Published in final edited form as:
Genet Med. 2013 January ; 15(1): 36–44. doi:10.1038/gim.2012.112.

An informatics approach to analyzing the incidentalome

Jonathan S. Berg[1,2,3], **Michael Adams**[1], **Nassib Nassar**[4], **Chris Bizon**[4], **Kristy Lee**[1], **Charles P. Schmitt**[4], **Kirk C. Wilhelmsen**[1,3,4], and **James P. Evans**[1,2,3]

[1]Department of Genetics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599

[2]Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599

[3]Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599

[4]The Renaissance Computing Institute, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599

Abstract

Purpose—Next-generation sequencing (NGS) has transformed genetic research and is poised to revolutionize clinical diagnosis. However, the vast amount of data and inevitable discovery of incidental findings require novel analytic approaches. We therefore implemented for the first time a strategy that utilizes an *a priori* structured framework and a conservative threshold for selecting clinically relevant incidental findings.

Methods—We categorized 2016 genes linked with Mendelian diseases into “bins” based on clinical utility and validity, and used a computational algorithm to analyze 80 whole genome sequences in order to explore the use of such an approach in a simulated real-world setting.

Results—The algorithm effectively reduced the number of variants requiring human review and identified incidental variants with likely clinical relevance. Incorporation of the Human Gene Mutation Database (HGMD) improved the yield for missense mutations, but also revealed that a substantial proportion of purported disease-causing mutations were misleading.

Conclusions—This approach is adaptable to any clinically relevant bin structure, scalable to the demands of a clinical laboratory workflow, and flexible with respect to advances in genomics. We anticipate that application of this strategy will facilitate pre-test informed consent, laboratory analysis, and post-test return of results in a clinical context.

Keywords

Whole genome sequencing; whole exome sequencing; clinical informatics; incidental findings; secondary findings

INTRODUCTION

The rapidly decreasing cost of whole genome sequencing (WGS) and its ability to simultaneously analyze all human genes make it an attractive technique for genetic diagnosis. Early anecdotal reports describing the use of WGS or whole exome sequencing (WES) have demonstrated the power of these new technologies to impact patient care.[1–3] However, there exist significant barriers to the widespread application of WGS/WES in

Correspondence: Dr. Jonathan S. Berg, MD/PhD, Department of Genetics, Room 5092 Genetic Medicine Building, Chapel Hill, NC 27599, 919-966-7043, jsberg@med.unc.edu.

clinical medicine. Technical hurdles are being addressed in the marketplace, where competition will lead to faster, cheaper, and more accurate sequencing.[4] Practical obstacles such as the time and effort required for analysis of clinically relevant variants, and return of complex results to patients, will require transition from traditional genetic testing approaches.

In a clinical environment, the most productive use of WGS/WES will likely be in the diagnosis of patients with distinctive features suggestive of a genetic disorder. In these individuals, there will also be genetic findings unrelated to the presenting symptoms, which are "incidental" or "secondary" findings, the aggregate of which has previously been termed the "incidentalome."[5] Arguably, the vast majority of an individual's genetic variants will be unrelated to the presenting symptoms. Thus, the problem of how to deal with incidental findings poses a formidable problem for clinicians and laboratorians.

In the pursuit of evidence-based genomic medicine, it will be vital to avoid overwhelming patients and physicians with genomic findings of dubious clinical value. Since the use of common single nucleotide polymorphisms (SNPs) for prediction of common disease risk is still of limited value clinically,[6] we have chosen to focus on monogenic disorders. Given that any individual has a very small *a priori* likelihood of being affected with an incidentally identified Mendelian disorder, few truly disease-causing genetic variants are expected per person. Thus, any attempt to sift through genomic data for clinically relevant incidental findings will benefit from the recognition that the vast majority of variants bear negligible clinical significance. In other words, the identification of incidental findings should maximize specificity.

The challenge, therefore, is to synthesize collective knowledge about the genetic causation of disease and implement a practical, clinically oriented approach to the analysis of genome-scale variant data. We recently described a conceptual strategy for classifying genes into "bins" to facilitate informed consent, analysis, and return of incidental findings in a clinical setting.[7] In our proposed system, the first step is to assign genes to bins according to features such as clinical utility/actionability (Bin 1) and clinical validity (Bin 2), and the potential to cause harm (Bin 2a, 2b, 2c; see Supplemental materials for details). The second step is to select the variants in a given individual that merit detailed review. The third step involves human review of the resulting subset of variants. Since a variant of uncertain significance (VUS), by definition, has no known clinical value, only known mutations or likely disease-causing novel mutations would be reported as incidental findings.

Variants identified by any sequencing method can be readily sorted based on their genomic location (whether they fall within a "binned" gene) and further annotated in terms of effect on the translated protein and predicted zygosity. For recessive disorders in which a single heterozygous mutation signifies the carrier state but is not considered disease causing, heterozygous variants would be moved into a separate category, "Bin R," for reproductive implications. Our binning approach thus attempts to capture clinical differences between genes and organize them into a succinct number of categories in order to facilitate the pre-test counseling and post-test reporting of suspected disease-causing variants when discovered incidentally during WGS/WES.

The goals of this endeavor were to evaluate the average number and type of potentially clinically relevant incidental findings and the impact of various "filters" on the output of the proposed analytic framework. We implemented a prototype of this strategy with an analysis of 80 whole genomes as a proof-of-concept, showing that multiple genomes can be efficiently analyzed to identify clinically relevant variants. This strategy can be refined with advances in our understanding of disease-causing and benign variants and offers an initial

means of structured clinical assessment of WGS/WES data in a practical and efficient manner.

SUBJECTS AND METHODS

Binning of Online Mendelian Inheritance in Man (OMIM) genes

OMIM files (accessed June, 2011) containing entries for 12,786 genes were scrutinized using OMIM, PubMed, Gene Reviews, and other resources. Genes were placed in Bin 3 (no clinical implications) if there was no indication of association with a Mendelian disorder, if only somatic mutations were reported, or if limited evidence of pathogenicity was available. Loci mapped by linkage, for which specific genes/mutations have not been documented, were also removed from consideration. A total of 2016 genes associated with Mendelian disorders were identified, and their respective inheritance patterns were determined.

We made two judgments about genes involved in Mendelian disorders: (1) Most genes do not have clinical utility/actionability in terms of definable preventive measures or treatment and (2) for most Mendelian disorders, the potential for psychosocial harm caused by their incidental discovery is neither trivial nor highly concerning. Thus, all 2016 genes were initially placed in Bin 2b. We then manually reviewed each gene and applied a first order approximation of the previously outlined criteria to provisionally place each gene into a bin. Genes for which there existed a reasonable suggestion of beneficial interventions were provisionally assigned to Bin 1. Genes having potentially significant risk of psychosocial harm were provisionally assigned to Bin 2c.

Genome sequences

WGS was performed by Complete Genomics (Mountain View, CA).[8] 19 genomes were from patients enrolled in an IRB-approved research study for genetic evaluation of hereditary cancer susceptibility. 61 genomes, representing presumably healthy individuals from diverse ethnic groups, were made publically available by Complete Genomics (http://www.completegenomics.com/sequence-data/download-data/). All genome coordinates are based on NCBI build 37.

Databases and Computational Analysis

Tables containing variant calls and annotations were stored in a PostgreSQL 8.4.3 database and joined with a table of allele frequencies generated from phase 1 consensus SNP sites (5/2/2011) from the 1000 Genomes project and small insertion/deletion calls from the 1000 Genomes pilot paper dataset (10/20/2010).[9] In order to address differences in allele frequency (AF) between different populations, we used the highest minor AF reported for a given variant in any of the phase 1 population groups. A local instance of the Human Gene Mutation Database (HGMD)[10] was created in another PostgreSQL database. Genomic coordinates for HGMD mutations were lifted over to NCBI Build 37 and converted to the Complete Genomics standard variant format. Variants matching with annotated disease mutations (“DM” variants) could then be readily identified in the 80 WGS samples.

A Python (2.6.5) script was written to iterate through variant files and select variants meeting the criteria outlined in the manuscript. Since Complete Genomics independently calls each allele, two separate lines in the variant file represent homozygous variants. The script collapses homozygous positions to a single line and indicates the variant’s zygosity in a separate field. For genes associated with autosomal recessive disorders, the script counts the number of variants meeting the predefined criteria and, if only one heterozygous variant exists, annotates that variant as signifying carrier status. The algorithm thus recognizes the

potential for biallelic mutations (although it is important to note that further investigation is required to adjudicate whether the mutations are in *cis* or in *trans*).

RESULTS

To demonstrate the applicability of the proposed analytic framework, we provisionally binned 2016 genes implicated in Mendelian disorders, implemented a computational analytic pipeline, and explored the output from 80 whole genome sequences. In this first attempt at binning the genome (Supplemental Table S1), 161 genes were assigned to Bin 1, 1798 genes were assigned to Bin 2b, and 57 genes were assigned to Bin 2c. We emphasize that the binning of genes used in this study is provisional and used for illustrative purposes; the final population of bins will change over time and the choices made by our group and others may well differ.

We then explored parameters (AF cut-offs and effect of the mutation) used to select variants for further manual review (Figure 1). The total number of variants selected (Figure 1A) is decreased 10–20 fold using AF filters of <5% or <1% (Figure 1B). Selecting for protein-altering variants (missense, nonsense, frameshift, and splice site) further decreases this number (Figure 1C). However, the resulting numbers are still incompatible with the small chance of an individual having a Mendelian disorder; thus, the vast majority of variants with <5% AF must have minimal clinical consequences. When selecting only predicted truncating (nonsense, frameshift, and splice site) variants, the number identified per patient is more consistent with realistic expectations (Figure 1D).

Clearly, the sensitivity of the algorithm is decreased by the exclusion of rare missense mutations. To address this issue we queried a local instance of HGMD for variants in these genes annotated as “DM” and identified 871 unique variants (771 missense) among the 80 whole genome sequences. On average there were 74 (range 61–106) “DM” variants per person (Figure 2A), which is strikingly similar to the report of the 1000 Genomes Project Consortium that individuals are heterozygous for 50–100 variants classified as disease causing in HGMD.[9] Nevertheless, this large number of putatively disease-causing mutations is surprising, given the very low probability of a Mendelian disorder truly being present in any of the subjects.

Since 88% of the unique “DM” variants were missense substitutions, we hypothesized that these variants could comprise a subset of the ~150 missense variants per person identified in Bins 1, 2b, and 2c with <5% AF (Figure 2A). Surprisingly, there was minimal overlap between the less common missense variants and “DM” variants detected in the 80 genomes (Figure 2B), and upon further review, 251 of the 871 unique “DM” variants (29%) had >5% AF. As a result, 78% of “DM” variants per person were >5% AF (Figure 2C). This finding is similar to a previous report that 74% of HGMD variants identified in 448 genes implicated in severe recessive diseases of childhood were variants with >5% AF.[11] Although some of these variants could represent recessive alleles that are relatively frequent in certain populations, this explanation cannot account for the vast majority of these variants.

To further assess the pervasiveness of misleading database errors, we queried the 1000 Genomes Project allele frequencies and found allele frequencies for 1811 out of 74,694 “DM” variants (mostly substitution variants). Of these, 1152 had <1% AF, 299 had 1–3% AF, 95 had 3–5% AF, and 265 (~0.35% of all “DM” variants) had >5% AF (Figure 2D). The small subset of variants with >5% AF comprised the majority of “DM” variants identified in a given genome sequence, simply because of the prevalence of these variants in

the general population; in subsequent analyses we restricted HGMD variants to those with <5% AF.

The final algorithm selected variants according to the following criteria: 1) presence in a binned gene, 2) <5% AF, and either 3) annotation as a disease-causing mutation ("DM") in HGMD or 4) predicted to be truncating. Variants were further analyzed for zygosity to assign single heterozygous variants in recessive genes to a separate category for carrier status (Bin R). When we applied this algorithm to the 80 genomes, a total of 1391 variants (906 unique variants) were selected. The per-person averages were 1.5 variants in Bin 1 genes, 6.4 variants in Bin 2b genes, 0.2 variants in Bin 2c genes, and 9.2 variants considered to imply carrier status for recessive disorders (Table 1 and Supplemental Table S2).

The variants selected by the algorithm were then manually reviewed using a combination of OMIM, PubMed, Google Scholar, UCSC genome browser, and locus-specific databases to assess the evidence for pathogenicity or to reclassify the variants selected from the 80 genomes. Variants were reclassified if two variants identified in an individual likely comprised a single complex substitution allele or comprised a single common haplotype. In many cases, variants annotated as "DM" in HGMD were reclassified as VUS or likely polymorphisms. In other cases, the type of variant or its location within a specific transcript was inconsistent with a pathogenic effect. Zygosity was reassessed when it was determined that two variants were likely to be in *cis* or that only one of the selected variants in a gene was likely to be pathogenic; in these cases, the remaining heterozygous variant was reassigned to Bin R. Table 2 shows examples of binned variants, reclassified variants, and variants removed from consideration. Several detailed examples are described in the Supplemental Materials. A list of binned variants from the 61 publically available genomes is available in Supplemental Table S3.

After review, 705 variants were removed from consideration and 71 were reassigned to carrier status. Differing percentages of variants were reclassified or removed from consideration in each bin category (Figure 3A) and lower proportions of novel variants were removed (Figure 3B) compared to HGMD "DM" variants (Figure 3C). In all, 279 of the 358 unique variants removed from consideration were HGMD "DM" variants. After the final analysis, the revised per-person averages were 0.3 variants in Bin 1 genes, 2.6 variants in Bin 2b genes, 0.06 variants in Bin 2c genes, and 5.5 variants considered to imply carrier status (Table 1 and Supplemental Table S2).

DISCUSSION

One barrier to the clinical use of WGS/WES is the legitimate concern that the burden of incidental findings will be overwhelming and lead to expensive and unnecessary follow-up despite little evidence that such variants have a strong role in causing disease.[5,12] The approach we describe here demonstrates that analysis of WGS/WES data for clinically significant incidental variants is a tractable problem and that manageable numbers of variants can be selected for manual review.

Predictive value of variants identified in an incidental context

These results indicate that a small number of potentially disease-causing variants can be readily identified using a relatively straightforward process consisting of *a priori* gene classification, computational filtering and database queries. As with any medical test, the analytic parameters used in this approach represent a trade-off between sensitivity and specificity. The choices outlined in our strategy reflect the impact of sensitivity and specificity on the calculation of the negative predictive value (NPV) and positive predictive value (PPV). When the prior probability of disease is very low (eg. the chance of having a

Mendelian disorder that would be discovered incidentally), a test with reduced specificity will yield results with poor PPV, whereas reduced sensitivity has negligible effect on the NPV. We have therefore chosen to set a threshold that emphasizes specificity, in order to enrich for incidental findings that have a high likelihood of representing truly disease causing mutations.

Since selection of rare missense variants in known disease genes results in a large number of VUS, which provide no "actionable intelligence" for a clinician or patient, we excluded missense variants unless annotated as "DM" in HGMD. Various algorithms are used in research to predict the likely functional consequences of missense variants,[13] but these programs are not clinically validated[14] and in the absence of other supporting data they are generally insufficient to upgrade the status of a missense variant from VUS to likely pathogenic.[15] The proposed framework also excludes synonymous variants as well as variants in the untranslated portions of the transcript and introns, which are most likely benign, but might alter expression of the transcript or cause splicing abnormalities. Although the exclusion of novel missense, synonymous, and noncoding variants decreases the sensitivity of the approach, the lack of any clinically validated means of selecting the true positive mutations from among the numerous variants of unknown (or no) clinical significance requires that we sacrifice some sensitivity in order to maintain high specificity. Inclusion of the HGMD substantially increased the sensitivity of the algorithm, but misannotated HGMD "DM" variants (which could represent errors in the medical literature or database curation errors) still constituted a major source of false positive results.

Since there is no gold standard against which to compare our results, we cannot definitively estimate the clinical sensitivity or specificity of this analytic framework. However, even after manual inspection, the numbers of variants selected per person (Table 1) indicate that a number of false positives remain. Some of the putative mutations identified in these 80 genomes could reflect sequencing artifacts, which would be revealed by follow-up Sanger sequencing. Many of the "DM" mutations remaining after manual curation may still represent VUS, or represent the milder end of the genotype-phenotype spectrum for a given disease. Perhaps more intriguingly, these findings could indicate a much greater degree of clinical variability and incomplete penetrance than has previously been appreciated in Mendelian disorders, which could dramatically impact the logistics of return of such information clinically. We anticipate that improvements in both clinical databases and predictive algorithms will allow us to further improve sensitivity and specificity over time.

Comparison to other reports

The average numbers of potentially clinically important variants identified in this manuscript differ substantially from previous efforts to quantify the burden of clinically important incidental findings and we feel that it represents a more realistic picture of what to expect from WGS in terms of clinical yield. These differences hinge largely on the assumptions made about disease causation and the framework we have chosen for identification of potentially clinically relevant variants. For example, while other groups have been inclined to report[1] and/or interpret the possible clinical significance[2] of variants that may modify risks for common diseases, we intentionally ignored common SNPs that are weakly associated with multifactorial diseases. This decision is based on the lack of validated models for incorporating such information into medical care[6] and the inconsistent interpretive results obtained in different labs,[16] although the framework described here could be readily modified to include multifactorial risk calculations if warranted by advances in medical genetics and genomics. Pharmacogenomic variants can also be accommodated in the binning framework but were not considered here.

Cassa and colleagues estimated that individuals harbor ~2100 substitution variants that might need to be returned to research subjects,[17] which is four orders of magnitude higher than the 0–2 likely deleterious Bin 1 variants per person identified in this study. Possible explanations for this striking difference are the stringency with which genes are categorized as having clinical utility, and the thresholds for reporting variants. We argue that a relatively high evidentiary standard should be applied in order for a gene to be placed in Bin 1, such that the expected benefits gained by improved medical management would outweigh the possible harms that could arise from the revelation of such a finding in an incidental context. Using these criteria, most known disease genes are placed in Bin 2, in which patient choice is paramount in determining whether such incidental findings should be returned. In addition, we believe that only variants that are known to be pathogenic or highly likely to be pathogenic should be returned in an incidental context. The vast majority (~96%) of variants included in the Cassa et al. study originated from the HGMD, and our current data demonstrate that many of these variants are likely to represent false positives. It is difficult to discern how many of the ~2100 substitution variants per person reported by Cassa et al. are actually benign common polymorphisms, although ~1/3 of these variants were homozygous (suggesting a general population AF substantially greater than 5%), indicating that the putative "reportable" variants identified by Cassa et al. include many variants that are not deleterious and should not be reported either in a research context or a clinical context.

MacArthur and colleagues reported a survey of loss-of-function variants in the 1000 Genomes Project data and identified many challenges of interpreting WGS/WES data with respect to generating annotations and predicting the effects of possibly truncating variants.[18] A number of known and likely disease-causing loss of function mutations were identified among the subjects analyzed, most of which would represent carrier status for autosomal recessive disorders. Again, however, these results point out the difficulty of predicting pathogenicity of a given variant and the importance of review by a clinical molecular diagnostician. Similar to our results, one putative disease causing mutation listed among the loss-of-function variants by MacArthur et al. was a nonsense mutation in *LRRK2*, which is of uncertain clinical significance since the reported mutations in *LRRK2*-related autosomal dominant Parkinson's disease are missense substitutions.[19]

Challenges and Future Directions

The bin assignments described here should be viewed as a first step in the development of the binning process. The central concept of Bin 1 is that these findings have sufficient clinical actionability that no preference would be elicited regarding their return (in effect, the "duty to warn" would supersede the patient's autonomy). This denial of the patient's "right not to know" requires us to set a very high threshold regarding the types of findings that are appropriate for this category. On the other hand, our strategy places the majority of disease genes within Bin 2, where the potential risk for harm is the organizing principle, and the concept of individual preference is paramount. Thus, we feel that our strategy strikes a balance regarding patient choice and medical paternalism. A possible future addition might be to subcategorize Bin 2b into disease groups (such as cancer, cardiovascular/sudden death, neurodegenerative, and "other" Mendelian disorders) that would allow a more refined choice in a clinical context. Of course, the disadvantage of introducing more and more categories is that the clinical decision-making could devolve into a gene-by-gene menu, which would impose prohibitive demands on clinicians and laboratories with respect to informed consent and analysis.

This provisional binning of genes is not meant to represent a final or definitive list, and we expect that there will be disagreement among experts about the criteria that define Bin 1 or Bin 2 genes, or the types of incidental findings that should routinely be returned to patients

(and how they should be returned) during the course of a genome-scale diagnostic test.[20] Furthermore, there may be differences of opinion regarding the classes of variants that should be reported to patients when discovered incidentally. Our evolving understanding of the genetic underpinnings of disease will necessitate a flexible approach to the structured clinical analysis of genome sequences, and an important future direction will be to establish more granular criteria for determining the novel variants that are selected for review, based on the reported spectrum of disease-causing mutations. It is likely that the large numbers of genomes currently being sequenced worldwide will greatly facilitate the clinical interpretation of variants that are found in known disease genes. Better estimates of penetrance will inform the contexts in which certain variants are reported, and many variants previously reported as disease-causing may need to be carefully scrutinized to separate those that are truly deleterious from those that simply reflect normal population variation. Thus, the value of a centralized and rigorously maintained clinical-grade database containing known variants and their significance cannot be overstated.

CONCLUSIONS

These results represent a proof-of-concept demonstration of a structured clinical analysis of incidental findings in genome-scale sequence data that can serve as a general model for assessment of WGS/WES incidental findings. This framework makes the identification clinically relevant incidental findings much more tractable, as it reduces the number of variants requiring hand curation to a manageable number (10–20), and it should prove robust to differing bin structures or gene assignments. We expect that consensus will be possible regarding the Bin assignment of many genes,[20] and we note that as of this publication there are ongoing discussions and debate among genetics professionals regarding these issues. Advances in medical genetics will also mandate a periodic re-evaluation of these Bin assignments. Nevertheless, we anticipate that assignment of genes to bins based on clinical utility and stratified based on the risk of psychosocial harm will enable efficient analysis of data as well as facilitating pre-test informed consent, post-test counseling and return of results as we enter the era of clinical genomics. Further research on the implementation of this analytic framework and the responses of individuals to incidental findings is underway.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the University Cancer Research Fund (http://unclineberger.org/ucrf) and the UNC Bryson Philanthropic Fund. J.S.B. received funding from the NC TraCS Institute, supported by grants UL1RR025747, KL2RR025746, and TLRR025745 from the NIH National Center for Research Resources. 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 DA030976-01. 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 Project allele frequencies.

References

1. Lupski JR, Reid JG, Gonzaga-Jauregui C, et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med. 2010; 362:1181–1191. [PubMed: 20220177]
2. Ashley EA, Butte AJ, Wheeler MT, et al. Clinical assessment incorporating a personal genome. Lancet. 2010; 375:1525–1535. [PubMed: 20435227]

3. Worthey EA, Mayer AN, Syverson GD, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med. 2011; 13:255–262. [PubMed: 21173700]
4. Bick D, Dimmock D. Whole exome and whole genome sequencing. Curr Opin Pediatr. 2011; 23:594–600. [PubMed: 21881504]
5. Kohane IS, Masys DR, Altman RB. The incidentalome: a threat to genomic medicine. JAMA. 2006; 296:212–215. [PubMed: 16835427]
6. Manolio TA. Genomewide Association Studies and Assessment of the Risk of Disease. N Engl J Med. 2010; 363:166–176. [PubMed: 20647212]
7. Berg JS, Khoury MJ, Evans JP. Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. Genet Med. 2011; 13:499–504. [PubMed: 21558861]
8. Drmanac R, Sparks AB, Callow MJ, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010; 327:78–81. [PubMed: 19892942]
9. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
10. Stenson PD, Ball EV, Howells K, Phillips AD, Mort M, Cooper DN. The Human Gene Mutation Database: 2008 update. Genome Med. 2009; 1:13. [PubMed: 19348700]
11. Bell CJ, Dinwiddie DL, Miller NA, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med. 2011; 3:65ra4.
12. McGuire AL, Burke W. An unwelcome side effect of direct-to-consumer personal genome testing: raiding the medical commons. JAMA. 2008; 300:2669–2671. [PubMed: 19066388]
13. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet. 2006; 7:61–80. [PubMed: 16824020]
14. Tchernitchko D, Goossens M, Wajcman H. In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics. Clin Chem. 2004; 50:1974–1978. [PubMed: 15502081]
15. Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB. IARC Unclassified Genetic Variants Working Group. In silico analysis of missense substitutions using sequence-alignment based methods. Hum Mutat. 2008; 29:1327–1336. [PubMed: 18951440]
16. Ng PC, Murray SS, Levy S, Venter JC. An agenda for personalized medicine. Nature. 2009; 461:724–726. [PubMed: 19812653]
17. Cassa CA, Savage SK, Taylor PL, Green RC, McGuire AL, Mandl KD. Disclosing pathogenic genetic variants to research participants: Quantifying an emerging ethical responsibility. Genome Res. 2012; 22:421–428. [PubMed: 22147367]
18. MacArthur DG, Balasubramanian S, Frankish A, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012; 335:823–828. [PubMed: 22344438]
19. Dächsel JC, Farrer MJ. LRRK2 and Parkinson disease. Arch Neurol. 2010; 67:542–547. [PubMed: 20457952]
20. Green RC, Berg JS, Berry GT, et al. Exploring concordance and discordance for return of incidental findings from clinical sequencing. Genet Med. 2012; 14:405–410. [PubMed: 22422049]

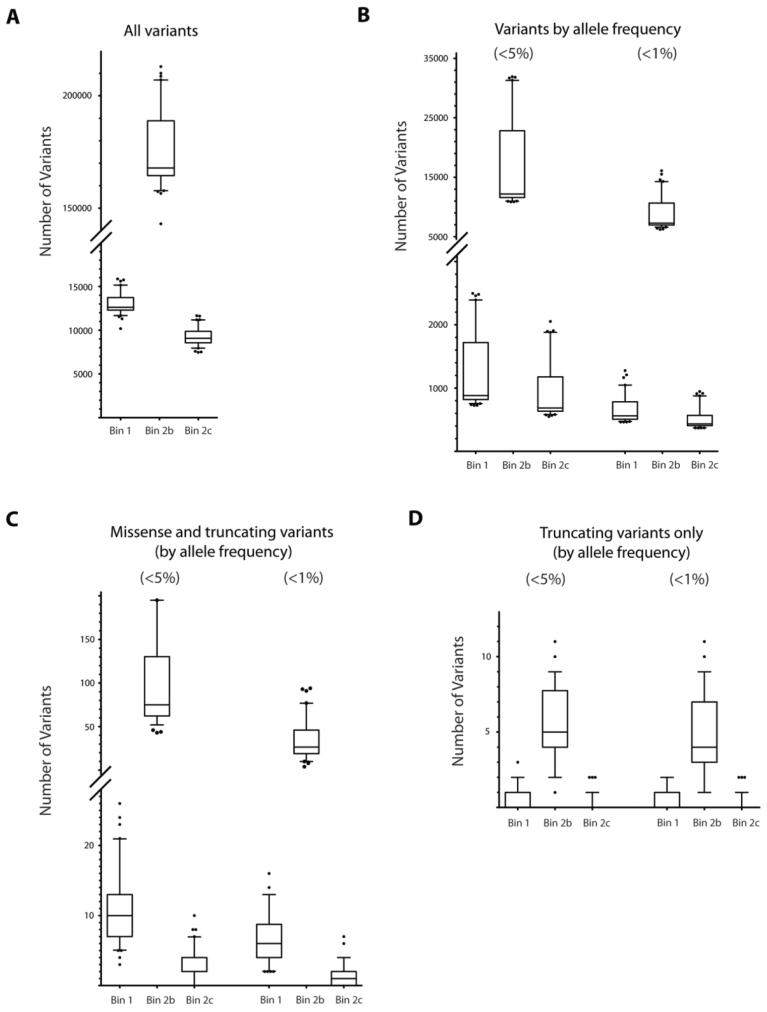

FIGURE 1. Selection of variants based on allele frequency and predicted effect on the translated protein

(A) The initial informatics analysis resulted in an average of ~13,000 variants per person in Bin 1 genes, ~175,000 variants per person in Bin 2b genes, and ~9000 variants per person in Bin 2c genes. (B) Limiting these variants to <5% AF or <1% AF reduces these counts ~10–15 fold, respectively. (C) Restricting to protein-coding variants (missense, nonsense, frameshift, splice site) at <5% AF results in ~10 variants per person in Bin 1 genes and 100–200 variants per person in Bin 2b genes. At <1% AF there were ~5 variants per person in bin 1 genes and 50–100 variants per person in Bin 2b genes. (D) Restricting only to truncating variants (nonsense, frameshift, splice site) results in only a small number of variants to be analyzed by the reviewer. Interestingly, the AF cut-off (<5% vs. <1%) does not dramatically affect the number of truncating variants that are selected.

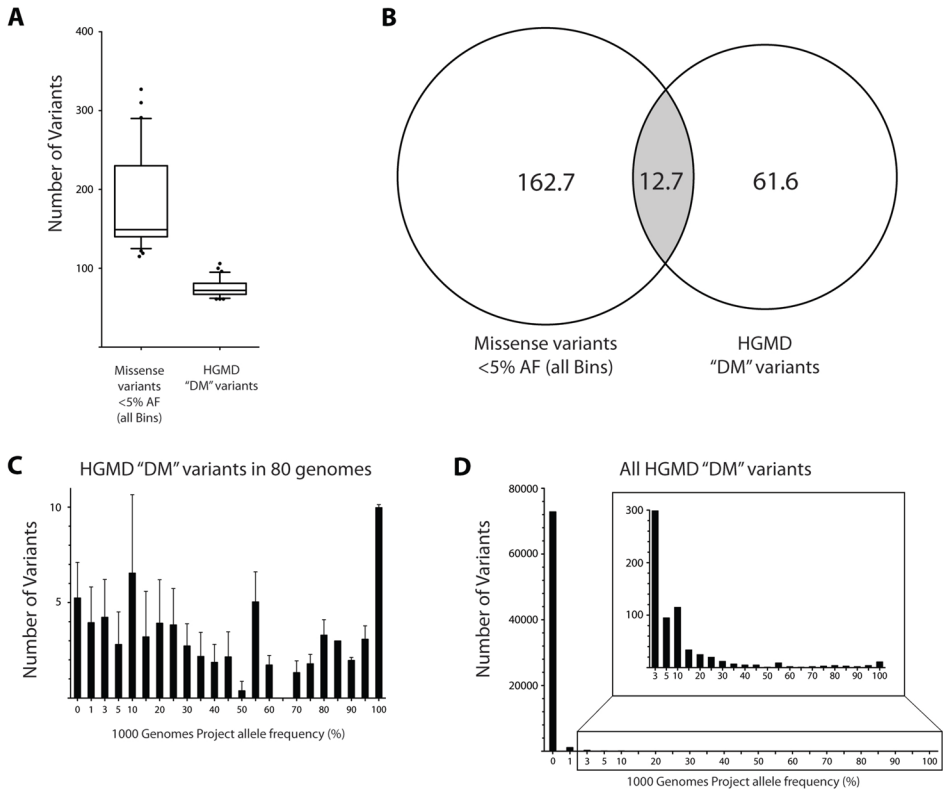

FIGURE 2. Analysis of mutations annotated as "DM" in HGMD

(A) All variants were queried against the HGMD to identify variants classified as "DM". The numbers of rare (<5% AF) missense variants in all bins and the numbers of HGMD "DM" variants per person are depicted as a box-whisker plot with whiskers indicating the 5^{th} and 95^{th} percentiles and outliers shown as filled circles. Homozygous variants are counted twice. (B) The overlap between the rare missense variants and "DM" variants is depicted as a Venn diagram. (C) The maximum 1000 Genomes allele frequencies were determined for each variant identified in the 80 whole genomes and histograms of allele frequencies were generated for each person. These histograms were then combined to depict the average number of variants per person within each range of allele frequencies (depicted as a bar plot with standard deviations). (D) The maximum 1000 Genomes allele frequencies were determined for all "DM" variants in HGMD and graphed as a histogram. The inset shows the distribution for "DM" variants with >1% allele frequencies.

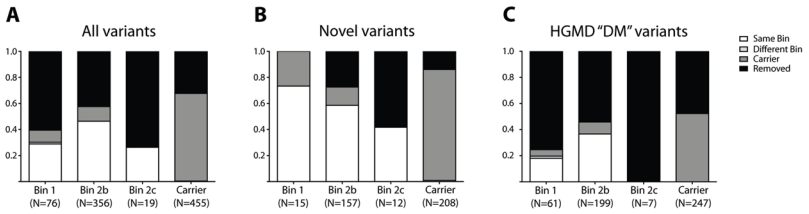

FIGURE 3. Results of the manual review of variants selected by the informatics algorithm

After individual review of the 906 unique variants returned by the final informatics algorithm, 45% were reassigned or removed from consideration. The graphs depict the variants initially selected within a given "bin" and the stacked segments represent the proportions of those variants that were confirmed, reassigned, or removed after review (see legend). Figure (A) shows all 906 unique variants, (B) shows the 392 rare truncating variants identified by the algorithm and (C) shows the 514 rare "DM" variants from HGMD. A higher proportion of "DM" variants in each bin category were removed from consideration compared to novel truncating variants.

Table 1

Numbers of variants selected by the informatics algorithm

	Bin 1	Bin 2b	Bin 2c	Bin R
Total variants per person	13129.7 (10268 – 15993)	174576.7 (144371 – 212760)	9251.6 (7472 – 11663)	N.D.
<5% AF	1219.8 (732 – 2532)	16362.1 (10845 – 31861)	915.5 (551 – 2053)	N.D.
<5% AF and either “DM” in HGMD or predicted truncating	3.0 (0 – 9)	14.2 (5 – 26)	0.45 (0 – 3)	N.D.
<5% AF and either “DM” in HGMD or predicted truncating, analyzed for zygosity	1.5 (0 – 5)	6.5 (2 – 14)	0.2 (0 – 2)	9.2 (0 – 17)
Revised after manual review	0.3 (0 – 2)	2.6 (0 – 8)	0.06 (0 – 1)	5.5 (0 – 12)

N.D. = not done

Table 2

Selected examples of selected variants, reclassified variants, and variants removed from consideration after human review

Subject	Gene (OMIM#)	NCBI 37 Location	Strand	Ref	Call	Impact	dbSNP	AF	HGMD	HGVS	Comments
Examples of predicted Bin 1 disease causing mutations											
NA18956	COL3A1 (120180)	chr2:189870930-189870930	+	−	insG	Splice site, frameshifting	N/A	N/A	N/A		Type IV Ehlers-Danlos syndrome (“vascular” subtype)
NA12877	FBN1 (134797)	chr15:48779379-48779381	−	GC	T	Frameshifting substitution	N/A	N/A	N/A		Marfan syndrome
NA18947	KCNH2 (152427)	chr7:150645539-150645540	−	G	A	Missense	N/A	N/A	CM085481	NM_000238.2c.2684C>T (T895M)	Long QT syndrome 2
NA12883	MSH2 (609309)	chr2:47637458-47637458	+	−	insG	Frameshifting insertion	rs63750786	N/A	CI041960	NM_000251.1c.592dupG	Lynch syndrome
Examples of predicted Bin 2c disease causing mutations											
NA18505	ITPR1 (147265)	chr3:4777025-4777025	+	−	insC	Frameshifting insertion	N/A	N/A	N/A		Spinocerebellar ataxia 15; alternatively spliced exon
NA12881 NA19025	PRKCG (176980)	chr19:54410082-54410084	+	GA	delGA	Frameshifting deletion	N/A	N/A	N/A		Spinocerebellar ataxia 14
Examples of variants reclassified to carrier status											
NA18504	MEFV (608107)	chr16:3299467-3299468	−	C	T	Missense	rs11466024	0.024	CM990838	NM_000243.1c.1223G>A (R408Q)	AR Familial Mediterranean Fever; R408Q and P369S reported in cis as single allele with highly variable clinical phenotype[11]
		chr163299585-3299586	−	G	A	Missense	rs11466023	0.041	CM990837	NM_000243.1c.1105C>T (P369S)	
NA18947	SLC34A1 (182309)	chr5:176815191-176815192	+	T	G	Splice site disrupt	N/A	N/A	N/A		AD Nephrolithiasis (heterozygous missense mutations); AR Fanconi renotubular syndrome (homozygous or compound heterozygous null mutations)
Examples of variants removed from consideration											
NA19238	APC (611731)	chr5:112173898-112173899	+	C	T	Missense	rs33974176	0.037	CM080070	NM_000038.3c.2608C>T (P870S)	AD Familial adenomatous polyposis; likely polymorphism
NA18505 NA19020 NA19025 NA19700	ATM (607585)	chr11:108121732-108121733	+	G	A	Missense	rs2235000	0.041	CM024583	NM_000051.3.1541G>A (G514D)	AR Ataxia telangiectasia; Presence of G514D and H1380Y in four unrelated individuals of African origin suggests a single allele with variants in cis; likely polymorphism
		chr11:108159731-108159732	+	C	T	Missense	rs3092856	0.044	CM021944	NM_000051.3c.4138C>T (H1380Y)	
NA19025 NA19239 NA19240	BRCA2 (600185)	chr13:32914838-32914839	+	A	G	Missense	rs55953736	0.004	CM022331	NM_000059.3c.6347A>G (H2116R)	AD Hereditary breast and ovarian cancer susceptibility; variant of uncertain clinical significance
NA18502	LRRK2 (609007)	chr12:40704361-40704362	+	C	T	Nonsense	rs114908017	0.001	N/A		AD Parkinsonism; variant of uncertain clinical significance since disease-causing mutations are typically missense
NA12877	HTT (613004)	chr4:3127346-3127347	+	G	delG	Frameshifting deletion	N/A	N/A	N/A		AD Huntington; variant of uncertain clinical significance since disease-causing mutations are typically triplet repeat expansions
NA12878	KIF1B	chr1:10425554-10425554	+	−	insC	Frameshifting insertion	N/A	N/A	N/A		AD CMT2A1; sequential variants that likely return to correct reading frame, resulting in two amino acid missense alterations
	KIF1B	chr1:10425557-10425558	+	G	delG	Frameshifting deletion	N/A	N/A	N/A		

AF = Highest minor AF among 1000 Genomes Project populations

HGVS = Human Gene Variation Society nomenclature

N/A = Not available

AD = Autosomal dominant

AR = Autosomal recessive