Genome interpretation and assembly—recent progress and next steps

Ten experts discuss innovations needed to analyze the millions of genomes soon to be sequenced.

With over 50,000 human genomes and exomes resequenced and >600 animal or plant genomes sequenced *de novo*, generating genome sequence data is becoming increasingly commonplace. The question is whether the tools and infrastructure to analyze these data are keeping up. *Nature Biotechnology* asked experts in academia and industry to share their thoughts on two of the sequencing field's key computational challenges—assembling genomes and developing pipelines to interpret genomes. The following edited compilation of their responses highlights the need for improved accuracy and centralized standards and the opportunities resulting from the rapid pace of innovation.

***Nature Biotechnology:* What problems do you commonly encounter with sequence-analysis pipelines and data interpretation?**

Shawn Baker: Keeping up with the changing algorithms and knowing which is the best to use. Beyond that, it would be dedicating the time necessary to get everything up and running and then keeping it running. And, if you're running a truly high-throughput system, finding enough storage for all of the data.

Shawn Baker, CSO, BlueSEQ

Anika Joecker, Senior Bioinformatics Specialist, CLC bio

Anika Joecker: There is a lack of validated sequencing data—that is, not many public datasets are available for which all genomic variants are known. Also, benchmarking analytic tools consumes a lot of time, which people try to save by using simulated data only. But the problem with this is that different technologies have different error sources, and algorithms have to be able to deal with all kinds of data. Therefore, before developing automatic pipelines for sequencing analysis, we have to show that on real data the pipeline actually does a good job.

George Church: Clinical interpretation needs high quality data, including haplotype phasing, to distinguish two mutations in one copy of a gene from mutations in both copies. Such methods (for example, Complete Genomics' long fragment read technology) also improve simple SNP accuracy 50-fold and help deal with repeats. We need to fix the 300 remaining gaps in the human reference genomes.

George Church, Professor of Genetics, Harvard Medical School

Michael Snyder & John West: Modularity is important because new and better algorithms appear regularly.

Steven Salzberg, Director, Center for Computational Biology, Johns Hopkins University School of Medicine

Steven Salzberg: For resequencing projects, a common pitfall in some biomedical projects is to expect too much from the sequencing report provided by the sequencing company. A number of companies offer to sequence the genome and report on all the SNPs and other variant calls, which may be a very good deal. However, the analysis is only just beginning when that report arrives. Researchers need to recognize that they need bioinformatics expertise to make sense of these reports, and they should be prepared to conduct much more computational analysis once the data arrive.

Is it feasible for most labs to develop their own genome-sequence analysis pipelines?

Elizabeth Worthey: Clinical groups could each develop their own pipelines, but this would not be the best outcome. There should be adoption of a limited number of tools that fundamentally support data exchange between different instances of the same tool and between different tools using well-defined data formats. No standardized clinical pipelines exist. There are research pipelines, but much work needs to be performed to bring these to clinical grade. There are advantages and disadvantages to both commercial and in-house solutions (**Table 1**).

Elizabeth Worthey, Assistant Professor, Medical College of Wisconsin

S.B.: It is feasible as a lot of open-source tools have been generated (such as Galaxy and GATK (Genome Analysis Toolkit)), but it does take some skill to set up and maintain. Although there are popular tools, they aren't exactly standardized, as people tend to have their own favorite specific pipeline and settings. The advantage of setting up your own pipeline (as opposed to using a 'black box' commercial solution) is that you have complete control over the process. A commercial pipeline can be a lot easier, but the downside is that it might let researchers avoid having to develop a true understanding of what's going on with the analysis, and this lack of understanding might lead to suboptimal results.

© 2012 Nature America, Inc. All rights reserved.

npg

Table 1 Advantages and disadvantages of commercial versus in-house genome data analysis

Commercial	In house
Can provide revenue for continued development.	Can be tailored to fit other in-house systems, such as laboratory information management systems and electronic health records.
Can force multiple groups to coalesce on formats and data requirements as they adhere to the inputs and outputs of purchased products.	Provides tailoring to specific questions and goals (and there are many approaches with next-generation sequencing at the moment).
Commercial partner can provide early funding for large development phase.	Requires substantial initial investment to get something that works.
May not meet the specific goals of each group, particularly in terms of integration with electronic health records, which are themselves diverse. This may push the actual solution to a professional service firm, which may negate some of the benefits of using commercial software.	Supports development of novel approaches or algorithms.
Practical if goal is to get something up and running in under 3 months.	Develops in-house expertise that can be used elsewhere in the organization at other times.

Courtesy of Elizabeth Worthey.

Todd Smith: For all but a few leading labs, developing independent analysis pipelines will result in largely incomplete solutions because of requirements for deep expertise and large infrastructure. Independent labs, however, can make significant contributions to larger efforts through their understanding of specific research problems.

Todd Smith, Senior Leader Research and Applications, PerkinElmer

A.J.: One factor to consider is that people in academia switch their workplaces quite often and pipelines (perhaps developed by a PhD student) are often not well documented and are difficult to maintain when the person who created them has left the lab.

What strategies are available for improving genome assemblies?

Jun Wang: At BGI, we have built *de novo* assembly maps for 416 animal genomes and 223 plant genomes, and we have sequenced 57,748 human genome samples (targeted region or whole genome) for medical research. Using current sequencing technology, we are able to do sequencing and assembly from single-cell and metagenomics samples. However, for single-cell sequencing, there are still a lot of technology limitations in regard to efficiently isolating a cell from tissue and amplifying DNA or RNA without bias. And for metagenomics studies, we still lack tools for isolating a single microorganism and sequencing its genome. Genomes with high repetitive content, with high heterozygosity or from polyploid species are difficult to sequence, but this can be done by using BAC (bacterial artificial chromosome)-to-BAC or fosmid-to-fosmid strategies in conjunction with next-generation sequencing [as was done for the oyster genome; Zhang, G. *et al.*, *Nature* **490**, 49–54, 2012).

Jun Wang, Executive Director, BGI

We believe improvements should be mainly targeted against two core limitations: gaps and lack of continuity. Longer reads help fill gaps. Mate-pair libraries, physical maps and genetic maps can be applied to improve continuity. We routinely generate 40-kb mate-pair libraries for Illumina (San Diego, CA) sequencing. We have also used a whole-genome mapping strategy (based on optical mapping) to improve the assembly of mammalian genomes. From our experience, restriction enzyme–associated DNA sequencing and genotyping-by-sequencing are both promising ways of obtaining high-density genetic maps for genome assembly.

S.S.: One strategy is simply to get deeper coverage: with bacterial genomes, we're able to get extremely deep coverage (200× or more) at very low cost, and the newest assembly software can take advantage of this. Coverage doesn't allow you to span longer repeats, though, so the second way to improve assemblies is to produce longer paired-end libraries. The third way to improve assemblies is through longer reads, which can have a dramatic effect on contig lengths. Read length is the area where the technology is changing the fastest, and if we can generate even light coverage (3–5×) with reads that are thousands of bases long, assemblies will be much better. In our experience, the quality of a genome assembly varies dramatically based on the choice of assembly software. We also have learned that quality needs to be gauged not just by the size of the assembled pieces (contigs) but by their correctness. Almost all genomes produced today are 'draft' genomes, and unless there is a closely related finished genome to compare against, we cannot easily evaluate correctness.

What are likely to be the major markets for new technologies in assembly and analysis?

G.C.: The research market probably is bigger now, but the clinical market is likely to be the biggest soon (billions of inherited and cancer whole-genome sequences). Ecological and environmental sequencing has plenty of room to grow, but funding resources are less clear than for clinical whole-genome sequencing.

Jun Wang: Currently, academic research is still the biggest driving force of new technologies development for sequencing. Having said that, it is hard to estimate the relative sizes of markets. I would say the ratio between academic and industrial use is about 1:1; but the ratio for the academic to clinical applications is at least 1:100.

T.S.: Clinical sequencing will have the greatest need for standards, data accuracy and precision. DNA sequencing is clearly important in agricultural biotech where it is used largely to improve plant breeding. In terms of market scale, historically human health has been the driver, with markets typically tenfold larger than in the agriculture sector. Over the coming years, this could change as we will have to get much smarter about how to feed people.

What key challenges do we face when analyzing human genomes for medical purposes?

G.C.: Two needs are, first, NIST-FDA (US National Institute of Standards–US Food and Drug Administration) standards for genome samples and, second, an open platform for minimal genome reports, such as http://evidence.personalgenomes.org/. Such a

© 2012 Nature America, Inc. All rights reserved.

npg

Box 1 The commercial angle

We asked interviewees how business models are evolving to address data analysis challenges. Their responses are provided below:

Michael Snyder, Professor and Chair of Genetics, Stanford

John West, CEO, Personalis

M.S. & John West: One business model is to just sell sequencers (for example, Illumina or Life Technologies); another is sequencing as a service (for example, Complete Genomics/BGI); and others purport to offer some level of genome interpretation (for example, Omicia or Knome) or combinations of sequencing and interpretation (for example, Knome). Some offer next-generation sequencing as a clinical diagnostic test (for example, Ambry or GeneDx). A variety of informatics companies that handle data and processing variants exist as well (for example, DNA Nexus). A careful, medically relevant interpretation of genomes using state-of-the-art information is where companies like Personalis fits in. Accuracy is important; a long list of variants is not useful to most people.

G.C.: For genome interpretation, there seems to be a significant fraction of clinical services and even research groups that do not want their data to leave their site, hence the development of 'interpretation-in-a-box' offerings, such as the Knosys 100.

T.S.: Software companies have tried many business models, including licensing, subscriptions and usage based. All have succeeded and struggled in different ways; none has broken through. Bioinformatics software continues to be heavily subsidized by venture capital, as evidenced by the genome-interpretation companies, which are largely service businesses. The most successful bioinformatics companies become part of a larger organization that can develop complete solutions from a large portfolio of assets.

S.B.: BlueSEQ recently surveyed the next-gen sequence analysis market and identified nearly 100 commercial solutions, and that's not taking into account the even larger number of open-source and academic tools. Needless to say, this is a dynamic market with a wide variety of business models, which range from specialized tools that address one specific need, to end-to-end pipelines that try to incorporate the entire workflow under a common user interface, to completely custom consultation services. The two hottest areas would be cloud-based analysis and genome interpretation. The cloud solutions are an attempt to simplify the whole process, setting up the entire analysis pipeline, data storage and data presentation with an easy-to-use interface hosted on the web. Genome interpretation focuses on turning the variant data into actionable medical information that can be used by the clinician. This is a growing market, with lots of new startups vying to establish themselves as a leader.

© 2012 Nature America, Inc. All rights reserved.

npg

platform can be augmented by each proprietary system, but we need a central coordinating site so that the FDA, GAO (US Government Accountability Office) and other agencies get consistent answers when they send out identical samples.

Jun Wang: We still need a good haplotype phasing technology to get a whole diploid genome. Nevertheless, sequencing a genome will not be a huge effort from now on. To precisely predict the function of variations in genome, a database of genome sequences from a large population should be built (for example, 1 million human genomes).

A.J.: A large database of genomes has to be publicly available and constantly updated. A better exchange of data and knowledge between different hospitals would help a lot. Furthermore, in all countries ethical guidelines have not been established on what should be reported back to the patient, on how sequencing data can be used and on how sequencing data should be stored.

E.W.: First, genotype and phenotype data need to be shared. If this cannot be achieved with everyone sending their genotype and phenotype data to a centralized resource (and it is likely that it cannot, owing to constraints on sharing of this type of highly sensitive and uniquely identifying data), then this will require the development of new tools for multisite querying of data. Second, an improved reference genome needs to be produced; this would ideally be a new type of construct that takes into account the variation present at each position.

S.B.: We need streamlined analysis and reporting so the clinician or physician sees only the information relevant to the condition being tested.

S.S.: The big challenge is a better understanding of biology, or rather a much more detailed understanding of how genotype relates to phenotype. I think we're only at the beginning of our understanding of how genetic variants affect our health.

M.S. & John West: We believe that existing technology already has value. However, higher accuracy, better coverage of difficult regions and better interpretation of the identified variants will all help medical sequencing. Perhaps most crucially, improved models for financing medical sequencing are very important.

Will clinical genome sequence analysis be done in-house at most large hospitals or outsourced?

Jeffrey G. Reid: Outsourcing is currently happening and is likely to be the model for the next several years. Until sequencing is more commoditized and there are appliances that make the technical aspects trivial, it is likely to stay that way, although eventually it will almost certainly be done in house at most large hospitals.

Jeffrey G. Reid, Research Assistant Professor, Baylor College of Medicine

E.W.: Clinical sequencing (whole-exome sequencing or whole-genome sequencing) will be carried out at mid-sized institutions. Once costs come down it will simply become another lab test. It will not be centralized because of such factors as delays with turnaround time, ability to control the pipeline (to select which cases needs a 2-day rather than 4-day turnaround, for example) and problems with movement of the data (everything else speeds up, but our bandwidth does not).

H. Craig Mak is Associate Editor, Nature Biotechnology