SPECIAL ARTICLE | Genetics inMedicine

Taxonomizing, sizing, and overcoming the incidentalome

Isaac S. Kohane, MD, PhD[1,2], Michael Hsing, PhD[1,2] and Sek Won Kong, MD[1,2]

Purpose: With the advent of whole-genome sequencing made clinically available, the number of incidental findings is likely to rise. The false-positive incidental findings are of particular clinical concern. We provide estimates on the size of these false-positive findings and classify them into four broad categories.

Methods: Whole-genome sequences (WGS) of nine individuals were scanned with several comprehensive public annotation databases and average estimates for the number of findings. These estimates were then evaluated in the perspective of various sources of false-positive annotation errors.

Results: At present there are four main sources of false-positive incidental findings: erroneous annotations, sequencing error, incorrect penetrance estimates, and multiple hypothesis testing. Of these, the first two are likely to be addressed in the near term. Conservatively, current methods deliver hundreds of false-positive incidental findings per individual.

Conclusion: The burden of false-positives in whole-genome sequence interpretation threatens current capabilities to deliver clinical-grade whole-genome clinical interpretation. A new generation of population studies and retooling of the clinical decision-support approach will be required to overcome this threat.

Genet Med 2012:14(4):399–404

Key Words: clinical interpretation; false-positives; incidental findings; whole-genome sequencing

INTRODUCTION

In the spring of 2008, 19 years after the start of the Human Genome Project, a publication described how James Watson's DNA had been fully sequenced for under 0.1% of the cost of the human genome using "next generation sequencing."[1] Less remarked at that time were the contents of **Table 3** of that publication listing the variants found in Dr Watson's genome that had been classically described as causing congenital diseases with Mendelian inheritance. Specifically, there were two variants for which he was homozygous. In other individuals, these two variants had previously been documented to be causal of Usher syndrome 1b (OMIM no. 276900) and Cockayne syndrome (OMIM no. 133540), diseases presenting typically at birth or early childhood. It seems very unlikely, based on what is publicly known, that Dr Watson suffers from these. Therefore, the publication of his genome might be regarded as a final warning of the deluge to come of incidental findings in genome-scale investigations—a downpour we have termed the incidentalome.[2] Even now, the number of false-positive findings is growing[3–5] and with the near-term availability of whole-genome sequencing for clinical diagnostics, these are likely to grow into a very large incidentalome. The scope of incidental findings addressed in this article is both broader and narrower than that defined by Wolf et al.[6] Of the set of findings "concerning an individual research participant that [have] potential health or reproductive importance and [are] discovered in the course of conducting research but [are] beyond the aims of the study" we focus on the false-positive incidental findings. False-positive incidental findings provide misleading and/or incorrect diagnostic or prognostic information and are, therefore, the most pernicious of the incidental findings; we focus exclusively on these. However, as we anticipate that whole-genome sequencing will become adopted in health-care delivery, the false-positive incidental findings obtained during a clinical care episode will also mushroom and, therefore, we include these incidental findings in the scope of this study.

The incidentalome can be taxonomized into four components. In order of increasing challenge, there is first, the substantial proportion of "textbook cases" of mutations documented to cause human disease in a highly penetrant Mendelian fashion, but they are incorrectly annotated in the databases. The second is the technical or measurement error rate in genome-scale sequencing. Third is the incorrect assignment of prior probabilities for much of our genetic and genomic knowledge. The fourth derives from testing multiple hypotheses across millions of variants. We will describe here the nature of these components, provide rough estimates for the magnitude of the problem, and point out existing approaches that will serve to control the growth of these aspects of the incidentalome. First, however, it is helpful to understand the magnitude of the interpretive challenge and the risks of false-positives by performing an example analysis of a whole-genome sequence (WGS) using the genomes of nine individuals of European descent sequenced by Complete Genomics.

[1]Children's Hospital Informatics Program, Children's Hospital, Boston, Massachusetts, USA; [2]Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA. Correspondence: Isaac S. Kohane (isaac_kohane@harvard.edu)

Submitted 6 September 2011; accepted 7 December 2011; advance online publication 9 February 2012. doi:10.1038/gim.2011.68

MATERIALS AND METHODS

Subjects and WGSs

We utilized the data set of publicly available genome sequences from nine unrelated HapMap individuals with European ancestry (NA06985, NA06994, NA07357, NA10851, NA12004, NA12889, NA12890, NA12891, and NA12892). These genomes were sequenced using a sequencing-by-ligation method at Complete Genomics.[7] The sequences were downloaded from http://www.completegenomics.com/sequence-data/download-data/ and imported into our in-house database for annotation and filtering as described later. The sequencing depths ranged from 64× to 88×. Each of the nine genomes was found to contain a total of 3.6–3.9 million genomic variants, the majority of which are single-nucleotide polymorphisms, with the other 12% consisting of insertions, deletions, or multiple-base substitutions. Among all the variants, ~0.6% were located within coding sequences.

Annotation and filtering of genomic variants

We developed a WGS analysis pipeline to annotate, filter, and analyze all the genomic variants presented in WGS. The pipeline, using MySQL database and PERL scripts, will be publicly available as a Web tool (manuscript in preparation). The pipeline focuses on two major annotation modules: (i) allele frequency (AF) recalculated from multiple large data sources, and (ii) functional impact estimation based on protein-coding genes and evolutionary sequence conservation. The combination of these two annotation modules allowed variant filtering and gene selection in the subsequent steps.

Each variant was annotated by the AFs calculated from three large database sources, the Single Nucleotide Polymorphism database (dbSNP) build 132,[8] the 1000 Genomes Project,[9] and 200 Exomes.[10] A total of 180 populations with European ancestry where each data set had ≥15 individuals were selected from dbSNP build 132 for AF calculations. If the same allele was reported from multiple populations, the AF from the largest population was used. The genotypes of 629 individuals were obtained from the 1000 Genomes Project website (November 2010 release). The VCF tool (version 0.1.4a) was used to calculate AFs from 261 of those 629 individuals with European ancestries including Utah residents with Northern and Western European ancestry, Finnish in Finland, Toscani in Italy, and British in England and Scotland.[11] The indel calls from the European subset of the same 629 individuals were obtained from the 1000 Genomes Project website (February 2011 release). In addition, AFs based on 200 Exomes of Danes were obtained from the SOAPsnp website.[10] To characterize the AFs, the four categories "common," "less common," "rare," and "novel" were used.[12] A "common" variant was defined by an AF ≥5% from any of the aforementioned three sources, and a "less common" variant was defined as an AF between 1% and 5%. A "rare" variant was the one present in any of the three sources but with <1% AF from all the three, and a novel variant was defined by its absence from all of the three sources.

Annotations of all variants in known protein-coding genes were based on the RefSeq gene model (March 2011 release for hg18). By comparing each variant sequence with the canonical transcript sequences from the RefSeq, variants with (i) synonymous, (ii) missense, (iii) insertion, (iv) deletion, (v) frameshift, (vi) nonsense, (vii) nonstop, (viii) misstart, or (ix) disruptive (at splice sites) impacts were identified. The variants of type ii–ix are considered nonsynonymous. The functional impacts of missense variants on proteins were obtained from dbNSFP,[13] which precomputed scale-invariant feature transform[14] and PolyPhen-2[15] scores on 75,931,005 possible nonsynonymous single-nucleotide polymorphisms based on CCDS, version 20090327. A weighted-voting method of the Condel was used to derive weighed average scores combining both scale-invariant feature transform and PolyPhen-2.[16]

The sequence conservation was estimated based on the Genomic Evolutionary Rate Profiling (GERP) scores that were calculated based on sequence alignments of up to 30 other mammals for each locus in hg18.[17] A higher GERP score indicates an evolutionarily conserved locus. The GERP scores for single-nucleotide polymorphisms in the nine genomes were obtained by mapping their genomic coordinates with those in the GERP tables. For indels and substitutions, the average GERP scores were calculated from the bases between the start and the end of each variant. A locus was considered "highly conserved" for the subsequent analysis if the GERP score was >2. The known genomic variants associated with human diseases were as indicated in the SafeGenes database,[18] which integrates annotations from the Human Gene Mutation Database,[19] OMIM,[20] genome-wide association studies, Pharmacogenetics Knowledge Base,[21] and dbSNP.[8]

After the aforementioned annotation steps, variants were filtered based on a combination of criteria that include (i) rare or novel, (ii) nonsynonymous, (iii) located at highly conserved loci, (iv) deleterious on protein function, (v) homozygous, and (vi) disease association. The number of protein-coding genes containing variants that met the criteria was reported.

RESULTS

The results of filtering all the known variants of the nine individuals' genomes by various annotations and filters available are shown in **Table 1** (see "Materials and Methods" section). For example, of the 3.8 million variants with respect to a reference genome found per individual, 3.1 million variants are known common variants, and 0.6 million variants are rare or novel variants. At the gene level, 400 genes per individual have rare or novel nonsynonymous variants at conserved loci. Of those, 136 are predicted to be deleterious. Also, 55 genes per individual have an average of 59 homozygous variants that are annotated as having an association, causal or purely statistical with diseases. Also, 65 genes presented with rare/novel nonsynonymous variants at conserved loci across all nine genomes we analyzed, as shown in **Figure 1**. It may be that many, perhaps most of these findings will be true-positive incidental findings; however, as described in the following

Table 1 Application of comprehensive annotation filters on nine putatively "normal" individuals of European descent with full-genome sequencing by Complete Genomics

Filtering	Number of genes/ individual	Number of variants/ individual
Genes with rare/novel, nonsynonymous mutations at highly conserved loci	400 ± 10	455 ± 9
Genes with rare/novel, nonsynonymous mutations at highly conserved loci and predicted to be deleterious	136 ± 6	147 ± 6
Genes with mutations implicated in disease	199 ± 3	226 ± 4
Genes with homozygous mutations implicated in disease	55 ± 2	59 ± 2
Genes with rare/novel mutations implicated in disease	3 ± 0	4 ± 0
Genes with rare/novel mutations implicated in disease and predicted to be deleterious	2 ± 0	2 ± 0

Shown here are the average numbers of genes per individual meeting the criteria specified in the table header/footnote and in the rightmost column the average number of variants per individual (i.e., there may be multiple variants per gene) meeting the criteria.

Nonsynonymous: (impact based on the RefSeq gene model), includes mutations that cause disruption, frameshift, in-frame deletion, in-frame insertion, missense, misstart, nonsense, and nonstop.

Conserved loci are those with a Genomic Evolutionary Rate Profiling score >2 (ref. 17). Deleterious mutations are those computationally predicted to affect the phenotype based on the consensus program Condel.[16] Disease association is based on annotations compiled in the SafeGenes database that includes the Human Gene Mutation Database, Online Mendelian Inheritance in Man (OMIM), genome-wide association studies, Pharmacogenetics Knowledge Base, and Single Nucleotide Polymorphism database annotations.[18]

there is good reason to believe that many of them are false-positive incidental findings.

DISCUSSION

Inaccurate variant annotations

A recent study[18] of all the mutations listed in several databases, including the OMIM,[20] Human Gene Mutation Database,[19] Pharmacogenetics Knowledge Base,[21] and dbSNP,[8] found that the frequency of unresolved mutation annotations varied widely among the databases, ranging from 4 to 23%. In these instances neither the reference nor the mutated sequence was present at the specified location of the genome. This relatively large number of errors is explained by a small number of annotation errors over the past few years and by a much larger number of coordinates and/or varied mutation descriptions that do not mesh with the modern state-of-the-art map of the genome. This is because many of these variants were discovered considerably before the first human genome map draft was even assembled. As a result, these early annotations are inaccurate; however, given the rarity of these variants, a large proportion of them remain referred to in this fashion. The magnitude of this component of the incidentalome is not likely to grow due to the increasingly rigorous adoption of nomenclature standards and specific reference to genome "build" versions in the annotation reports. Moreover, the emergence of several for- and nonprofit international efforts[22,23] to standardize the clinical annotations of the genome suggests that this aspect of the incidentalome will be soon resolved.

Sequencing errors

The accuracies of ultra-high-throughput genome sequencing are reported to range from 95% to 99.9%.[24–27] This level of performance is a technological *tour de force* given the millions of short reads of DNA that have to be assembled to obtain a WGS. Nonetheless, even with 99.9% accuracy, across a billion bases, this entails up to one million technical errors. It also appears that some of these errors are nonrandom and reflect particularities of the sequence context of specific parts of the genome.[24] Therefore, it is not surprising that one individual sequenced on two different sequencing platforms should appear different from the genomics perspective.[28] Further, let us presume that as in **Table 1** there are 455 variants across 400 genes found to be nonsynonymous, rare, and at highly conserved loci. Even with a generous estimate of global accuracy of 99%, whole-genome sequencing will nonetheless result in 4–5 variants erroneously reported. Fortunately, it seems likely that this aspect of the incidentalome will shrink in the near future. Whether through standardization in the follow-up of putative positives using alternative measurement means such as Sanger sequencing or genotyping of alleles, or through use of alternative "baits"[29] and primers for particularly problematic regions, or through continued rapid-paced advances and improvements in performance of the sequencing technology itself, sequencing error rates will drop dramatically. Nonetheless, until then, sequencing error is likely to be a significant contributor to the incidentalome.

Effect of genetic background and environment on penetrance

Fundamentally, the utility of the clinical annotation of a genomic variant is only as useful as its applicability to a patient. That is, if a variant were found to track with a disease in a specified group of patients, that annotation may in fact serve well if one belongs to that specific group of patients but serve rather poorly if one does not. A classic example of this is in hemochromatosis, in which >80% of individuals within a hereditary hemochromatosis clinic will have one of the known variants in the *HFE* gene.[30] However, if one tests the general population as was done with over 40,000 patients at Kaiser Permanente,[3] <1% of individuals who are homozygotes with the very same variants found in the hereditary hemochromatosis clinic will show clinical, biochemical, historical, or familial evidence of hemochromatosis. Why the discrepancy between <1% and 80%? Presumably, this is because the individuals for whom this disease was clinically ascertained (through history and physical exam, family history, or routine clinical laboratory tests) had shared genetic background

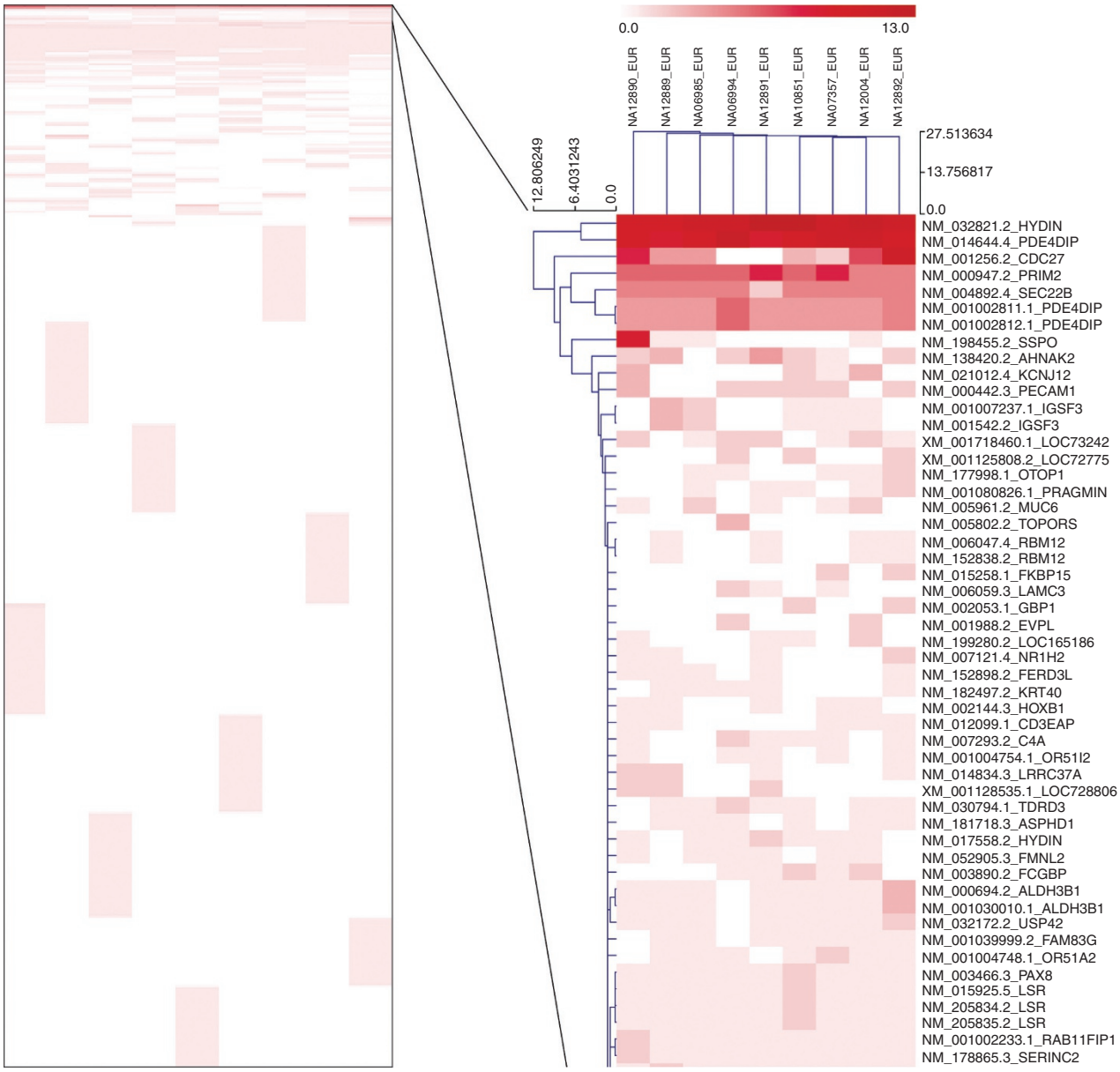

Figure 1 Overview of the genes with rare or novel nonsynonymous variants at conserved loci. Each row in the heatmap represents a unique transcript—a unique subsequence of the genome obtained during sequencing. The number of variants meeting the criteria of being nonsynonymous, rare, and conserved (see **Table 1** for definitions) was used to cluster transcripts and individuals. These numbers ranged from 0 (white) to the maximum of 13 (dark red) as shown in the color bar on the top right. Most variants were unique to each individual genome. Each individual presents an average of 222 such variants (range from 168 to 260) that are shown as red blocks on the heatmap (left). A total of 65 transcripts had more than one variant in all nine genomes. The top part of the heatmap was zoomed in on the right, which reveals that the genes such as *HYDIN, PDE4DIP, PRIM2*, and *SEC22B* tend to have more than one rare/novel nonsynonymous variant at a conserved locus, consistently in all genomes analyzed. The identification of these "hypervariable" genes can help to reduce the false-positive findings, but even a residual small false-positive rate (e.g., 1%) will incur a substantial population-wide reporting burden, as described in the text.

and/or shared environmental exposures. The effect of genetic background on penetrance (defined here as the conditional probability of disease given a genetic variant) has been well documented in mouse models where a disruption of a gene (e.g., *HFE*) will have significant effects in one strain and not in another.[31] With respect to hemochromatosis, the effect of environmental exposure on mutation effect also appears to be significant, such as the increased risk that comes from excessive alcohol consumption.[32] Similar phenomena explain why the reported penetrance of the *BRCA1/BRCA2* gene mutations has decreased markedly since the original publications[33,34] whereas the population to which this test has been applied has broadened. If the subject does not correspond well to the group studied for the originally reported finding, then these erroneous incidental findings are to be expected. This is particularly problematic as most of the mutations documented to

be highly penetrant (i.e., classical Mendelian genetics) are rare and have been found in a few families and the broader populations have not been genotyped for these variants. Therefore, of the hundreds of thousands of published disease-associated variants, an unknown but potentially large proportion will have a very different interpretation when applied to the general population. This challenge is particularly marked when the subject comes from a different ethnicity than the general population.[28] The most direct path to addressing this component of the incidentalome is the commoditization and subsequent widespread application of whole-genome sequencing to large populations.[35] Particularly, if these sequences are linked to detailed clinical phenotypes (e.g., from the electronic medical record[36]), we will have for the first time empirical estimates of the frequency of a large swath of mutations in clinical subpopulations of interest and in the general population and thereby will be able to accurately estimate penetrance for those populations. That is, we will be able to calculate data-driven positive predictive values, as we do for many clinical laboratory tests. This in turn will reduce the frequency with which these variants are falsely reported to increase the probability of a trait/disease. An immediate but controversial alternative would be to only perform (or report on) genetic variants when there is a clinical suspicion of the disease, whether through family history or clinical findings.

Genomic individualization reduces the availability of relevant comparison groups

We just described the challenge of comparing an individual to a group for which there is a known annotation that links a variant to disease. What about looking at all the individual's variants? Even if we obtain population-wide priors through extensive full-genome sequencing of entire populations, the very fact of a person's individuality when considering all variants in the genome will ensure that no comparison to any particular group will be perfectly appropriate. That is, when we perform multiple comparisons, as we will, when assessing each of hundreds of thousands to millions of variants for their clinical significance, we apply the knowledge of the meaning of each variant (i.e., the conditional probability of disease given the variant) with respect to a specific population that may or may not resemble the subject. If we treat each of these genetic variant–disease relationships as independent, then we should not be surprised that if we test each variant with 100% sensitivity and 99.9% specificity, merely 10,000 independent genetic variants associated with rare diseases will lead to more than half the entire population labeled with a false-positive risk/diagnosis.[2] The problems for clinical care then may well dwarf those entailed by incidental findings in research. As we originally articulated in our 2006 publication,[2] this proportion of false-positives will not only lead to concern and frustration on the part of consumers and health-care providers but ultimately lead insurers who are already reluctant to pay for genetic testing to object to payments for follow-up tests and investigations driven by such false-positive-finding-saturated approaches.

A purely statistical remedy to this multiple hypothesis–testing problem appears elusive. Even if an individual's phenotype and health state were fully determined by genetics, we could not expect to determine which of these multiple comparisons were most appropriate. How would we know which subset of genetic variants made one individual similar to a group of interest (e.g., a group with a specific disease)? Consider the simplest problem of determining which group an individual is most similar to, not across all variants but just pairs of them. Even if the entire global human population (7×10^9) were fully sequenced, this would be woefully inadequate to assess the relationship of all the pairs (10^{10}) of the one hundred thousand variants associated with disease. This suggests that our burgeoning but still fragmentary knowledge of molecular biology and the systems structure of genetic regulation will be required to overcome the multiple hypothesis–testing problem. In the interim, we may find that we have to purposely ignore most of the variants in the genome, focusing on small combinations of those that have the largest effects. In doing so, we will be recognizing what was realized early in the era of automated decision-making, long before the Human Genome Project: purely probabilistic reasoning approaches are too data hungry for even small clinical decision-making challenges.[37] When the probabilistic approach alone is inadequate, expert clinicians will complement probabilistic assessments with categorical and heuristic reasoning based on an assessment of how the patient fits known groups of pathophysiology based on their understanding of the patient's state and their understanding of physiology and its various pathobiologies. In this context, it would be quite ironic if one of the consequences of the genomic revolution and the surfeit of variables available to characterize patients would result in the revalorization of clinicians with deep pathophysiological knowledge and deep evidence-based expertise who are current with genomic literature and/or the genomic-database equivalents.

Conclusion

Whole-genome sequencing as it is performed today has a substantial burden of incidental findings that falsely report on the present or future state of the individual. As described earlier, the sources of these incidental findings—erroneous disease annotations of the genome, sequencing error, incorrect estimates of penetrance, and genomic individuality—can be addressed systematically over the coming years. At present, an unmediated, wholesale release of incidental finding interpretations to study subjects or patients[38,39] has clear and imminent potential for harm. We have argued previously[40] that interposing an expert body, informed by an understanding of the current accuracy of genome-scale testing, relevance of specific results, knowledge of patient-specific characteristics,[41] and respect for patient privacy and autonomy is required to safely communicate genome-scale interpretations. If, as we anticipate, the magnitude of false-positive incidental findings is reduced to a level that can be managed by well-trained clinicians and mechanisms are provided for patients and subjects alike to readily obtain additional clarification and personalized decision support, then the need

for mediated release of genomic incidental findings will correspondingly diminish.

ACKNOWLEDGMENTS

I.S.K. was funded in part by the NLM grant no. 5R01LM010125-02 and a presentation at the Incidental Findings conference was supported by National Institutes of Health, National Human Genome Research Institute grant no. 2-R01-HG003178 on "Managing Incidental Findings and Research Results in Genomic Biobanks & Archives" (S. Wolf, principal investigator). The authors are indebted to David Margulies, Alal Eran, Alvin Kho, and Arjun Manrai for their insightful comments.

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

1. Wheeler DA, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 2008;452:872–876.
2. Kohane IS, Masys DR, Altman RB. The incidentalome: a threat to genomic medicine. *JAMA* 2006;296:212–215.
3. Beutler E, Felitti VJ, Koziol JA, Ho NJ, Gelbart T. Penetrance of 845G–> A (C282Y) HFE hereditary haemochromatosis mutation in the USA. *Lancet* 2002;359:211–218.
4. Beutler E. Carrier screening for Gaucher disease: more harm than good? *JAMA* 2007;298:1329–1331.
5. Ng PC, Murray SS, Levy S, Venter JC. An agenda for personalized medicine. *Nature* 2009;461:724–726.
6. Wolf SM, Lawrenz FP, Nelson CA, et al. Managing incidental findings in human subjects research: analysis and recommendations. *J Law Med Ethics*. 2008;36(2):219–248, 211.
7. 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.
8. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;29:308–311.
9. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010;467(7319):1061–1073.
10. Li Y, Vinckenbosch N, Tian G, et al. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* 2010;42:969–972.
11. Danecek P, Auton A, Abecasis G, et al.; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics* 2011;27:2156–2158.
12. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010;11:415–425.
13. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* 2011;32:894–899.
14. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;4:1073–1081.
15. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7:248–249.
16. González-Pérez A, López-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* 2011;88:440–449.
17. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A; NISC Comparative Sequencing Program. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 2005;15:901–913.
18. Tong MY, Cassa CA, Kohane IS. Automated validation of genetic variants from large databases: ensuring that variant references refer to the same genomic locations. *Bioinformatics* 2011:27(6);891–893.
19. Cooper DN, Stenson PD, Chuzhanova NA. The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms. *Curr Protoc Bioinformatics* 2006;Chapter 1:Unit 1 13.
20. Pearson P, Francomano C, Foster P, Bocchini C, Li P, McKusick V. The status of online Mendelian inheritance in man (OMIM) medio 1994. *Nucleic Acids Res* 1994;22:3470–3473.
21. Hewett M, Oliver DE, Rubin DL, et al. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res* 2002;30:163–165.
22. Howard HJ, Horaitis O, Cotton RG, et al. The Human Variome Project (HVP) 2009 Forum "Towards Establishing Standards". *Hum Mutat* 2010;31(3):366–367.
23. Stokes T. Businesses ready whole-genome analysis services for researchers. *Nat Med* 2011;17:1161.
24. Harismendy O, Ng PC, Strausberg RL, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* 2009;10:R32.
25. Benaglio P, Rivolta C. Ultra high throughput sequencing in human DNA variation detection: a comparative study on the NDUFA3-PRPF31 region. *PLoS ONE* 2010;5:e13071.
26. Zhang W, Chen J, Yang Y, Tang Y, Shang J, Shen B. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PLoS ONE* 2011;6:e17915.
27. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;456:53–59.
28. Moore B, Hu H, Singleton M, Reese MG, De La Vega FM, Yandell M. Global analysis of disease-related DNA sequence variation in 10 healthy individuals: implications for whole genome-based clinical diagnostics. *Genet Med* 2011;13:210–217.
29. Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 2009;27:182–189.
30. Beutler E, Gelbart T, West C, et al. Mutation analysis in hereditary hemochromatosis. *Blood Cells Mol Dis* 1996;22:187–194; discussion 194a.
31. Fleming RE, Holden CC, Tomatsu S, et al. Mouse strain differences determine severity of iron accumulation in Hfe knockout model of hereditary hemochromatosis. *Proc Natl Acad Sci USA* 2001;98:2707–2711.
32. Scotet V, Mérour MC, Mercier AY, et al. Hereditary hemochromatosis: effect of excessive alcohol consumption on disease expression in patients homozygous for the C282Y mutation. *Am J Epidemiol* 2003;158:129–134.
33. Begg CB, Haile RW, Borg A, et al. Variation of breast cancer risk among BRCA1/2 carriers. *JAMA* 2008;299:194–201.
34. Domchek SM. Refining BRCA1 and BRCA2 penetrance estimates in the clinic. *Current Breast Cancer Reports* 2009;1:127–130.
35. 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.
36. Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet* 2011;12:417–428.
37. Szolovits P, Pauker SG. Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence in Medicine* 1978;11:115–144.
38. Altman RB. Direct-to-consumer genetic testing: failure is not an option. *Clin Pharmacol Ther* 2009;86:15–17.
39. Church GM. The personal genome project. *Mol Syst Biol* 2005;1:2005.0030.
40. Kohane IS, Mandl KD, Taylor PL, Holm IA, Nigrin DJ, Kunkel LM. Medicine. Reestablishing the researcher-patient compact. *Science* 2007;316:836–837.
41. Kohane IS, Taylor PL. Multidimensional results reporting to participants in genomic studies: getting it right. *Sci Transl Med* 2010;2(37):37cm19.