Chambers *et al. Genome Biology* 2011, **12**:310
http://genomebiology.com/2011/12/12/310

Genome **Biology**

MEETING REPORT

Opening sequence: computational genomics in the era of high-throughput sequencing

Emily V Chambers*, Alida S Kindt and Colin AM Semple

Abstract

A report on the 11th Cold Spring Harbor Laboratory/ Wellcome Trust conference on Genome Informatics, Cold Spring Harbor Laboratories, New York, USA, November 2-5, 2011.

Keywords Bioinformatics, genomics, transcriptomics.

It has now been a decade since the completion of the human genome project, but it is clear that much of its biomedical potential is still to be realized. With the advent of low-cost, high-throughput sequencing (HTS) technologies, there is now an abundance of genomic, transcriptomic and epigenomic data. Projects to sequence thousands of new human genomes (UK10K project; http://www.uk10k.org/) and the genomes of thousands of new species (Genome 10K; http://www.genome10k.org/) are already underway. The flood of new data is causing major shifts in bioinformatics. Only a few years ago computational biologists would spend much time and effort processing flawed, publicly available data collections often poorly suited to the tasks at hand. Many of us remember human gene number estimates of 100,000 based upon expressed sequence tag (EST) databases. Nowadays everyone seems to be wrestling with large, novel datasets; their storage, visualization and analysis, and of course attempting to understand what they mean. At the same time, we are still in the midst of a chronic skills shortage in bioinformatics. One of the organizers of the 11th Cold Spring Harbor Laboratory/Wellcome Trust conference on Genome Informatics announced that the number of participants at this meeting had swelled to an all time record of 300. (Apparently the meeting also achieved a record proportion of female participants: a quarter.) And herein lies the paradox of bioinformatics in 2011: unparalleled opportunities to produce enormous sequencing datasets, matched by a limited number of people capable of extracting knowledge from them. This quandary is now feeding software development, with the aim of providing computational environments to enable data analysis by the computationally challenged: invariably icon-driven or menu-driven interfaces to command line tools. Work describing such developments was a major theme at this meeting.

As we get these first glimpses of the universe of human genomic variation, much emphasis is on reliable sequence variant calling and discovery. There is a more neglected question, ‘Which disease variants detected are most functionally meaningful?’ This was addressed by Marc Fiume (University of Toronto, Canada), whose talk focused on the development of MedSavant (http://genomesavant.com/medsavant/), a downloadable tool that attempts to discriminate disease-causing genetic variants from others, via a user-friendly graphical interface. This tool is one of many dynamic and interactive platforms that can analyze, integrate and visualize data for a wider audience of researchers, and are designed to handle the unprecedented amount of new data now available. Several new applications featured in the meeting, as well as new features for more established platforms such as Galaxy (http://galaxy.psu.edu/). Jeremy Goecks (Emory University, USA) introduced the new Galaxy Track Browser (http://main.g2.bx.psu.edu/), which aims to integrate more powerful HTS data analysis tools for real-time manipulations of large datasets in Galaxy, while Ting Wang (Washington University, USA) and Christoph Bock (Broad Institute, USA) described the new interactive tools Human Epigenome Browser (http://epigenomegateway.wustl.edu/) and EpiExplorer (http://cosgen.bioinf.mpi-inf.mpg.de/welcome.php). The Human Epigenome Browser allows users to integrate their own data with datasets from the NIH Roadmap Epigenomics project (http://www.roadmapepigenomics.org), one of the current, large-scale efforts to comprehensively map the epigenomic landscape of human cells. EpiExplorer is a new web tool designed to search and integrate epigenomic and genomic annotations for analyses of custom datasets (again in real time), using tricks from Google search algorithms to speed things up. Although all of these tools

*Correspondence: emily.chambers@hgu.mrc.ac.uk
MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK

have different functions, they all have the same goal: to make analysis of data more accessible and manageable.

It is not only the genomic complexity of normal cells that is emerging, as the focus shifts to include variation among cell types and in aberrant states, such as cancers. Detecting variants in cancer usually involves mapping sequence reads from cancer cells to a reference sequence and comparing the results to mapped reads from 'normal' cell samples, preferably taken from the matched healthy tissue of the same individual. As with most assembly algorithms, misalignment can be the cause of false variant calls and real variants may be missed. Jared Simpson (Wellcome Trust Sanger Institute, UK) discussed new approaches to the identification of informative variants. His methods are based upon direct comparisons of normal and tumor samples rather than comparisons to the human genome reference sequence. This allows the calling of more complex structural variants between breast cancer and normal samples, since some classes of sequence variation are known to be under-represented in the reference sequence. A novel approach to studying cancer was described by Mamoru Kato (Cold Spring Harbor Laboratories, USA) who has applied population genetics to identify copy number variations (CNVs) under natural selection in breast cancer cell lines. It seems that most observed variations are selectively neutral, with a small number subject to selection during clonal evolution, and this suggests that such measures may be useful in flagging the most important alterations in cancer progression.

RNA was also a recurring topic of discussion, with several speakers focusing on transcriptomics. Mitchell Guttman (Broad Institute, USA) described recent work on long intergenic non-coding RNAs (lincRNAs), which seem to preferentially associate with chromatin regulatory proteins. The functionality of these lincRNAs is disputed, as previous knockdown experiments have shown few phenotypic consequences. However, knockdown of ESC-expressed lincRNAs has a strong effect on gene expression and pluripotency. This has led to the hypothesis that lincRNAs may interact with chromatin regulators to maintain the ESC state. No modern discussion of RNA would be complete without microRNAs, and Zhi-Qiang Du (Iowa University, USA) described the lineage-specific expansion of microRNA families during evolution and their influence on a myriad of complex traits such as reproduction and species-specific traits in pigs. Of course there are still many areas of RNA biology that are poorly understood or unexplored. Jakob Pedersen (Aarhus University, Denmark) introduced many new families of human regulatory RNA structures to the audience. A new comparative method, EvoFam (http://moma.ki.au.dk/prj/mammals/), has been used to identify constrained RNA structural families by their primary sequences and secondary structures. Further statistical analysis, such as gene ontology (GO) term enrichments, was used to generate hypotheses about potential functions for the new families identified.

Steven Salzberg (Johns Hopkins University School of Medicine, USA) reported the (only slightly exaggerated) mantra of genomics in 2011, 'Sequencing is free - so let's sequence everything'. However, sequencing everything introduces a lot of complexity to genome assembly. We lack an accurate reference genome for most species, and so there is enormous demand for new short-read assembly algorithms that are fast and memory efficient. The rapid evolution of sequencing technology means that older assemblers, used on longer sequence reads, are no longer appropriate. It is therefore important to continually develop and reassess assembling methods, especially with a view to ambitious projects such as the Genome 10K project, which aims to sequence and assemble 10,000 vertebrate genomes by 2015.

With several new assembler methods now available, there is considerable interest in comparing them to investigate which assembler is most appropriate for a particular application. Benedict Paten (University of California, Santa Cruz, USA) described the Assemblathon competition (http://www.assemblathon.org/), which compared several new assembly methods on current sequencing datasets. A total of 17 teams participated in the contest to assemble a complex genome from simulated data to examine, among other factors, their levels of coverage and contiguity. Salzberg presented the second comparison study: GAGE (Genome Assembly Gold-Standard Evaluations; http://gage.cbcb.umd.edu). This project compared several different assembly algorithms on sequencing datasets from a variety of species, including bacteria, invertebrates and vertebrates. They aimed to guide researchers planning sequencing projects in the extent of coverage needed, which assembler and parameters were most appropriate for their experimental designs, and species of interest. Both GAGE and Assemblathon highlight the large degree of disagreement between the different assemblers used, and suggest that genome assembly is far from a resolved issue. GAGE identified the assembler ALLPATHS-LG (http://www.broadinstitute.org/software/allpaths-lg/blog/) as a clear winner over all species that were tested, while the Assemblathon project identified SOAPdenovo (http://soap.genomics.org.cn/soapdenovo.html) as the best assembler, having the best overall score and highest coverage.

Keynote speaker Evan Eichler (Washington University, USA) discussed the occurrence of CNVs in the human genome; this is remarkably common (each of us carry several hundred variations greater than 5 kb long) and has been linked to a variety of diseases such as psoriasis and autism. In spite of this, only a small proportion of this variation is assayed by widely used SNP genotyping platforms. The human reference genome sequence necessarily provides an incomplete picture in regions harboring CNVs, as assembly algorithms will often collapse multiply duplicated regions into a single, smaller copy. Eichler described his and others' laborious efforts to fill these gaps using novel approaches. Longer sequencing reads are needed to reliably incorporate CNVs and repeats into the reference genome, and *de novo* human genome assemblies can be significantly shorter in length than the reference genome because of these missing sections. However, long reads are more costly to produce. In order to study gene CNV, Eichler has used fosmids, a sequencing vector capable of containing 40 kb genomic DNA, to investigate these previously inaccessible parts of the human genome. This in-depth re-sequencing of some loci has revealed exciting new regions subject to CNVs between human populations not observed in the reference sequence, containing genes involved in immunity and brain development. Eichler finished with a plea for a high-quality, high-coverage reference genome for every biomedically important species.

Abbreviations

CNV, copy number variation; EST, expressed sequence tag; GAGE, Genome Assembly Gold-Standard Evaluations; HTS, high-throughput sequencing; kb, kilobase; lincRNA, long intergenic non-coding RNA; SNP, single-nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

EVC and ASK are funded by MRC Capacity Building PhD Studentships; CAMS is funded by the MRC.

Published: 28 December 2011

doi:10.1186/gb-2011-12-12-310
Cite this article as: Chambers EV, *et al.*: Opening sequence: computational genomics in the era of high-throughput sequencing. *Genome Biology* 2011, 12:310.