Genome Biology

EDITORIAL

Next-generation human genetics

Jay Shendure*

Abstract

The field of human genetics is being reshaped by exome and genome sequencing. Several lessons are evident from observing the rapid development of this area over the past 2 years, and these may be instructive with respect to what we should expect from 'next-generation human genetics' in the next few years.

In 2005, two publications introduced methods for massively parallel DNA sequencing [1,2], marking the beginning of a dizzying free-fall in sequencing costs that continues today with no obvious end in sight. To enable the flexible application of these 'next-generation' technologies in the context of human genetics, our group and others have developed new methods for the parallel and programmable capture of complex subsets of the human genome at a cost and scale that is commensurate with the power of new sequencing technologies [3]. These methods facilitate the next-generation sequencing of specific subsets of the genome in many individuals for the same cost as whole-genome sequencing of a single individual. An effective compromise between the competing goals of genome-wide comprehensiveness and cost-control was realized in the concept of 'exome sequencing', that is, the capture and sequencing of the approximately 1% of the human genome that is protein coding [4,5].

The contents of this special issue of *Genome Biology*, as well as over 200 other publications since 2009 whose abstracts contain the term 'exome', confirm the success of exome sequencing as a new and effective technological paradigm within human genetics. Exome sequencing has proven useful for identifying the molecular defects underlying single gene disorders, as well as some genetically heterogeneous disorders; for identifying genes that are recurrently mutated in various cancers; and for new insights with respect to human evolution and population genetics. Furthermore, even though exome sequencing only became broadly accessible in late 2009, well over 10,000 exomes have been sequenced to date. Consequently, what has been published thus far is likely to represent only a small fraction of collective body of work in progress that applies exome sequencing in diverse contexts.

Today, the cost of whole-genome sequencing has fallen to a few thousand dollars, and exome sequencing is being declared in some quarters to be obsolete at the very moment when it has seemingly become pervasive. There is likely to be some truth to this. As the cost of whole-genome sequencing is falling to a level where it is broadly accessible, and as the cost differential between exome and genome sequencing is diminishing as well, there inevitably will be less motivation to bother with exome enrichment. However, although the 'exome versus genome' tension is of great practical relevance, I worry that it may distract us from other lessons that are evident from observing the rapid development of this field over the past 2 years. I attempt to summarize a few of these below, as they may be instructive with respect to what we should expect from 'next-generation human genetics' in the next few years.

High-yield genetics

Exome sequencing identifies approximately 20,000 variants [4], and genome sequencing identifies approximately 4,000,000 variants [6], per individual sequenced. New technologies have altered the nature of the starting point, but the fundamental problem for human geneticists remains the same: how to narrow to the single or few variants that are causal for a phenotype of interest. To date, nearly all successful studies applying exome sequencing to identify disease genes have adopted one of three paradigms for reducing search space. (1) For solving Mendelian disorders, a straightforward strategy initially proposed by our group involves exome sequencing of a small number of affected individuals, filtering of common variants by comparison to public SNP databases or unrelated controls, and prioritization of genes containing apparently rare, protein-altering variants in all or most affected individuals [4]. The major advantage of this approach is that it can be independent of linkage analysis, that is, it enables the identification of the

*Correspondence: shendure@uw.edu
University of Washington, Department of Genome Sciences, Foege Building S-250, Box 355065, 3720 15th Ave NE, Seattle, WA 98195-5065, USA

molecular basis of a Mendelian disorder without requiring access to pedigrees of sufficient size to properly map the locus, or any pedigrees, for that matter (though pedigree information can still be useful, especially for genetically heterogeneous disorders [7,8]). For recessive disorders, particularly those occurring in consanguineous families, exome sequencing of just a single individual (that is, $n = 2$ in terms of affected chromosomes) followed by filtering of common variants may be sufficient to narrow to one or a few candidate genes [9]. (2) An alternative strategy involves exome sequencing of parent-child trios to identify the (approximately) one *de novo* coding mutation occurring per generation [10]. This may be particularly effective for Mendelian disorders where a dominant mode of transmission is suspected and proband(s) with unaffected parents are available. More notably, however, this paradigm is being successfully applied to approach complex neuropsychiatric disorders, including intellectual disability [10], autism [11] and schizophrenia [12]. Although mutations in hundreds of genes may contribute to each of these genetically and phenotypically heterogeneous disorders, the fact that *de novo*, large-effect coding mutations appear to underlie a sizable proportion of sporadic cases provides a highly efficient means for identifying candidate genes. (3) For cancer, a straightforward approach involves the pairwise comparison of exome sequences of tumor and normal tissue from the same individual to distinguish the handful of somatic coding mutations from a large background of inherited variants. Exome sequencing of relatively modest numbers of matched tumor-normal pairs can yield the identification of novel, recurrent driver mutations for specific types of cancer [13,14].

A shared and compelling aspect of each of these strategies is that they represent 'high-yield genetics', that is, the unambiguous identification of a novel disease gene(s) with exome sequencing of a relatively small number of samples and a correspondingly modest investment of resources. There is clearly a lot of low-hanging fruit still to be had, and further decreasing costs and increasing analytical sophistication will only increase the productivity of these paradigms. Furthermore, as the broader field shifts from sequencing exomes to sequencing genomes, these same strategies may prove to be the most 'high yield' for ascertaining the contribution of non-coding mutations to Mendelian disorders as well as to at least some common diseases, for example, neuropsychiatric disorders and cancer.

Power to the people

Hundreds of independent research groups have successfully implemented exome sequencing in the past 2 years. At least five factors contributed to this being possible: (1) the widespread purchase of next-generation sequencing instruments since 2005; (2) the availability of excellent open-source software for data analysis, for example, *bwa* [15] and *samtools* [16]; (3) the rapid development and commercialization of effective reagents for exome capture, for example, Agilent SureSelect, Nimblegen SeqCap; (4) a relatively low cost per sample (that is, capture reagents and one sequencing lane) such that the entry point cost for exome sequencing was historically much more accessible than that of genome sequencing; (5) the fact that such a large number of groups have samples on hand on which they are highly motivated to perform exome sequencing. Why does this broad base of participation matter? First, the learning curve for new technologies can be substantial. As a consequence of the perceived effectiveness, simplicity and affordability of exome sequencing, a much larger group of researchers has engaged and become competent with next-generation sequencing than might otherwise have been the case. Second, the field itself benefits tremendously from this 'democratization' of access and participation, in the sense that much of the innovation and nearly all of the discoveries have come from small groups working with next-generation sequencing for the first time. Notably, there are very few discoveries made by whole-genome sequencing to date that could not have been made more cost effectively by exome sequencing. However, many fewer groups have thus far taken on whole-genome sequencing, and it is possible that broader participation - in terms of the researchers and their samples - remains the missing ingredient.

Challenges and opportunities

Even with the rapid maturation of this field, there are a number of areas that are still, to varying degrees, a work-in-progress; these are described as follows. (1) Exome sequencing fails to solve a substantial proportion of presumably Mendelian phenotypes, even in model organisms where the genetics are crystal clear [17]. If we are to conceive of solving all of the Mendelian disorders for which the causative gene(s) remains unknown, understanding the basis of these failures will be critical. Analogously, there are types of cancer where exome sequencing has not been that successful, due perhaps to marked genetic heterogeneity or the fact that many of the underlying driver mutations may be structural or non-coding. (2) There is tremendous interest in understanding the contribution of rare variation to the genetic basis of common diseases. Many such studies have been initiated using exome sequencing, but are still ongoing as they require large sample sizes to achieve power. These studies will set the stage for understanding the contribution of all rare variants, coding and non-coding, to these same diseases via whole-genome sequencing. (3) The discrete prioritization of all protein-altering variation over all other variation has clearly proven useful, but is undeniably crude. As we shift from exomes to genomes, we incur a 100-fold increase in noise for an unknown gain in signal. We are desperately in need of more sophisticated methods for assigning more appropriate 'priors' to coding and non-coding variants alike. (4) To date, attempts to interpret 'personal exomes' or 'personal genomes' for clinically relevant facts have been mostly disappointing. If we are to be successful in deploying these tools in a clinical setting, we have a very long way to go in terms of predicting phenotype from genotype.

We are only a few years into an incredible trajectory in which exome sequencing and genome sequencing are reshaping the landscape of human genetics. For some problems, it is clear that these technologies were exactly what were needed, and the application of high-yield paradigms by diverse research groups is leading to a plethora of rapid discoveries. For other problems, the removal of one rate-limiting step has only given way to a new rate-limiting step, and we are likely to have our work cut out for us for the foreseeable future.

Abbreviations

SNP, single-nucleotide polymorphism.

Competing interests

The author declares that they have no competing interests.

Published: 14 September 2011

References

1. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 2005, **309**:1728-1732.
2. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, *et al.*: Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005, **437**:376-380.
3. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ: Target-enrichment strategies for next-generation sequencing. *Nat Methods* 2010, **7**:111-118.
4. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J: Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009, **461**:272-276.
5. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR: Genome-wide *in situ* exon capture for selective resequencing. *Nat Genet* 2007, **39**:1522-1527.
6. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, *et al.*: Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008, **456**:53-59.
7. Walsh T, Shahin H, Elkan-Miller T, Lee MK, Thornton AM, Roeb W, Abu Rayyan A, Loulus S, Avraham KB, King MC, Kanaan M: Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. *Am J Hum Genet* 2010, **87**:90-94.
8. Johnson JO, Mandrioli J, Benatar M, Abramzon Y, Van Deerlin VM, Trojanowski JQ, Gibbs JR, Brunetti M, Gronka S, Wuu J, Ding J, McCluskey L, Martinez-Lage M, Falcone D, Hernandez DG, Arepalli S, Chong S, Schymick JC, Rothstein J, Landi F, Wang YD, Calvo A, Mora G, Sabatelli M, Monsurrò MR, Battistini S, Salvi F, Spataro R, Sola P, Borghero G; ITALSGEN Consortium, *et al.*: Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron* 2010, **68**:857-864.
9. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloğlu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP: Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* 2009, **106**:19096-19101.
10. Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, del Rosario M, van Bon BW, Hoischen A, de Vries BB, Brunner HG, Veltman JA: A de novo paradigm for mental retardation. *Nat Genet* 2010, **42**:1109-1112.
11. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, Karakoc E, Mackenzie AP, Ng SB, Baker C, Rieder MJ, Nickerson DA, Bernier R, Fisher SE, Shendure J, Eichler EE: Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nat Genet* 2011, **43**:585-589.
12. Girard SL, Gauthier J, Noreau A, Xiong L, Zhou S, Jouan L, Dionne-Laporte A, Spiegelman D, Henrion E, Diallo O, Thibodeau P, Bachand I, Bao JY, Tong AH, Lin CH, Millet B, Jaafari N, Joober R, Dion PA, Lok S, Krebs MO, Rouleau GA: Increased exonic *de novo* mutation rate in individuals with schizophrenia. *Nat Genet* 2011. doi: 10.1038/ng.886.
13. Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, Davies H, Jones D, Lin ML, Teague J, Bignell G, Butler A, Cho J, Dalgliesh GL, Galappaththige D, Greenman C, Hardy C, Jia M, Latimer C, Lau KW, Marshall J, McLaren S, Menzies A, Mudie L, Stebbings L, Largaespada DA, Wessels LF, Richard S, Kahnoski RJ, Anema J, *et al.*: Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* 2011, **469**:539-542.
14. Wei X, Walia V, Lin JC, Teer JK, Prickett TD, Gartner J, Davis S; NISC Comparative Sequencing Program, Stemke-Hale K, Davies MA, Gershenwald JE, Robinson W, Robinson S, Rosenberg SA, Samuels Y: Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat Genet* 2011, **43**:442-446.
15. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009, **25**:1754-1760.
16. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup: The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009, **25**:2078-2079.
17. Fairfield H, Gilbert GJ, Barter M, Corrigan RR, Curtain M, Ding Y, D'Ascenzo M, Gerhardt DJ, He C, Huang W, Richmond T, Rowe L, Probst F, Bergstrom DE, Murray SA, Bult C, Richardson J, Kile B, Gut I, Hager J, Sigurdsson S, Mauceli E, Di Palma F, Lindblad-Toh K, Cunningham ML, Cox TC, Justice MJ, Spector MS, Lowe SW, Albert T, *et al.*: Mutation discovery in mice by whole exome sequencing. *Genome Biol* 2011, in press.

doi:10.1186/gb-2011-12-9-408
Cite this article as: Shendure J: Next-generation human genetics. *Genome Biology* 2011, **12**:408.