

Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: lessons from the ClinSeq project

Leslie G. Biesecker, MD¹

Purpose: The debate surrounding the return of results from high-throughput genomic interrogation encompasses many important issues including ethics, law, economics, and social policy. As well, the debate is also informed by the molecular, genetic, and clinical foundations of the emerging field of clinical genomics, which is based on this new technology. This article outlines the main biomedical considerations of sequencing technologies and demonstrates some of the early clinical experiences with the technology to enable the debate to stay focused on real-world practicalities.

Methods: These experiences are based on early data from the ClinSeq project, which is a project to pilot the use of massively parallel sequencing in a clinical research context with a major aim to develop modes of returning results to individual subjects.

Results: The study has enrolled >900 subjects and generated exome sequence data on 572 subjects. These data are beginning to be interpreted and returned to the subjects, which provides examples of the potential usefulness and pitfalls of clinical genomics.

Conclusion: There are numerous genetic results that can be readily derived from a genome including rare, high-penetrance traits, and carrier states. However, much work needs to be done to develop the tools and resources for genomic interpretation. The main lesson learned is that a genome sequence may be better considered as a health-care resource, rather than a test, one that can be interpreted and used over the lifetime of the patient.

Genet Med 2012;14(4):393–398

Key Words: genetic mutation; incidental findings; inherited disease; massively parallel sequencing

INTRODUCTION

The experiences and data presented here derive primarily from the author's experience directing ClinSeq, a pilot clinical genomics research study.¹ The study was designed to develop infrastructure for generating whole-genome-scale data sets from individual patients, analyze and interpret those data, discover associations of genotype and phenotype, and return selected genomic results to individual subjects. The design of the study is similar in many ways to the informed cohort concept² in that the subjects are highly involved in the study, have ongoing interactions with the investigators, and are invited to undergo iterative clinical research investigations to allow genotype–phenotype associations. The study has enrolled more than 900 subjects and has generated more than 25 billion bases of sequence, and selected results have been returned to a growing number of the subjects.

The important attributes of massively parallel sequencing (MPS)

This technology is based on the recent development of what has been called next-generation or massively parallel sequencing (MPS). The sequencing of the first human genome by the publicly funded Human Genome Project took about 12 y and cost about \$3 billion (<http://www.genome.gov/11006943>). The

technology used in the Genome Project was based on capillary sequencing instruments that could process 96 sequencing reactions of about 700 bp of DNA each in about 10 h (about 3–4 million bp per day per instrument). A number of instruments have been developed in the past several years that can process hundreds of thousands of sequencing reactions of about 100 bp of DNA each—more than 60 billion bp per day. This massive increase in genomic capacity has led to an ~100,000-fold decrease in sequencing costs in the past 10 y.³ This phenomenal decrease in costs has led to the remarkable situation that an analytically valid (which in the United States means that the processes must conform to the Clinical Laboratory Improvement Amendments regulations of 1988) whole-genome sequence can currently be acquired for about \$10,000. With further cost decreases expected, this technology is well within the price range of many other clinically available tests and deserves serious evaluation as a clinical testing tool.

Although much of the debate surrounding high-throughput technology focuses on germline genomic sequencing, it is important to recognize that the instruments used for such sequencing have a wide range of purposes that include sequencing of tumors, sequencing of messenger RNA (complementary DNA) to assess gene expression, assaying methylation to assess gene regulation by epigenetic modifications, and others. These

¹Genetic Disease Research Branch and National Institutes of Health Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA. Correspondence: Leslie G. Biesecker (leslieb@helix.nih.gov)

Submitted 28 September 2011; accepted 7 December 2011; advance online publication 16 February 2012. doi:10.1038/gim.2011.78

other applications of genomic technology raise issues and questions that overlap with germline DNA sequencing; however, this analysis focuses on the direct assessment of germline DNA variation using MPS.

Current implementations of MPS include whole-genome (or so-called shotgun) sequencing and whole-exome sequencing (WES) (this section describes one commonly used current implementation of these techniques, which suffices for the purpose of this discussion. This should not be construed as an endorsement of any single technology or approach). It is important to understand the attributes and limitations of the technical approaches to sequencing when performing clinical interpretation of such data. Whole-genome sequencing (WGS) processes genomic DNA, typically purified from peripheral white blood cells, into small fragments that are then attached at both ends to short pieces of engineered DNA sequences that allow the DNA to attach to complementary DNA strands on a solid matrix, and then be sequenced. To compensate for the randomness of the process (some parts of the genome sequenced more than desired and some less), an excess of sequencing is performed (typically 30×). Then, a series of computational analyses are performed to deduce the sequence of the sample DNA, as compared to a reference DNA sequence. In this way, nearly the entire genome is interrogated, but the process adequately evaluates only about 90% of the genome (L.G. Biesecker, unpublished data). For this reason, the label of “whole” genome sequencing is literally inaccurate, and therefore comprises a less-than-complete genome interrogation.

WES is a variation of WGS that is based on the fact that protein-coding gene (exon) sequences comprise only about 1.5% of the genome yet most (probably >85%) of DNA variations that cause highly penetrant genetic disease lie in this small fraction of DNA. All of the exons of known genes are together called the exome. To target the exome, genomic DNA is hybridized to artificial DNA that represents most of the exome. The DNA that does not hybridize is washed away, and the remaining DNA is then sequenced and processed similar to the WGS method as described above. Because of the inefficiencies and unevenness of these processes, WES is also incomplete and typically covers 85–90% of the exome^{4,5} (and has no, or at best, poor coverage of other areas of the genome). Thus, both methods evaluate around 90% the protein-coding portions of the genome, and WGS in addition evaluates 85–90% of the rest of the genome.

Typically, WGS of an individual genome will detect about 3,000,000–4,000,000 sequence differences, when compared to the reference. Among these variants, 30,000–50,000 will reside in the protein-coding portions of the genes. A WES analysis focuses on these latter variants, generating a similar number as does WGS, but it interrogates very few of the non-protein-coding regions of the genome. The next major step in the analysis is to determine whether, and if so, how, and to what degree the detected sequence changes affect the proteins encoded by the genes and if these changes are likely to significantly affect health. Current methods of sequence analysis are quite effective for many kinds of variation in protein-coding portions of

genes but much less effective for non-protein-coding parts of the genome. The next section of this discussion centers on the analysis of the former class of variants, as this is more tractable and more useful for individual patient-care decision making.

Identification of variants known to cause human disease

The identification of a sequence variant that has been shown through prior genetic and clinical research to cause a disease is a practical application of clinical genomics. Nearly 40 y of human genetics research has generated more than 100,000 genetic variations that are known to cause or contribute to human disease.⁶ Here, it is important to distinguish high-penetrance from low-penetrance genetic variations. The former are genetic variants that are typically rare (found in fewer than 1% of individuals and ranging lower in frequency to those that may be found in only a handful of individuals) and are generally associated with rare or uncommon diseases, and the presence of such a variant in an individual predicts with relatively high likelihood (typically 75% to nearly 100%) that the individual has, or will develop, the disease. For a disease that affects 1/10,000 people in the general population, if a high-penetrance risk variant for this disease was converted to a relative risk, this would mean that the individual with the variant would be 7,500× to nearly 10,000× more likely to manifest the disease as compared with a person without the variant. These are called high-penetrance, rare variants, and MPS is the only practical method for detecting these variants in a comprehensive manner. It is easy to imagine how the detection of such a variant could be used to change the medical management of a patient with such a variant. One important further distinction is that some high-penetrance rare variants cause human disease in a recessive inheritance pattern, which means that an affected individual must harbor two such variants, one on each chromosome (or allele) of the gene.

In contrast to high-penetrance rare variants are the low-penetrance common variants. These variants can be quite common, with individual variants being found in up to 50% of the population, and they can be readily detected by both MPS and DNA chip genotyping methods or other methods of interrogating known single-nucleotide polymorphisms, these last implemented by a number of clinical and direct-to-consumer testing platforms. However, the probability that an individual with such a variant will have the associated disease is typically only 10–50% higher than the background rate of the disease. Again, this can be expressed as relative risk, where the person with the variant is 1.1× to 1.5× more likely than one without the variant to have the disease. The application of DNA testing to identify such modest relative risks (1.1× to 1.5×) for the purpose of making health-care decisions is questionable because the purported risks can be abrogated by other genetic variants or by environmental variation.

A clinical genome sequence (whole genome or exome) can therefore be evaluated to determine if it harbors any of these thousands of variants. Although the analysis tools are in early development stages, there are a number of implementations that allow this to be done in a reasonably effective manner. For example, we have shown that a cohort of patients with

high-throughput sequence can be interrogated for known disease-causing variants in several genes known to cause familial hypercholesterolemia,¹ a rare form of high cholesterol that is refractory to dietary management, whereas pharmacologic treatment can be lifesaving. This was done by computationally intersecting a list of thousands of variants found in the cohort with a database of mutations in two genes known to cause this disease. This resulted in identification of a number of individuals at high risk for this disease, some of which were previously diagnosed, although not all (the medical and genetic counseling issues of this scenario are discussed later). This experience demonstrates that genomic data can be used to identify and diagnose individuals with medically important diseases.

Yet there are many challenges to even this, the simplest form of genomic interrogation. It turns out that the databases that store these disease-causing mutations are far from perfect and include a substantial number of erroneously curated entries. We estimate that more than 10% of the entries in a commonly used genome-wide database of disease genes are erroneously curated, with most of the curation errors of the nature that a benign variant has been erroneously claimed to be pathogenic (L.G. Biesecker, unpublished data). Furthermore, the primary medical research upon which these curated database entries are based is of variable reliability. There are no widely accepted, uniform standards in the genetics research community for the determination of causality of a sequence variant. Based on these considerations, there is a risk that the WGS/WES sequencing results may generate more false-positive than true-positive results. However, the rarity of most of the dominant disorders that could be detected by this approach means that the number of detected variants will be small, and these can be individually, manually re-curated to reduce the number of false-positive results to levels commensurate with clinically available genetic testing.

The situation is much more challenging for recessive carrier mutations. The average person harbors 3–5 such variants and therefore every WGS/WES research study will include many such potential results. Therefore, manual re-curation of these variants will require a great deal of work and is therefore not practical for many research groups. Even were it not challenging because of the high frequency of the variants, the curation is intrinsically difficult because the large majority of persons with the variants are healthy carriers and thus one cannot use phenotypic data from them to determine causality. Instead, the causality has to be derived from affected homozygotes (having two copies of the same variant allele), who may be rare, or compound heterozygotes (having one copy each of two distinct mutant alleles), for whom it is challenging to dissect the causal contribution of each variant.

Identification of variants suspected to cause human disease

The earlier discussion presumed that the variant under consideration had been previously identified and assessed for pathogenicity. But this is not the case for most variants that are detected in a WGS or WES sequencing result. Most variants are

uncharacterized and some are novel—that is, never known to have been observed in a human sample. These variants can be much more challenging to interpret.

Some variants have attributes that allow for relatively facile and accurate clinical interpretation. For example, we have detected a DNA variant that predicts a particular amino acid change in the low-density lipoprotein receptor protein, which causes the aforementioned phenotype of familial hypercholesterolemia. Previously, a slightly different DNA variation that predicts the exact same amino acid change in the low-density lipoprotein receptor protein had been detected in several other patients with familial hypercholesterolemia.⁷ It is overwhelmingly likely that both DNA variants engender the same disease risk and thus the prior data can be generalized to the current, novel finding. As well, there is a large class of variants including nonsense, frameshift, and splice mutations for which inferences can be made from similar data, and application of these to the novel variants is highly likely to be correct.

However, for many, if not most, novel missense mutations, there can be little solid data upon which to base an assessment of causality. There are a number of tools to facilitate this, including bioinformatic analyses of evolutionary conservation, predicted structural changes, and many others. These tools are quite useful, but their predictive power is highly variable.

The ubiquitous “variants of uncertain significance”

Although it is clear that one can derive from an exome sequence results that are highly likely to be valid and medically actionable, this is not the case for the great majority of variants. Among the 30,000–40,000 variants, the typical subject will have three to eight actionable variants. In most subjects, most of these variant will relate to reproductive risks (i.e., heterozygous carrier alleles).⁸ But the remaining thousands are either highly likely to be completely benign, or of uncertain clinical significance. It can be as challenging to prove that a variant is benign as it is to prove it is pathogenic. Currently, nearly all of the variants among these tens of thousands must be considered to be of uncertain significance. Over time, clinical and molecular research will push many of these variants toward the ends of the causality spectrum (either convincingly benign or pathogenic), but it should be anticipated that there will remain a substantial number of variants in this category, including many novel variants, that will continue to be discovered with each additional exome or genome that is sequenced. In the future, as more MPS sequences are performed, the fraction of variants that are novel will decline. However, this will never go to zero as the normal mutation rate will continue to generate a nearly infinite spectrum of genomic variation. We should therefore anticipate that the challenge of variants of uncertain significance will persist for the foreseeable future.

Determining which variants can be returned to research subjects

From the earlier discussion, it is clear that variants from genome or exome and WES sequencing annotation range from those that are extremely likely to cause disease to those that are nearly

certain to be benign, and every gradation in between these extremes. What is also clear is that some of these variants can be not only highly predictive of disease, but their return can enable lifesaving treatment or preventive measures to be implemented. Finally, if one includes the recessive carrier variants as potentially clinically useful, then essentially every subject has results that could be returned—therefore, this is not a theoretical consideration but a practical and current challenge that must be addressed.⁹ The approach to the variants ought to be conditioned on the reliability of the finding, as discussed earlier, but investigators must be cognizant of the nearly continuous spectrum of causality, possible clinical utility, and psychosocial impact that are represented within the thousands of variants that are detected in each WGS/WES interrogation. No simple rules will suffice, as the variants will represent infinite combinations of attributes that defy categorization. The field of clinical genomics will need to develop evidence-based approaches to this challenge.

The clinical genomics research agenda

From the foregoing discussions, it is clear that there is a great deal of work to be done in refining and advancing WES and WGS data generation (so that the modifier of “whole” will be literally true), improving databases to increase the reliability of the determinations of pathogenicity, and developing automated (or at least semi-automated) software tools for annotating sequence data, and related molecular and bioinformatic processes.

In addition to solving the multitude of challenges described earlier, there is another set of considerations relating to results return that demand attention. Current clinical testing modes for inherited disease approach clinical testing primarily from a hypothesis-testing mode of evaluation. That is, clinicians gather a wide spectrum of data including presenting complaint, medical and family history, physical examination, imaging, and others to generate a differential diagnosis. This differential diagnosis is then converted into a unique diagnosis by performing one or a few specific molecular tests to distinguish among the various possibilities. In this mode of testing, the patient typically comes to the clinical juncture of the molecular test with a historical or experiential context of the disorder under consideration—they or their family members have lived and suffered, and some relatives may have even died from this disease. This existing disease context is both the strength and the weakness of current modes of diagnosis and molecular genetic testing. It is a strength because the presence of confirmed manifestations of the disease markedly raises the prior probability that the disease that the test is designed to detect is actually present, and the patient has a context into which the result can be fitted. It is a weakness because we are, in effect, requiring that a patient or their family members have existing morbidity and in some cases even mortality before we are willing to start the process—the antithesis of preventive medicine.

New modes of patient–test interactions

By using WGS/WES as an initial test or screen, clinicians have the opportunity to practice effective preventive medicine, but

it may not be so simple because of the loss of the context of disease. We have observed, in several of our research subjects, a degree of shock and incredulity regarding results of WES. It is not difficult to imagine two clinical scenarios accompanying the discovery of a highly penetrant variant predictive of a serious, late-onset disease inherited in a dominant pattern—a healthy person screened with WES and one with extensive family and/or personal manifestations of the disorder. The latter patient will undoubtedly be emotionally challenged by the finding, but to some degree that result may have been anticipated and the adaptation to the finding is more likely to be successful and prompt. It may be more challenging for the former patient to adapt to the finding. Again, based on our anecdotal experiences noted above, one may expect that the former patient may have a degree of shock and incredulity regarding the variant and the newly discovered disease risk. The medical and genetic counseling challenges of these novel scenarios are significant, and it will be essential to perform behavioral and counseling research to develop appropriate and effective clinical approaches to this challenge.

Although it is not possible for this article to address the topic of the entire workshop or the series of accompanying papers, our research group has developed an approach based on a few assumptions. First, we accept the conclusions of research showing that potential subjects strongly desire the return of individual results as a *quid pro quo* for participation in biomedical research.¹⁰ These data are highly consistent with what our subjects have described to us as one of, if not the major, motivation for their participation in our studies.¹¹ Second, when specifically asked, the subjects that we have enrolled for WES/WGS indicate a high degree of interest and motivation to learn their results—both primary (finding relevant to presenting disorder) and secondary (finding not relevant to presenting disorder). In fact, many of the subjects state that they wish to learn “all” of their results. Third, from our anecdotal experiences returning results in WGS, WES, and single-gene research, we have been impressed that the typical patient becomes overwhelmed and in effect, reaches a point of informational saturation after 20–40 min of counseling. Therefore, we recognize both that there is a high degree of interest in a wide range of results and that current modes of returning results are incompatible with return of large numbers of variants. It is clear from this paradox that novel approaches to this challenge need to be developed and that there is a pressing need for clinical research on standard and alternative modes of returning results.

Increasing the throughput of return of results

As noted previously, there are a huge number of potential variants that can be derived from WES or WGS results and not all of them have high clinical impact. Therefore, the assumption that all test results need to be returned in the context of a face-to-face clinical encounter can be questioned. For example, it would be difficult to argue that an expanded repeat in the Huntington disease gene and an allele of a gene that poses a 1.2× relative

risk of type II diabetes mellitus should be returned using similar clinical approaches. As well, there is a practical issue, which is that WEG/WGS sequencing results, if performed on large numbers of patients and returned using current clinical models, would consume enormous amounts of clinical resources. Given the widely variable clinical contexts of the results and the economic realities of clinical medicine, it is reasonable to propose that new methods of results return should be developed and tested. These may include returning results by mail or by interactive Internet tools whereby subjects can securely access their variant results and link out to information sources that allow them to gain background and context for the finding. As well, this mode of returning results has the potential advantage that it would allow the subjects to absorb results gradually, over a period of time that avoids informational overload. However, the potential disadvantage is that removing the results from the medical setting may excessively reduce the medical impact of the results to a degree that it limits the downstream use of the results. Therefore, these modes of results return warrant development, but they should be tested in rigorous clinical research studies that evaluate their effectiveness and appropriateness.

WES/WGS is a resource, not a test

As noted previously, a whole-genome or -exome result is overwhelming for both the clinician and the patient or research subject. The central problem here is that we are changing paradigms—and a key part of this paradigm shift is to disabuse ourselves of the concept that the WES/WGS is a unitary diagnostic test. The most likely driver of this testing is that the average lifetime costs of all potential individual germline DNA testing will exceed that of a single WGS/WES interrogation. Once this happens, it cannot be regarded as other than wasteful to use individual tests, and the WES/WGS will be the economical approach. Concomitantly, it will be essential to change our approach to the use of this interrogation. Just as the cost of the test should be amortized over the life of the individual, the burden and challenge of interpreting the potential results should be distributed over the lifetime of the patient/research subject. By changing this approach, one is freed from the apparent obligation to return the results of the entire assay in temporal proximity to its generation. This approach is attractive and acknowledges that the individual has different needs and uses for these data at distinct life stages. Therefore, it could be argued that the return of results should correspond to those life stages and the different needs of the patient over time and take into account the medical circumstances of the patients. Taken further, a WES or WGS dataset can be viewed as a health-care resource that can be interrogated by the patient and clinician in situations where it could be of potential use to the patient, when both agree to this use. This approach to treating the genome interrogation as a resource would require changing the mind-set of clinicians and patients—both may be likely to be uncomfortable with undisclosed, but available test results. Yet it is very attractive because the data will be available and ready for any potential use, with as little as a few

keystrokes by the clinician, at the request of the patient, delivering nearly instantaneous point-of-care genomic test results for many potential clinical scenarios.

Concluding remarks and future directions

From the experiences of the ClinSeq cohort¹ and other exome sequencing activities for patients with rare diseases,^{12–14} we have learned that subjects were highly motivated to learn results, that there were numerous medically actionable results that could be derived from WES and WGS, and that current modes of returning individual gene testing results do not scale to whole-genome testing. As well, because WGS and WES can identify susceptibility to disease in patients without an individual or family history of disease, the counseling can be challenging. The fact that these results are important and potentially useful challenges us to develop new modes of returning results and testing those modes of returning results to develop a rigorous evidence base for the exciting and challenging new practice of clinical genomics.

ACKNOWLEDGMENTS

The author acknowledges and thanks all members of both the ClinSeq and the rare disease research teams for their outstanding efforts in these programs, which have generated and pursued many of the ideas described here. The author is supported entirely by the Intramural Research Program of the National Human Genome Research Institute of the National Institutes of Health. The symposium at which the oral version of this article was presented 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).

DISCLOSURE

The author declares no conflict of interest.

REFERENCES

1. Biesecker LG, Mullikin JC, Facio FM, et al.; NISC Comparative Sequencing Program. The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. *Genome Res* 2009;19:1665–1674.
2. Kohane IS, Mandl KD, Taylor PL, Holm IA, Nigrin DJ, Kunkel LM. Medicine. Reestablishing the researcher-patient compact. *Science* 2007;316:836–837.
3. Green ED, Guyer MS; National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* 2011;470:204–213.
4. Teer JK, Bonnycastle LL, Chines PS, et al.; NISC Comparative Sequencing Program. Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Res* 2010;20:1420–1431.
5. Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet* 2010;19(R2):R145–R151.
6. Chen JM, Férec C, Cooper DN. Revealing the human mutome. *Clin Genet* 2010;78:310–320.
7. Ng D, Spaulding E, Mullikin JC, Biesecker LG; NIH Intramural Sequencing Centre. Identification of a novel LDLR mutation (c.261_262invGA, p.Trp87X): Importance of specifying DNA and protein mutations. *Atherosclerosis* 2010;211:397–398.
8. Speicher M, Antonarakis SE, Motulsky AG. Vogel and Motulsky's Human Genetics, 4th edn. Springer: Heidelberg, 2010.
9. Fabsitz RR, McQuire A, Sharp RR, et al.; National Heart, Lung, and Blood Institute working group. Ethical and practical guidelines for reporting

- genetic research results to study participants: updated guidelines from a National Heart, Lung, and Blood Institute working group. *Circ Cardiovasc Genet* 2010;3:574–580.
10. Murphy J, Scott J, Kaufman D, Geller G, LeRoy L, Hudson K. Public perspectives on informed consent for biobanking. *Am J Public Health* 2009;99:2128–2134.
 11. Facio FM, Brooks S, Loewenstein J, Green S, Biesecker LG, Biesecker BB. Motivators for participation in a whole-genome sequencing study: implications for translational genomics research. *Eur J Hum Genet* 2011;19:1213–1217.
 12. Lindhurst MJ, Sapp JC, Teer JK, et al. A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *N Engl J Med* 2011;365:611–619.
 13. Sloan JL, Johnston JJ, Manoli I, et al.; NIH Intramural Sequencing Center Group. Exome sequencing identifies ACSF3 as a cause of combined malonic and methylmalonic aciduria. *Nat Genet* 2011;43:883–886.
 14. Johnston JJ, Teer JK, Cherukuri PF, et al.; NIH Intramural Sequencing Center (NISC). Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. *Am J Hum Genet* 2010;86:743–748.