Laboratoire Écologie et Évolution des Interactions
FRANCE

Université de Perpignan Via Domitia

David Roquis, *Rémi Emans, Guillaume Mitta & Christoph Grunau*

Galaxy for the rest of us

Ph.D. student, 2nd year

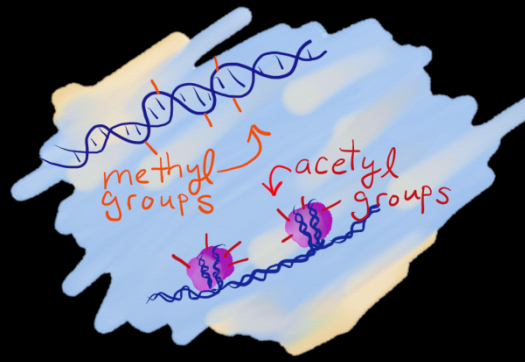
Galaxy Day
UPMC
December 4th, 2013

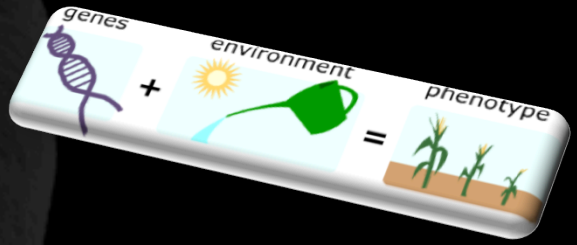
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Galaxy

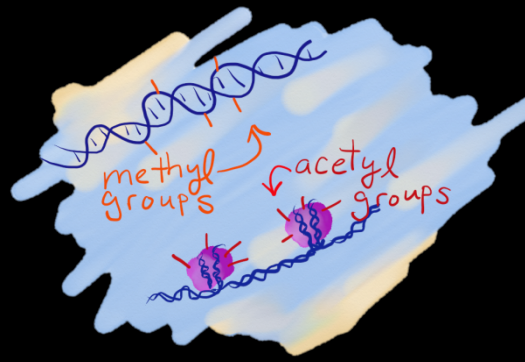
Our laboratory

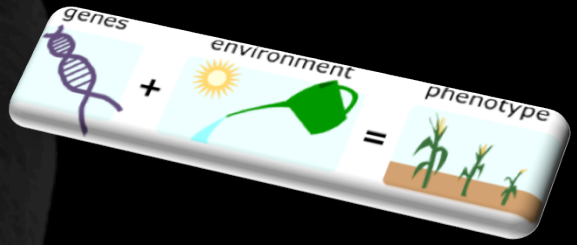
Our laboratory

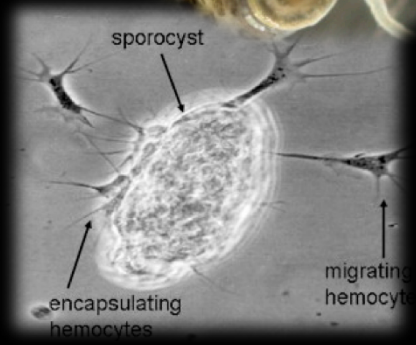




Our laboratory

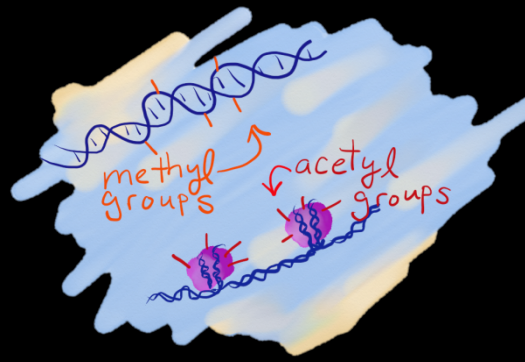


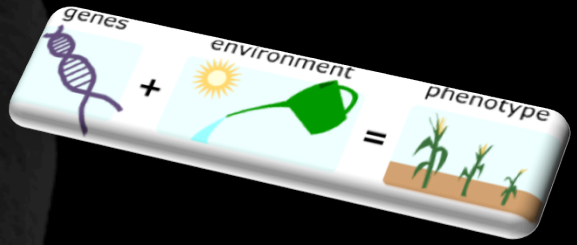


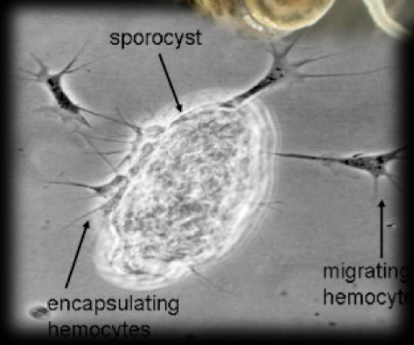


Our laboratory

Université de Perpignan Via Domitia







WORLD HEALTH ORGANIZATION
ORGANISATION MONDIALE DE LA SANTÉ

Our (non) models

A human parasite

Schistosoma mansoni

A freshwater snail

Biomphalaria glabrata

A tropical coral

Pocillopora damicornis

An insect

Dinocampus coccinellae

New Experimental Approaches

- Previously → Candidate gene approach.

New Experimental Approaches

- Previously → Candidate gene approach.
  - Next Generation Sequencing more **available** and **affordable**

New Experimental Approaches

- Previously → Candidate gene approach.
  - Next Generation Sequencing more **available** and **affordable**

  - Will to switch to genome-wide approaches to study **Genomes**, **Epigenomes**, **Transcriptomes**.

Drowed in Next-Gen Data

Université de Perpignan Via Domitia

+ No bioinformatician.
+ No dedicated platform or hardware on the campus.

+ But, a few people with some bioinformatics knowledge…

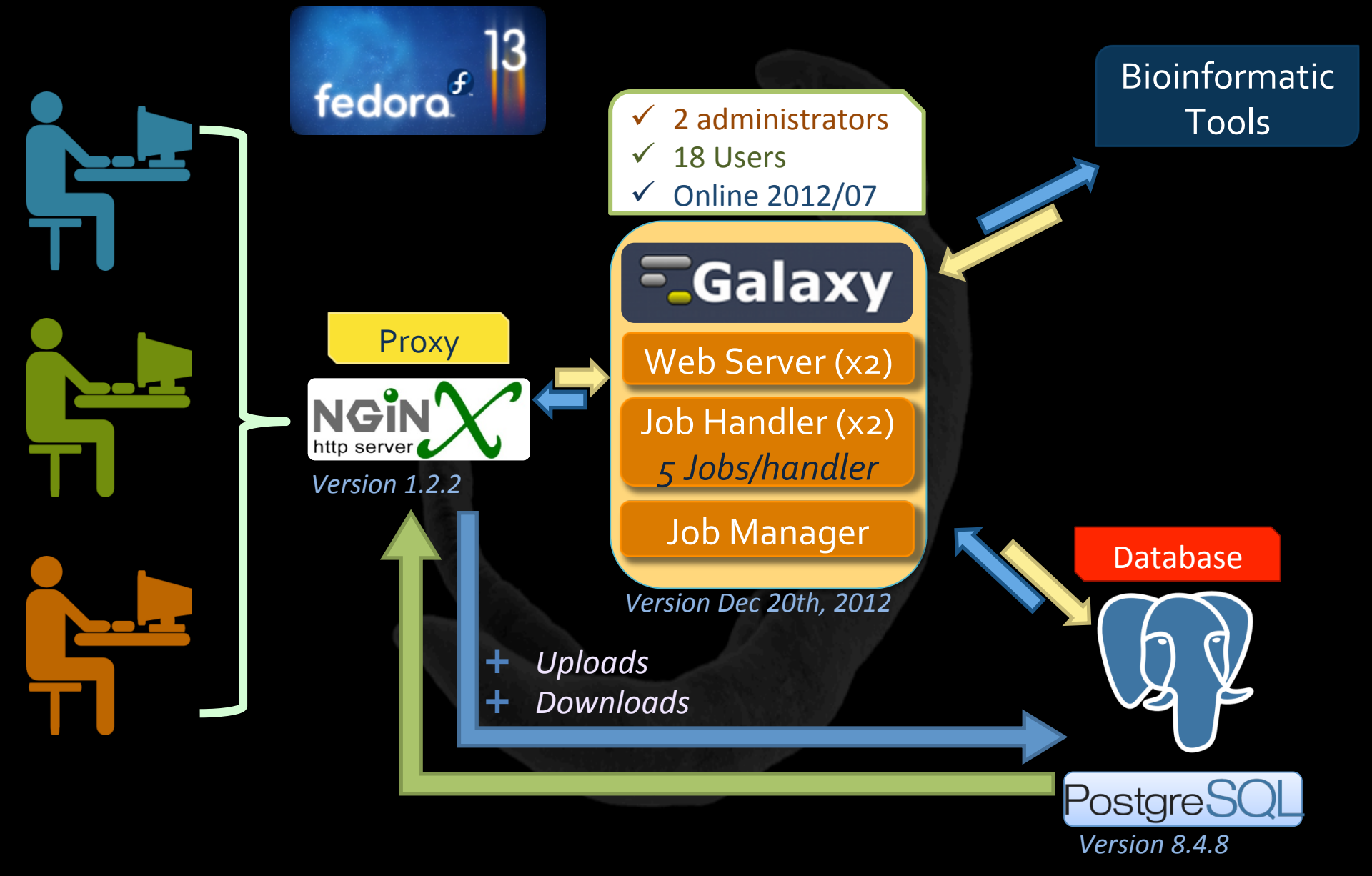
Simple, flexible solution?

Galaxy

But why a local instance?

+ Very different and specific needs.
+ Working on non-model organisms means that most of the tools need to be modified/adapted.
+ Flexibility.

Production mode config.



Hardware

Network

2 x 100 mb/sec

Dell PowerEdge R820
(~5 000 €)

- *16 cores, Xeon 2,4 GHz*
- *96 Gb RAM*
- *1.67 Tb hard drive*

Hardware

Network

2 x 100 mb/sec

Dell PowerEdge R820

(~5 000 €)

+ *16 cores, Xeon 2,4 GHz*
+ *96 Gb RAM*
+ *1.67 Tb hard drive*

Network

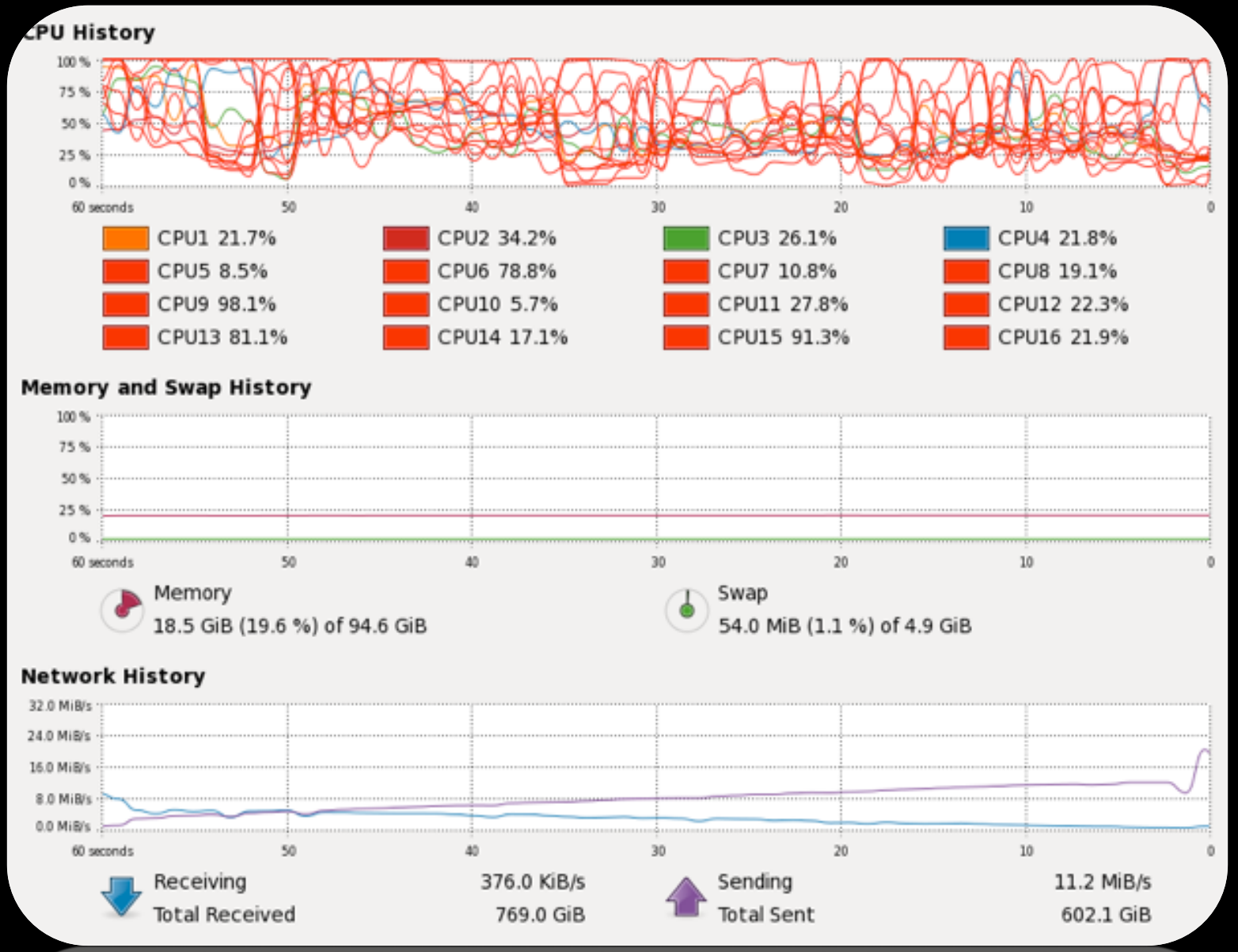
2 x 100 mb/sec

Buffalo TerraStation 5400 WSS

(~2 000 € x 2)

+ *20 Tb hard drive*
+ *Main and backup*

On a typical day



Data management

Data management

Local storage by project manager

Data management

Local storage by project manager

Data libraries by Galaxy Admin

Data Libraries

search dataset name, info, message, dbkey

Advanced Search

Data library name ↓	Data library description
B.glabrata annotation - genome	FASTA and gff files from VectorBase
B.glabrata annotation - results	B.glabrata analysis results
B.glabrata annotation - RNA-Seq I	RNA-Seq data provided by WashU (fastq f/r)
B.glabrata annotation - RNA-Seq II	RNA-Seq data provided by WashU (aligned BAM and assembled fasta)
B.glabrata annotation - RNA-Seq III	sorted BAM files of RNA-Seq suitable for upload to IGV (from G.Oliveira)
B.glabrata - assemblies	assemblies for read dependency
B.glabrata RNA-Seq (clean)	RNA-Seq without adaptors, quality clipped
B.glabrata RNA-Seq (paired)	fastq custom upload N.Dheilly 16/01/13

Data management

Local storage by project manager

Data libraries by Galaxy Admin

SFTP?

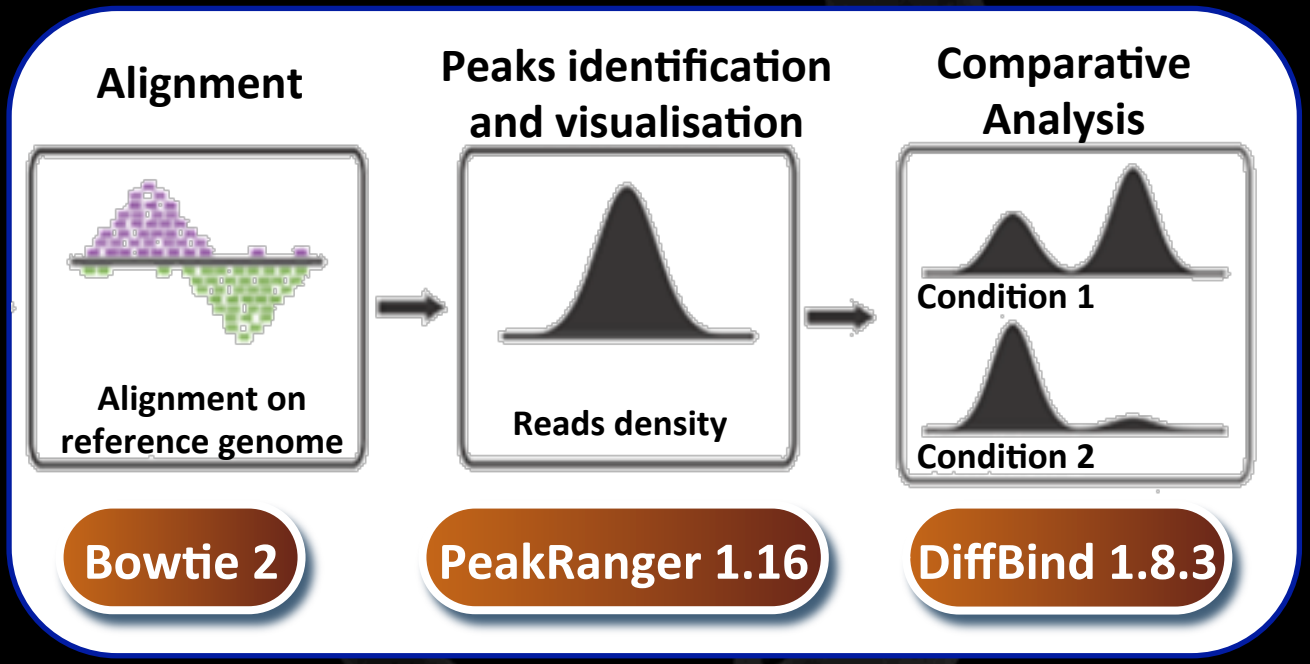
Data Libraries

search dataset name, info, message, dbkey

Advanced Search

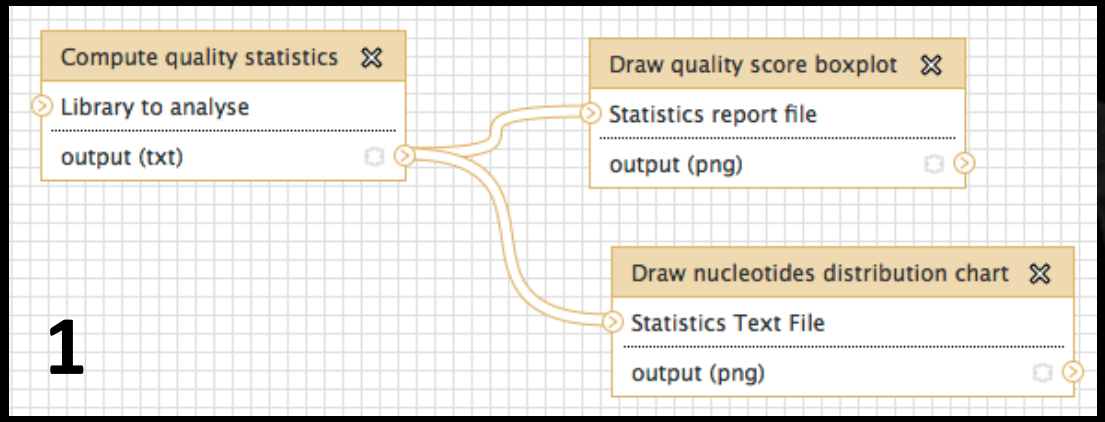
Data library name ↓	Data library description
B.glabrata annotation - genome	FASTA and gff files from VectorBase
B.glabrata annotation - results	B.glabrata analysis results
B.glabrata annotation - RNA-Seq I	RNA-Seq data provided by WashU (fastq f/r)
B.glabrata annotation - RNA-Seq II	RNA-Seq data provided by WashU (aligned BAM and assembled fasta)
B.glabrata annotation - RNA-Seq III	sorted BAM files of RNA-Seq suitable for upload to IGV (from G.Oliveira)
B.glabrata - assemblies	assemblies for read dependency
B.glabrata RNA-Seq (clean)	RNA-Seq without adaptors, quality clipped
B.glabrata RNA-Seq (paired)	fastq custom upload N.Dheilly 16/01/13

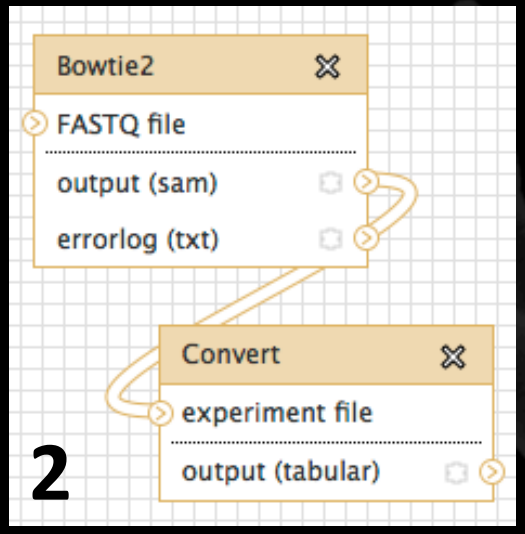
Example of application

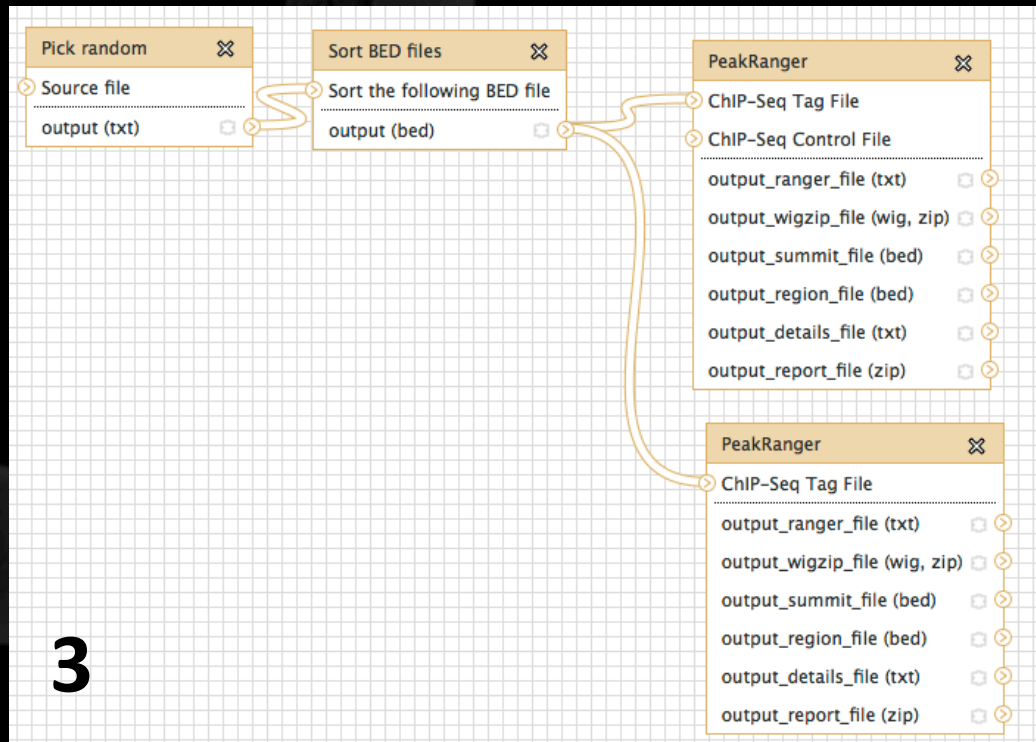


Adapted from Bardet et al. (2012)

Example of application







Example of application

Grooming
Fastq Groomer
+ custom script

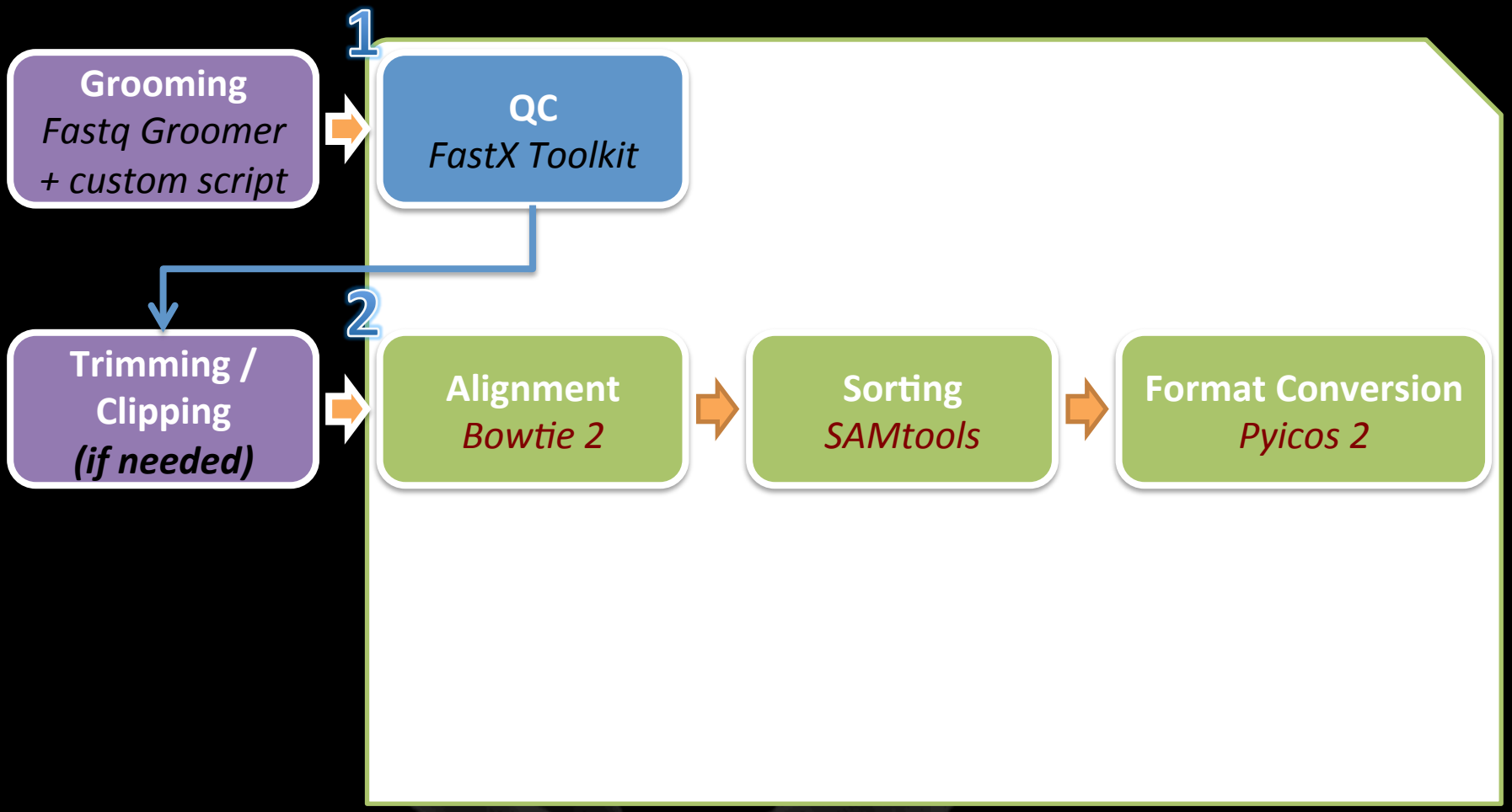
Example of application

Grooming
Fastq Groomer
+ custom script

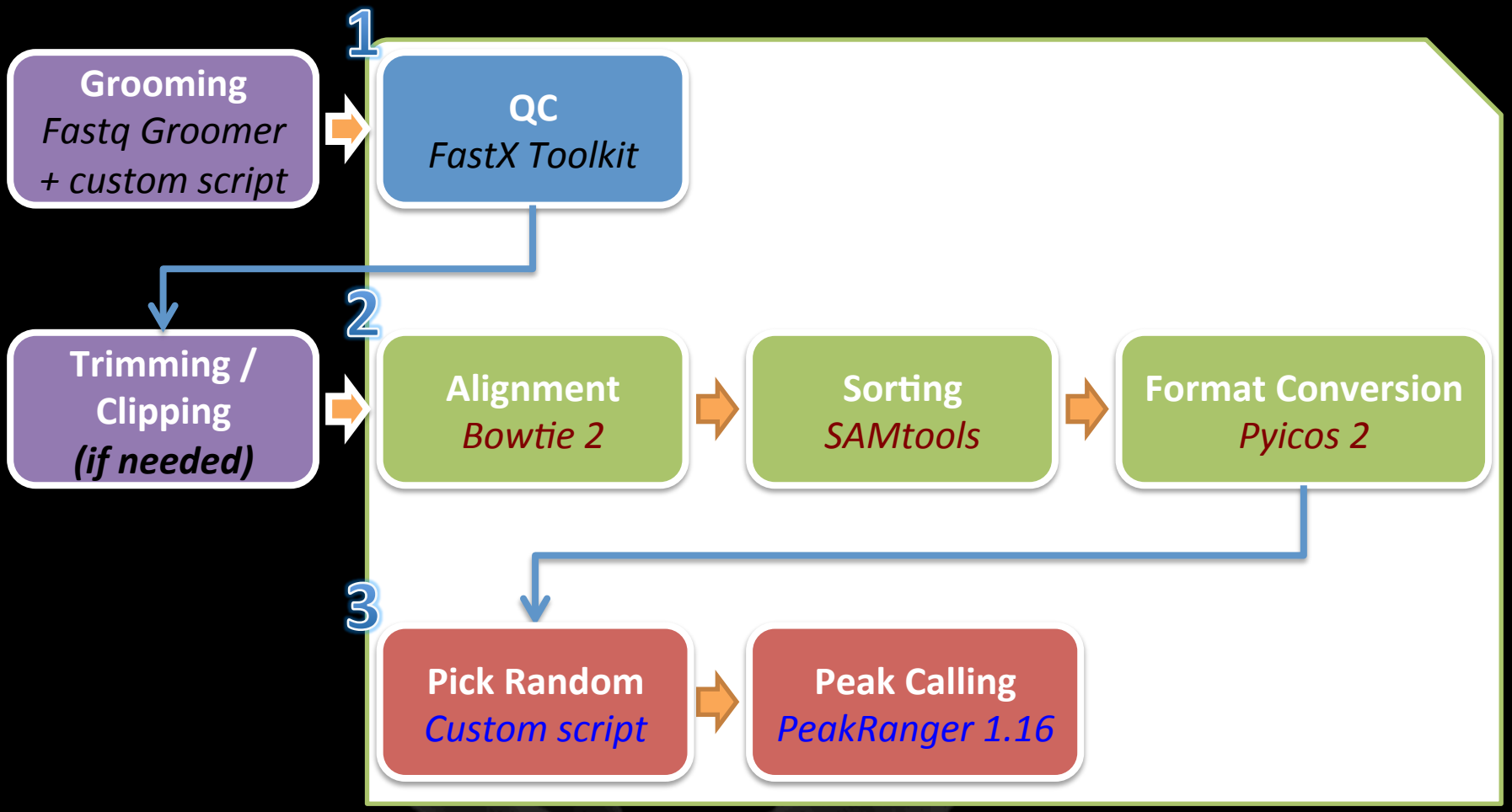
1

QC
FastX Toolkit

Example of application



Example of application



Example of application

Université de Perpignan Via Domitia

Grooming
Fastq Groomer
+ custom script

1

QC
FastX Toolkit

Trimming / Clipping
(if needed)

2

Alignment
Bowtie 2

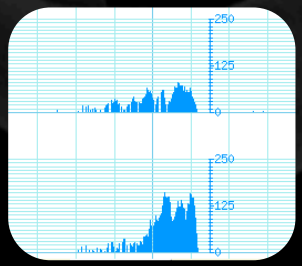
Sorting
SAMtools

Format Conversion
Pyicos 2

3

Pick Random
Custom script

Peak Calling
PeakRanger 1.16



Example of application

Université de Perpignan Via Domitia

Grooming
Fastq Groomer + custom script

1

QC
FastX Toolkit

Trimming / Clipping
(if needed)

2

Alignment
Bowtie 2

Sorting
SAMtools

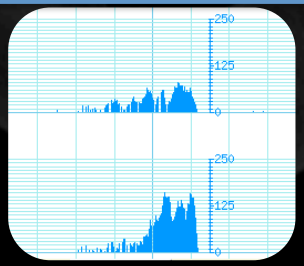
Format Conversion
Pyicos 2

3

Pick Random
Custom script

Peak Calling
PeakRanger 1.16

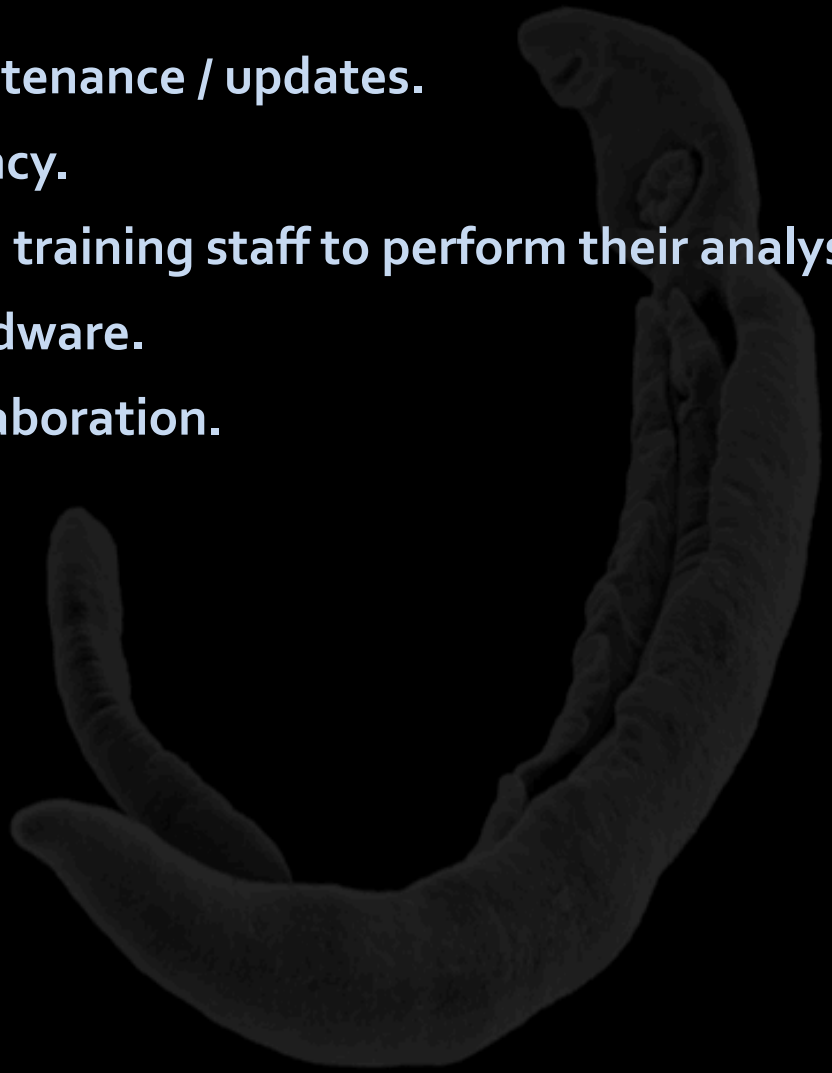
Peaks Comparison
Diffbind v1.8.3



Cons & Pros

Pros

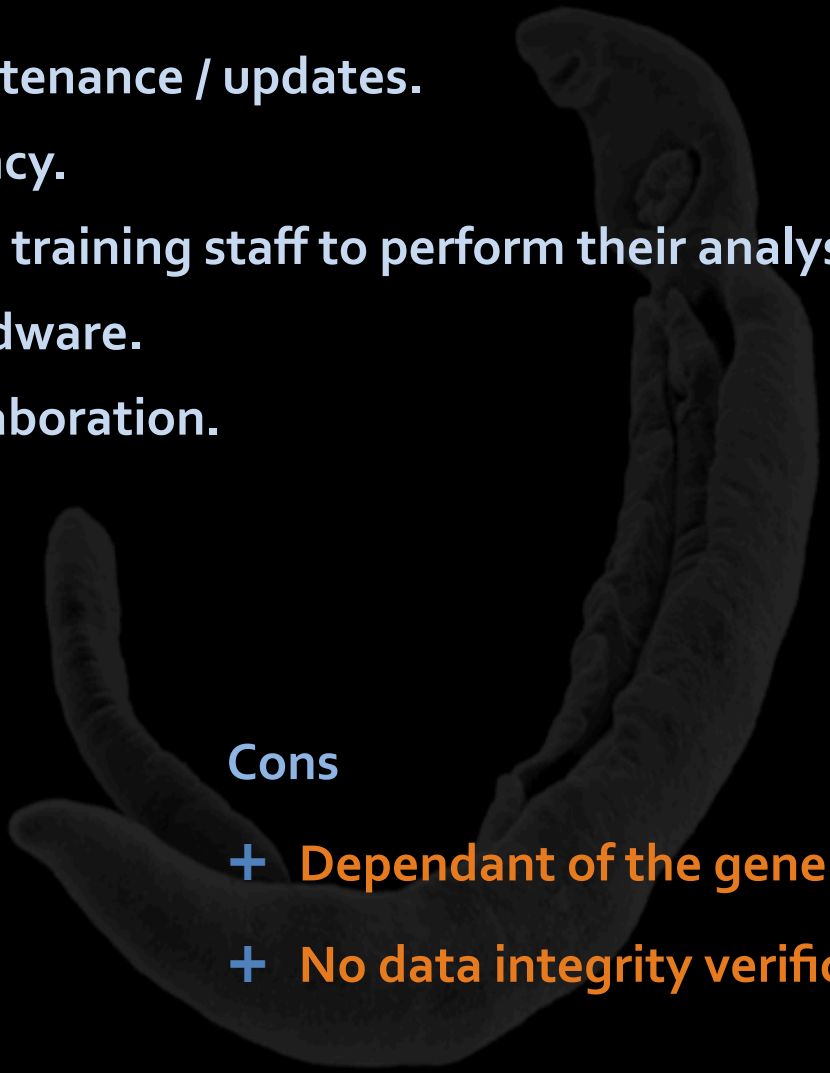
+ Very flexible maintenance / updates.
+ No data redundancy.
+ Less time spent in training staff to perform their analyses.
+ No expensive hardware.
+ International collaboration.



Cons & Pros

Pros

+ Very flexible maintenance / updates.
+ No data redundancy.
+ Less time spent in training staff to perform their analyses.
+ No expensive hardware.
+ International collaboration.



Cons

+ Dependant of the general network speed.
+ No data integrity verification.

Acknowledgments

System Administrator

Rémi Emans

Supervisor

Céline Cosseau

Supervisor

Christoph Grunau

Lab Director

Guillaume Mitta

AGENCE NATIONALE DE LA RECHERCHE

ANR

EPIGEVOL *ÉPIGÉnétique et EVOution*

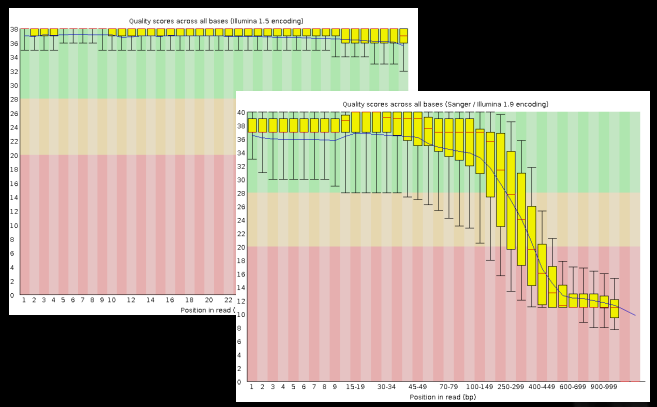
énergie environnement

Ecole Doctorale UPVD

Example of application

```
@HWI-ST609:177:D16HYACXX:3:1308:14623:59114 1:N:0:AGTTCC
TGAGACTAAGGCATATTATAATAATTGCCGCGAACTGTATAGTAAAAGCAC
+
@CCFFFFFGHHGHIJJIJJJJJJJIJIJIGGJIIIIGFIIIIGIJJIIBHI
```

+ *FastQC*



+ *Store clean data*

+ *Format Conversion*

+ *Align Reads*

+ *Format Conversion*

+ *Sorting*

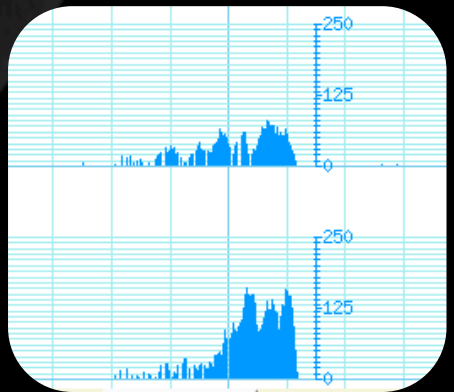
+ *Find Enriched Regions*

Data Libraries

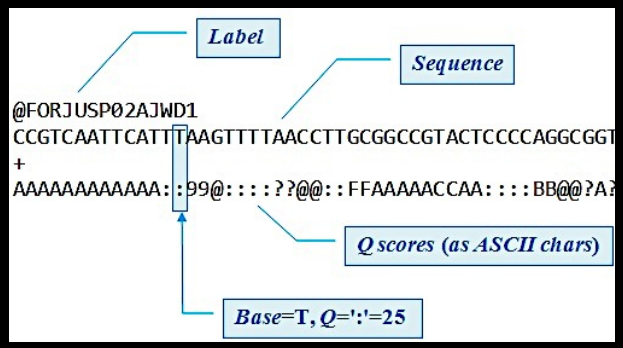
search dataset name, info, message, dbkey

Advanced Search

Data library name ↓	Data library description
B.glabrata annotation - genome	FASTA and gff files from VectorBase
B.glabrata annotation - results	B.glabrata analysis results
B.glabrata annotation - RNA-Seq I	RNA-Seq data provided by WashU (fastq f/r)
B.glabrata annotation - RNA-Seq II	RNA-Seq data provided by WashU (aligned BAM and assembled fasta)
B.glabrata annotation - RNA-Seq III	sorted BAM files of RNA-Seq suitable for upload to IGV (from G.Oliveira)
B.glabrata - assemblies	assemblies for read dependency
B.glabrata RNA-Seq (clean)	RNA-Seq without adaptors, quality clipped
B.glabrata RNA-Seq (paired)	fastq custom upload N.Dheilly 16/01/13



Example of application



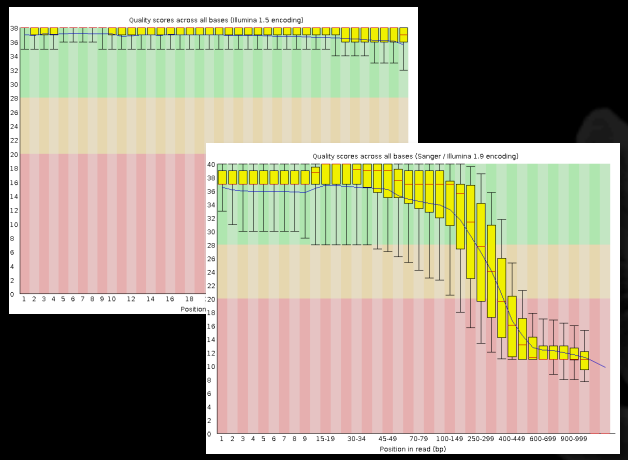
+ Format Conversion

+ Trimming depending on QC

Bow TIE

+ *Align Reads*

+ *FastQC*



+ Format Conversion

+ Sorting

PeakRanger

+ *Find Enriched Regions*