

Call-in Details please announce yourself when entering  
Monday, 9 February 2009, 11AM ET, 8am PT, 5pm Central Europe

## Topics:

- A. *support for short-read sequencing*
- B. *Distributing bioinformatics tools for Academic use via clouds?*

## Discussion about Support for Short-Read Sequencing

We discussed practical aspects of helping biologists access and make sense of data from their short-read sequencing experiments. We addressed two questions:

1. What standard processing ought we do beyond aligning reads to a genome? What kinds of additional processing can be standardized? What kinds of customized analyses should we offer and what is the best way to do so?
2. Can and should the short-read sequencing community develop a collaborative, open-source LIMS, rather than having each site roll its own?

People who participated (sorted by initials; apologies if I mis-identified any voices):

- Andy Arenson, Indiana University - Purdue University
- Andrea Hoerster
- Brent Richter, Partners Healthcare Systems
- Charlie Whitaker, MIT
- David Lapointe, University of Massachusetts Medical School
- David Sexton, Vanderbilt
- Dick Repasky, Indiana
- Fran Lewitter, MIT
- George Bell
- Hemant Kelkar, University of North Carolina at Chapel Hill
- Hershel Safer, Weizmann Institute of Science
- Jyothi Thimmapuram, University of Illinois
- Kip Bodi, Tufts
- Michael Rebhan, Friedrich Miescher Institute for Biomedical Research

## Processing for short-read sequencing

DS: We do processing through genome alignment as standard for everybody. Everything beyond this is custom, since everybody wants something different.

MR: Folks here find Galaxy useful for subsequent analysis.

HS: We offer the possibility of uploading non-standard regions to use as the basis for genome alignments, since some users are doing "odd" genomes. I use galaxy for analyzing ChIP-Seq data.

BR: A group at Brigham & Women's Hospital uses Galaxy as the core of a tool for analyzing association studies (genotyping).

HK: We're doing a lot of ChIP-Seq. We don't do any custom analyses, but we just hired another person and hope to start custom work soon.

KB: We are using CLC Genomics Workbench. It has a good, free viewer. Folks mostly use it for SNP hunting. It's a bit slow but is convenient for users.

HK, FL: We have found CLC to be of limited usefulness. It's not ready for prime time when working on human or other mammals.

HS: We hope to buy the Genomatix workbench for ChIP-Seq analysis. This will also give us a local copy of the entire Genomatix suite for work with transcription factor binding sites. But it's expensive -- Euro 80k to buy and Euro 25k annually for maintenance. We'll use it for consulting but users should be able to use it as well for literature searches, building networks, etc.

MR: Are any R packages available for use with short-read sequencing?

DL: A package called Bioseq is being developed as part of Bioconductor.

DL: Is anybody using Phred & Phrap for short-read sequencing? Can Consed display so many reads? Are other HPC tools available for free?

DS: Galaxy is probably the best free tool available.

BR: Consed version 18 works with large numbers of reads.

DS: We have scripted the Illumina Pipeline and alignment to be run automatically. It is integrated with our home-grown LIMS, which tracks samples and files.

## **LIMS for short-read sequencing**

DS: A collaborative LIMS would be a good idea. One difficult issue will be choice of language. We use Ruby on Rails, but I wonder if enough people know it or are willing to learn it.

KB: We also use Ruby on Rails. It is a great platform for creating a web-based LIMS.

AA: Is the objective of LIMS development a tool that can be used almost everywhere, or is it to create a basis with 80% of what's needed that can be easily modified for local use? An middle ground between open-source and commercial is community-source: it's developed collaboratively and then others pay for access.

DS: I am somewhat skeptical that folks will pay for access to something like this; they'll likely prefer to pay for their own. This also needs somebody to lead the project, administer it, and collect the money.

AA: This approach will only work if the project is big enough to make it worthwhile for somebody to take on the administration.

DS: Sanger is close to making its LIMS open-source. It's developed in Ruby on Rails. It is specific to sequencing, and is used for their Sanger and short-read sequencing projects.

HS: Folks who are doing short-read sequencing now already need LIMS systems. By the time a collaborative effort has something to show, everybody will already have one.

DS: The community would benefit if those who roll their own make them open-source.

### **What are folks giving to end users?**

DS: We're giving them what they ask for. In most cases, this is just the analysis results, though some ask for sequence files as well. We're also give wiggle files for display in the UCSC Genome Browser.

KB: Whatever they ask for. We start by providing seq files and do analysis if asked.

HK: We've had to provide entire datasets (everything except images) for people to submit to public repositories in order to publish (at least in Nature).

### **Discussion about ePHI and in the Cloud.**

One of the foundations of Bioinformatics research and development is the availability of open source tools. Dissemination of knowledge and encouragement of collective development has been a tremendous benefit for the community. Installing and deploying those tools, however, are often non-trivial tasks for even the most seasoned bioinformaticians. The major difficulty usually derives from technical dependencies of the tools on system and software libraries that may not be packaged with the distribution.

Today, with the wide use of Linux and Virtual Machine (VM) technologies, the possibility of distributing the tool as a VM, complete with all the necessary dependencies included in the OS, is possible.

### **Direct experiences with such systems? What are the utilities?**

DS: I have taken a look at Amazon S3 in offloading image files to the storage cloud for a couple of months. I quickly found the cost is more expensive than what is available internally at Vanderbilt.

DS: It comes to \$10K per year for 10TB. 10TB is \$1500 per month. Additionally, I have a lot of data going in and out of S3, this add an additional \$276 per month in I/O.

CW—Yes, it come to 15 cents per GB.

AA—VM's are cheaper than standalone servers, but does this help the investigator? it is true that it could help, but there is still a question as to who will maintain the system after deployment. With Amazon, every time you use a VM--does the backend change? It is stateless between sessions.

### **Does it make more sense to host a cloud in the core?**

DL: VM's are good to put up a website, One could also place processing on desktop. However, vmware is underpowered on desktops. With most researchers tending to place academic-ware on their desktops, the investigators may not be able to get the best use of a virtual build. It is best to put the builds on servers. Most investigators do not have access to these kinds of resources.

BR: An internal cloud makes sense here, then. At Partners we have an internal cloud for just these reasons—really out of necessity. With security concerns, the cost of transferring data back and forth, and providing a resource for investigators, a small internal cloud makes a lot of sense, especially if it is a service that is hosted in conjunction with an amazon cloud.