Development of a Snakemake “framework/templates” to study the whole-genome transcriptional profiling from a blood-flesh trait (bf) and non blood-flesh trait (non-bf) cultivars in Prunus persica

Laure Heurtevin, Thierry Pascal, Bénédicte Quilot-Turion, Martin-Magniette Marie-Laure, Jacques Lagnel

Development of a Snakemake "framework/templates" to study the whole-genome transcriptional profiling from a blood-flesh trait (bf) and non blood-flesh trait (non-bf) cultivars in *Prunus persica*

Laure Heurtevin[1], Thierry Pascal[1], Bénédicte Quilot-Turion[1], Marie Laure Martin Magniette[2] and Jacques Lagnel [1]

1-INRAE, GAFL, 84140, Avignon, France, 2-INRAE, IPS2, 91190, Gif-sur-Yvette, France

Introduction

Little is known about the mechanisms controlling anthocyanin biosynthesis in flesh of fruit. We explored the genetic pathways related to the elaboration of the blood-flesh trait in peach (*Prunus persica*). For this purpose, a comparative RNAseq study was carried out on flesh from a blood-flesh cultivar and a non blood-flesh cultivar at 4 fruit development stages, from 60 days after blooming up to fruit maturity. 40 libraries including biological replicates were sequenced by Illumina platform (Get-PlaGe) which generated 145Gbp (2.5Greads PE 150pb reads). The RNAseq pipeline was developed using Snakemake [1] and Singularity in order to ensure reproducibility and flexibility in the analysis, traceability of the samples, pipeline ease of use as well as facilitate the portability and the scalability to large data sets.
Here, we proposed a Snakemake "framework" based on a set of interoperable Snakemake rules as well as a set of templates (config, Slurm and samples sheet). This Snakemake framework/templates and Singularity recipes/images will be available on a public forge based on GitLab source code management software (https://forgemia.inra.fr/gafl).
Statistical analyses were performed by DiCoExpress (R workflow ML Martin-Magniette).

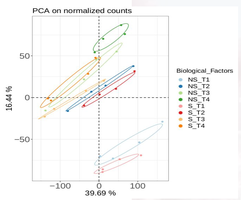
Biological Data and sequencing

Cultivars : 2 cultivars : blood-flesh trait and non blood-flesh trait
Samples : flesh from peach fruits
Reference genome: v2.0.1 (26873 genes)
Biological replicates : 2cultivarsx 4 stages x5 biological replicat
⇒ 5 biological replicates X 8 conditions (40librairies)
⇒ Illumina paired ends 2x150pb sequencing

Sequencing Raw data summary

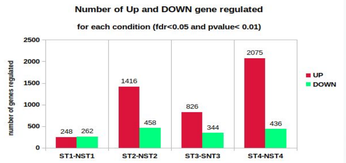
Total librairies	40
Average library reads count (Mreads)	30 +/- 5
Total Size (Gbp)	145
size of reads (bp)	2X150

Results



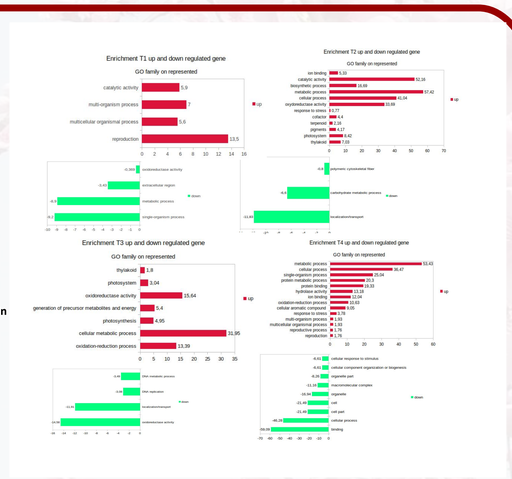
PCA of counts quality control

The PCA of counts shows that the biological replicates are well grouped together and separated between the different conditions. This quality control step allows us to validate our counts data in order to do the differential analysis.



Results of differential gene expression (DEG) profiles per time points

The histograms present the regulated genes level, up and down, between the blood-flesh fruits and non blood-flesh fruits in each time of development. We observe that despite restrictioves variables fdr and pvalue, we have more than 2000 genes differentially expressed in our experimental conditions.



Enrichment results on 4 times

The enrichment analysis was carried out with AgriGO software (http://systemsbiology.cau.edu.cn/agriGOv2/)
We observed at time 2 and 3 an overrepresentation of many genes linked to photosynthesis, pigmentation and oxidoreduction activity in the blood-flesh cultivars.

RNAseq pipeline

samples.csv

SampleName	fq1	fq2
sample1	s1.R1.fq.gz	s1.R2.fq.gz
sample2	s2.R1.fq.gz	s2.R2.fq.gz
….	…	….

config.yaml
-Output path
-Ref path
-parameters
-containers path
- ….

cluster.json
Slurm Ressouces reservation for each rule

RNA-seq SnakeMake workflow

https://forgemia.inra.fr/gafl/pipelines_snakemake/rnaseq

Preproc (fastp)
Reads QC (fastqc/multiqc)
Mapping star 2 pass (star2)
bam QC (multiqc)
Build count matrix

Counts.csv

Target.csv
Annotation.csv
Enrichment.csv

R analysis using DiCoExpress pipeline

DicoExpress v2.0 R 3.6.1

Quality control
Differential expression analysis (DEG)
CoExpression
GO Enrichment

Snakemake and Singularity framework

Snakemake

Snakemake module repository:
➢ **38 rules in 27 modules**
https://forgemia.inra.fr/gafl/snakemake_modules

fastp
multiqc
samtools
Features count
Utils
fastqc
STAR
….

Template files :
- **Snakefile, config.yaml**
- **Samples list in csv**
- **cluster.json**

Singularity

Singularity repository:
https://forgemia.inra.fr/gafl/singularity

Singularity repository:
➢ **80 containers recipes (versioned)**
➢ **gitlab CI/CD for images (.gitlab-ci.yml)**
➢ **Images accessible via OCI registry1**

STAR v2.7.3
fastp v0.20
fastqc v0.18
multiqc v1.7
samtools v1.9
R_base v3.6.1
…..
DicoExpress v2.0 R 3.6.1

Example STAR package:
https://forgemia.inra.fr/gafl/singularity/star

```
singularity pull star_v2.7.3.sif \
oras://registry.forgemia.inra.fr/gafl/singularity
/star/star:latest
```

Conclusion

The proposed Snakemake "framework" and singularity repository facilitate 1) the pipeline construction using interoperable modules, 2) the bio-analyses by non bioinformatician 3) the scalability. The parallelisation is fully automated using Slurm. In order to run the pipeline, the user only needs to provide sample files (csv) and set few parameters (config.yaml). Furthermore, the use of Snakemake workflow manager and Singularity containers increases the bioanalysis reproducibility and facilitates the deployment across HPC platforms. The only requirement for the HPC platform is to provide Singularity (>3.3) and use Slurm as ressource manager.

References
1 Köster J and Rahmann S. Snakemake – a scalable bioinformatics workflow engine. Bioinformatics, 28:2520-2522, 2012.
2 Lambert I, Paysant-Le Roux C, Colella S, Martin-Magniette ML- DiCoExpress: a workspace to process multifactorial RNAseq experiments from quality controls to co-expression analysis through differential analysis based on contrasts inside GLM models. Plantmethods,10.1186/s13007-020-00611-7 https://forgemia.inra.fr/GNet/dicoexpress

Acknowledgements: We are grateful to the Genotoul Bioinformatics platform of Toulouse for providing help and computing and storage resources, to Sylvain Santoni for the librairies constructions and the GeT-PlaGe platform for the RNAseq experiments.