Enabling Atlas2 Personal Genome Analysis on the Cloud

Uday S. Evani[1,4], Danny Challis[1], Jin Yu[1], Andrew R. Jackson[2,3], Sameer Paithankar[2,3], Matthew N. Bainbridge[1], Cristian Coarfa[2,3], Aleksandar Milosavljevic[2,3,4], Fuli Yu[1,3,4]

1. The Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA.
2. Bioinformatics Research Laboratory, Epigenome Center, Department of Molecular and Human Genetics.
3. Department of Molecular and Human Genetics, Baylor College of Medicine, TX 77030, USA.
4. Corresponding authors.

evani@bcm.edu, amilosav@bcm.edu, fyu@bcm.edu

***Abstract*-- Until recently, sequencing has primarily been carried out in large genome centers who also invested heavily in developing the computational infrastructure to enable post sequencing analysis. The recent advancements in sequencing technologies have lead to a wide dissemination of sequencing and we are now seeing many sequencing projects being undertaken in small laboratories. However, the limited accessibility to the computational infrastructure and high quality bioinformatic tools needed to enable analysis remains a serious road-block. The cloud computing and Software-as-a-Service (SaaS) technologies can help address this barrier. We deploy the Atlas2 Cloud Pipeline for personal genome analysis via the Genboree Workbench using software-as-a-service model. We report on a successful case study of personal genome analysis using this pipeline.**

1. BACKGROUND

The revolutionary development of massively parallel DNA sequencing has enabled identification of biomedically relevant genomic variants via whole genome [1] or exome resequencing [2]. Information relevant for personalized medicine such as assessment of longitudinal disease risks, and personalized treatment [3] are now within reach.

In a few very recent personal genomic studies, results have directly led to targeted treatment and dramatic improvement in the patients' quality of life [4]. These examples are paving the way to soon turn genomic sequencing into a routine diagnostic procedure and to enable personalized medicine.

Currently, analysis of sequencing data on a genomic scale requires bioinformatic expertise and access to extensive computational resources, presenting a significant barrier. Most cutting-edge genome analysis applications [5, 6] are still limited to a command line interface and require at least moderate informatics expertise to operate. In addition, large scale genomic data analysis requires routine access to a high performance compute cluster. Such requirements are entirely unsuitable for the operational models of smaller research/diagnostic laboratories due to the excessive investment requirements of computing infrastructure and personnel.

The deployment of genomic analysis software as a service within a cloud computing framework offers a unique solution for these problems. The concept behind cloud computing is to outsource computation to third-party servers or clusters at a remote location. This allows small laboratories to take advantage of external computational resources without having to maintain an in-house compute cluster. This software as a service model removes the upfront investment requirement and any delays associated with building local computing infrastructure. Earlier solutions involved providing access to analysis pipeline on a large cloud service provider like Amazon EC2 [7]. Though this solution eliminates the need for access to large compute clusters, the user would still need to have considerable bioinformatic skills due to the steep learning curve involved in working with clusters.

Following the "software-as-a-service" model, we integrated our genomic variant analysis pipeline – Atlas2 Suite [8] – onto a local "Cloud" using the Genboree Workbench(www.genboree.org). We also performed a case study using this pipeline as a proof of concept to demonstrate the potential of personal genome analysis on the cloud. Our cloud analysis pipeline has a web browser-based drag and drop interface, allowing users to interact with the software through their browser at any location, and making it practical for the software to be used by non-bioinformaticians. Our cloud pipeline is actively maintained by our team, which also removes the need for users to update the software.

2. RESULTS

Deploying our personal genomic analysis pipeline "Atlas2" via the Genboree Workbench

The Atlas2 Suite is a variant detection software package optimized for variant discovery on all the three next generation sequencing platforms. The suite consists of Atlas-SNP2 for calling Single Nucleotide Polymorphisms (SNPs) and Atlas-Indel2 for calling short insertions and deletions (INDELs) (http://www.hgsc.bcm.tmc.edu/cascade-tech-software-ti.hgsc). These tools have been available for command line usage, and applied to a number of large scale projects including the International 1000 Genomes Project [9], The Cancer Genome Atlas Project (TCGA), and follow-up resequencing in the context of disease genome wide association studies.

Genboree Workbench is a platform for deploying genomic tools as a service and is deployed at Baylor College of Medicine (www.genboree.org) and at Rackspace, a commercial cloud computing provider (www.genboree.com). With extensive utilities supporting genomic analysis, Genboree allows rapid integration and deployment of tools such as Atlas2 and provides a useful model for rapid cloud deployment and testing of the Atlas2 suite in the software-as-a-service mode.

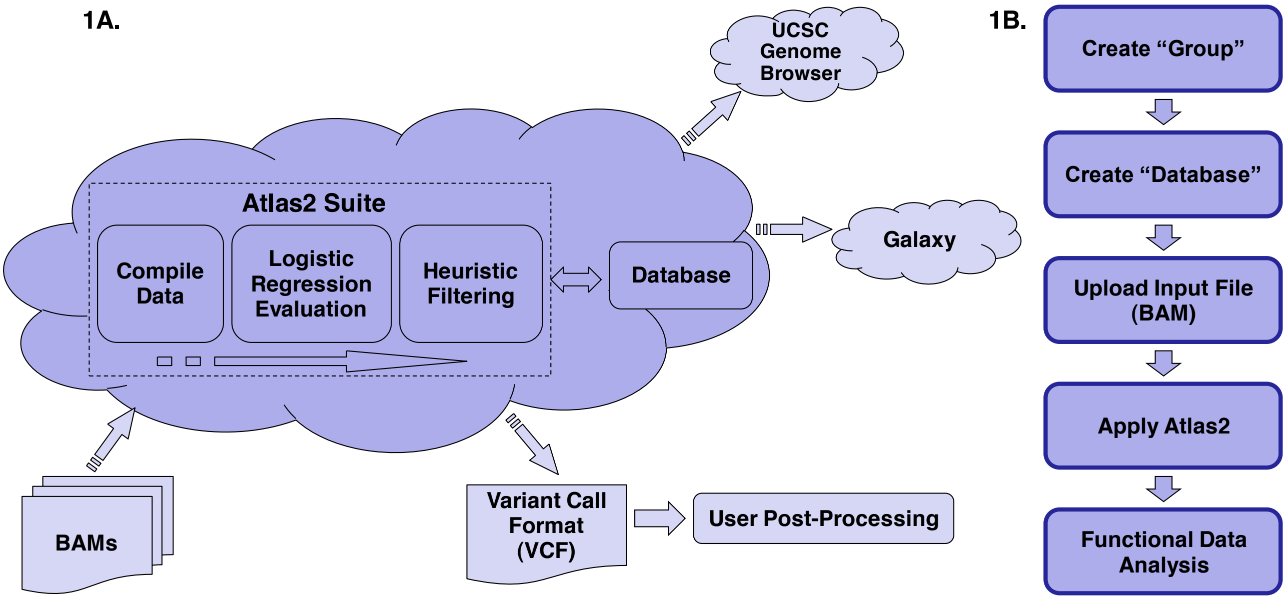


Figure 1A: High level representation of the Atlas2 cloud pipeline. Figure 1B: Specific steps involved in running Atlas2 pipeline on Genboree.

The Genboree Workbench Graphical User Interfaces (GUI) extensively relies on Ext-JS, a JavaScript library. Tools within the workbench make API (Application Programming Interface) calls to the REST (Representational State Transfer) API which is hosted on a thin server. This is done asynchromously using Asynchronous Javascript and XML (AJAX). Since Genboree System uses REST style of architecture to communicate between the server and the client, it allowed us to easily integrate Atlas2 within a couple of weeks. Genboree is backed by a small cluster of nodes which are managed by TORQUE resource manager (an open source tool) and Maui (developed by Adaptive Computing) to schedule jobs.

Applying Atlas2 cloud pipeline

The Atlas2 cloud pipeline can be accessed as a Genboree Workbench Toolset. Users from external groups with access to a web browser can 1) upload data onto the cloud, 2) run Atlas2 for variant analysis, and 3) visualize the variant calling results using different genome browsers such as Genboree Browser or University of California, Santa Cruz Genome Browser (Figure 1A). The Atlas2 cloud pipeline has a web-interface with hierarchical click-through steps. The self-explanatory nature of the web-interface eases the usage overhead. The workflow illustrated in Figure 1B shows the specific steps in running the Atlas2 Suite on the Genboree System.

Genboree Data Selector

The Genboree Workbench organizes data in a hierarchal tree. Before using the Atlas2 Suite users must define a group and create a database. Within the database are the "Files" and "Tracks" subdirectories. Files contain input files uploaded by the user and output files generated by Atlas2. Tracks contain processed output files which can be used for visualization on the Genboree browser. This hierarchical representation is shown in a screenshot of the Genboree workbench in Figure 2A.

Uploading data onto Atlas2 cloud pipeline

The Atlas2 cloud pipeline accepts Binary sequence Alignment/Mapping format (BAM) files as input. Files are uploaded onto the Atlas2 cloud pipeline by dragging the destination database from the Data Selector to the Output Targets box and selecting "transfer files" under the data tab in the menu. A prompt window allows users to select an input BAM file from their local computer and upload it to the cloud servers. A 24 GB BAM file took approximately one hour to upload on a 50 Mb/sec bandwidth connection.

Variant calling

The Atlas2 Suite may be run by simply assigning the desired input and output and selecting the appropriate tool (Figure 2A). The Atlas2 cloud pipeline allows users to specify parameter cutoffs in the job parameter-setting window (Figure 2B). Here one can choose from the three different sequencing platforms and tune the parameters.

The tool produces two output files, an LFF file and a Variant Call Format (VCF) [10] file which are stored under the files section inside of the database specified in the output target box. The LFF format is adapted from the LDAS upload format used to store variants and annotations (http://www.genboree.org/java-bin/showHelp.jsp?topic=lffFileFormat). Both the files can be downloaded by selecting the specific file and clicking on the download file option.

Genboree system allows integration with third party tools

Cloud deployment may produce “silos” of integration where extension of analysis pipelines and addition of analysis steps beyond those offered as a service may be hard to accomplish. To overcome this problem, Genboree system provides application programming interfaces for programmatic access to all the data and tools. Also data is accessible in formats that can be readily fed into a variety of ancillary tools. The interfaces and data format compatibilities enable mixing-and-matching of tools required in specific steps such as visualization in various genome browsers including UCSC genome browser[11], invocation of pipelines such as Galaxy[12], and integration with custom or third-party variant analysis and annotation tools such as ANNOVAR[13]. As described next, we successfully tested all three types of integration.

Visualizing variants with genome browsers

1. Genboree browser

The variant calls can be readily viewed in the Genboree genome browser. After going into the browser, variants can be visualized by selecting the appropriate database. Genboree browser supports looking at variants from multiple samples simultaneously.

2. UCSC genome browser

The variants called by Atlas2 cloud pipeline on Genboree can be directly exported to UCSC genome browser [11] for further viewing, annotation and analysis. The variants can be exported by converting our variants file into a BigBed format file (http://genome.ucsc.edu/goldenPath/help/bigBed.html) via the cloud file conversion functionality.

Integration with Galaxy

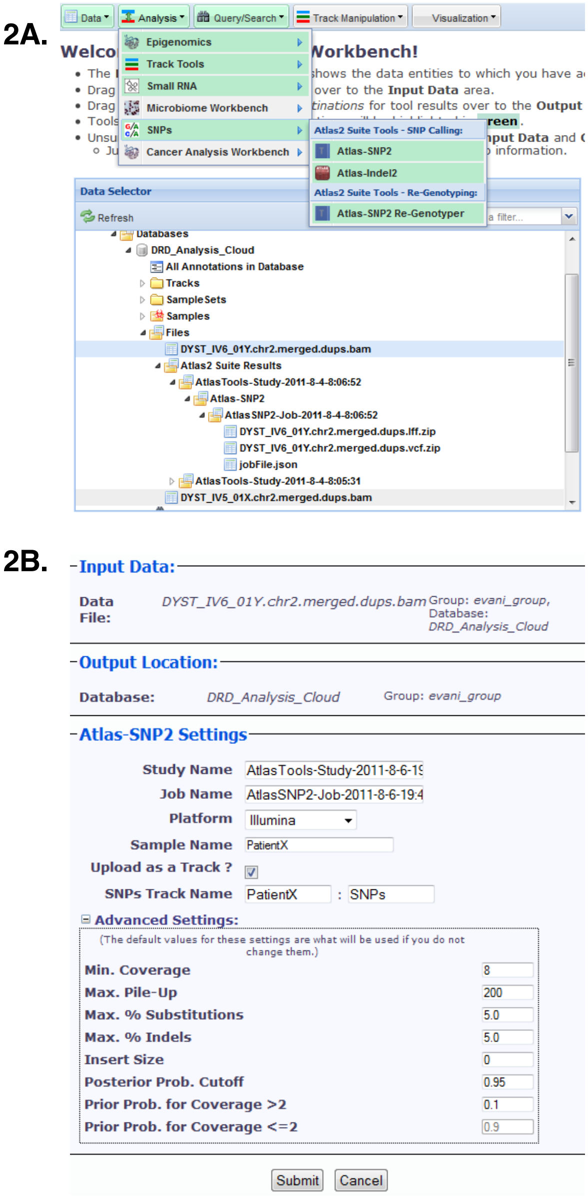
As our initial trial, we were able to upload our raw VCF file downloaded from Genboree without post-processing onto Galaxy and convert the VCF file into a multiple alignment format (MAF) custom track using the VCF to MAF custom track function with Graph/Display data.

Post-processing with third party variant annotation tools

The VCF file downloaded from Genboree was annotated and filtered using ANNOVAR. ANNOVAR categorizes variants into intronic, exonic, splicing, non-coding RNA, 5` untranslated region, 3` untranslated region, upstream, downstream and intergenic. The exonic variants are further categorized into synonymous, nonsynonymous, stop gain (gain of stop function), stop lost (loss of stop function), and frameshift or non-frameshift changes caused by insertions, deletions or block substitutions. ANNOVAR can also be used to filter out variants found in dbSNP.

Application of the Altas2 Cloud pipeline to a case of personal genome study

We next tested the Atlas2 cloud pipeline by performing an analysis on a recently published personal whole genome sequencing data set [4]. We examined the resource usage metrics and reproducibility in variant analysis, and examined the challenges related to integrating multiple tools required for variant detection, visualization, and analysis.


2A.
Data
Analysis
Query/Search
Track Manipulation
Visualization
Welco
Workbench!
The
Drag
Drag
Tools
Unsu
Ju
shows the data entities to which you have a
over to the Input Data area.
tinations for tool results over to the Output
reen.
nput Data and
information.
Epigenomics
Track Tools
Small RNA
Microbiome Workbench
SNPs
Cancer Analysis Workbench
Atlas2 Suite Tools - SNP Calling:
Atlas-SNP2
Atlas-Indel2
Atlas2 Suite Tools - Re-Genotyping:
Atlas-SNP2 Re-Genotyper
Data Selector
Refresh
a filter...
Databases
DRD_Analysis_Cloud
All Annotations in Database
Tracks
SampleSets
Samples
Files
DYST_IV6_01Y.chr2.merged.dups.bam
Atlas2 Suite Results
AtlasTools-Study-2011-8-4-8:06:52
Atlas-SNP2
AtlasSNP2-Job-2011-8-4-8:06:52
DYST_IV6_01Y.chr2.merged.dups.lff.zip
DYST_IV6_01Y.chr2.merged.dups.vcf.zip
jobFile.json
AtlasTools-Study-2011-8-4-8:05:31
DYST_IV5_01X.chr2.merged.dups.bam
2B.
Submit
Cancel

Input Data:

Data File: *DYST_IV6_01Y.chr2.merged.dups.bam* Group: *evani_group*, Database: *DRD_Analysis_Cloud*

Output Location:

Database: *DRD_Analysis_Cloud* Group: *evani_group*

Atlas-SNP2 Settings

Study Name: AtlasTools-Study-2011-8-6-1
Job Name: AtlasSNP2-Job-2011-8-6-19:4
Platform: Illumina
Sample Name: PatientX
Upload as a Track ? ☑
SNPs Track Name: PatientX : SNPs

⊟ Advanced Settings:

(The default values for these settings are what will be used if you do not change them.)

Min. Coverage	8
Max. Pile-Up	200
Max. % Substitutions	5.0
Max. % Indels	5.0
Insert Size	0
Posterior Prob. Cutoff	0.95
Prior Prob. for Coverage >2	0.1
Prior Prob. for Coverage <=2	0.9

Figure 2A: Shows how various folders are represented within Genboree, and how to navigate the menu bar to get to Atlas2 Suite.
Figure 2B:Atlas2 Suite customization window.

Description of the personal genome data set

Bainbridge et al.[4] employed the SOLiD 4 next-generation sequencing platform, and sequenced the complete genomes of a 14-year-old fraternal twin pair, one female (patient_x) and one male (patient_y) diagnosed with dopa (3,4-dihydrophenylalanine)-responsive dystonia (DRD). DRD is a genetically heterogeneous and clinically complex movement disorder with parkinsonian features that is usually treated with L-dopa. After identifying six heterozygous autosomal mutations in three genes, a new clinical intervention was prescribed that has dramatically improved the quality of life of both twins.

Table 1: Summarizes the amount of computation and time required to get the data (Chr 2 and Chr 19) on the cloud and to run through the variant calling steps.

	Resource usage (Chr2 / Chr 19)	
	Patient_X	**Patient_Y**
Size of BAM file (GB)	22 /5	24 /6
Time to upload (Min)	70 /12	85 /13
Atlas2 Runtime (Min)	390 /27	420 / 33
Atlas2 Memory Usage (MB)	1196 /275	1192 /270

Variant analysis using Atlas2 Cloud capability

We analyzed Chromosomes 2 and 19 since all six mutations were on these two chromosomes. Uploading the BAM files, ~ 22 GB in size, took ~70 mins using the Genboree workbench interface. We ran Atlas2 using the SNP default settings for the SOLiD platform. It took an average of ~6 hours to run Chromosome 2 and ~ 30 minutes to run Chromosome 19 (Table 1). The average memory usage on the cloud node was ~800 MB. Detailed numbers of time taken to upload, run Atlas-SNP2 and memory usage by sample are summarized in Table 1. The VCF file generated this way on the cloud was then downloaded for further analysis.

Table 2: Summarizes the total number of raw variants found in chromosome 2 and 19 of the two patients. Raw variants were then filtered with dbSNP (ver 129) and annotated with genetic information.

	Variant calls	
Nucleotide Variants	**Patient_X**	**Patient_Y**
All Variants	229484	235450
%dbsnp	87.51	87.66
Coding	1867	2062
Nonsynonymous	921	983
Coding (novel)	170	198
Nonsynonymous (novel)	124	129
Candidate genes	5	6

Combining results from the Chromosome 2 and 19, Atlas2 called 229,484 and 235,450 high confidence single nucleotide variants (SNV) in patient_x and patient_y, respectively. Annotating the VCF file with dbSNP (version129) using ANNOVAR we found 87.9% and 88.6% of SNV called in patient_x and patient_y respectively overlapped with dbSNP, which is very similar to what has been found by Bainbridge et al. (88.1% and 88.7%) [4]. The annotation was completed by running ANNOVAR, with this step serving as a filtering step to further narrow down the variants such that the we can get to novel nonsynonymous SNVs are more likely to be causal (Table 2).

Our Atlas2 pipeline successfully called all the six variants relevant in three genes in patient_y whereas only five of six variants were called in patient_x. Information regarding the three genes, variants and whether it was called is summarized in Table3. The one undetected mutation by our pipeline was in SPR gene at position 72969094 (A>G) causing a change from Arginine to Glycine. The reason Atlas-SNP2 was not able to call this SNV was due to a default heuristic filter which requires at least two high quality reads with variants. After examining the raw BAM file, we found that only one such variant read was found at this locus. In cases such as this, in order to lower the detection threshold, users can go back to the setting window and lower our heuristic cutoffs to achieve much higher sensitivity.

3. DISCUSSION AND FUTURE WORK

If personal genomic studies are to become a routine part of personalized diagnostics and medical management that is accessible to small research and clinical laboratories, advanced bioinformatic analysis must be made accessible both in terms of computational resources and usability. We have demonstrated the suitability of deploying existing analysis tools onto a cloud resource to address these issues, and demonstrated its utility by duplicating a real-world case study. Integrating this and many other analysis pipelines with Genboree and other cloud deployment platforms and resources such as UCSC genome browser and Galaxy to create custom pipelines is a plausible approach to removing the resource and usage barrier to personal genome analysis.

Table 3: Following three genes were found to contain two or more predicted amino acid altering hetrozygous mutation in both the patients. The table shows how many of those mutations did we reproduce.

				Reproducibility	
Chr	**Position**	**Reference/Variant**	**Gene**	**Patient_X**	**Patient_Y**
2	72972139	A/T	SPR	Found	Found
2	72969094	A/G	SPR	Not Found	Found
19	63464322	A/G	ZNF544	Found	Found
19	63464133	C/A	ZNF544	Found	Found
2	27657528	C/T	C2orf16	Found	Found
2	27655701	G/C	C2orf16	Found	Found

ACKNOWLEDGEMENTS

USE, DC, JY, MNB, and FY were supported by the National Human Genome Research Institute, National Institutes of Health, under grants 5U54HG003273 and 1U01HG005211-0109. AM, SP, ARJ, and CC were supported by NIH grants R01HG004009 and U01DA025956.

REFERENCES

[1] T. Tucker, *et al.*, "Massively parallel sequencing: the next big thing in genetic medicine," *Am J Hum Genet*, vol. 85, pp. 142-54, Aug 2009.

[2] S. B. Ng, *et al.*, "Targeted capture and massively parallel sequencing of 12 human exomes," *Nature*, vol. 461, pp. 272-6, Sep 10 2009.

[3] E. A. Ashley, *et al.*, "Clinical assessment incorporating a personal genome," *Lancet*, vol. 375, pp. 1525-35, May 1 2010.

[4] M. N. Bainbridge, *et al.*, "Whole-genome sequencing for optimized patient management," *Sci Transl Med*, vol. 3, p. 87re3, Jun 15 2011.

[5] A. McKenna, *et al.*, "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Res*, vol. 20, pp. 1297-303, Sep 2010.

[6] H. Li, *et al.*, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, pp. 2078-9, Aug 15 2009.

[7] B. Langmead, *et al.*, "Searching for SNPs with cloud computing," *Genome Biol*, vol. 10, p. R134, 2009.

[8] Danny Challis, Jin Yu, Uday Evani, Andrew R. Jackson, Sameer Paithankar, Cristian and A. M. Coarfa, Richard A. Gibbs, Fuli Yu, "An integrative variant analysis suite for whole exome next-generation sequencing data," *In Press*, 2011.

[9] N. Siva, "1000 Genomes project," *Nat Biotechnol*, vol. 26, p. 256, Mar 2008.

[10] P. Danecek, *et al.*, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, pp. 2156-2158, Aug 1 2011.

[11] P. A. Fujita, *et al.*, "The UCSC Genome Browser database: update 2011," *Nucleic Acids Res*, vol. 39, pp. D876-82, Jan 2011.

[12] D. Blankenberg, *et al.*, "Galaxy: a web-based genome analysis tool for experimentalists," *Curr Protoc Mol Biol*, vol. Chapter 19, pp. Unit 19 10 1-21, Jan 2010.

[13] K. Wang, *et al.*, "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Res*, vol. 38, p. e164, Sep 2010.