Towards a Pushbutton Human Genome

Barry Merriman[1,2], Brian O'Connor[3,4], & Timothy Harkins[1]

1 Life Technologies, Inc., Carlsbad, CA, USA
2 UCLA, Department of Human Genetics, Los Angeles, CA, USA
3 University of North Carolina at Chapel Hill, Lineberger Cancer Center, Chapel Hill, NC, USA
4 Nimbus Informatics LLC, Carrboro, NC, USA

life technologies™ http://lifetechnologies.com

nimbus INFORMATICS http://nimbusinformatics.com

Abstract

The current operating costs of next-gen sequencers enable routine generation of the *amount of data* needed to resequence a human genome. However, the informatics infrastructure needed to efficiently process this data into *useful information* is still missing, and this becomes the major barrier to adoption of genome sequencing in the research and clinical communities. An ideal "pushbutton" solution would eliminate both the burden of maintaining a computer installation, and of retaining a team of informaticians. Here we describe one such solution, designed on the principles of cloud computing, open source software, data security, automated analysis, and databasing of results. Relying on commercial cloud computing removes the computer hardware and system administration burdens, open source software allows for transparency and leverages the largest development community, data security provides the IRB and HIPAA compliance researchers and clinicians require, a fully automated pipeline gives fast, consistent and auditable processing, and placing the results in a scalable variant database supports interactive querying of the results, aggregation of information across subjects, and diverse downstream analyses. The project uses the Amazon Elastic Compute Cloud (EC2) for the computational substrate, the SeqWare[1] open source tools for the software infrastructure, and extensive encryption of data for IRB and HIPPA compliance. From the users' perspective, it provides a pushbutton workflow for human genome or exome sequencing, currently optimized for the SOLiD platform: starting from sending an encrypted hard drive of raw data to EC2, it performs alignment, re-alignment, variant detection, databasing, and variant annotation. We present an overview of the system design, an analysis of the costs of processing genomes in this way, and examples from collaborative projects that show the power for variant discovery research. Users are free to create their own instance of the open source system on EC2, or utilize a full-service, commercially supported instance[2].

1 SeqWare Query Engine: storing and searching sequence data in the cloud. Brian O'Connor, Barry Merriman and Stanley Nelson. BMC Bioinformatics 2010, 11(Suppl 12):S2

2 Nimbus Informatics, LLC sponsors the open source SeqWare project and provides commercial support, data hosting, and push-button utilization of SeqWare through an easy-to-use web interface. This service is currently in private beta, see http://nimbusinformatics.com for more information.

SeqWare Architecture

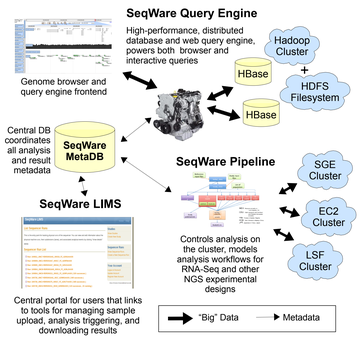


Figure 1: An overview of the SeqWare Architecture

The SeqWare open source project consists of four sub-projects. The **SeqWare MetaDB** provides a central repository to track samples and analysis. It is visualized using the **SeqWare LIMS** which displays sample annotations and analytical results. The **SeqWare Pipeline** is based on the cluster-agnostic Pegasus workflow engine and is responsible for automated data processing. Finally, **SeqWare Query Engine** is the interactive gateway for results and links to genome browsers and tools. These components form the basis of a comprehensive sequence analysis platform.

Cloud Security & Encryption

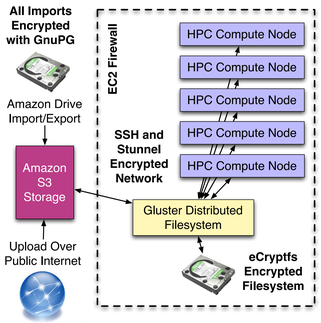
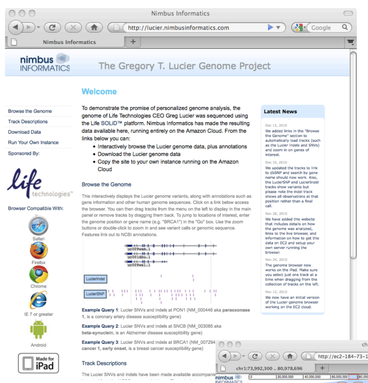


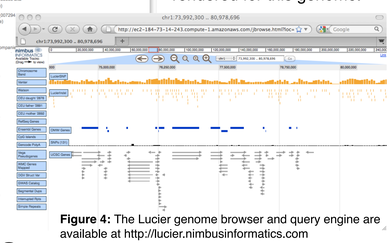
Figure 3: Security in the Amazon Cloud

The import of sequence data on a private cluster install of SeqWare is straightforward. However, maintaining IRB and HIPAA compliance on public clouds such as EC2 requires careful security procedures. SeqWare provides tools to facilitate the encryption of data into and out of the cloud. This includes encryption of sequence data sent by users and results sent back, encrypted network connections and filesystems on compute nodes, and following Linux security best practices.

Reporting & Querying



The **SeqWare Query Engine** allows for the databasing of annotated variants and querying in real time. For example, variants can be filtered by consequence, depth, pathway membership, and other attributes. Results can be loaded in a genome browser or saved as a file for use with other tools. The **Lucier browser** is an example of a complex report generated by the Query Engine in which a collection of JBrowse tracks have been rendered for this genome.



Outputs of the Query Engine, such as a full genome browser, can be cloned to other EC2 instances allowing for both scalability and a means for sharing results with others.

Figure 4: The Lucier genome browser and query engine are available at http://lucier.nimbusinformatics.com

Cloud Computing Costs

Genome

Step	Current Timing	Current Cost	Projected Timing	Projected Costs
Data Import (drive import)	~2-4 days	$100	~2-4 days	$100
Temp Data Storage	~1 week	$300	~3 days	$100
HPC Computation Time	5 days, 8 nodes	$1,500	3 days, 8 nodes	$900
Total	~8 days	$1,900	~6 days	$1,100

Item	Period	Cost
Recuring Storage Costs	per month	$5
Recuring Archive Costs	per month	$45
Recuring Hosting Costs	per month	$15
Total	per month	$65

Exome

Step	Current Timing	Current Cost	Projected Timing	Projected Costs
Data Import (upload)	~1 days	$10	~1 days	$10
Temp Data Storage	~3 days	$30	~1 days	$10
HPC Computation Time	2 days, 2 nodes	$150	1 days, 2 nodes	$75
Total	~2 days	$190	~1 day	$95

Item	Period	Cost
Recuring Storage Costs	per month	$5
Recuring Archive Costs	per month	$5
Recuring Hosting Costs	per month	$15
Total	per month	$25

In addition to security, cost is a major concern for researchers looking towards the cloud for computational resources. We have found the costs of EC2 to be exceptionally reasonable especially when local computational resources are unavailable. Currently it costs under **$2,000 to analyze a whole human genome on EC2** (100GBase) and, with optimizations to SeqWare, we expect that number to decrease to approximately **$1,000 per genome**.

Open Source Science

The SeqWare project is hosted on SourceForge (http://seqware.sf.net) and brings together the best tools for cloud computing and sequence analysis. For scientists and clinicians, the open source nature of SeqWare translates to greater flexibility and transparency than commercial, closed source offerings. Many of the tools built or employed by SeqWare are published, or used in published studies, and this allows for a clear understanding of the performance characteristics and capabilities of each algorithm. The use of open source also means anyone is able to extend and contribute to the project by writing new analysis modules and workflows. This freedom to extend the system is coupled with the ability to run SeqWare in a variety of environments: deployed on a private cluster, on Amazon's cloud (EC2), or through a web-based interface hosted by Nimbus Informatics. These last two options provides a low-cost, on-demand ability to scale analysis as needed with the hosted option requiring no technical expertise. SeqWare is used by several groups including the Cancer Genome Atlas group at UNC, Life Technologies' Collaborations group, and UCLA's Genomics Core.

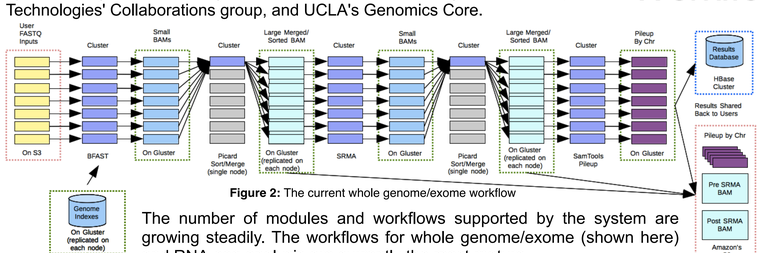


Figure 2: The current whole genome/exome workflow

The number of modules and workflows supported by the system are growing steadily. The workflows for whole genome/exome (shown here) and RNA-seq analysis are currently the most mature.

Workflow Automation

A key feature of SeqWare is the highly integrated nature of the various subprojects. For example SeqWare MetaDB stores sample and sequencer run annotations making it possible to create rules to automate the launching of workflows. UNC's deployment of SeqWare includes scripts that periodically look for new RNAseq lanes and, when found, launch the RNAseq workflow. This system has processed over **1.4TBase** of sequence in the last 6 months. Similarly SeqWare, as deployed by Nimbus Informatics, allows users to link uploaded sequence to workflows in a "pushbutton" fashion. This system has currently processed **6 whole human genomes on EC2** and is supporting a Life Collaborations' project to analyze HuRef at **310x coverage** on the SOLiD platform.

Conclusions

Massively parallel sequencing technologies have opened the door to a future where thousands of genomes will be sequenced on a regular basis. A considerable challenge exists for how to analyze and query the massive amount of data produced. Here we have presented SeqWare, a flexible open source framework designed to run both on a local environment and the cloud, as a potential solution. The project is fully open source and available from **http://seqware.sf.net** for deployment locally or on EC2. Nimbus Informatics LLC provides a commercially-supported, easy-to-use web-based version of SeqWare backed by Amazon's EC2 cloud (currently in private beta, please see **http://nimbusinformatics.com**). SeqWare development is supported by Nimbus Informatics LLC, Life Technologies Inc., users including UNC, UCLA, and the Cancer Genome Atlas, and the open source community.