Genome Biology

MEETING REPORT

Eastern genomics promises

Scott C Edmunds*

Abstract

A report on the Bio-IT World Asia meeting, Marina Bay Sands, Singapore, 6-8 June 2012.

In this era of bigger and more globalized biology, Bio-IT World has spread this year - the 10th anniversary of its conference series - from its usual Boston base to Singapore for its first Asian meeting. Substituting the usual New England lobster for chili crab, the conference's proximity to the Singapore Biopolis and booming Chinese and Asian biotech and research sector, and the charms of the Marina Bay Sands resort were a winning combination in getting an impressive mix of scientists and tech evangelists from around the Asia-Pacific region and beyond to attend and present.

Following the usual focus of Bio-IT World meetings on the application of information technology (IT) and informatics to biomedical research and drug discovery, the many talks on pharmacological and personalized genomics usefully concentrated on Asia-specific diseases and pharmacogenomic and genetic variation that can be understudied in the rest of the world. The meeting's other focus on big data may have not been quite so geographically tailored: many of the talks on data transfer and cloud services were presented from a US perspective without addressing the special challenges that lie across the Pacific in cloud vendors' edge locations and behind 'great firewalls.'

With conference tracks on IT infrastructure and the cloud, drug discovery informatics, bioinformatics, and data management and interpretation for next-generation sequencing, there were several themes that crossed most of the sessions.

The inevitable talks on big data

Chris Dagdigian (Bioteam), in an opening keynote entitled 'Bio-IT trends from the trenches', summarized the current strategies for handling big data and set the scene for the whole meeting. Predictably, however, the main message to come from this and other talks was issues with scale. As increasing numbers of petabyte-scale data systems are deployed and the latest network-based data movement using Aspera and GridFTP show impressive gains over physical (postal or hand delivery) data movement (at least using US networks), research centers are just about keeping their heads above water in dealing with data growth. With chemistries changing faster than data centers and research IT infrastructure can be refreshed, Dagdigian was understandably pessimistic about sustainability, but the work on CRAM compression algorithms from the European Bioinformatics Institute (http://www.ebi.ac.uk/ena/about/cram_toolkit) was cited as one possible hope for the future. He was also dismissive of any cloud computing vendors lacking compatibility with the Amazon application programming interface (API). With the bold claim that these API-less 'cloud pretenders', as well as platforms lacking self-service, do not equal a cloud, Dagdigian left the many following cloud-oriented talks with a tough act to follow. One subsequent speaker took these conclusions to heart and had to admit that what he was presenting was 'more of a fine mist' than a cloud.

Several cloud-oriented vendor talks still attempted to rise to this challenge. Representatives from Amazon Web Services, IBM, BT and Appistry highlighted their latest services, and Xing Xu presented on the new Easy Genomics cloud-based bioinformatics platform from the BGI Cloud team. The trial version of Easy Genomics has an attractive data analysis platform and an Aspera connection, and it includes six graphics processing unit-based and cloud-optimized sequencing tools, including the new Hadoop-optimized version of BGI's popular SOAP genome assembly suite (http://www.genomics.cn/FlexLab/html/gaea.html).

Closing the reproducibility gap

Of the many critical issues coming out of the data-rich universe that we now find ourselves in, James Taylor (Emory University) focused his talk on what feels is the main crisis in genomics research reproducibility. With the life-sciences increasingly reliant on computational and data-driven approaches, access to the supporting data and tools and accessibility in using computational resources has not kept pace. With this in mind, Taylor honed in on the ways in which the popular Galaxy

*Correspondence: scott@gigasciencejournal.com
GigaScience, BGI Hong Kong Ltd, 16 Dai Fu Street, Tai Po Industrial Estate, NT, Hong Kong

workflow environment (http://galaxy.psu.edu/) is working to address the problem.

The successes and challenges faced by Galaxy mirror on a smaller scale those of genomics as a whole: the 600 TB of data they host and tens of thousands of analysis jobs a month they are now handling is causing inevitable strains to their infrastructure. Taylor outlined how these pressures should be eased as increasing numbers of users use Galaxy Cloudman (http://usegalaxy.org/cloud), where they can take advantage of the elasticity and suitability as a reproducibility platform that a cloud-based platform with pre-configured software and a user-friendly interface can provide. The challenges for the Galaxy team in keeping on top of the increasing numbers of bioinformatics tools available has also been aided by Galaxy users wrapping and adapting tools for the Galaxy Toolshed, a directory of nearly 2,000 tools - over 10 times greater than those available through the main Galaxy site.

The growing popularity of Galaxy was apparent from the number of talks presenting work using it. For example, Andrew Lonie of the Victorian Life Sciences Computation Center (Melbourne) spoke about the Australian Genomics Virtual Library, which uses Galaxy and Bio-Linux in Australia's national research cloud, NeCTAR (http://www.nectar.org.au/). William Bartlett described Galaxy's integration into the computational and data management architecture at the National Center for Genome Analysis Support in Indiana, and Tin-Lap Lee also promoted the Galaxy platform that he and collaborators are putting together at the Chinese University of Hong Kong to handle data from the BGI and its *GigaScience* journal and database. Despite Galaxy being an open-source platform, even speakers from the pharmaceutical industry demonstrated its use, with Yaron Turpaz from Astra Zeneca highlighting the Cistrome platform recently published in *Genome Biology* (http://cistrome.org/ap).

Genomes, genomes, genomes

Several talks highlighted the wealth of genomes that are now publicly available. Yaron Turpaz presented the Asian Cancer Research Group's work on cancers prevalent in Asia and the recent public release of 88 paired hepatocellular carcinoma and normal genomes (available from *Giga*DB: http://dx.doi.org/10.5524/100034). Richard Tearle from Complete Genomics outlined their publicly available datasets, including 69 individuals and two matched tumor-normal pairs (http://www.completegenomics.com/public-data/).

In light of the meeting's location, it was refreshing to see so much Asian genomic data on display and being publicly released. Jong Bhak (SNU/Theragen) provided an example of this in the rapid growth in the number of publically available Korean genomes: from Seong-Jin Kim's genome - the first Korean to be sequenced, in 2010, and incidentally also a speaker at the meeting - via 20 individuals in 2011, to the current 38 individuals, and on to the goal of sequencing 10,000 genomes over the next 3 years to capture all of the genetic variation in the relatively homogeneous Korean population. Being an open data advocate and influenced by his time in the laboratory of George Church, Bhak has made the data available from the Korean node of the Personal Genome Project (http://opengenome.net). Use of these data is available under his totally public domain BioLicense waiver (http://biolicense.org/), pioneered years before the recent interest in open data licenses and portable consent for personal genomics data.

A more genetically heterogeneous example was presented by Stephen Rudd of MGRC Malaysia, who presented on the MyGenome Malaysian genome project, which has so far sequenced 26 genomes from 6 of the many diverse ethnicities making up the Malaysian population. With the proliferation of new '-ome' words showing no signs of abating, the MGRC has delivered their own contribution to the field: the 'corporate pan-genome'. Following a company-wide 'spit party', all 50 employees of the sequencing company were sequenced at 2x coverage. Rudd used this fun project as a demonstrative starting point to discuss topics as diverse as contamination of saliva with someone's breakfast to the unique organizational occupational health insights that can be gleaned from this work. Although currently this may be an unusual team-building exercise, it is a window into the types of creative projects that are likely to be increasingly feasible as we move towards the 1,000 ringgit genome.

The scale and quality of the research presented were of a high international standard, but this conference still managed to impart some refreshing local flavor. The recurring themes of the promise and challenges associated with large-scale biological data are clearly global, but given that the Bio-IT World conference now spans three continents (including an upcoming European meeting in Vienna), the series provides opportunities to follow how researchers are tackling these issues from distinct regional perspectives. The addition of an Asia-Pacific meeting hopefully gives this previously underrepresented research community an opportunity to make their voices better heard and to work more closely on issues and genetic models that are more relevant to the region.

Competing interests
The author is an employee of the BGI and collaborates with a number of speakers at the meeting.

Acknowledgements
SCE would like to thank Alexandra Basford for her feedback and comments.

Published: 17 July 2012

doi:10.1186/gb-2012-13-7-317
Cite this article as: Edmunds SC: Eastern genomics promises. *Genome Biology* 2012, **13**:317.