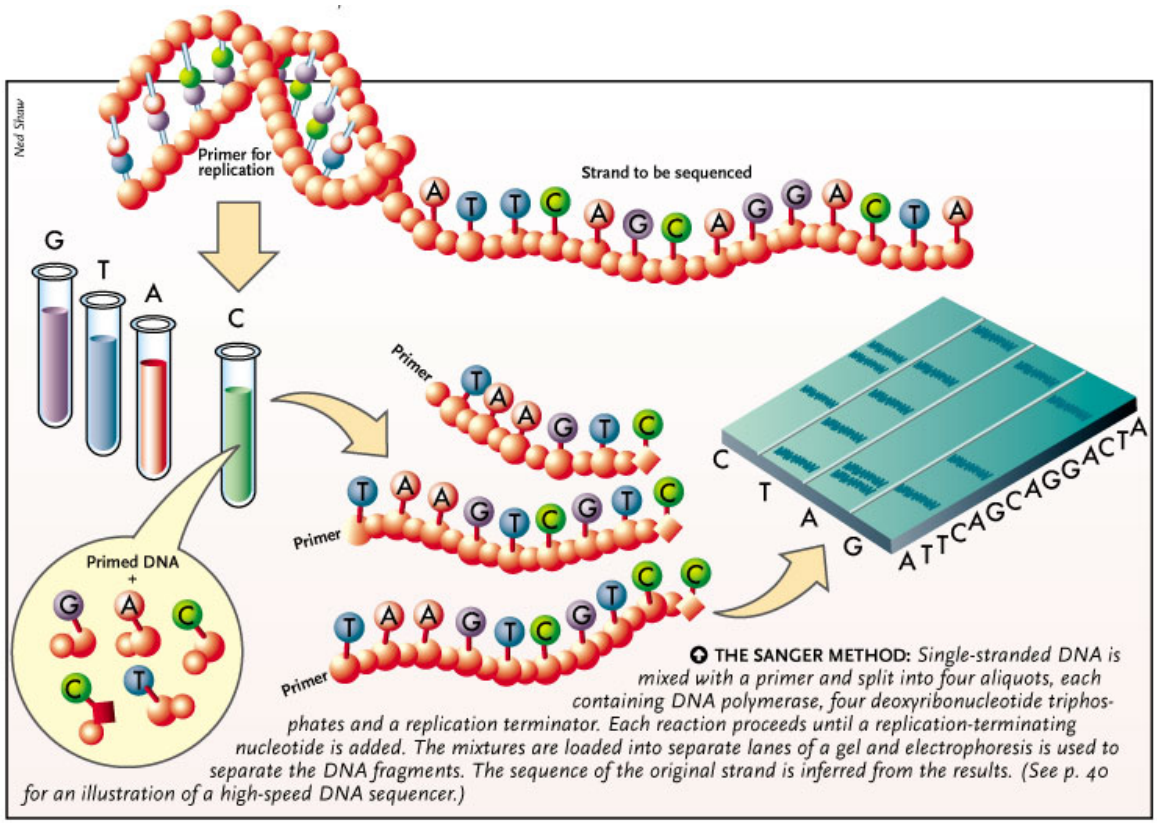
Eoulsan
Analyse du séquençage à haut débit dans le cloud et sur la grille

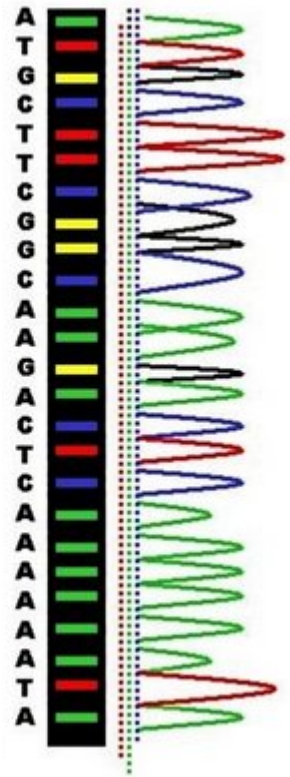
Journées SUCCES

Stéphane Le Crom (UPMC – IBENS)
stephane.le_crom@upmc.fr

The Sanger DNA sequencing method

- **Sequencing by synthesis** by Frédérick Sanger (**1977**, chemistry nobel price 1980).

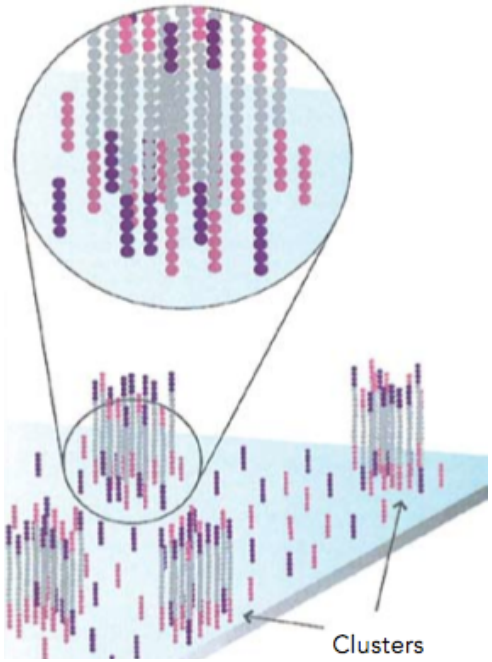


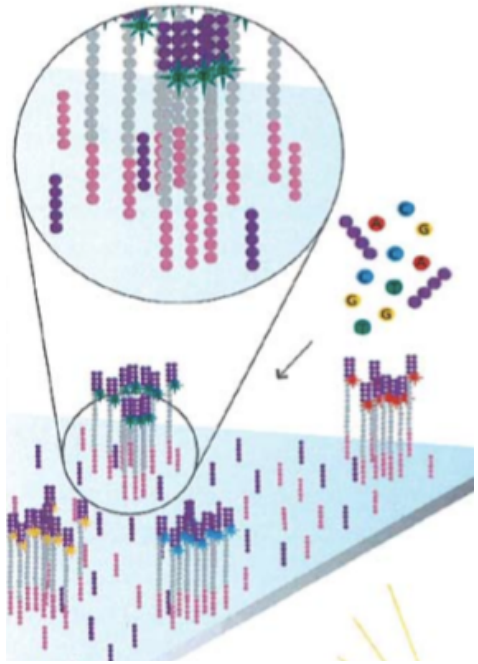


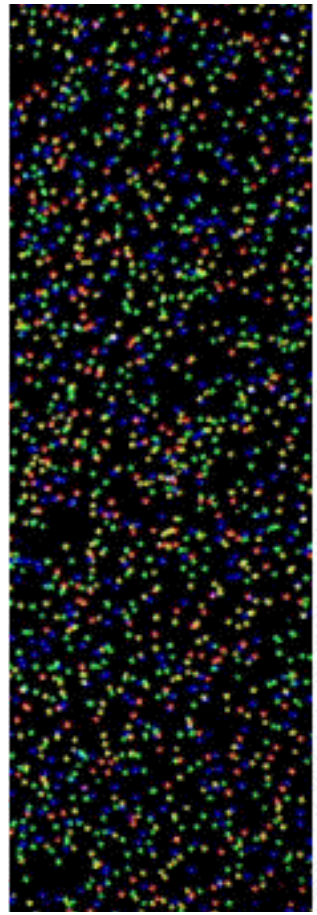
- **126 capillary sequencers** can sequence **one human genome in 12 days**.

From The Scientist

The second generation of sequencing device





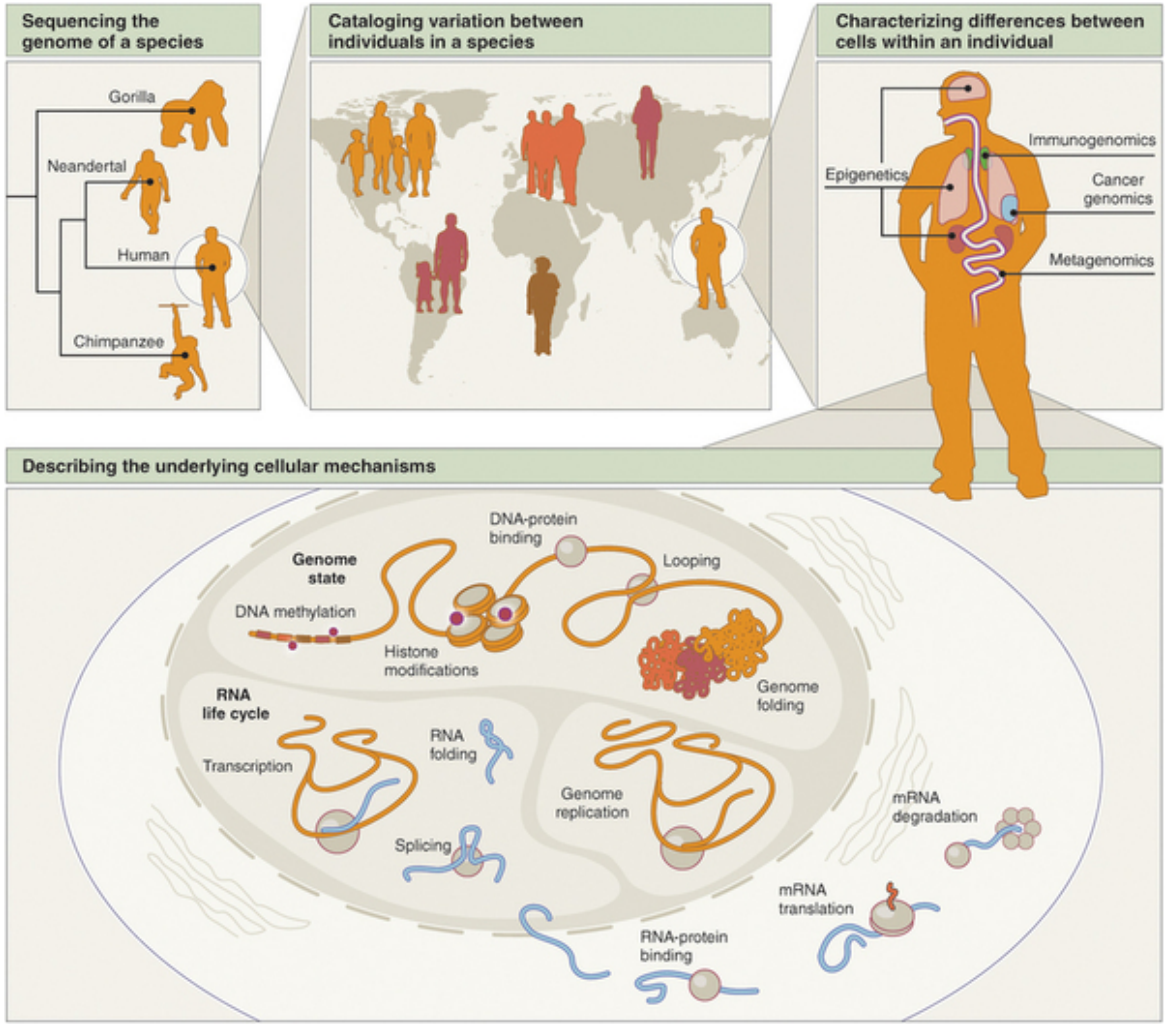


- **One** Illumina **HiSeq** can sequence **two human genomes** in **8 days**.

From htttp://www.illumina.com

Applications cover a lot of biological fields

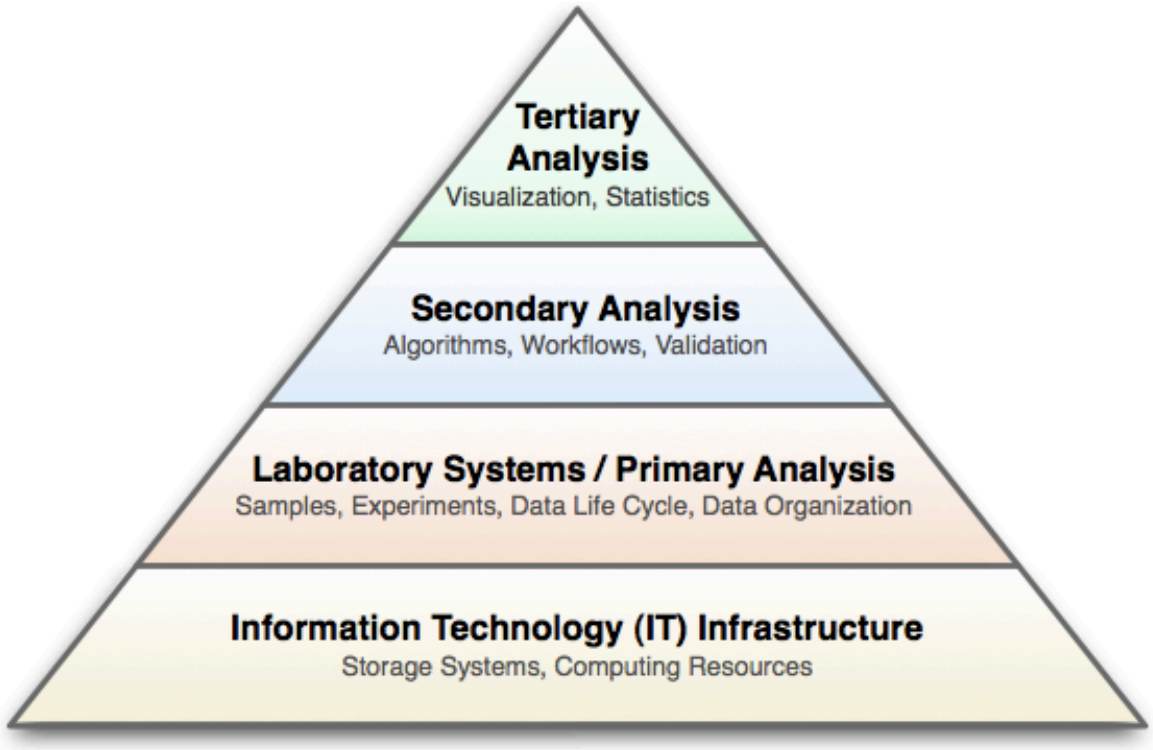
- De novo sequencing;
- Resequencing;
- Functional analysis;
- Metagenomic;
- Diagnostics;
- Personal genomic…



Shendure & Aiden (2012) Nat. Biotech.

Data analysis levels

- Sequencer outputs produce **To of data for each run** and large raw result files.
- Data analysis on big genomes **require calculation power and memory**.

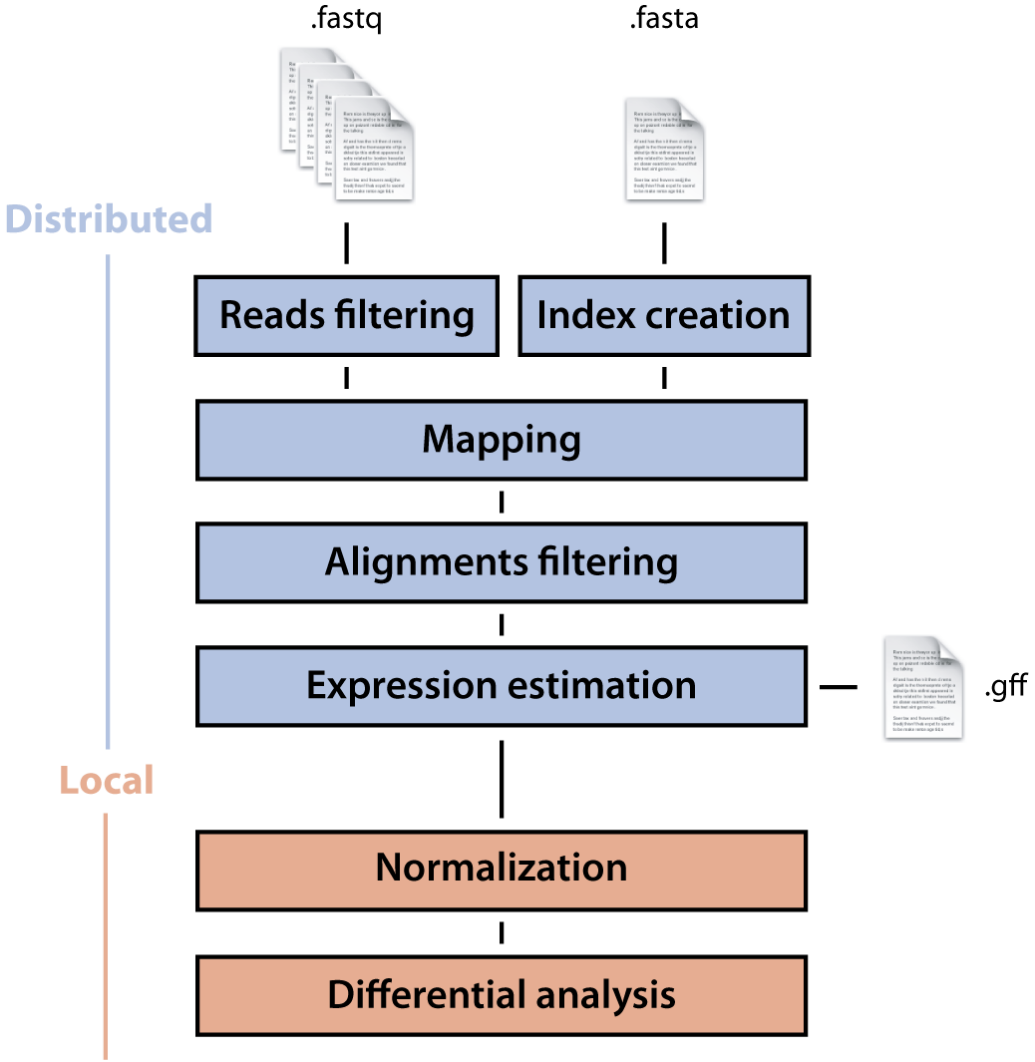


http://www.geospiza.com/finchtalk/labels/Next Generation Sequencing.html

Eoulsan: our RNA-Seq data analysis workflow

- Eoulsan works **from raw sequencer outputs** (Illumina) **to the list of differentially expressed genes**.
- Its **modular** and **flexible** analysis framework runs with various already available analysis solutions (SOAP2, Bowtie, Bowtie2, BWA, GSNAP).
- New **function extension** through external java plug-in.
- **Distributed calculation** to speed up the analysis.

Eoulsan

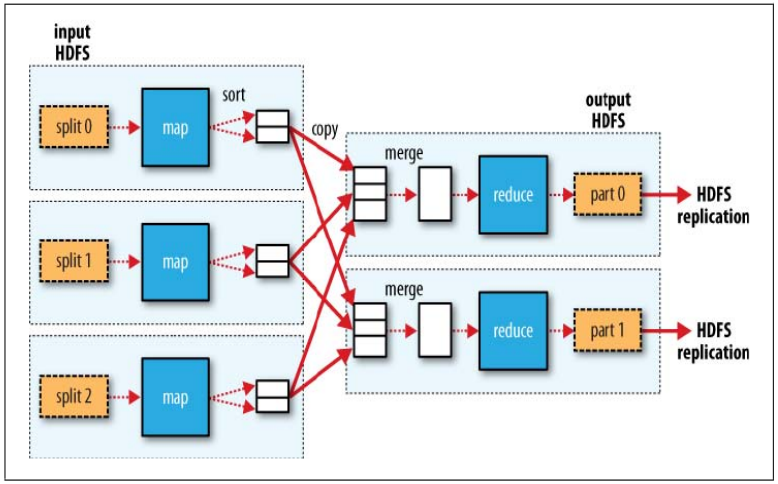


Jourdren (2012) *Bioinformatics*

http:/transcriptome.ens.fr/eoulsan/

We use MapReduce to increase analysis speed

- **MapReduce** is used for parallel computation and **automatically handles duties**, such as job scheduling, fault tolerance and distributed aggregation.



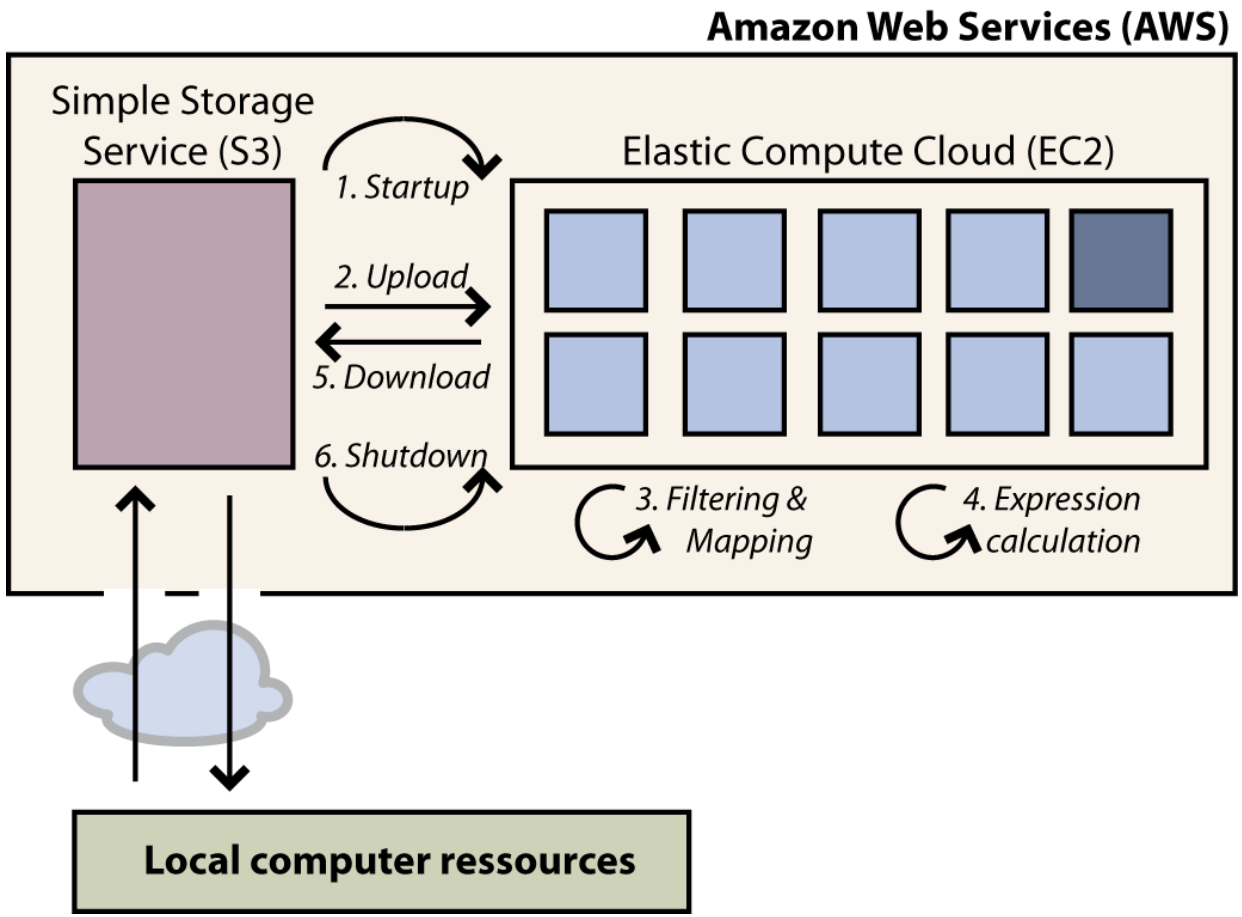
White (2009) *O'Reilly Media*

```
Map(id_alignment, alignment)
    → list(id_exon, 1)

Sort(id_exon,1)

Reduce(id_exon, list(1,1...1))
    → list(id_exon, count)
```

- **Hadoop** is a popular (Twiter, facebook, eBay …) **open-source implementation** of the MapReduce framework as the original Google implementation is not public.

- Hadoop is a **Java framework** that can be executed on any cluster.

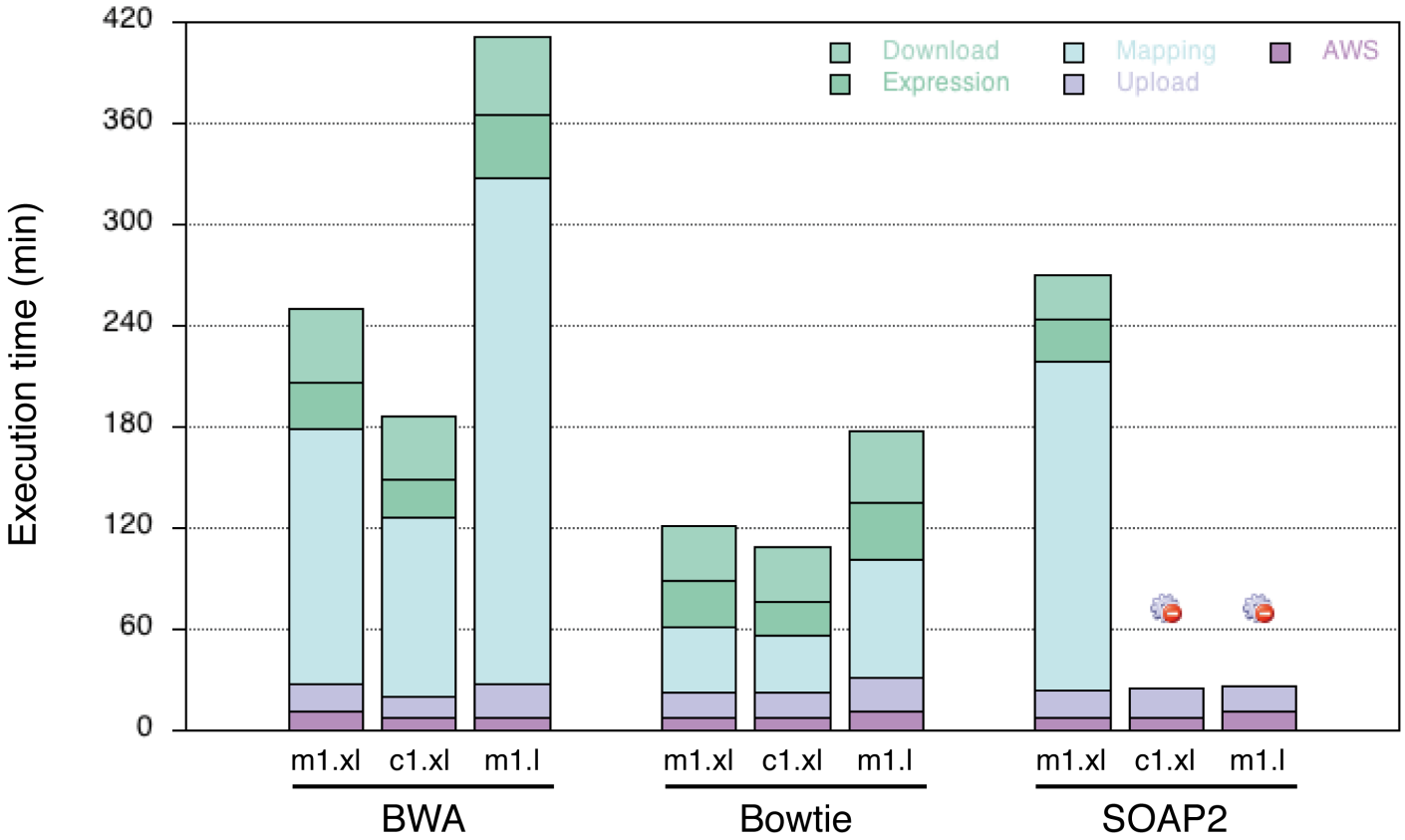
Eoulsan workflow on Amazon Web Services




```
$ eoulsan.sh -conf conf-aws.txt awsexec -d "Job name" param-aws.xml design.txt
s3://sgdb-test/demo
```

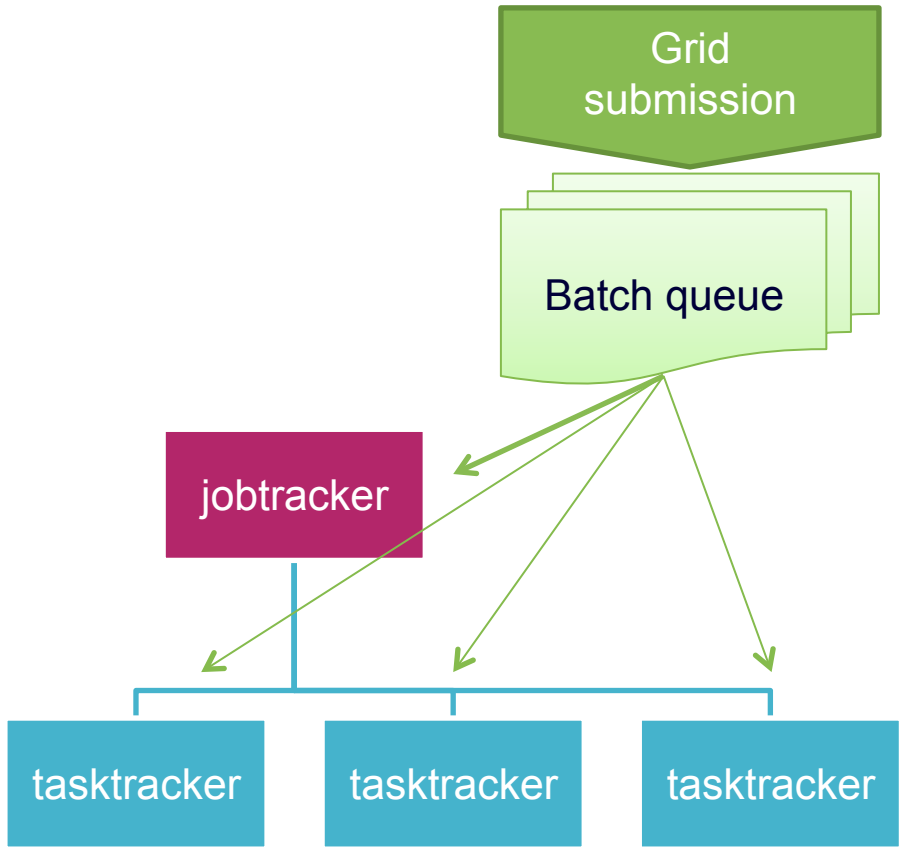
Tests for instance and mapper types

Instance	Memory (Go)	CPU (EC2 unit)	I/O performance	Price USD/hour
m1.large	7.5	4	high	$0.44
m1.xlarge	16.0	8	high	$0.88
c1.xlarge	7.0	20	high	$0.88



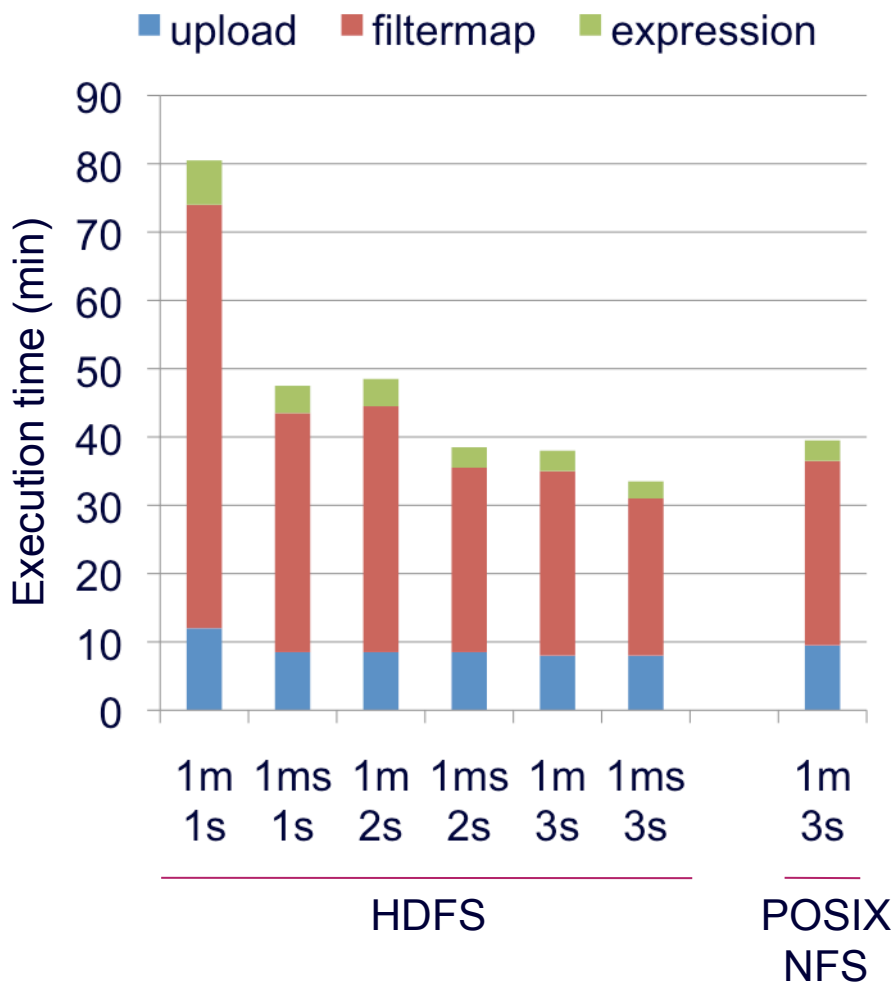
Can we run Eoulsan on a grid infrastructure?

- **Is Hadoop compatible with grid cluster?**
- Several technical solutions need to be addressed.
- **Task management**
  - Master and slaves on the grid cluster;
  - Or, only slaves on the grid cluster.
- **Data storage**
  - HDFS storage available on site;
  - POSIX/NFS storage to build HDFS;
  - HDFS cluster run locally on nodes.



First results

- Tests using a **R&D grid cluster at LLR**.
- **4 identical nodes**
  - E5520 2.27GHz 16 virtual cores;
  - 48GB RAM;
  - 250GB hard drive;
  - 1Gb/s Ethernet connection;
  - OS Scientific Linux 6.3 64bit;
- hadoop-1.0.4-1.x86_64
- **Storage**
  - Dedicated HDFS cluster;
  - Or POSIX/NFS storage.
- **Same** Mouse RNA-Seq **data** than the one we use on **AWS**.

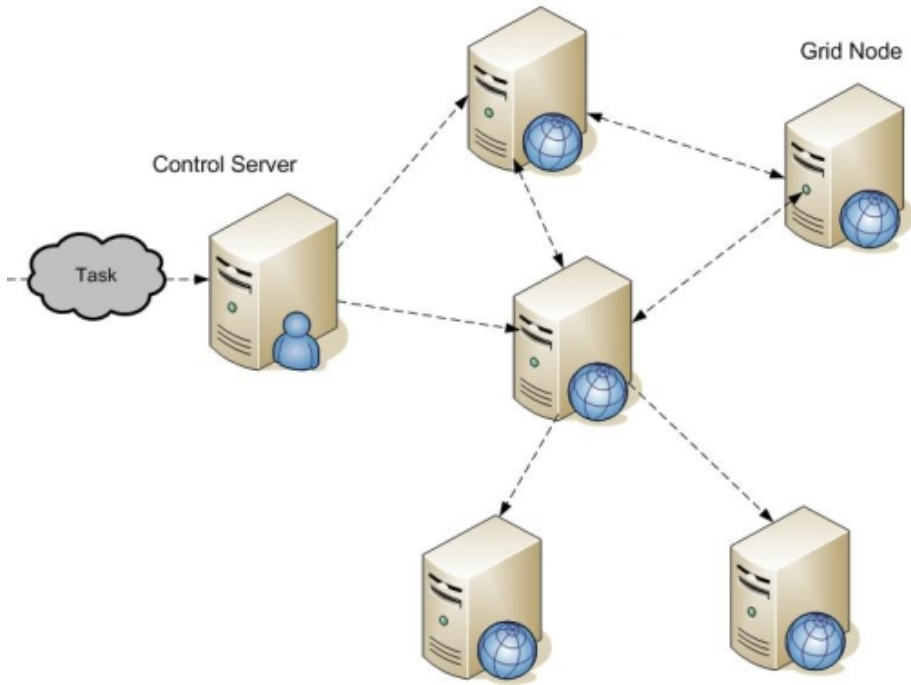


Conclusion

- The **first tests are very promising**: Hadoop can work on grid cluster.
- Eoulsan delivers **quicker results** than using Amazon Web Services.
- **Storage choice as no influence** on analysis performance.
- There is a **maximum number of slaves** after which the gain is minimal (5 for the task we used for the tests).
- **Interface for downloading/uploading** data from/to HDFS/NFS/HTML/Grid storage is available.
- The workflow heavily uses **whole node scheduling**. This is now a standard tool on the GRID but it may be nontrivial to effectively implement it.

Future work

- **Benchmark larger datasets** to compare performances between grid and AWS.
- Allow Eoulsan to remotely run on the grid.
- **Final goal**: make the **choice** of the computing infrastructure **transparent for final users**
- By **creating generic VM** to run Hadoop everywhere:
  - CNRS Research Cloud;
  - GenoCloud (BioGenOuest);
  - Stratuslab cloud…



Acknowledgments

Laurent Jourdren

Stéphane Le Crom

Maria Bernard

Claire Wallon

Vivien Deshaies

Sophie Lemoine

Sandrine Perrin

Andrea Sartirana

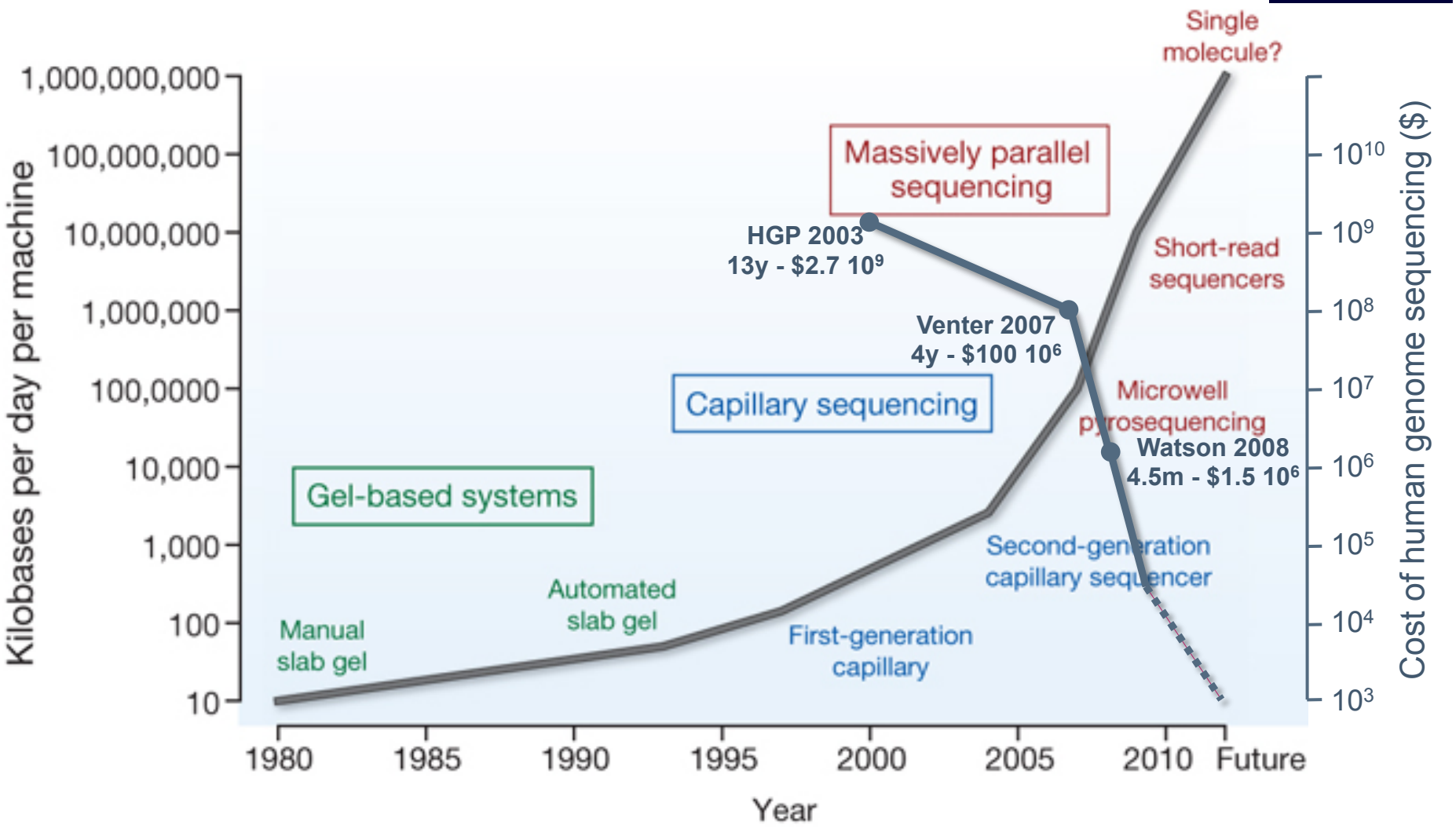
Philippe Busson

David Chamont

Pascale Hennion

Paulo Mora de Freitas

Evolution of sequencing technologies



Stratton (2009) *Nature*

A ready-to-use software solution

- **Aim:** to automate the analysis of a large number of samples at once.
- With **minimal file requirements**.
  - Data: several Fastq files (.bz2)
    1 reference genome (.fasta)
    1 annotation file (.gff3)
  - Set up: 1 XML parameter file
    1 design file

```
<!-- Filter reads -->
<step skip="false">
        <name>filterreads</name>
        <parameters>
                <parameter>
                        <name>lengthThreshold</name>
                        <value>11</value>
                </parameter>
                <parameter>
                        <name>qualityThreshold</name>
                        <value>12</value>
                </parameter>
        </parameters>
</step>
```

- A design file inspired from the limma R package.

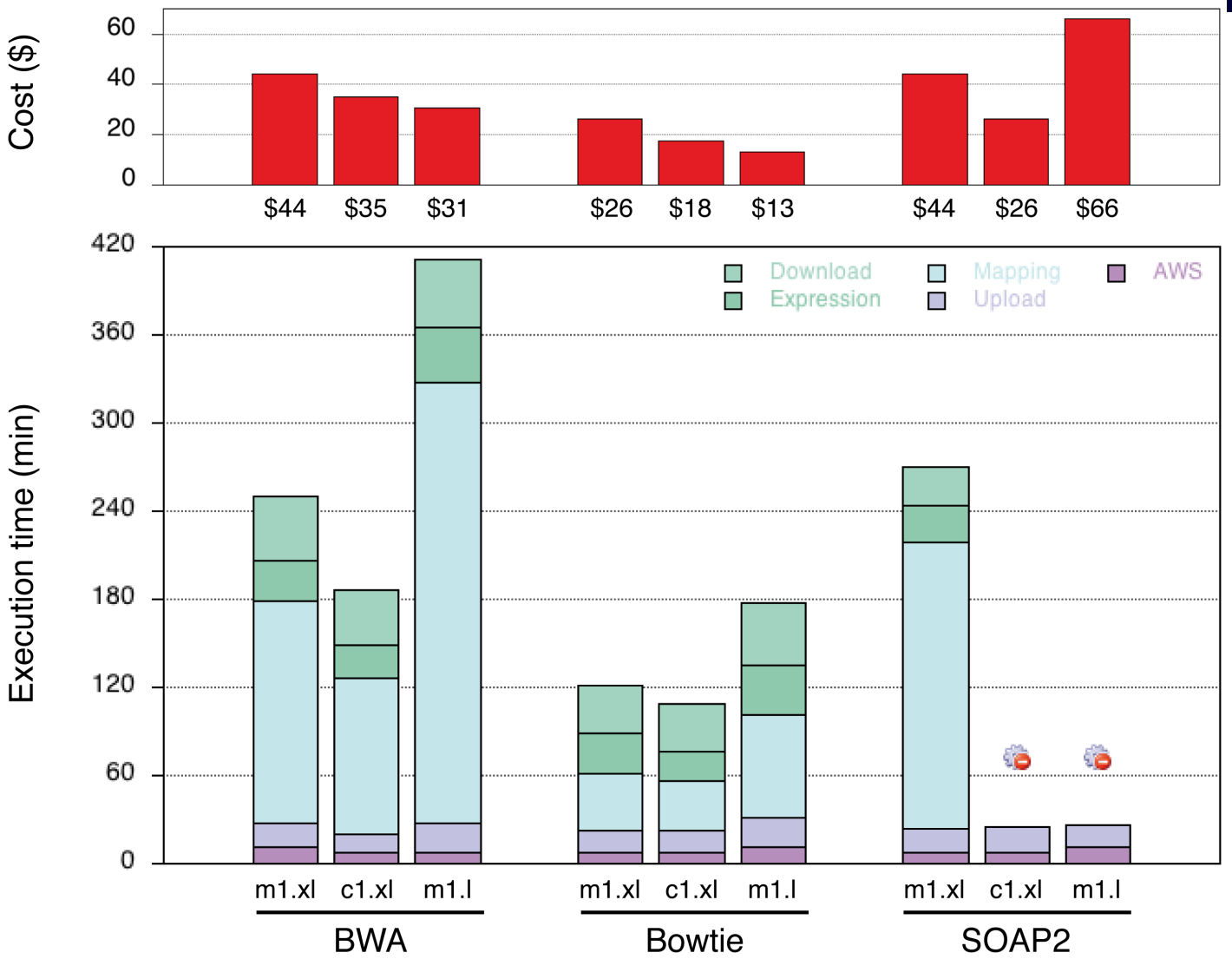
SampleNumber	Name	FileName	Genome	Annotation	Condition	ReplicateType
1	s1	s1.fq	mouse_build37.fasta	mouse_build37.gff	c1	B
2	s2	s2.fq	mouse_build37.fasta	mouse_build37.gff	c2	B
3	s3	s3.fq	mouse_build37.fasta	mouse_build37.gff	c1	B
4	s4	s4.fq	mouse_build37.fasta	mouse_build37.gff	c2	B

- And **one command line** to launch the whole process.

```
$ eoulsan.sh exec parameter_file.xml design_file.txt
```

http:/transcriptome.ens.fr/eoulsan/

Tests for instance and mapper types



Conclusion

- Eoulsan **automates** the analysis of **a large number of samples at once**;
- Its **modular** and **flexible** analysis **framework** runs with various already **available analysis solutions**;
- On **several computer infrastructure** types.

- **Final goal**: make the **choice** of the computing infrastructure **transparent** for final users.
- By **creating generic VM** to run Hadoop everywhere (Private Cloud, Stratuslab)

http:/transcriptome.ens.fr/eoulsan/

Standalone version

Distributed version

Local clusters

Cloud Computing

Grids