

Monday, 14th January 2013

Initiation to Galaxy

“Variant detection into chromosome 4 of the human genome”

N. Choisne, S. Gallina, O. Inizan, A. Lermine, D. Naquin, J. van Helden, M. Zytnicki



CNRS UPMC INSU

Station Biologique
Roscoff

aviesan

alliance nationale
pour les sciences de la vie et de la santé

1. Get data

- Go to Shared Data/Data Libraries
- Go the library named "TP Initiation"
- Select chr4.fastq and chr4.fa and import them to the current History
- Clic on customTrack.txt and select "Download this dataset"

2. Quality control on raw data

2.1 QC

- Run **Fastqc** tool on chr4.fastq file

FastQC:Read QC (version 0.51)

Short read data from your current history:
1: chr4.fastq

Title for the output file - to remind you what the job was for:
FastQC
Letters and numbers only please - other characters will be removed

Contaminant list:
Selection is Optional
tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Execute

- Look at the generated html file (clik on eye)
 - * How many reads do you have?
 - * What is the read size?
 - * What is the mean quality of most of the reads?

2.2 Fastq filtering by quality

- Run **Filter FASTQ** tool

Filter FASTQ (version 1.0.0)

FASTQ File:
Requires groomed data: if your data does not appear here try using the FASTQ groomer.

Minimum Size:
0

Maximum Size:
0
A maximum size less than 1 indicates no limit.

Minimum Quality:
0.0

Maximum Quality:
0.0
A maximum quality less than 1 indicates no limit.

Maximum number of bases allowed outside of quality range:
0

This is paired end data:

Quality Filter on a Range of Bases
The above settings do not apply to these filters.
Add new Quality Filter on a Range of Bases

Execute

- Galaxy doesn't offer you any fastq file. It's due to the quality values which are not normalized (C.f. Fastq format presentation). In order to normalize these values, you will use the **FASTQ groomer** tool:

FASTQ Groomer (version 1.0.4)

File to groom:
1: chr4.fastq

Input FASTQ quality scores type:
Sanger

Advanced Options:
Hide Advanced Options

Execute

- Then rerun **Filter FASTQ** (with minimum quality at 20)

Filter FASTQ (version 1.0.0)

FASTQ File:
4: FASTQ Groomer on data 1

Requires groomed data: if your data does not appear here try using the FASTQ groomer.

Minimum Size:
0

Maximum Size:
0
A maximum size less than 1 indicates no limit.

Minimum Quality:
20.0

Maximum Quality:
0.0
A maximum quality less than 1 indicates no limit.

Maximum number of bases allowed outside of quality range:
0

This is paired end data:

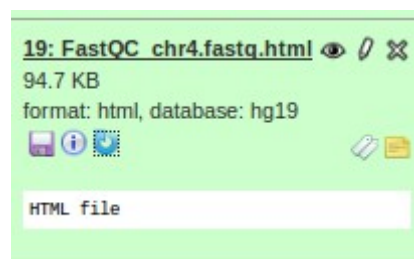
Quality Filter on a Range of Bases
The above settings do not apply to these filters.

Add new Quality Filter on a Range of Bases

Execute

- Run **Fastqc** on groomed and filtered fastq

****To rerun easily tools, just clic on the blue arrow in dataset box****



- Just change the input file by selecting the groomed and filtered fastq file

FastQC:Read QC (version 0.51)

Short read data from your current history:

5: Filter FASTQ on data 4

Title for the output file - to remind you what the job was for:

FastQC

Letters and numbers only please - other characters will be removed

Contaminant list:

Selection is Optional

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Execute

- * How many reads do you have?
- * What is the read size?
- * What is the mean quality of most of the reads?

3. Alignment of raw reads with Bowtie

- Run the **Map with Bowtie for Illumina** tool (look at the parameters on screenshot)

Map with Bowtie for Illumina (version 1.1.2)

Will you select a reference genome from your history or use a built-in index?
 ↕
Built-ins were indexed using default options

Select the reference genome:
 ↕

Choose whether to use Default options for building indices or to Set your own:
 ↕
These settings are ignored when using a prebuilt index

Is this library mate-paired?:
 ↕

FASTQ file:
 ↕
Must have ASCII encoded quality scores

Bowtie settings to use:
 ↕
For most mapping needs use Commonly used settings. If you want full control use Full parameter list

Skip the first n reads (-s):

Only align the first n reads (-u):

-1 for off

Trim n bases from high-quality (left) end of each read before alignment (-5):

Trim n bases from low-quality (right) end of each read before alignment (-3):

Maximum number of mismatches permitted in the seed (-n):

May be 0, 1, 2, or 3

Maximum permitted total of quality values at mismatched read positions (-e):

Seed length (-l):

Minimum value is 5

Whether or not to round to the nearest 10 and saturating at 30 (--nomaqround):
 ↕

Number of mismatches for SOAP-like alignment policy (-v):

-1 for default MAQ-like alignment policy

Whether or not to try as hard as possible to find valid alignments when they exist (-y):
 ↕
Tryhard mode is much slower than regular mode

Report up to n valid alignments per read (-k):

Whether or not to report all valid alignments per read (-a):
 ↕

Suppress all alignments for a read if more than n reportable alignments exist (-m):

-1 for no limit

Write all reads with a number of valid alignments exceeding the limit set with the -m option to a file (--max):

Write all reads that could not be aligned to a file (--un):

Whether or not to make Bowtie guarantee that reported singleton alignments are 'best' in terms of stratum and in terms of the quality values at the mismatched positions (--best):

Use best

Removes all strand bias. Only affects which alignments are reported by Bowtie. Runs slower with best option

Maximum number of backtracks permitted when aligning a read (--maxbts):

800

Whether or not to report only those alignments that fall in the best stratum if many valid alignments exist and are reportable (--strata):

Use strata option

Override the offset of the index to n (-o):

-1

-1 for default

Seed for pseudo-random number generator (--seed):

-1

-1 for default

Suppress the header in the output SAM file:

Bowtie produces SAM with several lines of header information by default

Execute

- * What is the output format?
- * How can you distinguish the mapped reads from the unmapped reads?
- * What mean 0, 4 et 16 in second row?
- **Visit <http://picard.sourceforge.net/explain-flags.html>**
- * How can you see that you have single end reads?

4. SAM to BAM conversion

- Run **SAM-to-BAM** tool

SAM-to-BAM (version 1.1.2)

Choose the source for the reference list:

History

Convert SAM file:

7: Map with Bowtie f...apped reads

Using reference file:

2: chr4.fa

Execute

5. Mapping statistics

- Run the **flagstat** tool

flagstat (version 1.0.0)

BAM File to Convert:

8: SAM-to-BAM on dat...nverted BAM

Execute

- * How many reads are present ?
- * How many reads are mapped?
- * How many reads are unmapped?
- * How many reads are duplicates?

6. Duplicates detection

- Run the **Mark Duplicates reads tool** (don't use Remove duplicates from output file)

Mark Duplicate reads (version 1.56.0)

SAM/BAM dataset to mark duplicates in:
8: SAM-to-BAM on dat..nverted BAM ↕
If empty, upload or import a SAM/BAM dataset.

Title for the output file:
Dupes Marked
Use this remind you what the job was for

Remove duplicates from output file:

If true do not write duplicates to the output file instead of writing them with appropriate flags set.

Assume reads are already ordered:

If true assume input data are already sorted (most Galaxy SAM/BAM should be).

Regular expression that can be used to parse read names in the incoming SAM file:
[a-zA-Z0-9]+:[0-9]:([0-9]+):([0-9]+):([0-9]+).*

Names are parsed to extract: tile/region, x coordinate and y coordinate, to estimate optical duplication rate

The maximum offset between two duplicate clusters in order to consider them optical duplicates.:
100
e.g. 5-10 pixels. Later Illumina software versions multiply pixel values by 10, in which case 50-100.

Execute

- Run the **flagstat** tool on the output bam file from last step

flagstat (version 1.0.0)

BAM File to Convert:
10: MarkDups_Dupes Marked.bam ↕

Execute

* How many reads are duplicates?

7. Variant Calling

- Run the **Mpileup** tool

MPileup (version 0.0.1)

Choose the source for the reference list:
History ▾

BAM files

BAM file 1
BAM file:
10: MarkDups_Dupes Marked.bam ▾
Remove BAM file 1

Add new BAM file

Using reference file:
2: chr4.fa ▾

Genotype Likelihood Computation:
Do not perform genotype likelihood computation ▾

Set advanced options:
Advanced ▾

Do not skip anomalous read pairs in variant calling:

Disable probabilistic realignment for the computation of base alignment quality (BAQ):

Coefficient for downgrading mapping quality for reads containing excessive mismatches:
0

Max reads per BAM:
250

Extended BAQ computation:

List of regions or sites on which to operate:
Selection is Optional ▾

Minimum mapping quality for an alignment to be used:
20

Minimum base quality for a base to be considered:
20

Only generate pileup in region:

Output per-sample read depth:

Output per-sample Phred-scaled strand bias P-value:

Execute

* What is the meaning of row 3, 4 et 5?

* How can we read the content of row 5?

Visit <http://samtools.sourceforge.net/pileup.shtml>

8. Pileup file filtering

- Run the **Filter Pileup** tool

Filter pileup (version 1.0.2)

Select dataset:

which contains:

See "Types of pileup datasets" below for examples

Do not consider read bases with quality lower than:

No variants with quality below this value will be reported

Do not report positions with coverage lower than:

Pileup lines with coverage lower than this value will be skipped

Only report variants?:

See "Examples 1 and 2" below for explanation

Convert coordinates to intervals?:

See "Output format" below for explanation

Print total number of differences?:

See "Example 3" below for explanation

Print quality and base string?:

See "Example 4" below for explanation

- * How many variant have you?
- * What is the meaning of row 4 et 12?

- Run the **Compute** tool

Compute (version 1.1.0)

Add expression:

as a new column to:

Dataset missing? See TIP below

Round result?:

- * Whats the meaning of the calculated value?
- * How can we use this value to select the most likely variants?

- Run the **Filter** tool

Filter (version 1.1.0)

Filter:
16: Compute on data 15 Dataset missing? See TIP below.

With following condition:
c13>=0.5

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Execute

- * How many good variants do you have?
- * What are the characteristics of this(ese) variant(s)?
 - Chromosome position
 - Nucleotide from the reference?
 - Nucleotide from the variant?
 - Variation frequency?

9. Identified variant visualization

- Go to this [page](#)
- Clic on choose file and select the customTrack.txt file downloaded at the beginning
- Clic on go to genome browser
- Go to the variant chromosome and position

10. Construct the corresponding workflow

- Clic on the history parameters button
- Select Extract Workflow
- Give it a name and clic on Create Workflow
- You can edit your workflow by clicking on Workflow/You workflow name/Edit

11. Results publication

- Use the galaxy pages system to publish your results (Answers to the questions and integration of the created datasets).