

NGS data analysis with Galaxy

CRG Master Course 2014 - 10th October 2014

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

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. “

Tutorial – 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](#). As any member of the CRG you can access to galaxy.crg.es with your CRG credentials.

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://tritypdb.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 SKTPYITELLQMLNSNSNIDLSYFLSHSKIYVAVDERNRLKSRTYDLCSDAIEVVVKMSRIYM 255
L.donovani VL      192 -----C----- 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 RealingerTargetCreator 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” > “LdoX36 SNP Calling”
 - 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
- a. (T) Go to “NGS: Mapping” > “Map with BWA for Illumina”
 - b. (D) A reference genome is already selected, check it is IdoX36
 - c. (D) Select the dataset “Filter FASTQ on CL-se”
 - d. (D) Select the full parameter list
 - e. (D) Set option “aln -l” to “23”
 - f. (D) Set “sampe/samse -r” to “Yes”
 - g. (D) Type “CL-se” in the box corresponding to “ID”, “LB” and “SM”
 - h. (D) Type “ILLUMINA” in the box corresponding to “PL”
 - i. (D) Execute
 - j. (H) Use  to call back the tool,
 - k. (D) Select the filtered FASTQ of VL-se and change the value of “ID”, “LB” and “SM” accordingly
 - l. (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
- a. (H) Use  from “BWA CL-se SAM” to call back the tool




- b. (D) Select paired end
- c. (D) Select the filtered FASTQ of CL-pe
- d. (D) Set the option sampe -a to "1000"
- e. (D) Change the value of "ID", "LB" and "SM" accordingly
- f. (D) Execute
- g. Repeat a to f for VL-pe
- h. 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
 - a. (T) Go to "NGS: SAM Tools" > "Filter SAM or BAM"
 - b. (D) Click on  and select the two single-end SAM alignments
 - c. (D) Set "Filter on bitwise flag" to "yes"
 - d. (D) In the panel "Skip alignments with any of these flag bits set" check the box "The read is unmapped"
 - e. (D) Execute
 - f. (H) Change the dataset names to "Filter SAM CL-se" and "Filter SAM VL-se"
 - g. (H+D) Display content of "BWA CL-se SAM". The second column for read "SRR1254939.18077" is 4
 - h. Open a new Firefox tab and go to <http://picard.sourceforge.net/explain-flags.html>
 - i. 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 default 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_IdonovaniBPK282A1_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, as expected, the two samples from CL have a mutation G/A leading to an amino acid substitution R231C as annotated by snpEFF

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