

NGS data analysis with Galaxy

EMBO Global Exchange Lecture Course – Tunis – Tunisia

High-throughput next generation sequencing applied to infectious diseases

Jean-François Taly – [Bioinformatics Core Facility](#) – [CRG](#) – Barcelona – Spain

THURSDAY 18TH SEPTEMBER		
10:45-12:30	Lecture 9: Fundamentals of NGS data analysis using Galaxy	1h45
12:30-14:00	Lunch	
14:00-15:45	Tutorial 4: Creating a DNA-seq Galaxy workflow	1h45
15:45-16:15	Coffee break	
16:15-18:00	Tutorial 5: Creating an RNA-seq Galaxy workflow	1h45

Introduction

Galaxy Rationale

The following text has been extracted from [Goeks et al. Genome Biol. 2010](#).

“Computation has become an essential tool in life science research. This is exemplified in genomics, where first microarrays and now massively parallel DNA sequencing have enabled a variety of genome-wide functional assays, such as ChIP-seq and RNA-seq (and many others), that require increasingly complex analysis tools. However, sudden reliance on computation has created an 'informatics crisis' for life science researchers: computational resources can be difficult to use, and ensuring that computational experiments are communicated well and hence reproducible is challenging. Galaxy helps to address this crisis by providing an open, web-based platform for performing accessible, reproducible, and transparent genomic science. “

About the teacher

Jean-François Taly is the head of the Bioinformatics core facility at the Center for Genomic Regulation in Barcelona, Spain. Originally trained in biochemistry, he obtained his PhD in structural bioinformatics. As a postdoctoral researcher, he worked in the field of multiple sequence alignments and phylogeny. Since he joined the core facilities, he provides bioinformatics trainings and services, working closely with investigators in the experimental design and analysis of genome-scale research using next-generation sequencing technologies. Additionally, he supports the CRG residents for any issues related to scientific software usage or databases and web sites development. Convinced by the necessity to bring bioinformatics to experimental biologists, he deployed in his institute a customized local instance of the Galaxy platform.

Lecture Outline

This lecture is about the Galaxy platform and not about describing bioinformatics methods for NGS analysis.

- Intro
 - When it started? Who is developing?
 - User friendly GUI to command line software
 - No need of Linux expertise
 - But need help from an administrator for set up
 - Galaxy and nanoPORE will bring high throughput sequencing to our laptops
 - Reproducibility and collaborative work
 - Froze an environment in a virtual machine or Docker
 - Distribute a workflow with your articles
 - Easy sharing of results and workflows
- Galaxy community:
 - Overview of Galaxy public servers
 - Galaxy is developed by an active community
- Galaxy admin: Local implementation and Custom Wrappers
 - Pros and cons for deploying a local Galaxy
 - Fast overview of installation process
 - Wrapper Vs Software
- Demo of the General Functions
 - Tools
 - 3 different up-load systems
 - File editing: Merge files, Extract columns, sort and filter
 - Tool shed
 - Writing your wrappers
 - History
 - Keep track of what you did
 - Share data and analysis
 - Workflow
 - Transform a group of analysis into an automatic procedure
 - Edit and re-use
 - Share your workflow
 - Visualization with Trackster
 - Display your data in the local Galaxy genome browser
- Demo of NGS in Galaxy on a RNA-Seq example
 - Alignments
 - Peak calling
 - Genome browser
- Presentation of the two hands-on tutorials
 - DNA-seq
 - RNA-seq

Virtual Machine

We use for this workshop a virtual machine set up with all software and databases necessary. You need for using it to install Virtualbox (<http://virtualbox.org/>) on your machine. We used as a starting point the virtual machine prepared for the [tutorials](#) of the Galaxy Community Conference [GCC2014](#). After you installed Virtualbox, just download the following image and double-click on it. Start the virtual machine and log as user “galaxy” with password “galaxy”.

- Image (5Gb): http://public-docs.crg.es/biocore/JFT/EMBO-Tunis/EMBO_course_Tunis2014.ova
- User: galaxy
- Password: galaxy

Starting the server and open Galaxy

Once started the machine, follow this protocol:

1. Right-click on the folder “galaxy” situated in your Desktop and select “open a terminal here”.
2. In the terminal type “sh run.sh”.
3. Open Firefox (There is a link in the Desktop)
4. Click on the “galaxy” bookmark
5. Galaxy user name is admin@pasteur.rns.tn and password is “galaxy”

You will find in the desktop a file called README where are given the command lines used for configuring the server and preparing the input data.

Some additional configuration in order to adapt to the tutorial computers

Change keyboard settings

1. Open a terminal
2. Type “sudo dpkg-reconfigure keyboard-configuration”
3. Keep the selected generic keyboard and press “Enter”
4. Select the French language and press “Enter”
5. Keep default for all and press “Enter”

Configure the server

1. Go to the directory “Desktop/galaxy”
2. Open the file “universe_wsgi.ini” with a text editor
3. Erase the “#” symbol in front of “host = 127.0.0.1”
4. Save and Exit

Configure Firefox

1. Go to Edit>Preferences>Advanced>Network>Settings>Proxy
2. Check the box “manual proxy configuration”
3. Type 172.20.4.1

Tutorial 4 – DNA-seq

Introduction

This tutorial aim to reproduce a part of the results obtained in the article of Zhang *et al.* “Genetic analysis of *Leishmania donovani* tropism using a naturally attenuated cutaneous strain”, [Plos Pathogen 2014](#).

Material

Note that we will work only on a subset of the initial raw data (SRA accession number: SRS484822 and SRS484824). Moreover, for this tutorial, we will work only on the chromosome 36 of *Leishmania donovani* BPK282A1. The genomic sequence in FASTA format as well as the cognate annotations in GFF have been downloaded from the TritypDB server:

<http://tritrypdb.org/common/downloads/release-8.0/LdonovaniBPK282A1>.

Results expected

Among the different outcome of this paper is the fact that the strain CL is showing a specific mutation in the coding sequence gene Ras-like GTPase. The supplementary table S1 gives the following information:

Chrom	Position	Change	Gene ID	Size	Site	Gene product
36	2280763	G/A	LdBPK_366140.1	364	R231C	Ras-like small GTPases, putative

The Figure 2 gives a multiple sequence alignment of close orthologs, showing that the mutation is found in a very conserved region (“-” indicates sequence identity).

```

L.donovani 282A1    192 SKTPYITELLQMLNSNSNIDLSYFLSHSKIYVAVDERNRRLKSRTYDLCSDAIEVVVKMSRIYM 255
L.donovani VL      192 -----R----- 255
L.donovani CL      192 -----C----- 255
L. major          192 -----G----- 255
L. mexicana       192 ----- 255
L. braziliensis   192 C-M-----T 255
L. guyanensis     192 C-M-----T 255
T.cruzi           192 PQS-----V-----R---FL-L---S-A-L---G-----E-MM---Q--T 255
T.brucei          192 LQI-----I-----L-----R---FLS---V-T---EI-----MMG-N---S 255
Strigomonas culicis 273 PRLQ--N---E---GTC-AVY-----R---L-SA-H-V-----D--I-T-D--T 336
Angomonas deanei   93  H---N-----V-----R---L-A---A-----G---EV-R 156
  
```

Method given in the article

Alignments - text extracted from the supplementary file Page 1

“Paired-end reads were aligned against the reference genome using BWA (version 0.7.3a-r367), configured to allow a maximum of two mismatches (-k 2) in a 23 bases long seed sequence (-l 23) and its sampe module, with expected maximum insert size of 1000 bases (-a 1000), was used to generate the alignments in sam format. The default values were accepted for all other BWA parameters. Single-end reads were aligned using samse module with similar alignment stringency. SAMtools (v0.1.18) was used to convert sam files into binary format (samtools view), sort (samtools sort) and index (samtools index) bam files.”

Variant Prediction calling - Text extracted from the supplementary file Page 2


"The RealignerTargetCreator script from the GATK suite was used to identify all intervals from bam files that contained indels. Then IndelRealigner script extracted reads from these regions and performed a local alignment (smith-waterman based) creating a new bam file with realigned indel regions. Variants were called by inputting alignment files (bam) from all the libraries together into the UnifideGenotyper script from the GATK suite. Both single nucleotide polymorphisms and small indels were called simultaneously using the `-genotype_likelihood_models=BOTH` option. Identification of high quality variants was enforced by (a) restricting our identification to regions with a combined coverage of 250 (all 4 libraries; `-dcov=250`), (b) bases with a phred-scaled quality score of 21 or more (`-min_base_quality_score=21`), (c) variants with a phred-scaled confidence of at least 30, (`-stand_emit_conf=30`) and (d) `--sample_ploidy=8`."

Method employ in this tutorial

1. Input data quality check with FASTQC
2. Convert FASTQ to the proper format with FASTQ groomer
3. Filter for bad quality reads
4. Aligned reads to the reference genome with BWA
5. Call SNPS with GATK
6. Annotate SNPs with snpEFF
7. Align protein sequences with ClustalW
















Step-by-step

Loading input data

1. The main screen ("Analyze Data") is divided in 4 blocks. The letters B, T, H and D will give the location of the item described in the text.
 - a. Header bar (B)
 - b. Tool panel on the left (T)
 - c. History panel on the right (H)
 - d. Display box in the center (D)
2. Getting the FASTQ files from the ENA web sites
 - a. Note that we won't work on this dataset
 - b. In the tool panel (T). GOTO: "Get Data" > "EBI SRA"
 - c. In the display box (D). Type "SRS484822" in the box asking for an ENA accession number and click on search.
 - d. (D) In the "Read Files" box, click on the link "File 1" in the line corresponding to the run "SRR1254937" and the column "FASTQ files (galaxy)"
 - e. In the history panel (H), a new dataset appears in grey. The dataset will turn yellow while downloading and change color to green when finished.
 - f. (H) Click on  to stop the download and delete the dataset in order to avoid wasting the limited resources of the server.
3. Import a published history. The inputs FASTQ of this tutorial have been pre-loaded into Galaxy.




- a. In the header bar (B). Go to “Shared Data” > “Published Histories” > “EMBO DNA-seq Start”
- b. (D) 7 datasets are displayed:
 - i. Two for the paired end (pe) data from the Cutaneous Leishmaniasis (CL) sample
 - ii. Two for the paired end (pe) data from the Visceral Leishmaniasis (VL) sample
 - iii. One for the single end (se) data from the Cutaneous Leishmaniasis (CL) sample
 - iv. One for the single end (se) data from the Visceral Leishmaniasis (VL) sample
 - v. One bed files containing the coordinates of the genes annotated in the chromosome 36
- c. (D) Click on import this history

Basic dataset functions






4. Basic function for datasets (H)
 - a. (H) 3 symbols for each dataset for view , edit  and delete 
 - b. (H) Click on  of dataset 1 “CL-pe1.fastq” to view the content
 - i. Notice the [FASTQ](#) format
 - c. (H) Click on  to edit the attributes
 - i. (D) Change the info as you wish
 - ii. (D) Notice the list of available genome builds
 - iii. (D) Save
 - iv. (D) Click on “Datatype”
 - v. (D) Click on the drop down list
 - vi. (D) Notice the different format available
 - vii. (D) Click on Permissions to share the dataset with users
 - d. (H) Click on the dataset name “CL-pe1.fastq”
 - i. (H) Notice info about dataset size, format and genome build
 - ii. (H) Notice the preview
 - iii. (H) Notice 2 boxes of functions
 1. Save , get info , re-do  and visualize 
 2. Edit tags  and annotate 
 - iv. (H) Click on , the file is downloaded on your local machine
 - v. (H) Click on 
 1. Notice all information about the dataset and how it has been generated
 - vi. We will see redo and visualize later
 - vii. (H) Change tag by clicking on 
 - viii. (H) Annotate the dataset by clicking on 
 - e. (H) : Get information about the BED file
 - i. What is the size of the BED dataset?
 - ii. Display its content on the screen
 - iii. Notice the [BED6](#) format


Input data quality check and pre-processing

5. Perform the first analysis: Quality Check
 - a. (T) Go to “NGS: QC and manipulation” > “FastQC:Read QC”
 - b. (D) Click of the drop list of available datasets under “Short read data from your current history”
 - c. (D) Select the dataset “VL-se.fastq”
 - d. (D) Click on “Execute”
 - e. (H) Notice that a new dataset appears. The new dataset successively is colored in grey, yellow and green when the job is pending, running and finished respectively.
 - f. (H+D) Display the content of the dataset
 - g. (D) Notice the different criteria evaluated
 - h. (D) What is [PHRED](#) quality format of the file?
6. Quality Check of all other FASTQ files in a row
 - a. (T) Go back to the tool, “NGS: QC and manipulation” > “FastQC:Read QC”
 - b. (D) Click on next to “Short read data from your current history”
 - c. (D) Select a pool of datasets with shift+click
 - d. (D) Execute
 - e. (H) Several analysis are ran in parallel
 - f. (H+D) Display the content of the FASTQC output for CL-pe1
 - g. (D) What is PHRED quality format of the file?
7. Convert the two single end datasets to the default sanger format
 - a. (T) Go to “NGS: QC and manipulation” > “FASTQ Groomer”
 - b. (D) Click on next to “File to groom”
 - c. (D) Select the two single end datasets with ctrl+click
 - d. (D) Check options and select “Illumina 1.3-1.7” for “Input FASTQ quality score”
 - e. (D) Select “Show Advanced Options” in the cognate menu
 - f. (D) Check options but leave the Output FASTQ quality scores type to “Sanger (recommended)”
 - g. (D) Leave other options to their defaults
 - h. (D) Execute
 - i. (H) After execution, click on the name of one of the new dataset
 - j. (H) Click on
 - k. (D) Click on the link “stdout” to check the logs of the tool
 - l. (H) For both datasets, click on and change the name to the cognate value (“FASTQ Groomer on CL-se” or “FASTQ Groomer on VL-se”)
8. Join the left and right reads of paired-end data. This is an obligatory step when dealing with paired-end reads in Galaxy
 - a. (T) Go to “NGS: QC and manipulation” > “FASTQ Joiner”
 - b. (D) Select “CL-pe1.fastq” as Left-hand Reads
 - c. (D) Select “CL-pe2.fastq” as Right-hand Reads
 - d. (D) Leave “FASTQ Header Style” to “old”
 - e. (D) Execute

- f. (H) Click on  of the new dataset “FASTQ joiner on data2 and data1”.
- g. (D) Notice that the two end-reads have been fused into a unique one
- h. (H) Click on the name of the new dataset “FASTQ joiner on data2 and data1”
- i. (H) Click on the “re-do” symbol 
- j. (D) The tool has been re-called with the parameters used for generating the dataset
- k. (D) Select the two paired-end datasets of the VL sample
- l. (D) Execute
- m. (H) For both datasets, click on  and change the name to the cognate value (“FASTQ joiner on CL-pe” or “FASTQ joiner on VL-pe”)



Remove from FASTQ files the bad quality reads

- 9. Filter the single end reads per average quality
 - a. (T) Go to “NGS: QC and manipulation” > “Filter FASTQ”
 - b. (D) Click on  and select the two FASTQ Groomer datasets
 - c. (D) Note the different filtering criteria but leave them to default
 - d. (D) Click on “Add a Quality Filter on a Range of Bases”
 - e. (D) Select “mean of scores” in the menu “Aggregate read score for specified range”
 - f. (D) Leave the next option to “>=”
 - g. (D) Type 30.0 in the box “Quality Score”
 - h. (D) Execute
 - i. (H) Click on the name of one of the new dataset
 - j. (H) In the log is given the number of reads and the percentage of reads kept after filtering
 - k. (H) For both datasets, click on  and change the name to the cognate value (“Filter FASTQ on CL-se” or “Filter FASTQ on VL-se”)
- 10. Filter the paired-end reads per average quality
 - a. (T) Go to “NGS: QC and manipulation” > “Filter FASTQ”
 - b. (D) Click on  and select the two FASTQ joiner datasets
 - c. (D) Check the box “This is paired end data”
 - d. (D) Click on “Add a Quality Filter on a Range of Bases”
 - e. (D) Select “mean of scores” in the menu “Aggregate read score for specified range”
 - f. (D) Leave the next option to “>=”
 - g. (D) Type 30.0 in the box “Quality Score”
 - h. (D) Execute
 - i. (H) For both datasets, click on  and change the name to the cognate value (“Filter FASTQ on CL-pe” or “Filter FASTQ on VL-pe”)
- 11. Split the fused paired-end reads
 - a. (T) Go to “NGS: QC and manipulation” > “FASTQ splitter”
 - b. (D) Click on  and select the two filtered paired-end FASTQ datasets
 - c. (D) Execute
 - d. (H) Click on the name of a new dataset
 - e. (H) In the preview box, note the characters “/1” or “/2” at the end of the first read name


- f. (H) For all datasets, click on  and change the name to the cognate value ("FASTQ splitter on CL-pe1" or "FASTQ splitter on CL-pe1")

Align the reads to a reference genome

12. Align single end reads to the reference genome


- (T) Go to "NGS: Mapping" > "Map with BWA for Illumina"
- (D) A reference genome is already selected, check it is IdoX36
- (D) Select the dataset "Filter FASTQ on CL-se"
- (D) Select the full parameter list
- (D) Set option "aln -l" to "23"
- (D) Set "sampe/samse -r" to "Yes"
- (D) Type "CL-se" in the box corresponding to "ID", "LB" and "SM"
- (D) Type "ILLUMINA" in the box corresponding to "PL"
- (D) Execute
- (H) Use  to call back the tool,
- (D) Select the filtered FASTQ of VL-se and change the value of "ID", "LB" and "SM" accordingly
- (H) For all datasets, click on  and change the name to the cognate value ("BWA CL-se SAM" or "BWA VL-se SAM")

13. Align paired-end reads to the reference genome

- (H) Use  from "BWA CL-se SAM" to call back the tool
- (D) Select paired end
- (D) Select the filtered FASTQ of CL-pe
- (D) Set the option sampe -a to "1000"
- (D) Change the value of "ID", "LB" and "SM" accordingly
- (D) Execute
- Repeat a to f for VL-pe
- Change the dataset names to BWA "CL-pe SAM" and BWA "VL-pe SAM"

Filter out bad not aligned reads and SAM/BAM manipulation

14. Filter-out the not-aligned single end reads

- (T) Go to "NGS: SAM Tools" > "Filter SAM or BAM"
- (D) Click on  and select the two single-end SAM alignments
- (D) Set "Filter on bitwise flag" to "yes"
- (D) In the panel "Skip alignments with any of these flag bits set" check the box "The read is unmapped"
- (D) Execute
- (H) Change the dataset names to "Filter SAM CL-se" and "Filter SAM VL-se"
- (H+D) Display content of "BWA CL-se SAM". The second column for read "SRR1254939.20781" is 4
- Open a new Firefox tab and go to <http://picard.sourceforge.net/explain-flags.html>
- What is the annotation associated to 4?

- j. (H+D) Display content of “Filter SAM CL-se”. The read “SRR1254939.20781” has been filtered out.
15. Filter-out the not-aligned paired-end reads
 - a. (H) Use from “Filter SAM CL-se” to call back the tool
 - b. (D) Click on and select the two paired-end SAM alignments
 - c. (D) In the panel “Only output alignments with all of these flag bits set” check the box “Read is mapped in proper pair”
 - d. (D) Execute
 - e. (H) Change the dataset names to “Filter SAM CL-pe” and “Filter SAM VL-pe”
16. Convert SAM in BAM.
 - a. (T) Go to “NGS: SAM Tools” > “SAM-to-BAM”
 - b. (D) Click on and select all filtered SAM alignments
 - c. (D) The genome is already selected
 - d. (D) Execute
 - e. (H) Change the names to “Filter BAM XX-XX”
 - f. (H) Compare the size of the CL-pe SAM and BAM files.

Visualize reads in the Galaxy genome browser (Trackster)



17. Visualize your reads on the local genome browser
 - a. (B) Go to “Visualization” > “New Track Browser”
 - b. (D) Give your visualization a name and select the genome IdoX36
 - c. (D) Click on “Add Datasets to this Visualization”
 - d. (D) Check the box corresponding to the four BAM files
 - e. (D) Four lines appear and turn to yellow while Galaxy is converting the BAM file to [BigWig](#) and [BedGraph](#)
 - f. (D) Zoom to the region with most coverage: in the coordinates bar, click and slide along the bar while maintaining the click.
 - g. (D) The aligned reads appear
 - h. (D) Put the mouse cursor to the first track, a list of options appear next to the name
 - i. (D) Click on and change the display mode to “coverage”
 - j. (D) Try the different display mode and set it to “coverage” for all tracks
 - k. (D) Click on to save your visualization
 - l. (B) Go back to “Analyze Data”

Variant calling with GATK

18. Use GATK to realign the reads around INDELS with an alignment method more suitable for insertion and deletions. The first tool detect the INDELS and create a files with their coordinates. The second tool realign the regions specified. Useful tutorial are given at the web site of GATK: <https://www.broadinstitute.org/gatk>
 - a. (T) Go to “NGS: GATK variant analysis” > “Realigner Target Creator”
 - b. (D) Click on and select all filtered BAM alignments
 - c. (D) Execute with default values

- d. (H) Rename the interval files to “Realigner Target Creator XX-XX”
 - e. (T) Go to “NGS: GATK variant analysis” > “Indel Realigner”
 - f. (D) For each sample, select the cognate BAM and interval files and Execute with default values
 - g. (H) Rename the new BAM datasets to “Indel Realigner on XX-XX”
19. Use GATK Unified Genotype to call variants.
- a. (T) Go to “NGS: GATK variant analysis” > “Unified Genotyper”
 - b. (D) With the button “Add new BAM file”, insert the four realigned BAM datasets
 - c. (D) Execute with options ...
 - d. (H+D) Look at the content of the [VCF](#) file

Annotate SNPs with snpEFF

20. Annotate the variants. This step aim to predict the effect of the SNPs on the annotated genes. This tool use a precompiled databases build from the FASTA sequence of the genome and the cognate GFF annotation. More details <http://snpeff.sourceforge.net>
- a. (T) Go to “NGS: snpEFF variant analysis” > “SnpEff”
 - b. (D) The VCF files from GATK is already selected as well as the database.
 - c. (D) Execute with default values
 - d. (D) Two outputs are generated: an HTML output giving a detailed summary and another VCF files with additional snpEFF annotations. See the snpEFF manual for more explanation: http://snpeff.sourceforge.net/SnpEff_manual.html#input
21. Filter the annotated variants to keep only the one occurring in coding sequence and corresponding to a non-synonymous mutation.
- a. (T) Go to “NGS: snpEFF variant analysis” > “SnpSift Filter”
 - b. (D) Select the SnpEff VCF dataset
 - c. (D) Type "(EFF[*].EFFECT = 'NON_SYNONYMOUS_CODING')"
 - d. (D) Execute
 - e. (D) Display the content of the output VCF file. Only 16 SNPs have been retained
 - f. (D) For the SNP at position 2280763, note the annotation
“NON_SYNONYMOUS_CODING(MODERATE|MISSENSE|Cgc/Tgc|R231C|LdBPK_366140”
 - g. (D) Put the mouse cursor on any of the SNP line, click on the visualization symbol appearing 
 - h. (D) Select “View in saved visualization”
 - i. (D) Select your visualization
 - j. (D) Click on  too add the BED file containing annotated genes
 - k. (D) Check the box corresponding to “TritypDB-8.0_lدونوانيBPK282A1_genes_X36.bed”
 - l. (D) Click on Add
 - m. (D) Click on the coordinates box and type “Ld36_v01s1:2280746-2280776”
 - n. (D) Note that the two samples from CL have a mutation G/A leading to an amino acid substitution R231C as annotated by snpEFF.

Tutorial 5 –RNA seq

Introduction

This tutorial is based on the article “Iron regulates differentiation in Leishmania” by Mittra *et al.* published in J. Exp. Med. 2013: <http://www.ncbi.nlm.nih.gov/pubmed/?term=23382545/>

Material

Note that we will work only on a random selection of 10% of the initial raw data (SRA accession number: SRP016502). The genomic sequence in FASTA format as well as the cognate annotations in GFF has been downloaded from the TritypDB server:

<http://tritrypdb.org/common/downloads/release-8.0/LmexicanaMHOMGT2001U1103/>

Results Expected, from the article p404

“Analysis of the results revealed that 11 genes were up-regulated by more than threefold in two independent replicates, whereas 21 genes were down-regulated by at least threefold after growth in iron-depleted medium. Another 59 and 109 genes showed consistent twofold up- or down-regulation, respectively (selected examples are shown in Table S2). As expected, genes up-regulated more than threefold include the ferrous iron transporter LIT1 (LmxM.30.3070; Huynh et al., 2006) and the ferric iron reductase LFR1 (LmxM.29.1610; Flannery et al., 2011), whereas several genes that encode iron-dependent proteins were down-regulated (LmxM.34.1540, LmxM.08.0290, and LmxM.18.0510).”

Method given in the article P 412-413















“Reads containing a TTG tag at the 5’ end, indicating the presence of authentic SL sequences, were separated from non-TTG-containing reads. The first three bases (TTG) of the reads were trimmed, and the remaining sequence was aligned against the L. mexicana genome (TriTrypDB version 4.0; Aslett et al., 2010) using bowtie (Langmead et al., 2009) configured to allow roughly one to three mismatches depending on quality of base where the mismatch occurred. Alignments were processed using SAMtools (Li et al., 2009), and internally developed Perl scripts were used to calculate the number of reads aligned against each gene (including the CDS and the 5’ upstream intergenic region). edgeR software (Robinson et al., 2010) was used for differential expression (DE) analysis of count data from all four libraries. Genes that did not have one read per million (cpm, counts per million) aligned reads in all four samples were excluded from further DE analysis. The biological coefficient of variation within the biological replicates was estimated using edgeR and used in DE calculations. Based on a negative binomial error model, edgeR was used to fit the count data to a generalized linear model and to calculate DE for each gene. P-values were adjusted for multiple comparisons as described in Klipper-Aurbach et al. (1995). The SL RNA-seq data have been submitted to the GEO database under accession no. GSE41641.”

Method employed

1. Quality check with FASTQC
2. Remove the splice leader sequence with CutAdapt
3. Align with Bowtie2
4. Assemble aligned reads with CuffLinks
5. Merge assembled UTRs annotation into a unique one with CuffMerge
6. Measure differential expression with CuffDiff and the new annotation
7. Filter the output to keep only significant results
8. Visualize differentially expressed genes with the Galaxy genome browser Trackster

Step-by-step

Loading input data

1. “Analyze Data” screen can be divided in 4 blocks
 - a. Header bar (B)
 - b. Tool panel on the left (T)
 - c. History panel on the right (H)
 - d. Display box in the center (D)
2. Getting the FASTQ files from the ENA web sites
 - a. Note that we won’t work on this dataset
 - b. In the tool panel (T). GOTO: “Get Data” > “EBI SRA”
 - c. In the display box (D). Type “SRP016502” in the box asking for an ENA accession number and click on search.
 - d. (D) In the “Read Files” box, click on the link “File 1” in the line corresponding to the run “SRR594631” and the column “FASTQ files (galaxy)”
 - e. In the history panel (H), a new dataset appears in grey. The dataset will turn yellow while downloading and change color to green when finished.
 - f. (H) Click on  to stop the download and delete the dataset in order to avoid wasting the limited resources of the server.
3. Import a published history
 - a. Click on “Shared Data” (B)
 - b. Click on Published Histories
 - c. Click on “EMBO RNA-seq START”
 - d. Import History
4. Basic function for datasets (H)
 - a. 3 symbols for each dataset for view , edit  and delete 
 - b. Click on  of file “Iron” to view the content
 - i. Notice the fastq format
 - c. Click on  to edit the attributes
 - i. Change the name as you wish
 - ii. Notice the list of available genome builds
 - iii. Save
 - d. Click on the dataset “IronMinus_Rep1_SRR594632.fastq”
 - i. Notice info about dataset size, format and genome build
 - ii. Notice the preview
 - iii. Notice 2 boxes of functions
 1. Save , get info , re-do  and visualize 
 2. Edit tags  and annotate 
 - iv. Click on , the file is downloaded on your local machine
 - v. Click on 
 1. Notice all information about the dataset and how it has been generated

- vi. We will see redo and visualize later
- vii. Change tag by clicking on
- viii. Annotate the dataset by clicking on
- e. (H) : Get information about the GFF file
 - i. What is the size of the GFF dataset?
 - ii. Display its content on the screen
 - iii. Notice the [GFF](#) format

Input data quality check and pre-processing

5. Perform the first analysis: Quality Check
 - a. (T) Go to “NGS: QC and manipulation” > “FastQC:Read QC”
 - b. (D) Click of the drop list of available datasets under “Short read data from your current history”
 - c. (D) Select the first dataset “IronMinus_Rep1_SRR594632.fastq”
 - d. (D) Click on “Execute”
 - e. (H) Notice that a new dataset appears
 - f. (H) The new dataset successively is colored in grey, yellow and green when the job is pending, running and finished respectively.
 - g. (H+D) Display the content of the dataset
 - h. (D) Notice the different criteria evaluated
6. Quality Check of all other FASTQ files in a row
 - a. (T) Go back to the tool, “NGS: QC and manipulation” > “FastQC:Read QC”
 - b. (D) Click on next to “Short read data from your current history”
 - c. (D) Select a pool of datasets with shift+click
 - d. (D) Execute
7. Convert all FASTQs to the default sanger format
 - a. (T) Go to “NGS: QC and manipulation” > “FASTQ Groomer”
 - b. (D) Click on next to “File to groom”
 - c. (D) Select all FASTQ datasets with ctrl+click
 - d. (D) Check options and select “Illumina 1.3-1.7” for “Input FASTQ quality score”
 - e. (D) Select “Show Advanced Options” in the cognate menu
 - f. (D) Check options but leave the Output FASTQ quality scores type to “Sanger (recommended)”
 - g. (D) Leave other options to their defaults
 - h. (D) Execute
 - i. (H) After execution, click on the name of one of the new dataset
 - j. (H) Click on
 - k. (D) Click on the link “stdout” to check the logs of the tool
 - l. (H) For all datasets, click on and change the name to the cognate value (“FASTQ Groomer on IronMinus/Plus Rep1/2”)

Filter out bad quality reads

8. Remove the splice leader sequence

- a. (T) Go to “NGS: QC and manipulation” > “Cutadapt”
 - b. (D) Click on and select all FASTQ datasets
 - c. (D) Click on the box “Add new 5’ (Front) Adapters”
 - d. (D) From the “source” drop-down list “select Enter custom sequence”
 - e. (D) Replace the default custom sequence by “^TTG”
 - f. (D) Change the maximum error rate to 0.0
 - g. (D) Click on the box “Match Read Wildcards”
 - h. (D) From the “Additional OutPut options” menu, select “Additional Output Files”
 - i. (D) Check the box “Untrimmed Reads”
 - j. (D) Execute
 - k. (H+D) Look at report
 - l. (H) Change the name of the FASTQ files corresponding to selected reads
9. Hide some datasets for a better clarity of your history
- a. (H) Click on ☒, click on “All” and de-select the cutadapt FASTQs and the BED file
 - b. (H) Click on “For all selected ...” and on “Hide”
 - c. (H) Click again on ☒ to finish the process
 - d. (H) To see your hidden datasets, click on and “Include Hidden Datasets”

Align reads onto the reference genome

10. Align the reads onto the reference genome with Bowtie2
- a. (T) Go to “NGS: Read Mapping” > “Bowtie2”
 - b. (D) Do not click on ! Run one by one all FASTQ datasets
 - c. (D) Select the *Leishmania mexicana* as reference genome
 - d. (D) Execute with default parameters
 - e. (H) Rename the BAM output datasets
11. Evaluate alignments statistics
- a. (T) Go to “NGS: SAM Tools” > “flagstat”
 - b. (D) Click on and select all BAM datasets
 - c. (H+D) Look at the output. What is the percentage of aligned reads on IronPlus Rep1?
12. Align the reads onto the reference genome with Bowtie2 with faster algorithm
- a. (T) Go to “NGS: Read Mapping” > “Bowtie2”
 - b. (D) Select only “IronPlus Rep1” FASTQ
 - c. (D) Select the *Leishmania mexicana* as reference genome
 - d. (D) Select “Full parameter list” in the drop-down menu “Parameter Settings”
 - e. (D) Select the “Very Sensitive” option of the “Preset” drop-down list
 - f. (D) Execute
 - g. (D) Use flagstat to know the percentage of mapped reads
 - h. (D) Delete the two last datasets with

Visualize reads in the Galaxy genome browser (Trackster)

13. Visualize your reads on the local genome browser
- o. (B) Go to “Visualization” > “New Track Browser”
 - p. (D) Give your visualization a name and select the genome Imex

- q. (D) Click on “Add Datasets to this Visualization”
- r. (D) Check the box corresponding to the four BAM files and the BED file
- s. (D) Four lines appear and turn to yellow while Galaxy is converting the BAM file to [BigWig](#) and [BedGraph](#)
- t. (D) Zoom to the region with most coverage: in the coordinates bar, click and slide along the bar while maintaining the click.
- u. (D) The aligned reads appear
- v. (D) Put the mouse cursor to the first track, a list of options appear next to the name
- w. (D) Click on and change the display mode to “coverage”
- x. (D) Try the different display mode and set it to “coverage” for all tracks
- y. (D) Note that the reads fall outside of the annotated genes. We cannot then use the GFF annotation to measure differential expression.
- z. (D) Click on to save your visualization
- aa. (B) Go back to “Analyze Data”

Create a new annotation from aligned reads

- 14. Assemble the aligned reads into transcripts
 - a. (T) Go to “NGS: RNA analysis” > “Cufflinks”
 - b. (D) Click on and select all BAM datasets
 - c. (D) Execute with default values
 - d. (H) Cufflinks generates one GTF and two expression files
- 15. Merge the four generated GTF files
 - a. (T) Go to “NGS: RNA analysis” > “Cuffmerge”
 - b. (D) Use the button “Add Additional GTF Input Files” to select the four “assembled transcripts” files
 - c. (D) Execute with default values

Measure differential Expression

- 16. Measure differential expression with the new annotation
 - a. (T) Go to “NGS: RNA analysis” > “Cuffdiff”
 - b. (D) Select the Cuffmerge annotation
 - c. (D) Give the name “Iron+” to the first condition
 - d. (D) Use the button “Add replicates” to select the two corresponding BAM datasets
 - e. (D) Give the name “Iron-” to the second condition and
 - f. (D) Use the button “Add replicates” to select the two corresponding BAM datasets
- 17. Sort the gene expression output by decreasing values of log₂(Fold Change)
 - a. (T) Go to “Filter and Sort” > “Sort”
 - b. (D) Select the dataset “Gene Differential Expression Testing”
 - c. (D) Type “10” in the column box
 - d. (D) Execute
- 18. Filter the sorted output to keep only significant values
 - a. (T) Go to “Filter and Sort” > “Filter”
 - b. (D) Select the sorted dataset

- c. (D) Type “c14==‘yes’” in the condition box
 - d. (D) Execute
 - e. (D) Only 20 lines remain
19. Annotate the list of deregulated with Trackster
- a. Open your visualization in a new Firefox tab
 - b. (D) Copy the coordinates of the most overexpress gene
 - c. (D) Paste it in coordinates box of Trackster
20. Compare with published Table S2

Gene ID	Gene Name	Locus	log2(FC) _{article}	log2(FC) _{galaxy}
LmxM.08_29.0620	ABC transporter, putative	LmxM.08:915717-915860	2.6	2.55634
LmxM.14.1360	myo-inositol-1-phosphate synthase	LmxM.14:561059-561202	2.15	2.13183
LmxM.19.0985	4-coumarate:coa ligase-like protein	LmxM.19:375283-375332	NA	1.58612
LmxM.27.1260	T-complex protein 1, beta subunit	LmxM.27:511109-511314	NA	-1.51378
LmxM.23.0050	Cyclophilin, putative	LmxM.23:12571-12690	NA	-1.80997
LmxM.36.2360	Tyrosine aminotransferase	LmxM.20:934474-934561	-1.77	-1.85951
LmxM.07.0060	Cytochrome C1	LmxM.07:21802-21923	NA	-1.87449
LmxM.32.1140	Hypothetical protein	LmxM.32:413889-414222	-2.38	-2.49446
LmxM.36.6330	Hypothetical protein	LmxM.20:2412004-2412160	NA	-2.66887
LmxM.30.1200	Hypothetical protein	LmxM.30:459936-460021	-2.6	-2.79336
LmxM.02.0460	Hypothetical protein	LmxM.02:172598-172710	-3.45	-3.02987