Genome Biology

MEETING REPORT

Moving beyond genome sequencing into personalized genomic medicine: biological and computing challenges

David M Kristensen*

Abstract

A report of the second annual Beyond the Genome conference held on the 19-22 September 2011 at The Universities at Shady Grove, Rockville, Maryland, USA, where increases in computing that may help make personal genomics a reality were a major focus.

The conference started with a Genome Informatics 'pre-meeting', which focused on the increased use of computing technologies in genomics research. This set the tone for the main conference, which discussed recent progress in cancer genomics, exome sequencing, and understanding the microbiome. The increasing reliance on computers has permeated the field to the point where a researcher must be aware of the biology and the computer science behind even routine tasks. Although sequencing technologies improve, the analysis of these data continues to lag far behind. If personal genomics is ever to become a viable mainstream treatment option for patients, then improvements throughout the process will be necessary.

Mike Schatz (Cold Spring Harbor Laboratory, USA), who introduced and chaired the pre-meeting, along other speakers predict the reversal of the trend for most genomics research to become concentrated in the hands of large sequencing centers and collaborations involving hundreds of people. As technology becomes more affordable, sequencing will once again become feasible for an individual or small lab (James Taylor (Emory University, USA) referred to this as the 'democratization of sequencing'). Key to this effort is the automation of routine tasks, as computer algorithms become more sophisticated, intelligent, and easy to use. Schatz highlighted three major themes relating to how technologies are changing to meet the current demands of genomics:

Smarter algorithms and more focused approaches

Algorithms that can correct for typical sequencing errors are increasing the reliability of data and decreasing overall cost (requiring less manual interaction). Not all errors are technical in nature, for example rarely occurring mutations or cell lines that have been passaged many times, some since the 1970s, as pointed out in the sponsored technical presentation by Complete Genomics. Other algorithms focus on increasing throughput, for example Schatz mentioned that the slowest indexed approach (Burrows–Wheeler transform) to map reads onto an existing genome is faster than the fastest approach using traditional Smith–Waterman alignments. David Jaffe (The Broad Institute, USA) discussed improvements for generating complete, cheap and foolproof assemblies. He presented two options for assembly of small and large genomes using a range of libraries for both, because sequencing from libraries with a spectrum of size inserts improves assembly. He demonstrated his protocol for assembling large genomes on previously generated data, using ALLPATHS-LG. Such improvements are crucial if the focus of genome research is to move from large sequencing centers to individual labs, a point that was also made by the winners of the poster competition. Another way to reduce costs is to simply avoid "unnecessary" work. Exome rather than full-genome sequencing is a prime example of this as it avoids sequencing the non-protein-coding 99% of the genome.

Parallel computing

Sequencing data are currently growing at least 4-fold faster than computing power (and often referred to as the "data tsunami"). If data analysis is to keep pace with this increased throughput, parallelization is required, and often the task of a smarter algorithm is to spread the computing load across hundreds or thousands of processors.

*Correspondence: David.Kristensen@nih.gov
National Center for Biotechnology Information, NLM, NIH, Bethesda, MD 20894, USA

Ben Langmead (Johns Hopkins University, USA) illustrated some of the challenges faced in parallelizing his work on the Bowtie, Crossbow, and Myrna software. One emerging technology that aims to make it easier to perform tasks in parallel is "cloud computing", as described by Matt Wood (Amazon Web Services, USA). One advantage of this approach is that researchers can share a 'computing instance', complete with software and data pre-installed. Yinguri Li (BGI, China) added that making data publicly available does not simply mean making it downloadable – it needs to be usable. On a similar note, James Taylor (Emroy University, USA) is committed to providing the computing services of Galaxy free of charge, making the data readily accessible and reproducible, through the use of common tools. Taylor questioned whether it does any good to make a protocol that captures the 'best practices' in an area, if nobody in the community follows that protocol afterwards (e.g. most people citing the 1000 Genomes project do not use the protocols developed by it)?

Data availability, storage and transfer

Researchers are increasingly forced to consider the limitations of hardware and I/O bandwidth. Lincoln Stein (Ontario Institute for Cancer Research, Canada) gave examples of how data storage and transfer are extremely limiting. He used as an example the International Cancer Genome Consortium, a coordinated effort that uses a federated database system that can only provide interpreted data to the public, because the raw data require several petabytes of storage capability. The consensus at the end of this meeting was that no "magic formula" exists to deal with this issue, although lossy compression schemes, which discard the least-interesting data, are being explored for future use (such as those that map reads onto an existing reference genome and store only the differences). One question fielded during the panel discussion combines all three issues: why don't we take advantage of graphics processors (GPUs) on existing machines? The consensus was that it was tried and failed because: (i) they are special-purpose devices that do not always work well for the task at hand; (ii) smarter algorithms and increased use of parallelization are driving the field, not more efficient use of a single processor; and (iii) even when a GPU could perform the computations faster than a traditional processor, the data I/O limitations reduce its speed

Cancer and personalized medicine

The main conference comprised three themes over three days. The first of these was cancer genomics and the session was chaired and introduced by Elaine Mardis (The Genome Institute at Washington University, USA), who began by discussing how the cancer genome evolves as acute myelogenous leukemia develops from 'myodysplasia syndromes' (MDS) (a form of preleukemia). By sequencing and clustering cells according to the frequency of mutant alleles, she showed it is possible to track the evolution from MDS to AML, with populations beautifully clustering on graphs. Peter Campbell (The Wellcome Trust Sanger Institute, UK) also focused in on MDS, but used an exome sequencing approach to identify further 'driver' mutations in this clonal stem cell disease. Sam Aparicio (BC Cancer Agency, Canada) discussed some results from METABRIC (a large breast cancer study). He associated the genomic mutations in breast cancer with changes in genome expression, and showed that such changes can be detected in *trans* and in *cis.* Mathew Ellis (Washington University in St. Louis, USA) described how cancer is currently being treated 'backwards'; that is, we are sequencing genomes to determine patient eligibility for existing therapies, rather than designing a treatment for the particular cancer that a patient has, which would require real-time genomics. Mark Boguski (Harvard Medical School, USA) discussed the importance of cost in patient decision-making, reminding us that to make genomic medicine a reality, we need clinical-grade results that are trusted and understandable by non-technical experts, and available at a reasonable price. Lincoln Stein (Ontario Institute for Cancer Research, Canada), who spoke at the pre-meeting, provided us the salient reminder that cancer chemotherapy is currently 'one-size-fits-all', even though every cancer is different, which needs to be remedied in the future.

Exome sequencing

The exome session was introduced and chaired by Jay Shendure (University of Washington, USA), who discussed the use of exome sequencing to identify causative variants of Kabuki syndrome and autism. Through the Kabuki example, he emphasized the need for better pipelines for variant discovery and the heterogeneity of disease. Strong pipelines for variant discovery are particularly important as regards the discovery of *de novo* mutations. Joris Veltman (Radbound University, The Netherlands) discussed *de novo* mutation discovery in intellectual disability by exome sequencing of patient trios. Such mutations are now being discovered owing to the availability of exome data . However, there are still many questions that need addressing. Are there *de novo* mutation hotspots in the genome? What is the best way to distinguish between pathogenic and non-pathogenic mutations? The wealth of exomic data becoming available from large consortia, such as the Exome Sequencing Project (the main aim of which is to identify the causes of heart, lung and blood disorders), gives us the means to understand population evolution. Timothy O'Connor,

(University of Washington, USA) presented the genetic variability in the Exome Sequencing project. He showed very few variants that are shared by individual, which is a consequence of recent population expansion; also the number of variants is a function of sample size. Interestingly, when analyzing for purifying selection even in synonymous mutations, there is a significant proportion of mutations under selection. Thus the exomic and genomic data explosion is allowing insight into both medical and population genetics.

Human microbiome

The human genome is only one of the genomes that are contained within our bodies. Studying the genomes of the various microbes living within us (the 'microbiome') will present interesting challenges. For example, it is increasingly clear that while the species in a microbiome can change, the metabolic profile remains more constant. Karen Nelson (J Craig Venter Institute, USA) introduced and chaired this session and discussed the creation of a catalog of reference genomes – necessary because only 1% of bacteria have been previously cultivated – and how the microbiome present in the human 'superorganism' can change over as a disease progresses. Mihai Pop (University of Maryland, USA) mentioned that assembly is an almost impossible task; it is more challenging to assemble the gut microbiome than the microbiome of the teeth and vagina. Ian Wilson (The Scripps Research Institute, USA) described how high-throughput structural genomics is aiding in the understanding of the microbiome. Maria Giovanni (National Institute of Allergy and Infectious Disease (NIAID), USA) and Sarah Highlander (Baylor College of Medicine, USA) described the roles of NIAID and the Human Microbiome Project in providing funding and leadership in developing bioinformatics approaches to help address the challenges faced in this field, for example filling in gaps of the phylogenetic tree depicting the 'normal flora' of the human body.

Conclusions

Although modern researchers must increasingly be mindful of the computer resources necessary to solve current biological problems (including internet routers, disk storage devices, number of processors available to work in parallel, etc.), as routine tasks become automated, we can look forward to moving towards real-life, practical uses of genomic technologies in clinical and other settings. Until then, as Schatz mentioned during his presentation – "A word of caution: new technologies are new."

Competing interests

The author declares that he has no competing interests.

Acknowledgements

I thank Ben Busby (NCBI, NLM, NIH, Bethesda MD) and Linda Kristensen for helpful feedback.

Published: 24 October 2011

doi:10.1186/gb-2011-12-10-308
Cite this article as: Kristensen DM: Moving beyond genome sequencing into personalized genomic medicine: biological and computing challenges. *Genome Biology* 2011, **12**:308.