
the IT architects

# Whitepaper

## Genomic Processing in the Cloud

### Development of a Cloud-based solution for the Spanish National Cancer Research Institute (CNIO)

Author: Gerardo Viedma, Senior Consultant, The Server Labs

December 2010. info@theserverlabs.com

## Introduction

Over the last decade, a new trend has manifested itself in the field of genomic processing. With the advent of the new generation of DNA sequencers has come an explosion in the throughput of DNA sequencing, resulting in the cost per base of generated sequences falling dramatically. Consequently, the bottleneck in sequencing projects around the world has shifted from obtaining DNA reads to the alignment and post-processing of the huge amount of read data now available. To minimize both processing time and memory requirements, specialized algorithms have been developed that trade off speed and memory requirements with the sensitivity of their alignments. Examples include the ultrafast memory-efficient short read aligners BWA and Bowtie, both based on the Burrows-Wheeler transform.

The need to analyse increasingly large amounts of genomics and proteomics data has meant that research labs allocate an increasing part of their time and budget provisioning, managing and maintaining their computational infrastructure. A possible solution to meet their needs for on-demand computational power is to take advantage of the public cloud. With its on-demand operational model, labs can benefit from considerable cost-savings by only paying for hardware when needed for their experiments. Procurement of new hardware is also simplified and more easily justified, without the need to expand in-house resources.

Not only does the cloud reduce the time to provision new hardware; it also provides significant time-savings by automating the installation and customization of the software that runs on top of the hardware. A controlled computational environment for the post-processing of experiments allows results to be more easily reproduced, a key objective to researchers across all disciplines. Results can also be easily shared among researchers, as cloud-based services facilitate the publishing of data over the internet, while allowing researches control over their access. Finally, data storage in the cloud was designed from the ground-up with high-availability and durability as key objectives. By storing their experiment data in the cloud, researchers can ensure their data is safely replicated among data centres. These advantages free researchers from time-consuming operational concerns, such as in-house backups and the provisioning and management of servers from which to share their experiment results.

Given the vast potential benefits of the cloud, The Server Labs is working with the Bioinformatics Department at the Spanish National Cancer Research Institute (CNIO) to develop a cloud-based solution that would meet their genomic processing needs.

## An Environment for Genomic Processing in the Cloud

The first step towards carrying out genomic processing in the cloud is identifying a suitable computational environment, including hardware architecture, operating system and genomic processing tools. CNIO helped us identify the following software packages employed in their typical genomic processing workflows:

- **Burrows-Wheeler Alignment Tool** (BWA, http://bio-bwa.sourceforge.net/bwa.shtml): BWA aligns short DNA sequences (reads) to a sequence database such as the human reference genome.
- **Novoalign** (http://www.novocraft.com/download.html): Novoalign is a DNA short read mapper implemented by Novocraft Technologies. The tool uses spaced-seed indexing to align either single or paired-end reads by means of Needleman-Wunsch algorithm. The source code is not available for download. However, anybody may download and use these programs free of charge for their research and any other non-profit activities as long as results are published in open journals.
- **SAM tools** (http://samtools.sourceforge.net): After reads alignment, one might want to call variants or view the alignments against the reference genome. SAM tools is an open-source package of software aplications which includes an alignments viewer and a consensus base caller tool to provide lists of variants (somatic mutations, SNPs and indels).
- **BEDTools** (http://code.google.com/p/bedtools/): This software facilitates common genomics tasks for the comparison, manipulation and annotation of genomic features in Browser Extensible Data (.BED) format. BEDTools supports the comparison of sequence alignments allowing the user to compare next-generation sequencing data with both public and custom genome annotation tracks. BEDTools source code in freely available.

Note that, except for Novoalign, all software packages listed above are open source and freely available.

One of the requirements of these tools is that the underlying hardware architecture is 64-bit. For our initial proof of concept, we decided to run a base image with Ubuntu 9.10 for 64-bit on an Amazon EC2 large instance with 7.5 GB of memory, 4 EC2 Compute Units (2 virtual cores with 2 EC2 Compute Units each) and 850 GB of local instance storage. Once we had selected the base image and instance type to run on the cloud, we proceeded to automate the installation and configuration of the software packages. Automating this step ensures that no additional setup tasks are required when launching new instances in the cloud, and provides a controlled and reproducible environment for genomic processing.

By using the RightScale cloud-management platform, we were able to separate out the selection of instance type and base image from the installation and configuration of software specific to genomic processing. First, we created a Server definition for the instance type and operating system specified above. We then scripted the installation and configuration of genomic processing tools, as well as any OS customizations, so that these steps can be executed automatically after new instances are first booted.

Once our new instances were up and running and the software environment finalized, we executed some typical genomic workflows suggested to us by CNIO. We found that for their typical workflow with a raw data input between 3 and 20 GB, the total processing time on the cloud would range between 1 and 4 hours, depending on the size of the raw data and whether the type of alignment was single or paired-end. With an EC2 instance pricing at 38 cents per hour for large instances, and ignoring additional time required for customization of the workflow, **the cost of pure processing tasks totaled less than $2 for a single experiment**.

We also **found the processing times to be comparable to running the same workflow in-house** on similar hardware. However, when processing on the cloud, we found that transferring the raw input data from the lab to the cloud data centre could become a bottleneck, depending on the bandwidth available. We were able to work around this limitation by processing our data on Amazon's European data centre and avoiding peak-hours for our data uploads.

Below, we include an example workflow for paired-end alignment that we successfully carried out in the Amazon cloud:

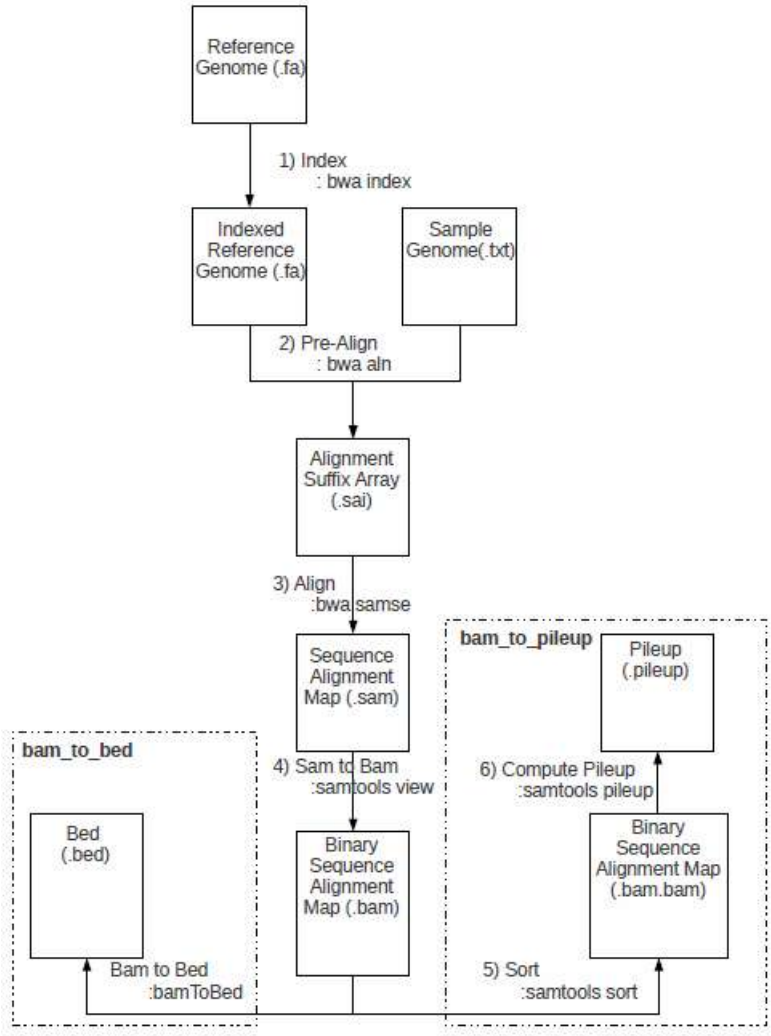


*Figure 1: Example single-end alignment workflow executed in the cloud.*

## Maximizing the Advantages of the Cloud

In the first phase, we demonstrated that genomic processing in the cloud is feasible and cost-effective, while providing a performance on par with in-house hardware. To truly realize the benefits of the cloud, however, what we need is an architecture that allows tens or hundreds of experiment jobs to be processed in parallel. This would allow researchers, for instance, to run algorithms with slightly different parameters to analyse the impact on their experiment results. At the same time, we would like a framework which incorporates all of the strengths of the cloud, in particular data durability, publishing mechanisms and audit trails to make experiment results reproducible.

To meet these goals, The Server Labs is developing a genomic processing platform which builds on top of RightScale's RightGrid batch processing system. Our platform facilitates the processing of a large number of jobs by leveraging Amazon's EC2, SQS, and S3 web services in a scalable and cost efficient manner to match demand. The framework also takes care of scheduling, load management, and data transport, so that the genomic workflow can be executed locally on experiment data available to the EC2 worker instance. By using S3 to store the data, we ensure that any input and result data is highly available and persisted between data centres, freeing our users from the need to backup their data. It also ensures that data can be more easily shared with the appropriate level of access control among institutions and

researchers. In addition, the monitoring and logging of job submissions provides a convenient mechanism for the production of audit trails for all processing tasks.

The following diagram illustrates the main components of The Server Labs Genomic Processing Cloud Framework:

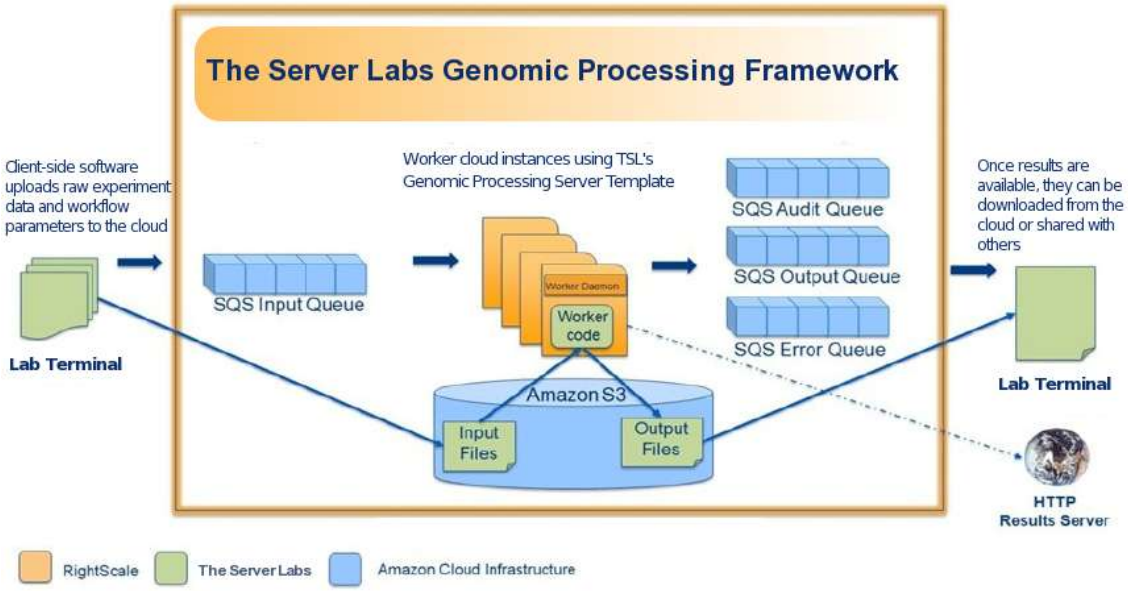


*Figure 2: The Server Labs Genomics Processing Cloud Framework*

The Worker Daemon is based on The Server Labs Genomic Processing Server Template, which provides the software stack for genomic processing. It automatically pulls experiment tasks from the SQS input queue along with input data from S3 to launch the genomic processing workflow with the appropriate configuration parameters. Any intermediate and final results are stored in S3 and completed jobs are stored in SQS for auditing.

## Cost Analysis

Given a RightGrid-based solution for genomic processing, we would like to analyse how much it would cost to run CNIO's workflows on the Amazon cloud. Let us assume for the sake of our analysis that CNIO runs 10 experiments in the average month, each of which generate an average of 10 GB of raw input data and produce an additional 20 GB of result data. For each of these experiments, CNIO wishes to run 4 different workflows, with an average running time of 2 hours on a large EC2 instance. In addition, we assume that the experiment results are downloaded once to the CNIO in-house data center. We also assume that the customer already has a RightScale account, the cost of which is not included in the analysis.

| **Amazon Service** | **Cost** |
|---|---|
| SQS | Negligible |
| S3 | Data transfer in: $0.10 per GB * 10 GB per workflow = $1.00 per workflow<br>Data transfer out: 1 download per workflow * 20 GB per download * $0.15 per GB = $3 per workflow<br>Storage: 30 GB per workflow * $0.14 per GB = $4.20 per workflow<br>Total cost: $8.20 per workflow |
| EC2 | $0.38 per hour * 2 hours per workflow = $0.76 per workflow |

| Amazon Service | Cost |
|---|---|
| All services | **Total cost: $8.20 + $0.76 = $8.96 per workflow** |

**Total Cost:**
10 experiments per month * 4 workflows per experiment * $8.96 per workflow =
**$358.40 per month or $4300.80 per year**

## Towards an On-demand Genomic Processing Service

By building on the RightGrid framework, The Server Labs is able to offer a robust cloud-based platform on which to perform on-demand genomic processing tasks, at the same time enabling experiment results to be more easily reproduced, stored and published. To make genomic processing even simpler on the cloud, the on-demand model can be taken even one step forward by providing a pay-as-you-go software as a service. In such a model, researchers are agnostic to the fact that the processing of their data is done in the cloud. Instead, they would interact with the platform via a web interface, where they would be able to upload their experiment's raw input data, select their workflow of choice, and choose whether or not to share their result data. They would then be notified asynchronously via email once the processing of their experiment data has been completed.

## References

Low Cost, Scalable Proteomics Data Analysis Using Amazon's Cloud Computing Services and Open Source Search Algorithms: http://pubs.acs.org/doi/abs/10.1021/pr800970z

How to map billions of short reads onto genomes:
http://www.cbcb.umd.edu/publications/files/ShortReadMappingPrimer-reprint.pdf

The RightGrid batch-processing framework: http://support.rightscale.com/12-Guides/RightGrid_User_Guide/01-RightGrid_Overview

### The Server Labs

The Server Labs (TSL) is a specialist IT Consultancy and Development Company and a leading authority in Cloud Computing services. Founded in 2004, The Server Labs focuses on the design and implementation of IT architectures and advanced software engineering projects, working with the most advanced solutions and technologies and offering its clients cost-effective, scalable and high performance solutions.

TSL's customer groups are predominantly large and medium-sized corporations, which share a growing need for cost effective and scalable IT solutions. TSL has offices in Spain, Germany and the UK. TSL is partnering with Amazon and RightScale to facilitate the adoption of Cloud Computing in Europe.

More information about The Server Labs is available on www.theserverlabs.com