

CAGI slide pack guidelines

I agree to my slides being used according to the following terms:

Lipika R. Pal

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 2013. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.





SickKids5 Clinical Genomes Challenge Predictions

Moult Lab

Lipika Ray Pal

Institute for Bioscience and Biotechnology Research
University of Maryland at College Park, Rockville, MD

July 7, 2018

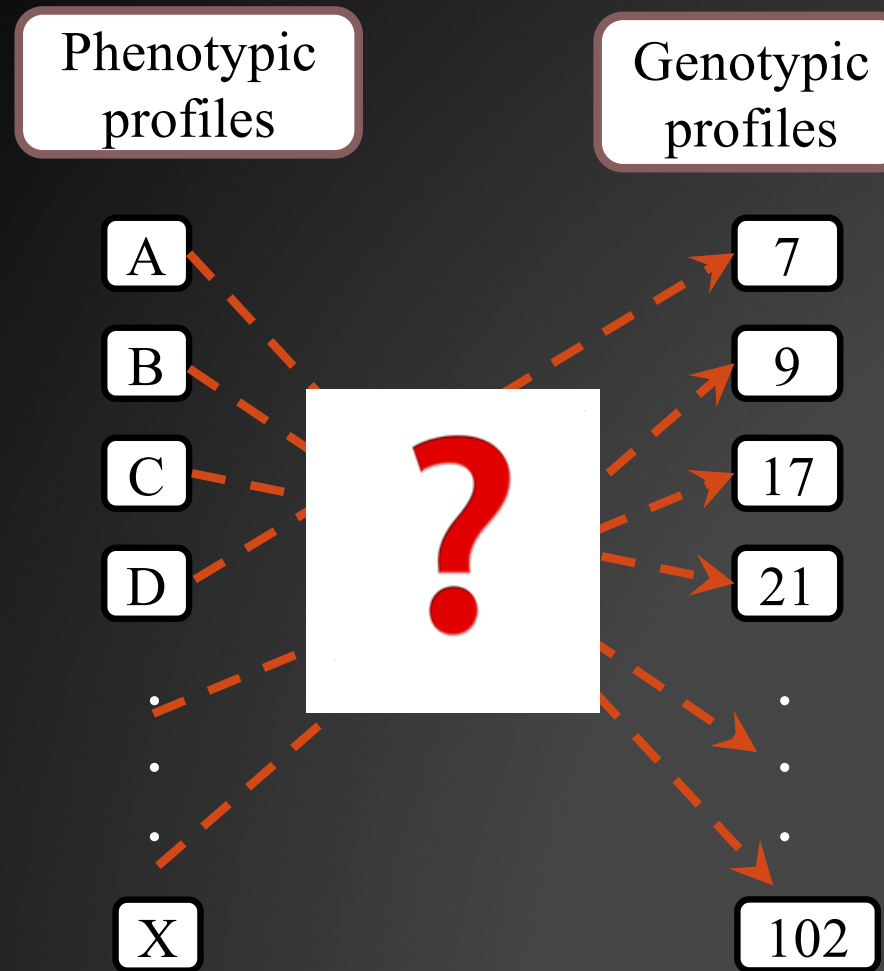


UNIVERSITY OF MARYLAND | NIST
**INSTITUTE FOR BIOSCIENCE
& BIOTECHNOLOGY RESEARCH**



The challenge is to predict:

- (1) What class of disease is associated with each genome
- (2) Which genome corresponds to which specific clinical description



CAGI5 SickKids data processing for 24 children:

Whole genome Illumina
HiSeq data

VCF files with SNVs and indels. Annotation by Varant and Annovar tools.

Categorization of variants based on the strength of the evidence for disease relevance and different frequency bins at less than 1% population frequency.

SC1: Clinical variants – HGMD & ClinVar

SC2: Nonsense, Frameshift, Splice sites. Non-synonymous variants with consensus impact prediction rate > 0.6 by SNPs3D profile & stability, SIFT, Polyphen2, Vest, Revel, CADD.

SC3: Other impact Non-synonymous variants with consensus prediction rate < 0.6.

SC4: Benign non-synonymous variants.

SC5: Close to splice acceptor or donor site.

SC6: UTR and intronic impact variants by any of CADD, GERP++ and Eigen

Variants from each category are divided into 4 frequency bins according to their highest population frequency found in any of 1000 genome, ExAC, GnomAD databases.

f0: Novel variants, **f01:** Frequency > 0 and < 0.001

f05: Frequency > 0.001 and < 0.005, **f1:** Frequency > 0.005 and < 0.01.

Clinical
phenotype
description

Genelist:
HPO,
dbNSFP,
RetNet
database

6239 Genes
from 213
phenotypes
(about 9
phenotypes
per child)



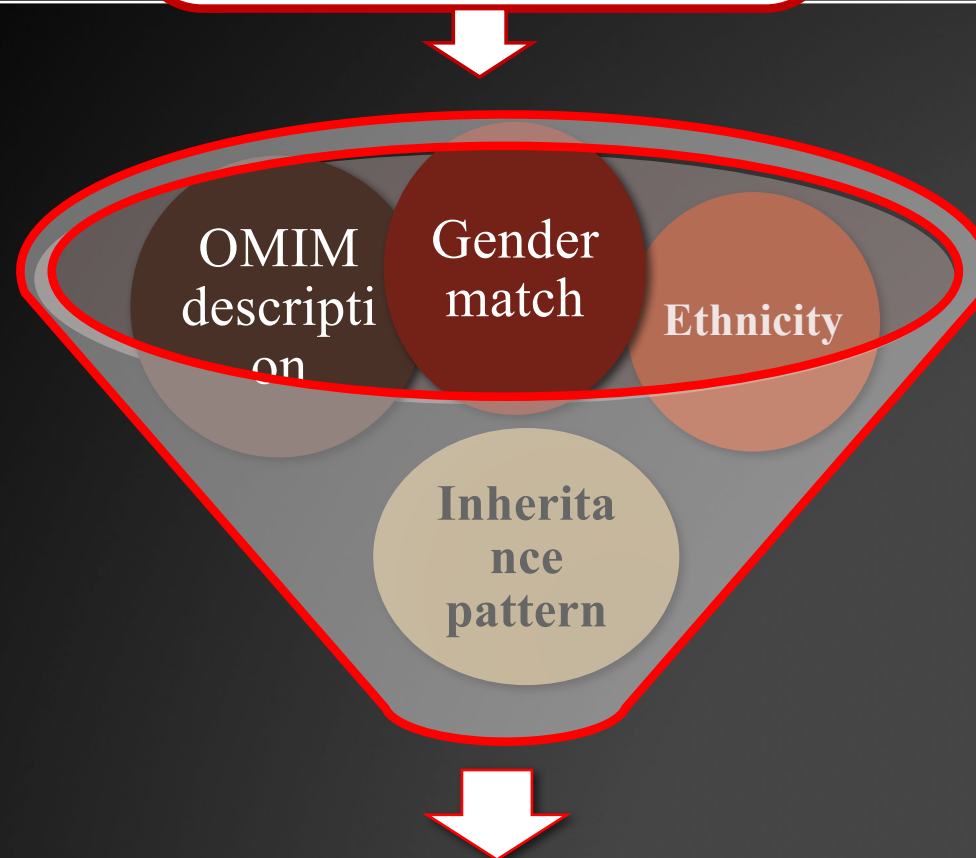
SCORING SCHEME

For each phenotypic profile, each phenotypic term (T) was assigned a weight (W) from 0 to 1, according to its importance (1 = most important, 0 = least important) in that clinical profile. Each variant (V) in a genome was scored according to the total number (n) of phenotypic terms associated with the corresponding gene when comparing with a particular clinical profile.

$$V_i = \sum_{j=1}^n W_{T_j}$$



**Top 5 scoring variants in
a genome for a given
clinical profile and
among all clinical
profiles**



**Probable diagnostic
variant(s)**



Disease class for 24 kids:

Eye disorder: 6

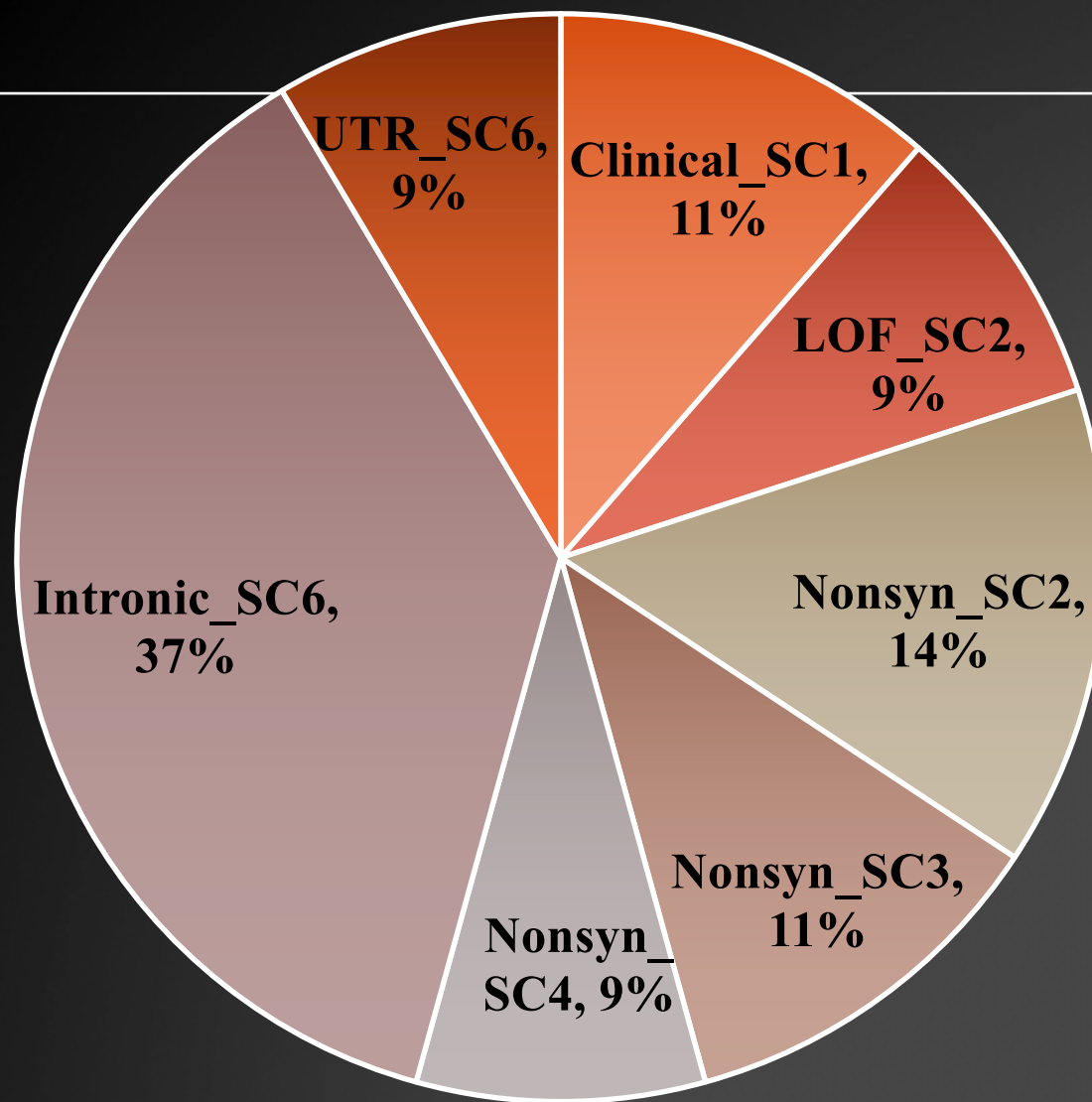
Neurological disorder: 7

Connective Tissue disorder: 11

Overall Performance Self - assessment

- (1) Genomes assigned the correct disease class : 12
- (2) Genomes assigned the correct clinical profile: 5

Selected Diagnostic Variants





Exact Match Cases

Clinical Profile	Genotypic Profile	Disease class	Diagnostic Variants
B	099	Neurologic	UTR variant in PIGT
C	095	Eye	Compound het intronic variants in USH2A
F	093	Connective	Novel missense variant in EP300
H	017	Eye	Compound het variants, one clinical and another intronic in ABCA4
N	056	Connective	UTR and intronic variants in FBN1

Data supporting chr15:48849771 (rs16961144)

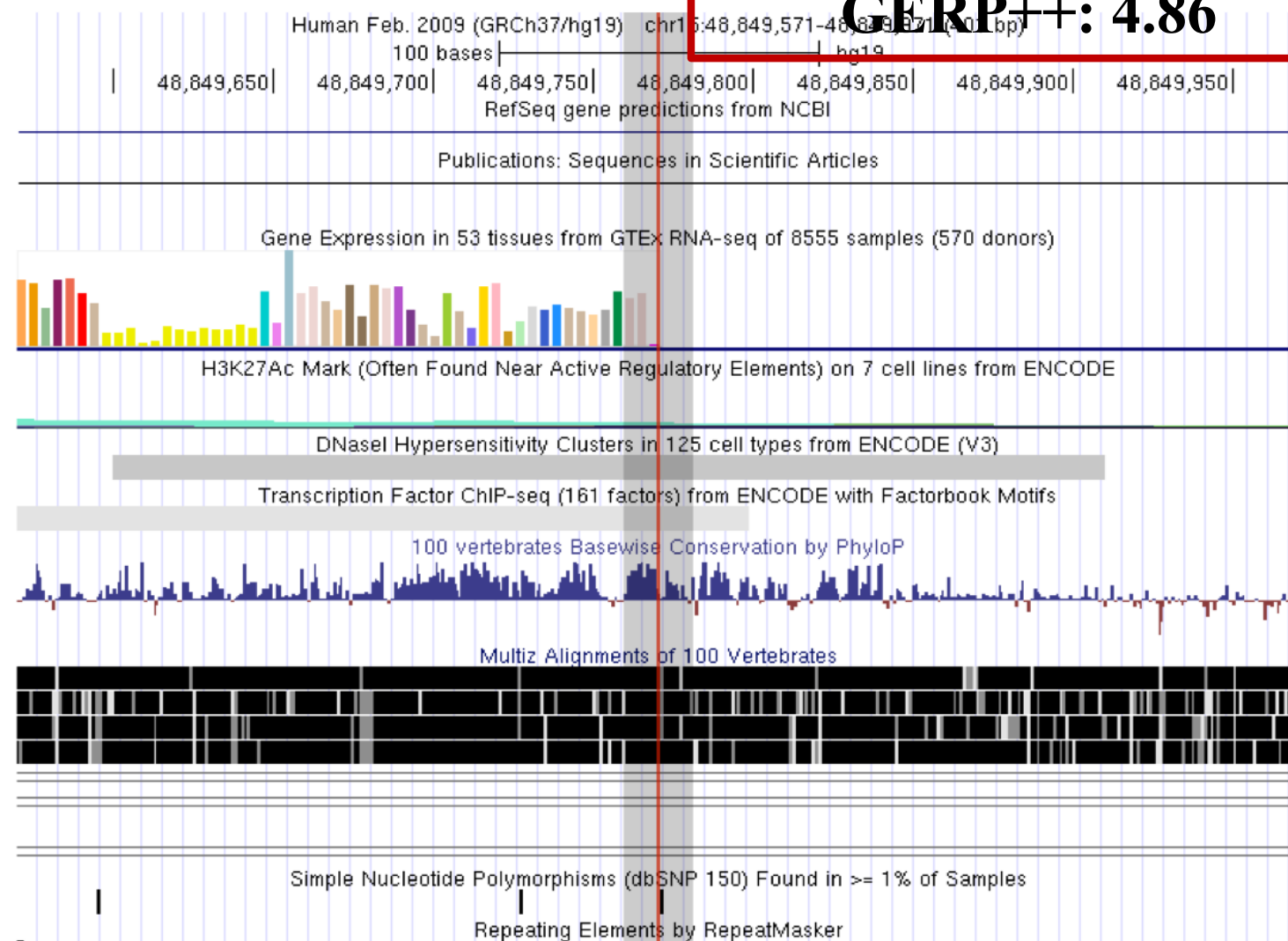
Score: 4

Minimal binding evidence

FBN1 - chr15:48849792

CADD: 22.1

GERP++: 4.86





Unsuccessful Puzzling Cases

Wrong Clinical Profile	Wrong Genotypic Profile	Disease class	Wrong Diagnostic Variants
J	091	Neurologic	Clinical variant and NonFrameShift Delete in POLG, ClinVar Mitochondrial variant
D	076	Connective	FrameShift insert in COL5A1
I	092	Connective	NonFrameShift Delete in ELN
I	081	Connective	UTR5 variant in ELN



Predictive Secondary Variants

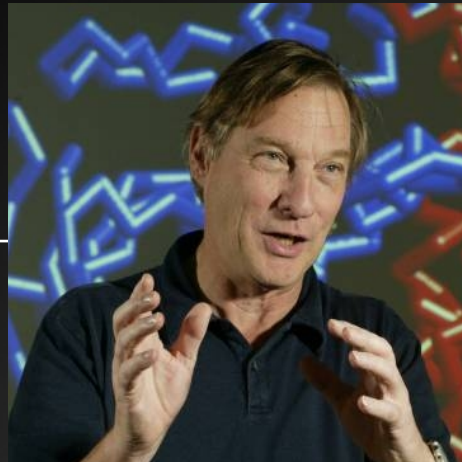
- ❑ Predictive secondary variants according to 2017 ACMG guidelines are present in 6 children out of 24.
- ❑ Searched for clinical variants and loss of function variants as recommended in the guideline for 59 genes.
- ❑ We found a total of 6 clinical variants (3 of which are novel) and 2 Frameshift deletions (all novel) as predictive secondary variants.



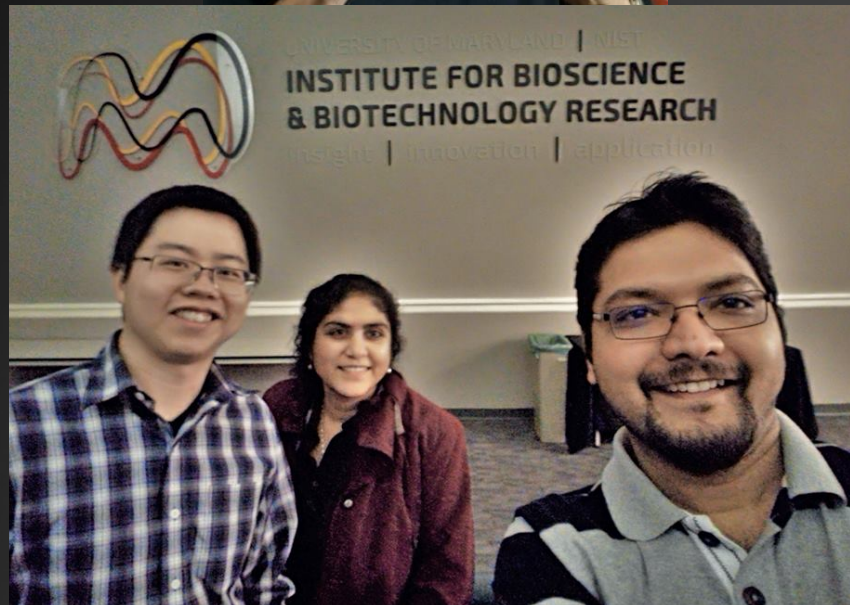
Conclusion

- ❑ 46% of our predicted diagnostic variants are intronic and UTR. 46% of all variants are non-synonymous. The other 9% are loss of function variants.
- ❑ Disease class identification is more feasible than exact match. We found 5 exact match cases (21%) and 12 cases (50%) with correct disease class.
- ❑ Need to understand more about the mechanism underlying the disease as why LOF variants are not related to patient's phenotype.

Acknowledgement:



Prof. John Moults



Dr. Yizhou Yin

Kunal Kundu

CAGI Organizers
Gaia Andreoletti

Dr. Steven Meyn
All the genome donors