High-throuput workflow for RNAseq data treatment linking laboratory data server and remote parallel calculation platform

Marie Garavillon-Tournayre, David D. Lopez, Jean-Stéphane J.-S. Venisse, Marion Landi, Corentin Hochart, Boris B. Fumanal, Aurelie A. Gousset, Philippe Label

High-throuput workflow for RNAseq data treatment linking laboratory data server and remote parallel calculation platform

Marie GARAVILLON-TOURNAYRE, David LOPEZ, Marion LANDI, Corentin HOCHART, Jean-Stéphane VENISSE, Boris FUMANAL, Aurélie GOUSSET-DUPONT, and Philippe LABEL

philippe.label@clermont.inra.fr 33+(0)473 407 922

INRA SCIENCE & IMPACT
UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND
PIAF UMR-547

PROBLEM

1. **RNAseq** softwares using graphical user interfaces are **slow** and data volume **limited**
2. We want to **maximize** the use of **calculation** resources and speed-up **results**

SOLUTION

local server
32 cpu - 128Gb
PIAFdb

genotoul bioinfo
remote cluster
4856 cpu - 34.6Tb

Tasks automatization through scripted workflow

Illumina → fastq → storage

1. **sequences** stored and secured

Perl + Python + Bash → scripts

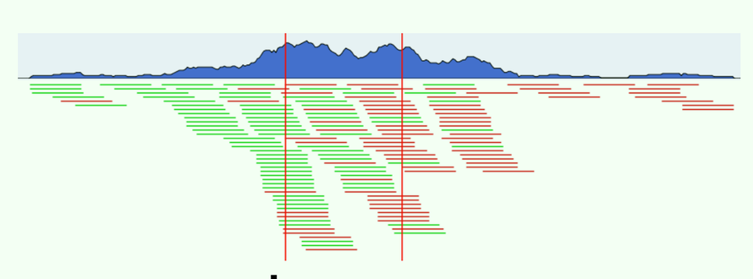
2. **scripts** automatically prepared locally

fastq+scripts → network → storage

3. **transfert** of sequences and scripts

quality, bases, length, poly-A, repeats, contamination, mapping → calculations

4. calculations massively **parallelized**

cluster →  bam → PostgreSQL → database

5. **mapping** results download

DEseq → calculations

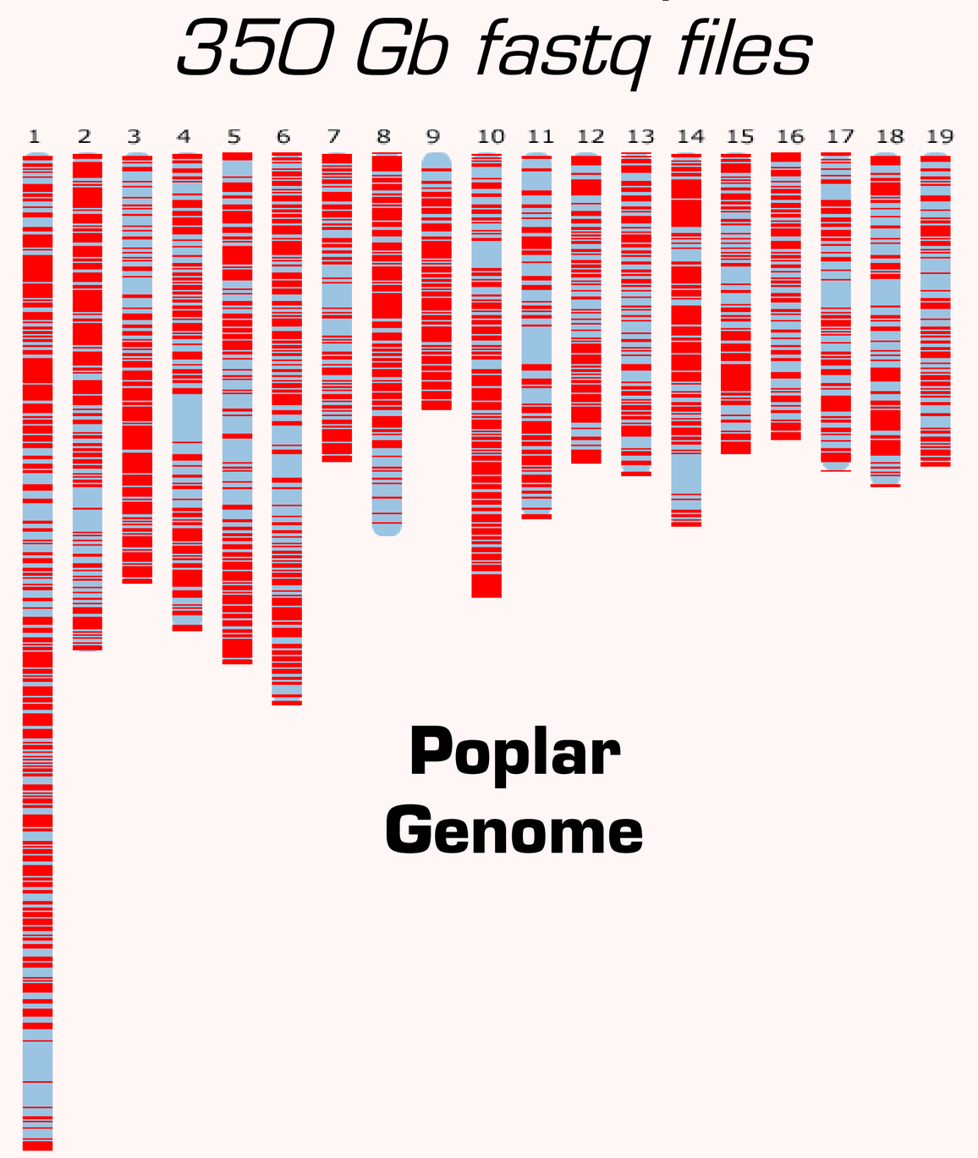
6. **biostatistics** performed locally

EXAMPLE

Black poplar
Populus nigra

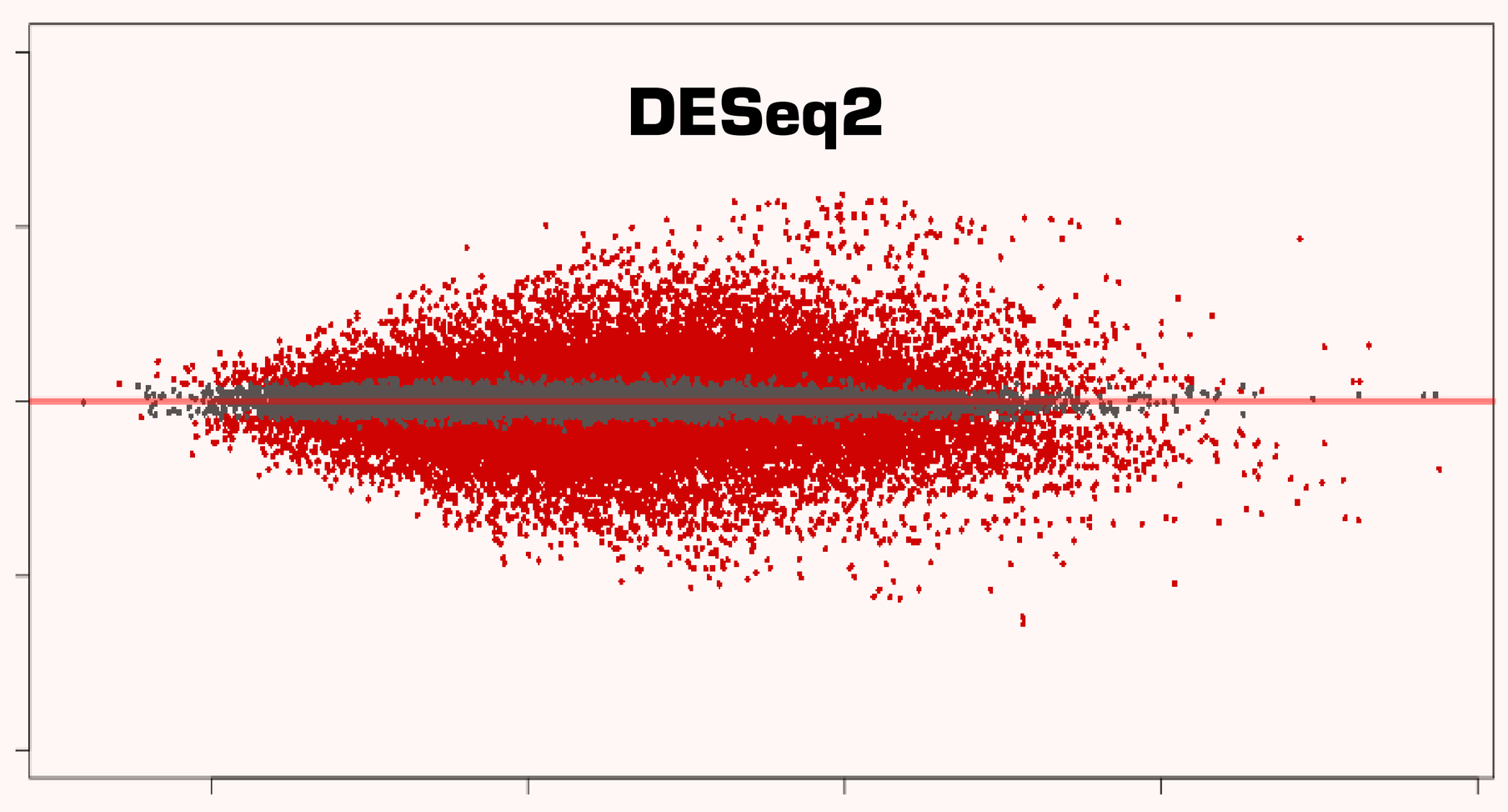
INPUT

1,4 Gseq
paired-end 2 x 100 bp
48 samples, 6 genotypes,
2 conditions, 4 repetitions
350 Gb fastq files



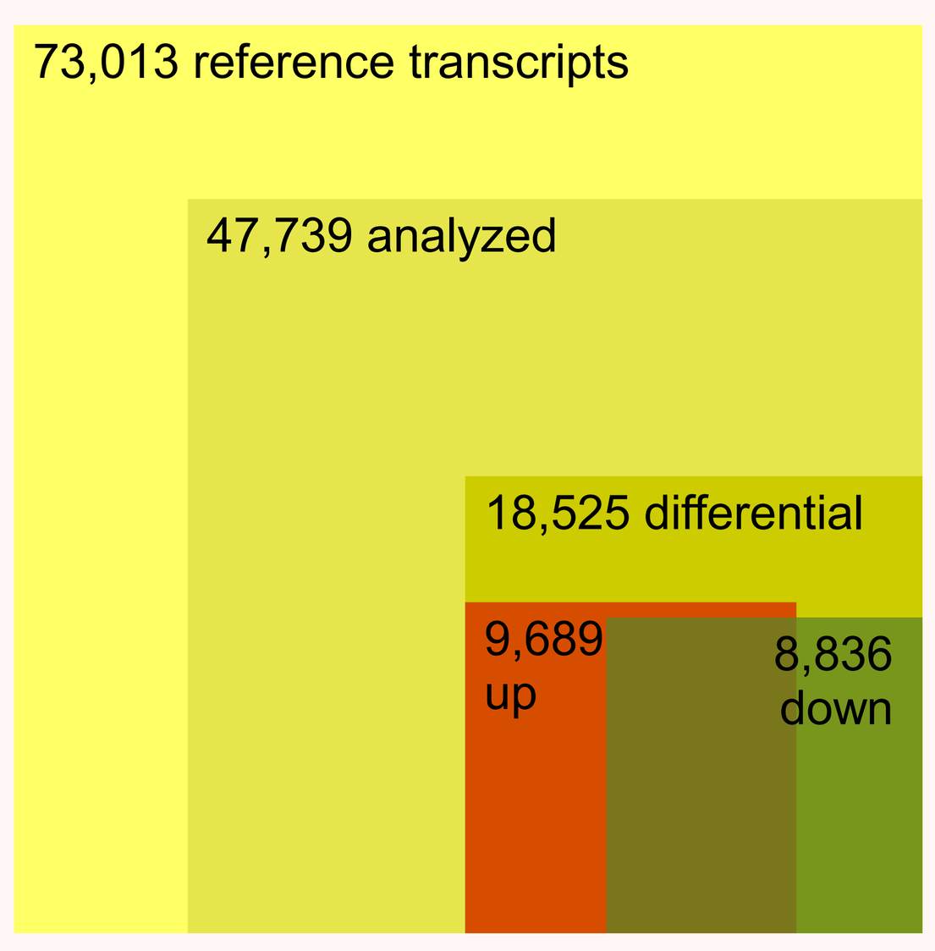
OUTPUT

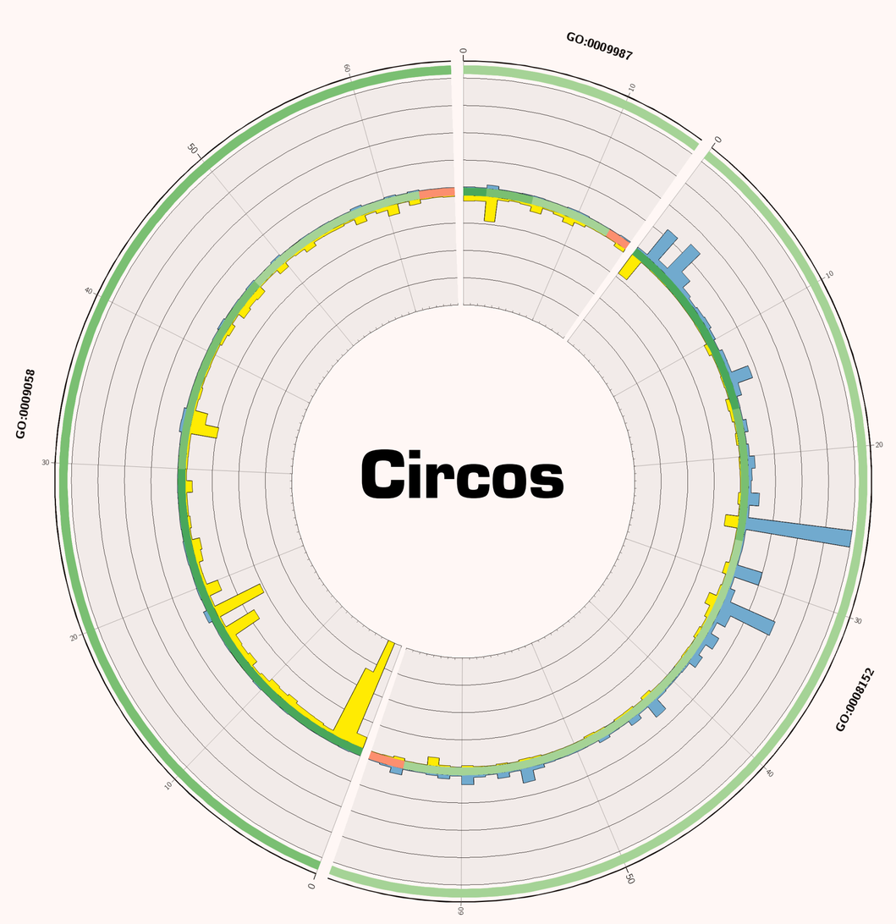
7% sequences filtered out
83% mapped on reference
66 Gb bam files
129 Mb count files



PROCESS

404 batched files
9,159 cpu used
10h fastq transfer → 140 Mseq.h^{-1}
3h calculations → 470 Mseq.h^{-1}





ATGCGCGTAGCATCGATCGGTCGACTAGCTAGCAATCGCATCGTTCGCTA 0101011101010101

JOBIM 2015, 6-9 july