An Informatics Pipeline for Identifying Clinically Relevant Incidental Variants in Whole Genome Sequences

Jonathan S. Berg[1], Michael Adams[1], Nassib Nassar[2], Chris Bizon[2], Kristy Lee[1], Charles P. Schmitt[2], Kirk C. Wilhelmsen[1,2], and James P. Evans[1]

1. Department of Genetics, UNC-Chapel Hill 2. Renaissance Computing Institute, Chapel Hill, NC

Introduction

Next generation sequencing (NGS) has transformed medical genetics research and appears poised to revolutionize clinical diagnosis of genetic diseases.

However, the vast amounts of data and inevitable discovery of clinically relevant incidental findings pose challenges to the adoption of these techniques in the clinic, necessitating novel analytic approaches.

We developed an informatics pipeline that can identify potential disease-causing variants in genes implicated in over 2000 Mendelian diseases. We intend for this approach to be used in the ***incidental*** analysis of genome sequences; as such, the parameters used to decide which variants to report reflect a low a priori likelihood that an individual has a Mendelian disease.

Methods

We recently described a conceptual strategy for classifying genes into three "bins" to facilitate informed consent, analysis, and return of incidental findings in a clinical setting.

- Bin 1 contains genes with clinical utility, in which a mutation would trigger specific medical action with defined benefit in morbidity/mortality.
- Bin 2 contains genes known to be associated with human diseases for which evidence does not support any specific action; bin 2 is further stratified based on the potential for psychosocial distress or harm.
- Bin 3 contains genes with no known clinical relevance.

2016 genes linked with Mendelian diseases were categorized into Bin 1, Bin 2b, Bin 2c.

	Criteria:	Clinical Utility	Clinical Validity				Unknown Clinical Implications
Genes	Bins:	**Bin 1 Medically actionable incidental information**	**Bin 2A Low risk incidental information**	**Bin 2B Medium risk incidental information**	**Bin 2C High risk incidental information**	**Bin R Carrier status for recessive disorders**	**Bin 3 All other loci**
	Examples:	BRCA1/2, MLH1, MSH2, FBN1, NF1	PGx variants and common risk SNPs	APOE	Prion diseases ALS (SOD1) SCA	Heterozygous carriers of autosomal recessive or X-linked diseases	
	Number of genes/loci:	161		1798	57		~20,000

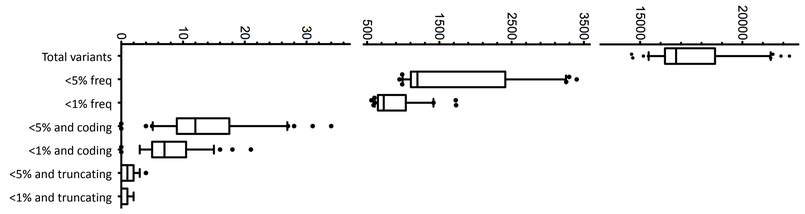
Alleles that would be reportable (YES) or not reportable (NO) in a clinical context

Variants							
	Known deleterious	YES	YES/NO[1]	YES/NO[1]	YES/NO[1]	YES/NO[1]	N/A[2]
	Presumed deleterious	YES	N/A[3]	YES/NO[1]	YES/NO[1]	YES/NO[1]	NO[4]
	VUS	NO	N/A[3]	NO	NO	NO	NO[4]
	Presumed benign	NO	N/A[3]	NO	NO	NO	NO
	Known benign	NO	NO	NO	NO	NO	NO

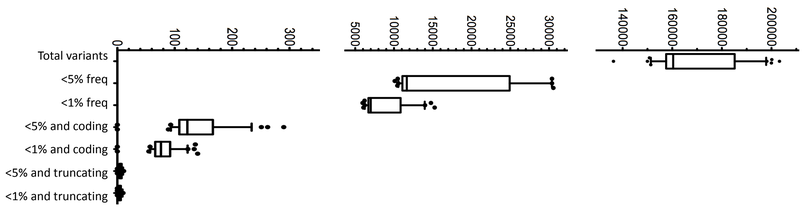
N/A: not applicable; VUS: Variant of uncertain significance
[1] Reporting through decision making with an appropriate provider if elected by the patient
[2] By definition, variants in genes with unknown implications could not be considered deleterious
[3] By definition, SNPs or PGx variants will either be present or absent
[4] Variants in genes with unknown clinical implications would not be reported; however they may serve as an important substrate for research, potentially uncovering new disease genes

We used a python script to bin every variant from 80 whole genome sequences. To assist with this, we used allele frequencies (AF) from the 1000 Genomes project.

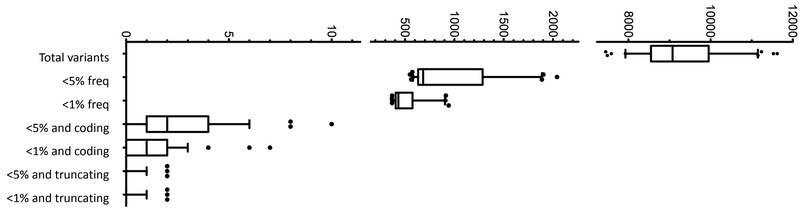
Results—Bin 1 genes



Results—Bin 2b genes

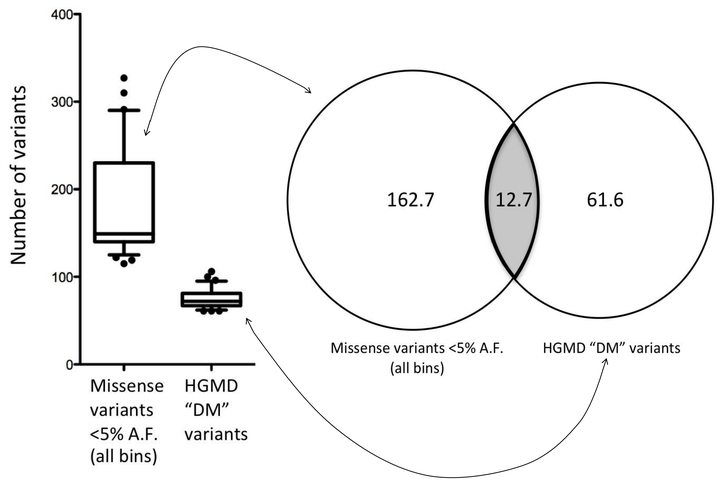


Results—Bin 2c genes



The algorithm effectively identified incidental variants of clinical relevance. Restricting analysis to rare (<5% allele frequency) truncating variants drastically reduced the number of variants in each bin. However, there remained too many rare missense variants (most of which would be classified as VUS) for effective manual curation, necessitating that these be ignored.

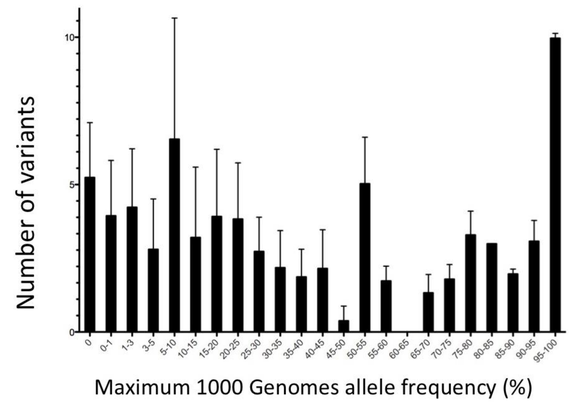
To improve sensitivity to possible disease-causing missense variants, we queried a local instance of the Human Gene Mutation Database for variants that were annotated as "DM". We found minimal overlap between the ~150 missense variants per person and the DM variants detected in the 80 genomes.



The HGMD query resulted in 871 unique "DM" variants across 80 genomes.

- 251 of 871 unique "DM" variants had AF >5%
- 78% of "DM" variants per genome had AF >5%

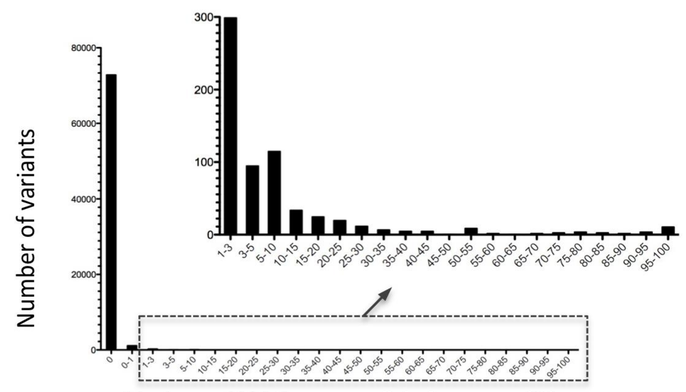
HGMD "DM" variants in 80 genomes



Considering all HGMD "DM" variants:

- Vast majority of variants had AF <1% (as expected)
- 1546 (~2%) of variants had AF 1-5%
- 265 (~0.4%) of variants had AF >5%

All HGMD "DM" variants



Conclusions

- Caution should be used when interpreting HGMD "DM" variants, because a significant proportion are likely not pathogenic
- Computer-assisted "binning" of the genome facilitated the discovery of known disease-causing mutations and novel, predicted deleterious mutations.
- This method is an efficient and practical way identify clinically relevant incidental findings.
- It is readily adaptable to other types of clinical genomic analyses, scalable to the demands of a clinical laboratory workflow, and flexible with respect to advances in medical genetics and genomics.

This work was supported by the University Cancer Research Fund (http://unclineberger.org/ucrf) and the UNC Bryson Philanthropic Fund. J.P.E is supported by the UNC Center for Genomics and Society (NHGRI 5-P50-HG004488-03) and a UNC Clinical Translational Science Award (1-UL1-RR025747-01). K.C.W. is supported by NIDA 1R01 DA03097601. The authors would like to acknowledge Kristy Crooks, Jessica Booker, and Karen Weck for thoughtful discussions regarding "binning" and Erik Scott and Guifeng Jin for assistance with processing the Complete Genomics data and 1000 Genomes allele frequencies.