BACTERIAL TRANSCRIPTION

EMBL 2010
Bioconductor Developer Meeting

Leonardo Collado Torres
Winter Genomics
OUTLINE

- Background information
  - Work team
  - Developer team
- Biology
- Goals
- Our work dynamic
- What we’ve done
- To do list
A DIVERSE WORK TEAM

- A benchwork lab (Morett’s at iBT UNAM)
  - Developer of new transcription start sites mapping techniques and maintainer of the UUSMD (local seq. facility)
- A bioinformatics lab (Collado-Vides’s at CCG UNAM)
  - Transcriptional bacterial regulation and maintainer of RegulonDB
- A new bioinformatics company (Winter Genomics)
  - New high throughput sequencing bioinformatics service company

- Undergraduate Program on Genomic Sciences (LCG) UNAM
  - All of the developers come from this program (graduated and current students)
DYNAMIC DEVELOPER TEAM

- April – June 2010
  - (5th) Alejandro Reyes Quiroz
  - (5th) Victor Moreno Mayar
  - (5th) Gabriel Cuellar Partida

- Aug 2010 – currently
  - (3rd) Carlos Vargas Chavez
  - (6th) Melvin Noe Gonzalez
  - (6th) Mayela Soto
  - (6th) Daniela Garcia Sorano

- Other programmers
  - Veronica Jimenez Jacinto
  - Leticia Vega Alvarado
  - Blanca Taboada
BIOLOGY

- Understand the transcriptional landscape at the genomic level

- Transcription Start Sites (TSSs)
  - Sites where the mRNA begins its transcription
  - Identify all active TSSs in a given condition
  - Unexpectedly high variability!

- Transcription Units (TUs)
  - One or multiple genes transcribed in the mRNA
  - Overlapping genes lead to complex cases!

- TSSs vs TUs correspondence
TSSgrams for different regulatory frontiers (Adapt 1)

- Convergent
- Intra-Orf
- Upstream

- Divergent
- Divergent_Forward
- Divergent_Reverse

% of positions used vs. Relative position (in bp)
TSSs Overlaps in E. coli

Y axis: Number of Unique Positions

- 1: Mono
- 2: MonoTri
- 3: TriAda
- 4: TriExo

Present only in:
- 1
- 2
- 3
- 4
- 1+2
- 1+3
- 1+4
- 2+3
- 2+4
- 3+4
- 1+2+3
- 1+2+4
- 1+3+4
- 2+3+4
- 1+2+3+4
TSSs Overlaps in E. coli
Y axis: Number of Reads

1: Mono
2: MonoTri
3: TriAda
4: TriExo

Present only in
1 2 3 4 1+2 1+3 1+4 2+3 2+4 3+4 1+2+3 1+2+4 1+3+4 2+3+4 1+2+3+4
Last possible initial position of read in TU 1

First possible initial position of read in TU 2
PROJECT GOALS

- Guarantee reproducibility
- Complete proposal on how to analyze this kind of data
  - Including working software!
- Facilitate future similar analyses from other bacteria
- Create easy (straight forward) to use software

- Learn more about BioC
WHAT WE’VE DONE SO FAR

- A pile of ideas 😊
- Most of the code for the TSSs is ready
  - Granges
  - List (up to 3) of SimpleRleList
- Prototypes for the 3 TU methods
- Summary information
  - GRanges // data.frame
  - Plots mostly using lattice
- Trained undergrads in R / BioC ^_^
TO DO

- Define how to evaluate the TU methods
- Evaluate them

- DOCUMENTATION!

- Feedback on objects that would be less prone to being broken by users
- Check the SummarizedExperiment class

- Aim: getting done prior to the next release
LINKS

- Collado-Vides’ lab http://www.ccg.unam.mx/en/ComputationalGenomics
- Winter Genomics http://www.wintergenomics.com/
- UUSMD http://uusmd.unam.mx/
- LCG http://www.lcg.unam.mx/