Genome Biology

MEETING REPORT

Out of the sequencer and into the wiki as we face new challenges in genome informatics

Zemin Ning[1]* and Stephen B Montgomery[2]

Abstract

A report on the joint Cold Spring Harbor Laboratory/ Wellcome Trust Conference 'Genome Informatics', 15-19 September 2010, Hinxton, Cambridge, UK.

Next generation sequencing (NGS) analysis, open-source software, cloud computing and wiki-style genomics were among the hot topics and discussions at the recent Genome Informatics meeting at the Wellcome Trust Genome Campus, Cambridge, UK. Here we summarize some highlights of the meeting.

Accuracy of polymorphism detection

Comparison of related genomes can generate a wealth of knowledge about genome evolution and function. Recent advances in NGS technologies have greatly increased the scale and scope with which we can interrogate novel genomes and uncover genetic variation. However, for variation detection and statistical analysis, there are false positive errors for various reasons, notably incompleteness of reference genomes, read mapping errors or limitations, and sequencing-induced features. Benjamin Dickins (Penn State University, University Park, USA) discussed an approach to estimate polymorphism accuracy from NGS data by deeply sequencing a small plasmid genome and comparing it with Sanger sequencing.

Elliott Margulies (National Human Genome Research Institute, Bethesda, USA) gave an enticing presentation on this topic entitled 'Analysis of identical twins' genomes reveals sources of false-positive variation detection'. With 55X and 50X depth of read coverage from each twin's sample, they initially identified 83,538 discordant genotype calls across 97.6% of the human reference genome. Through inspection of a random set of discordantly genotyped positions, he revealed that a majority occurred in regions with poorly aligned reads. Margulies noted that he would be highly suspicious of genotype calls in regions with high coverage but with low mapping scores. When these events were filtered out, the number dropped to 13,140, a reduction of 84%. By then further introducing other filtering mechanisms, such as incorrect alignments of short reads across indels, Q20 (99% confidence) evidence in the other twin and 10% allele frequency, Margulies' final number of discordant genotype calls was only in the range of 500 to 1,000.

Margulies' suspicion was certainly shared by Richard Durbin (Wellcome Trust Sanger Institute, Cambridge, UK), who told the audience, "If someone tells me that his accuracy on variant detection is 99%, I should be cautious". Durbin described that in the pilot phase of the 1000 Genomes Project, the sequencing strategy with low coverage (2X to 4X) worked well, and it can efficiently find shared sequence variation with good accuracy. The consortium has now started sequencing 2,500 people in five different ethnic groups. Furthermore, Durbin highlighted the recently launched UK 10K project, which, in collaboration with clinical investigators, will involve sequencing 4,000 cohort samples with rich phenotypes. With massive sample collections under way, further advances in informatics pipelines to reduce errors in variation detection will be indispensable.

Genome assembly and read alignment

NGS data provide challenging but exciting prospects for *de novo* assembly. There were several presentations that focused on development and evaluation of genome assemblers using such data. David Jaffe (Broad Institute, Cambridge, USA) outlined a practical and general laboratory/computational method for generating high-quality *de novo* genome assemblies. Three types of data were produced with the Illumina platform: 180 bp short-insert reads; 3,000 bp mate-pair reads; and 40,000 bp fosmid ends. The short-insert data were mainly intended to ensure contig base accuracy, whereas the longer jumping fragments provide potential for long-range connectivity. Jaffe generated data for mouse C57BL/6 and human Yoruban NA18507 and then compared the

*Correspondence: zn1@sanger.ac.uk
[1]Sequencing Informatics, The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK
Full list of author information is available at the end of the article

short-read assembler, ALLPATHS, with the reference assembly that used the capillary long-read assembler, Arachne. From the comparison of the results, the quality of short-read assemblies was found to be close to that of capillary assemblies, in terms of both base accuracy and contig connectivity.

Transcriptome data also offer great value and unique challenges in helping to identify and assemble gene sets for multiple species. Daniel Zerbino (University of California, Santa Cruz, USA) presented a new methodology, Oasis, which is built in with the popular short-read assembler Velvet. This new program takes in a preliminary assembly produced by Velvet and exploits the read sequence and pairing information to produce transcript isoforms. When possible, it also detects and reports standard alternative splicing events. It is specifically designed to get around the issues of unequal expression levels and alternative splicing breakpoints.

Short-read alignment saw further new developments at the meeting. New tools include: SMALT by Hannes Ponstingl (Wellcome Trust Sanger Institute, Cambridge, UK); NextGenMap by Fritz Sedlazeck (Max F Perutz Laboratories, Vienna, Austria); and YASRAT by Masahiro Kasahara (The University of Tokyo, Kashiwa, Japan). One distinguishing feature in these new alignment tools is that they all try to address the cases in which there is a higher level (over 2%) of base-pair errors in the raw reads.

Databases and data visualization

As sequence availability has increased, data access, representation, analysis and visualizations pose significant challenges in the community. Enis Afgan (Emory University, Atlanta, USA) has developed a solution that allows experimentalists to perform large-scale analysis using cloud-computing resources (http://usegalaxy.org/cloud). Because the solution is built on the open-source Galaxy framework, analyses using this service are accessible, transparent and reproducible. Furthermore, popular tools and workflows for sequence analysis from various types of experiment are already built in and ready to run. Marc Fiume (University of Toronto, Canada) presented the Sequence Annotation Visualization and ANalysis Tool (Savant), a fast and interactive genome browser that can display sequence, read alignment, single nucleotide polymorphism and other genomic datasets stored in standard file formats. Harminder Sehra (European Bioinformatics Institute, Hinxton, UK) described how the well established UniProt Knowledgebase (UniProKB) has developed from manual to automatic annotation. Marcin Piechota (Institute of Pharmacology, Krakow, Poland) described genes2mind.org, an online resource for the genomic profiling of psychoactive drugs. The database contains comparisons of the effects of various classes of psychoactive drugs on transcriptional alternations of about 20,000 genes in the mouse brain. And Maximillian Hauessler (University of Manchester, UK) presented text2genome, a method for annotating the genome using sequences that are reported in scientific publications.

Although NGS has required the development and improvement of many novel visualization tools, Martin Krzywinski (British Columbia Cancer Agency, Vancouver, Canada), the author of the Circos tool, took us back to basics by redefining network diagrams in his compelling talk in which he introduced a technique for turning network diagrams from incomprehensible 'hairballs' into information-rich schematics.

New insights from sequencing repetitive elements

Repetitive element sequencing is revealing some interesting new biology. Karen Hayden (Duke University, Durham, USA) demonstrated unique patterns of repeats in human centromeres, which could facilitate completion and understanding of specific biology in these hard-to-assemble regions. Kateryna Makova (Penn State University) highlighted the lifecycle of microsatellites and their potential functional impact near genes. And Mark Batzer (Louisiana State University, Baton Rouge, USA) showed from 37 orangutans clear differences in retrotransposon polymorphism, which could support the separation of the Sumatran from the Bornean orangutan.

Repeats are not always a good thing. There were also some presentations of genome analysis on large plant genomes. The highly repetitive wheat genome, with a size of 17 GB, poses significant challenges using the current sequencing platforms, in physical mapping as well as *de novo* assembly. Frederic Choulet (Institut National de la Recherche Agronomique, Clermont-Ferrand, France) described the effort to sequence and analyze wheat chromosome 3B. Contiguous bacterial artificial chromosome pools were sequenced by combining Roche 454 and Illumina platform sequencing. Shiran Pasternak (Cold Spring Harbor Laboratory, Cold Spring Harbor, USA) described the current progress on assembling the wheat D-genome, and Rachel Brenchley (University of Liverpool, UK) highlighted the advantages of comparative genome analysis in distinguishing genic loci from the abundance of repetitive content in the wheat genome.

RNA-seq and epigenomics

There was one dedicated session on RNA sequencing (RNA-Seq) and its application to genome annotation, in which various issues were discussed. Specifically, identifying the underlying biases in the RNA-Seq presents a significant analytical challenge. Projects such as the RNA-Seq Genome Annotation Assessment Project (RGASP) are examples of successful collaborations that are pushing the development of algorithms that can infer

accurate transcript quantification. RGASP, led by Jen Harrow (Wellcome Trust Sanger Institute, Cambridge, UK), and Felix Scheslinger (Cold Spring Harbor Laboratory) showed the utility of spike-ins in RNA-Seq experiments to achieve adequate normalization. Many talks also highlighted how RNA-Seq was used for novel interrogating of the impact of genetic variation, such as that of Emilie Graison (University of Lausanne, Switzerland) who demonstrated how RNA-Seq can be used to uncover interesting connections between copy number variants and transcript diversity.

Chromatin immunoprecipitation sequencing (ChIP-Seq) provides an extraordinary window into the dynamics of protein-DNA binding. Xin Feng (Stony Brook University, New York, USA) presented PeakRanger, a new algorithm to call peaks from ChIP-Seq data and explore complex peak structures in regions of interest. It works by identifying broad regions of enriched binding and integrates topological clues to detect narrow peaks.

Wikis can change the way we do science

The growth of sequencing data is far outpacing the rate of curation. Previously successful models are not meeting the demands of genome biology and, in many historical cases, the majority of annotations were from a small number of contributors. A view that was heard in the conference was that it is necessary to incorporate the community to a greater extent into the annotation process using tools such as wikis. Although lack of recognition and credit may prevent scientists from actively participating, this may be only part of the story. Reasons for not giving back to the community need to be investigated. In his keynote speech, Alex Bateman (Wellcome Trust Sanger Institute, Hinxton, UK) described new approaches to engage the community of RNA biologists to improve the annotation of the Rfam RNA database as well as more generally contribute to the annotation of RNA in Wikipedia. The journal *RNA Biology*, in collaboration with Rfam, has pioneered a new model of scientific publication where scientists are required to write a Wikipedia article to go alongside their manuscript paper describing new families of non-coding RNAs. At the same time, the Wikipedia article will also be under a full peer review process. Complementing this approach, Daniel Renfro (Texas A&M University, College Station, USA) presented community-focused annotation resources using the Mediawiki software. Currently this includes two model bacterial species: *Escherichia coli* (EcoliWiki) and *Bacillus subtilis* (SubtilisWiki).

In recent years, we have seen significant changes in the bioinformatics community: from massive data flow to complicated pipelines; from big computational farms to cloud computing; from a single genome or chromosome to many thousand genomes; from different sequencing platforms to various data types and error models. The themes reviewed here reflect the fact that the field of genome informatics is constantly under change, responding with new solutions, new algorithms and new technologies. The future presents us with a new challenge in how we transform the scale of data into collective knowledge, and it is likely that, as open-source software is accelerating bioinformatics and cloud computing is helping us to scale up, the growing online collaboration offers a chance to tackle this.

Acknowledgements
SBM is funded by the Louis-Jeantet Foundation.

Author details
[1]Sequencing Informatics, The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [2]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva 1211, Switzerland.

Published: 28 October 2010

doi:10.1186/gb-2010-11-10-308
Cite this article as: Ning Z, Montgomery SB: Out of the sequencer and into the wiki as we face new challenges in genome informatics. *Genome Biology* 2010, **11**:308.