Instituts thématiques

Inserm

Institut national de la santé et de la recherche médicale

Aix*Marseille université

TAGC Inserm U1090

technological advances for genomics and clinics

Développement de workflows pour le NGS

Exécution dans un environnement virtuel

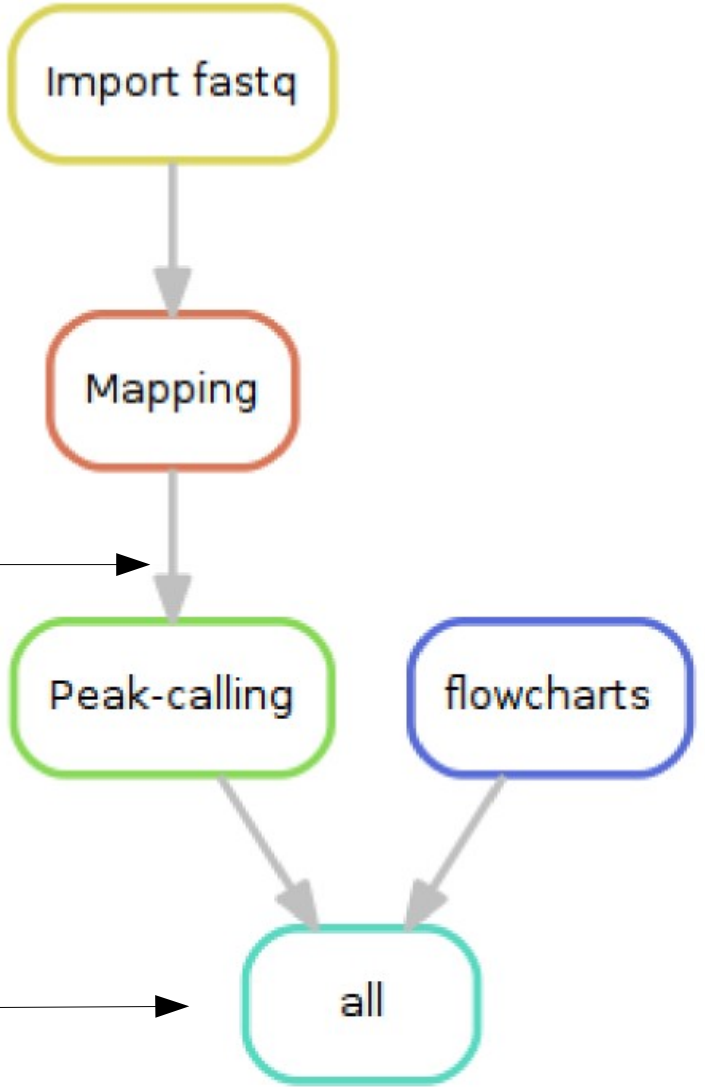
Claire Rioualen

Principes de Snakemake

Règle

Chaînage des opérations
défini par les inputs et outputs

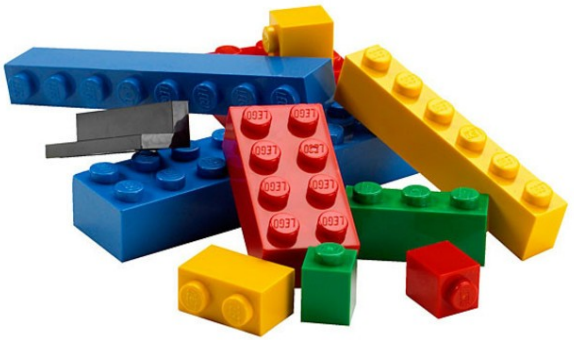
Cible finale

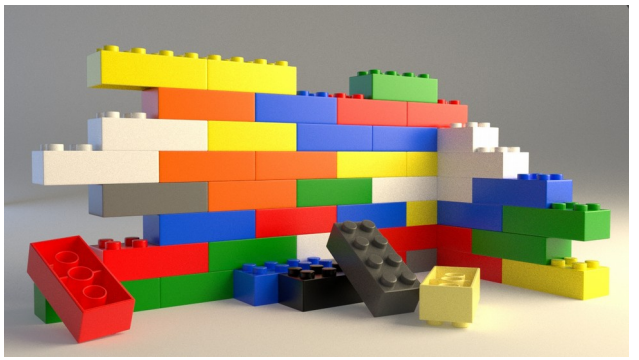


Workflow

Gene-regulation : développement de pipelines flexibles et reproductibles pour le NGS

- Pool de règles recyclables et interchangeables
  - Mapping (bowtie, subread, tophat...)
  - Peak-calling (macs2, homer, swembl...)
  - ...





- Workflows recyclables et customisables
  - ChIP-seq
  - RNA-seq

Choix d'organisation des workflows

- Organisation des métadonnées par projet ou dataset :
  - 1 fichier “workflow.py”

```
## Includes
RULES = os.path.join("gene-regulation/scripts/snakefiles/rules")

include: os.path.join(RULES, "bowtie2_index.rules")
include: os.path.join(RULES, "bowtie2.rules")
include: os.path.join(RULES, "fastqc.rules")
include: os.path.join(RULES, "homer.rules")
include: os.path.join(RULES, "macs2.rules")
include: os.path.join(RULES, "spp.rules")

## Samples
SAMPLE_IDS = read_table(config["metadata"]["samples"])['ID']

## Experimental design
DESIGN = read_table(config["metadata"]["design"])
TREATMENT = DESIGN['treatment']
CONTROL = DESIGN['control']

#============================================================#
#                          Workflow                          #
#============================================================#

## Data import
IMPORT = expand(FASTQ_DIR + "/{samples}/{samples}.fastq", samples=SAMPLE_IDS)

## Quality report
QC = expand(FASTQ_DIR + "/{samples}/{samples}_fastqc/{samples}_fastqc.html", samples=SAMPLE_IDS)

## Mapping
ALIGNER    = config["tools"]["mapping"].split()
INDEX      = expand(GENOME_DIR + "/{aligner}/" + config["genome"]["fasta_file"], aligner=ALIGNER)
MAPPING    = expand(SAMPLE_DIR + "/{samples}/{samples}_{aligner}.bam", samples=SAMPLE_IDS, aligner=ALIGNER)

## Peak-calling
PEAKCALLER = config["tools"]["peakcalling"].split()
PEAKS      = expand(expand(PEAKS_DIR + "/{treat}_vs_{control}/{{peakcaller}}.bed", zip, treat=TREATMENT, control=CONTROL), peakcaller=PEAKCALLER)

rule all:
    input: \
           IMPORT, \
           QC, \
           INDEX, \
           MAPPING, \
           PEAKS
```

Choix d'organisation des workflows

- Organisation des métadonnées par projet ou dataset :
  - 1 fichier “workflow.py”
  - 1 fichier “config.yml”

```yaml
---
author: "Claire Rioualen"
qsub: "-V -m a -d ."
genome:
    organism: "Saccharomyces cerevisiae"
    fasta_file: "sacCer2.fa"
    gff3_file: "sacCer2.gff3"
    gtf_file: "sacCer2.gtf"

metadata:
    samples: "metadata/samples.tab"
    design: "metadata/design.tab"
    configfile: "metadata/config.yml"
    seq_type: "se"

dir:
    fastq: "fastq"
    genome: "genome"
    results: "results"

tools:
    trimming: "sickle"
    mapping: "bowtie2"
    peakcalling: "macs2 homer macs14 spp swembl"

sickle:
    threshold: "20"

bowtie2:
    threads: "5"
    max_mismatches: "1"

macs14:
    mfold: "5,30"
    keep_dup: "auto"
    bandwidth: "300"
```

Choix d'organisation des workflows

- Organisation des métadonnées par projet ou dataset :
  - 1 fichier “workflow.py”
  - 1 fichier “config.yml”
  - 1 fichier “samples.tab”

```
1 ; Sample description file
2 ; Organism: Saccharomyces cerevisiae
3 ; Experiment type: ChIP-seq
4 ;
5 ID	Condition	ENA_experiment	Scan Name
6 GSM521934	input	SRX021359	SRR051930
7 GSM521935	chip	SRX021358	SRR051929
8
9
```

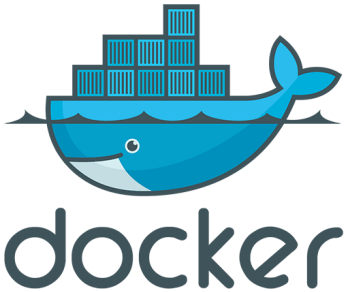
Choix d'organisation des workflows

- Organisation des métadonnées par projet ou dataset :
  - 1 fichier “workflow.py”
  - 1 fichier “config.yml”
  - 1 fichier “samples.tab”
  - 1 fichier “design.tab”

```
1 ; Design for the peak-calling of ChIP-seq analysis
2 ;
3 treatment → control
4 GSM521935 → GSM521934
```

Développement d'environnements virtuels

- Contrôler l'environnement d'exécution d'un workflow
- Fournir une solution “clés en main” avec tous les programmes et dépendances inclus
  - Système d'exploitation (Ubuntu 14.04)
  - Librairies et packages divers
  - Outils d'analyse NGS
  - Dépôt git gene-regulation
- Tutoriels
  - Gene-regulation tutorials



rioualen / **gene-regulation**

PUBLIC | AUTOMATED BUILD

rioualen/gene-regulation ☆

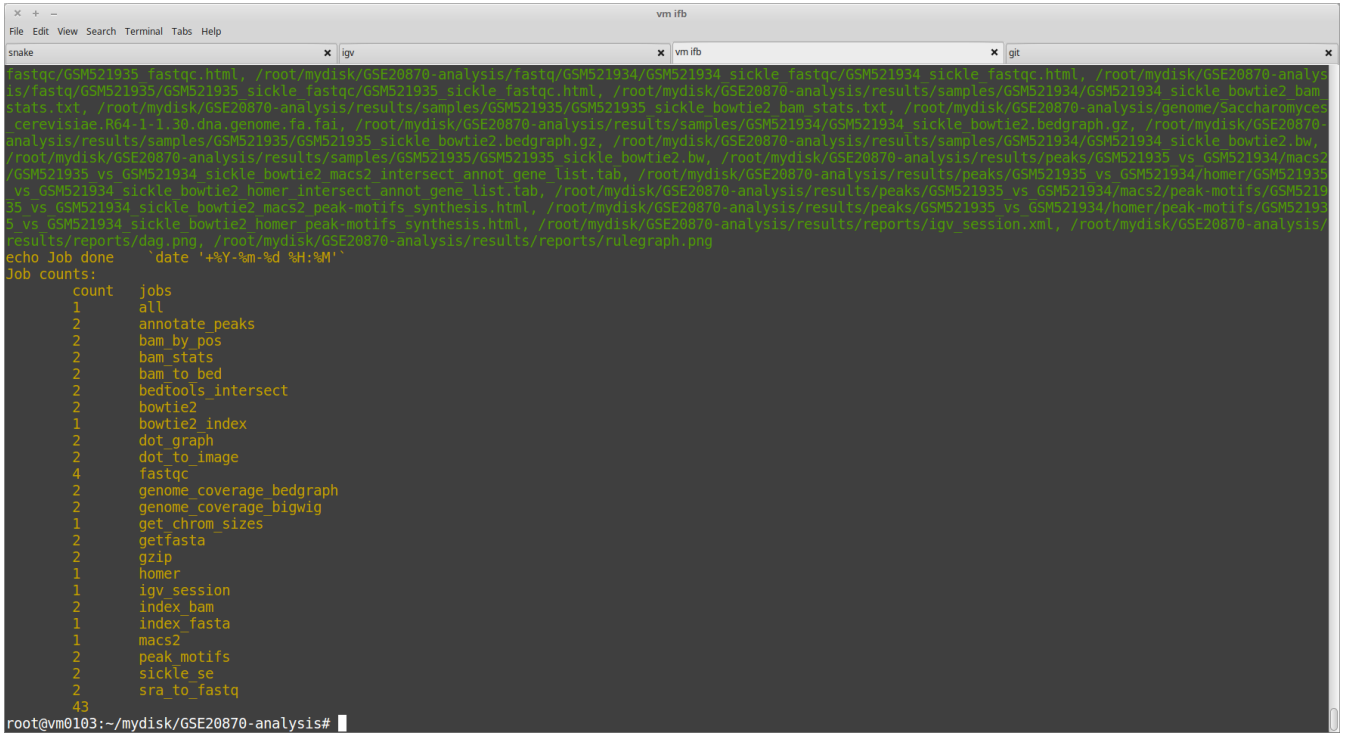
Exécution dans un environnement virtuel

- Création d'une *appliance* Gene-regulation sur le cloud IFB
- Création d'une *instance* de Gene-regulation
- Connexion d'un disque virtuel :

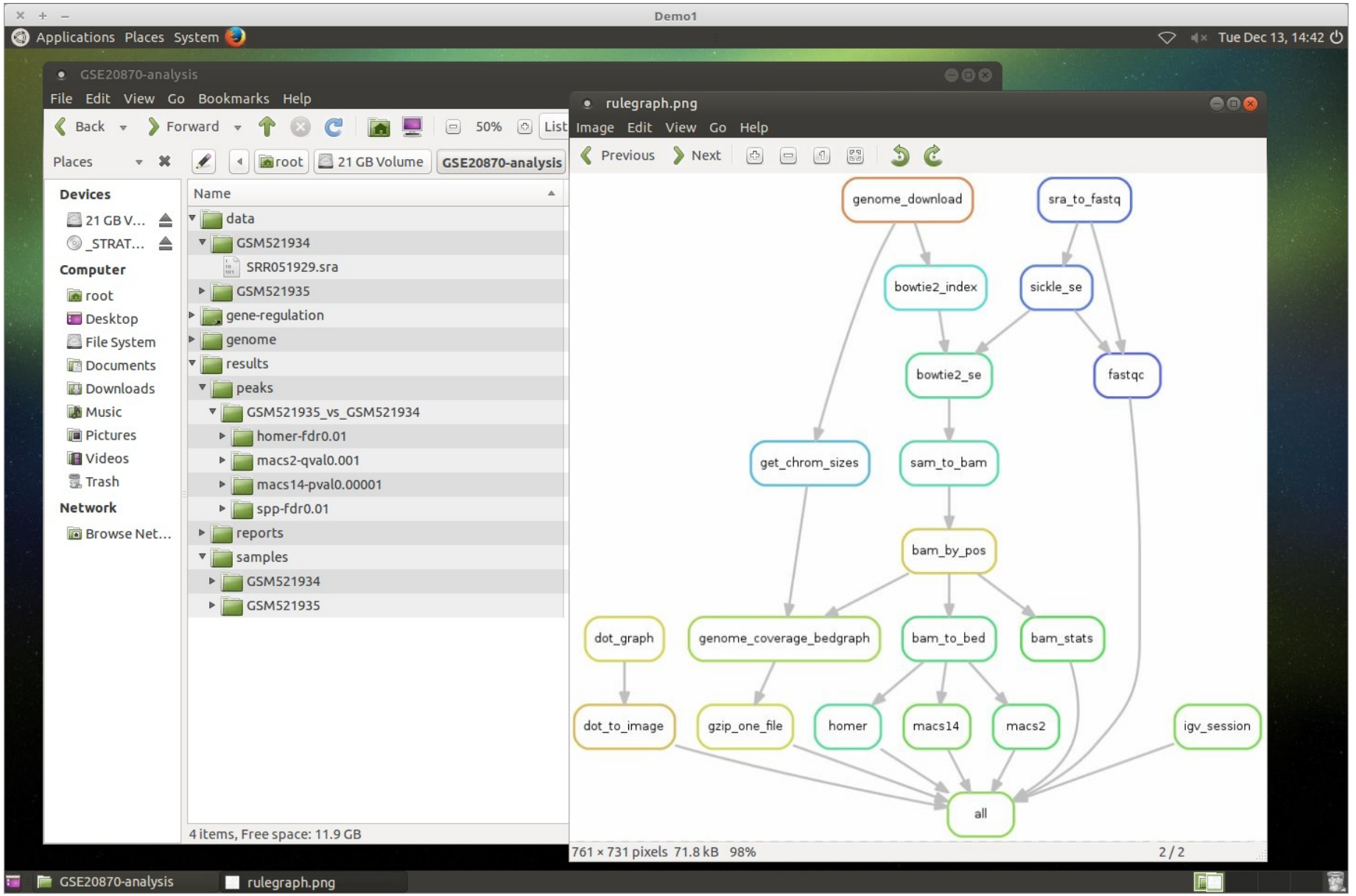
```
root@vm0103:~/mydisk/GSE20870-analysis# ll -R
.:
total 16
drwxr-xr-x 4 root root 4096 Dec  1 16:01 ./
drwxr-xr-x 3 root root 4096 Jun  2 13:32 ../
drwxr-xr-x 4 root root 4096 Jun  2 12:52 data/
lrwxrwxrwx 1 root root   21 Dec  1 15:59 gene-regulation -> /root/gene-regulation/
drwxr-xr-x 2 root root 4096 Dec  1 16:02 genome/

./data:
total 16
drwxr-xr-x 4 root root 4096 Jun  2 12:52 ./
drwxr-xr-x 4 root root 4096 Dec  1 16:01 ../
drwxr-xr-x 2 root root 4096 Jun  2 12:52 GSM521934/
drwxr-xr-x 2 root root 4096 Jun  2 12:52 GSM521935/

./data/GSM521934:
total 81192
drwxr-xr-x 2 root root     4096 Jun  2 12:52 ./
drwxr-xr-x 4 root root     4096 Jun  2 12:52 ../
-rw-r--r-- 1 root root 83043426 Jun  2 12:52 SRR051929.sra

./data/GSM521935:
total 94124
drwxr-xr-x 2 root root     4096 Jun  2 12:52 ./
drwxr-xr-x 4 root root     4096 Jun  2 12:52 ../
-rw-r--r-- 1 root root 96275406 Jun  2 12:52 SRR051930.sra

./genome:
total 29652
drwxr-xr-x 2 root root     4096 Dec  1 16:02 ./
drwxr-xr-x 4 root root     4096 Dec  1 16:01 ../
-rw-r--r-- 1 root root 12360636 Jun  2 13:33 Saccharomyces_cerevisiae.R64-1-1.30.dna.genome.fa
-rw-r--r-- 1 root root  6673011 Jun  2 13:38 Saccharomyces_cerevisiae.R64-1-1.30.gff3
-rw-r--r-- 1 root root 11268860 Jun  2 13:38 Saccharomyces_cerevisiae.R64-1-1.30.gtf
root@vm0103:~/mydisk/GSE20870-analysis#
```

Exécution dans un environnement virtuel

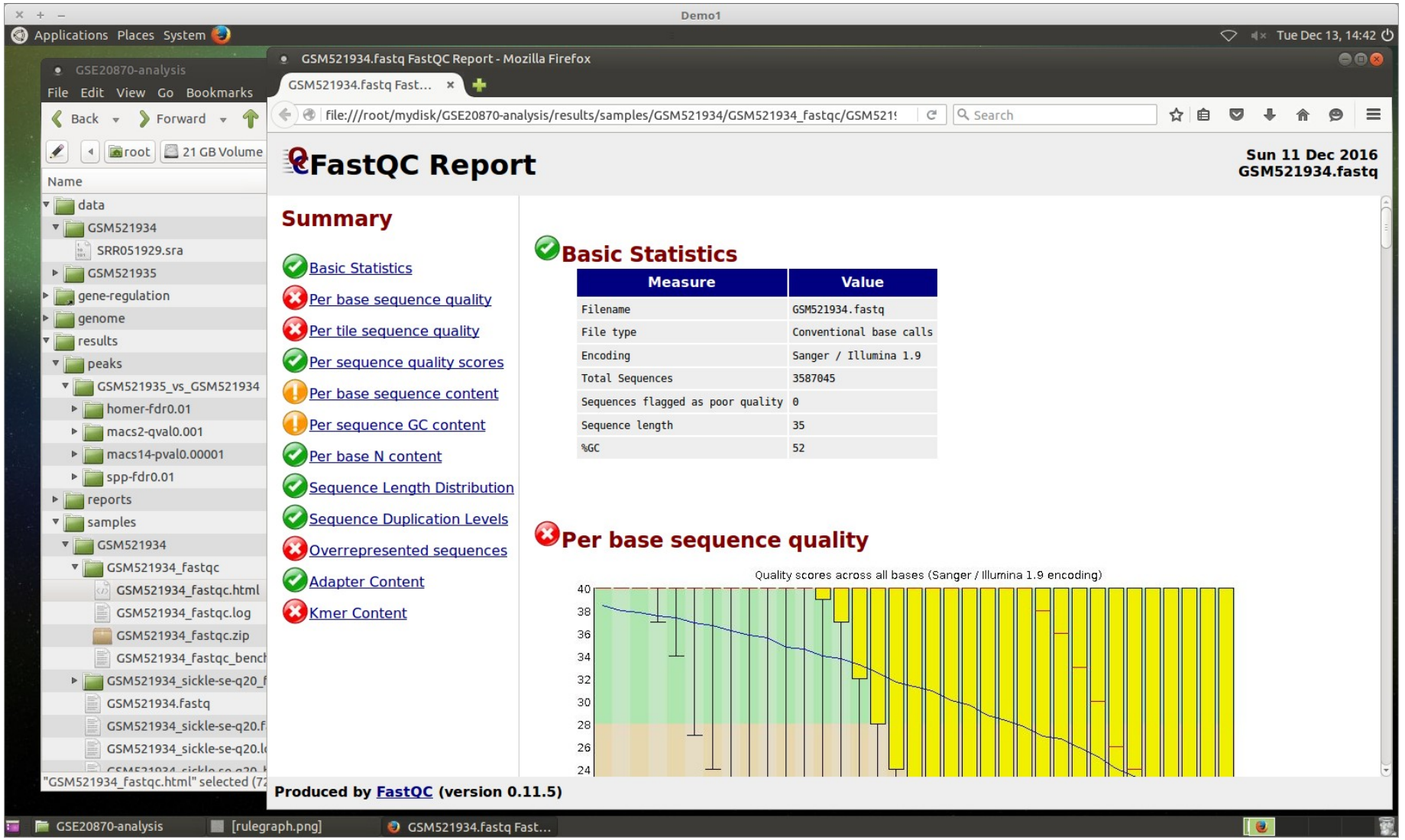
- Création d'une *appliance* sur le cloud IFB
- Création d'une *instance* de Gene-regulation
- Connexion d'un disque virtuel
- Exécution du workflow:



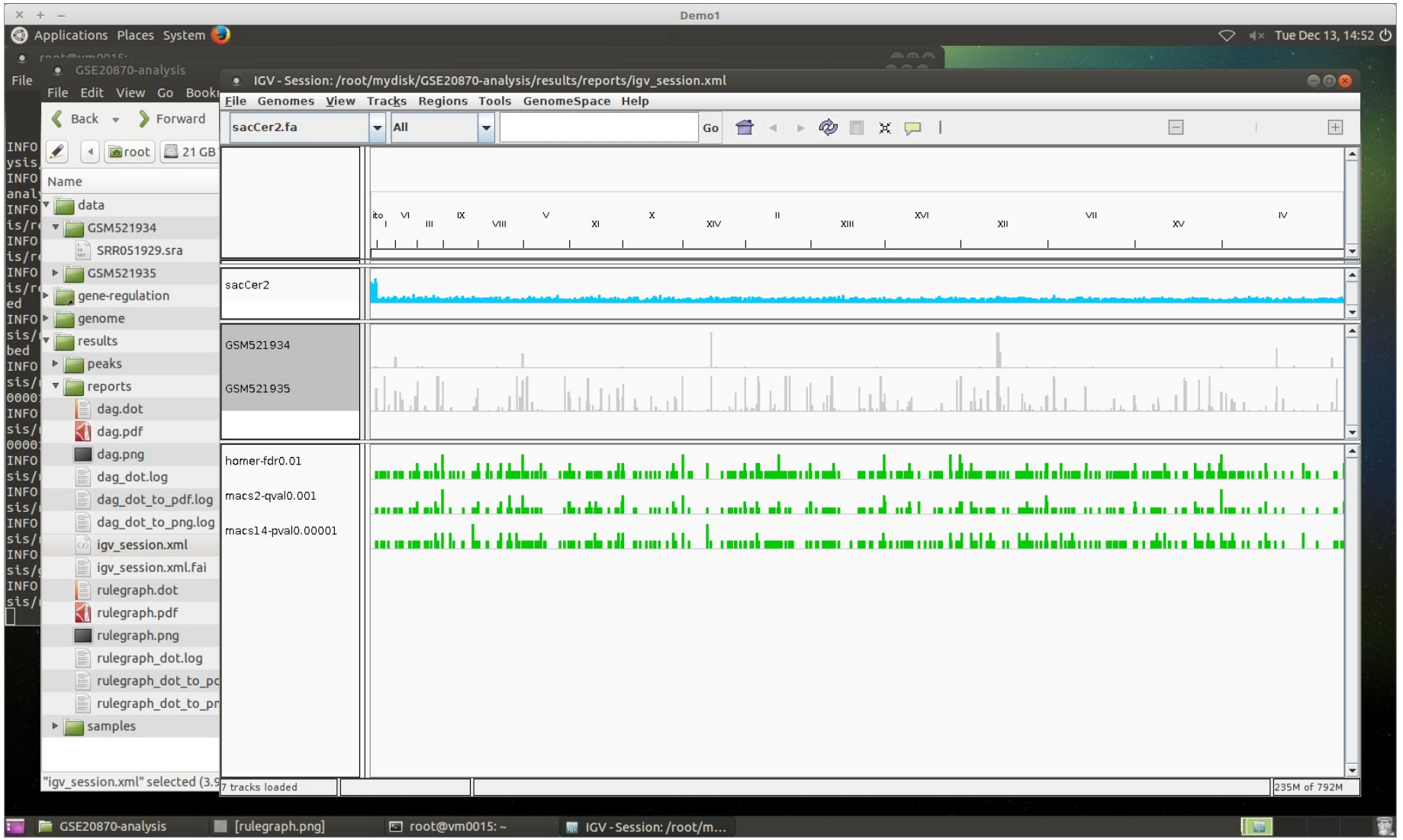
Aperçu des résultats (bureau distant)



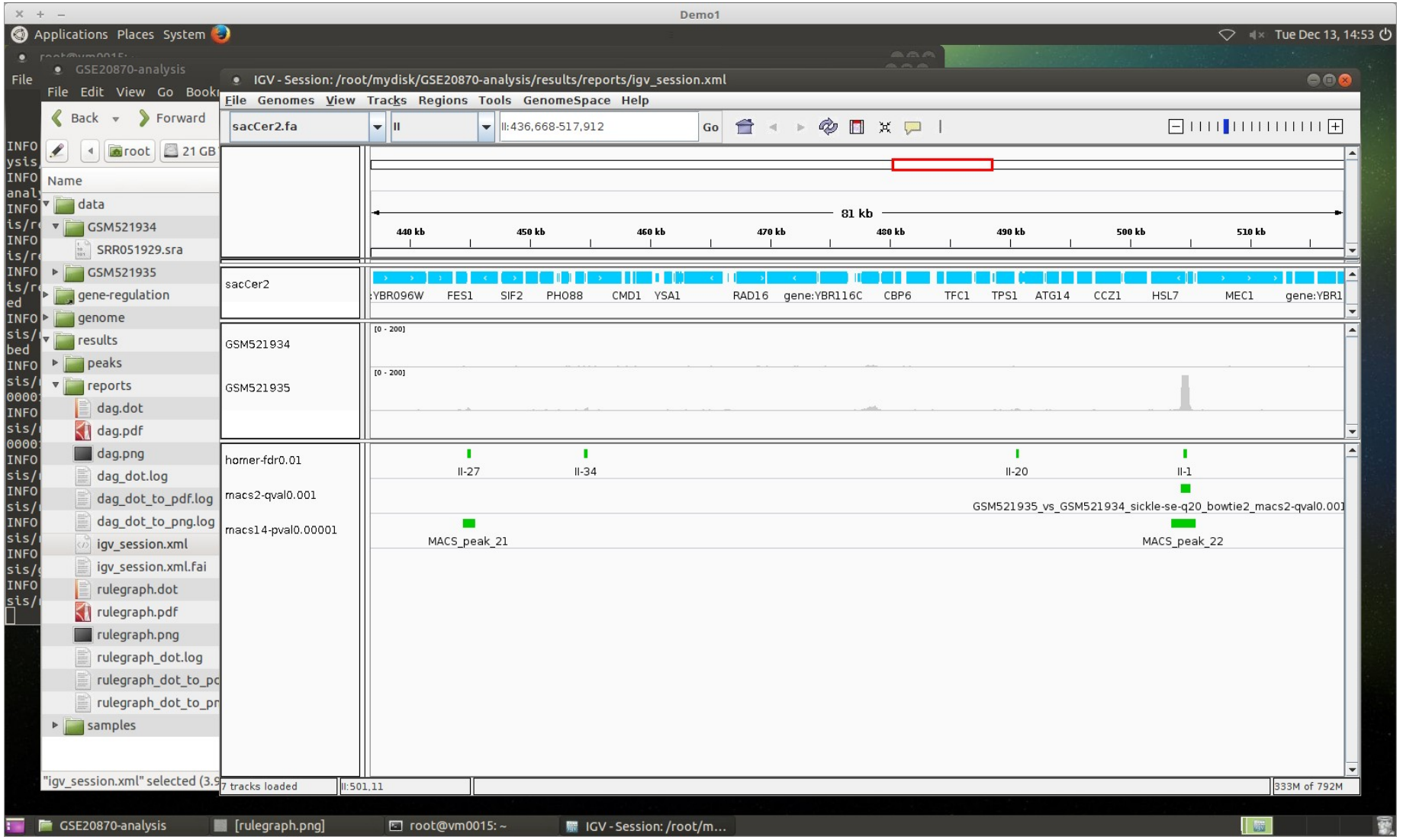
Aperçu des résultats (bureau distant)



Aperçu des résultats (bureau distant)



Aperçu des résultats (bureau distant)



- Jacques van Helden
- Denis Puthier
- Lucie Khamvongsa-Charbonnier
- Jaime Castro-Mondragon

- Julio Collado-Vides
- Alberto Santos-Zavaleta
- Mishael Sánchez-Pérez

- Jocelyn Brayet