JCAD 2018

Utilisation des ressources CIMENT dans le cadre du projet epimed:

Les besoins spécifiques de la bioinformatique pour l'analyse des données d'exome.

Quentin Testard, Julien Thevenon, Laure Raymond, Jean-François Taly

25/10/18

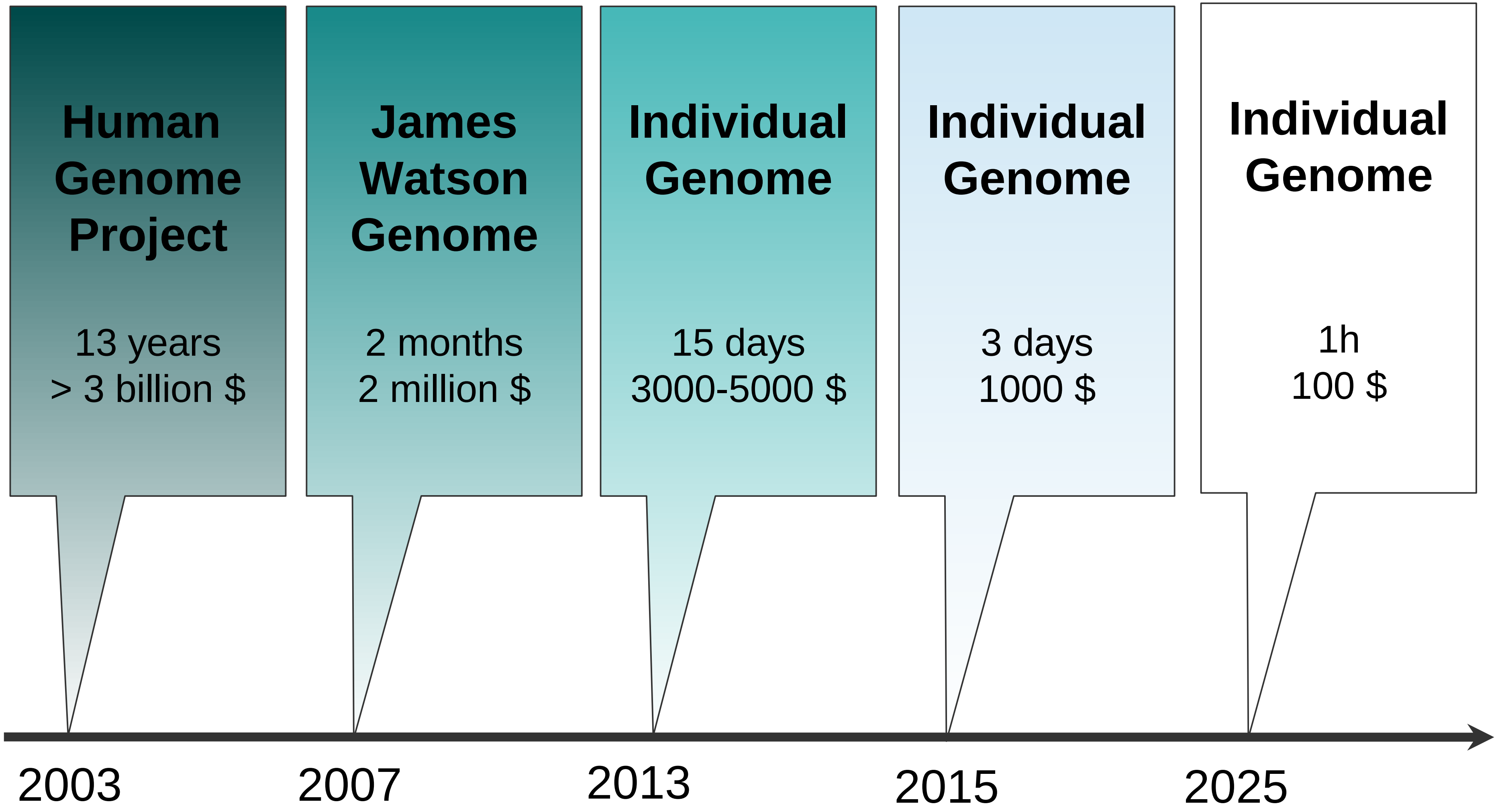
My thesis objectives

- **Industrialized** processes for exome analysis ⇒ **Workflow**
- **Speed up** / use **novels approaches** for exome analysis
- **Master** pipeline execution

⇒ **Containers**

- **Control** pipeline possible failures ⇒ **Pipeline certification**
- **Will be used in routine health diagnosis activity**

DNA sequencing



Human Genome Project

13 years
> 3 billion $

James Watson Genome

2 months
2 million $

Individual Genome

15 days
3000-5000 $

Individual Genome

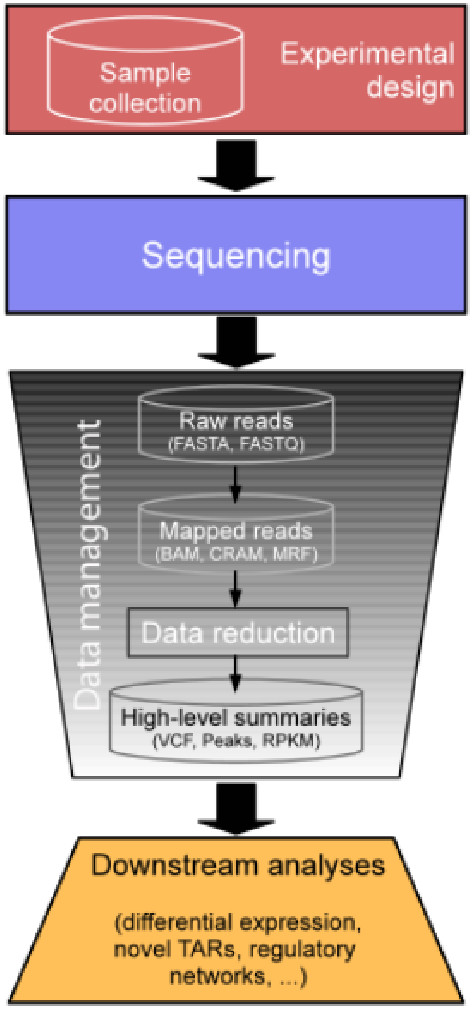
3 days
1000 $

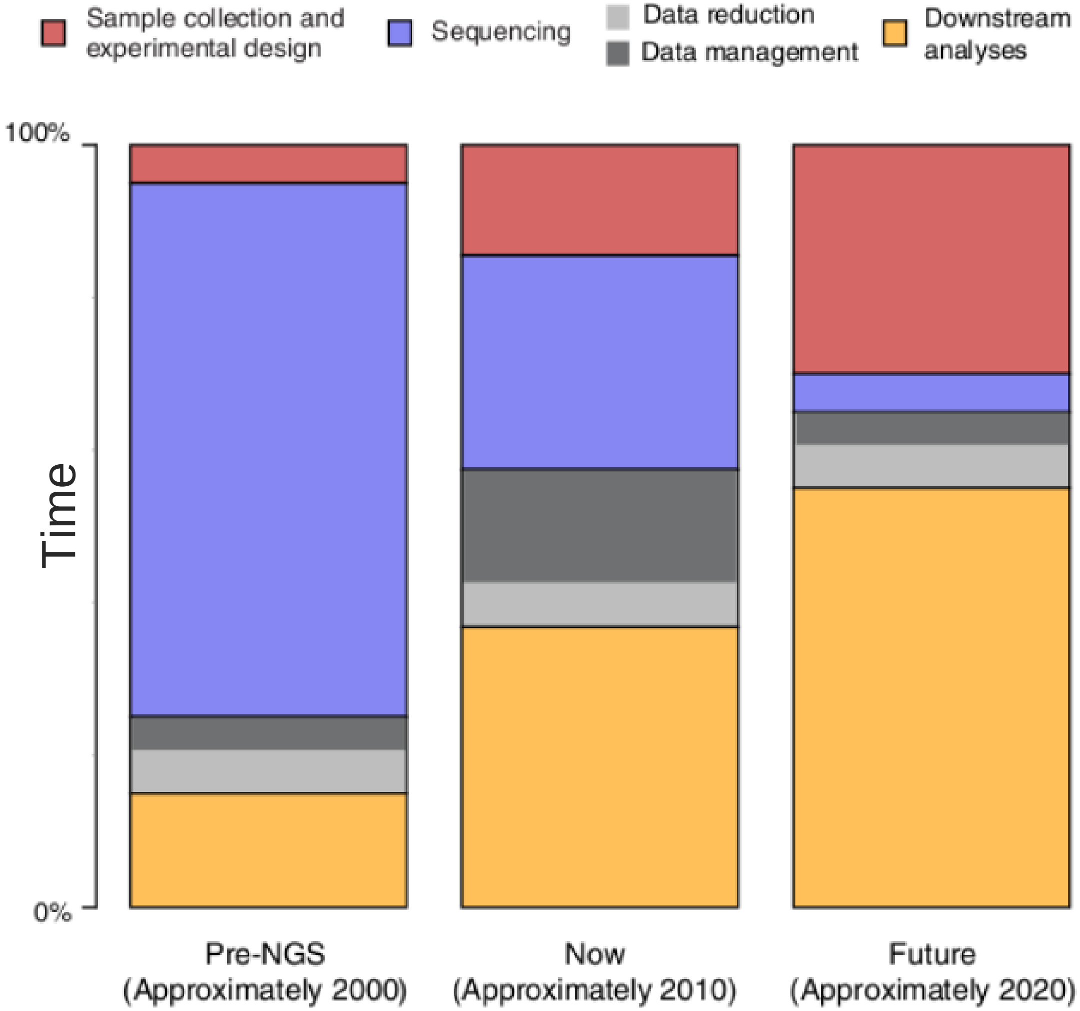
Individual Genome

1h
100 $

2003 2007 2013 2015 2025

Bioinformatics analysis


Sample collection
Experimental design
Sequencing
Data management
Raw reads (FASTA, FASTQ)
Mapped reads (BAM, CRAM, MRF)
Data reduction
High-level summaries (VCF, Peaks, RPKM)
Downstream analyses
(differential expression, novel TARs, regulatory networks, ...)



The Epimed initiative

Medical Epigenetics and Bioinformatics

- **Objectives :**

Facilitate a translational research in epigenetics, between the fundamental research teams and the medical teams.

Analyze large-scale whole-genome data that is essential to understand the epigenome.

Organize an interactive **database** associating **“omics” data** with **biological and clinical data.**

Community

Collaborative network of about **50 people**:

- IAB, CHU, TIMC-IMAG, LJK, CEA
- International collaborations

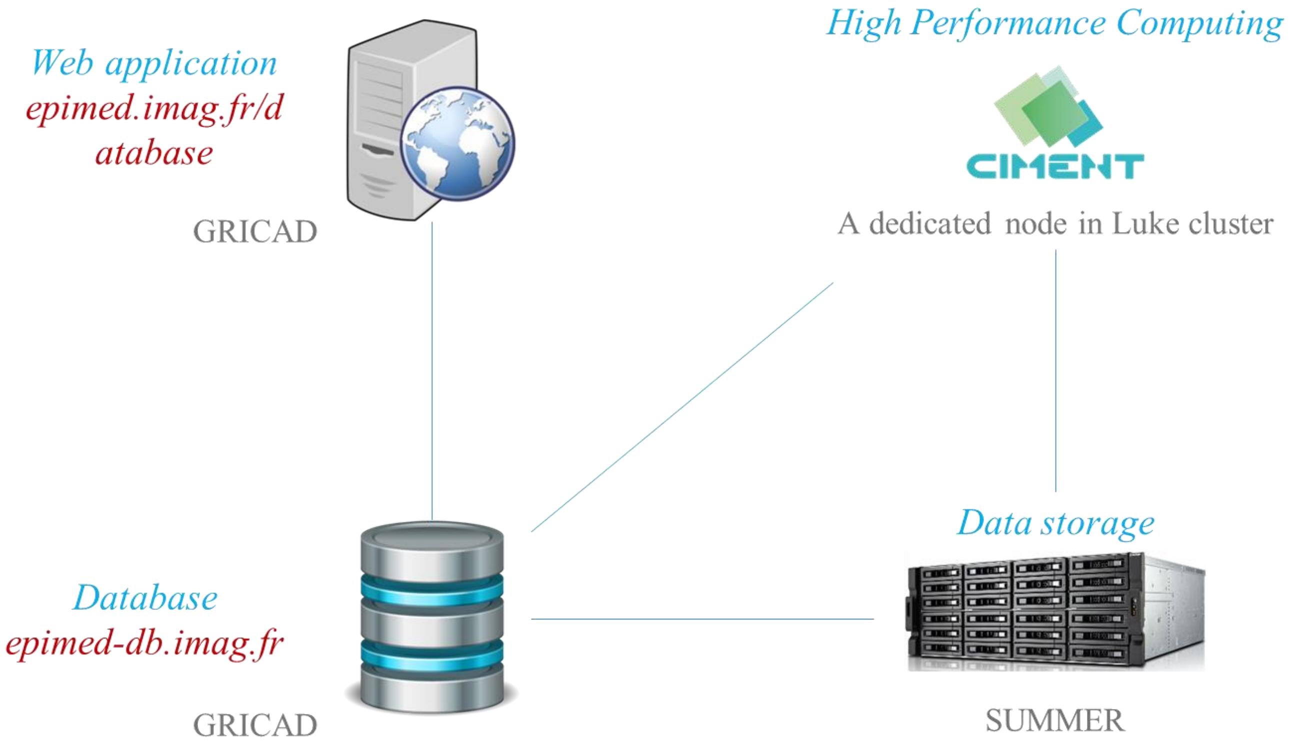
Sophie Rousseaux (DR)
epigenetics,
medical research

Anne-Laure Vitte (IE)
molecular biology

Florent Chuffart (IR)
statistics, computing

Ekaterina Flin (IR)
databases, web systems

The Epimed infrastructure



Web application
epimed.imag.fr/database
GRICAD

High Performance Computing
A dedicated node in Luke cluster

Database
epimed-db.imag.fr
GRICAD

Data storage
SUMMER

iAB
EPIGENETICS
CHRONIC DISEASES
CANCER

CHU
GRENOBLE
ALPES

GRICAD

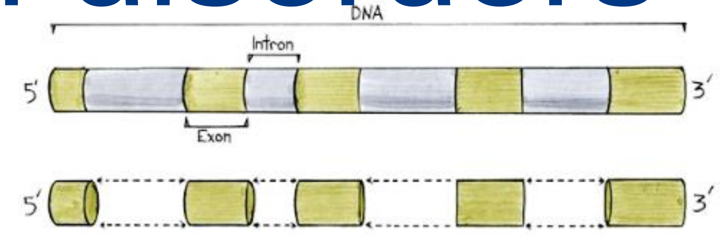
GRENOBLE ALPES
RECHERCHE

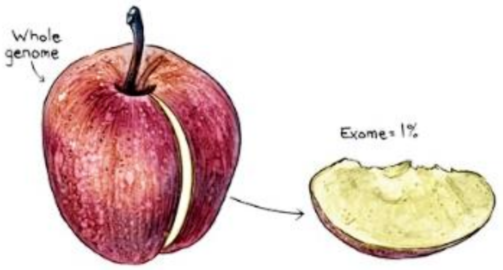
INFRASTRUCTURE DE
CALCUL INTENSIF
ET DE DONNÉES

Exome and rare Mendelian disorders

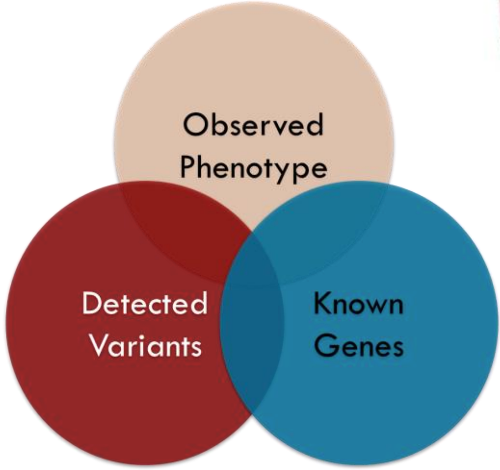
- Mostly **genetic related**
- Mostly **monogenic/mendelian diseases**
- Quite **difficult** to **diagnose** :
  - Often **long** to **diagnose** => 25 % between 5 to 25 years with **traditional methods**
- **Severe :**
  - Child **mortality rates** around **30%** under **5 y/o**
- **Numerous :**
  - Around **7000 different diseases**
  - **7-8% of total worlwide population (around 30M people in Europe)**

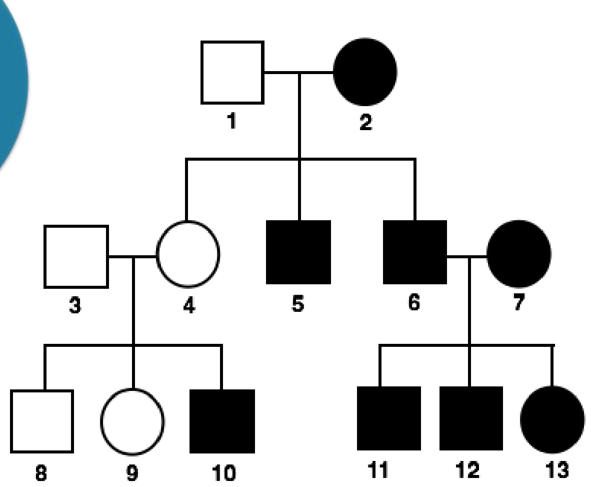
- **Exome analysis :**
  - **Compare sample DNA material to a human « reference »**
  - Find **discordance** between both
  - Relate difference to **patient clinical features** (**phenotype**)
  - **Resolve 40% of undiagnosed case**



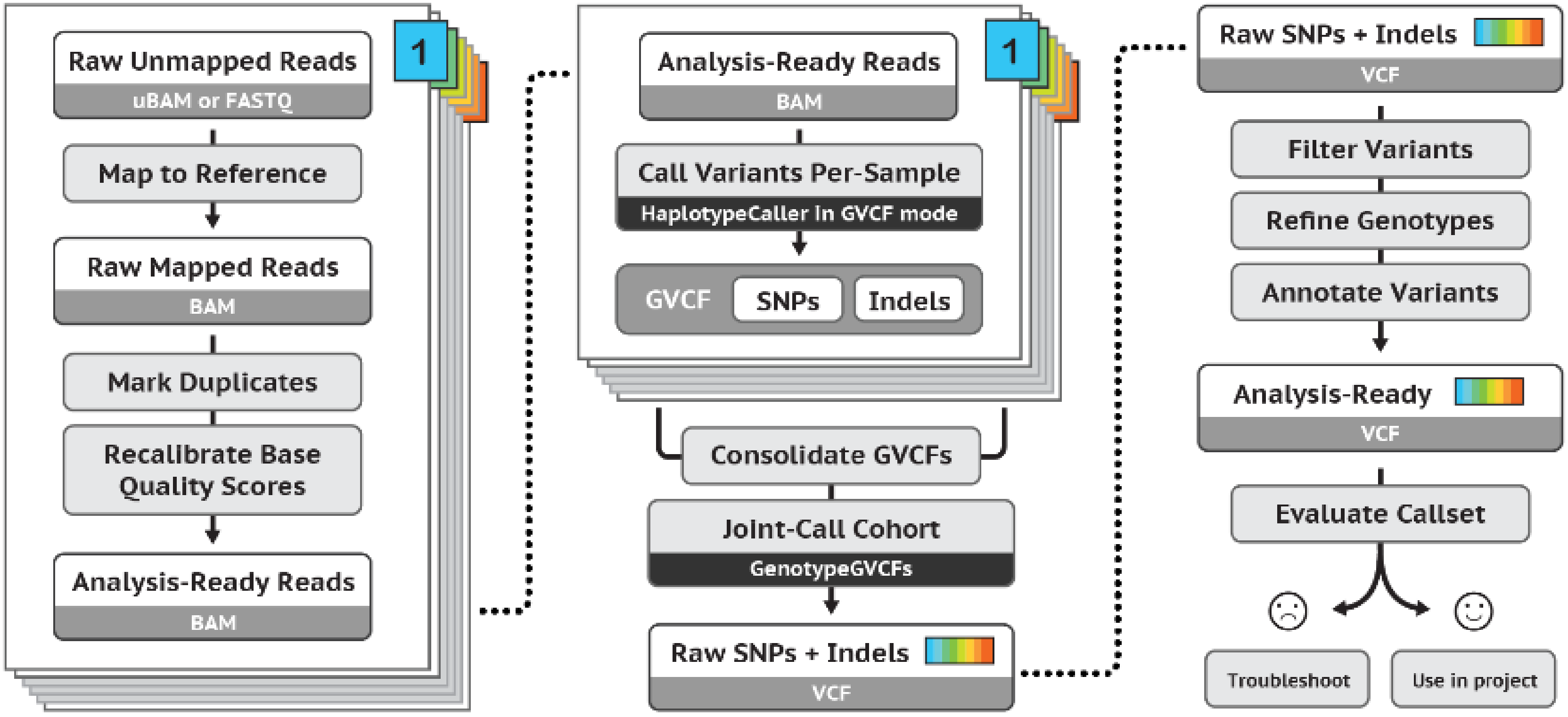


Copyright © 2012 University of Washington





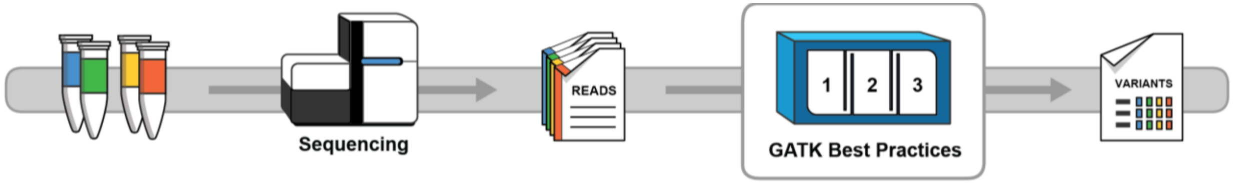
Bioinformatics analysis



Raw Unmapped Reads
uBAM or FASTQ
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM
1

Analysis-Ready Reads
BAM
Call Variants Per-Sample
HaplotypeCaller in GVCF mode
GVCF
SNPs
Indels
1
Consolidate GVCFs
Joint-Call Cohort
GenotypeGVCFs
Raw SNPs + Indels
VCF

Raw SNPs + Indels
VCF
Filter Variants
Refine Genotypes
Annotate Variants
Analysis-Ready
VCF
Evaluate Callset
Troubleshoot
Use in project



Sequencing
READS
1
2
3
GATK Best Practices
VARIANTS

https://software.broadinstitute.org/gatk/

Curr Protoc Bioinformatics. 2013;43:11.10.1-33. doi: 10.1002/0471250953.bi1110s43.

From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.

Van der Auwera GA[1], Carneiro MO[1], Hartl C[1], Poplin R[1], Del Angel G[1], Levy-Moonshine A[1], Jordan T[1], Shakir K[1], Roazen D[1], Thibault J[1], Banks E[1], Garimella KV[2], Altshuler D[1], Gabriel S[1], DePristo MA[1].

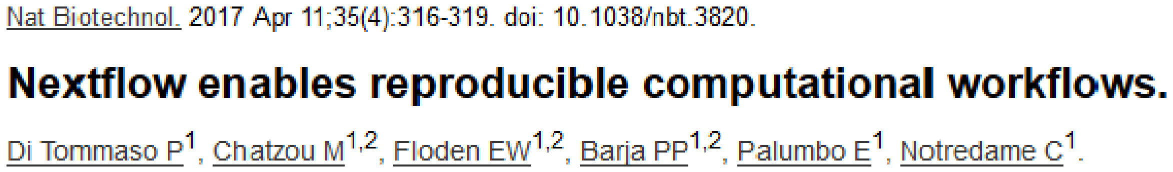
« Required » ressources

- **Cluster Dahu CIMENT :**
  - 72 Dell C6420 (2304 cores)
  - 2x Intel Skylake Gold 6130
  - 192 GB RAM
  - SSD 240GB + SSD 446GB + HDD 4TB

Analysis step	Input	Time	RAM	CPU	Size output	Read disk
Mapping	≈ 2 * 4-5 Gb	≈ 50 – 60 min	≈ 24 Gb	≈ 16000%	≈ 8 - 10 Gb	≈ 30 - 35 Gb
Mark Duplicates	≈ 8 - 10 Gb	≈ 20 - 25 min	≈ 36 - 38 Gb	≈ 10000%	≈ 10 - 12 Gb	≈ 20 - 21 Gb
Base Recalibration 1	≈ 10 - 12 Gb	≈ 55 - 60 min	≈ 36 - 38 Gb	≈ 10000%	≈ 18 - 20 Gb	≈ 45 - 50 Gb
Base Recalibration 2	≈ 18 - 20 Gb	≈ 60 - 70 min	≈ 36 - 38 Gb	≈ 10000%	≈ 18 - 20 Gb	≈ 55 - 65 Gb
Haplotypecaller	≈ 18 - 20 Gb	≈ 30 - 40 min	≈ 36 - 38 Gb	≈ 10000%	≈ 50 - 90 Mb	≈ 30 - 40 Gb
CombineGVCFs	X * 50 - 90 Mb + 0,5 - 3 Gb	≈ 10 - 40 min	≈ 36 - 38 Gb	≈ 10000%	24 * 0,5 - 3 Gb	≈ 10 - 40 Gb
GenotypeGVCFs	24 * 0,5 - 3 Gb	≈ 1 - 30 min	≈ 36 - 38 Gb	≈ 10000%	X * 4 - 5 Mb	≈ 1 - 12 Gb
Annotation	X * 4 - 5 Mb	≈ 18 - 20 min	≈ 48 - 50 Gb	≈ 8000%	X * 30 - 40 Mb	≈ 14 - 20 Gb

Used tools

nextflow

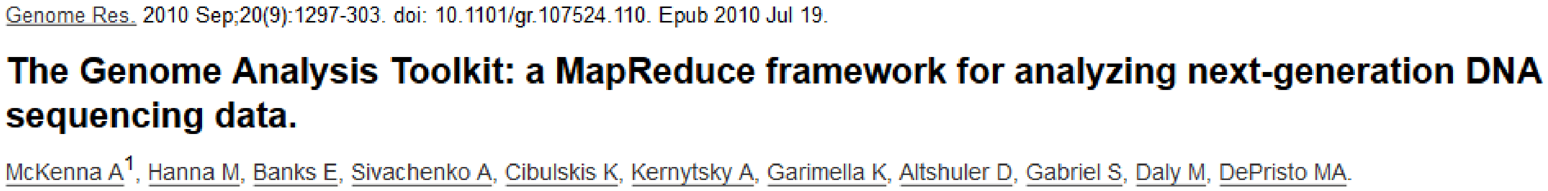


Nat Biotechnol. 2017 Apr 11;35(4):316-319. doi: 10.1038/nbt.3820.

Nextflow enables reproducible computational workflows.

Di Tommaso P[1], Chatzou M[1,2], Floden EW[1,2], Barja PP[1,2], Palumbo E[1], Notredame C[1].

gatk



Genome Res. 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110. Epub 2010 Jul 19.

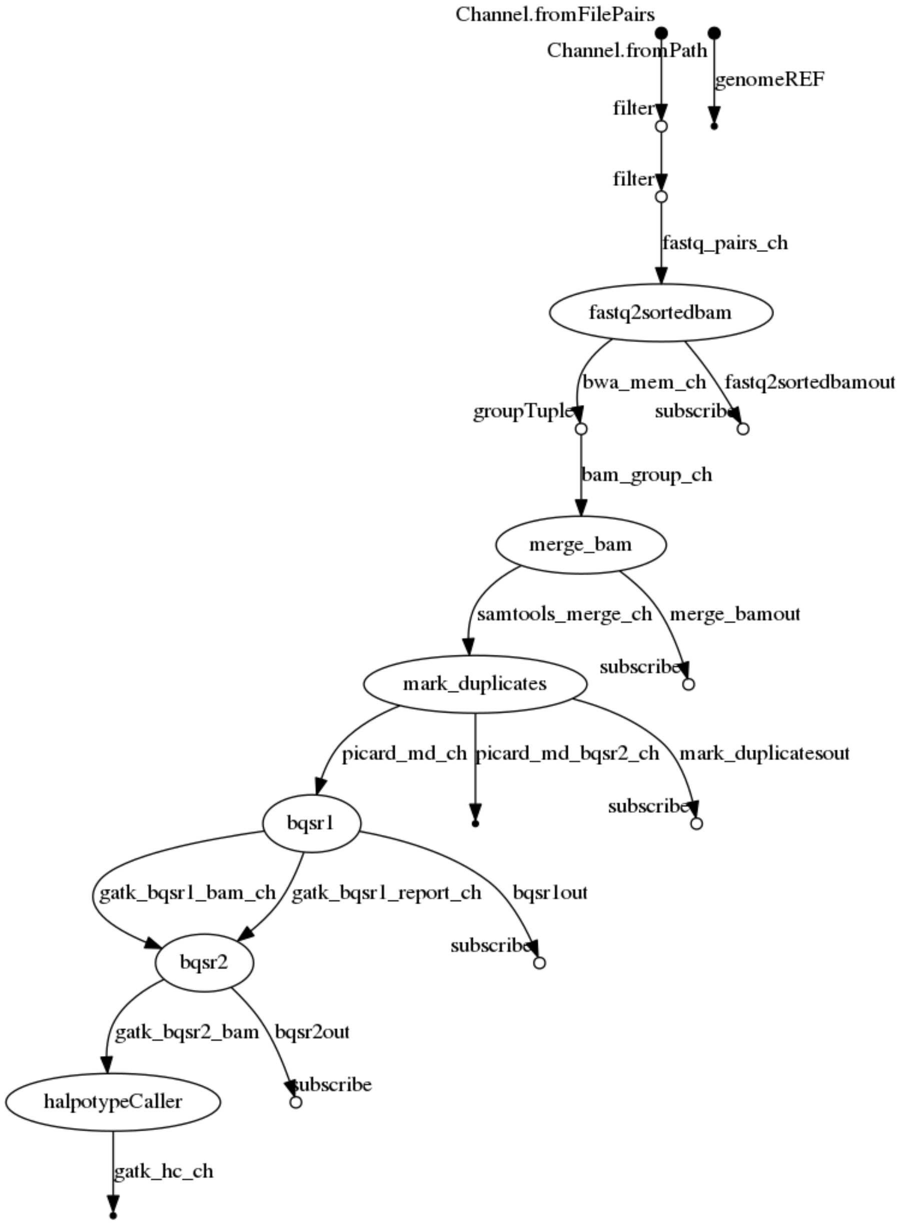
The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.

McKenna A[1], Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA.

Singularity

docker

nextflow



- https://www.nextflow.io/
- **Dataflow manager**
- « **Crash** » **management** and « **resume** »**fonction**
- **Automatic pipeline launching**
- « **Compatible** » with **OAR**, CIMENT batch scheduler
- **Summary reports**

process

```
process fastqc {
  publishDir "${params.resultDir}/fastqc", mode: 'copy'

  input:
  file fastq from fastq_list

  output:
  file '*.zip' into fastq_out_zip
  file '*.html' into fastq_out_html
  val "${params.resultDir}" into fastqc_rep

  """
  fastqc $fastq
  """
}

$fastqc {container = "${params.singularityDir}/fastqc-0.11.15.img"}
```

channels

containers

My « future » pipeline

- **Robust** (managed to process 387 samples during a single week-end)

- Every pipeline version will be « **Vmized** » with all its dependencies (reproductibility for each version)

- **Fully fonctionnal** and **documented exome version** coming **Q1 2019**

- **Coming next, genome version !**

And soon … The Human Genome !

- We need **novel approaches** to **speed up** the analysis :

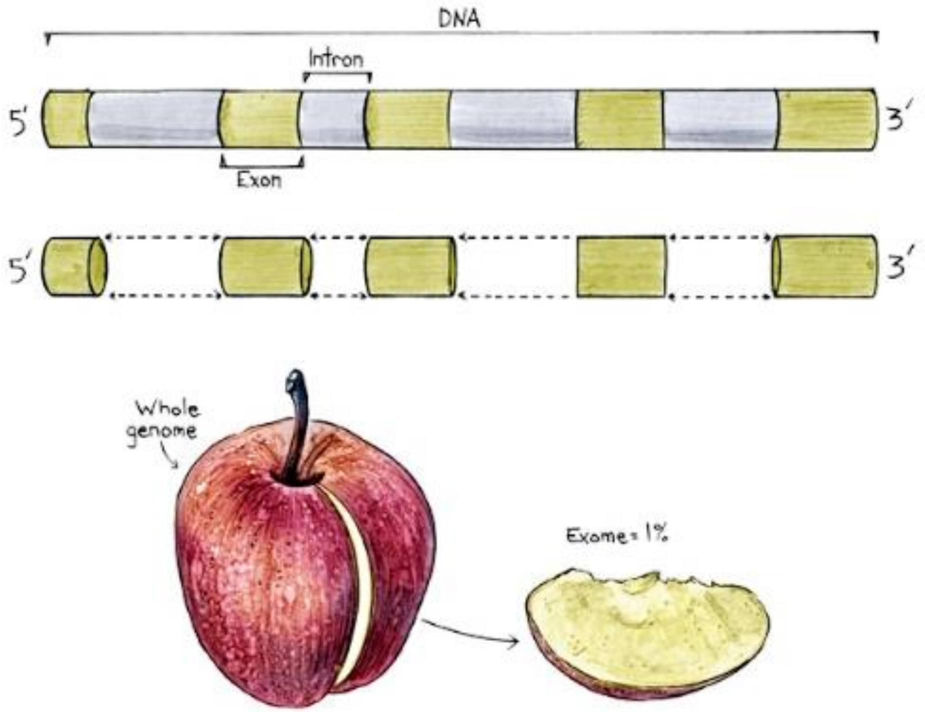
  - **Splitting** operation per **logical unit** (chromosomes) :

    => Reformatting my **code**

  - Speeding up by **massively parallelizing processes** :

    => GPU programming

    => MapReduce Spark



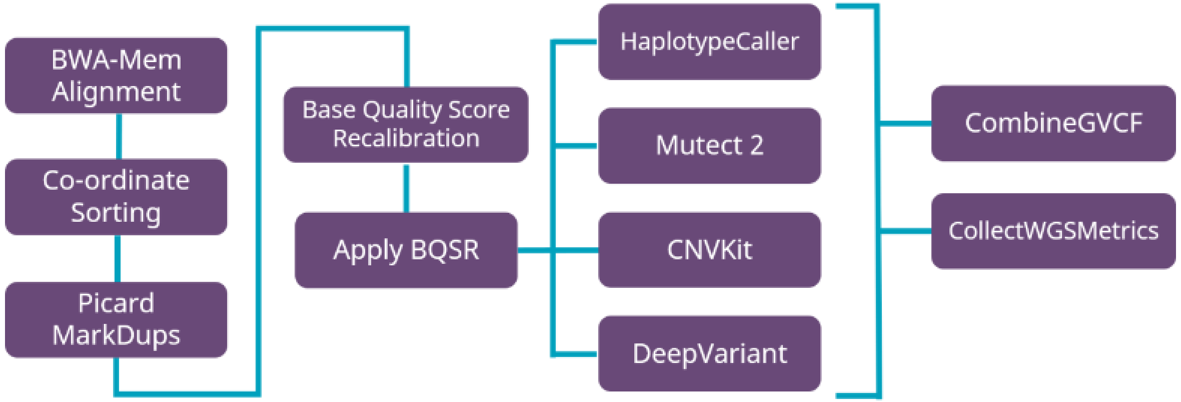
Copyright © 2012 University of Washington

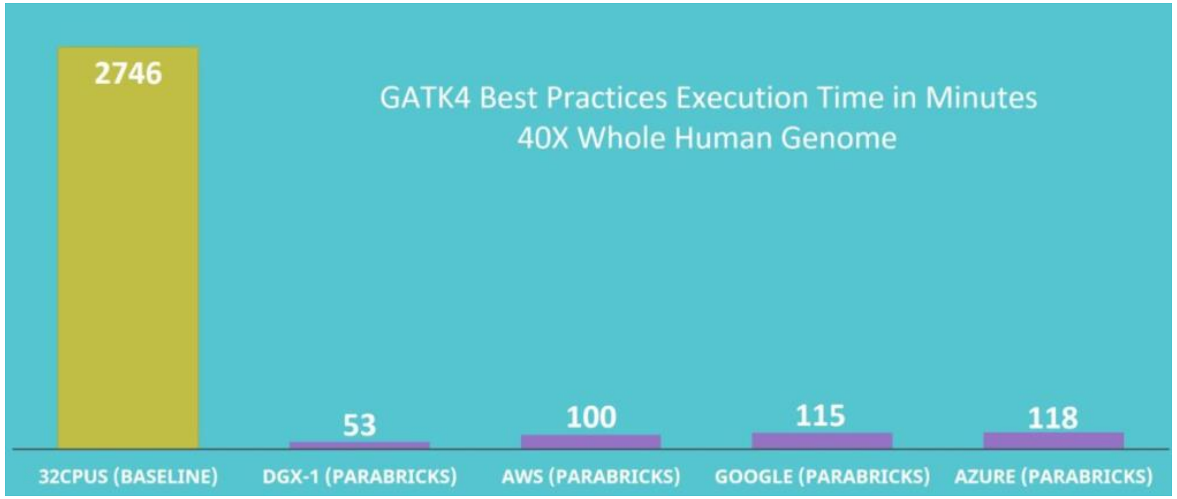
APACHE Spark™

Tools already exist !

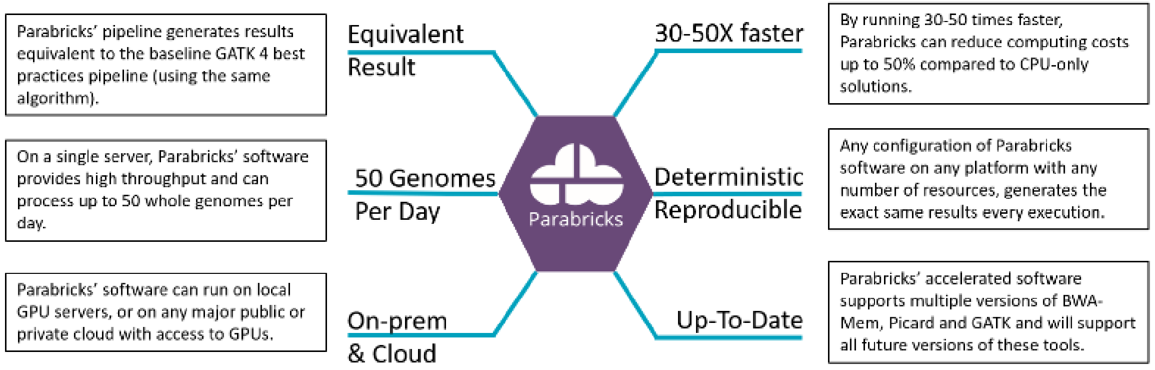
GPU

Parabricks





The Parabricks Advantage



Equivalent Result — Parabricks' pipeline generates results equivalent to the baseline GATK 4 best practices pipeline (using the same algorithm).

50 Genomes Per Day — On a single server, Parabricks' software provides high throughput and can process up to 50 whole genomes per day.

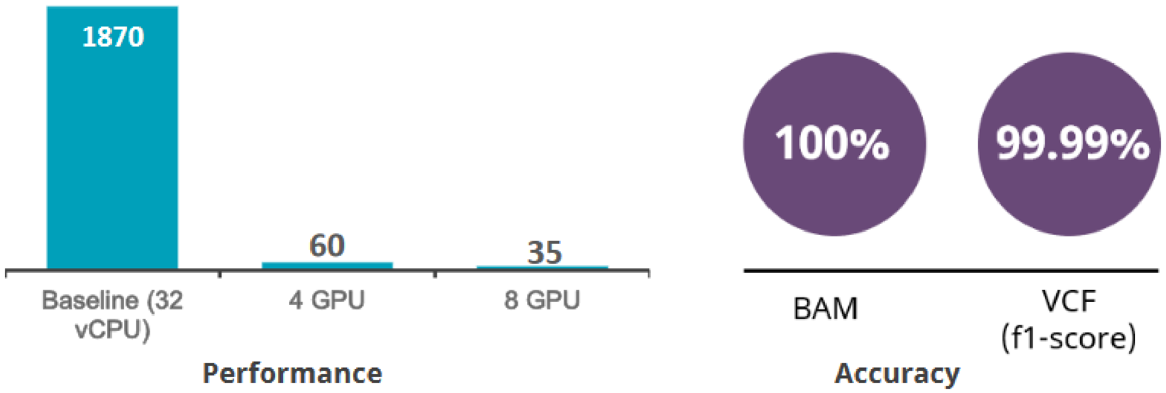
On-prem & Cloud — Parabricks' software can run on local GPU servers, or on any major public or private cloud with access to GPUs.

30-50X faster — By running 30-50 times faster, Parabricks can reduce computing costs up to 50% compared to CPU-only solutions.

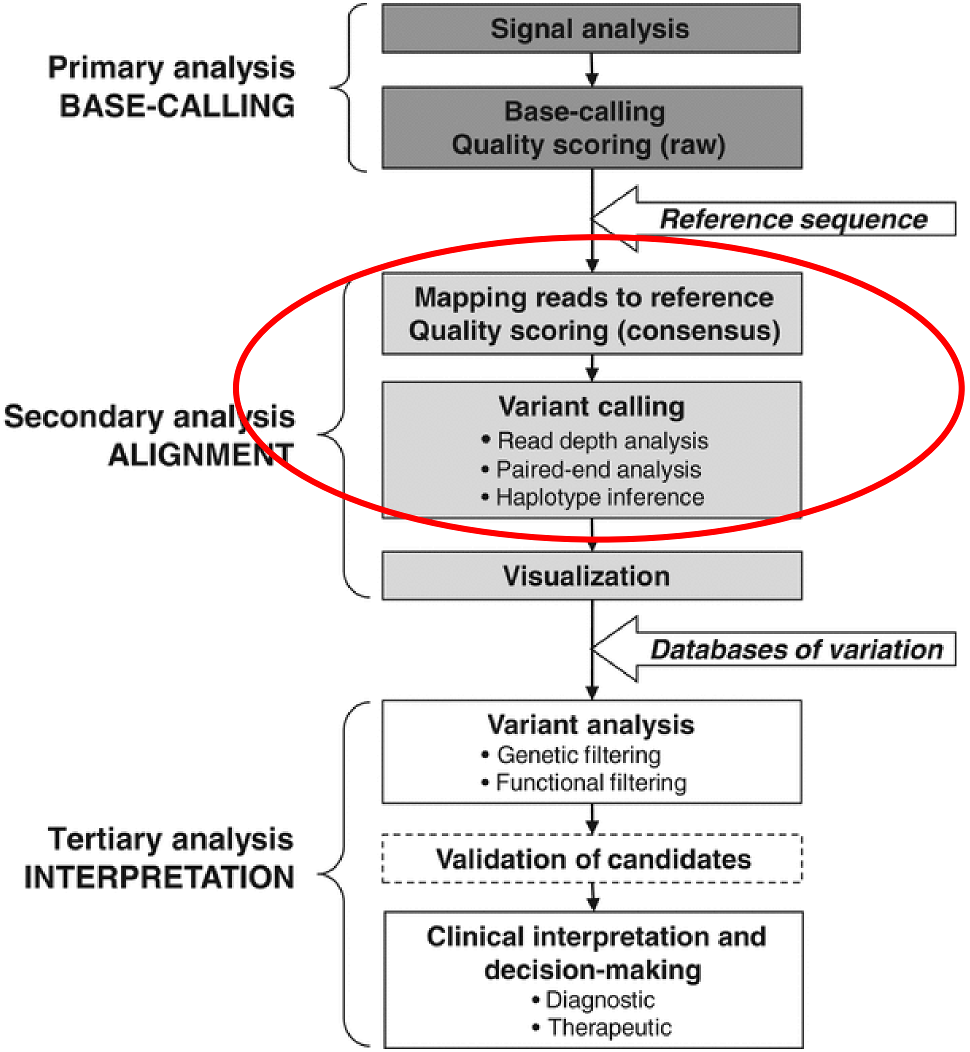
Deterministic Reproducible — Any configuration of Parabricks software on any platform with any number of resources, generates the exact same results every execution.

Up-To-Date — Parabricks' accelerated software supports multiple versions of BWA-Mem, Picard and GATK and will support all future versions of these tools.

Performance Comparison



GATK4 and Spark



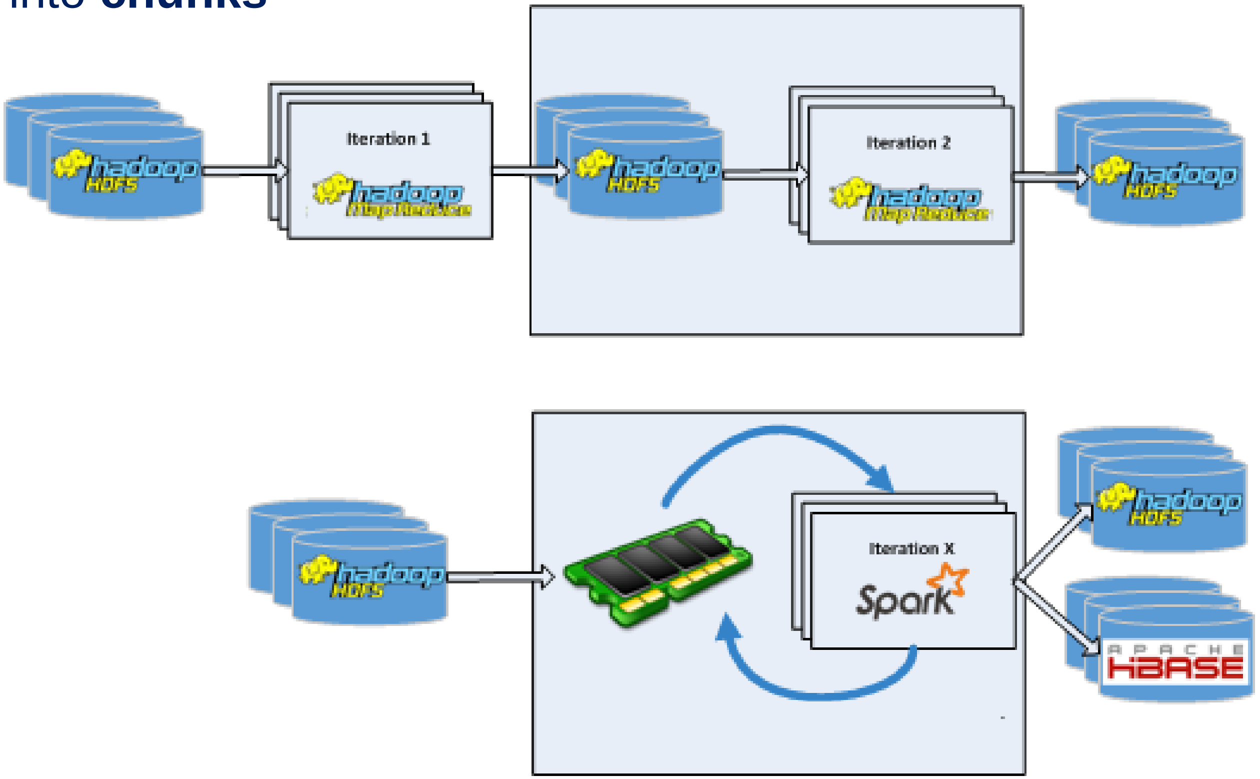
Native Spark tools are **included** into **GATK4** for **mapping** and **variant calling** (two of the most paralellizable steps)

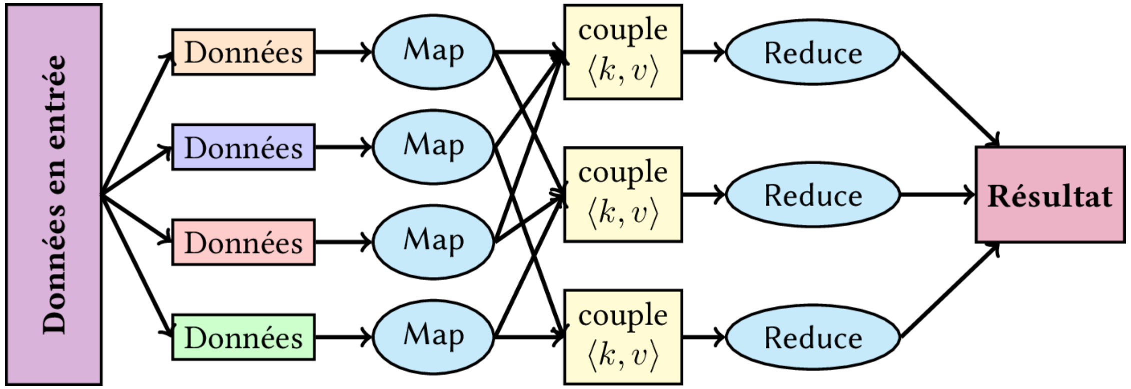
Why Spark?

- Permit to **massively parralelize** processes

- Spark Bioinformatics tools « cut » **biological data** into **chunks**

and do « **bioinformatic** » **operations** on it



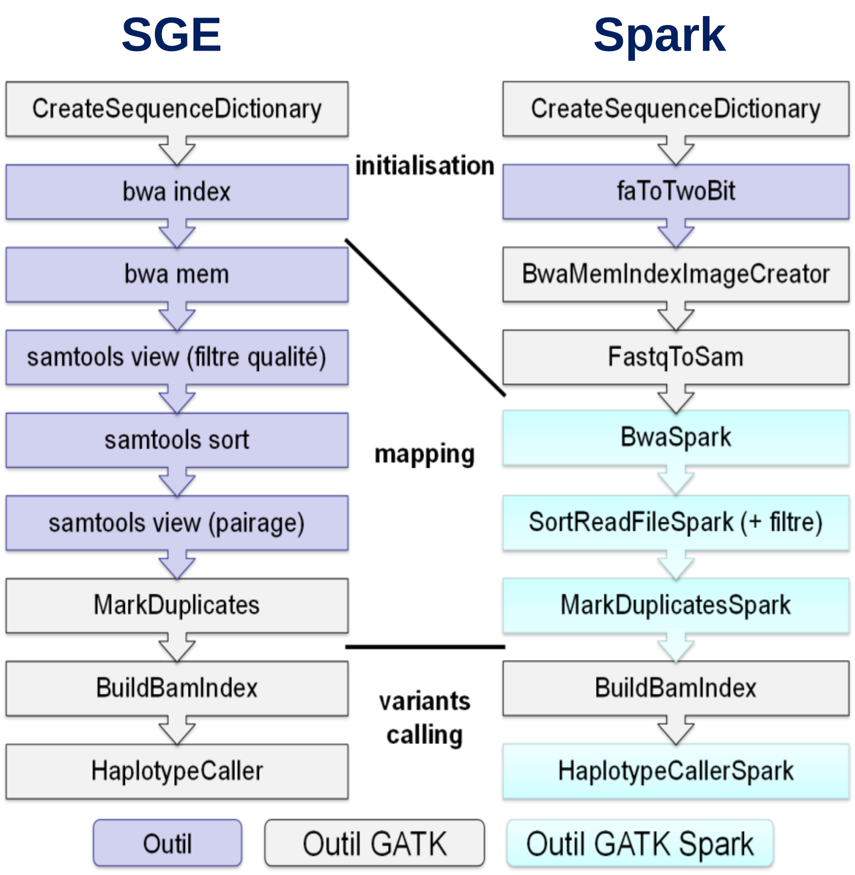


Données en entrée
Données
Données
Données
Données
Map
Map
Map
Map
couple $\langle k, v \rangle$
couple $\langle k, v \rangle$
couple $\langle k, v \rangle$
Reduce
Reduce
Reduce
Résultat

MapReduce

Hadoop MapReduce and Spark

GATK4 Spark benchmarking



- **Toulouse LIPM Cluster** :
  - 8 physical nodes
  - 252 cores
  - 642 Go RAM memory
  - 34To HDD

- One supplementary frontal node to **distribute Spark jobs**

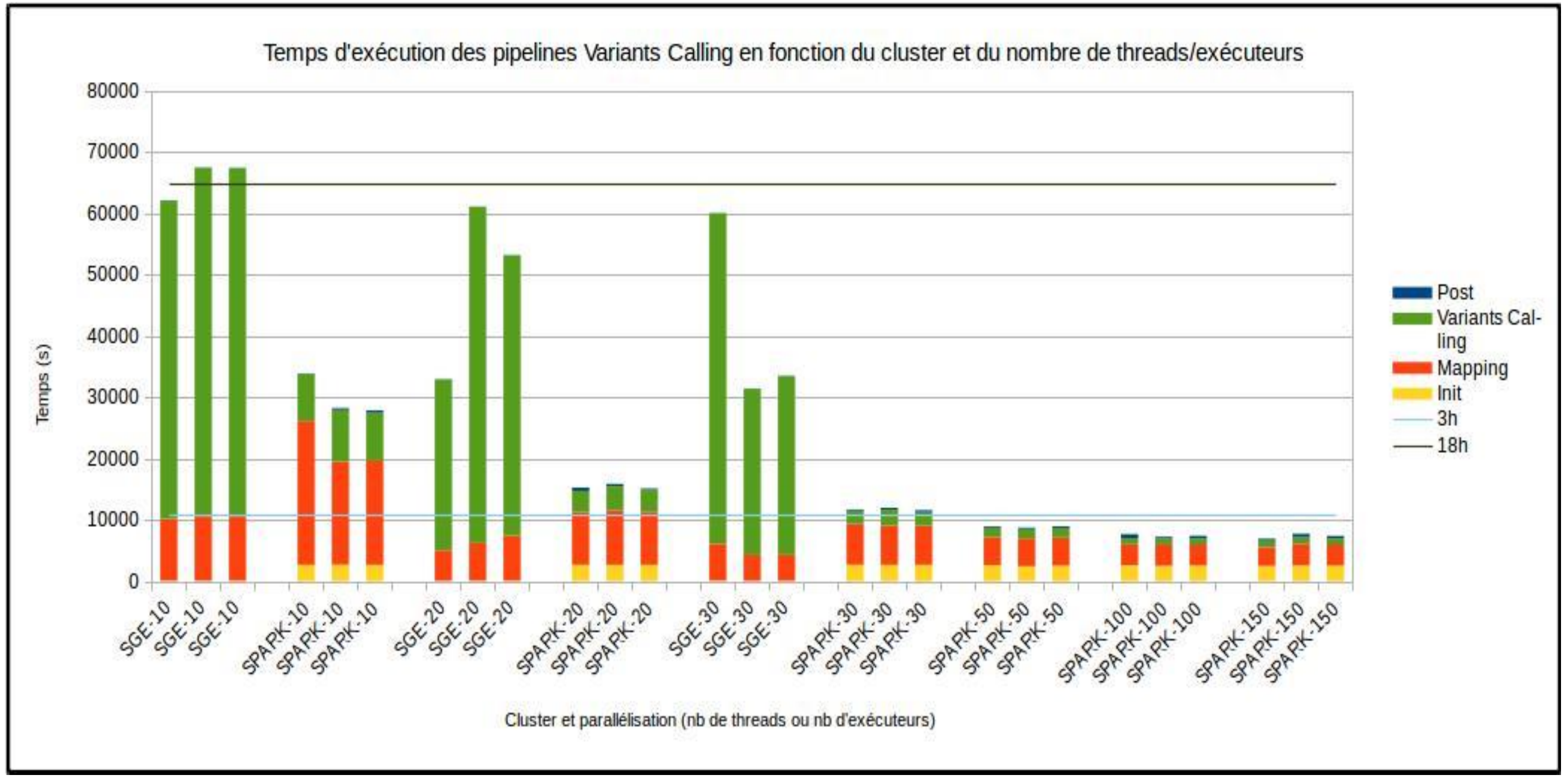
- The **same cluster** have been used for **Spark** and **SGE jobs**, but **not at the same time**

Spark benchmarking

- The **GATK4 non-Spark tools** are **poorly parallelized**

- The **GATK4 Spark tools** are still in **Beta**

- There is a **significative speed improvemen**t with **Spark tools**



Genome : Plasmopara halstedii (75 Mbases), fastq.gz : 2x12Go, 3 repeats

Source : Axel Verdier INRA UMR LIPM Toulouse

Conclusion

- Currently **CIMENT luke/Dahu** HPC nodes are **sufficient** for our **exome analyses**

- But for **high-scale genome analysis** we will need to **step up** :
  - Produce a **better code**
  - Use **bigger HPC infrastructures**
  - Use **Spark** or **GPU** programmed **tools** and **infrastructures**

- It's not that hard ! **Tools already exist** and are indeed **speeding up genome analyses**

- We need to **proceed** with **caution**. We are working with **health datasets**. **No mistake allowed**.

eurofins | Biomnis

iAB
EPIGENETICS
CHRONIC DISEASES
CANCER

Thank you!

UNIVERSITÉ
Grenoble
Alpes

CHU
GRENOBLE
ALPES