Differential Isoform Analysis orchestrated by Snakemake

Thibault Dayris

December 2016

Introduction

- Definitions
- Pipeline overview

Challenges encountered

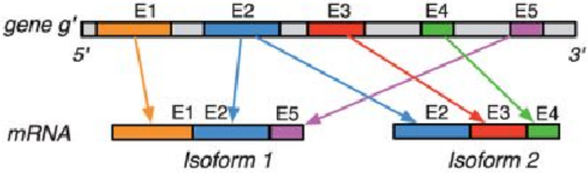
- Expanded inputs
- File closure delay on computing clusters
- External module importation

Future improvements

- Rules versionning
- Docker-isation

A gene isoform

Two **genes isoforms** are mRNA produced from the same gene locus, they have different transcription start sites (**TSS**), different coding DNA sequences (**CDS**) and/or different untranslated regions (**UTR**). This may - or not - change their function or efficience.

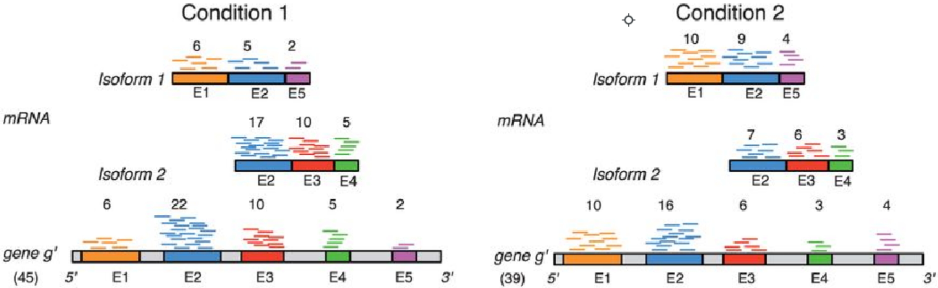


Hypothetical isoforms on a gene[1]

[1] Leng, Ning, et al. "EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments." Bioinformatics 29.8 (2013): 1035-1043.

Differential expression of isoforms

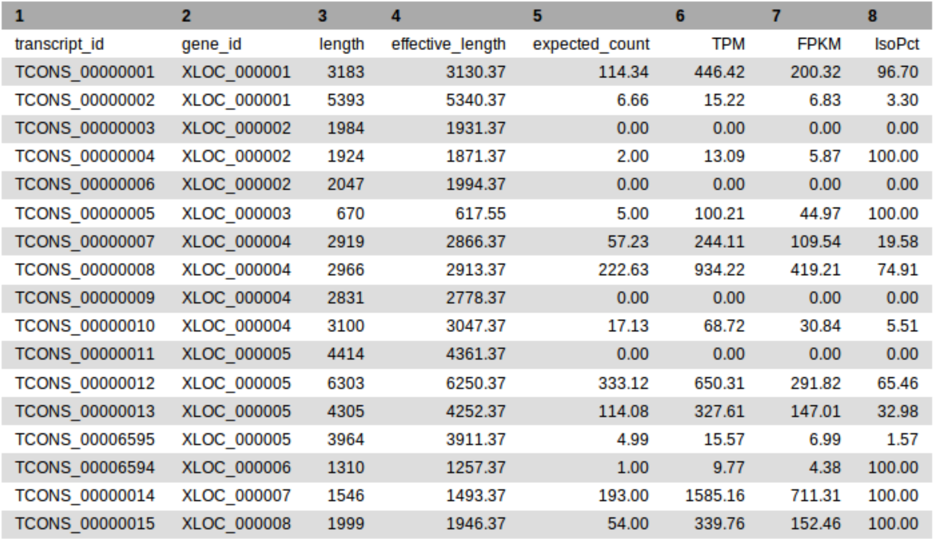
The differential analysis is a technic that aims to evaluate the factors that are different or unique among possible alternatives - *i.e.* to know if some gene is not processed identically among a given set of biological condition.



Hypothetical read alignment on isoforms between two conditions[2]

[2] Leng, Ning, et al. "EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments." Bioinformatics 29.8 (2013): 1035-1043.

Beyond the table and its numbers (1/2)



1	2	3	4	5	6	7	8
transcript_id	gene_id	length	effective_length	expected_count	TPM	FPKM	IsoPct
TCONS_00000001	XLOC_000001	3183	3130.37	114.34	446.42	200.32	96.70
TCONS_00000002	XLOC_000001	5393	5340.37	6.66	15.22	6.83	3.30
TCONS_00000003	XLOC_000002	1984	1931.37	0.00	0.00	0.00	0.00
TCONS_00000004	XLOC_000002	1924	1871.37	2.00	13.09	5.87	100.00
TCONS_00000006	XLOC_000002	2047	1994.37	0.00	0.00	0.00	0.00
TCONS_00000005	XLOC_000003	670	617.55	5.00	100.21	44.97	100.00
TCONS_00000007	XLOC_000004	2919	2866.37	57.23	244.11	109.54	19.58
TCONS_00000008	XLOC_000004	2966	2913.37	222.63	934.22	419.21	74.91
TCONS_00000009	XLOC_000004	2831	2778.37	0.00	0.00	0.00	0.00
TCONS_00000010	XLOC_000004	3100	3047.37	17.13	68.72	30.84	5.51
TCONS_00000011	XLOC_000005	4414	4361.37	0.00	0.00	0.00	0.00
TCONS_00000012	XLOC_000005	6303	6250.37	333.12	650.31	291.82	65.46
TCONS_00000013	XLOC_000005	4305	4252.37	114.08	327.61	147.01	32.98
TCONS_00006595	XLOC_000005	3964	3911.37	4.99	15.57	6.99	1.57
TCONS_00006594	XLOC_000006	1310	1257.37	1.00	9.77	4.38	100.00
TCONS_00000014	XLOC_000007	1546	1493.37	193.00	1585.16	711.31	100.00
TCONS_00000015	XLOC_000008	1999	1946.37	54.00	339.76	152.46	100.00

Beyond the table and its numbers (2/2)

Minimal analysis

- Isoform quantification → Salmon[a]
- Isoform differential analysis → Sleuth[bc]

[a] https://salmon.readthedocs.io/en/latest/index.html
[b] https://rawgit.com/pachterlab/sleuth/master/inst/doc/intro.html
[c] https://github.com/COMBINE-lab/wasabi

We can do more !

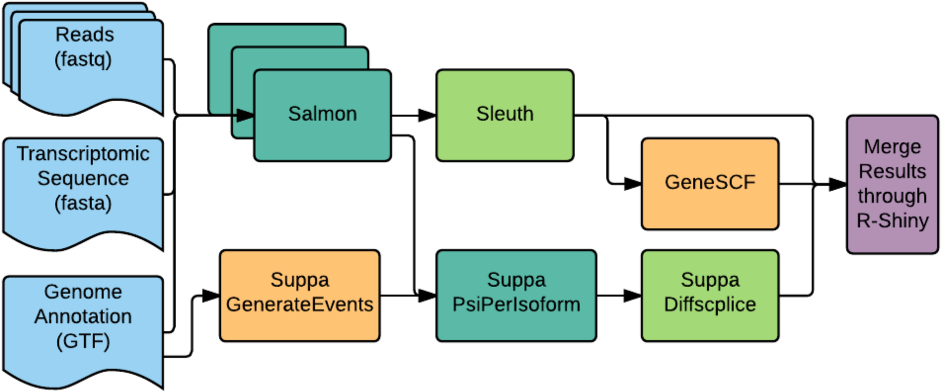
- Alternative splicing event caracterisation → Suppa[a]
- Metabolic pathways affected → geneSCF[b]

[a] https://bitbucket.org/regulatorygenomicsupf/suppa
[b] http://genescf.kandurilab.org/index.php

Snakemake

Snakemake is a workflow management system[3]. It herits from the GNU-Make rules philosophy, but provides an execution environment based on input/output description through wildcards, and python3.

[3] https://bitbucket.org/snakemake/snakemake/wiki/Home


Reads (fastq)

Transcriptomic Sequence (fasta)

Genome Annotation (GTF)

Salmon

Sleuth

GeneSCF

Suppa GenerateEvents

Suppa PsiPerIsoform

Suppa Diffscplice

Merge Results through R-Shiny

- Introduction
  - Definitions
  - Pipeline overview
- Challenges encountered
  - Expanded inputs
  - File closure delay on computing clusters
  - External module importation
- Future improvements
  - Rules versionning
  - Docker-isation

The expanded inputs

Snakemake "expand" key-workd

Used to represent a pythonic list and geather multiple datasets produced by a previous rule

$expand("\{quantified\_sample\}.tab, quantified\_sample = ["sample1", "sample2", ...])$

The required python[4] list, which is build by a **manual parsing** of the samples names, instead of letting Snakemake doing it by itself

[4] `https://www.python.org/`

File closure delay on computing clusters

Deployment through Torque

Torque Ressource Manager[a] handles process submission on a computing cluster

[a] http://www.adaptivecomputing.com/products/open-source/torque

File closure delay on computing clusters

Deployment through Torque

Torque Ressource Manager[a] handles process submission on a computing cluster

Variable file system latency due to the cluster charge

[a] http://www.adaptivecomputing.com/products/open-source/torque

File closure delay on computing clusters

Deployment through Torque

Torque Ressource Manager[a] handles process submission on a computing cluster

Variable file system latency due to the cluster charge

[a] http://www.adaptivecomputing.com/products/open-source/torque

Current workaround : - -**latency-wait** and - -**wait-for-files**, specially designed for cluster and file system latency

Rules that do not produce any output

Module loading example

On computing cluster, modules are akin path exportation in working session. It does not produce any output files, but are required to call system tools.

Current workaround: Snakemake provides both **temp** and **touch** as keywords

Introduction

Definitions

Pipeline overview

Challenges encountered

Expanded inputs

File closure delay on computing clusters

External module importation

Future improvements

Rules versionning

Docker-isation

Rules versionning

Different versions of a same tool

Using the same pipeline, which adapts through several versions of a given tool, would be a deployment asset.

Actually, it seems possible with **version** keyword, the - -**allowed-rules** command line option, or through the **bash command** itself

Docker for lazy deployment

What is Docker?

Docker is a technology designed to ”wrap a piece of software in a complete filesystem that contains everything needed to run [...] This guarantees that the software will always run the same, regardless of its environment.”[a]

[a] https://www.docker.com/

This would reduce the cost of the pipeline deployment, and would ensure, that used tools are up to date.

Special thanks

Fot their help and support, a special thanks to:

- Claire Toffano-Nioche
- Emilie Drouineau
- Pierre Bertin

Also, thank you for your attention