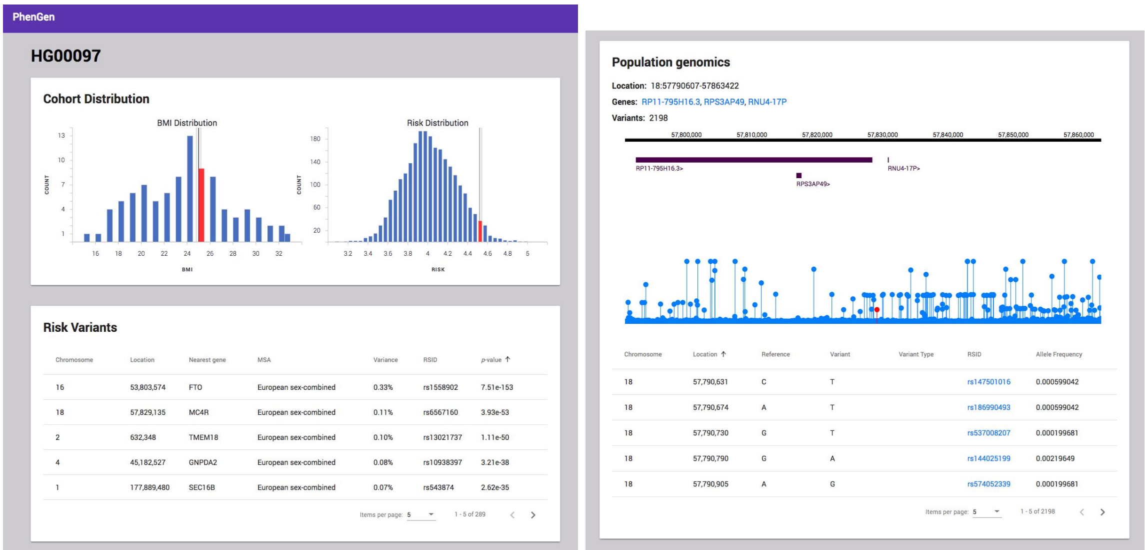
Going serverless for genomic and medical Big Data

Serverless architecture or Function-as-a-Service (Faas) has been on the rise and it is easy to see why when browsing through the Amazon Web Services (AWS) Serverless Application Repository which holds examples ranging from simple Alexa skills to complicated web services.

“AWS services have become so versatile and interoperable that it is now possible to set up complex research workflows as serverless web services,” says Dr Denis Bauer from the Commonwealth Scientific and Industrial Research Organization (CSIRO). Dr Bauer’s team’s bioinformatics research tool, GT-Scan2, was featured in this Blog two years ago as one of the first examples of serverless architecture catering for compute-intensive tasks.

Dr Bauer’s team, in collaboration with John Pearson’s Genome Informatics team from QIMR Berghofer Medical Research Institute, is now taking on one of the biggest challenges in bioinformatics -- making sense of the human genome -- and they are planning to do this using serverless infrastructure, demonstrating that serverless architecture can handle data-intensive tasks as well as compute-intensive tasks.



HG00097

Cohort Distribution

Risk Variants

Chromosome	Location	Nearest gene	MSA	Variance	RSID	*p*-value ↑
16	53,803,574	FTO	European sex-combined	0.33%	rs1558902	7.51e-153
18	57,829,135	MC4R	European sex-combined	0.11%	rs6567160	3.93e-53
2	632,348	TMEM18	European sex-combined	0.10%	rs13021737	1.11e-50
4	45,182,527	GNPDA2	European sex-combined	0.08%	rs10938397	3.21e-38
1	177,889,480	SEC16B	European sex-combined	0.07%	rs543874	2.62e-35

Items per page: 5 1 - 5 of 289

Population genomics

Location: 18:57790607-57863422

Genes: RP11-795H16.3, RPS3AP49, RNU4-17P

Variants: 2198

Chromosome	Location ↑	Reference	Variant	Variant Type	RSID	Allele Frequency
18	57,790,631	C	T		rs147501016	0.000599042
18	57,790,674	A	T		rs186990493	0.000599042
18	57,790,730	G	T		rs537008207	0.000199681
18	57,790,790	G	A		rs144025199	0.00219649
18	57,790,905	A	G		rs574052339	0.000199681

Items per page: 5 1 - 5 of 2198

Fig 1 Screenshot of the PhenGen tool showing a patient’s genetic risk for obesity and comparing his/her BMI against that of a peer-group with similar risk.

What is in your genes?

Genomic information is increasingly being used in clinical practice, however in many cases we still have an incomplete understanding of which genomic changes (variants) cause disease. Better integration of genomics with clinical data from patient records will increase our power to link genomic variants to disease, however the datasets are increasingly large and the compute power needed for the queries are immense. This challenge requires new ways to store, query and link data -- which is where serverless architectures come in.

“Working out which variants cause disease is not a needle-in-a-haystack problem, it’s a needle-in-a-bucket-of-needles problem,” says John Pearson, Manager of the Genome Informatics Group at QIMR Berghofer Medical Research Institute and co-founder of genomiQa.

Finding disease genes with Serverless functions

The teams at CSIRO and QIMR Berghofer have now joined forces to develop a prototype serverless genotype-phenotype (gen-phen) database implemented on AWS. This tool allows researchers and clinicians to combine information from multiple data silos to virtually create vast clinical datasets. They can then filter the data in real-time to identify medically similar patient cohorts and compare their genomes in the search for causative changes.

For the medical information the team uses the emerging Fast Healthcare Interoperability Resources (FHIR) standard along with clinical terminology to describe the clinical phenotype, or Phenome.

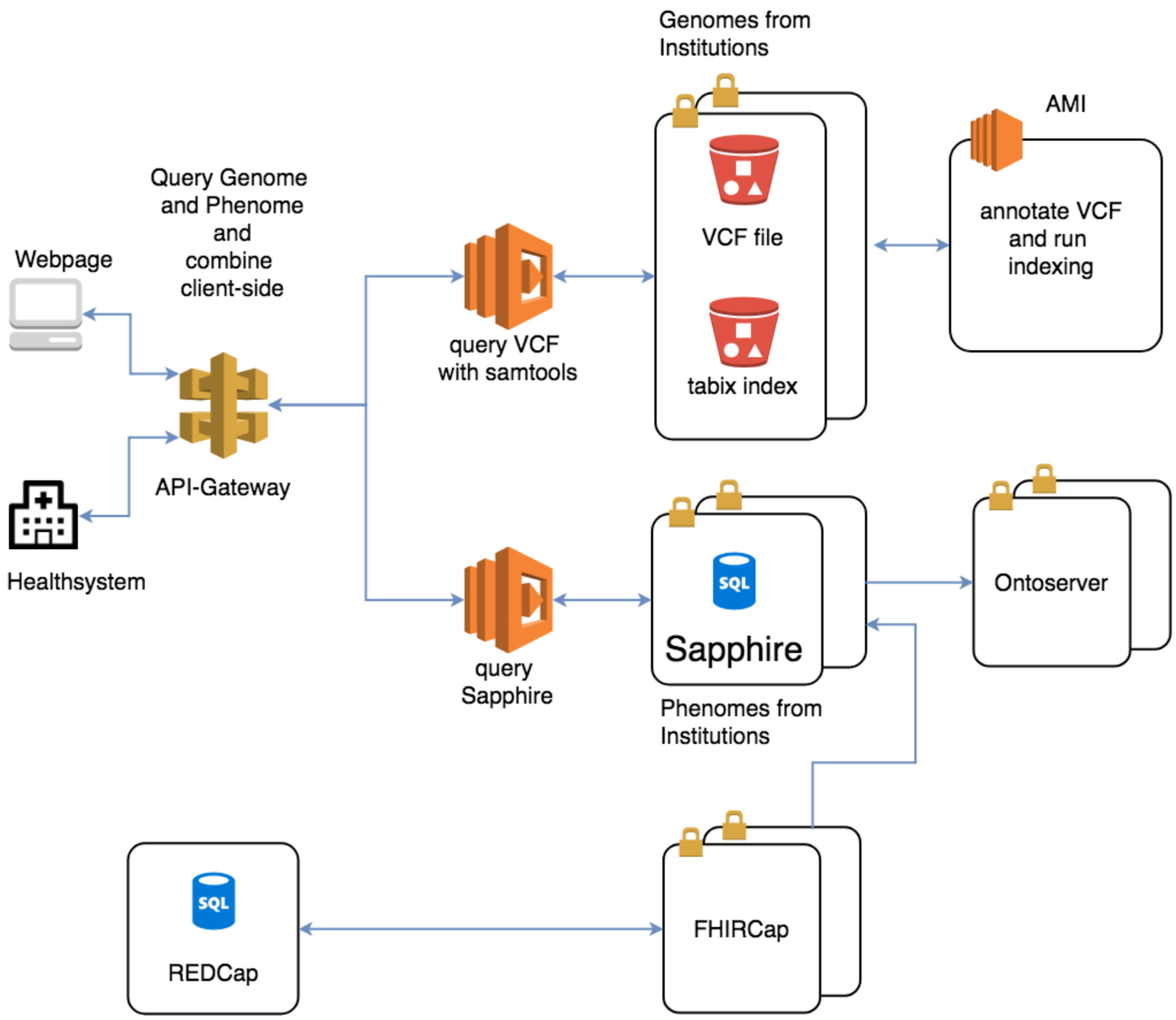
“FHIR is essentially a standard for exchanging healthcare information that allows querying clinical data in a standard way,” says Dr Alejandro Metke Jimenez (CSIRO), who set up the Phenome component of the workflow.

In the implementation an AWS Lambda function receives a query from the user, e.g. “all patients diagnosed with Melanoma under the age of 30” and converts it into standard FHIR queries, which are then sent to the CSIRO FHIR server that contains the clinical data.

The bigger challenge was the genomic information which is frequently multiple gigabytes per subject and quickly scales to terabytes when thousands of subjects are involved. “Traditional database schemas don’t scale to cater for such large cohort sizes and are certainly not economical enough to handle bursting workloads sustainably,” says Aidan O’Brien (CSIRO), one of the key developers of this new serverless Gen-Phen system.

The team uses AWS Lambda functions to orchestrate fast access to genomic variant data stored in S3 buckets. The Lambda functions run a python library to access samtools, a bioinformatic tool specifically designed for fast access to genomic data (VCF files). For this architecture, the team pre-processes the VCF through an EC2 instance to index (tabix) and block-gzip it allowing random access to byte ranges within the larger file and minimizing traffic from S3.

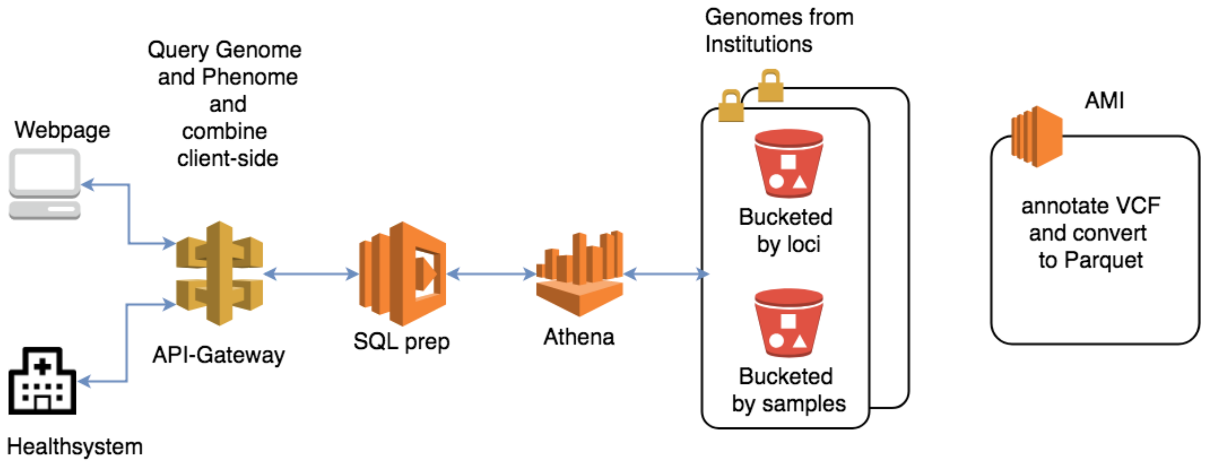
The medical and genomic information is then merged client-side using AngularJS, which allows customized filtering to cater for the bespoke use-case of each clinician.


Genomes from Institutions
AMI
VCF file
tabix index
annotate VCF and run indexing
Query Genome and Phenome and combine client-side
Webpage
query VCF with samtools
API-Gateway
Healthsystem
SQL
Sapphire
Ontoserver
query Sapphire
Phenomes from Institutions
SQL
REDCap
FHIRCap

Benchmarking access time

“To arrive at the current optimal architecture we investigated several alternatives specifically focused on speeding up the genomic file access,” adds Dr Conrad Leonard (QIMR Berghofer) who developed the AWS Lambda-based variant query architecture. “While our current architecture is the fastest we could come up with for now, we did identify new technologies that may allow for even faster solutions in the future.”

A particularly promising candidate was Athena, which is an AWS serverless query service allowing real-time scalability to cater for the drastically varying traffic of clinical workflows. Here, VCF files are converted to Parquet, which allows the data to be stored efficiently in AWS S3, and accessed seamlessly through Athena queries. The conversion step makes use of technologies like Apache Spark and Hive and so is able to complete the one-off processing of several terabytes of data in minutes.


Genomes from Institutions
Query Genome and Phenome and combine client-side
Webpage
AMI
Bucketed by loci
annotate VCF and convert to Parquet
API-Gateway
SQL prep
Athena
Bucketed by samples
Healthsystem

“So far, samtools has proved more performant when it comes to tailored genomics workflows like returning genomic regions and patient IDs,” explains Aidan O’Brien. While samtools takes around 600ms (pre-warmed Lambda) to query a 10,000 base region in the genome of 2000 patients, Athena takes more than 3.6s (with the same setup as described above).

“So samtools is probably the method of choice for specialized genomic retrieval applications,” says Conrad Leonard. “However, Athena’s ability to perform more complicated data transformations keeps it in the mix as a contender for future applications.”

The team were able to measured the runtimes of these two different setups using Epsagon, which allows workflow-based time measurements for serverless architecture.

How can a serverless Phen-Gen database help in clinical practice?

“Traditionally, Phen-Gen databases focus on either the medical or the genomics data and optimize accordingly, which limits the type of questions that can be asked,” explains Dr Oscar Luo (CSIRO). “By going serverless and embracing FHIR we are bringing the two worlds together.”

The team hence envisions the tool to be scalable to thousands of simultaneous users while remaining cost-effective because the solution does not pay for compute infrastructure to sit idle during periods of low or no use. “This really sets the CSIRO-QIMR Berghofer tool apart from the current database- and/or server-based systems,” says Angel Pizarro (AWS), who advises the NIH Cancer moonshot initiative. “Those tools are limited by the pre-defined database schema and demand ever increasing persistent resources as data volumes increase.”

“The system avoids data silos while preserving data ownership and patient privacy,” adds Brian Thorne of CSIRO's Confidential Computing Group. “It does this by maintaining a separation between patient's medical information and their genomic information, as well as isolating data across separate S3 buckets, which allows institutions to tightly control information release."

"We have to look beyond research-scale solutions if we are going to realise the promise of genomics-driven precision medicine," Pearson says.

CSIRO's Australian e-health Research Centre (AEHRC) and QIMR Berghofer are responsible for delivering the analysis and data management capability for the Queensland Genomics Health Alliance (QGHA).

"Innovations such as this are exactly the sort of technology that will support the implementation of genomics into clinical care to scale from the local to the national level," Dr David Hansen, AEHRC's CEO says.