Commentary

New strategies and emerging technologies for massively parallel sequencing: applications in medical research

Elaine R Mardis

Address: The Genome Center at Washington University School of Medicine, Department of Genetics, 4444 Forest Park Boulevard, St Louis, MO 63108, USA. Email: emardis@wustl.edu

Published: 17 April 2009

Genome Medicine 2009, **1**:40 (doi:10.1186/gm40)

The electronic version of this article is the complete one and can be found online at http://genomemedicine.com/content/1/4/40

© 2009 BioMed Central Ltd

Abstract

A variety of techniques that specifically target human gene sequences for differential capture from a genomic sample, coupled with next-generation, massively parallel DNA sequencing instruments, is rapidly supplanting the combination of polymerase chain reaction and capillary sequencing to discover coding variants in medically relevant samples. These studies are most appropriate for the sample numbers necessary to identify both common and rare single nucleotide variants, as well as small insertion or deletion events, which may cause complex inherited diseases. The same massively parallel sequencers are simultaneously being used for whole-genome resequencing and comprehensive, genome-wide variant discovery in studies of somatic diseases such as cancer. Viral and microbial researchers are using next-generation sequences to identify unknown etiologic agents in human diseases, to study the viral and microbial species that occupy surfaces of the human body, and to inform the clinical management of chronic infectious diseases such as human immunodeficiency virus (HIV). Taken together, these approaches are dramatically accelerating the pace of human disease research and are already impacting patient care.

The human genome lies at the core of research into human disease. New technologies for obtaining genome sequence data are being combined with novel bioinformatics analyses to characterize disease samples of many types, in the hope of enhancing our fundamental understanding of susceptibility and onset for inherited diseases, of the somatic changes that take place to initiate cancers and cause metastatic disease, and of the identity and allelic spectra of pathogenic and commensal microbes that infect humans. These sequencing-based discoveries will have a major impact on medical practice, including the development of diagnostic and prognostic assays, the identification of altered proteins to which targeted therapies may be developed, the ability to predict onset and severity of disease, and an improved capability to predict our range of responses to pathogenic agents. They will also create large datasets that effectively identify each patient by their sequence information, establishing the potential of linking a patient to a disease and heightening the need to safeguard the privacy of these data through legislation against genetic discrimination.

Inherited complex diseases have proved the most pervasive yet recalcitrant examples of human disease to reveal their genomic secrets. From a standpoint of statistical significance, studying inherited disease at the genomic level requires large numbers (ideally thousands) of cases (affected) versus controls (unaffected) to uncover initial findings, as well as the replication of any primary discoveries

in other case-control cohorts to solidify the association of a given genomic variant(s) with disease. Although genome-wide association studies (GWAS) have been broadly applied across the spectrum of hypertension, diabetes, autism and other diseases, the identification of disease-associated genes by GWAS has so far identified mainly genes of low effect size or within regions of the genome that do not contain annotated genes, hence making it difficult to assign functionality or even putative causality. With the development and publication of several methods that can selectively isolate sequences of interest from a genomic DNA sample, there are now high-throughput methods available for variant discovery within genomic regions identified by GWAS or by candidate gene approaches. These methods use a variety of solid-phase [1,2] or solution-phase [3] strategies to capture the desired loci, typically isolating DNA fragments that represent hundreds to thousands of genes in a single experiment, which are then sequenced using next-generation sequencing technology. Downstream variant discovery aligns the sequences obtained to the targeted regions or genes, and then identifies high-quality sequence differences. A secondary level of interpretation can identify those variants encoding a different amino acid or a premature stop codon likely to impact the structure and function of the proteins containing them. This exercise can provide clues about which modified proteins may be contributing to the disease biology. Combining the variant information obtained across affected individuals with an analysis of those cellular pathways in which the altered proteins participate can then enable higher-level concepts to emerge about disease biology.

All targeted capture methods are somewhat limited by the fact that they rarely yield 100% coverage of the sequence from any targeted region. As such, where coverage gaps exist variants cannot be discovered, but this also occurs with polymerase chain reaction (PCR)-based approaches. By contrast, the combined targeted capture plus next-generation sequencing generates data much faster, is more scalable in terms of genes targeted and the ability to combine patient samples into a single capture experiment, and is cheaper than the conventional approach of PCR and capillary sequencing. Moreover, because each next-generation sequence read represents data from a single DNA strand, the ability to discover sequence variants is greatly facilitated over that of diploid PCR product sequences (both alleles represented in the same reaction) obtained from a capillary read. Targeted capture approaches are now being applied to GWAS peaks for many inherited disease studies. The resulting data will reveal the spectrum of rare single nucleotide and insertion-deletion variants, and hopefully will shed additional light on the genomic predisposition to the disease of interest. Given the operational scale possible for some of these methods (for example, solution-phase capture can be carried out readily in a 96-well plate format) and the massively parallel scale of next-generation sequencing throughput, we will soon know whether this predicted efficacy yields the promised rare variant discoveries. However, the limited representation of the human genome on single nucleotide polymorphism (SNP) arrays will not provide a complete picture of case-specific variation via GWAS. Rather, as the price of whole-genome resequencing falls, so will the desire to fully characterize the genomes of inherited, disease-affected individuals by whole genome resequencing so that variant discovery is unbiased.

In contrast to inherited disease, somatic diseases such as cancer do not typically require large sample numbers to provide significant discovery potential. Both targeted capture and whole-genome resequencing are currently being used to study cancers of various types. Here, the list of targeted genes typically is derived by a candidate-based approach, which includes known cancer-associated genes from studies of the particular tumor type, and may possibly include genes lying under genomic regions determined to have altered copy number (amplification or deletion) [4]. Studies to date using PCR and capillary sequencing have essentially revealed that although many hundreds of genes are sequenced, only a handful attain significant levels of mutation across all samples studied. For example, in a recent study of glioblastoma multiforme, The Cancer Genome Atlas consortium reported that of 601 candidate genes sequenced, only 223 revealed a mutation in at least one of the 91 tumor samples studied and only eight genes could be classified as significantly mutated [5]. Similarly, when whole-genome resequencing was used to study an acute myeloid leukemia (AML) genome and the genome of its matched normal (skin) sample, the eight single-nucleotide non-synonymous variants found were in genes that essentially would never have been on a candidate gene list for AML [6]. This early result has made a strong case for whole-genome resequencing as an unbiased approach by which medical research can pursue the genomic basis of cancer. To date, we have been hindered by the cost of whole-genome resequencing, but the falling cost of next-generation sequencing experiments is quickly eliminating this barrier, and hundreds of cancers will likely be sequenced during 2009.

One interesting concept to note is that, using a variety of DNA and RNA preparatory methods combined with next-generation sequencing, one can produce sequence-based characterizations of tumors that reveal the spectrum of variation across the genome, the 'methylome' (that is, cancer-specific changes in DNA methylation patterns), and the transcriptome (expression levels of mRNA plus other non-coding RNAs) in comparison to non-cancerous matched tissue [7]. This wealth of data, when coupled with clinically relevant information about the cancer (age of onset, treatment history, outcome, family history/susceptibility, and so on), provides the potential to more fully shape our understanding of the disease biology. With similar analyses

in hand for many samples of the same cancer type or subtype, a correlated spectrum of affected genes, pathways, treatment options and outcomes, including findings with translational impacts on patient care, will begin to emerge. Only by performing many such studies will we begin to understand how individualized each cancer is, and hence shed light on the importance of whole-genome versus targeted analysis of cancers as a diagnostic or prognostic measure.

While cancer and inherited diseases are important disorders to study using genomic data, the combination of viral, bacterial and fungal pathogens adversely impacts many more lives worldwide per year. Given the difficulty or impossibility of culturing most pathogenic species, the use of next-generation sequencing has greatly enabled their identification. This typically is accomplished by isolating all nucleic acids from a human sample (such as feces, tooth or skin scrapings), sequencing the isolated DNA mixture, and then assembling the non-human sequence reads to reconstruct segments of the pathogenic genomes carried in the sample. Next, these DNA sequences are translated into amino acid sequences to help identify the novel pathogen(s) and to characterize gene content and other attributes (such as antibiotic resistance). Such 'metagenomic' studies comprise some of the most exciting discovery-based science happening in medical research today.

In terms of known pathogens, the exquisite sensitivity of next-generation sequencing is already being applied to disease management. One brilliant example of this is in human immunodeficiency virus (HIV) disease management, where newly diagnosed patients' sera are input to PCRs that target specific viral genome regions, allowing the clinician to assess the mutational status of each viral population as a precursor to developing a patient-specific drug cocktail. In essence, while there are over 20 HIV-specific therapies available, only by sequencing each patient's viral population can therapies to which viral resistance already exists be avoided [8].

It also remains unclear whether and how healthy individuals interact with their natural commensal bacterial, viral and eukaryotic species ('microbiome'), or how changes to the normal prevalence of various species (or strains) may influence our relative wellness. The NIH Roadmap has set up a Human Microbiome Project [9] and provided funding to begin to address such questions, as well as to dramatically improve the census of sequenced human microbiome species available in public databases. These genomes will provide a tremendously enhanced database for microbiome- and metagenomic-based queries, furthering medical research in the process of answering the question 'Who's there?' for different healthy and disease states.

Cumulatively, the impact of next-generation sequencing on medical research is beginning to gain momentum, and we can already predict a point at which discoveries resulting from the kind of inquiries described in this commentary may overwhelm our abilities to translate them into clinical applications. After all, sequencing data will provide a multitude of clues, but few direct answers. To meet this challenge, functional screening approaches must scale up to high throughput and our characterization and annotation of human genome functional elements must accelerate. We need to find ways to engage colleagues familiar with biochemistry, cell biology, medicine, pharmacology, and other relevant fields to help interpret discoveries from next-generation sequencing studies and to think about the next steps. And, in an era of genomic data accessibility, we need to ensure the privacy of would-be study volunteers so our work proceeds without the delays that will occur if there is a dearth of properly consented samples.

Abbreviations

AML, acute myeloid leukemia; GWAS, genome-wide association studies; HIV, human immunodeficiency virus; PCR, polymerase chain reaction; SNP, single nucleotide polymorphism.

Competing interests

The author was a director of the former Applera Corporation (now Life Technologies) from 2007 until 2008. Dr Mardis currently serves on the Scientific Advisory Board of Pacific Biosciences Corporation.

Acknowledgements

I would like to thank my colleagues, Rick Wilson, Tim Ley and George Weinstock for critical comments and input.

References

1. 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.
2. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA: **Direct selection of human genomic loci by microarray hybridization**. *Nat Methods* 2007, 4:903-905.
3. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum, C: **Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.** *Nat Biotechnol* 2009, 27:182-189.
4. Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, Lin WM, Province MA, Kraja A, Johnson LA, Shah K, Sato M, Thomas RK, Barletta JA, Borecki IB, Broderick S, Chang AC, Chiang DY, Chirleac LR, Cho J, Fujii Y, Gazdar AF, Gioradano T, Greulich H, Hanna M, Johnson BE, Kris MG, Lash A, Lin L, Lindeman N, et al.: **Characterizing the cancer genome in lung adenocarcinoma.** *Nature* 2007, 450:893-898.

5. The Cancer Genome Atlas Consortium: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 2008, 455:1061-1068.
6. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J., Conyers J, Sander N, Shi X, Osborne JR, Minx P, *et al.*: **DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.** *Nature* 2008, 456:66-72.
7. Mardis ER: **The impact of next-generation sequencing technology on genetics.** *Trends Genet* 2008, 24:133-141.
8. Kozal MJ: **Drug-resistant human immunodefiency virus.** *Clin Microbiol Infect* 2009, 15(Suppl 1):69-73.
9. **NIH Roadmap for Medical Research** [http://nihroadmap.nih.gov/hmp/].