OptraHealth
Engineering Digital Healthcare

WHITEPAPER

NEXTGEN GENOME INFORMATICS

iPhronesis™

NEXTGEN GENOME INFORMATICS

OVERVIEW

The aim of this document is to outline our experience in the development of highly scalable infrastructure for next generation genome sequencing analysis. A part of this solution was developed for a reputed biotechnology company associated with Genomics England. Other modules discussed here are developed for blue chip league clients.

The document covers our domain expertise of genomics and years of bioinformatics software development experience that were seamlessly integrated towards the development of powerful NGS modules executed on an equivalent powerful cloud infrastructure.

CHALLENGE

With the advent of Next Generation Sequencing(NGS) technologies and the falling cost of high throughput genome sequencing biotech organizations have adopted genomics as a fundamental part of their core research and development efforts. Genomics will be a useful technology only when we are able to truly discover the hidden wealth of knowledge in genomic datasets, and to accomplish this goal genomics is largely dependent upon powerful and intelligent software systems that allow for continually integrating newer analysis methods plus unique visualizations and instantly report findings. Our solution is designed to bridge the gap and ease the bottlenecks of genome analysis, visualizations and reporting by leveraging established and newly developed analytics methods on a compute intensive backbone with ease of generating customizable industrial grade reports.

Clients generated massive amounts experimental data in the form of FASTQ and BAM files which required analysis of all sorts. Analyzed data was then used for variety of downstream applications. In one case analyzed datasets are used for clinical reporting whereas in another scenario analyzed datasets are used in understanding drug efficacy using gene regulatory networks across population scale studies. Such wide variation of NGS data presented a clear challenge in front of us and based on this we developed modules for addressing the broad requirements summarized as follows:

Scalable Infrastructure

- Scalable Infrastructure
- Powerful Analytics
- Genome Intelligence
- Mutation Analysis

The next section discusses each of the above offerings in terms of its designs, deployment and support.

SOLUTION

The high level approach to providing scalable & extensible analysis was the development of a platform that hosts a powerful execution core tightly coupled to an analytics framework. We adopted a modular approach of development where all requirements of backend such as hosting to frontend analytics were addressed.

Genomic processing requires immense computational power and storage which becomes more difficult often with minimal IT expertise or support. What's more, complex integration and performance tuning of the infrastructure can take months. Through our integrated genomic processing infrastructure and platform, we can help you :

- Concentrate resources on R&D, not complex infrastructure
- Optimize genomic pipelines for quicker results
- Overcome obstacles to mainstream product viability
- Identify treatments in clinically relevant timeframes
- Enable cost-effective bioinformatics centers
- Maintain compliance and protect confidential data using secure, in-house resources

The platform with its infrastructure can deliver high-throughput and fast turnaround of genomic workflows. It for Next Generation Sequencing (NGS) writes sequenced data directly into the system's computational scratch space for processing. Output data and user files are then can be network-accessed by researchers for further investigation.

Some characteristics of the platform with its integrated genomic processing infrastructure are :

- Capability of processing up to 25 genomes per day using less than 7.5kWh/genome
- Can help in reducing lengthy implementation timelines from months to weeks.
- Ability to process 100GB of BAM data in 4 mins in a 32 core machine with 64GB RAM for CNV analysis.
- Easy execution of whole genome pipelines with characteristics quality control, alignment, variant calling, and annotation for samples of paired-end FASTQ files (~80 GBases)
- The infrastructure also hosts our in-house developed bioinformatics tools and databases.
- Pre-installed software required for analysis – BWA, Bowtie, MAQ, SAMtools, Annovar, Picard, GATK, Galaxy, Hadoop, Yarn, slurm, mesos, Hive, Spark.
- Complemented with workflow tools, big data analysis packages such as Spark, web applications/services, scripting APIs to assist in overall pipeline executions and research analytics.
- Hadoop-BAM Java library to act as an integration layer between analysis applications and BAM files stored in the Hadoop Distributed File System (HDFS). The library exposes a Picard compatible Java API to programmers. Hence, Hadoop code can be written without considering the issues of BGZF compression, block boundary detection, BAM record boundary detection, or parsing of raw binary data.

Cloud enabled computing solution that permits genome analysis pipelines to execute efficiently and cost-effectively. This solution addresses a range of problems that otherwise hinder use of cloud computing for NGS analysis. In particular, it supports the configuration of compute resources with application software that matches the versions required by analysis pipelines; access from cloud resources to pipeline code and input and reference data; and in order to minimize cost and provide efficient execution, on demand provisioning of compute resources when required, and the release of those resources when they are not in use.

SCALABLE INFRASTRUCTURE

CLOUD INFRASTRUCTURE

We power analytics using the latest software technologies to deliver high performance. The platform is designed for BigData scale analysis and provides an application layering that allows for simple user interactions.

The solution being hosted on commercial Amazon cloud (deployable easily to other cloud like Azure/Google) resources with Elastic Scaling M2.4xlarge enables reliable and highly scalable execution of NGS analysis workflows. Integrated data management capabilities address the challenges associated with managing the movement of big data from acquisition through analysis and storage.

The analysis core is powered by a scalable and extensible backbone that comprises of BigData environment using Hadoop, and Yarn. All modules are data agnostic and work with both structured and unstructured data such as accessioned from experimental databases and primary literature. The execution core implements Python code and wrappers that recursively call other modules. The core execution model is multi-threaded and multi-processed by design and in a recent benchmark we processed 100GB of BAM data in 4 mins in a 32 core machine with 64GB RAM for CNV analysis. All modules are available as REST services and are invoked from client applications or seamlessly integrated into your existing workflows. Admins can easily scale application servers automatically using our cloud scaler and job management modules.

One typical deployment primarily contains :

STANDARD WORKER NODES	FAT MEMORY WORKER NODES	Cluster
Each Node – 2 CPU with 12 Core	Each Node – 2 CPU with 32 Core	Number of compute Node - 30
Xeon® Processor	Xeon® Processor	Number of Core – 560
256 GB of memory	1 TB of memory	Memory – 16 TB
1x800GB SSD disk for OS	1x800GB SSD disk for OS.	10G Ethernet

Benchmarking

To evaluate the execution of the pipeline we executed 20 concurrent pipelines using 10 different input sequence files (~30GB each). A reference human genome "hg19.fa" is locally cached, and following time (hrs) benchmarking has been achieved :

Instance	Hours
Single M1.large	150.5
Single M2.4x large	112.3
Elastic Scaling M2.4xlarge	4.5 (24 core)

DATA MANAGEMENT

The data to be analyzed was provided in different forms such as from different vendors, different reference genome used during alignment and from end applications such as whole genome vs. whole exome vs. targeted sequencing. To address this requirement, the first item addressed was cloud storage. This allowed to extend the storage capacity at a massive scale. For efficient data management we developed a management module that allows users and system admins to add/ edit and remove files as required but also to generate metadata in the form of annotations. Such annotations were stored in a file management database for later access and meta-analysis of genomic data.

As the data was from different providers we developed agnostic pipelines that allows for specific conversion taking into account the nuances of each sequencing platform provider. Additionally, some files were provided in BAM format and this pipeline also converted from aligned BAM to basic FASTQ files. While this may seem accomplishable, it is ridden with some very level specifics that have to applied such as the correct selection of aligners and most important carefully selecting the reference genome. To execute this pipeline, we used the compute power of our cloud instances which accommodated for extreme memory requirements and fast processing.

The data can be located in any data center. The infrastructure provides an environment in which researchers do not need to worry about the physical location of the data. With respect to users rights, queries will be sent to each remote server. The host will process the request and return the results back to the main server where all the privacy limitations are controlled for the data. Once the results are ready, the end user can see the desired information. Depending on the type of query, results will be divided into two parts, the first part is related to the samples to which the user has authorized access, and for which the users can see all details. The second part contains results for the whole population, for which the user has only access to some aggregate statistics without details.

POWERFUL ANALYTICS

CORE PROCESSING

Modules developed within this scope are mainly to work on massive datasets such as for entire cohorts. We developed several scalable and extensible modules mainly in Python as this language is specifically designed for BigData like work. One of the requirements was the accurate implementation of GATK Best Practices workflow, this was developed in using Python that calls GATK and other associated programs such as Picard. Scalability in genome analysis is of paramount importance and to reduce the compute time we adopted a multi-core, multi-process approach. This was accomplished by using a whole genome region BED file, using this we split BAM file into smaller BAMs that are region specific. Each BAM file is then forked on an individual process. At the end of the process all the VCF files are collated and using native commands that allows for even more computational speed. Individual processing of BAM files is a job in our ecosystem and to support this functionality we developed a cloud job scheduler and task manager allowing admins and users to receive real-time updates about their tasks.

SCALABLE VARIANT STORE

The final outcome of NGS analysis is the widely adopted VCF file format. Our solutions work with both gVCF and VCF files. Variant detection algorithms may be changed depending up on the use case for e.g. for population studies where joint haplotype calling is required we used GATK which generates gVCF whereas in another use case where the aim was to generate VCF to detect Mendelian disorders we used GATK and SamTools. Our pipelines allow for integrating any such algorithms. The gVCF/ VCF are also annotated using standard tools and we also support custom annotators. The end result of this process is a very large VCF file. Since VCF file is just the beginning of variant analysis downstream we adopted a scalable variant store approach. Instead of storing each VCF as a file we loaded the VCF into Hive, which is preferred databases from the BigData ecosystem. This allows for loading many, many VCF files that contained millions of variants across many, many samples. The added advantage of this approach is it also allowed for executing SQL queries over using native commands to find variants.

MACHINE LEARNING

The biomedical industry has adopted machine learning type programming, wherein the programs "understand" the action through a self-learning iterative process. We have implemented machine learning in some of our modules as well. For e.g. GATK tool itself user machine learning when it performs the base quality recalibration. In addition to that we used machine learning for concordance/ discordance analysis across data sets specifically to identify missing values and understand trends and patterns in the data. Another application of machine learning is in one of our EMR connector module that performs regression and harmonization of medical data performing actions such as named entity recognition, control of vocabulary and co-relations with ontologies. Machine learning algorithms are generally available as Python libraries, these libraries are part of our platform, some of which we modified at source.

VISUALIZATIONS

The magnitude of genome information makes it an ideal candidate for using new visualization techniques. We have developed modules specifically designed for working with genomic datasets. This allows a user to easily switch view such as a user may prefer to view data as a Manhattan plot for one type of data but immediately view a concept plot for viewing relationship between several entities. We developed a module for understanding gene interactions and drug efficacy for an oncology application and concept plot was used there.

GENOME INTELLIGENCE

The process of variants analysis for requires a deep systematic review of causal variants before labelling them as pathological, in a clinical setting. Often this task is arduous and takes much valuable time. We developed a knowledge automation module to address this need. This module ingests disparate data such as scientific literature, medical notes, EMR records and experimental data and generates a corpus. This corpus is the underlying base of a machine learning powered query engine named BioNLP (Bio Natural Language Processing). BioNLP allows users to compose complex queries in natural language and combines data from multiple sources and provides assertions, insights and alerts the users to areas of interest during the course of clinical reporting. Thus, the knowledge automation module provides quick contextualization of variants and provides an accurate phenotype- genotype mapping backed with evidences such as statistic values, pathogenicity, reproducibility and helps in downstream analysis processes. Full customization of the underlying corpus and rules framework to match your specific workflow is easily accomplished.

MUTATION ANALYSIS

A requirement of our client was mutation class identifications for clinical applications, which is the emerging application of NGS technologies. To address this requirement, we developed a python based pipeline which included secondary and tertiary analysis and variant annotation. The end result of this pipeline are variants identified from a set of input genome data files. In the next step the module classifies the variant as CNA/ CNV, SNP or SV and provides genome annotation data. Identifying the variants is only one- half of the solution and the next step is to classify the variant, pathogenic, likely pathogenic, uncertain significance, likely benign and benign. In the final step variants identified are correlated with databases such as ClinVar, dbSNP, CiVIC etc. to provide evidences for reporting. Client used this module in their clinical testing pipeline.

BENEFITS

Our platform uses data agnostic technologies and we develop all modules with performance and end user application as a part of the design process at the outset. This enables us to deliver mission critical applications that easily scale on demand to meet time and resource needs of an organization. Following is a list of the most powerful features of this platform.

SCALABILITY
Scale from a single server to many servers automatically to balance load and reduce analysis time.

ON- DEMAND
Enable/ disable modules for your users on the fly.

COMPLEX VISUALIZATIONS
View complex datasets using the latest technological tools and with recommendations of widgets for data.

INTEGRATION
All modules are available as REST APIs allowing you to integrate with your solutions.

EXTENSIBILITY
Use any code that you have developed and call that with our agnostic modules.

DOMAIN WRAPPERS
We provide ready end-to-end modules for e.g. for tumor- normal analysis, drug interactions, population studies etc.

KNOWLEDGE AUTOMATION
Turn data into knowledge by leveraging our BioNLP powered cloud data repository.

For more information on Optra Health, **please visit iphronesis.com**