Genome Medicine

MUSINGS

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis*

Having recently attended the Personal Genomes meeting at Cold Spring Harbor Laboratories (I was an organizer this year), I was struck by the number of talks that described the use of whole-genome sequencing and analysis to reveal the genetic basis of disease in patients. These patients included a child with irritable bowel disease, a child with severe combined immunodeficiency, two siblings affected with Miller syndrome, and several with cancers of different types. Although each presenter emphasized the rapidity with which these data can now be generated using next-generation sequencing instruments, they also listed the large number of people involved in the analysis of these datasets. The required expertise to 'solve' each case included molecular and computational biologists, geneticists, pathologists and physicians with exquisite knowledge of the disease and of treatment modalities, research nurses, genetic counselors, and IT and systems support specialists, among others. While much of the attendant effort was focused on the absolute importance of obtaining the correct diagnosis, the large number of specialists was critical for the completion of the data analysis, the annotation of variants, the interpretive 'filtering' necessary to deduce the causative or 'actionable' variants, the clinical verification of these variants, and the communication of results and their ramifications to the treating physician, and ultimately to the patient. At the end of the day, although the idea of clinical whole-genome sequencing for diagnosis is exciting and potentially life-changing for these patients, one does wonder how, in the clinical translation required for this practice to become commonplace, such a 'dream team' of specialists would be assembled for each case. In other words, even if the cost and speed of generating sequencing data continue their precipitous decreases, the cost of 'team' analysis seems unlikely to immediately follow suit. However, rather than predicting from this reasoning that widespread diagnosis by sequencing is unlikely to occur widely, it is perhaps more fruitful to predict, in my opinion, what is probably required for it to occur. I therefore offer the following as food for thought.

One source of difficulty in using resequencing approaches for diagnosis centers on the need to improve the quality and completeness of the human reference genome. In terms of quality, it is clear that the clone-based methods used to map, assign a minimal tiling path, and sequence the human reference genome did not yield a properly assembled or contiguous sequence equally across all loci. Lack of proper assembly is often due to collapsing of sequence within repetitive regions, such as segmental duplications, wherein genes can be found once the correct clones are identified and sequenced. At some loci, the current reference contains a single nucleotide polymorphism (SNP) that occurs at the minor allele frequency rather than being the major allele. In addition, some loci cannot be represented by a single tiling path and require multiple clone tiling paths to capture all of the sequence variations. All of these deficiencies and others not cited provide a less-than-optimal alignment target for next-generation sequencing data and can confound the analytical validity of variants necessary to properly interpret patient-derived data. Hence, although it is difficult work to perform, the ongoing efforts of the Genome Resource Consortium [1] to improve the overall completeness and correctness of the human reference genome should be enhanced.

Along these lines, although projects such as the early SNP Consortium [2], the subsequent HapMap projects [3-5], and more recently the 1,000 Genomes Project [6] have identified millions of SNPs in multiple ethnic groups, there is much more diversity to the human genome than single base differences. In some ways, the broader scope of 'beyond SNP' diversity of the genome across human populations remains mysterious, including common copy number polymorphisms, large insertions and deletions, and inversions. Mining the 1,000 Genomes data using methods to identify genome-wide structural variation should augment this considerably [7], with validation playing an important role, as many methods are still nascent. Lastly, devising clever ways to provide all such classes of variants as a 'searchable space' for sequence data alignment remains a significant challenge, as does the development of sequence alignment algorithms that facilitate the analysis of structurally complex loci.

*Correspondence: emardis@wustl.edu
The Genome Center at Washington University School of Medicine, 4444 Forest Park Blvd, St Louis, MO 63108, USA

How well do we understand the functions encoded by our genome? Certainly, comprehensive functional information about proteins, including the impact of mutations, is complete for relatively few genes. The development of high-throughput systems for biochemistry and enzymology could have a dramatic impact on this deficiency and would add vitality to these areas of scientific endeavor. Efforts that annotate regulatory protein binding sites, sites of RNA-mediated regulatory mechanisms, and other motifs that contribute to transcriptional regulation in the human genome must continue. Improved understanding of these regions, and thus their annotation, will require the power of model-organism-based systems to identify and characterize functional proteins or mechanisms that are shared with humans. We also must transfer these findings into human cell experimental systems that allow researchers to examine the impact of the mutations or other alterations of the genome on cellular pathways and the resulting disease biology. With functional consequences in hand, we will begin to understand and associate the clinical validity of genomic variants, effectively enabling the correlation of variant(s) with the resultant phenotype(s).

If our efforts to improve the human reference sequence quality, variation, and annotation are successful, how do we avoid the pitfall of having cheap human genome resequencing but complex and expensive manual analysis to make clinical sense out of the data? One approach would emphasize the development of 'clinical grade' interpretational analysis pipelines to perform much of the initial discovery from datasets derived from massively parallel sequencing [8]. Although such pipelines already exist in the research setting [9], manual checks and orthogonal validation of variants are required because of the ongoing development of the analytical approaches. Towards patient diagnoses, such validation could initially be performed in a clinical laboratory medicine setting, but ultimately we must develop sophisticated analytical approaches and quality filters that enable high-confidence variant detection solely from the primary data. All discovered variants would then be interpreted in the context of the ever-improving human genome annotation and evaluated in the contexts of medical genetics, of demonstrated clinical validity, and of the pharmaceutical databases (when appropriate), to identify causative or therapeutically actionable genes. Ultimately, as in medicine today, the results will require interpretation by a physician, which raises a separate but equally important issue: the significant need to develop and implement training programs in genomics for medical professionals. Pathologists and genetic counselors will be the first in line for training programs focused on genomic diagnostics, and improving the genomics education of medical students will also be a first priority. More challenging will be the genomics education of practicing physicians and other medical professionals, many of whom do not require genetics to perform their valuable role in health care daily, but who will be confronted in the near term by increasingly well informed patients who expect their doctors to be as well versed as they are about genome-guided diagnosis and treatment.

A final word on the important topic of patient access to genome-guided medicine seems necessary and appropriate. The current high cost of whole-genome sequencing and analysis relative to most clinical diagnostic assays, coupled with the fact that these costs are not currently reimbursed by insurers, might mean that only those with the means to pay for the test will be allowed access. Perhaps worse, those with the fattest wallets might pay extra for a place higher in the queue, denying earlier access to patients who more desperately need the information. Although there are no easy answers here, one plausible solution might be the establishment of funds at major medical centers, where genome-guided medicine is likely to be practiced first, that pay for the genomic sequencing, diagnosis and associated costs and thus allow equitable access to this new assay.

Competing interests

The author declares that they have no competing interests.

Acknowledgements

I thank Deanna Church, Timothy Ley and W Richard McCombie for their critical reading and suggestions.

Published: 26 November 2010

References

1. Genome Resource Consortium: Human Genome Overview [http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/index.shtml]
2. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, *et al.*: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.** *Nature* 2001, **409:**928-933.
3. International HapMap Consortium: **The International HapMap Project.** *Nature* 2003, **426:**789-796.
4. International HapMap Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437:**1299-1320.
5. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, *et al.*: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449:**851-861.
6. International HapMap 3 Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, *et al.*: **Integrating common and rare genetic variation in diverse human populations.** *Nature* 2010, **467:**52-58.
7. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J; 1000 Genomes Project, Eichler EE: **Diversity of human copy number variation and multicopy genes.** *Science* 2010, **330:**641-646.
8. Boguski MS, Arnaout R, Hill C: **Customized care 2020: how medical**

sequencing and network biology will enable personalized medicine. *F1000 Biol Rep* 2009, **1**:73.

9. Ding L, Wendl MC, Koboldt DC, Mardis ER: **Analysis of next-generation genomic data in cancer: accomplishments and challenges.** *Hum Mol Genet* 2010, **19:**R188-R196.

doi:10.1186/gm205
Cite this article as: Mardis ER: The $1,000 genome, the $100,000 analysis? *Genome Medicine* 2010, **2:**84.