Decrypting your genome data privately in the cloud

Marc Sitges
Data Manager@Made of Genes

@madeofgenes
The Human Genome

3.200 M (x2) ~20,000 genes (~30%)
Base pairs (bp) (Exons ~1%)
Sequencing cost decrease
NEXT GENERATION SEQUENCING

Large DNA molecule

Fragmentation

Sequenced

Assembly of overlapping DNA sequencing

Assembled sequence

(1st) – 800bp / poor parallelization
(2nd) – 32 to 400bp / Extremely parallelized

[http://www.micronautomata.com/bioinformatics]
The Pipeline

SEQUENCING

MAPPING

VARIANT CALLING

Big data files generated (not BigData though)

High computing requirements

Results variability

Reference CCGTTAGAGTTACAATTGCA
Read 2 TTAGAGTAACAA
Read 3 TCGTTAGAGTAA
Read 4 TTAGAATTCGA
Read 5 TAGAAACAA
Read 6 TTAGAATACAAAT
Analysis steps

- quality control
- trimming + adaptor removal
- mapping paired
- mapping unpaired
- merge
- sort by reference
- remove duplicates
- indel realignment
- base recalibration

Natively parallel

- merge
- remove duplicates
- indel realignment

- snp calling
- snp annotation
- snp filtering
## Analysis steps

<table>
<thead>
<tr>
<th>Step</th>
<th>Type of computation</th>
<th>Memory Usage</th>
<th>I/O</th>
<th>Disk Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality control</td>
<td>online processing, by input file</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence filtering</td>
<td>online processing by paired files, inherently parallel by read</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mapping (lane)</td>
<td>exponential problem solved approximately, inherently parallel by read</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alignment reordering (lane)</td>
<td>sorting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Remove duplicates (lane)</td>
<td>processing by similar chunks, unbalanced trivial parallelization by chromosome</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Realignment (lane)</td>
<td>cost dependant on the target region, unbalanced trivial parallelization by chromosome</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recalibration (lane)</td>
<td>cost dependant on the target region, unbalanced trivial parallelization by chromosome</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNP calling + basic annotation</td>
<td>processing by regions, statistical analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNP hard filtering</td>
<td>online processing</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example: exome 30x, input fastq = 2 x 1.5 GB (ascii compressed), sam = 10 GB (ascii format), bam = 2 GB (binary format), reference=3GB (ascii format), vcf=20 MB (ascii format)
Current implementation

- Slurm DB
- Data Staging
- Profiling
- In-house JSON defined workflows
- Bash wrapper
- Storage
- Cluster File System
- Computing nodes
- easybuild
- Login node

shiftc
Aspera
Current implementation pitfalls

- Adhoc in-house workflow design: no portability
- Different steps have different computing requirements
- Difficult to install/upgrade or switch between software versions and setups
- Even if limited pipeline configurability, different libraries already required
- Not easy restart in case of failure

We need a flexible a scalable environment!!
1. User sequence data (genome, exome).

2. We protect your data indefinitely.

3. User can transfer his or her data for clinical trials.

4. User chooses to purchase an analysis + visit or a medical visit.

5. The health professional can ask for and receive a complementary diagnostic test.

6. The user receives personalized medical care while contributing to society.

- **Researcher**
- **Hybrid Cloud LOPD compliance Informed consent**
- **3rd-party solutions Analytics / Products / Reports**
- **Expert doctors**
- **Pharma & Industry**
- **Healthcare providers**
- **User (genome owner)**
First step: workflow definition

Main goal: standardize the way we describe analysis tools and workflows to enable portability and reproducibility of data-intensive jobs.

Implementations:
- Common Workflow Language (CWL)
- Workflow Description Language (WDL)
- Nextflow's model
A container ships an application with all the environment (including libraries and other dependencies) that the applications needs to run, but shares kernel and more general bins/libraries with the system.

Provide close to native performance (IBM Reserach Report, RC25482)
User-defined images

Software framework which allows to include portable and customized OS environments that accompany the application.

- Isolate HPC platform from dependencies
- Easier development cycle and software maintenance
- Increased flexibility and portability
- Better version control and compatibility
- Facilitate reproducible results, pipeline sharing and collaboration
- Image sharing in public repositories (DockerHub, Quay, Dockstore)
<table>
<thead>
<tr>
<th>Step</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quality control</strong></td>
<td>fastqc-0.11</td>
</tr>
<tr>
<td><strong>Trimming adapter removal</strong></td>
<td>trimmomatic-0.33, cutadapt</td>
</tr>
<tr>
<td><strong>Mapping</strong></td>
<td>bwa-mem-0.7.12, samtools-1.2, gem, bowtie, samtools-1.2</td>
</tr>
<tr>
<td><strong>Merge</strong></td>
<td>samtools-1.2 merge, samtools-1.2 sort</td>
</tr>
<tr>
<td><strong>Sort by reference</strong></td>
<td>samtools-1.2 sort</td>
</tr>
<tr>
<td><strong>Remove duplicates</strong></td>
<td>picard mark_duplicates</td>
</tr>
<tr>
<td><strong>Indel realignment</strong></td>
<td>gatk-1.6 RealignerTargetCreator, gatk-1.6 IndelRealigner, gatk-3.5 RealignerTargetCreator, gatk-3.5 IndelRealigner</td>
</tr>
<tr>
<td><strong>Base recalibration</strong></td>
<td>gatk-1.6 CountCovariates, gatk-1.6 TableRecalibration, gatk-3.5 BaseRecalibrator, gatk-3.5 PrintReads</td>
</tr>
<tr>
<td><strong>Snp calling</strong></td>
<td>samtools-1.2 mpileup, bcftools-1.2 call, gatk-1.6 UnifiedGenotyper, gatk-3.5 UnifiedGenotyper</td>
</tr>
<tr>
<td><strong>Snp annotation</strong></td>
<td>SnpSift-4.1</td>
</tr>
<tr>
<td><strong>Snp filtering</strong></td>
<td>bcftools-1.2 filter, gatk-1.6 SelectVariants, gatk-3.5 VariantRecalibrator</td>
</tr>
</tbody>
</table>

3rd-party solutions
- Analytics / Products / Reports

Dockerized pipeline

- quality control
- trimming adapter removal
- mapping
- merge
- sort by reference
- remove duplicates
- indel realignment
- base recalibration
- snp calling
- snp annotation
- snp filtering
**Dockerized pipeline**

- **quality control**
  - fastqc-0.11
  - fastqc-0.15

- **trimming adaptor removal**
  - trimmomatic-0.33
  - cutadapt

- **mapping**
  - bwa-mem-0.7.12
  - samtools-1.2
  - gem
  - bowtie

- **merge**
  - samtools-1.2 merge
  - samtools-1.2 sort

- **sort by reference**
  - samtools-1.2 sort

- **remove duplicates**
  - picard mark_duplicates

- **indel realignment**
  - gatk-1.6 RealignerTargetCreator
  - gatk-1.6 IndelRealigner

- **base recalibration**
  - gatk-1.6 CountCovariates
  - gatk-1.6 TableRecalibration

- **snp calling**
  - samtools-1.2 mpileup
  - bcftools-1.2 call
  - gatk-1.6 UnifiedGenotyper

- **snp annotation**
  - SnpSift-4.1

- **snp filtering**
  - bcftools-1.2 filter
  - gatk-1.6 SelectVariants

- **3rd-party solutions**
  - Analytics / Products / Reports
Dockerized pipeline

- **quality control**
  - fastqc-0.11
  - fastqc-0.15

- **trimming / adaptor removal**
  - trimmomatic-0.33
  - cutadapt

- **mapping**
  - bwa-mem-0.7.12
  - samtools-1.2
  - gem
  - bowtie
  - samtools-1.2

- **merge**
  - samtools-1.2 merge
  - samtools-1.2 sort

- **sort by reference**
  - samtools-1.2 sort

- **remove duplicates**
  - picard mark_duplicates

- **indel realignment**
  - gatk-1.6 RealignerTargetCreator
  - gatk-1.6 IndelRealigner
  - gatk-3.5 RealignerTargetCreator
  - gatk-3.5 IndelRealigner

- **base recalibration**
  - gatk-1.6 CountCovariates
  - gatk-1.6 TableRecalibration
  - gatk-3.5 BaseRecalibrator
  - gatk-3.5 PrintReads

- **snp calling**
  - samtools-1.2 mpileup
  - bcftools-1.2 call
  - gatk-1.6 UnifiedGenotyper
  - gatk-3.5 UnifiedGenotyper

- **snp annotation**
  - SnpSift-4.1

- **snp filtering**
  - bcftools-1.2 filter
  - gatk-1.6 SelectVariants
  - gatk-3.5 VariantRecalibrator

---

3rd-party solutions
Analytics / Products / Reports
quality control
trimming adaptor removal
mapping
merge
sort by reference
remove duplicates
indel realignment
base recalibration
snp calling
snp annotation
snp filtering

quality control: fastqc-0.11, fastqc-0.15
trimming adaptor removal: trimmomatic-0.33, cutadapat
mapping: bwa-mem-0.7.12, samtools-1.2
merge: samtools-1.2 merge
sort by reference: samtools-1.2 sort
remove duplicates: picard mark_duplicates
indel realignment: gatk-1.6 RealignerTargetCreator, gatk-1.6 IndelRealigner
base recalibration: gatk-1.6 CountCovariates, gatk-1.6 TableRecalibration
snp calling: samtools-1.2 mpileup, bcftools-1.2 call
snp annotation: SnpSift-4.1
snp filtering: bcftools-1.2 filter

Dockerized pipeline

3rd-party solutions
Analytics / Products / Reports
Dockerized pipeline

quality control
- fastqc-0.11
- fastqc-0.15

trimming
- trimmomatic-0.33
- cutadapt

adaptor removal
- bwa-mem-0.7.12
- samtools-1.2

mapping
- gem
- samtools-1.2
- bowtie

merge
- samtools-1.2 merge
- samtools-1.2 sort

sort by reference
- picard mark_duplicates

remove duplicates
- gatk-1.6 RealignerTargetCreator
- gatk-1.6 IndelRealigner

indel realignment
- gatk-1.6 CountCovariates
- gatk-1.6 TableRecalibration

base recalibration
- gatk-3.5 BaseRecalibrator
- gatk-3.5 PrintReads

.snp calling
- samtools-1.2 mpileup
- bcftools-1.2 call

.snp calling
- gatk-1.6 UnifiedGenotyper

.snp annotation
- SnpSift-4.1

.snp annotation
- gatk-1.6 SelectVariants

.snp filtering
- bcftools-1.2 filter

.snp filtering
- gatk-3.5 VariantRecalibrator

3rd-party solutions

Analytics / Products / Reports

RUN!
Managing containers in an HPC environment

Provide an abstraction layer between the pipeline process and the cluster resource manager.
Cloud computing

Resources limitation or underuse

On demand resources. Pay per use
Open vs Personal
Open vs Personal
Open vs Personal

Gender (optional)
Phenotype (optional)
Disease_site (optional)
Sample_Type (optional)
Open vs Personal

Gender (optional)
Phenotype (optional)
Disease_site (optional)
Sample_Type (optional)
Open vs Personal

Gender (optional)
Phenotype (optional)
Disease_site (optional)
Sample_Type (optional)
Open vs Personal

Gender (optional)
Phenotype (optional)
Disease_site (optional)
Sample_Type (optional)
Open vs Personal

• Open data is publicly accessible, but anonymized
  • Phenotypes and metadata is registered on submission
    • Omics datasets and Metadata are (usually) stored in different formats/servers (ie, BAMs on EGA but phenotype in Excel as Supplementary file in journal)
  • Lack of evolution tracking on patients

• Personal data allows identification of patients
  • Phenotypical data can be collected on-demand
  • Evolution of patients can be followed
Sharing vs Compliance
Sharing vs Compliance
Sharing vs Compliance

• All in all, the problem is the ownership of the data
  • NO, you cannot do whatever you want with the patient’s data
  • YES, the patient can do whatever they want with it

• According Patient’s Autonomy Law, every single clinical record generated in an hospital belongs to the patient.
  • The problem is that 3rd parties should not decide what to do with recollected data, because, at the end this data belongs to a patient.
Sharing vs Compliance

• Clinical personal data cannot be stored in the cloud
  • But the cloud was convenient and cheap!
  • You should have your own servers in a physically controlled location

• Data cannot be handed over without acceptance of informed consent
  • But we want other people to collaborate with us!
Hybrid cloud
Hybrid cloud

- De-identified copies of the raw and processed genomic data are stored in the public cloud instance.

- Clinical data and personal identifiers are stored in encrypted databases in the private cloud.
- Raw and processed genomic data is processed and stored in the private storage.
**SECURITY**

- Data is encrypted in transit and at rest using AES-256.
- Databases containing PHI and identifiers are also encrypted.
- Analysts access encrypted data using a host-prof method.
- Access control mechanism in private cloud.

---

### De-identified and Encrypted Genomic Data

- **Raw data**
  - QC
  - Trimming
  - Mapping

- **Aligned reads**
  - Recalibration
  - Remove duplicates
  - Variant calling

- **Processed variants**
  - Private Health Information
  - Customer identifiers

- **Processed data**
  - 10-100 GB

---

**Hybrid Cloud**

- **Data Warehouse**
  - 1-10^3 MB

---

**Private Health Information**
Hybrid Cloud

WebApp
API
Genomic data upload service
Database
DMZ

Genomic data storage
Computing nodes
SLURM workload management system
Databases with critical information

Genomic data storage
Computing nodes
On demand
Cloud bursting

VM images repository

Could Provider API

SGE, LSF, others…
Cloud bursting

VM images repository

Cloud Provider API

SGE, LSF, others…

slurm
workload manager

GRID ENGINE
Cloud bursting

VM images repository

Cloud Provider API

SGE, LSF, others…

slurm
workload manager

GRID ENGINE
Cloud bursting

VM images repository

Cloud Provider API

SGE, LSF, others…

slurm

GRID ENGINE
Cloud bursting

VM images repository

SGE, LSF, others…

Could Provider API

SOFTLAYER®
MADE OF GENES

All genomics. All yours.

www.madeofgenes.com
@ we.are@madeofgenes.com
Facebook /madeofgenes
Twitter @madeofgenes