PDCB topic: High throughput sequencing analysis with Bioconductor August-December 2010

Leonardo Collado Torres

`lcollado@wintergenomics.com` and `lcollado@ibt.unam.mx`

`www.lcg.unam.mx/~lcollado/`

August 18, 2010

Abstract

The PDCB topic on high throughput sequencing analysis with Bioconductor will be taught at the *Centro de Ciencias Genómicas* at classroom 01 on Wednesdays from 17:00 to 19:00 hrs.

This topic will cover aspects of the high throughput sequencing (HTS) technology, Bioconductor, input/output and infrastructure packages, visualization and integration packages, and finally pipelinespecific packages.

The oficial page of the topic is `http://cursos.ccg.unam.mx`. There you can find the presentations, code associated to the presentations, exercises, expected answers, supporting material, and the data sets that we'll use.

Contents

1 Objectives

- Learn how to use BioC packages for HTS data analysis.

- Learn how to build a BioC package for future use.

- Realize the actual picture of HTS data analysis: it's not as pretty as you might think.

- Understand the statistics behind the use-case specific packages.

2 Sections

1. HTS tech: understanding the HTS technology with special emphasis on the Illumina Genome Analyzer.

2. BioC: understanding Bioconductor from the point of view of a package developper. This will enable you to know where to find the documentation and write your own packages!

3. Infra-IO: using the infrastructure and input-output BioC packages.

4. Vis-Int: using the visualization and integration BioC packages.

5. Advanced: going deep into the statistics and usage of pipeline specific packages such as those for RNA-seq and ChIP-seq.

3 Evaluation

Although it might seem very strict, we prefer to leave it as clear as possible at the beginning. As long as you hand in a minimum of 10 homeworks and your project, your grade will depend on four factors:

Attendance 20 %

Simply attending all topic sessions.

Participation 30 %

Your contributions to the sessions. It would be great if you found something interesting to share from sources such as the BioC mailing list.

Exercises 50 %

Each session will include some practical exercises. The next week, I'll publish the solution to the exercises and you'll have to grade them yourself. I will just make sure that your grades are coherent :)

4   *Tentative* session calendar

18 Aug Session I

R Reviewing R

1. Tentative session calendar
2. R review: from vectors to lists

25 Aug Session II

HTS tech Illumina

1. Understanding the tech.
2. Probable errors

1 Sept Session III

HTS tech   1. Dealing with tech. errors

8 Sept Session IV

BioC Introduction

1. Package Overview

15 Sept Session V

Session canceled

22 Sept Session VI

BioC Writing a vignette

1. Latex
2. Sweave
3. Stangle
4. weaver package

29 Sept Session VII

Infra-IO Reading an HTS file

1. ShortRead
2. Rsamtools

6 Oct Session VIII

Infra-IO Manipulating large data sets

1. IRanges
2. GenomicRanges

13 Oct Session IX

Infra-IO Infra-IO exercices

1. A series of exercices to practice your IO skills
2. Exercices where you need to manipulate large data sets

20 Oct Session X

Vis-Int Visualizing HTS data

1. GenomeGraphs
2. Rtracklayer

27 Oct Session XI

Vis-Int Integrating with other dbs

1. biomaRt
2. GenomicFeatures

3 Nov Session XII

Vis-Int Vis-Int exercises

1. Polishing your visualization and integration skills

10 Nov Session XIII

Advanced Statistics behind DE

1. RPKM definition
2. edgeR paper
3. DESeq paper

17 Nov Session XIV

Advanced DE exercises

1. edgeR
2. DESeq

24 Nov Session XV

Advanced Statistics behind ChIP-seq

1. Stats behind BayesPeak, CSAR, PICS and related packages

1 Dec Session XVI

Advanced ChIP-seq exercises

1. Calling ChIP-seq peaks
2. Associating with annotation via ChIPpeakAnno

8 Dec Session XVII

Advanced New R 2.12 packages

1. We'll review some of the new packages that will be released under Bioconductor 2.7

15 Dec Session XVIII

Advanced Topic round up

1. Filling in gaps that emerged during the course