UNIVERSITÁ DEGLI STUDI DI SALERNO

DOTTORATO IN MANAGEMENT & INFORMATION TECHNOLOGY

CURRICULUM: INFORMATION SECURITY & INNOVATION SYSTEMS

COORDINATORE: Ch.mo. Prof. Antonelli Valerio

Ciclo XVII N.S.

Novel tools for reproducible Next Generation Sequencing data analysis and integration

Relatori

Ch.mo. Prof. Tagliaferri Roberto

Ch.mo. Prof. Angelini Claudia

Candidato

Righelli Dario

Matr. 8800800010

ANNO ACCADEMICO 2017/2018

How to reach a goal?
Without haste but without rest
Goethe

Add acknowledgements here

Write your abstract here

Contents

Chapter 1

Introduction

1.1 Biological Background

1.2 Sequencing Techniques

1.3 Computational Aspects

Chapter 2

TiCoRSe - Time Course RNA-Seq data analysis

2.1 Introduction

2.1.1 Time Course RNA-Seq

2.2 Methods

2.2.1 General Approach

2.2.2 Time Course Methods

2.2.3 Other Methods

2.2.4 Additional Features

2.3 Results

Chapter 3

DEScan2 - Differential Enriched Scan 2

few words on integration of epigenomic with transcriptomic

To investigate and to answer a epigenetic biological questions we decided to create an instrument useful for analysing the epigenomic data. Very often the biological questions, as for the RNA-Seq, to be answered, need the comparison of two or more different biological conditions. On this basis we designed *Differential Enriched Scan 2* (*DEScan2*), a software for helping the analysis of epigenomic data.

3.1 Introduction

The *Differential Enriched Scan 2* is an R [1] tool developed for detecting epigenomic signal in order to facilitate the Differential Enrichment of the signal between two or more biological conditions.

It is available on Bioconductor [2] since the version 3.7 and it's organized in three main steps.

A peak caller, which is a standard moving window scan that compares the counts within a sliding window to the counts in a larger region outside the window, using a simple Poisson likelihood (no overdispersion estimation) and providing a final score for each detected peak.

The filtering step is aimed to determine if a peak is a "true peak" on the basis of its replicability in other samples. To do so, this step is based on a double user-defined threshold, one on the peak score and one on the number of samples.

Finally, the third step produces a counts matrix where each column is a sample and each row a filtered peak computed in the filtering step. The value of the matrix cell is the number of reads for the peak in the sample.

The so produced counts matrix, as illustrated in the figure 3.1.1, is useful both for doing differential enrichment between the conditions and for integrating the epigenomic data with other -omic data types.

3.2 Methods

3.2.1 Peak Caller

However, the package can work with any external peak caller returning results in terms of bed files, indeed the package provides additional functionalities to load bed files of peaks and handle them as GenomicRanges [3] structures.

3.2.2 Peak Filtering and Alignment

Basing on this idea, the filtering step is developed to filter out those peaks not present in at least a user-defined number of samples. In the light of this, the user can decide the min- imum number of samples where each peak has to be detected. A further threshold can be used over the peak score.

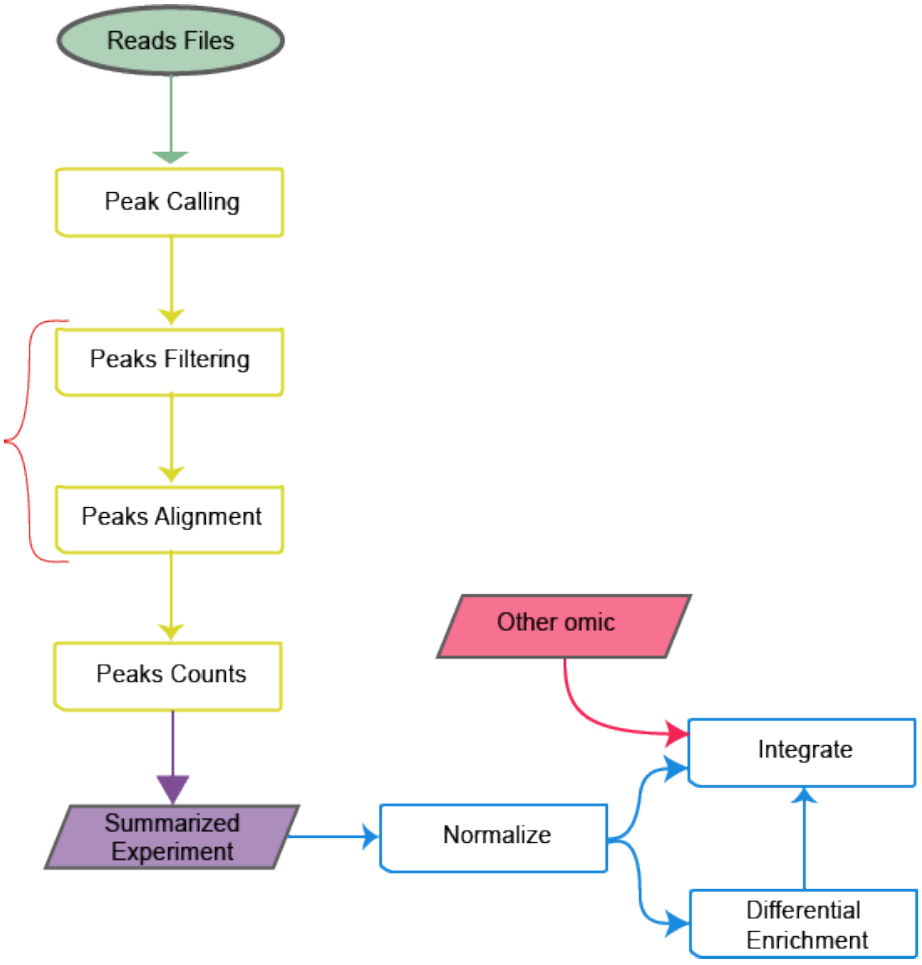
3.2.3 Peak Counts

3.2.4 Additional Features

Furthermore, our package provides several functionalities for GenomicRanges data structure handling. One over the others gives the possibility to split a GenomicRanges over the chromosomes to speed-up the computations parallelizing them over the chromosomes.

3.3 Results

Figure 3.1.1: A differential enrichement flow representation. DEScan2 steps are highlighed in yellow.



Chapter 4

IntegrHO - Integration of High-Throughtput Omics data

4.1 Introduction

4.2 Methods

4.2.1 Single Omic Approach

4.2.2 Multi Omic Approach

Low Level Itegration

High Level Itegration

4.3 Implementation Aspects

4.4 Reproducible Computational Research

4.5 Results

Chapter 5

Conclusions & Future Works

Appendices

.1 R Language

.2 R Markdown Language

Chapter 6

Bibliography

1. Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*. ISSN: 15372715. doi:10.1080/10618600.1996.10474713. arXiv: arXiv:1011.1669v3 (1996).
2. Gentleman, R. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*. ISSN: 1465-6914. doi:10.1186/gb-2004-5-10-r80 (2004).
3. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology* **9** (ed Prlic, A.) e1003118. ISSN: 1553-7358 (2013).

List of Figures

List of Tables