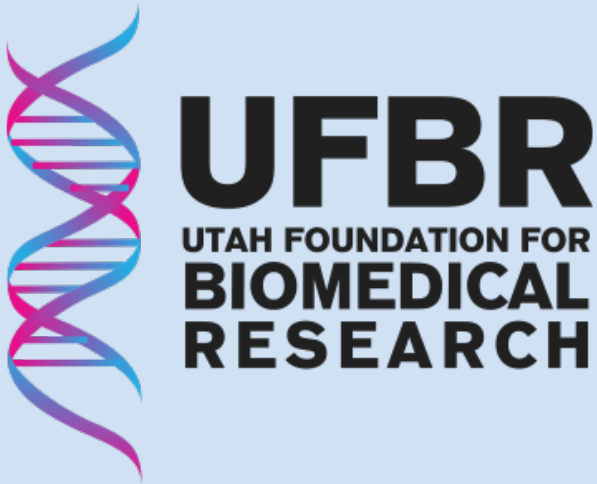
Increasing Accuracy for Exome and Whole Genome Sequencing

Gholson Lyon, M.D. Ph.D.



Acknowledgments

CSH STANLEY INSTITUTE FOR COGNITIVE GENOMICS
COLD SPRING HARBOR LABORATORY

Jason O'Rawe
Yiyang Wu
Han Fang
Michael Schatz
Giuseppe Narzisi

VBI VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

David Mittelman
Gareth Highman

our study families

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*]

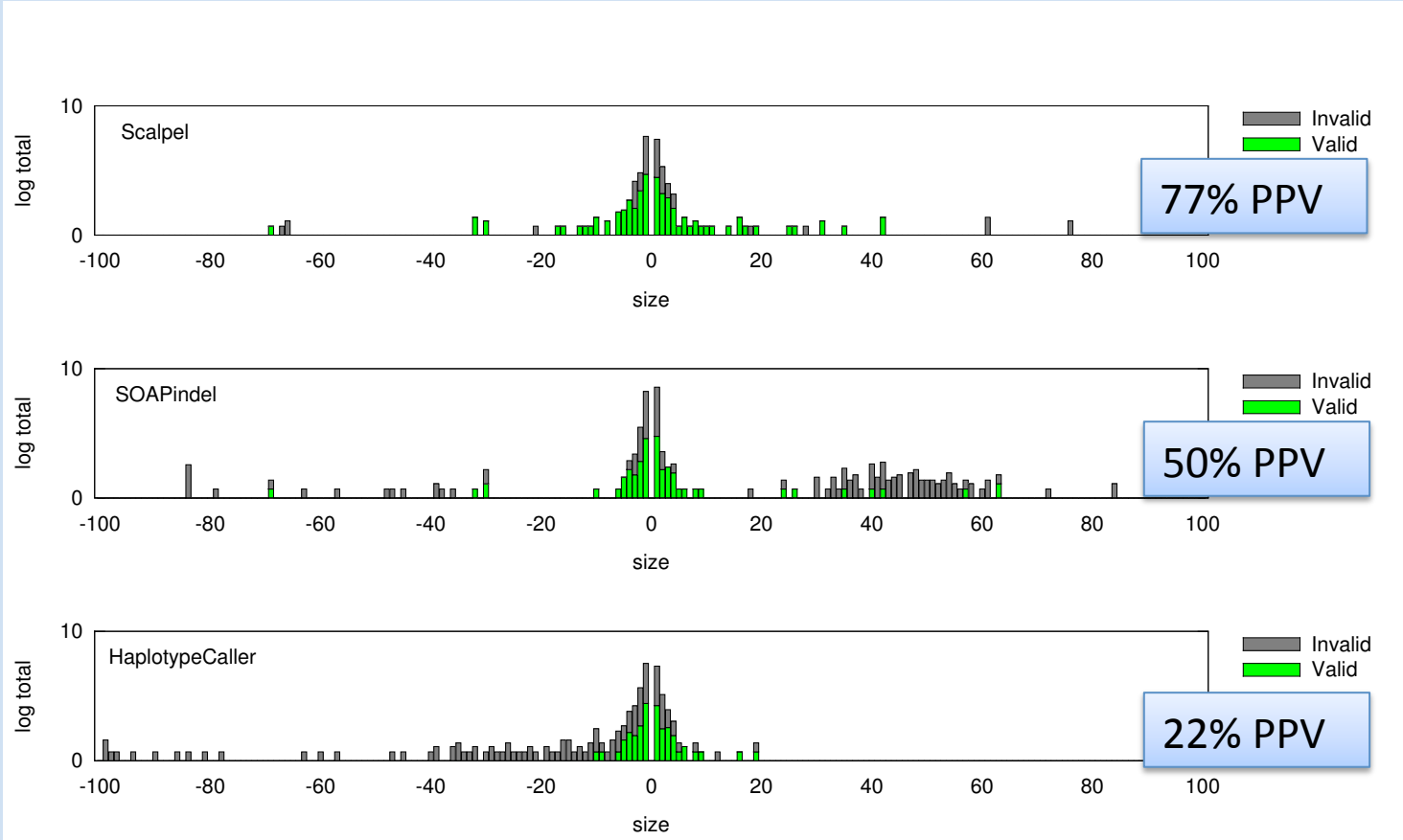
SCALPEL

Micro-Assembly Approach to detect INDELs within Exome-Capture data

Jason O'Rawe
Yiyang Wu
Giuseppe Narzisi, PhD
Michael Schatz, PhD

CSH Cold Spring Harbor Laboratory

1000 INDELs for Experimental Validation



Tool	Valid (all)	Invalid (all)	PPV (all)	Valid (≥30bp)	Invalid (≥30bp)	PPV(%) (≥30bp)
Scalpel	145	43	77.1	13	1	92.8
SOAPindel	101	99	50.5	8	129	5.8
HaplotypeCaller	45	155	22.5	7	62	11.3

Variant Analysis Pipeline

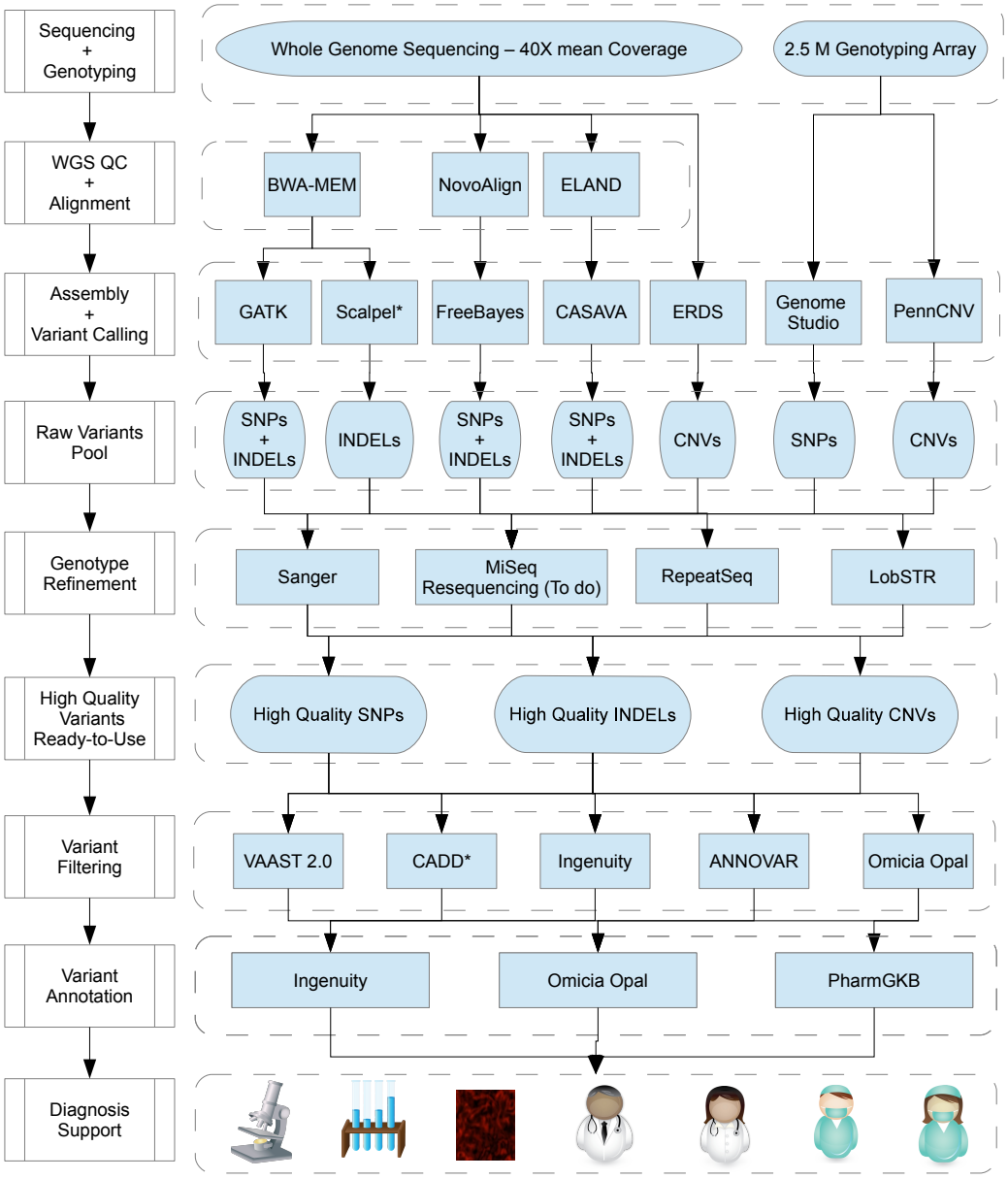
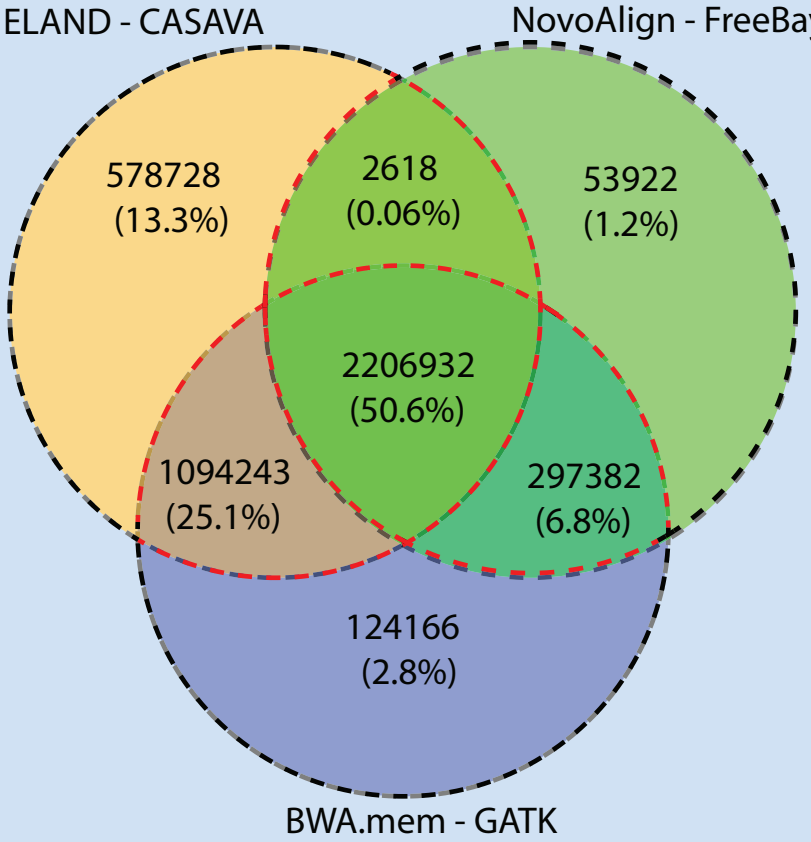


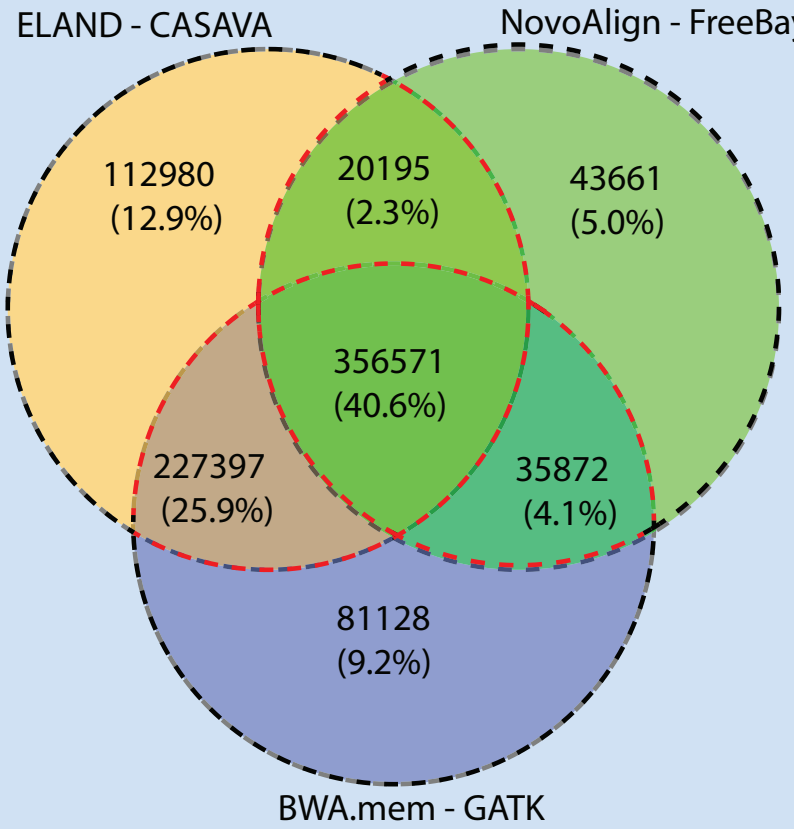
Figure 2. Flow chart of our variant analysis pipeline.
* Both Scalpel and CADD are still inpress. For CADD, see http://cadd.gs.washington.edu/

Whole Genome variant calling on 2x100 bp paired end reads- HiSeq2000

SNV calling

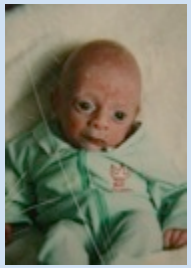


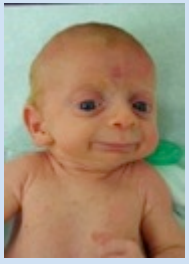
Indelscalling

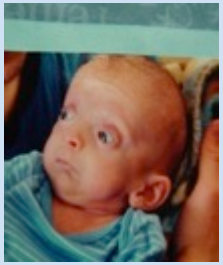


If we focus on variants called by at least two pipelines, i.e. the central region surrounded by red dash lines, we can reduce algorithm-induced error and achieve a significant higher power, 82.6% for SNPs and 72.9% for INDELs, respectively.

Vignette #1: Variable expressivity in any disease, including in this one: Ogden Syndrome in Ogden, Utah.





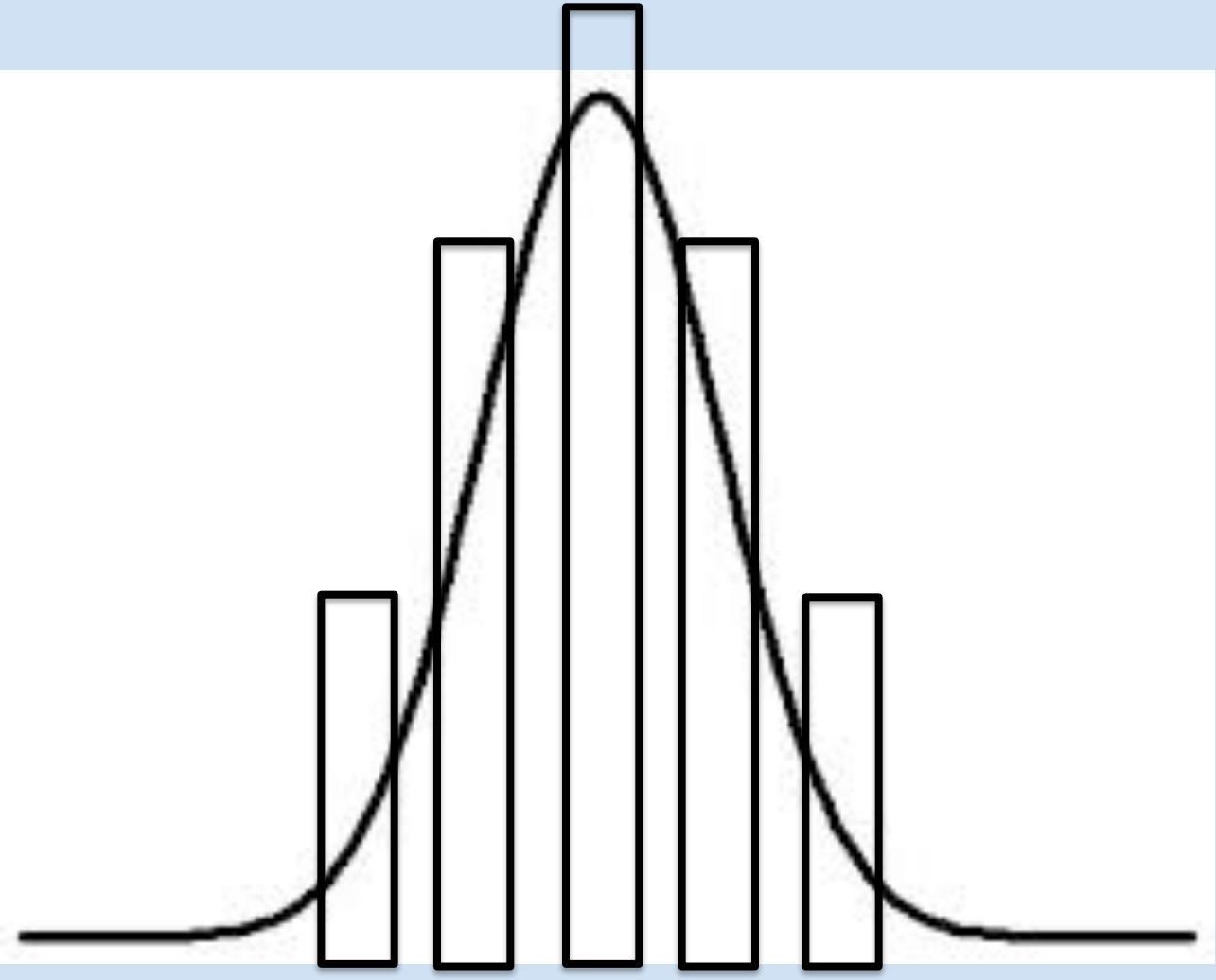


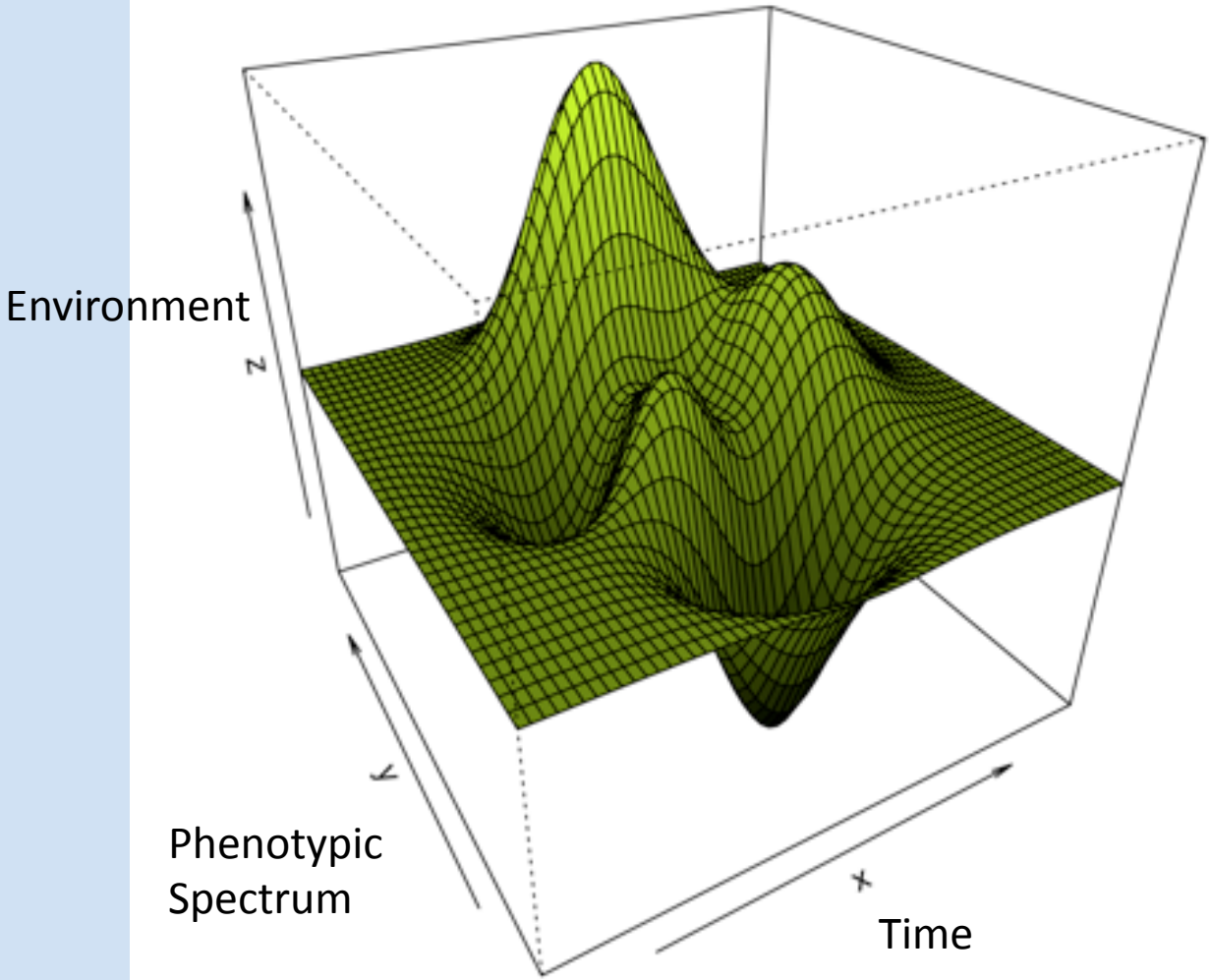
Am J Hum Genet. 2011 Jul 15;89(1):28-43.

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.	

Categorical Thinking Misses Complexity





A conceptual model of genotype-phenotype correlations. The *y* plane represents a phenotypic spectrum, the *x* plane represents the canalized progression of development through time, and the *z* plane represents environmental fluctuations.

bioRxiv beta

THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOU

Search

1 Results for term "gholson"

Results/page 10 **Order by** Best Match

Clinical genetics of neurodevelopmental disorders

Gholson J Lyon, Jason O'Rawe

bioRxiv doi: 10.1101/000687

New Results

> ...as described at http://creativecommons.org/licenses/by/3.0/ Clinical genetics of neurodevelopmental disorders **Gholson** J Lyon 1 3 glyon@cshl.edu , http://lyonlab.cshl.edu/ Jason O'Rawe 2 jazon33y@gmail.com * Corresponding author...

Expression Issues

- We do not really know the expression of pretty much ALL mutations in **humans**, as we have not systematically sequenced or karyotyped any genetic alteration in **Thousands to Millions** of **randomly** selected people, nor categorized into ethnic classes, i.e. clans.
- Complexity, or “The False Negative Problem”

Summary from Vignette #1

Genotype ≠ Phenotype

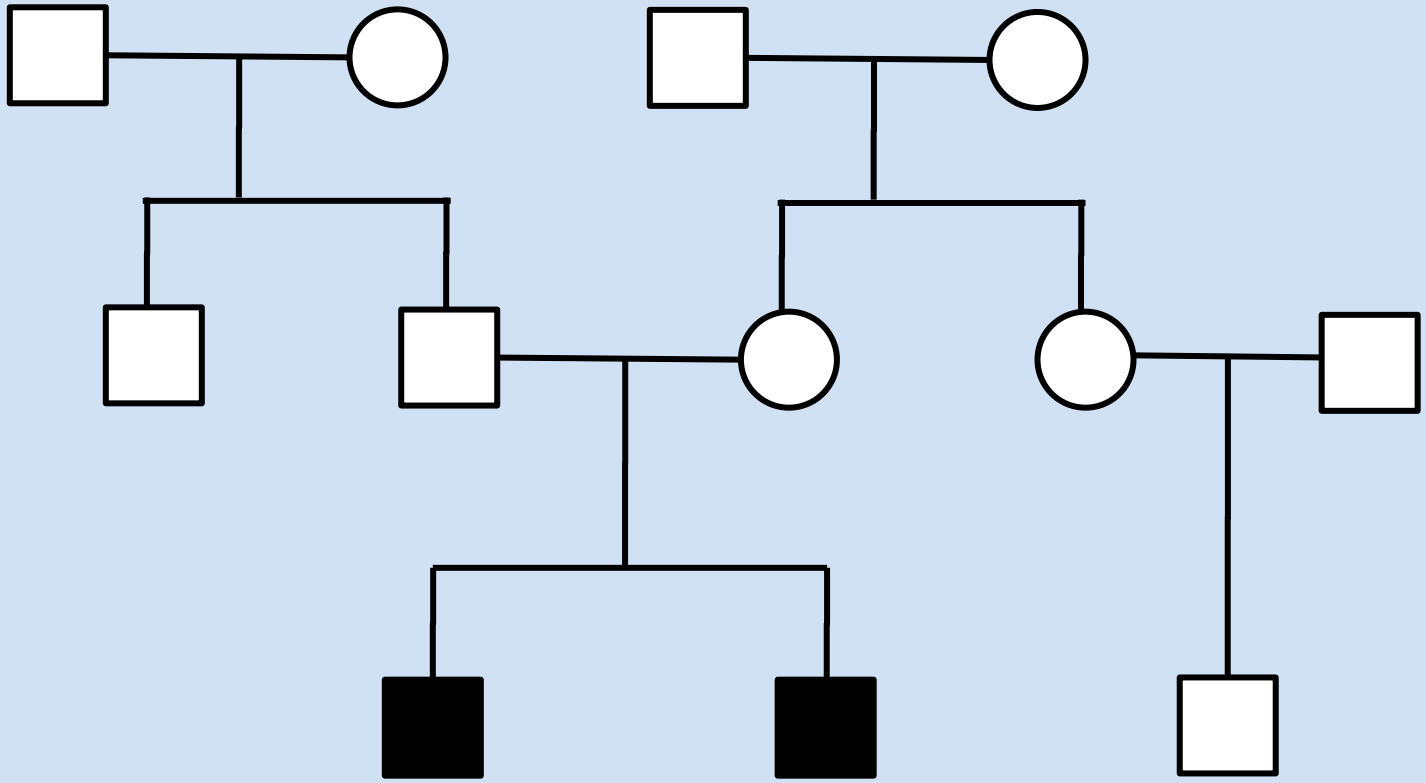
Environment matters!

Ancestry matters!

Genomic background matters!

Longitudinal course matters!

Vignette #2: Another family in Utah: New Syndrome with Intellectual Disability, “Autism”, “ADHD”



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

1.5 years old

3.5 years old

3 years old

5 years old

Dysmorphic
Mental Retardation
“autism”
“ADHD”
Hearing difficulties

Workup Ongoing for past 10 years

- Numerous genetic tests negative, including negative for Fragile X and MANY candidate genes.

- Whole genome sequencing was performed using :
  - Complete Genomics sequencing and analysis pipeline v2.0
  - Ilumina HiSeq 2000 sequencing platform.
    - Illumina reads were mapped to the hg19 reference genome using BWA v. 0.6.2-r126
    - Variant detection was performed using the GATK v. 2.4-9.
    - A second analytical pipeline was used to map reads to the hg19 reference genome using Novoalign, and variants were also detected using the FreeBayes caller.

- Standard approaches can then be used to identify potentially deleterious mutations conforming to classical disease models for genetic disorders.

- We subset the full dataset to evaluate differences between raw numbers of mutations detected between different data sets:

  – WGS data from the nuclear family,
  – WGS from a larger portion of the family.

Using only nuclear family:

55195 Variants were found to be *de-novo* in the two affected boys

122 were coding :

107 non-synonymous missense

4 splicing

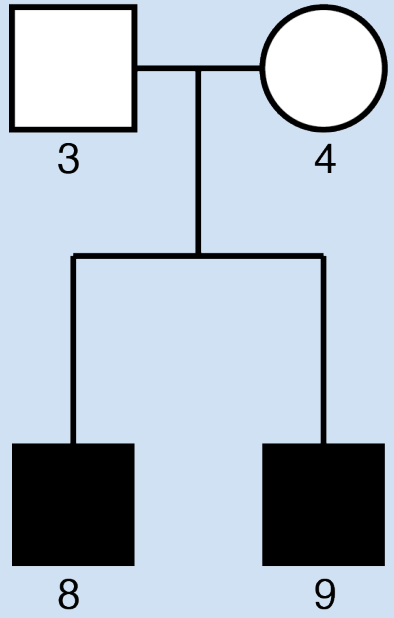
3 frame-shift deletions

3 frame-shift insertions

2 frame-shift substitutions

2 stop-gain

1 stop-loss



26514 Variants were found to conform to an X-linked disease model

28 were coding:

27 non-synonymous missense

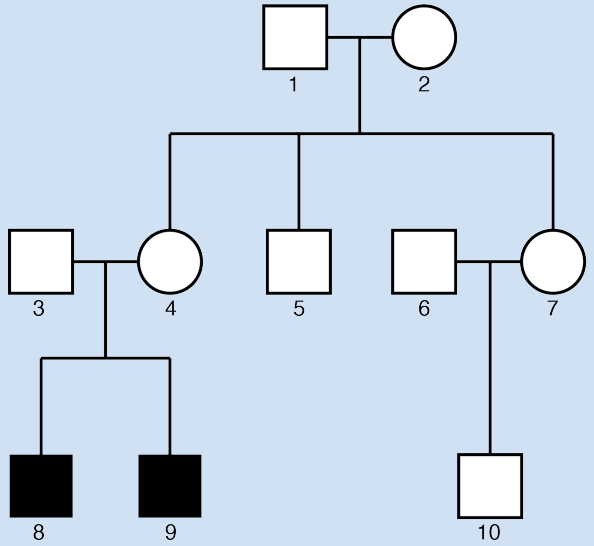
1 splicing

Using information from a greater portion of the family structure:

17726 Variants were found to be *de-novo* in the two affected boys

40 were coding :

- 32 non-synonymous missense
- 3 splicing
- 2 frame-shift deletions
- 1 stop-loss
- 1 frame-shift insertion
- 1 frame-shift substitution

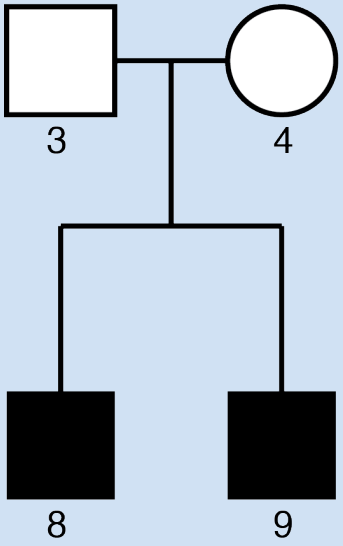


2824 Variants were found to conform to an X-linked disease model

4 were coding:

- 3 non-synonymous missense
- 1 splicing

- The numbers of mutations differ as expected between these two sets of analyses:
  - More mutations are filtered when a greater portion of the family is incorporated into the analysis.
  - This is likely due to false positive and false negative rates across sequencing and informatics platforms.



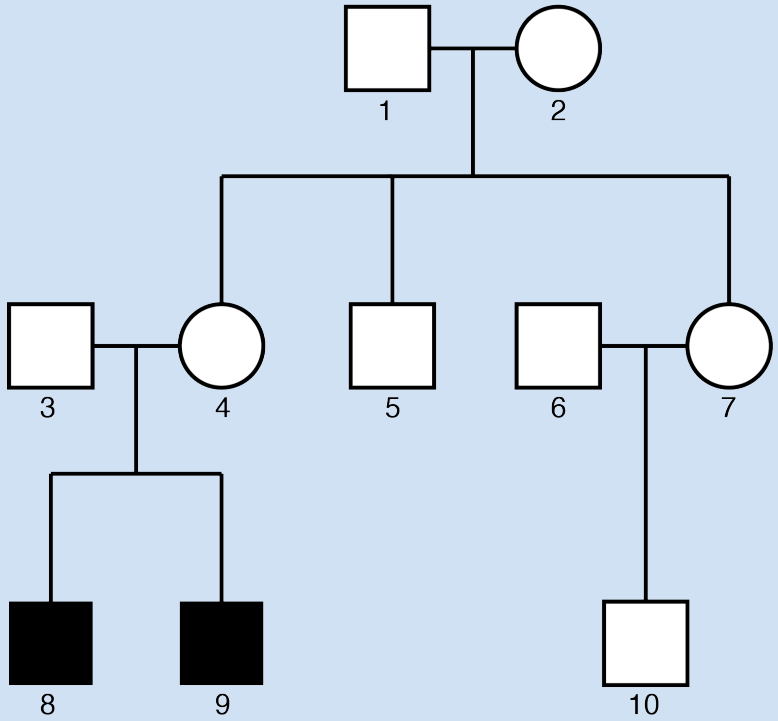
Using only nuclear family:

De-novo ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
1	PRAMEF4	0.00192	0.00144,0.00265	13.13	chr1:12939476;13.13;G->C;N->K;0,1
2	PRAMEF10	0.00318	0.00243,0.00417	20.77	chr1:12954852;20.77;T->C;H->R;3,2
3	LOC440563	0.00523	0.00416,0.00653	9.89	chr1:13183056;9.89;T->C;N->D;0,1

X-linked ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
1	ASB12	0.000898	0.000898,0.00119	18.7	chrX:63444792;18.70;C->A;G->C;0,1
2	TAF1	0.00153	0.00117,0.00214	14.59	chrX:70621541;14.59;T->C;I->T;0,1
3	ZNF41	0.002	0.0015,0.00275	12.9	chrX:47307978;12.90;G->T;D->E;0,1



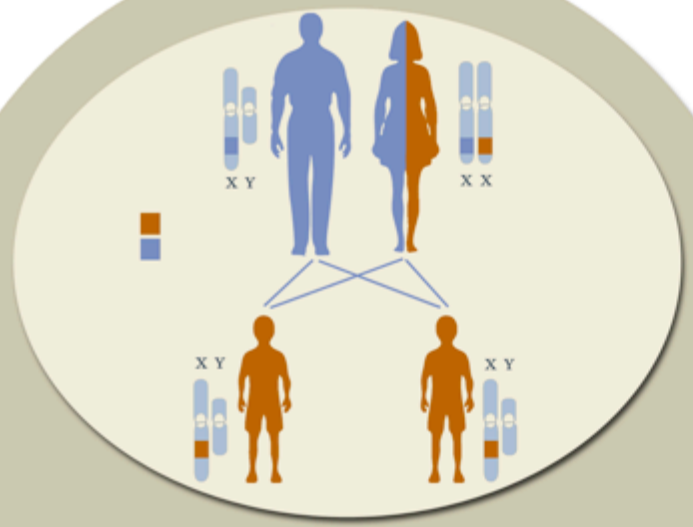
Using information from a greater portion of the family structure:

De-novo ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
1	PRAMEF10	0.00342	0.00262,0.00445	20.77	chr1:12954852;20.77;T->C;H->R;3,2

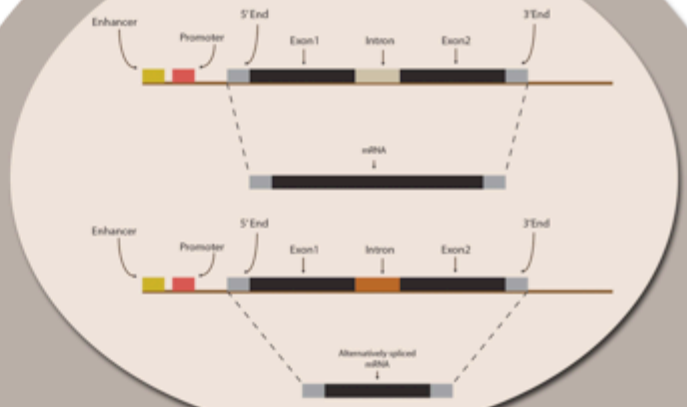
X-linked ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
1	TAF1	0.002	0.0015,0.00275	14.59	chrX:70621541;14.59;T->C;I->T;0,1



X-linked

Gene	Locus	Exon	Protein
ZNF41	X:47307978	5	p.Asp397Glu
ASB12	X:63444792	2	p.Gly247Cys
TAF1	X:70621541	25	p.Ile1337Thr



Non-coding

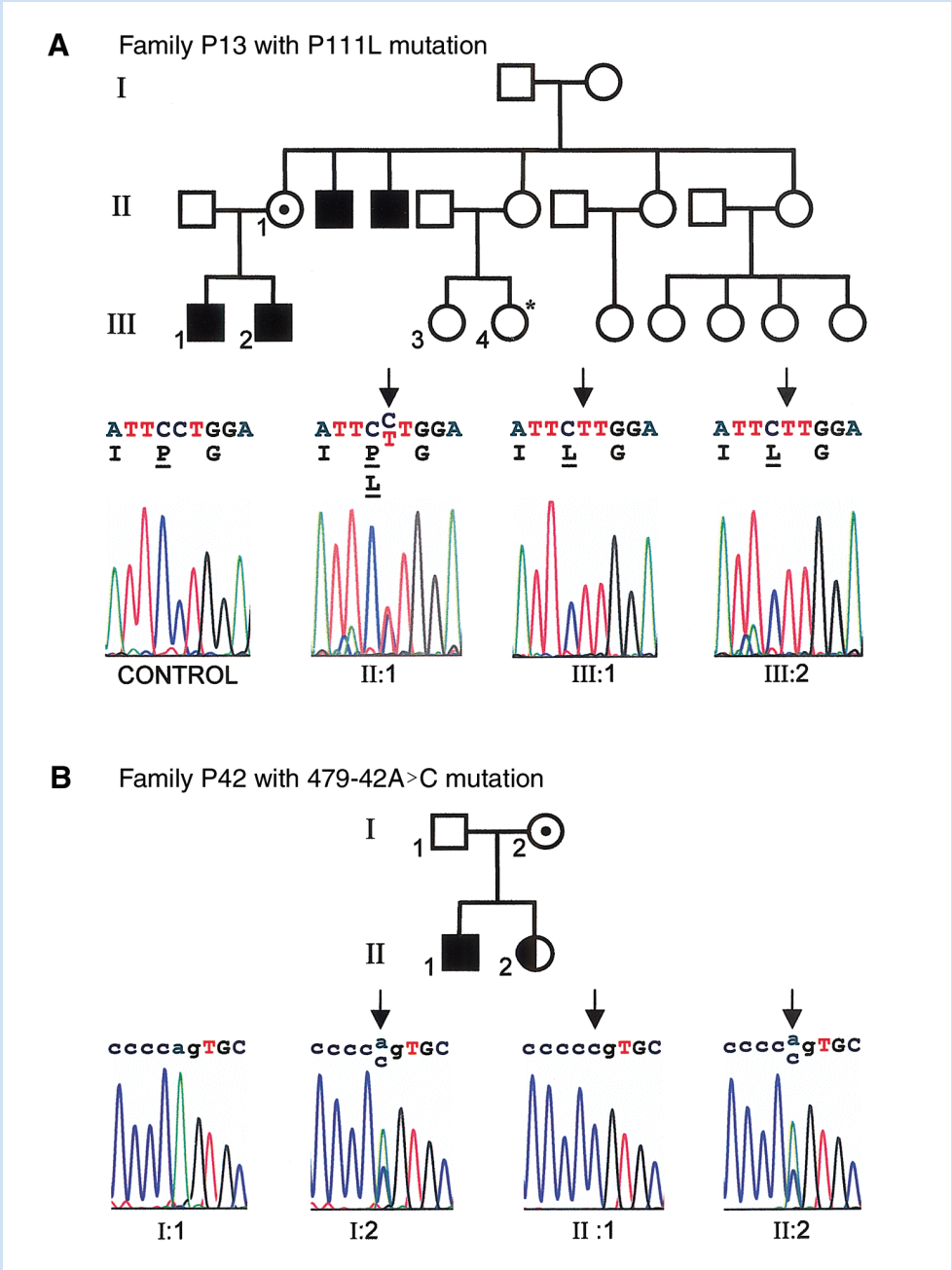
Gene	Locus	Exon	Protein
UTR3 AR	X:66945414	-------	-------
FAM155B (*dist=271971*)	X:68453113	-------	-------
MIR221 (*dist=35606*)	X:45569979	-------	-------
DMD-AS2 *intronic*	X:31284835	-------	-------
MID1 (*dist=30252*)	X:10383096	-------	-------

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]

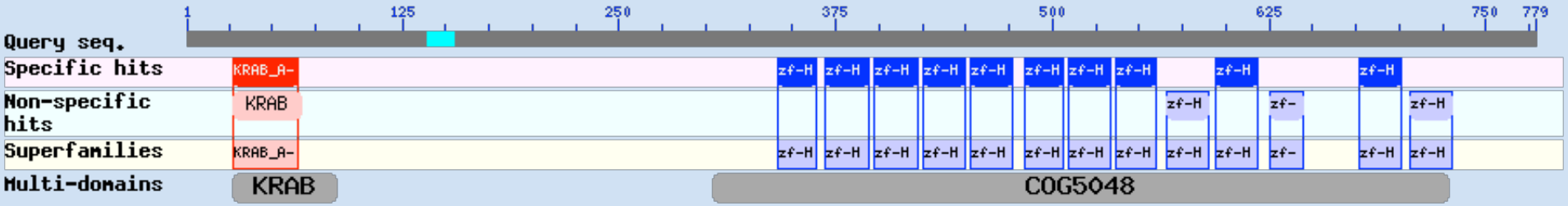
[1]Max-Planck-Institute for Molecular Genetics, Berlin; [2]Institute of Human Genetics, University of Erlangen-Nuremberg, Erlangen-Nuremberg; [3]Centre Hospitalier Universitaire de Montpellier, Hôpital Saint-Eloi, Montpellier, France, [4]Hunter Genetics and University of Newcastle, Waratah, Australia; [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



The two brothers with the P111L mutations reported in the prior paper do have mental deficiency, hyperkinesia, no motor or neurologic sign except for the delay, and slight dysmorphic facial anomalies: large low-set ears, thin upper lip, slight downward palpebral slants, but no upturned nose, and a short philtrum. The mother was normal in appearance.

ZNF41



- 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.

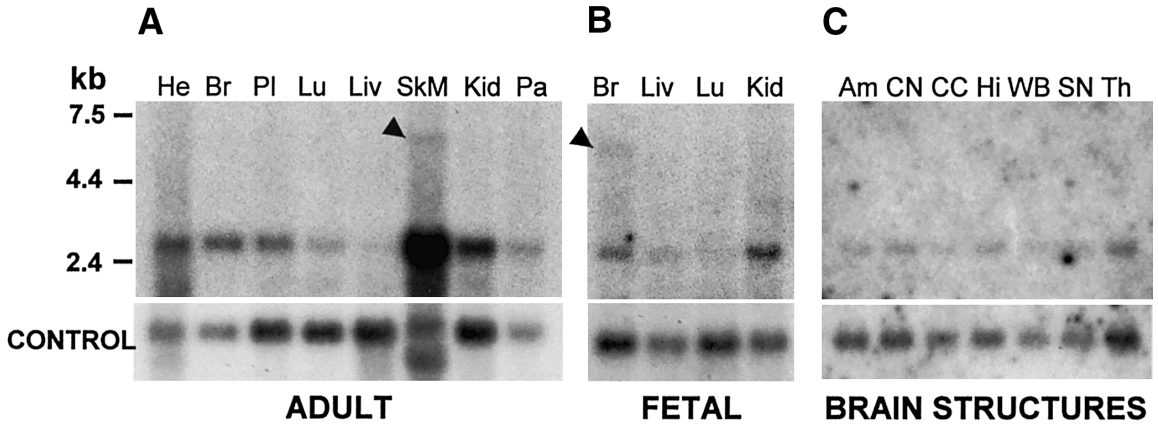
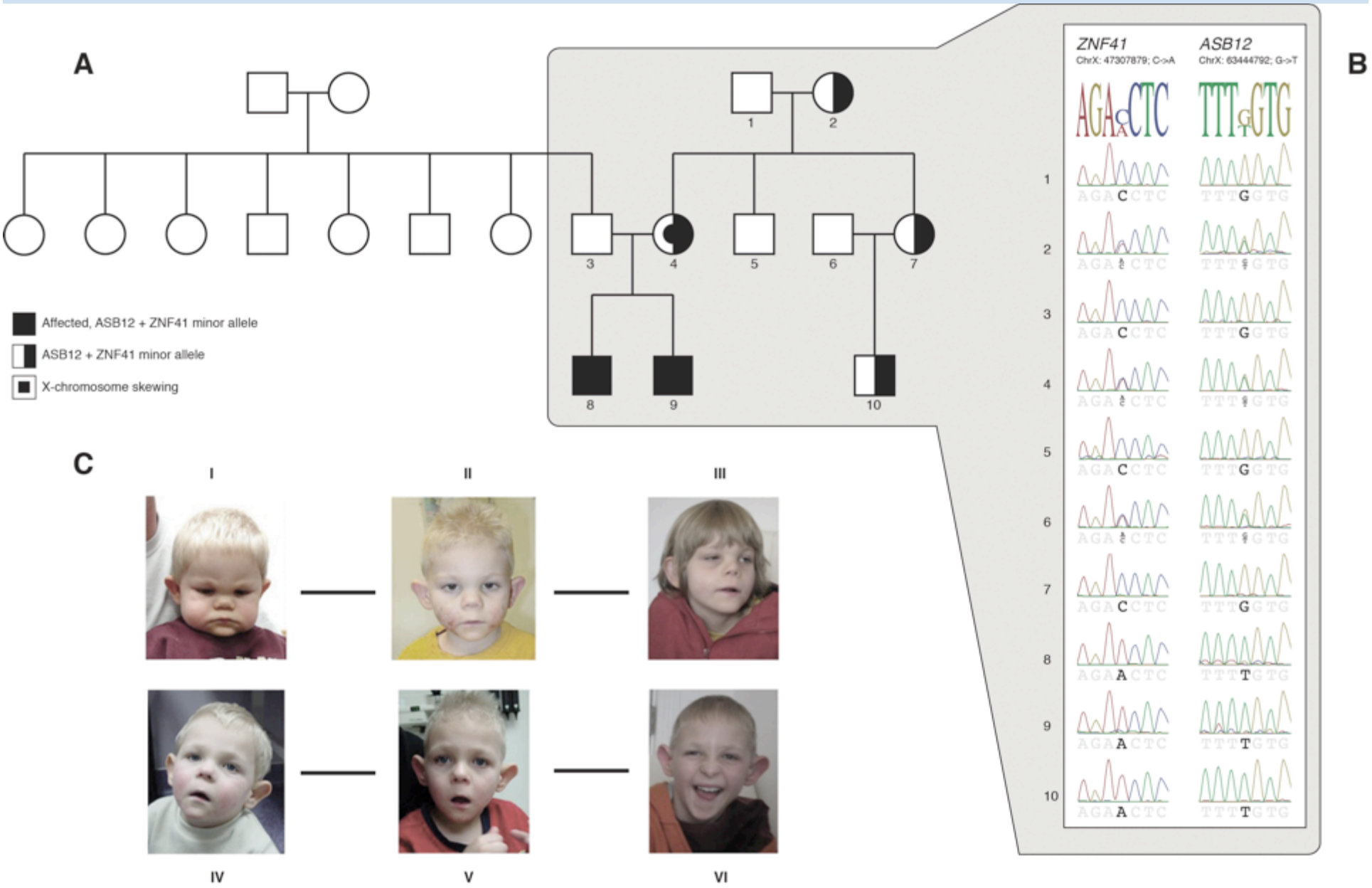


Figure 6 Northern blot hybridization of *ZNF41*, by use of a probe corresponding to nucleotides 621–1099 of *ZNF41* transcript variant 1. *A*, Adult tissues (left to right): heart, brain, placenta, lung, liver, skeletal muscle, kidney, and pancreas. *B*, Fetal tissues (left to right): brain, lung, liver,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

Proving the relevance of this mutation

- 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.


A
1
2
3
4
5
6
7
8
9
10
Affected, ASB12 + ZNF41 minor allele
ASB12 + ZNF41 minor allele
X-chromosome skewing
B
ZNF41
ChrX: 47307879; C->A
ASB12
ChrX: 63444792; G->T
C
I
II
III
IV
V
VI

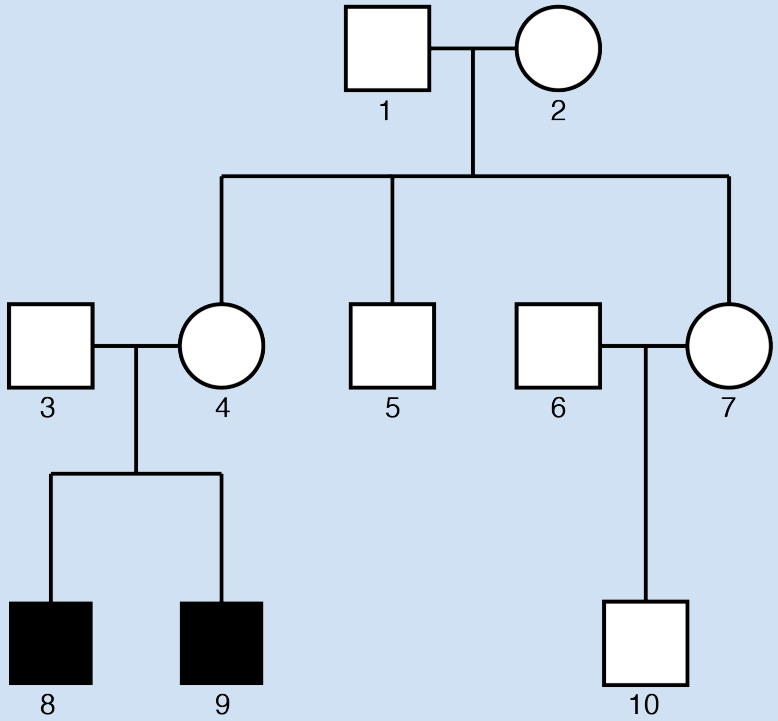
- Previously reported P111L change in the ZNF41 protein has now also been found in two "male controls" (EVS server, ESP6500), and furthermore, there are two rare, likely heterozygous ZNF41 frameshift mutations and one heterozygous stop-gained mutation reported in control individuals (ESP6500) (personal communication from Dr. Vera Kalscheuer).

XLID-Causing Mutations and Associated Genes Challenged in Light of Data From Large-Scale Human Exome Sequencing

Amélie Piton,[1,2,4,*] Claire Redin,[1,2,4] and Jean-Louis Mandel[1,2,3,*]

Because of the unbalanced sex ratio (1.3–1.4 to 1) observed in intellectual disability (ID) and the identification of large ID-affected families showing X-linked segregation, much attention has been focused on the genetics of X-linked ID (XLID). Mutations causing monogenic XLID have now been reported in over 100 genes, most of which are commonly included in XLID diagnostic gene panels. Nonetheless, the boundary between true mutations and rare non-disease-causing variants often remains elusive. The sequencing of a large number of control X chromosomes, required for avoiding false-positive results, was not systematically possible in the past. Such information is now available thanks to large-scale sequencing projects such as the National Heart, Lung, and Blood (NHLBI) Exome Sequencing Project, which provides variation information on 10,563 X chromosomes from the general population. We used this NHLBI cohort to systematically reassess the implication of 106 genes proposed to be involved in monogenic forms of XLID. We particularly question the implication in XLID of ten of them (*AGTR2*, *MAGT1, ZNF674, SRPX2, ATP6AP2, ARHGEF6, NXF5, ZCCHC12*, *ZNF41*, and *ZNF81*), in which truncating variants or previously published mutations are observed at a relatively high frequency within this cohort. We also highlight 15 other genes (*CCDC22*, *CLIC2*, *CNKSR2*, *FRMPD4*, *HCFC1*, *IGBP1*, *KIAA2022*, *KLF8*, *MAOA, NAA10*, *NLGN3*, *RPL10, SHROOM4*, *ZDHHC15*, and *ZNF261*) for which replication studies are warranted. We propose that similar reassessment of reported mutations (and genes) with the use of data from large-scale human exome sequencing would be relevant for a wide range of other genetic diseases.

The American Journal of Human Genetics *93*, 1–16, August 8, 2013



Using information from a greater portion of the family structure:

De-novo ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
1	PRAMEF10	0.00342	0.00262,0.00445	20.77	chr1:12954852;20.77;T->C;H->R;3,2

X-linked ranked genes:

RANK	Gene	p-value	p-value-ci	Score	Variants
1	TAF1	0.002	0.0015,0.00275	14.59	chrX:70621541;14.59;T->C;I->T;0,1

Summary from Vignette #2

Need to reduce the "False Negative Problem"

Genotype ≠ Phenotype

Environment matters!
Ancestry matters!
Genomic background matters!
Longitudinal course matters!

We can only begin to really understand this if we utilize the power of intense networking via internet-enabled archiving and distribution of data.

Down Syndrome

ARTICLE

An Excess of Deleterious Variants in VEGF-A Pathway Genes in Down-Syndrome-Associated Atrioventricular Septal Defects

Christine Ackerman,[1] Adam E. Locke,[2,8] Eleanor Feingold,[3] Benjamin Reshey,[1] Karina Espana,[1] Janita Thusberg,[4] Sean Mooney,[4] Lora J.H. Bean,[2] Kenneth J. Dooley,[5] Clifford L. Cua,[6] Roger H. Reeves,[7] Stephanie L. Sherman,[2] and Cheryl L. Maslen[1,*]

About half of people with trisomy 21 have a congenital heart defect (CHD), whereas the remainder have a structurally normal heart, demonstrating that trisomy 21 is a significant risk factor but is not causal for abnormal heart development. Atrioventricular septal defects (AVSD) are the most commonly occurring heart defects in Down syndrome (DS), and ~65% of all AVSD is associated with DS. We used a candidate-gene approach among individuals with DS and complete AVSD (cases = 141) and DS with no CHD (controls = 141) to determine whether rare genetic variants in genes involved in atrioventricular valvuloseptal morphogenesis contribute to AVSD in this sensitized population. We found a significant excess ($p < 0.0001$) of variants predicted to be deleterious in cases compared to controls. At the most stringent level of filtering, we found potentially damaging variants in nearly 20% of cases but fewer than 3% of controls. The variants with the highest probability of being damaging in cases only were found in six genes: *COL6A1*, *COL6A2*, *CRELD1*, *FBLN2*, *FRZB*, and *GATA5*. Several of the case-specific variants were recurrent in unrelated individuals, occurring in 10% of cases studied. No variants with an equal probability of being damaging were found in controls, demonstrating a highly specific association with AVSD. Of note, all of these genes are in the VEGF-A pathway, even though the candidate genes analyzed in this study represented numerous biochemical and developmental pathways, suggesting that rare variants in the VEGF-A pathway might contribute to the genetic underpinnings of AVSD in humans.

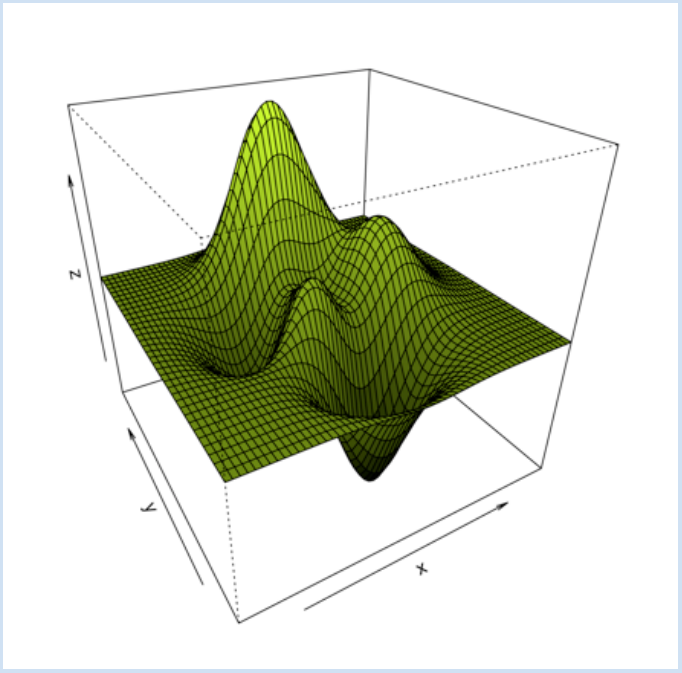
Creating a Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data

June 3, 2013

Organizing Committee

David Altshuler	Broad Institute of Harvard and MIT, MGH
Peter Goodhand	Ontario Institute for Cancer Research
David Haussler	HHMI/University of California, Santa Cruz
Thomas Hudson	Ontario Institute for Cancer Research
Brad Margus	A-T Children's Project
Betsy Nabel*	Brigham and Women's Hospital
Charles Sawyers	HHMI / Memorial Sloan-Kettering
Michael Stratton*	Wellcome Trust Sanger Institute

Need sharing of thousands to millions of “individuated genomes” – credit Nathaniel Pearson
http://www.slideshare.net/NathanielPearson/pearsontcgc2013


z
y
x

The End