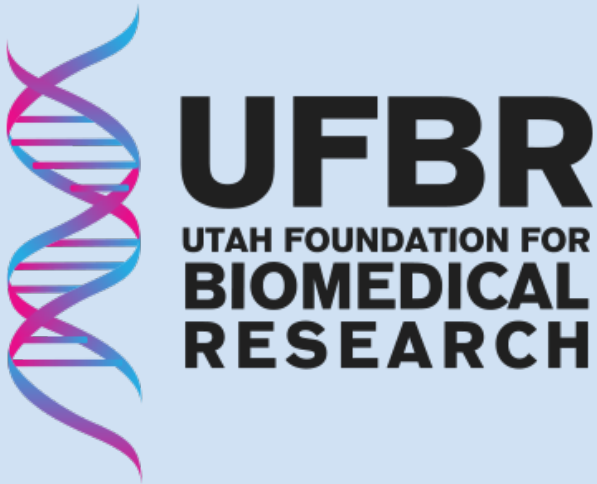
Human Genetics and Orphan Diseases

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



Jason O’Rawe

Yiyang Wu

Han Fang

Laura Jimenez Barron

Max Doerfel

UFBR
UTAH FOUNDATION FOR BIOMEDICAL RESEARCH

Reid Robison

THE UNIVERSITY OF UTAH

Barry Moore
Alan Rope
Jeffrey J Swensen
Lynn Jorde
Mark Yandell

USC TROJANS

Kai Wang

CSH STANLEY INSTITUTE FOR COGNITIVE GENOMICS
COLD SPRING HARBOR LABORATORY

Lyon Lab
Jason O'Rawe
Yiyang Wu
Han Fang
Max Doerfel
Laura Jimenez Barron
Jillian Ho
Constantine Hartofilis
Noah Davis

Past members:
Jake Weiser
Syndi Barish
Prashant Kota
Michael Klingener

Mt. Sinai: Sunita D'Souza

our study families and many others

Other CSHL labs

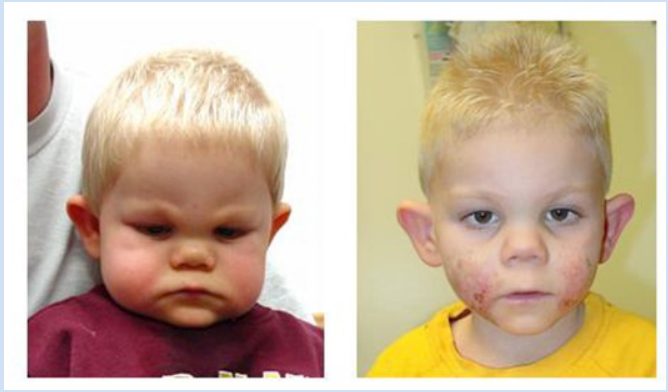
Michael Schatz
Giuseppe Narzisi
Darryl Pappin
Keith Rivera

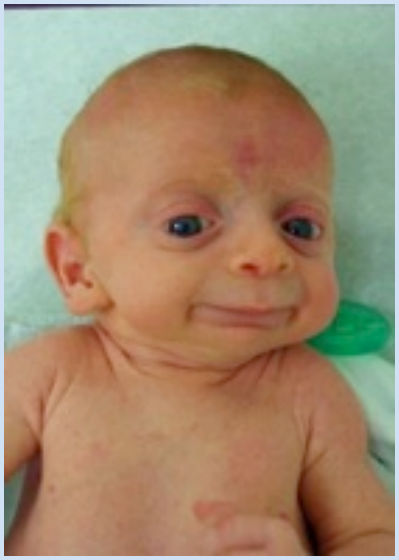
UNIVERSITAS BERGENSIS

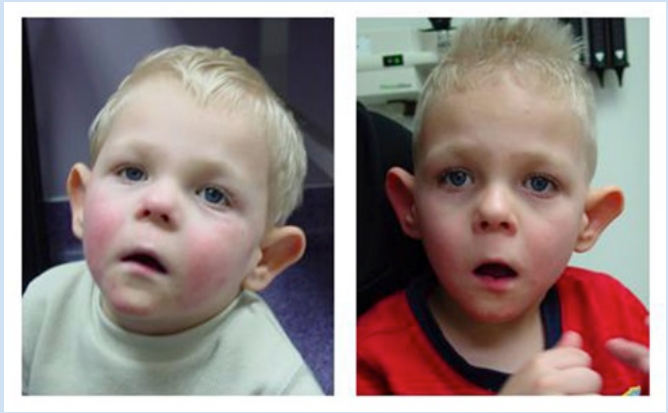
Thomas Arnesen
Nathalie Reuter
Line Myklebust

Ghent, Belgium
Petra Van Damme
Kris Gevaert

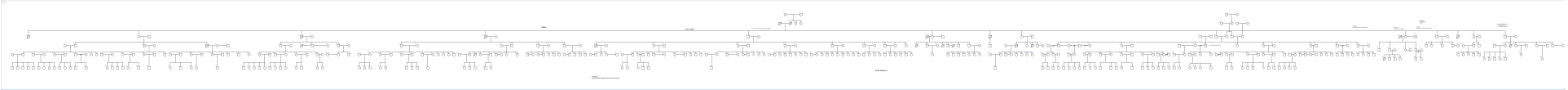
Rare Diseases and Forward Genetics

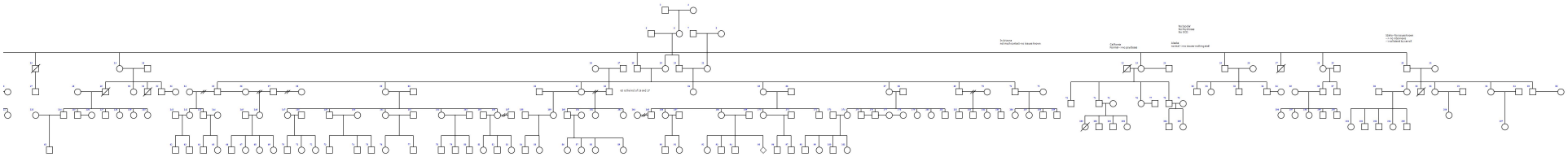


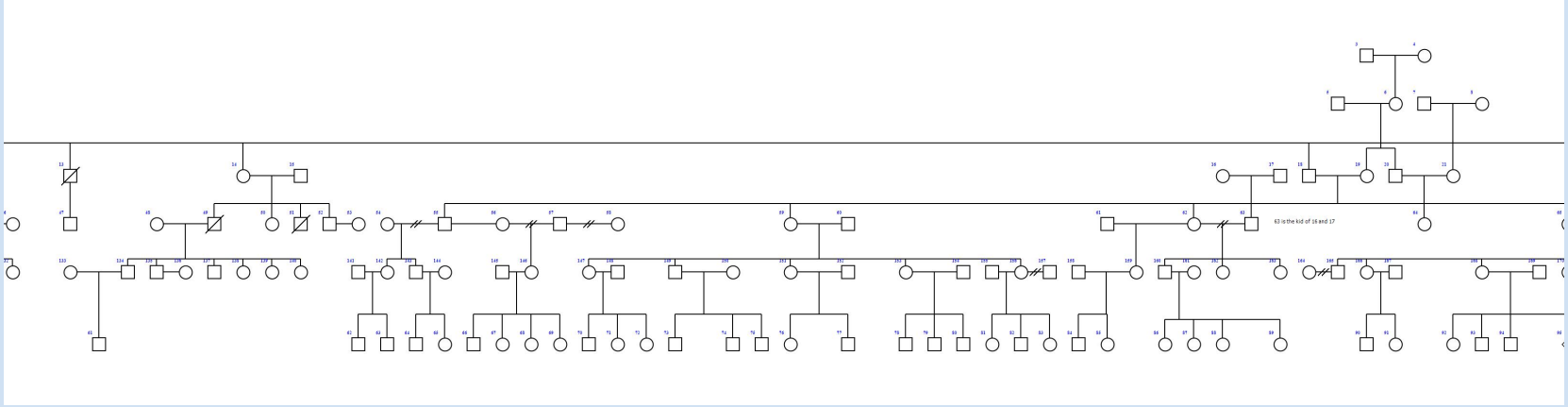




Utah Genome Project







Large pedigrees in human sequencing studies: toward a more resolved and accurate picture of genetic disease

Jason O'Rawe[1], Yiyang Wu[1,2], Alan Rope[3], Laura T. Jimenez Barrón[1,4], Jeffrey Swensen[5], Han Fang[1], David Mittelman[6], Gareth Highnam[6], Reid Robison[7], Kai Wang[7,8], Gholson J. Lyon[1,2,7]

[1]Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, NY, USA; [2]Stony Brook University, Stony Brook, NY, USA; [3]Department of Medical Genetics, Northwest Kaiser Permanente, Portland, OR, USA; [4]Centro de Ciencias Genomicas, Universidad Nacional Autonoma de Mexico, Cuernavaca, Morelos, MX; [5]Caris Life Sciences, Phoenix, Arizona, USA; [6]Gene by Gene, Ltd., Hou-ston, TX, USA ; [7]Utah Foundation for Biomedical Research, Salt Lake City, UT, USA, [8]Zilkha Neu-rogenetic Institute, Department of Psychiatry and Preventive Medicine, University of Southern California, Los Angeles, CA, USA

Abstract

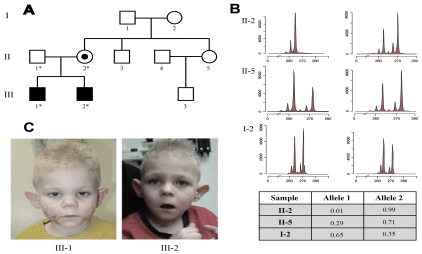
Background: We describe a comprehensive whole genome sequencing (WGS) study using the Illumina and Complete Genomics (CG) sequencing platforms for one family containing two affected male brothers, aged 10 and 12, with severe intellectual disability and very distinctive facial features. High accuracy and sensitivity is of particular importance in the context of detecting or discovering the genetic influencers of human diseases.

Methods: WGS was performed on ten members of this family using the Il-lumina HiSeq2000 platform, with four (the two affected boys and their parents) being additionally sequenced using the CG WGS platform. CG data analysis was performed by CG, using their version 2.0 pipeline. Multiple variant calling pipelines were used to detect SNVs, INDELs, STRs and CNVs. Disease variant prioritization was performed using ANNOVAR, Golden Helix SVS v8.1.4 and GEMINI v0.9.1, and VAAST v2.0.

Results: CG WGS covered >85% of the genome and >90% of the exome, both with 20 or more reads. Illumina WGS covered >90% of the genome with 30 reads or more and with >80% of the bases having a quality score of >30. On average, we find a 2.4 to 14.0 mean fold difference in the number of variants detected as being relevant for various disease models when using different sets of sequencing data and analysis pipelines. We found a number of putative genetic variants and archive them here.

Presentation of the phenotype

- The two affected male brothers have severe intellectual disability, autism-like behavior, attention deficit issues, and very distinctive facial features (**Fig. 1C**), including broad, upturned nose, sagging cheeks, downward sloping palpebral fissures, relative hypertelorism, high-arched palate, and prominent ears.
- The mother of the two affected (**Fig. 1A, II-2**) was shown to have 99:1 x-chromosome inactivation (**Fig. 1B**).



Sample	Allele 1	Allele 2
II-2	0.01	0.99
II-5	0.29	0.71
I-2	0.65	0.35

Fig. 1 (**A**) Pedigree structure of all individuals in the family that were sequenced during the course of this study. Individuals with a star next to their number indicates that their whole genomes were sequenced with both the Complete Genomics sequencing and analysis pipeline as well as with Illumina sequencing.

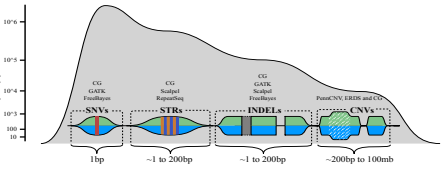
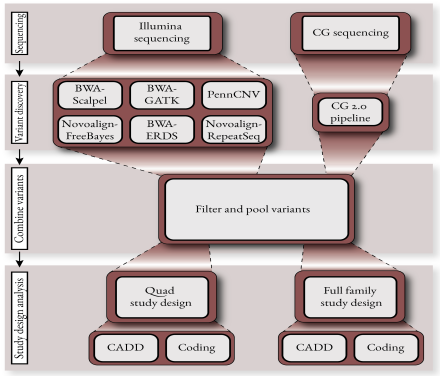
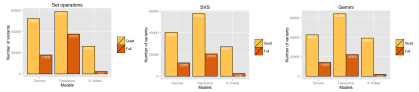


Fig. 2 A conceptual map of human sequence variation, and a list of the bioinformatics programs we used during the course of our study.



Bioinformatics analysis

- Human sequence variation ranges in manifestation from differences that can be detected at the single nucleotide level, to whole chromosome differences (**Fig 2**).
- We sought to identify variants following de-novo, autosomal recessive and x-linked models of transmission that may be contributing, together or alone, to the disease phenotype. We used several methods to prioritize and identify possible disease-contributory germ-line variants, in-cluding VAAST, Golden Helix SVS v8.1.4, ANNOVAR (2013Aug23 version), and GEMINI v0.9.1.



- We found a 2.4 to 14.0 mean fold difference in the number of variants detected as being relevant for various disease models when using different sets of sequencing data.

Results

- 14 unique INDELS and SNVs were discovered using two different prioritization (**Fig. 3A**).
- The TAF1 variant arose in this family as a de-novo variant on the X-chromosome of the mother (**Fig. 1A, II-2**) of the two affected children (as it is not found in any of the other members of the family) and was then transmitted to both of them. The mother also is the only female in the family to exhibit extreme X-chromosome skewing.
- The transcription factor initiation complex TF11D has recently been implicated in playing a role in intellectual disability and developmental delay, and TAF1 represents the largest known subunit of this multi-protein complex, and this variant falls within a conserved region of the protein (**Fig 3B**).

A

Model	Location	Ref	Alt	Variant Caller	Annotation	Function	Scheme
Recessive	chr1:210851705	TT	T	CG, GATK, FreeBayes, RapaorSeq	ANNOVAR, GEMINI, SVS	KCNH1:UTR3	CADD, score:27.5
Recessive	chr1:224772440	AATAATTTG	TA	CG, GATK, FreeBayes	GEMINI	intergenic	CADD, score:22.1
Recessive	chr2:60537356	TTTTATTT	ATTATTA	CG, FreeBayes, GATK, RapaorSeq	GEMINI	intergenic	CADD, score:22.3
Recessive	chr8:109098066	AT	A	CG, FreeBayes, GATK, RapaorSeq	GEMINI	intergenic	CADD, score:24.6
Recessive	chr15:66786022	ACAAA	A	FreeBayes, GATK	GEMINI	SNAPC5:intronic	CADD, score:23.6
Recessive	chr16:49061346	TA	T	CG, FreeBayes, GATK	ANNOVAR, GEMINI	intergenic	CADD, score:25.3
Recessive	chr16:49612367	GAC	G	CG, FreeBayes, GATK	GEMINI, SVS	ZNF423:intronic	CADD, score:20.5
Recessive	chr10:135438929	T	G	CG, FreeBayes, GATK	ANNOVAR, GEMINI, SVS	NM_001080998: I171L	Coding, gene:FRG2B
Recessive	chr10:135438951	GGCCC	AGCCT	FreeBayes, Scalpel	GEMINI, SVS	NM_001080998:	Coding, gene:FRG2B
Recessive	chr10:135438967	C	T	GATK, FreeBayes	GEMINI, SVS	NM_001080998: K158Q	Coding, gene:FRG2B
Recessive	chr15:85438314	C	CTTG	CG, FreeBayes, GATK, Scalpel	GEMINI	NM_201635: K141delinsIE	Coding, gene:SLC28A1
De-novo	chr1:53925373	G	GCCGCCC	FreeBayes, CG, Scalpel	GEMINI, SVS	NM_033067: A83delinsAAP	Coding, gene:DMRTB1
X-linked	chrX:34961492	T	C	CG, FreeBayes, GATK	GEMINI	NM_182631: V182H	Coding, gene:FAM47B
X-linked	chrX:70621541	T	C	CG, FreeBayes, GATK	ANNOVAR, GEMINI, SVS	NM_004606:I133T	Coding, gene:TAF1; CADD, score:22.9

B

```
H.sapiens       1332  D--MEELRKVEGTKIVL----------  1346
M.mulatta       1244  D--MEELRKVEGTKIVL----------  1258
C.lupus         1332  D--MEELRKVEGTKIVL----------  1346
B.taurus        1316  D--MEELRKVEGTKIVL----------  1330
M.musculus      1343  D--MEELRKVEGTKIVL----------  1357
R.norvegicus    1332  D--MEELRKVEGTKIVL----------  1346
G.gallus        1357  D--MEELRKVEGTKIVL----------  1371
X.tropicalis    1377  D--MEELRKVEGTKIVL----------  1391
D.rerio         1396  D--D-DLRKVDGTKVIL----------  1409
D.melanogaster  1419  D--EGDLRKVDGTKVIL----------  1433
A.gambiae       1263  SLVADPAMQVDGTKVKPNLNFAEIRKEQNREEKELRKLAKMAEAAVRERQ  1312
C.elegans       1294  D--MEELRKVEGTKIVL----------  1308
```

Fig. 3

- *Conclusions:* Analyzing multi-generational pedigrees using multiple orthogonal bioinformatics pipelines using two sequencing platforms can reliably reveal human sequence variants that may be important in rare disease. We have found a number of sequence variants that may play a role in the rare disease described here and highlight a variant in TAF1. Our findings are consistent with the literature on the importance of the TF11D complex in developmental delay and ID.

Reducing INDEL calling errors in whole genome and exome sequencing data

Han Fang[1,2,3], Yiyang Wu[1,2], Giuseppe Narzisi[3,4], Jason A. O'Rawe[1,2], Laura T. Jimenez Barrón[1,5], Julie Rosenbaum[3], Michael Ronemus[3], Ivan Iossifov[3], Michael C. Schatz[3§], Gholson J. Lyon[1,2§]

[1] Stanley Institute for Cognitive Genomics, One Bungtown Road, Cold Spring Harbor Laboratory, NY, USA; [2] Stony Brook University, 100 Nicolls Rd, Stony Brook, NY, USA; [3] Simons Center for Quantitative Biology, One Bungtown Road, Cold Spring Harbor Laboratory, NY, USA; [4] New York Genome Center, New York, NY; [5] Centro de Ciencias Genomicas, Universidad Nacional Autonoma de Mexico, Cuernavaca, Morelos, MX;

INDELs, especially those disrupting protein-coding regions of the genome, have been strongly associated with human diseases. However, there are still many errors with INDEL variant calling, driven by library preparation, sequencing biases, and algorithm artifacts. We characterized whole genome sequencing (WGS), whole exome sequencing (WES), and PCR-free sequencing data from the same samples to investigate the sources of INDEL errors. We also developed a classification scheme based on the coverage and composition to rank high and low quality INDEL calls. We performed a validation experiment on 600 loci, and find high-quality INDELs to have a substantially lower error rate than low quality INDELs (7% vs. 51%).

Simulation and experimental data show that assembly based callers are significantly more sensitive and robust for detecting large INDELs (>5bp) than alignment based callers, consistent with published data. The concordance of INDEL detection between WGS and WES is low (52%), and WGS data uniquely identifies 10.8-fold more high-quality INDELs. The validation rate for WGS-specific INDELs is also much higher than that for WES-specific INDELs (84% vs. 57%), and WES misses many large INDELs. In addition, the concordance for INDEL detection between standard WGS and PCR-free sequencing is 71%, and standard WGS data uniquely identifies 6.3-fold more low-quality INDELs. Furthermore, accurate detection with Scalpel of heterozygous INDELs requires 1.2-fold higher coverage than that for homozygous INDELs.

Lastly, homopolymer A/T INDELs are a major source of low-quality INDEL calls, and they are highly enriched in the WES data. Overall, we show that accuracy of INDEL detection with WGS is much greater than WES even in the targeted region. We calculated that 60X WGS depth of coverage from the HiSeq platform is needed to recover 95% of INDELs detected by Scalpel. While this is higher than current sequencing practice, the deeper coverage may save total project costs because of the greater accuracy and sensitivity. Finally, we investigate sources of INDEL errors (e.g. capture deficiency, PCR amplification, homopolymers) with various data that will serve as a guideline to effectively reduce INDEL errors in genome sequencing.

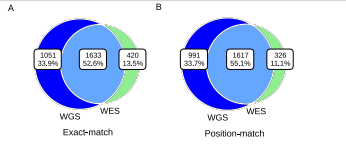


Figure 1. Concordance of INDELs over eight samples between WGS (blue) and WES (green) data.

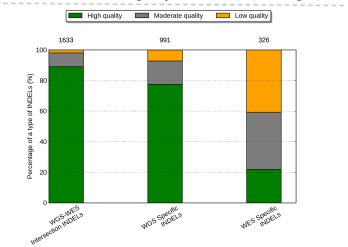


Figure 2. Percentage of high quality, moderate quality and low quality INDELs in three call set.

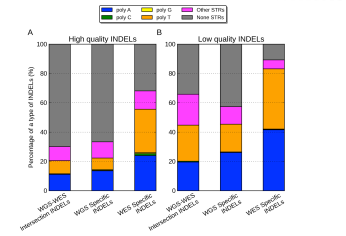


Figure 3. Percentage of poly-A, poly-C, poly-G, poly-T, other-STR, and non-STR in three call set.

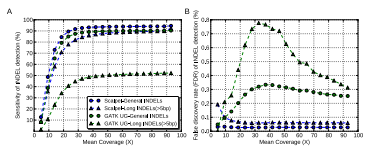


Figure 4. Performance comparison between the Scalpel and GATK-UnifiedGenotyper in terms of sensitivity (A) and false discovery rate (B) at different coverage (simulation data).

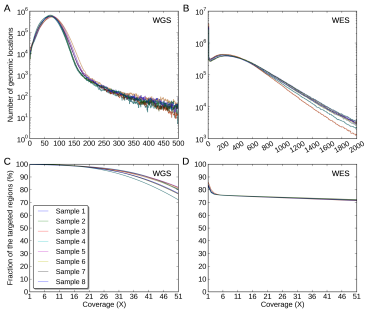


Figure 5. Coverage distributions of the exonic targeted regions in (A) the WGS data, (B) the WES data.

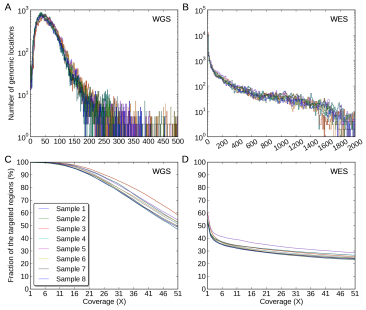


Figure 6. Coverage distributions of the WGS-specific INDELs regions in (A) the WGS data, (B) the WES data.

Table 2. Validation rates of WGS-WES intersection INDELs, WGS-specific, and WES-specific INDELs. We also calculated the validation rates of large INDELs (>5 bp) in each category. The validation rate, positive predictive value (PPV), is computed by the following: PPV=#TP/(#TP+#FP), where #TP is the number of true-positive calls and #FP is the number of false-positive calls.

	INDELs	Valid	PPV	INDELs (>5bp)	Valid (>5bp)	PPV (>5bp)
WGS-WES intersection	160	152	95.0%	18	18	100%
WGS-specific	145	122	84.1%	33	25	75.8%
WES-specific	161	91	56.5%	1	1	100%

Table 3. Number and fraction of large INDELs in the following INDEL categories: 1) WGS-WES intersection INDELs, 2) WGS-specific, and WES-specific.

	All INDELs	Large INDELs (>5bps)	Fraction of large INDELs (>5bp)
WGS-WES intersection	2009	176	8.8%
WGS-specific	494	104	21.1%
WES-specific	674	10	1.5%

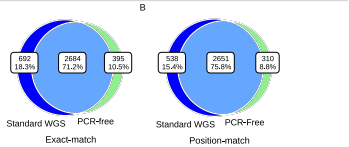


Figure 7. Concordance of INDEL detection between PCR-free and standard WGS data on NA12878.

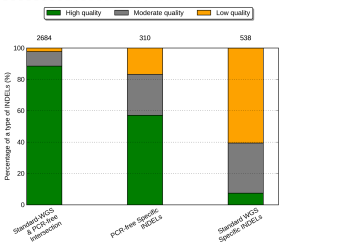


Figure 8. Percentage of high quality, moderate quality and low quality INDELs in two datasets.

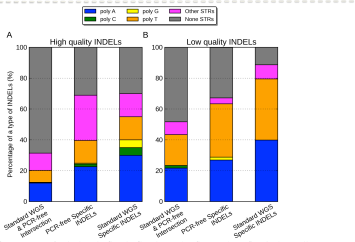


Figure 9. Percentage of poly-A, poly-C, poly-G, poly-T, other-STR, and non-STR in three call sets.

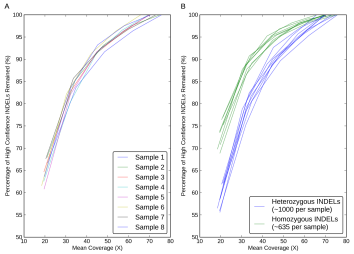


Figure 10. Sensitivity performance of INDEL detection with eight WGS datasets at different mean coverages on Illumina HiSeq2000 platform.

Genome Wide Variant Analysis of families with Autism Spectrum Disorder (ASD) using an Integrative Bioinformatics Pipeline

Laura T Jimenez-Barron[1,2], Han Fang[1], Jason O'Rawe[1], Ivan Iossifov[1], Gholson J Lyon[1,3]

[1]Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics, New York, NY, [2]Universidad Nacional Autonoma de Mexico, Centro de Ciencias Genomicas, Cuernavaca, Mexico, [3]Utah Foundation for Biomedical Research, Salt Lake City, UT.

Introduction. Autism spectrum disorders (ASD) are a group of developmental disabilities that affect social interaction, communication and are characterized by repetitive behaviors. There is now a large body of evidence that suggests a complex role of genetics in ASD, in which many different loci are involved. Although many current population scale studies have been demonstrably useful, these studies generally focus on analyzing a limited part of the genome or use a limited set of bioinformatics tools. These limitations make it difficult to see the complete and panoramic picture of each ASD case. To address this problem, here we describe an integrative bioinformatics pipeline used to get a more complete and reliable set of candidate ASD-variants for validation and further functional analysis.

Methods. We studied three simplex Autism Families, two of which belong to the Simon's Simplex Collection (SSC), and all probands and families were clinically evaluated and extensively phenotyped. The third family, recruited at the Utah Foundation for Biomedical Research, had extensive clinical evaluations performed, along with fragile X and Chromosomal Microarray Analysis (CMA) on the proband and mother, with no obvious disease-contributory mutations found. All family members were genotyped using an Illumina Omni2.5 Array and/or WGS was performed using the Illumina HiSeq 2000 to ~40-75X coverage. WGS reads were aligned to the GRCh37/hg19 human reference genome using BWA-MEM software, with variant calling for SNVs and INDELs using the GATK HaplotypeCaller and FreeBayes. To better support de novo calls, we used Scalpel for INDEL detection and the Multinomial Analyzer. The ERDS software was used to call CNVs from WGS data. Microarray data were used to call CNVs with the software package PennCNV using the joint-calling algorithm.

Results. The resulting set of candidate variants include three small heterozygous CNVs (~22, ~36 and ~50 Kb). All of the CNVs were only found by ERDS, and despite the fact that the K21 pedigree had microarray data, PennCNV did not detect any CNV in those regions. A heterozygous *de novo* nonsense mutation in *MYBBP1A* was found in one of the quads (K21) located within exon 1, and a second *de novo* variant was also among the final results from another quad (SSC_2), this time a missense mutation in *LAMB3*, which also has not yet been observed in any other ASD proband.

Having established a more comprehensive WGS pipeline, we are moving to implement our framework for the analysis and study of families from Utah and from the SSC .

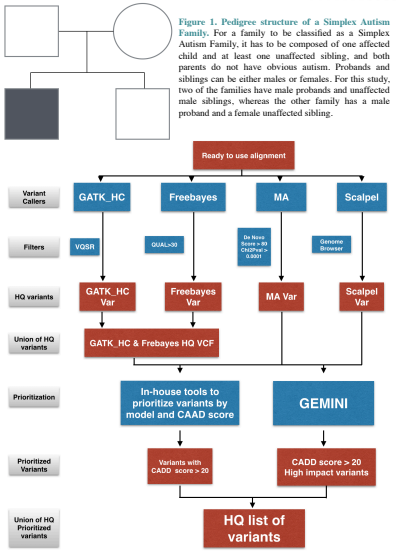


Figure 1. Pedigree structure of a Simplex Autism Family. For a family to be classified as a Simplex Autism Family, it has to be composed of one affected child and at least one unaffected sibling, and both parents do not have obvious autism. Probands and siblings can be either males or females. For this study, two of the families have male probands and unaffected male siblings, whereas the other family has a male proband and a female unaffected sibling.

Figure 2. Variant calling pipeline. After aligning the genomes with BWA-MEM 0.7.5a-r405, the resulting alignments were converted to binary format, then sorted and indexed using SAMtools version 0.1.19-44428cd. Duplicated reads were marked and read groups were assigned to each lane using Picard tools **v1.84**. The GATK Indel realigner v3.0-0 was used to correct initial mapping artifacts due to reads aligning to the edges of INDELs, which often map with mismatching bases that may look like evidence for SNPs, while they are not. The GATK Base Quality Score Recalibrator was also used to correct known systematic errors of sequencing technologies. Finally all lanes were merged by sample with Picard tools to generate a ready-to-use alignment. Various algorithms were used to call SNPs and Indels, all resulting variants were filtered and prioritized with different methods.

Table 1. Final set of Small Variants.

Model	Ref->Alt/ Effect	Location	Affected Gene	Algorithms that called the variant	Pedigree ID	ExAC Allele Frequency	CAAD score
De Novo	sub(C->T) missense	chr1: 209823359	LAMB3	Freebayes, Multinomial Analyzer, GATK	SSC_2	0	22.7
De Novo	sub(G-A)nonsense	chr17: 4458481	MYBBP1A	Freebayes, Multinomial Analyzer, GATK	K21	1/74014=0.00001351	40

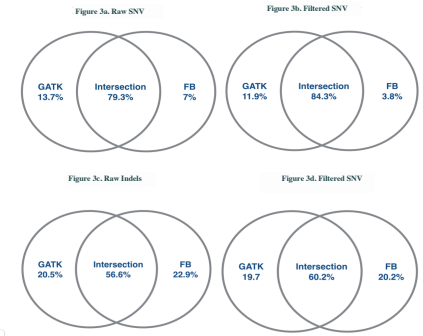


Figure 3. Algorithm concordance. GATK and Freebayes were compared as they are algorithms that call both SNVs and INDELs with a comparable number of calls. The intersection grows when comparing filtered SNVs and Indels. The Multinomial Analyzer and Scalpel were only used to call de novo SNPs and Indels respectively.

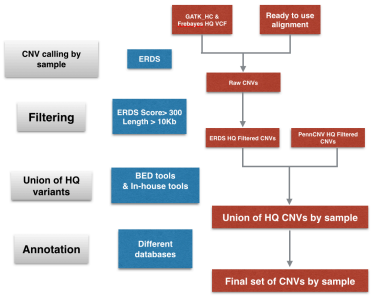


Figure 4. Copy Number Variant calling pipeline. Using the same ready to use alignment described in Figure 2 plus the union of variants called by Freebayes and GATK, the Estimation by Read Depth with Single Nucleotide Variants (ERDS) software was used to call CNVs. PennCNV was used in the samples where Microarray data was available and both calls sets were compared.

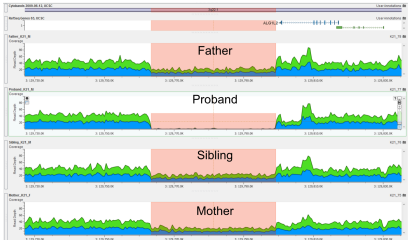


Figure 5c. Genome Browser Screen cut for the Read Depths in the ~ 36 Kb intergenic CNV on 3q22.1 (Pedigree K21).

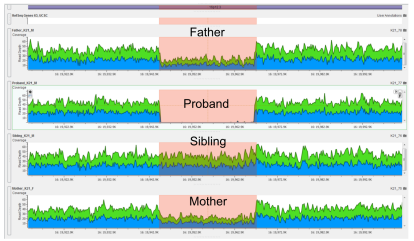
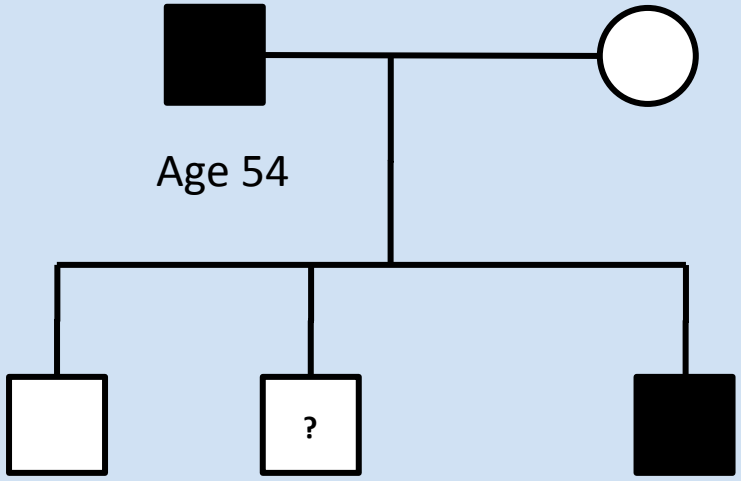


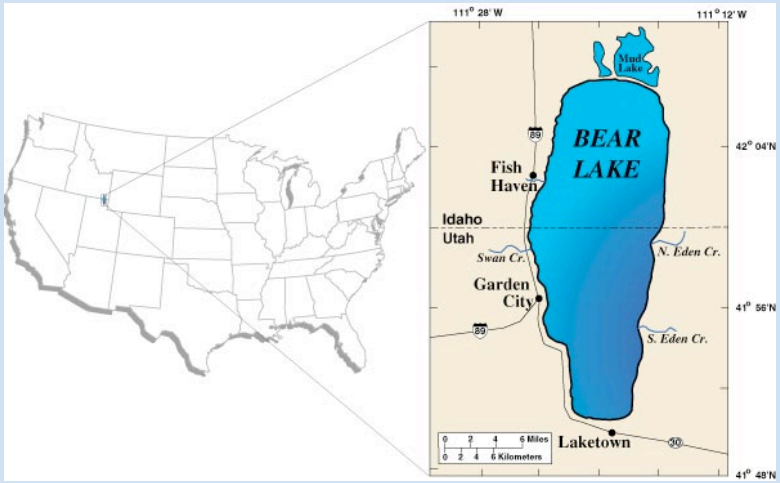
Figure 5c. Genome Browser Screen cut for the Read Depths in the ~ 22 Kb intergenic CNV on 16p12.3 (Pedigree K21).

Figure 5c. Genome Browser Screen cut for the Read Depths in the ~ 50 Kb intergenic CNV on 4p16.3 (Pedigree SSC_2).

- Mark Yandell will visit CSHL from Utah and give a seminar about the Utah Genome Project on Wednesday, December 10th at 4 PM in Hawkins.

Expanding the Pedigree –K8101





Collected ~100 DNA samples from the extended family, due to very large excess of major depression, bipolar, Tourette and OCD.

Case Presentation

- Male, age 55 currently.
- Psychotic break at age 20 with bipolar features.
- Evolution into schizoaffective disorder over next 25 years.
- Also with severe obsessive compulsive disorder and severe Tourette Syndrome
- At least two very severe suicide attempts at age 22, including throwing self under a truck one time and then driving head-on into another car (with death of two passengers in other car, found not guilty by reason of insanity).

- Extensive medication trials over many years, along with anterior capsulotomy with very little effect for the OCD.

- Current meds:

Klonopin	Lithium
Nicotinamide	Seroquel
Lunesta	Lamictal
Ativan	Luvox

Randomness – *sluchainost'*

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

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]

Contents lists available at SciVerse ScienceDirect

Applied & Translational Genomics

journal homepage: www.elsevier.com/locate/atg

Practical, ethical and regulatory considerations for the evolving medical and research genomics landscape

Gholson J. Lyon [a,b,*], Jeremy P. Segal [c,**]

[a] *Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, NY, United States*
[b] *Utah Foundation for Biomedical Research, Salt Lake City, UT, United States*
[c] *New York Genome Center, New York City, NY, United States*

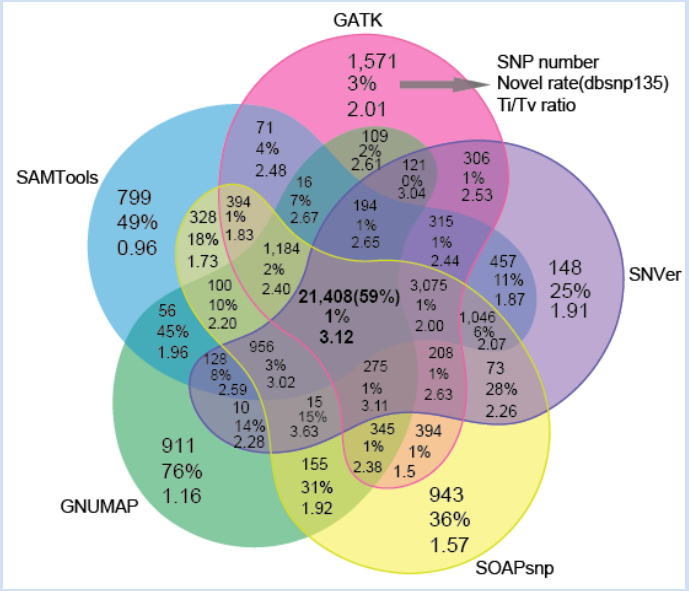
O'Rawe *et al. Genome Medicine* 2013, **5**:28
http://genomemedicine.com/content/5/3/28

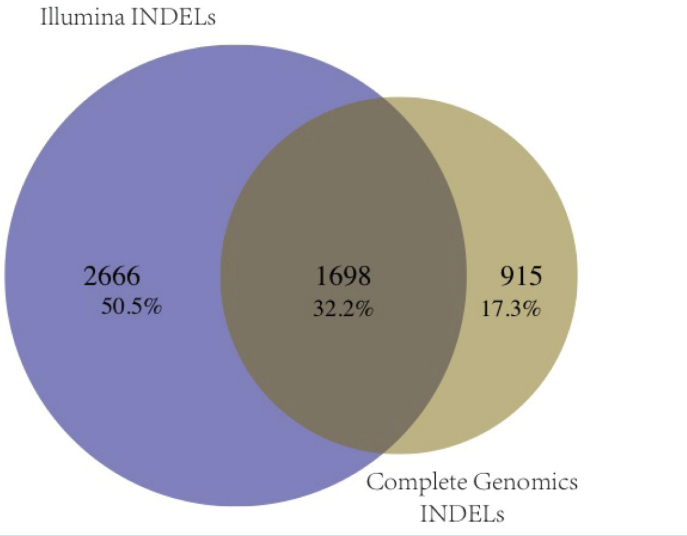
Genome Medicine

RESEARCH Open Access

Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing

Jason O'Rawe[1,2], Tao Jiang[3], Guangqing Sun[3], Yiyang Wu[1,2], Wei Wang[4], Jingchu Hu[3], Paul Bodily[5], Lifeng Tian[6], Hakon Hakonarson[6], W Evan Johnson[7], Zhi Wei[4], Kai Wang[8,9*] and Gholson J Lyon[1,2,9*]





Fang *et al. Genome Medicine* 2014, **6**:89
http://genomemedicine.com/content/6/11/89

Genome Medicine

RESEARCH Open Access

Reducing INDEL calling errors in whole genome and exome sequencing data

Han Fang[1,2,3], Yiyang Wu[1,2], Giuseppe Narzisi[3,4], Jason A O'Rawe[1,2], Laura T Jimenez Barrón[1,5], Julie Rosenbaum[3], Michael Ronemus[3], Ivan Iossifov[3], Michael C Schatz[3*] and Gholson J Lyon[1,2*]

Accurate *de novo* and transmitted indel detection in exome-capture data using microassembly

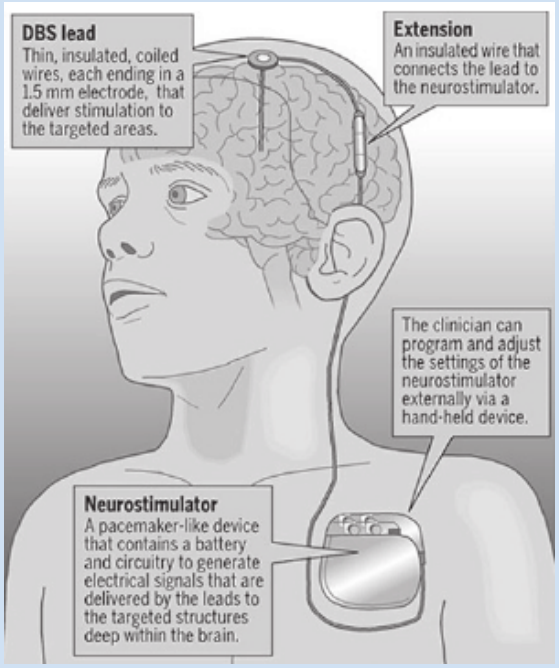
Giuseppe Narzisi[1,2], Jason A O'Rawe[3,4], Ivan Iossifov[1], Han Fang[3,4], Yoon-ha Lee[1], Zihua Wang[1], Yiyang Wu[3,4], Gholson J Lyon[3,4], Michael Wigler[1] & Michael C Schatz[1]

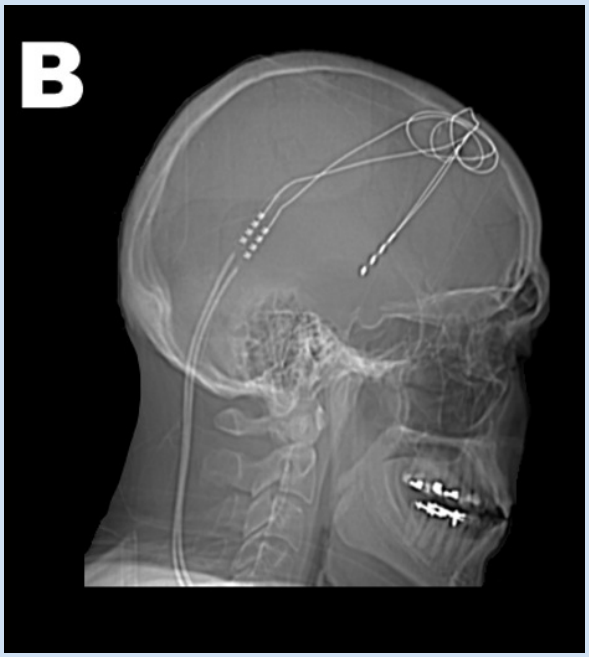
PeerJ

Integrating precision medicine in the study and clinical treatment of a severely mentally ill person

Jason A. O'Rawe[1,2], Han Fang[1,2], Shawn Rynearson[3], Reid Robison[4], Edward S. Kiruluta[5], Gerald Higgins[6], Karen Eilbeck[3], Martin G. Reese[5] and Gholson J. Lyon[1,2,4]

[1] Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, NY, USA
[2] Stony Brook University, Stony Brook, NY, USA
[3] Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA
[4] Utah Foundation for Biomedical Research, Salt Lake City, UT, USA
[5] Omicia Inc., Emeryville, CA, USA
[6] AssureRx Health, Inc., Mason, OH, USA





Cold Spring Harbor Laboratory

BioRxiv is great.

bioRxiv beta

THE PREPRINT SERVER FOR BIOLOGY

Human genetics and clinical aspects of neurodevelopmental disorders

Gholson J Lyon and Jason O'Rawe

bioRxiv first posted online November 18, 2013
Access the most recent version at doi: http://dx.doi.org/10.1101/000687

"There are ~12 billion nucleotides in every cell of the human body, and there are ~25-100 trillion cells in each human body. Given somatic mosaicism, epigenetic changes and environmental differences, no two human beings are the same, particularly as there are only ~7 billion people on the planet".

PubPeer

The online journal club

Search by DOI, PMID, arXiv ID, keyword, author, etc.

The PubPeer database contains all articles. Search results return articles with comments.
To leave a new comment on a specific article, paste a unique identifier such as a DOI, PubMed ID, or arXiv ID into the search bar.

Search Publications

PubPeer comments on PubMed and journal websites with our browser extension!

Follow @PubPeer 4,036 followers

Textbook chapter soon to be published in:

The Genetics of Neurodevelopmental Disorders

K. Mitchell

ISBN: 978-1-118-52488-6

374 pages
May 2015, Wiley-Blackwell

http://www.wiley.com/WileyCDA/WileyTitle/productCd-1118524888.html

Involvement with Industry

Advisory Boards

genepeeks

micia
UNLOCKING INDIVIDUALIZED MEDICINE

Other non-paid consulting:

GOLDEN HELIX
Accelerating the Quest for Significance™

TUTE GENOMICS

Discovering idiopathic orphan diseases: Ogden Syndrome and the Nt-acetylation of proteins.

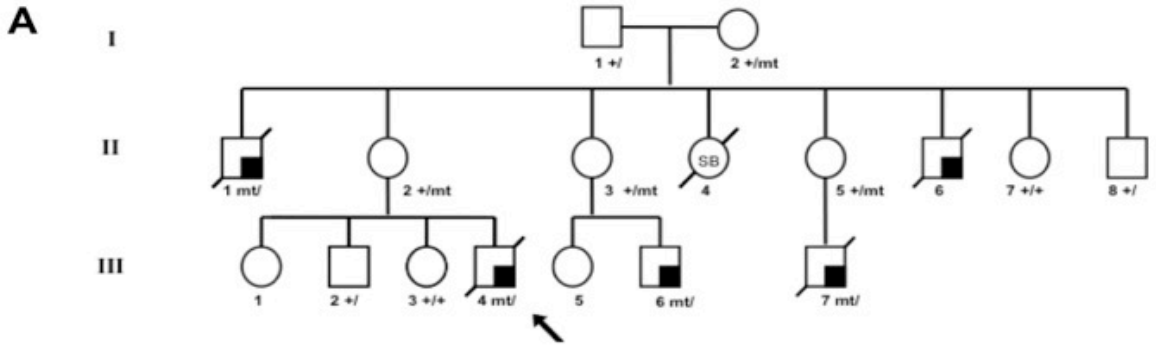
ARTICLE

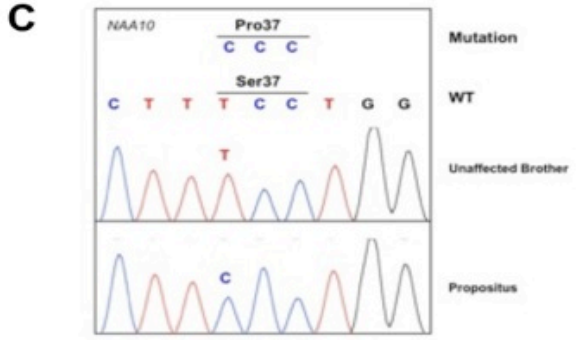
Using VAAST to Identify an X-Linked Disorder Resulting in Lethality in Male Infants Due to N-Terminal Acetyltransferase Deficiency

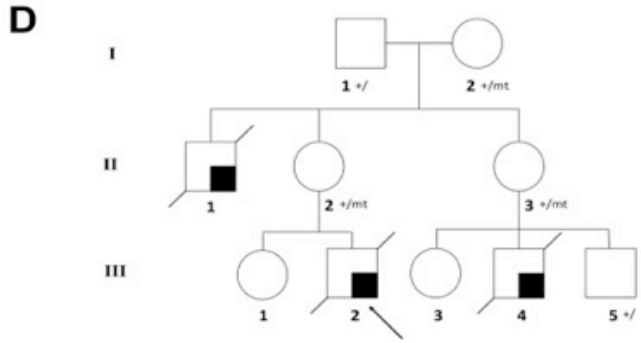
Alan F. Rope,[1] Kai Wang,[2,19] Rune Evjenth,[3] Jinchuan Xing,[4] Jennifer J. Johnston,[5] Jeffrey J. Swensen,[6,7] W. Evan Johnson,[8] Barry Moore,[4] Chad D. Huff,[4] Lynne M. Bird,[9] John C. Carey,[1] John M. Opitz,[1,4,6,10,11] Cathy A. Stevens,[12] Tao Jiang,[13,14] Christa Schank,[8] Heidi Deborah 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

Ogden Syndrome





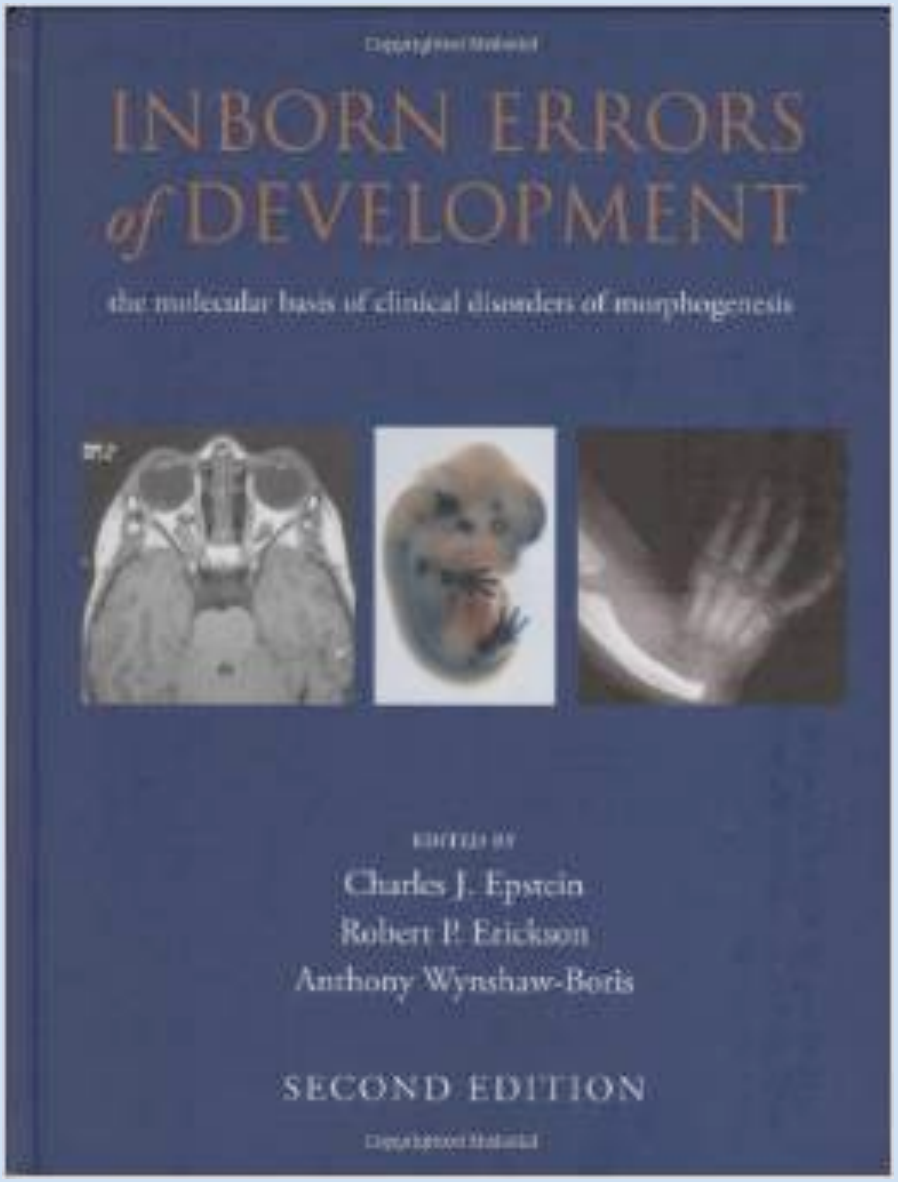


We found the SAME mutation in two unrelated families, with a very similar phenotype in both families, helping prove that this genotype contributes to the phenotype observed.

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.	

Textbook Chapter on Ogden Syndrome in pending 3rd Edition



COLD SPRING HARBOR
Molecular Case Studies
A journal of precision medicine

CSH Molecular Case Studies is an open-access, peer-reviewed, international journal in the field of precision medicine. Articles in the journal present genomic and molecular analyses of individuals or cohorts alongside their clinical presentations and phenotypic information. The journal's purpose is to rapidly share insights into disease development and treatment gained by application of genomics, proteomics, metabolomics, biomarker analysis, and other approaches.

The journal covers the fields of cancer, complex diseases, monogenic disorders, neurological conditions, orphan diseases, infectious disease, and pharmacogenomics. It has a rapid peer-review process that is based on technical evaluation of the analyses performed, not the novelty of findings, and offers a swift, clear path to publication.

The journal publishes:

- **Research Reports** presenting detailed case studies of individuals and small cohorts
- **Research Articles** describing more extensive work using larger cohorts and/or functional analyses
- **Follow-up Reports** linked to previous observations
- Plus **Review Articles**, **Editorials**, and **Position Statements** on best practices for research in precision medicine

Coming soon from Cold Spring Harbor Laboratory

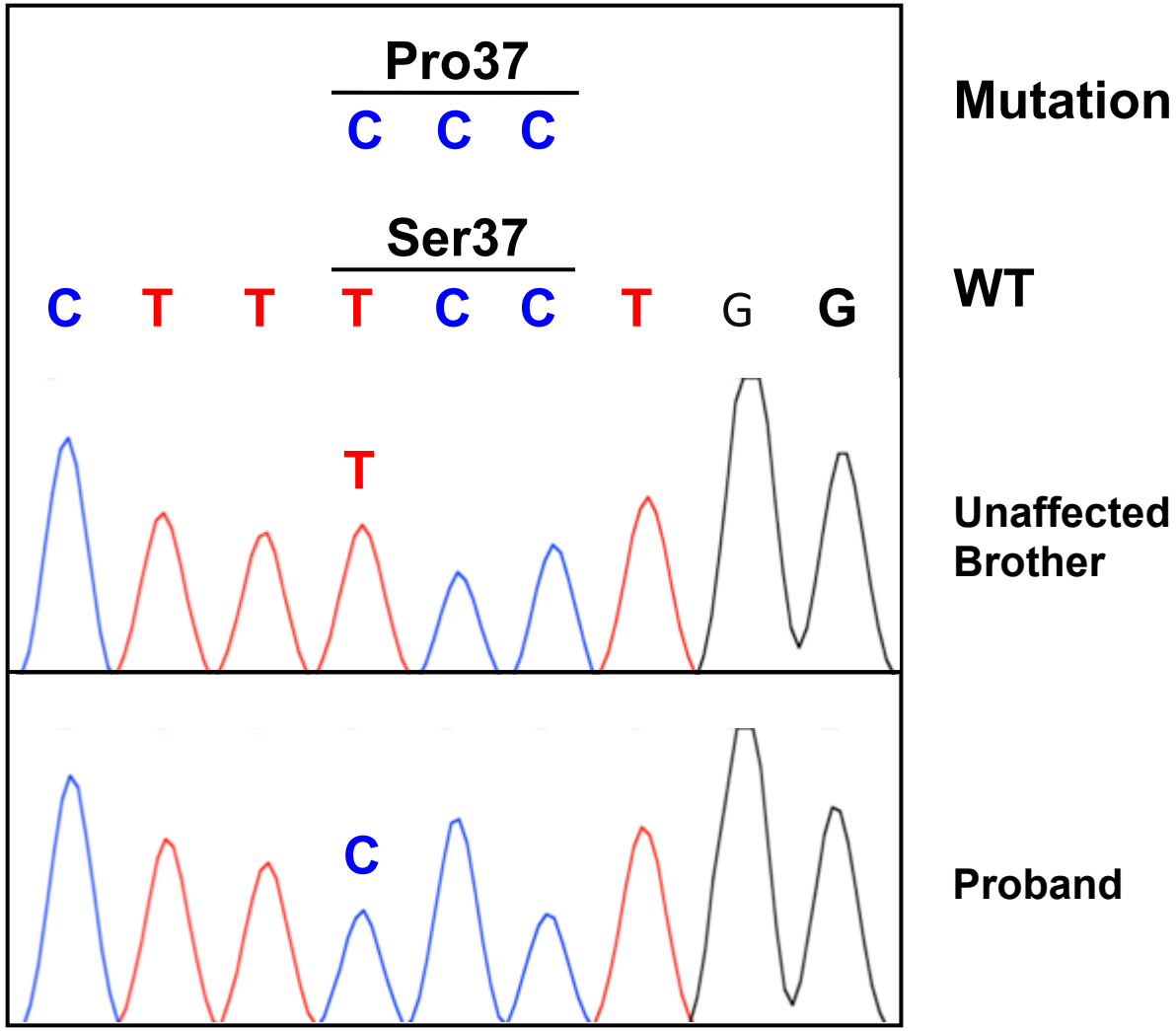
Keep informed! Enter your email address for the lastest news, updates, and author information.

Enter your email address *

Submit

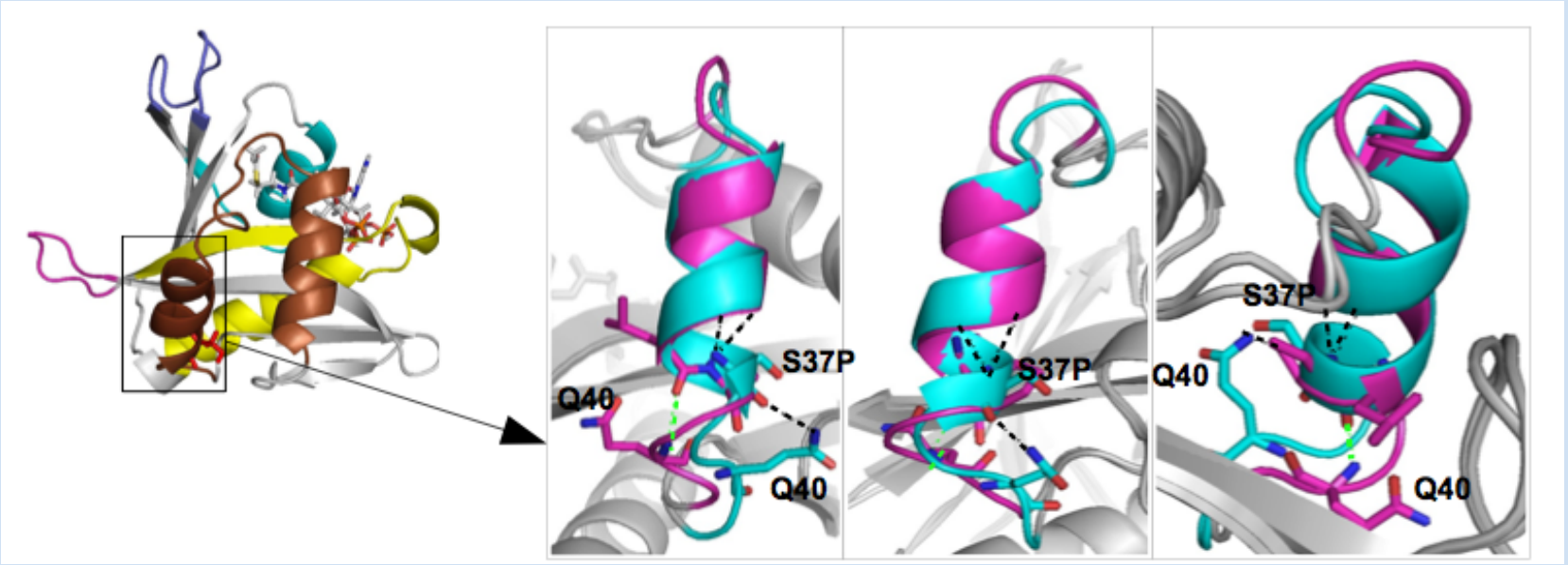
Note: Cold Spring Harbor Laboratory Press will not publish, share, or sell your email address to anyone.

This is the mutation we found… one nucleotide change out of 6 billion nucleotides in a diploid genome.

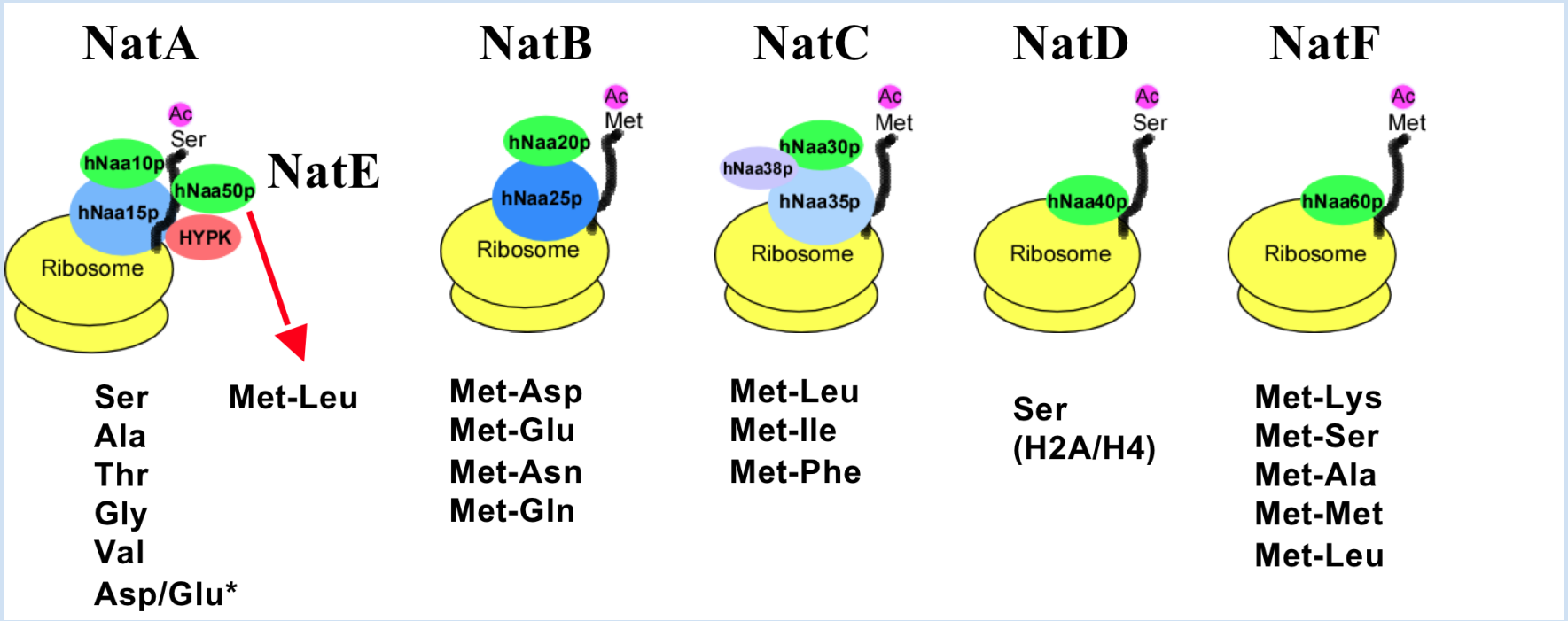


The mutation is a missense resulting in Serine to Proline change in Naa10p

- Ser 37 is conserved from yeast to human
- Ser37Pro is predicted to affect functionality (SIFT and other prediction programs)
- Structural modelling of hNaa10p wt (cyan) and S37P (pink)



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



Slide courtesy of Thomas Arnesen

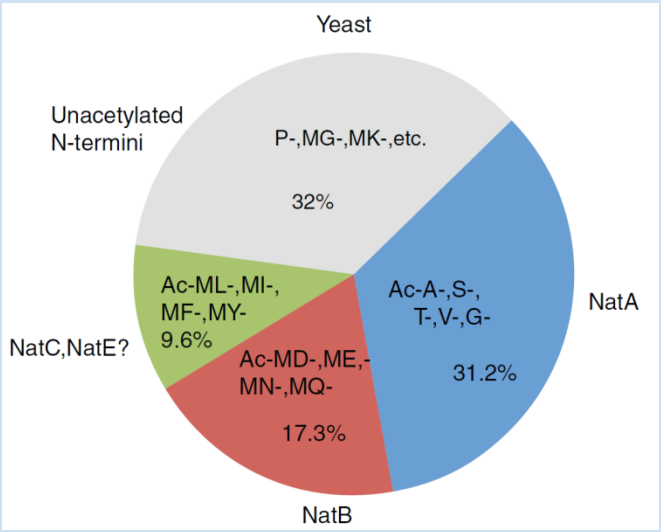
function of N$^{\alpha}$-terminal acetylation

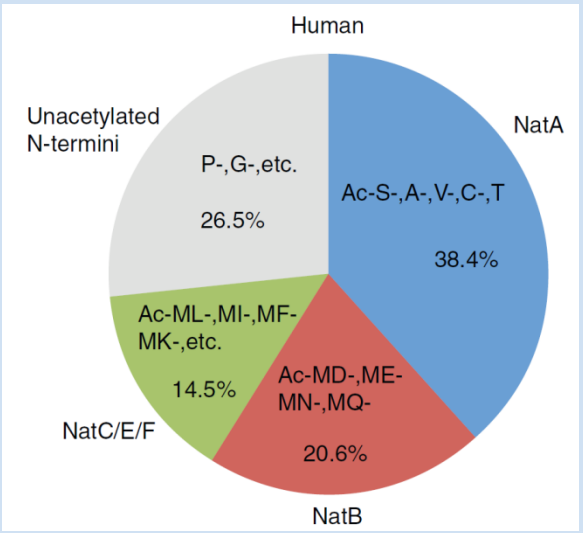
general

- most abundant protein modification in eukaryotes
- NatA is the major NAT

NAT function

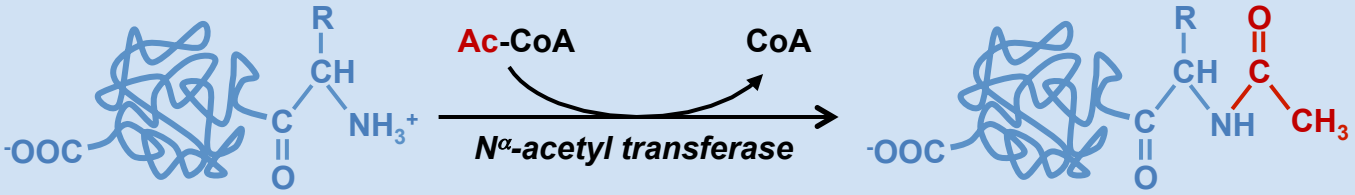
- protein function (hemoglobin, actin/tropomyosin…)
- protein stability
- proteasomal degradation via ubiquitin ligase Doa10
- avidity enhancer
- protein targeting to ER





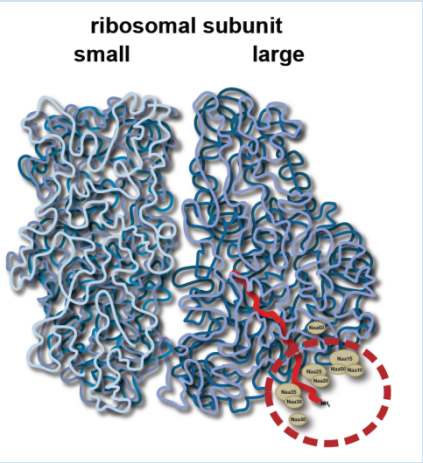
Trends in Biochemical Sciences April 2012, Vol. 37, No. 4

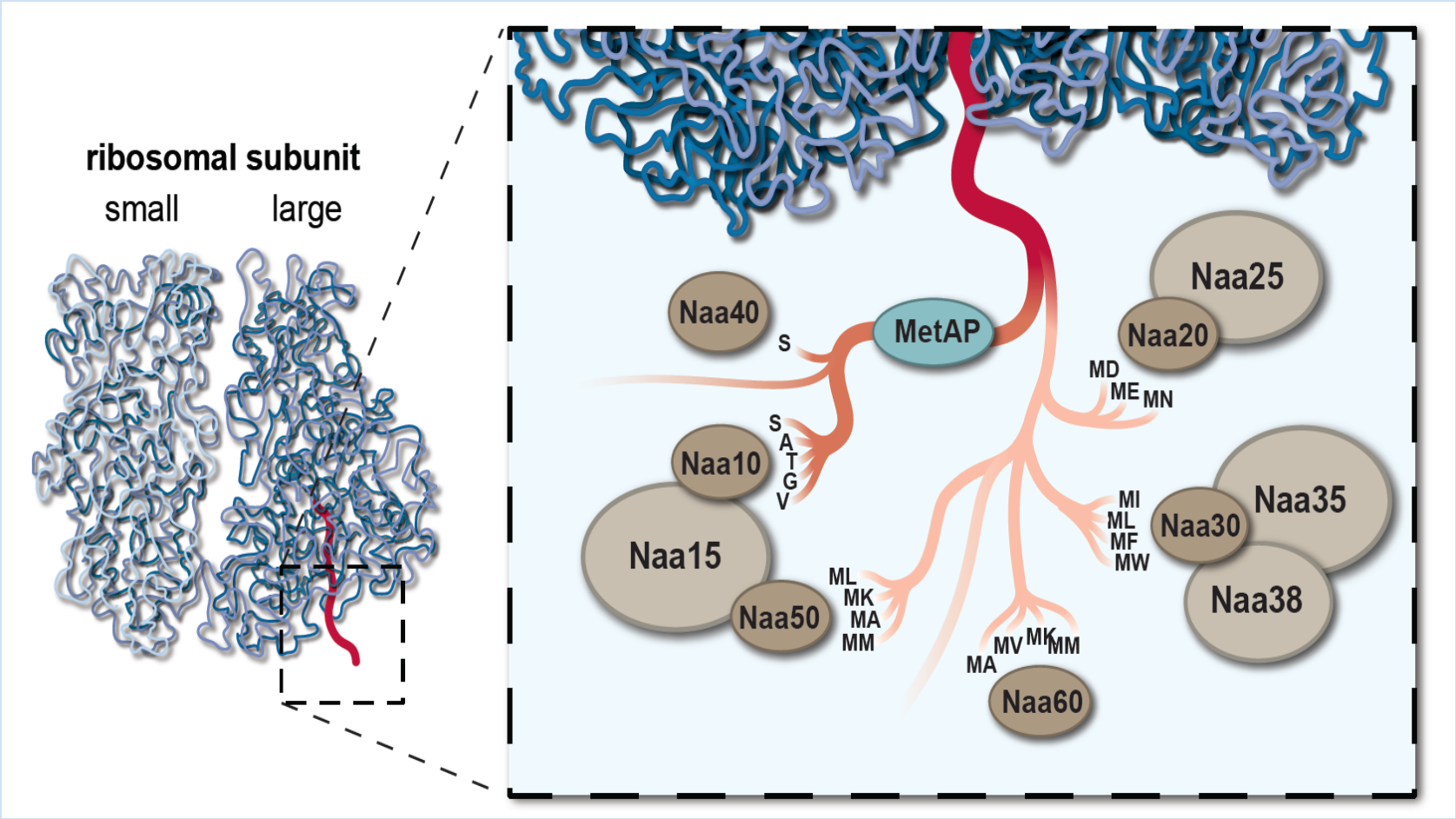
N$^{\alpha}$-terminal acetyltransferases

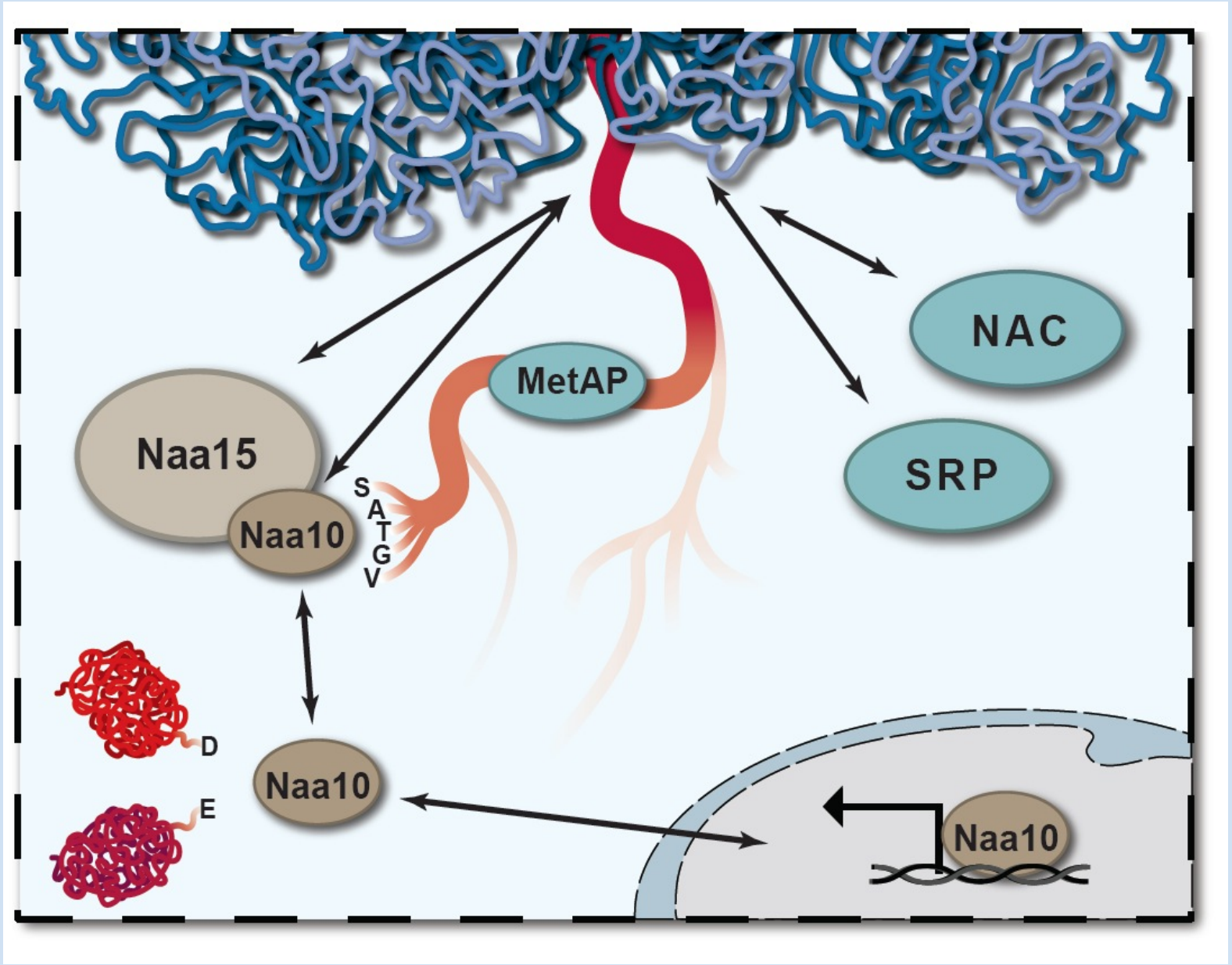


human NATs

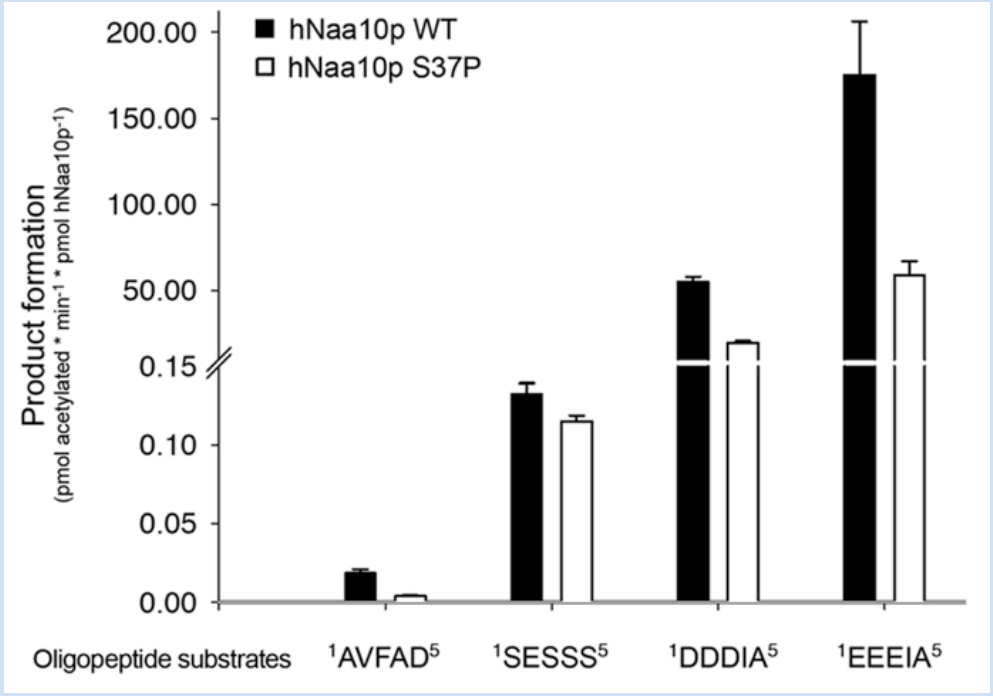
- NatA-NatF
- associated with ribosome
- act co-translationally
- distinct substrate specificity

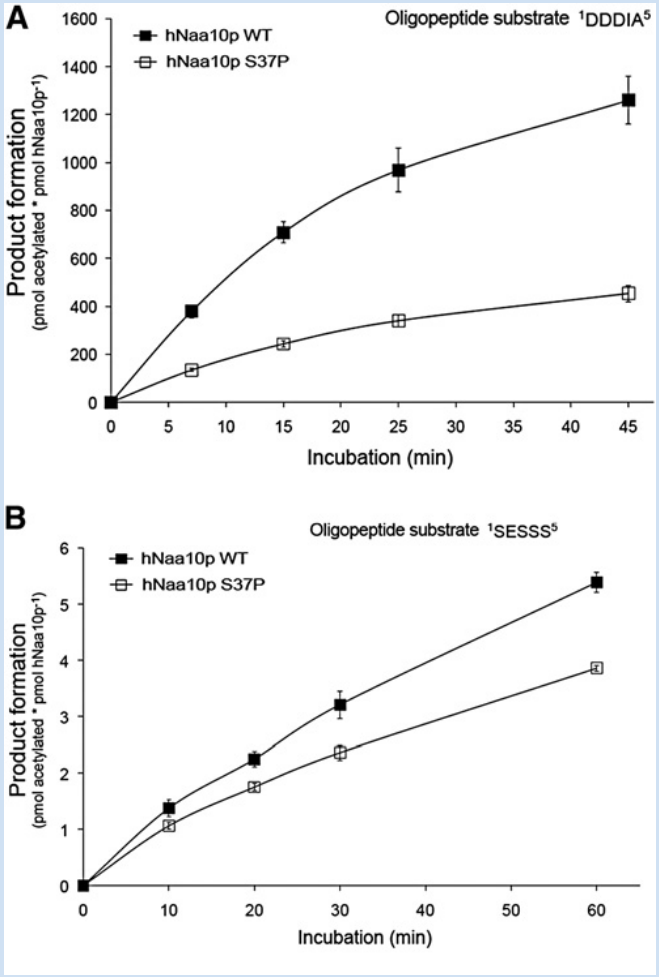


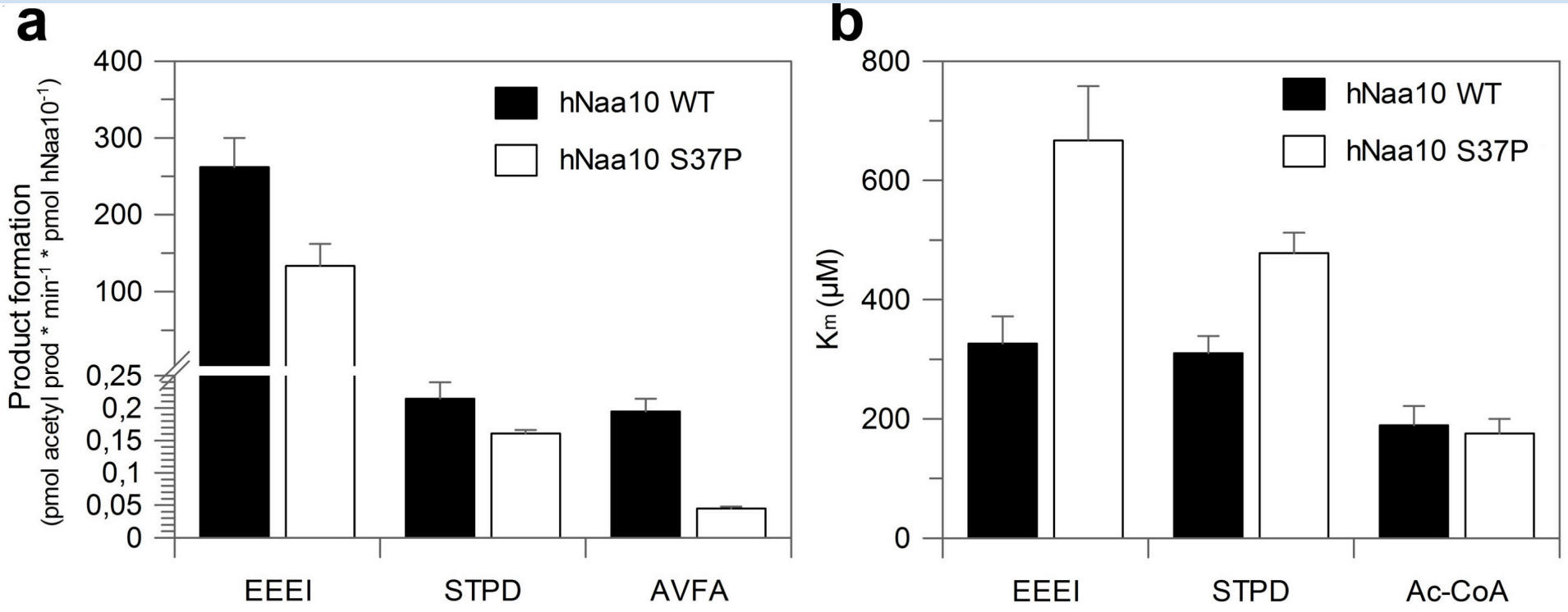

ribosomal subunit
small
large
Naa40
S
MetAP
Naa25
Naa20
MD
ME
MN
S
A
T
G
V
Naa10
Naa15
Naa50
ML
MK
MA
MM
MI
ML
MF
MW
Naa30
Naa35
Naa38
MA
MV
MK
MM
Naa60


NAC
MetAP
Naa15
SRP
Naa10
S
A
T
G
V
D
E
Naa10
Naa10

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







C

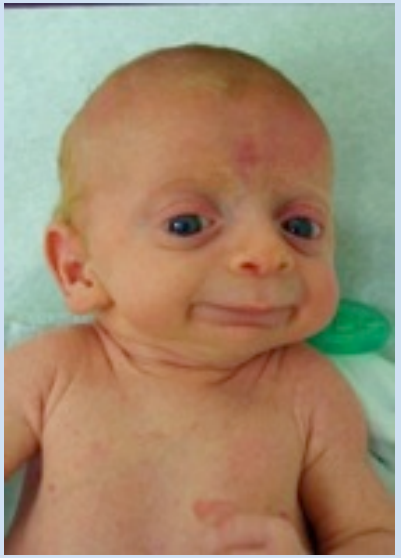
	EEEI		STPD		Ac-CoA	
	WT	S37P	WT	S37P	WT	S37P
Km (µM)	326,5 ± 46,4	667,0 ± 91,5	310,9 ± 28,9	478,3 ± 34,7	189,2 ± 31,8	175,6 ± 24,1
kcat/Km ($M^{-1}* s^{-1}$)	$13,1*10^3$ $\pm 2,0*10^3$	$3,2*10^3$ $\pm 0,5*10^3$	6,3 ± 0,5	1,8 ± 0,2	$28,6*10^3$ $\pm 5,3*10^3$	$15,4*10^3$ $\pm 2,2*10^3$

Manuscript accepted at Human Molecular Genetics

Big Questions though:

Simulated structure of S37P mutant

?



What is the molecular basis of Ogden syndrome?

- Naa10/Naa15 complex
- Naa10 localisation
- Naa10 function

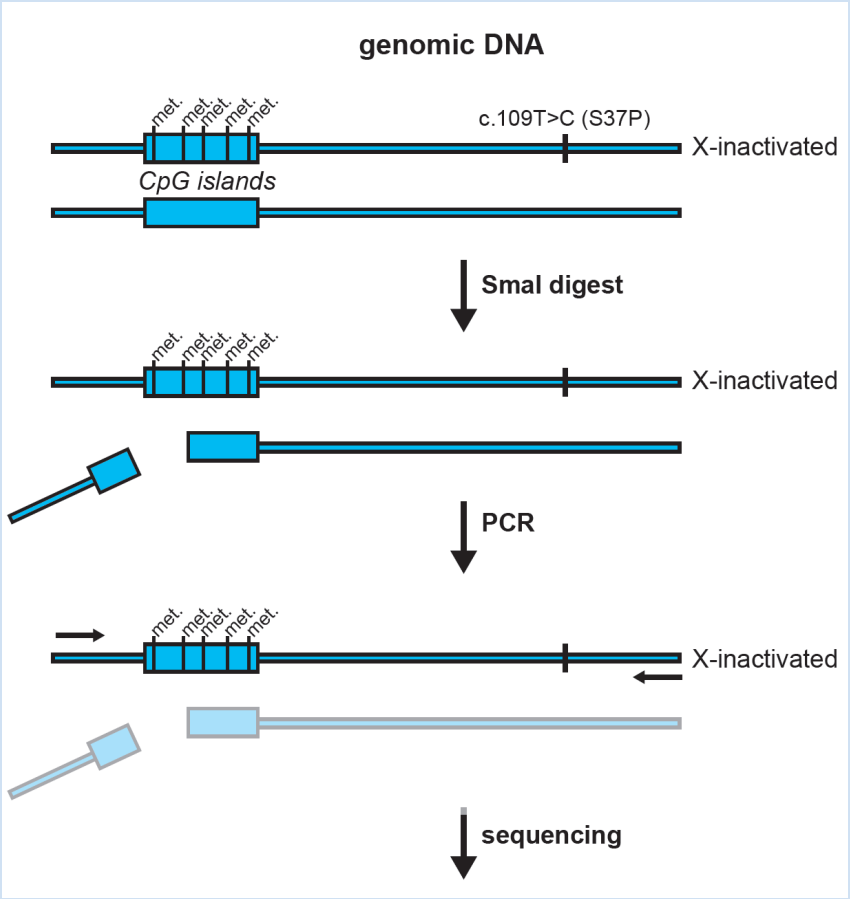
what can we learn from Ogden syndrome?

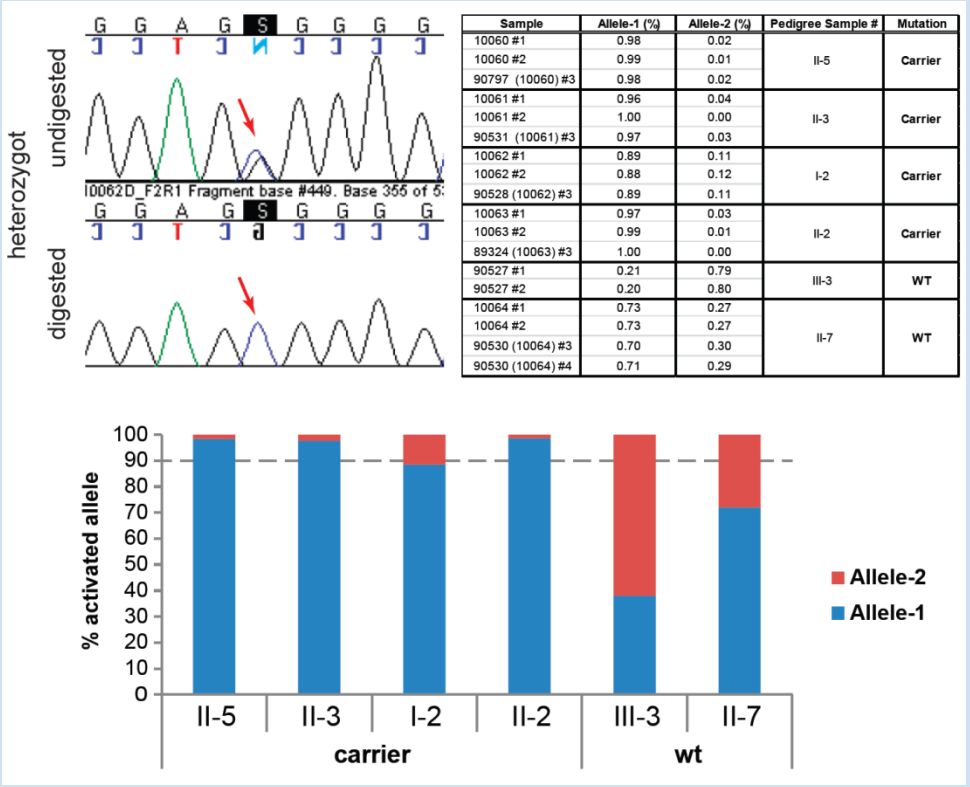
- characterizing different model systems (fibroblasts, iPS cells, yeast, mouse*)*

X-chromosome skewing

skewing analysis on B-cells

- from wt and S37P carriers females



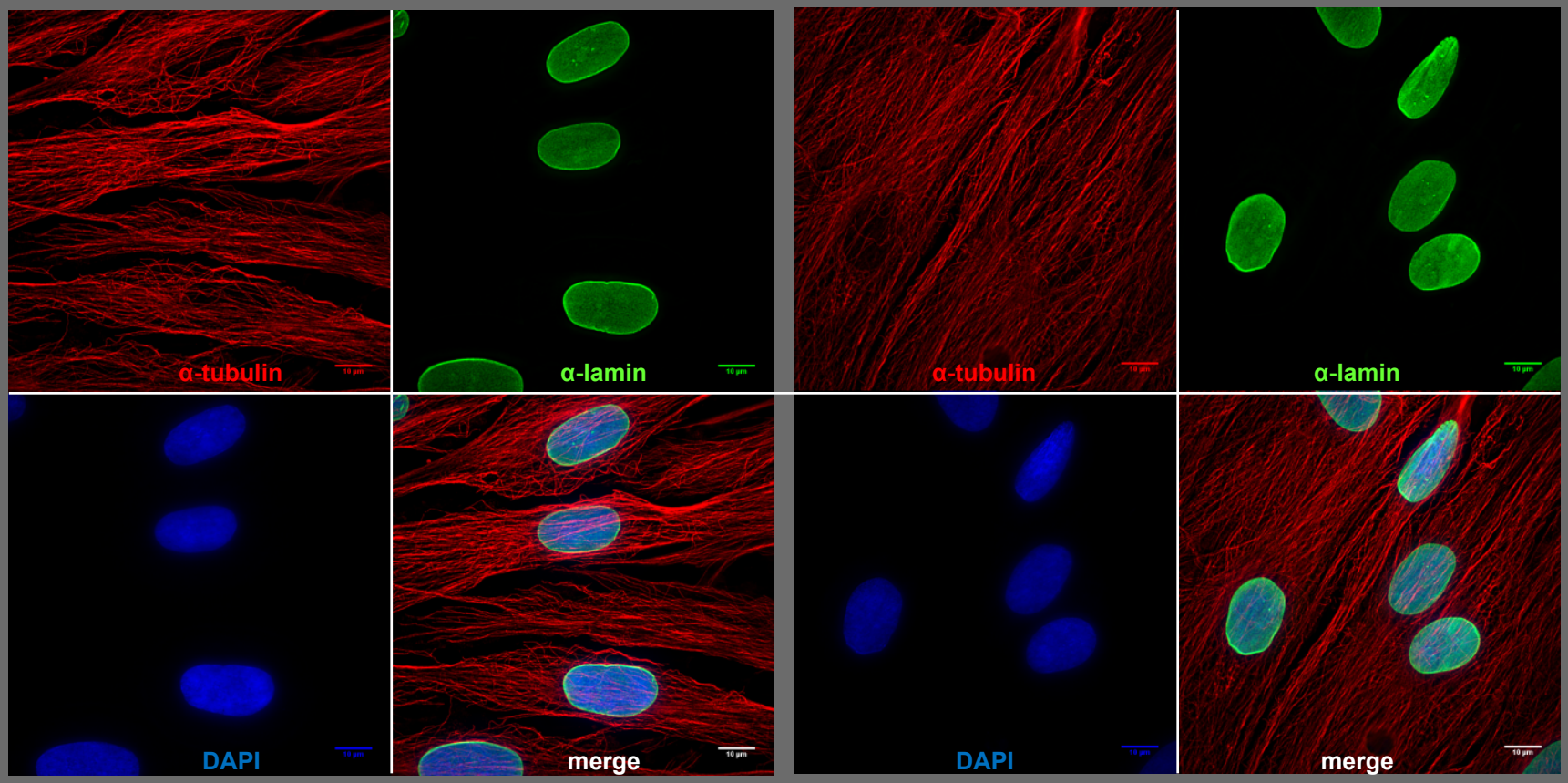


Sample	Allele-1 (%)	Allele-2 (%)	Pedigree Sample #	Mutation
10060 #1	0.98	0.02	II-5	Carrier
10060 #2	0.99	0.01		
90797 (10060) #3	0.98	0.02		
10061 #1	0.96	0.04	II-3	Carrier
10061 #2	1.00	0.00		
90531 (10061) #3	0.97	0.03		
10062 #1	0.89	0.11	I-2	Carrier
10062 #2	0.88	0.12		
90528 (10062) #3	0.89	0.11		
10063 #1	0.97	0.03	II-2	Carrier
10063 #2	0.99	0.01		
89324 (10063) #3	1.00	0.00		
90527 #1	0.21	0.79	III-3	WT
90527 #2	0.20	0.80		
10064 #1	0.73	0.27	II-7	WT
10064 #2	0.73	0.27		
90530 (10064) #3	0.70	0.30		
90530 (10064) #4	0.71	0.29		

Manuscript accepted at Human Molecular Genetics

Lamin and tubulin staining in primary fibroblasts

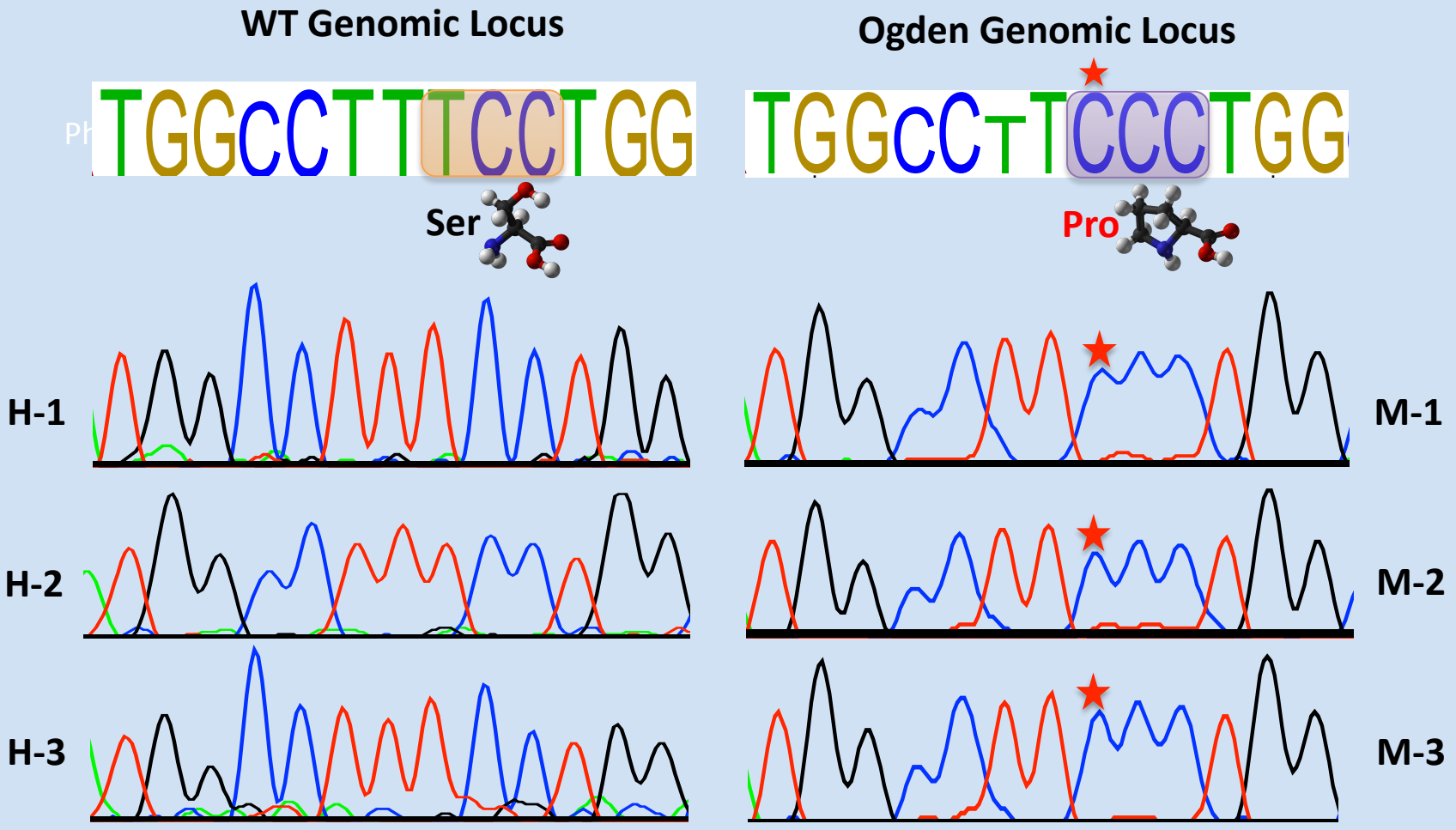
- staining for lamin and tubulin
- pictures were taken on Applied Precision DeltaVision (wide-field fluorescence microscope with deconvolution)



control boy (BJ)

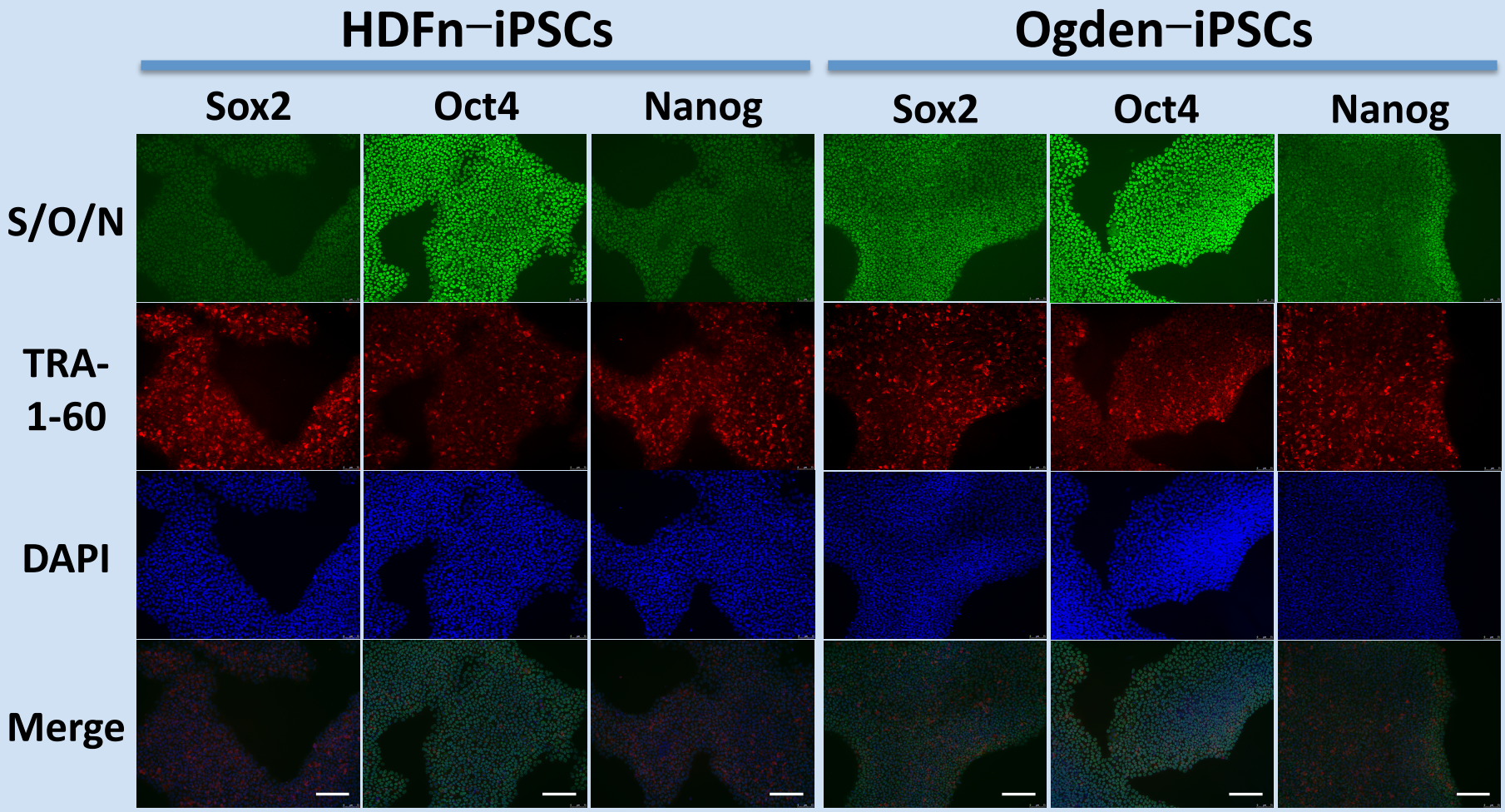
Ogden boy (MC)

Naa10 S37P Mutation Validation in iPSCs



Sanger sequencing results of genomic DNA from WT HDFn-iPSCs (left) and Ogden MC-iPSCs (right). The c.109T>C mutation in exon 2 of *Naa10* is indicated by the red asterisk, and the codon change from Serine to Proline is illustrated respectively.

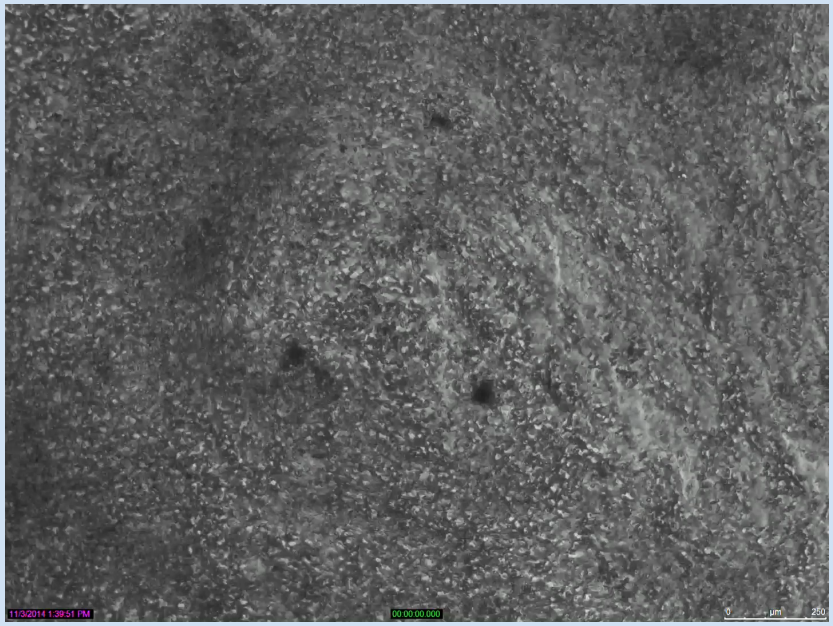
Characterization of iPSCs



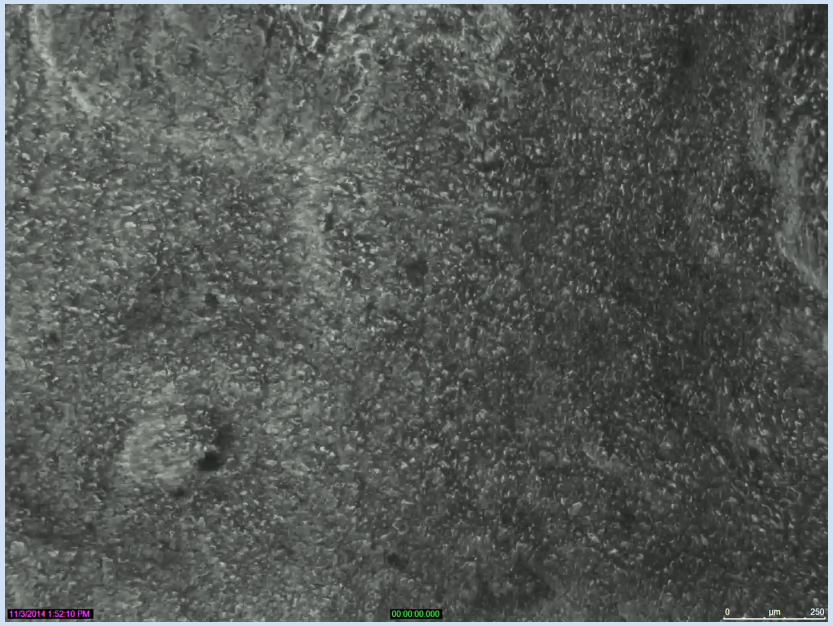
Immunofluorescence analysis of the indicated pluripotent markers in HDFn-iPSCs (left) and Ogden-iPSCs (right). Nuclei were visualized with DAPI stain (blue). Scale bar, 150 µm.

Cardiac Lineage Differentiation

WT iCM on Day 41

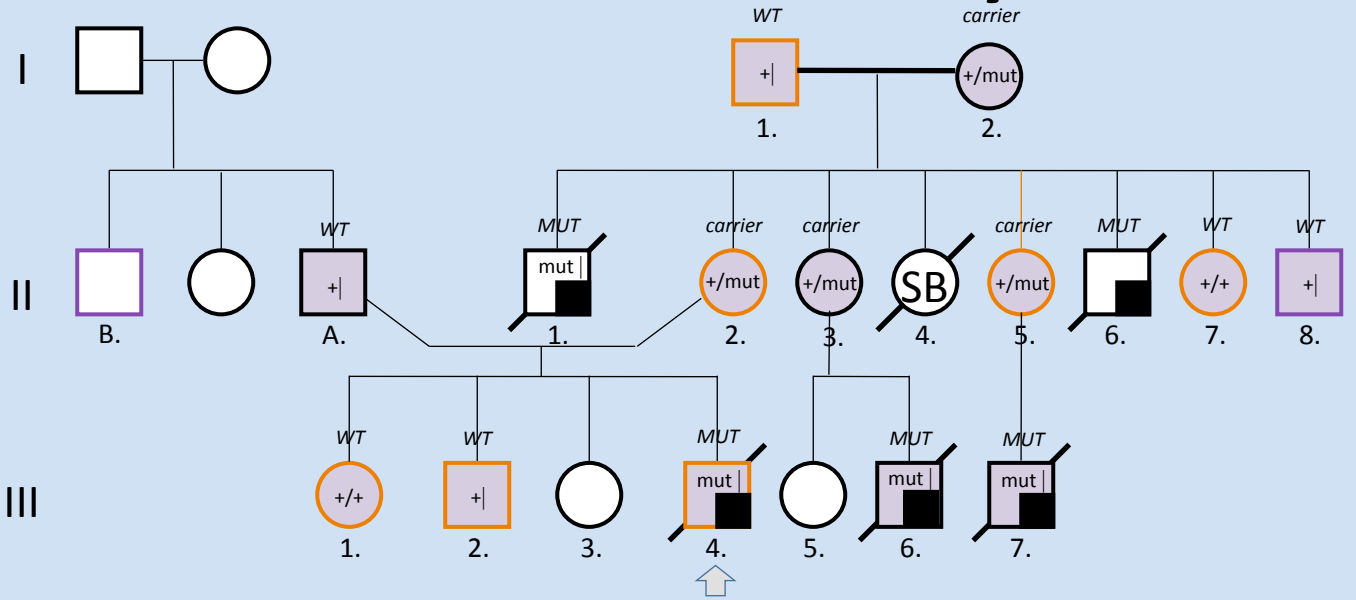


Ogden iCM on Day 38



Initial cardiac differentiation assay showed wave-like beating sheets of cells derived from WT and Ogden iPSCs when plated on matrigel on Day 41 and Day 38, respectively. Scale bar, 250 µm.

Proteomics Analysis of EBV-transformed cell lines and fibroblasts from family members



= male ♂

= female ♀

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

SB = stillborn

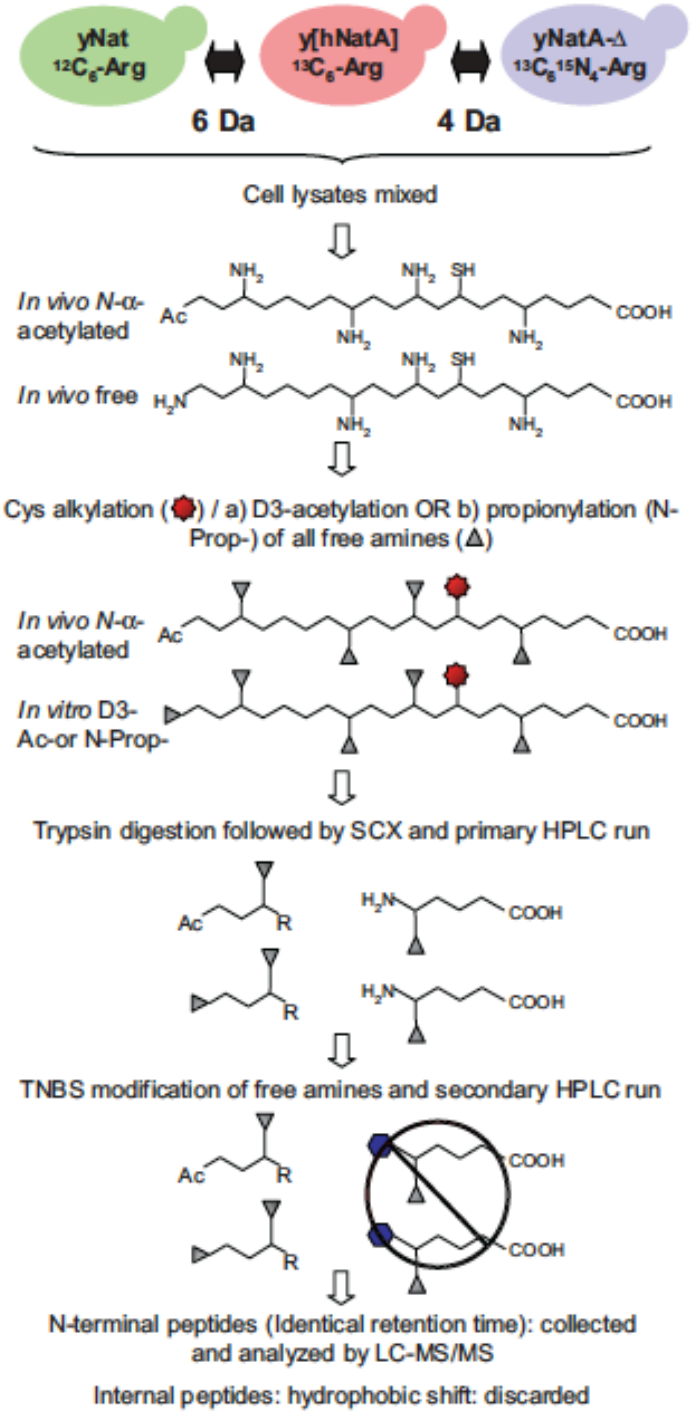
= proband

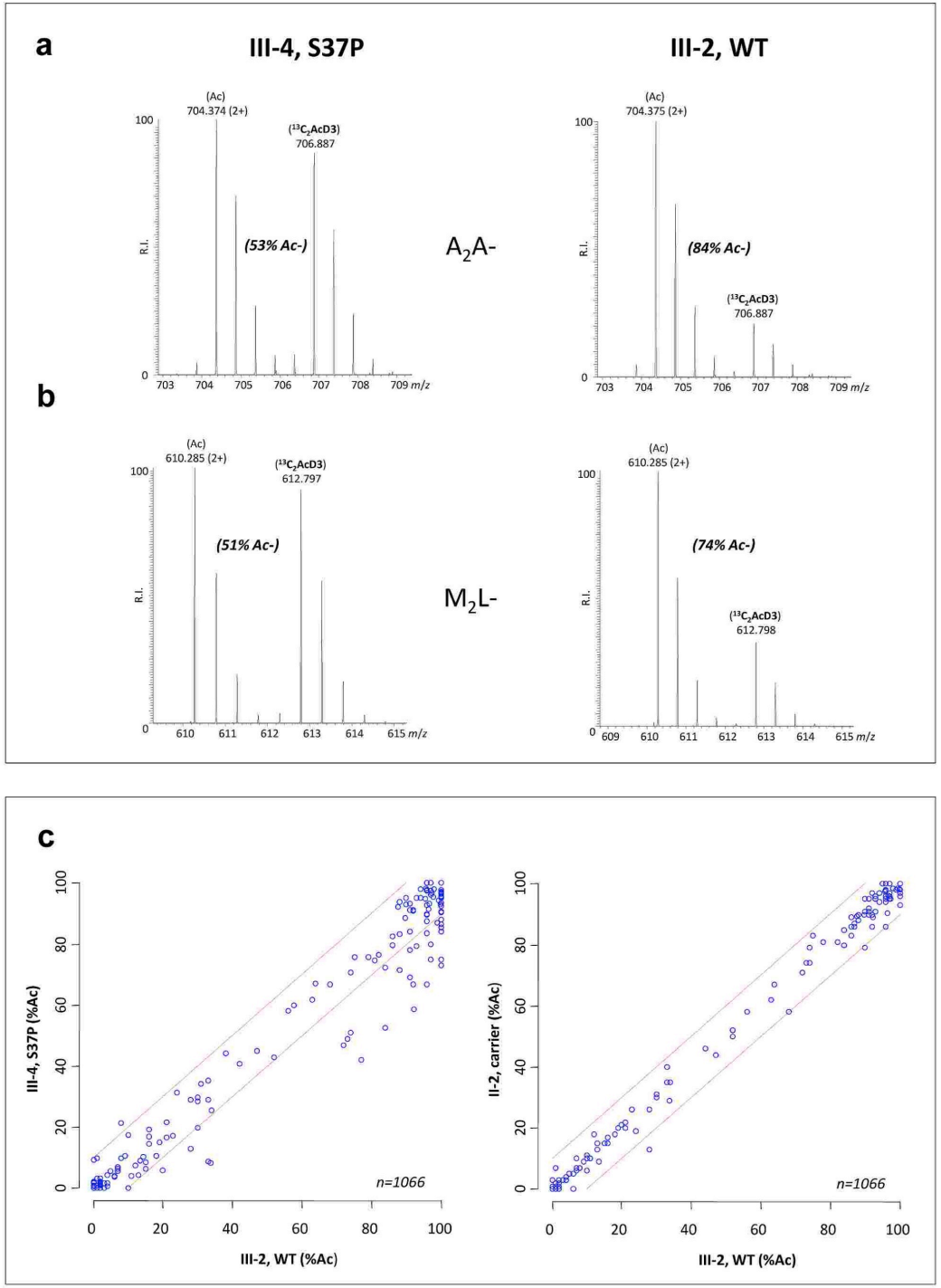
= patient samples analyzed by N-terminal COFRADIC analyses **(#1 to #5)**

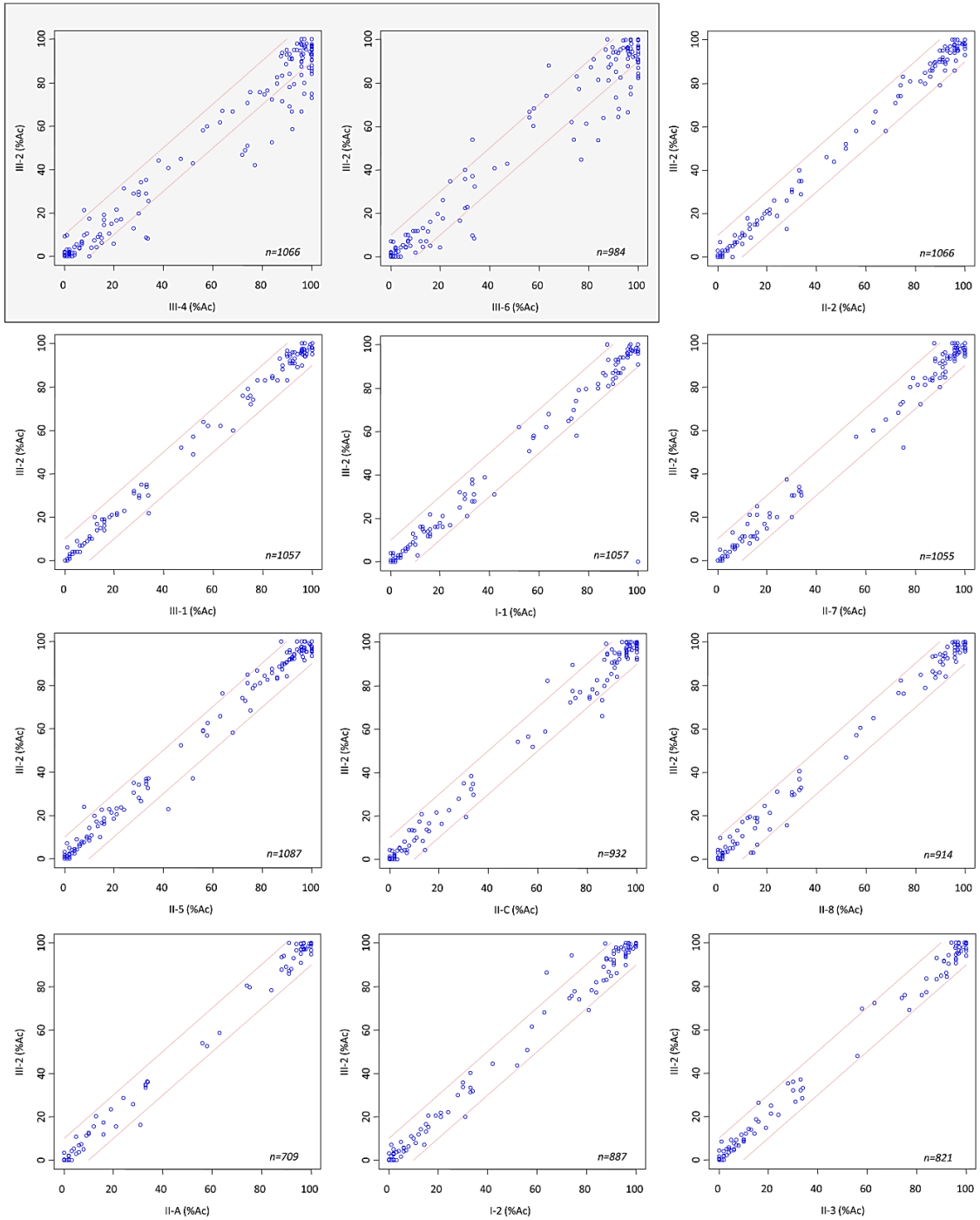
= patient samples prepared for N-terminal COFRADIC analyses (but still to be analyzed) **(#8 and #9)**

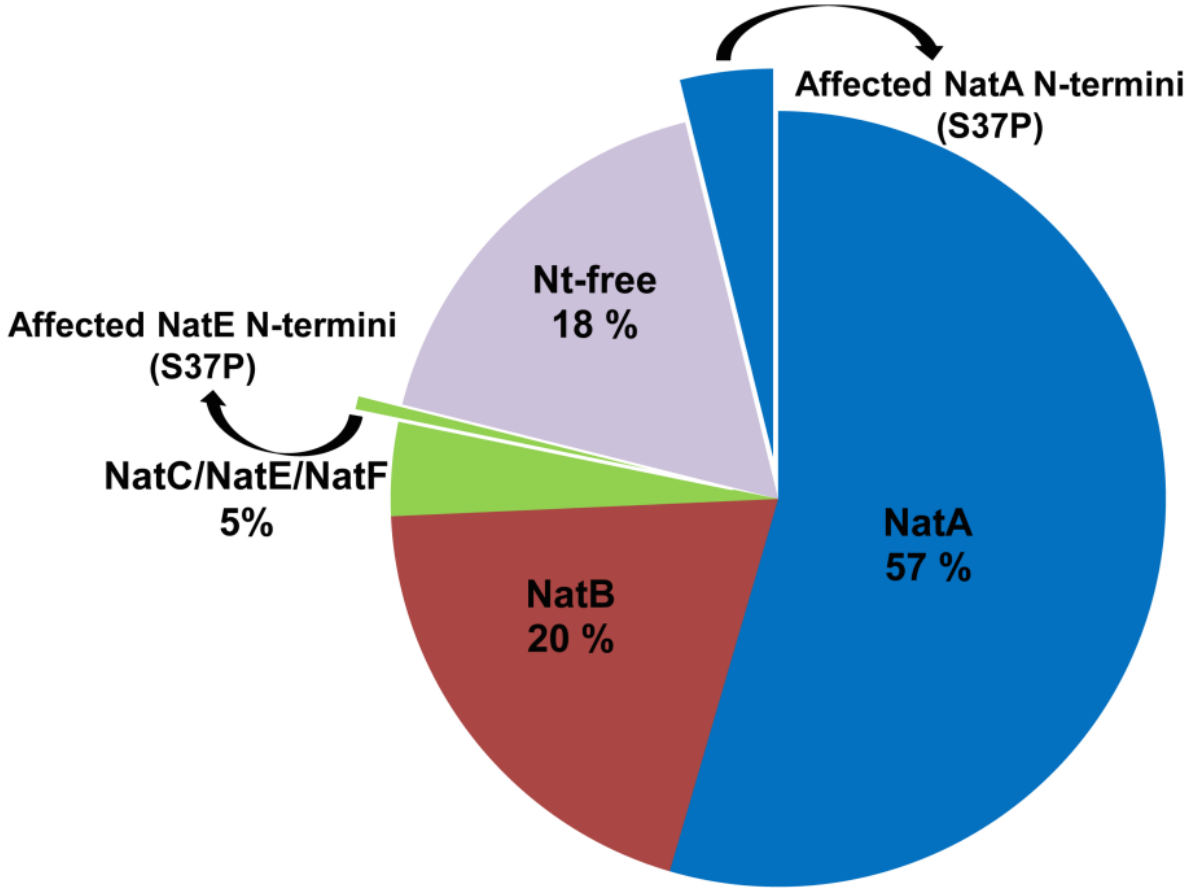
III.4. proband hemizygous, mutant (89323) **(#1a)** (#1b)
II.2. mother of proband, carrier (89324) **(#2)**
II.A. married-in father of proband, WT(89325)
III.2. brother of proband, WT(90526) **(#3)**
III.1. sister of proband, WT (90527) **(#4)**
I.2. grandmother of proband, carrier (90528)
I.1. married-in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

Proteomics Strategy With Thomas Arnesen, Petra van Damme And Kris Gevaert




a
III-4, S37P
III-2, WT
(Ac)
704.374 (2+)
(13C2AcD3)
706.887
(53% Ac-)
(Ac)
704.375 (2+)
(84% Ac-)
(13C2AcD3)
706.887
A2A-
b
(Ac)
610.285 (2+)
(13C2AcD3)
612.797
(51% Ac-)
(Ac)
610.285 (2+)
(74% Ac-)
(13C2AcD3)
612.798
M2L-
R.I.
m/z
c
III-4, S37P (%Ac)
III-2, WT (%Ac)
n=1066
II-2, carrier (%Ac)
III-2, WT (%Ac)
n=1066


III-2 (%Ac)
III-4 (%Ac)
n=1066
III-6 (%Ac)
n=984
II-2 (%Ac)
n=1066
III-1 (%Ac)
n=1057
I-1 (%Ac)
n=1057
II-7 (%Ac)
n=1055
II-5 (%Ac)
n=1087
II-C (%Ac)
n=932
II-8 (%Ac)
n=914
II-A (%Ac)
n=709
I-2 (%Ac)
n=887
II-3 (%Ac)
n=821

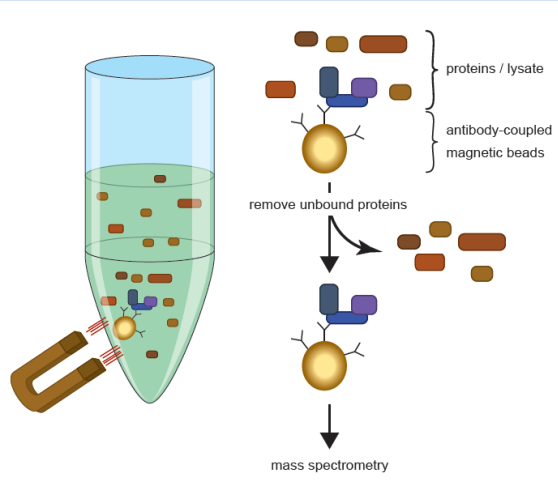


Manuscript accepted at Human Molecular Genetics

Table 1: Overview of N-termini less acetylated in Naa10-S37P B-cells, fibroblasts and siNatA HeLa cells.

NAT type	P1	P1'	P2'	significant B-cells	significant fibroblasts	siNatA (HeLa)	Description
NatA	M	V	N	√	√	√	**Peptidyl-prolyl cis-trans isomerase A**
NatA	M	A	A	√	√	√	**Translational activator GCN1**
NatA	M	A	A	√	√	√	**Transcription elongation factor B polypeptide 3**
NatA	M	A	V	√	√	√	**Ribonuclease P protein subunit p30**
NatA	M	G	A	√	√		**THO complex subunit 7 homolog**
NatA	M	S	A	√	√		**Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit DAD1**
NatA	M	T	M	√	√		**14-3-3 protein beta/alpha**
NatA	M	A	G	√	√		**39S ribosomal protein L15, mitochondrial**
NatA	M	A	A	√	√		**E3 ubiquitin-protein ligase RNF5**
NatA	M	T	K	√	√		**Leucine-rich repeat-containing protein 59**
NatA	M	A	V	√	√		**Ras GTPase-activating protein 3**
other	-	M	V	√	√		**Peptidyl-prolyl cis-trans isomerase A**
other	-	M	V	√	√		**SUMO-activating enzyme subunit 1**
NatA	M	A	E	√		√	**60S ribosomal protein L13a**
NatA	M	V	E	√		√	**SUMO-activating enzyme subunit 1**
NatA	M	A	L	√			**26S protease regulatory subunit 8**
NatA	M	A	Q	√			**Serum response factor-binding protein 1**
NatA	M	V	E	√			**Protein LCHN**
NatA	M	S	G	√			**Transmembrane protein 50A**
NatA	M	T	A	√			**Transmembrane protein 85**
NatC or other	-	M	L	√			**Kinesin-like protein KIF21A**
NatC or other	-	M	L	√			**p53 and DNA damage-regulated protein 1**
other	-	M	V	√			**Deoxyhypusine hydroxylase**
other	-	M	M	√			**Uncharacterized protein C11orf46**
NatA	M	A	A	√	√		**Epidermal growth factor receptor substrate 15**
NatA	M	S	T		√		**Mediator of RNA polymerase II transcription subunit 30**
NatA	M	A	A		√		**AN1-type zinc finger protein 5**
NatA	M	G	A		√		RNA-binding protein 7

IP and mass spec of NAA10



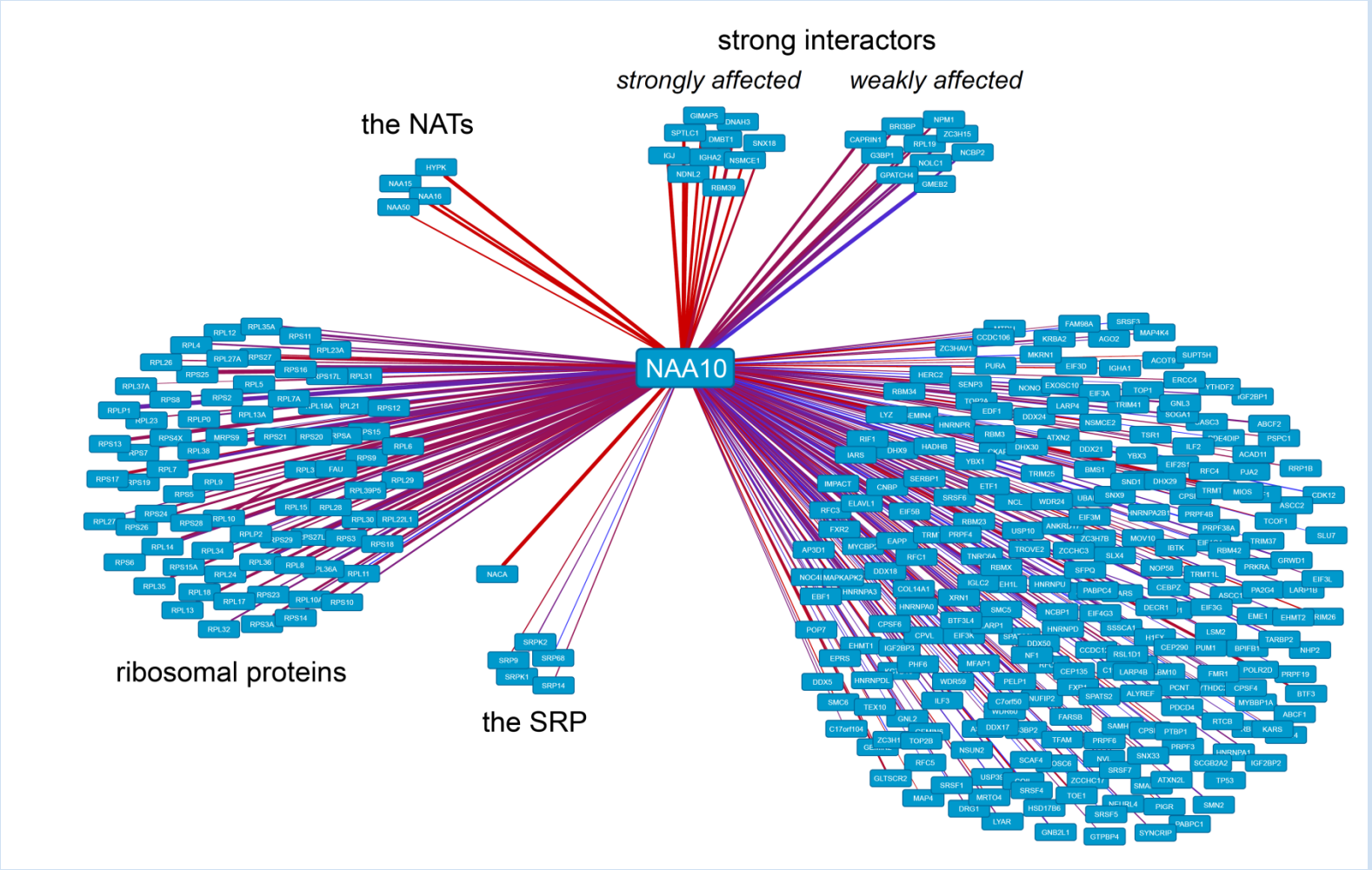
5 x 106 HEK293 cells were seeded in 10 cm dish and transfected after 24 h with pcDNA3.1 V5/His hNaa10 wt, pcDNA3.1 V5/His hNaa10 S37P or corresponding empty vector. Cells were lysed after 48 h .

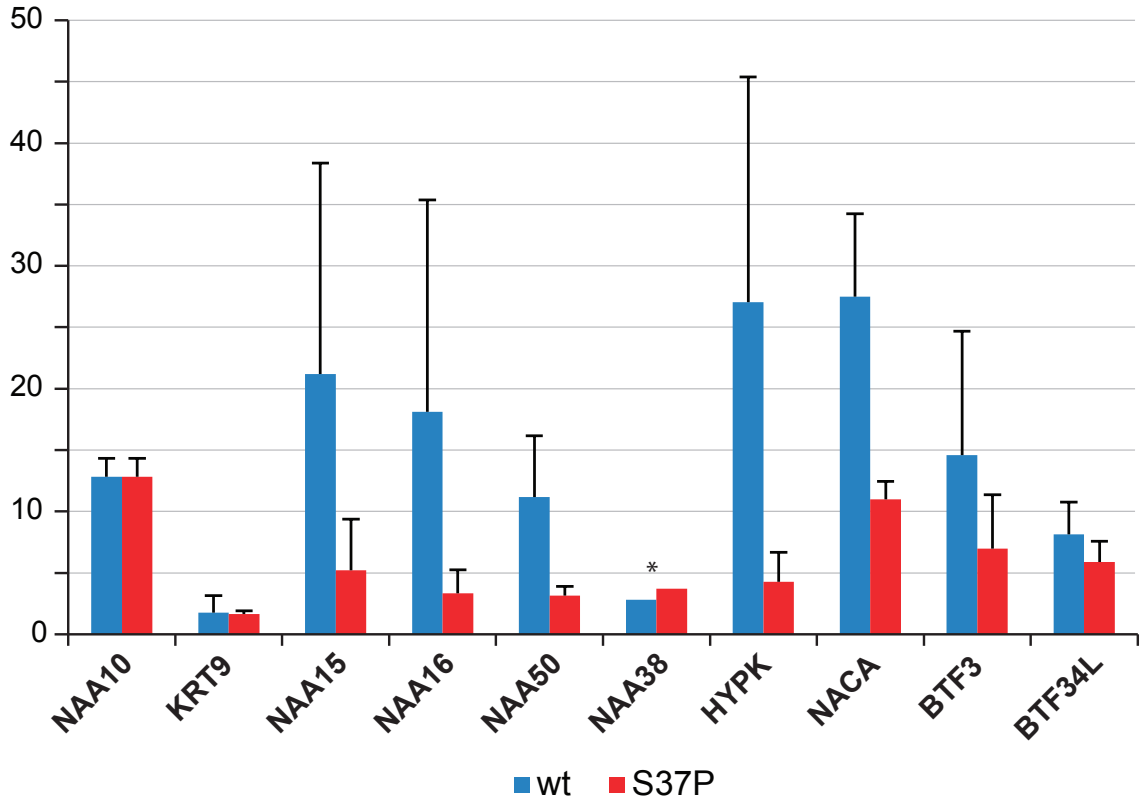
For IP, 40 mg total protein were incubated with 400 µl anti-V5-coupled magnetic beads (Invitrogen) for 2 h at 4°C under constant agitation.

Proteins were digested with trypsin off the beads and labelled with distinct isobaric iTRAQ reagents. The samples were combined and subjected to standard 2D MudPIT LCMS and analyzed using a Thermo Velos Orbitrap mass spectrometer.

The relative enrichment in the samples were calculated as a ratio of the intensities between the samples and the empty-vector control. The whole experiment was done twice.

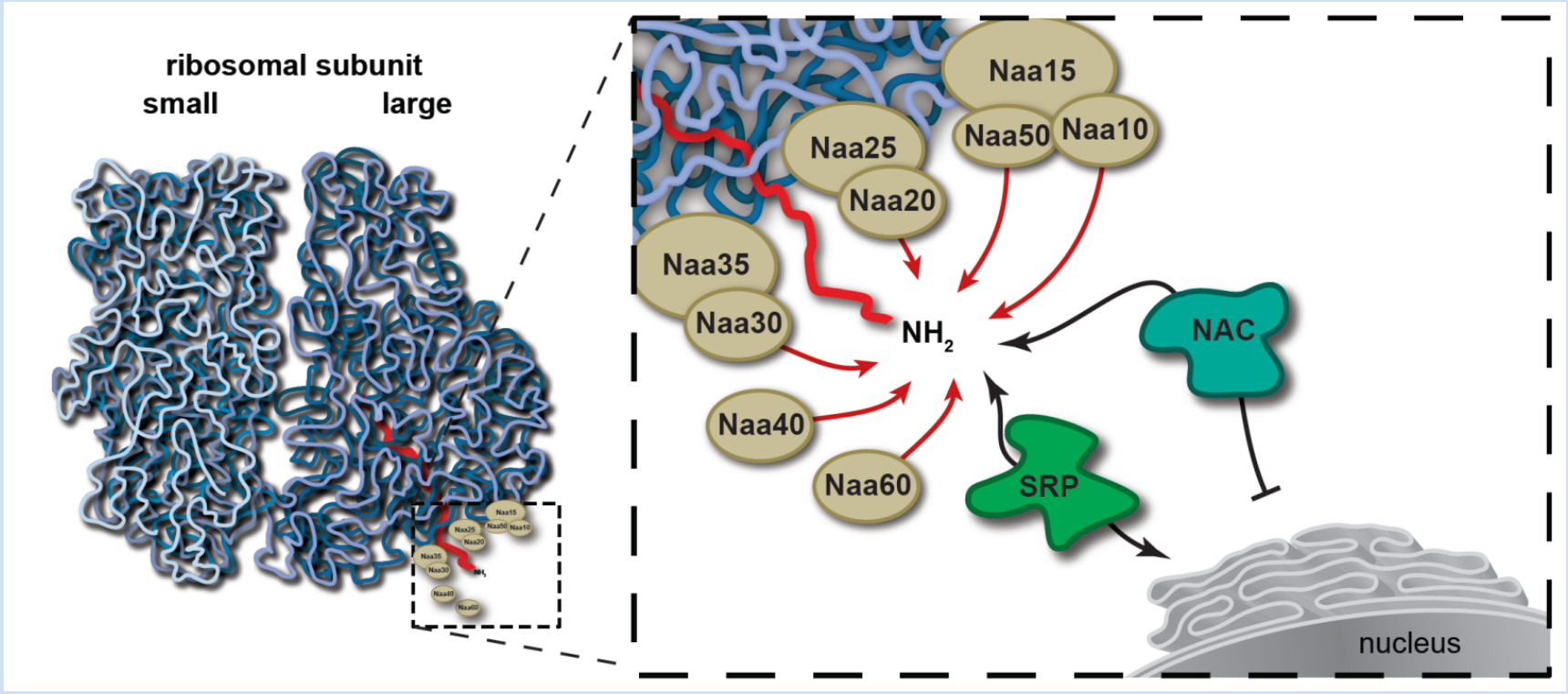
slides IP & mass spec

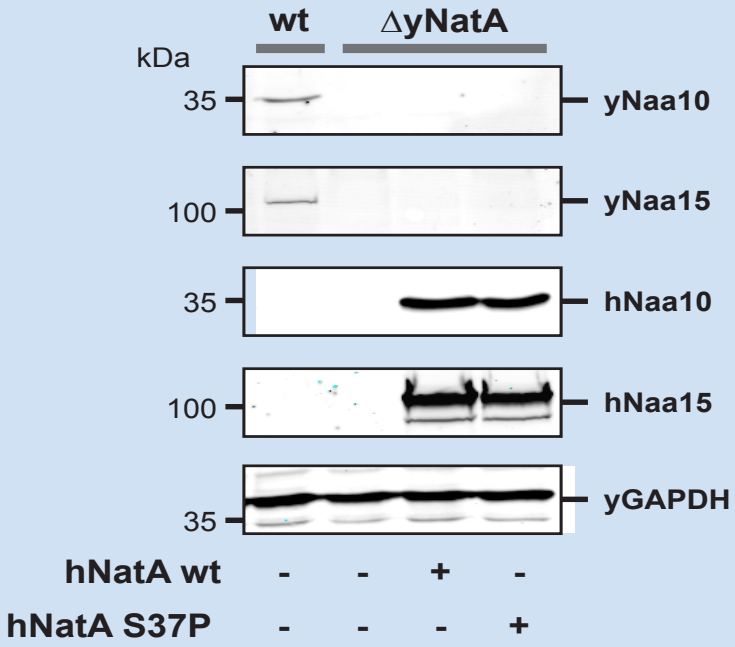


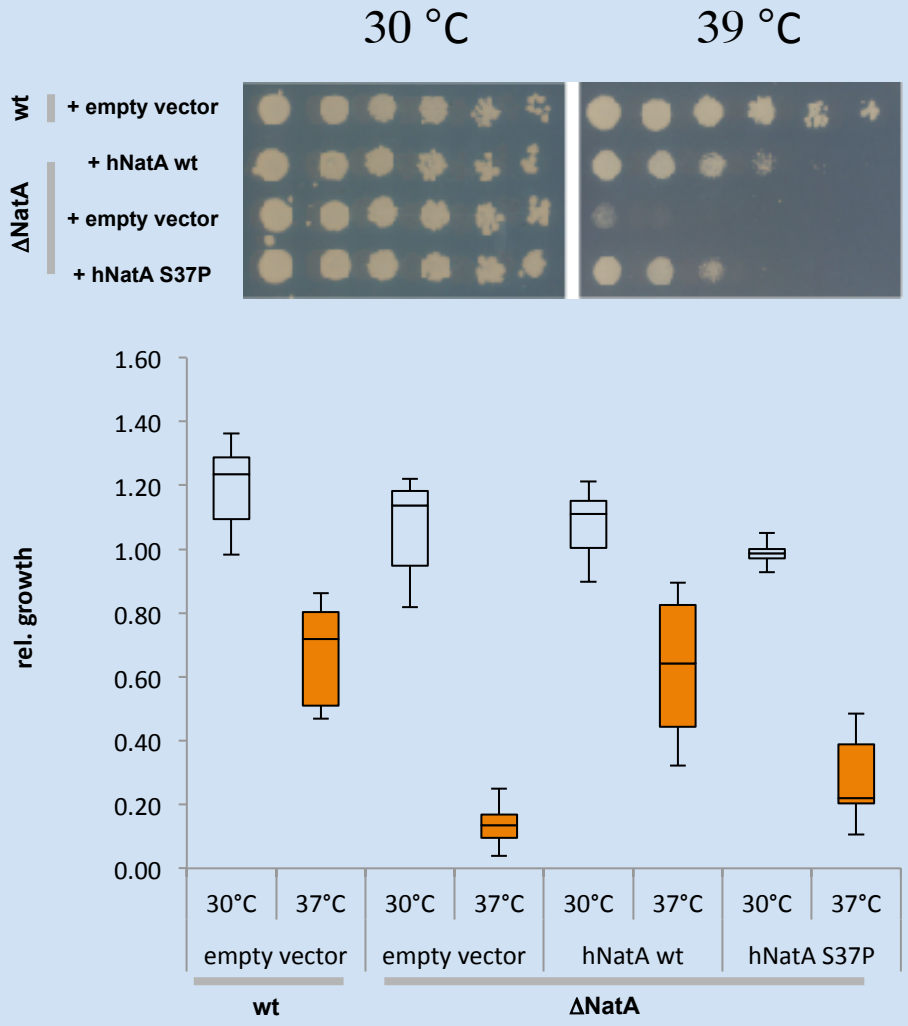


* Naa38 was only found in one set

NAC


ribosomal subunit
small
large
Naa15
Naa25
Naa50
Naa10
Naa20
Naa35
Naa30
NH2
NAC
Naa40
Naa60
SRP
nucleus

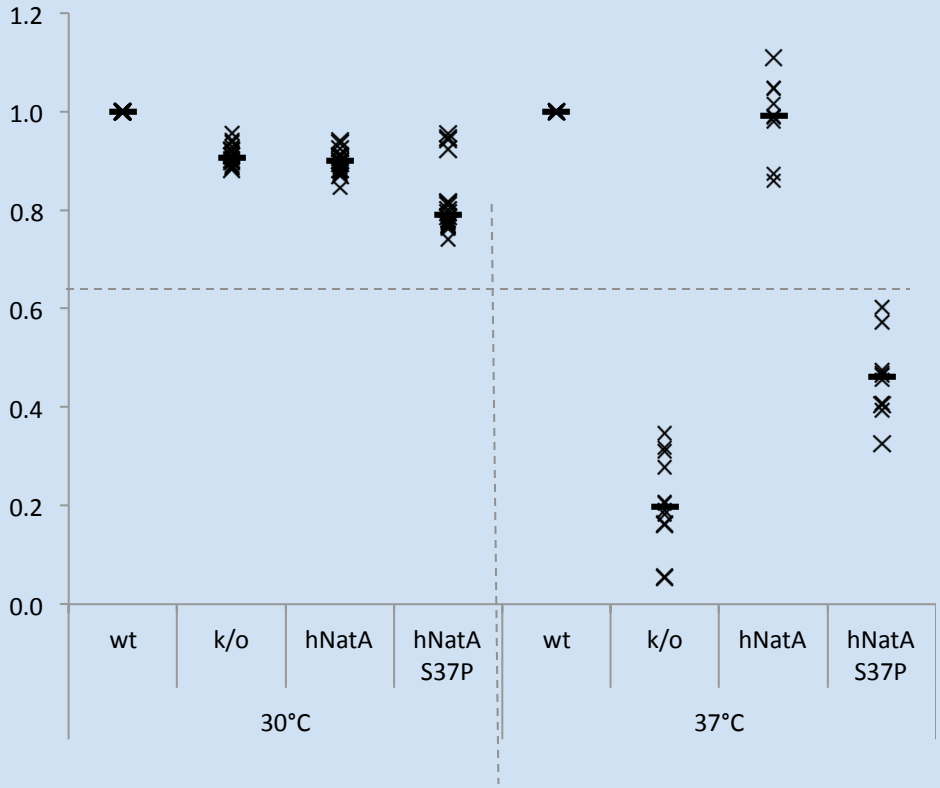

wt
ΔyNatA
kDa
35
yNaa10
100
yNaa15
35
hNaa10
100
hNaa15
35
yGAPDH
hNatA wt - - + -
hNatA S37P - - - +


30 °C
39 °C
wt
+ empty vector
ΔNatA
+ hNatA wt
+ empty vector
+ hNatA S37P
rel. growth
1.60
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
30°C
37°C
30°C
37°C
30°C
37°C
30°C
37°C
empty vector
empty vector
hNatA wt
hNatA S37P
wt
ΔNatA

Same data, just presented differently.

yeast growth
YPDA media (Clonetech, #630464)
Yeast minimal SD base (Clonetech, #630411) supplemented with drop out mix –Ura (Clonetech, #630416)

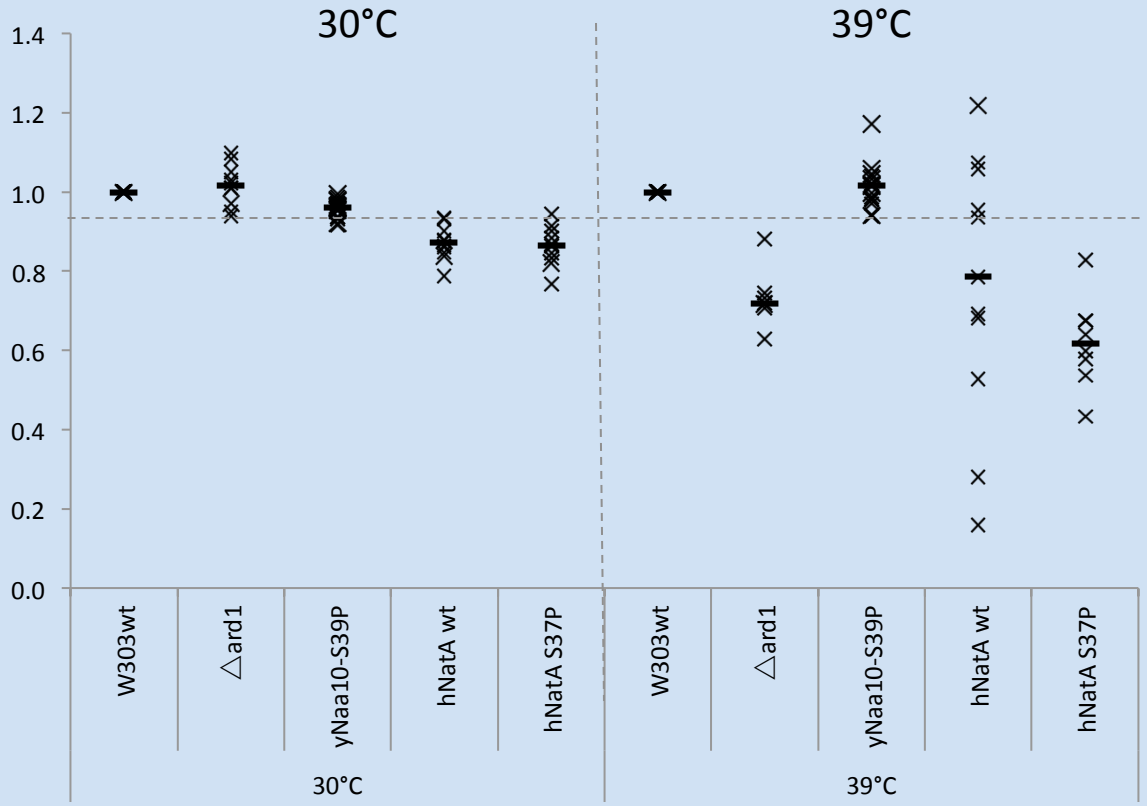
A 5 ml overnight culture was grown in SD^{-URA} at 30°C. Cells were diluted to an OD_{600} of 0.1 and either spotted in 1:5 serial dilutions on plates for 48 h (upper panel) or grown in 2 ml cultures at 30°C or 39°C under constant agitation for 24 h (lower panel). Optical density was plotted
n=11



Endogenous, single-copy genes in yeast.

Optical density as a measure of growth was normalized to the W303 wt strain for every independent experiment and plotted (X). The median of all experiments is shown as a short line

n=22 for S39P
n=11 for all other strains



Conclusions

- Expanding Utah Genome Project significantly.
- Making good progress on many new rare, orphan diseases.
- Working toward highly accurate whole genome sequencing.
- Elaborating the mechanistic basis of Ogden Syndrome in molecular detail.

The End