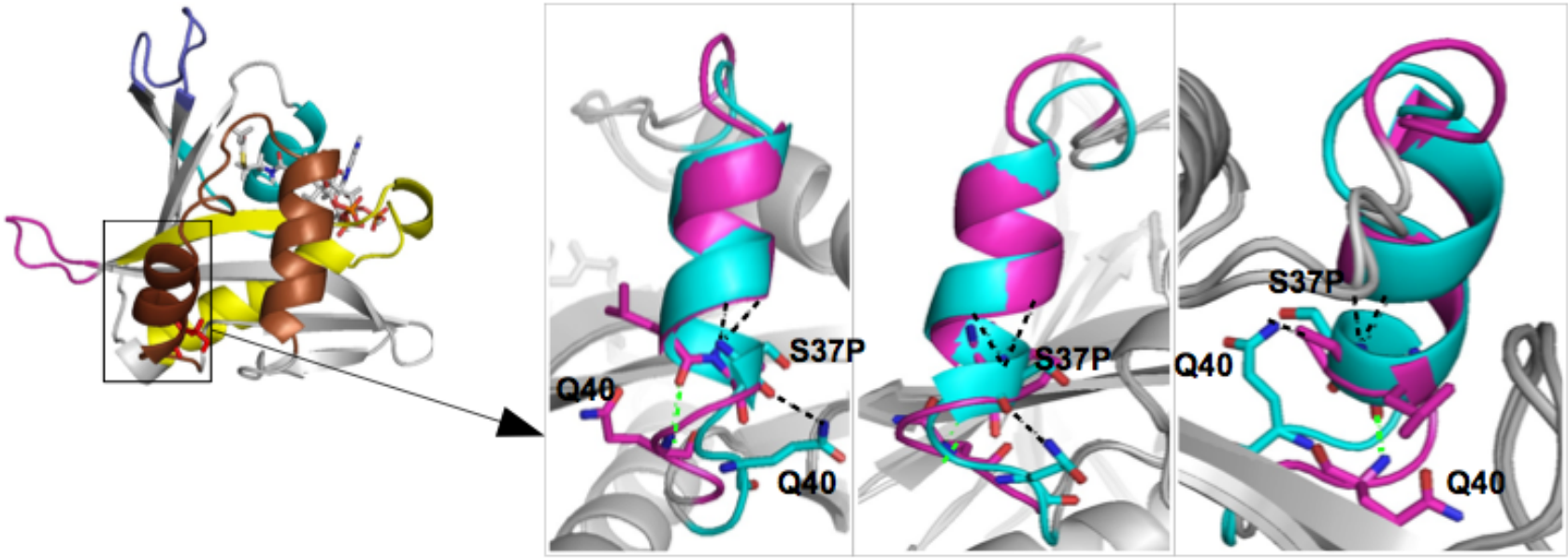
We analyzed the data with 3 pipelines.

Table 3: The SNV count, non-synonymous coding SNV count and Ti/Tv ratio for each individual for each variant analysis pipeline in Family 1.

Sample	Pipeline	SNV Count	Non-synonymous	Ti/Tv
III-4	samtools	1499	114	2.0
	GATK	1546+236[a]	146+6 (nonsyn+frame)	2.0
	GNUMAP	2168	155	2.0
II-2	samtools	2512	219	1.6
	GATK	1999+270[a]	168+8	2.1
	GNUMAP	2893	183	2.0
III-2	samtools	1491	106	2.0
	GATK	1509+252[a]	134+10	2.0
	GNUMAP	2062	131	2.0
I-2	samtools	2637	229	1.5
	GATK	2032+278[a]	160+10	2.0
	GNUMAP	2920	183	1.9
II-8	samtools	1513	108	1.9
	GATK	1572+243[a]	136+8	1.9
	GNUMAP	1924	139	2.0

[a]microindels ascertained with GATK pipeline

Structural modelling of hNaa10p wt (cyan) and S37P (pink)

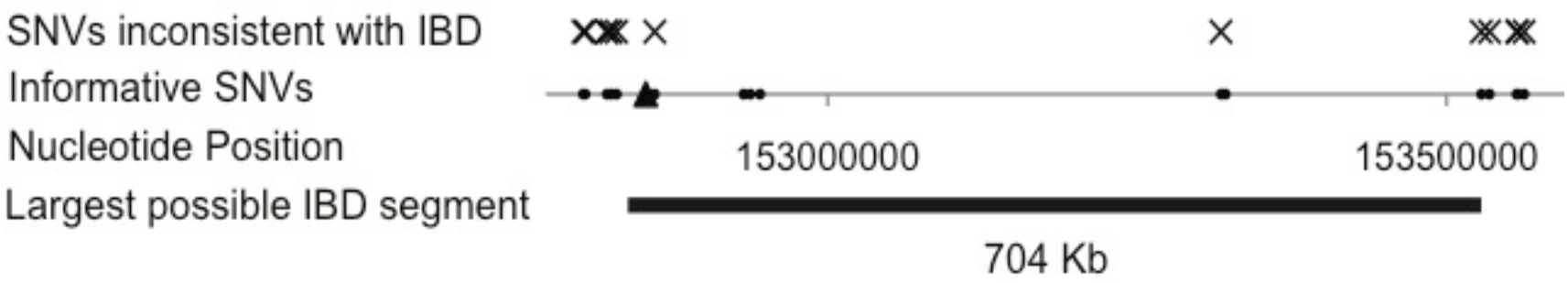


The mutation causes Naa10p S37P.

- Ser 37 is very conserved

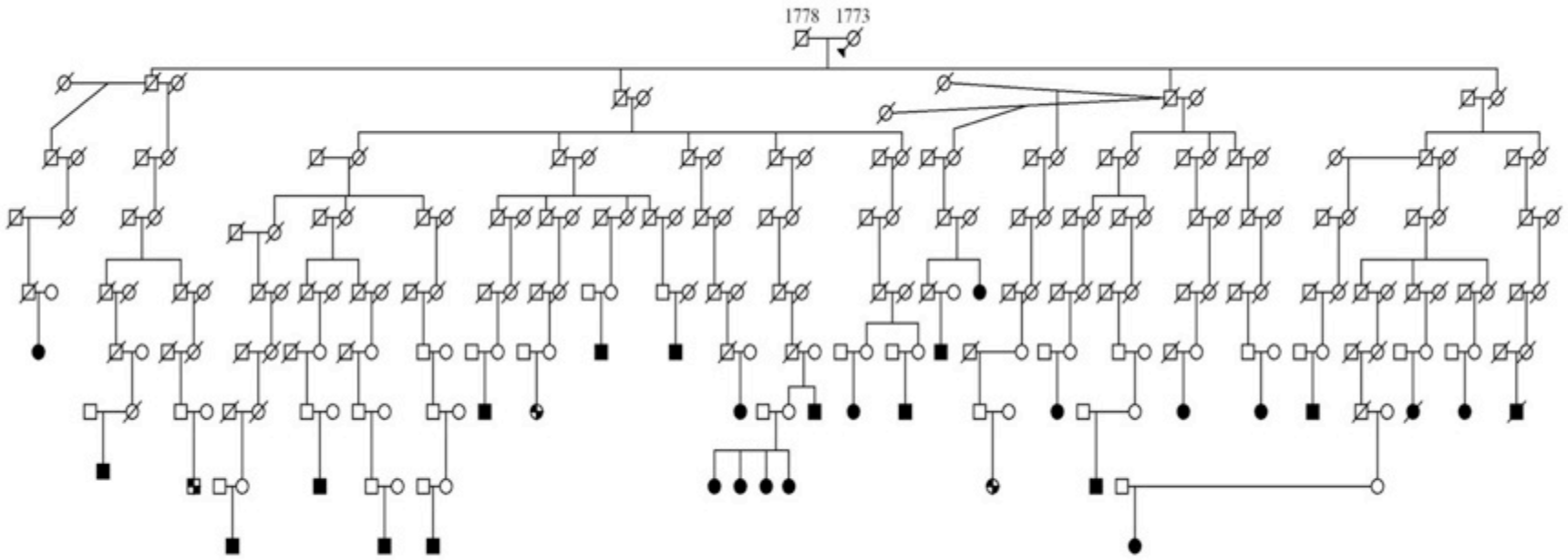
- Ser37Pro is predicted to affect functionality (SIFT-analysis)

These two families are UNRELATED, i.e. no common founder.



Courtesy of Chad Huff and Lynn Jorde

This is a typical Utah disease pedigree



UTAH Foundation for BIOMEDICAL research

Stephen Guthery, MD

All Clinical Diagnostic Tests are regulated in America:

Clinical Laboratory Improvement Amendments (CLIA)-certified lab

And sometimes:

College of American Pathology (CAP) certification

Technically, CLIA-certification for DNA tests means the following:

1) **Blood drawn in a CLIA-certified facility by a phlebotomist covered by that facility.**

2) **DNA isolated in a CLIA-certified facility.**

3) **Sequencing performed on CLIA-certified machines with CLIA-certified reagents in a CLIA-certified facility.**

4) **Analysis performed with a CLIA-certified bioinformatics pipeline.**

5) **DNA results returned to patients by a licensed genetic counselor, signed off on by a board-certified medical geneticist.**

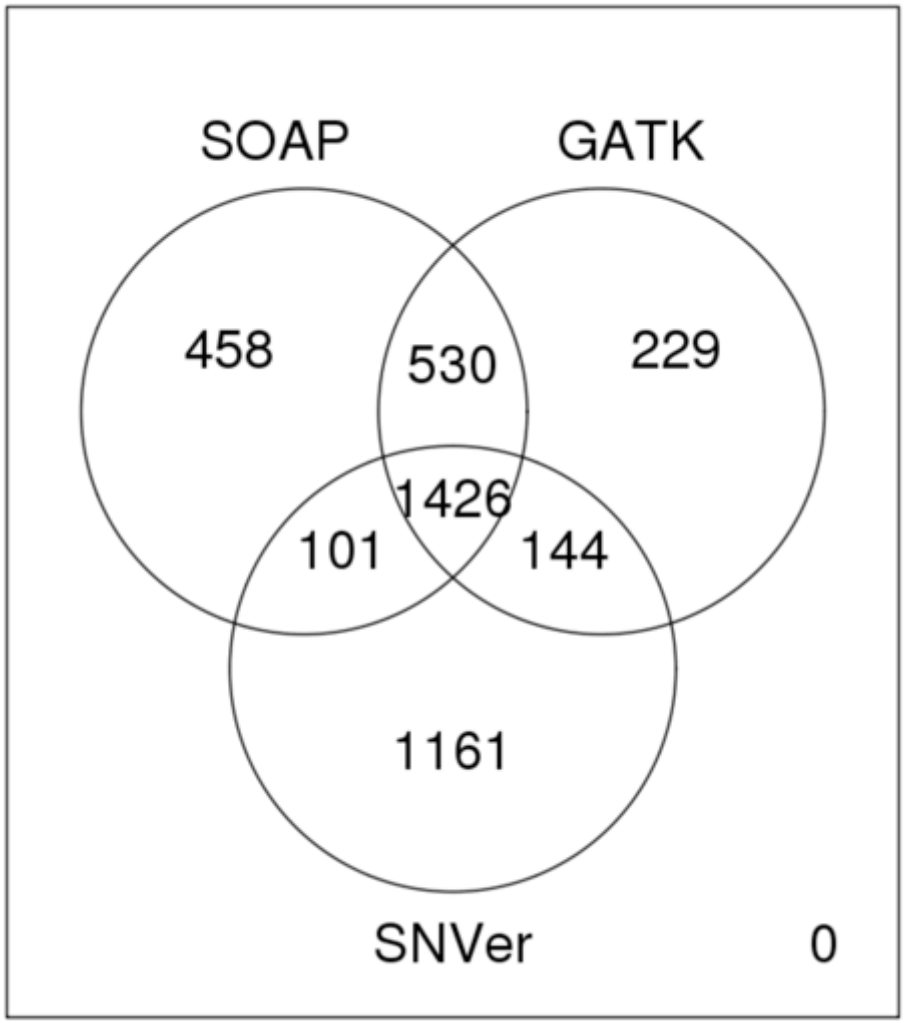
- Any clinical test at a major diagnostic lab first has to go through a validation process by the R&D group (known samples are run through the test in a specific way, then the results are documented in a formal validation packet that is approved by management). All results are reviewed by medical directors.

- After the validation, accounting has to decide how the test will be billed, genetic counselors have to write the information that will appear on the clinical reports and the test has to be formally transferred into the clinical lab (which involves putting paperwork in place, getting reagents in stock and training employees).

- The test usually also has to be integrated into the computer system and test directory, although some of this can be bypassed if one requires that the test be specially ordered.

- THIS TOOK ABOUT SIX MONTHS!!! and one diagnostic lab only agreed to do this once a second family with this condition was found, given economic considerations.

Intersection of variants. We show here the variants identified by the three main pipelines as being present in the three males with ADHD, but not present in the unaffected mother.



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

Bioinformatics Analysis for ADHD pedigree

Table 1. Summary of SNVs for exome capture samples				
ExomeCapture	84060 (child 1)	84615 (child 2)	92157 (father)	88962 (mother)
Sequencing platform	GA IIX	GAIIX	GAIIX	HiSeq 2000
Reads property	76bp PE	76bp PE	76bp PE	90bp PE
Number of SNVs (Method 1: SOAP)	19825	19270	20430	22294
Ti/Tv ratio	2.8	2.7	2.9	2.8
Number of SNVs+indels (Method 2: BWA+GATK)	19655+947	18892+955	20100+916	21572+513
Ti/Tv ratio	2.9	2.9	3.0	2.9
Number of SNVs (Method 3: Shrimp2+SNVer)	16063	16704	18253	23917
Ti/Tv ratio	2.7	2.6	2.7	2.4
*We have not yet analyzed the mother's exome with the 4th method (GNUMAP), so we have omitted this method from the table.				