



High-throughput 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

► To cite this version:

Marie Garavillon-Tournayre, David D. Lopez, Jean-Stéphane J.-S. Venisse, Marion Landi, Corentin Hochart, et al.. High-throughput workflow for RNAseq data treatment linking laboratory data server and remote parallel calculation platform. JOBIM 2015, Jul 2015, Clermont-Ferrand, France. , pp.1, 2015. hal-02796330

HAL Id: hal-02796330

<https://hal.inrae.fr/hal-02796330>

Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

High-throughput 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

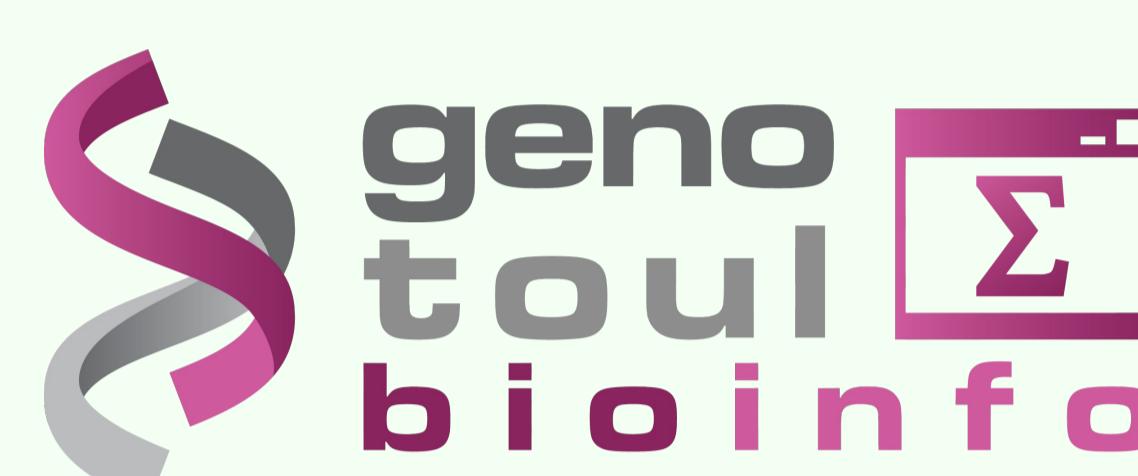


33+(0)473 407 922

PIAF UMB-547

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

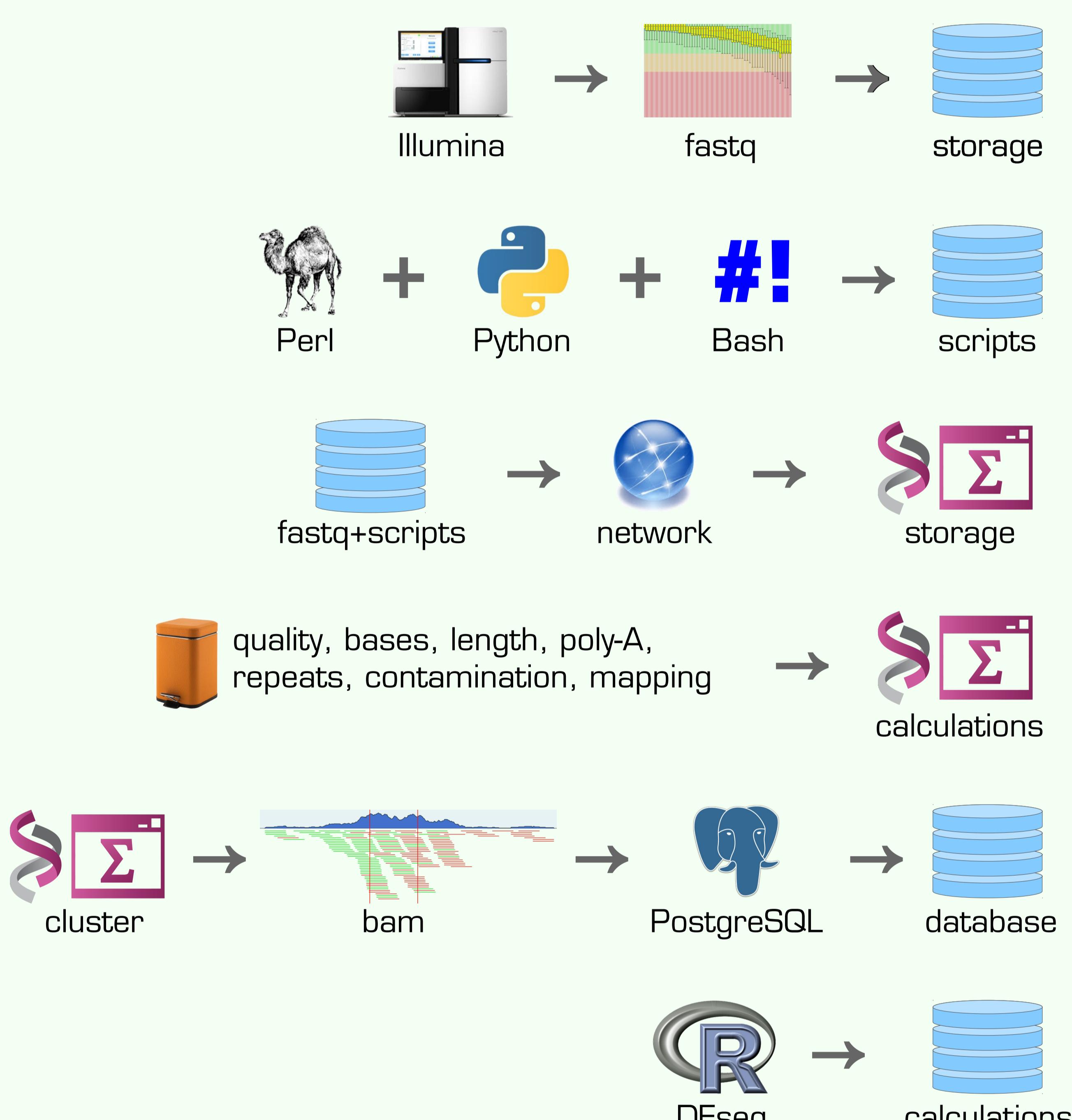
A 3D white stick figure stands in the foreground, holding a large, bright green checkmark with both hands. The figure is positioned in front of a light blue rectangular background. Diagonally across the background, the word "SOLUTION" is written in large, bold, green capital letters, partially obscured by dashed green lines.



remote cluster

4856 cpu - 34.6Tb

Tasks automatization through scripted workflow



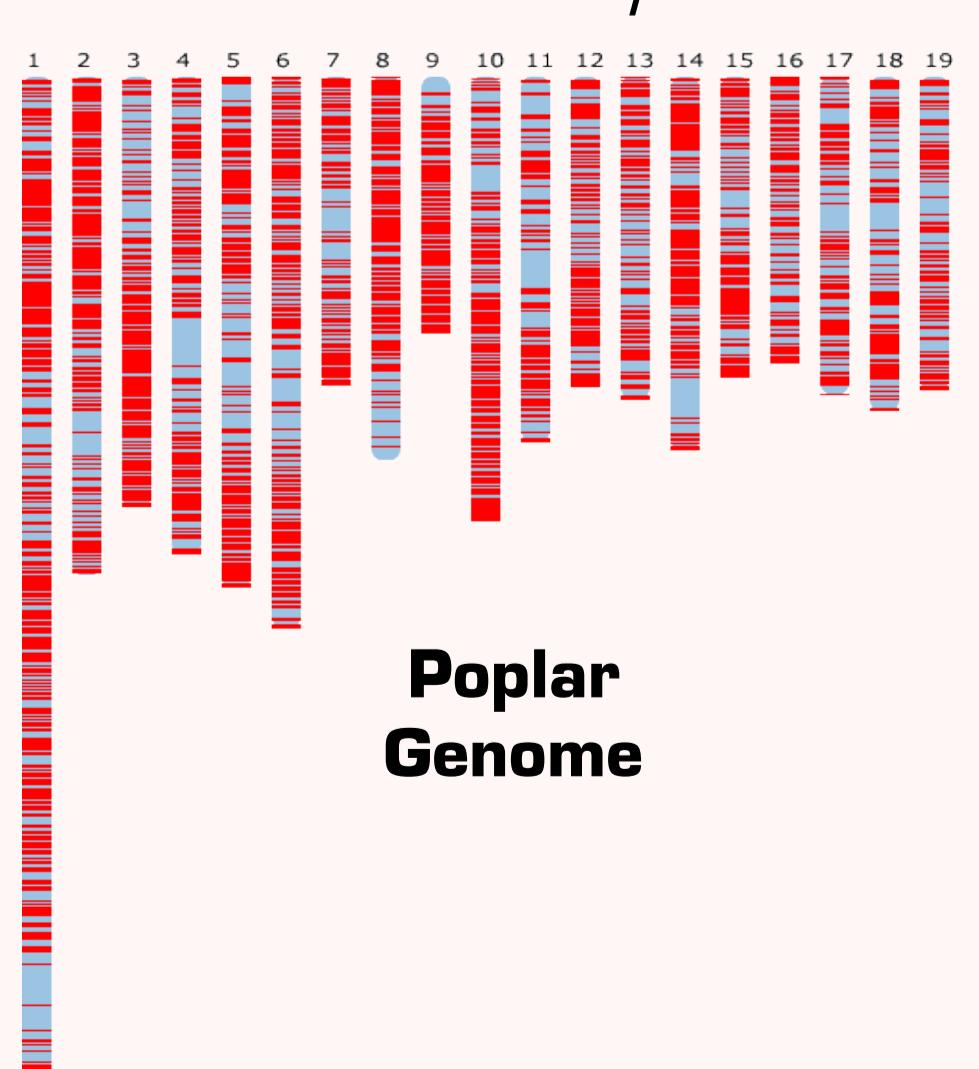
1. **sequences stored and secured**
 2. **scripts automatically prepared locally**
 3. **transfert of sequences and scripts**
 4. **calculations massively parallelized**
 5. **mapping results download**
 6. **biostatistics performed locally**

A large, mature black poplar tree (Populus nigra) stands prominently on a stack of three light-colored wooden boards. The tree has a thick trunk and a dense canopy of green leaves. A large, red, rectangular stamp with the word "EXAMPLE" in capital letters is positioned at the top left of the image, angled diagonally.



INPUT

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



Black poplar

Populus nigra



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**

