GRAPE RNAseq Analysis Pipeline Environment

David González, Maik Röder

Roderic Guigó’s lab (CRG)

June 12, 2012
Outline

1. Aim
2. Overview
3. Details
4. Installation
Outline

1. Aim
2. Overview
3. Details
4. Installation
Aim of GRAPE

- Automate the processing of the initial analysis of RNAseq data analysis.
- Record the exact environment used to ensure reproducibility.
- Store and organize the processed results in an easily accessible manner.
- Provide a visual interface to the data

---

Maik Röder
- Easy to use
- Easy to add new steps
Outline

1. Aim
2. Overview
3. Details
4. Installation
Two main steps

- Configuration
- Execution
Configuration

- Prepare accession file: sequence files and metainformation associated with them
- Prepare profile file: parameters used for the analysis (mismatches, genome file, annotation...)
- Create directory structure: This is done using the buildout tool.
Execution

- Execution up to the specified point. The execution script is based on the PIP workflow manager\textsuperscript{2}
- There are also a few scripts that can be used in common preprocessing and postprocessing tasks such as trimming and transforming or collecting the data from different pipeline runs.
- Data visualization using RAISIN.

\textsuperscript{2}Conery et al. 2005
Outline

1. Aim
2. Overview
3. Details
4. Installation
Quality Control

Aim

Overview

Details

Installation

Quality Control

Genome

Annotation

Reads

Transcriptome mapping

Quality summary

Average quality per position

Ambiguous bases (N's) per position

Read redundancy
Analyses

- Overall expression
  - Expressed clusters
    - Intronic, intergenic
    - Exonic fraction

- Junction Expression
  - Known
    - Novel
    - EJEI

- Exon expression
  - Exon RPKM
    - Exon inclusion level

- Gene expression
  - Gene RPKM
    - Gene readcount

- Transcript expression
  - Transcript RPKM
    - Fusion transcripts
Internally the pipeline is organized as a graph

- Each step depends on one or more previous steps.
- Each of these steps knows which are its dependencies.
- When a step is called it will check any necessary prerequisites and execute them only if needed in order to run successfully.
Mapping steps overview

1. Start
2. Read stats
3. Genome mapping
4. Junctions mapping
5. Transcriptome mapping
6. Unmapped reads
7. Split mapping
8. Junction Inclusion
9. Fusion transcripts
10. Recursive mapping
11. Still unmapped
12. Transcriptome unique maps
13. Genome unique maps
14. Junctions unique maps
15. Annotation tables and files
16. Indices
17. Exon seqs
18. Transcript seqs
19. Junction seqs
20. Genes RPKM
21. Gene mRPKM
22. Exon RPKM
23. Exon inclusion
24. Initial clusters
25. Junction/Split-map Classification tables

David González, Maik Röder
GRAPE
Step execution

- Start
- Read stats
- Junctions
- Exon seqs
- Transcript seqs
- Genome mapping
- Junctions mapping
- Transcriptome mapping
- Unmapped reads
- Split mapping
- Fusion transcripts
- Still unmapped
- Recursive mapping
- Initial clusters
- Junction/Split-map Classification tables
- Annotation tables and files
- Gene RPKM
- Gene mRPKM
- Exon RPKM
- Exon inclusion
Step execution

Start ➔ Junctions ➔ Junction seqs ➔ Indices ➔ Annotation tables and files

Genome mapping ➔ Junctions mapping ➔ Transcriptome mapping ➔ Unmapped reads ➔ Split mapping ➔ Junction Inclusion

Genome unique maps ➔ Junctions unique maps ➔ Transcriptome unique maps ➔ Split unique maps ➔ Initial clusters ➔ Junction/Split-map Classification tables

Gene RPKM ➔ Gene mRPKM ➔ Exon RPKM ➔ Exon inclusion

Fusion transcripts ➔ Still unmapped ➔ Recursive mapping
RAISIN, a web frontend for browsing the data that can be accessed at rnaseq.crg.cat.
Outline

1. Aim
2. Overview
3. Details
4. Installation
Prerequisites

- mysql
- R
- GEM
- Flux Capacitor
- Samtools
- Bioperl
- Perl and Python
Available at the crg svn server

svn://svn.crg.es/big/grape/grape.buildout

or at:

http://big.crg.cat/services/grape
Questions?