

Preliminary Processing of NGS Data

Mark Dunning: Sheffield Bioinformatics Core Director
(sbc.shef.ac.uk)

23 January 2018

Practical setup

Go to one of following Galaxy servers

- ▶ Surnames: A-K
 - ▶ <https://bioinf-galaxian.erasmusmc.nl/galaxy/>
- ▶ Surnames: L-N
 - ▶ <https://galaxy.hidelab.org/>
- ▶ Surnames: O-Z
 - ▶ <http://services.cbib.u-bordeaux.fr/galaxy/>

Data Upload

► Go *Get Data*

- *Upload File* (may be a different place on the menu depending which server you connect to!)

The screenshot displays the Galaxy/Galaxian web interface. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'Login'. A light blue banner at the top contains the text: 'Please refrain from running large jobs on teaching days (see news on main page)'. On the left, a 'Tools' sidebar is visible, with a search bar and a 'Get Data' section. A red arrow points to the 'Upload File from your computer' option in the 'Get Data' list. The main content area features a 'Welcome to Galaxian the EMC Galaxy Server!' message. Below this, the Erasmus MC University Medical Center Rotterdam logo is shown, along with a DNA double helix graphic and the text 'Erasmus Bioinformatics'. To the right, a 'Training Materials' section lists various resources, including 'Galaxy 101 Manual Slides (HTML) Slides (PDF)', 'Variation Analysis (DNA) Manual Slides (HTML) Slides (PDF)', 'RNA-Seq Differential Gene Expression Manual Answers Slides', and 'Cancer Analysis and Reporting Manual Slides (HTML) Slides (PDF)'.

Galaxy / Galaxian

Analyze Data Workflow Shared Data Visualization Help Login

Please refrain from running large jobs on teaching days (see news on main page)

Tools

search tools

Get Data

- Upload File from your computer
- EGA Download streamer data from the European Genome-phenome Archive in a secure manner
- Download Import import a file from owncloud (works best in Chrome)
- UCSC Main table browser
- UCSC Test table browser
- UCSC Archive table browser
- EBI SRA ENA SRA
- Get Microbial Data
- BioMart Ensembl server
- BioMart Test server

Welcome to Galaxian the EMC Galaxy Server!

Erasmus MC University Medical Center Rotterdam

Bioinformatics

Training Materials

The Galaxy Training Network (GTN) is a network of people and groups around the world that present Galaxy and Galaxy-based training. Check out some of the GTN's materials [here](#).

At the EMC we frequently offer Galaxy training courses on various subjects. All materials for these workshops can be found here:

- Galaxy 101 Manual Slides (HTML) Slides (PDF)
- Variation Analysis (DNA) Manual Slides (HTML) Slides (PDF)
- RNA-Seq Differential Gene Expression Manual Answers Slides
- Cancer Analysis and Reporting Manual Slides (HTML) Slides (PDF)

Figure 1

Data Upload


- ▶ *Choose local file*
 - ▶ Select JoeBlogsBRCAPanel_R1.fastq, JoeBlogsBRCAPanel_R2.fastq and click **Start**

Download from web or upload from disk

Regular Composite Collection

Drop files here

Type (set all): 🔍 Genome (set all):



Outline

- ▶ The *Fastq* format for sequencing reads
- ▶ Quality assessment of *fastq* files
- ▶ The *bam* format for representing *aligned* reads
- ▶ Stage 1 of an analysis *pipeline*

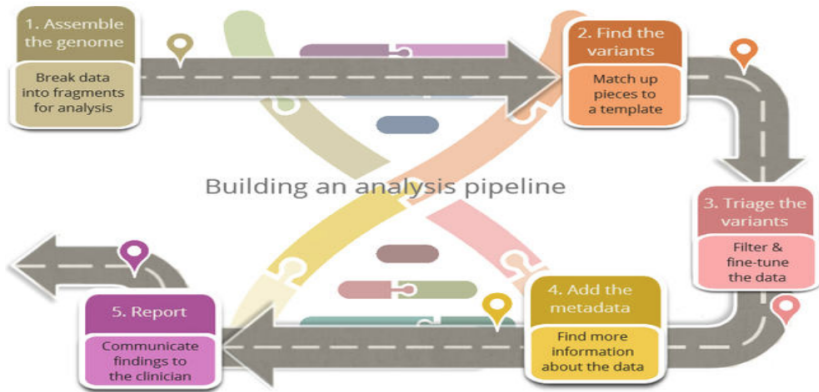
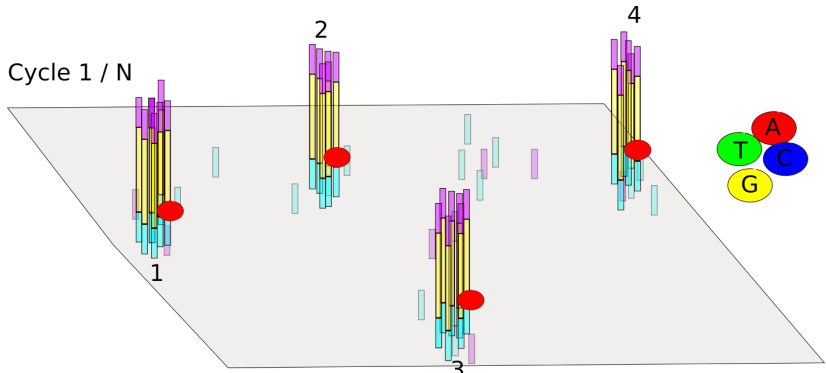


Figure 3

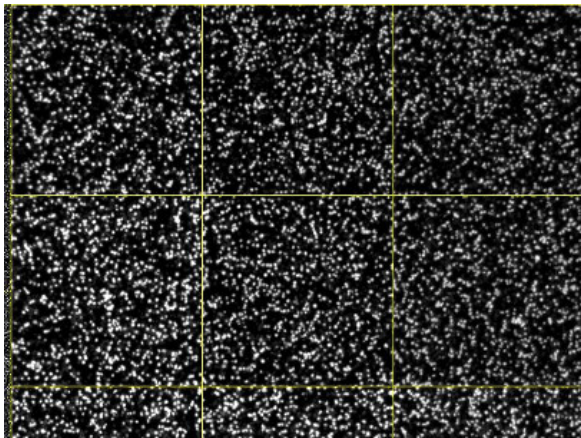
The sequencing process

- ▶ <https://www.youtube.com/embed/HMyCqWhwB8E> for a nice video
- ▶ sequencing consists of a series of *cycles* (e.g. 100), at each cycle we try and incorporate different bases (A, T, C, G)
- ▶ the base that is successfully added will illuminate brightly in a particular colour



Imaging

- ▶ much of the sequencing time is spent taking images of the flowcell
- ▶ it is these images that are used to discover what fragments of DNA were sequenced
- ▶ this process is not perfect and can introduce *uncertainty*



Scale of data

Instrument	No. of Reads	Size
Ion Torrent PGM	5 million reads	1Gb
MiSeq	25 million reads	6GB
HiSeq rapid run	600 million	150GB (*)
HiSeq high-output	4 billion	1 TB

- ▶ Equivalent to **40** HD movies
- ▶ File sizes are for 100 bp reads, unzipped
- ▶ Number of reads from thermofisher.com and illumina.com

Fastq format

- ▶ a text file
- ▶ can be *compressed* as a gz file
- ▶ four lines per read
- ▶ the sequenced is most interesting, there are two other lines that we potentially investigate

The diagram shows a single Fastq record consisting of four lines. Red arrows point from labels to specific parts of the record: 'Sequence ID' points to the first line, 'Sequenced Read' points to the second line, 'Blank' points to the third line, and 'Quality Score' points to the fourth line.

```
@SRR081708.237049.1  
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG  
+  
!>:<9>>>:=;>?<:@?>;==@@@>?=AAA<=A@?6>4B=<>>.@?<@;?#####
```

Figure 6

Sequence ID

Can contain the following

- ▶ name of sequencer
- ▶ flow cell *lane*
- ▶ coordinates of the read on the flow cell
- ▶ whether this is a *paired* read and whether it is read 1, or 2
- ▶ which biological sample the read came from (if multiple samples were sequenced on the same lane)

Quality scores

- ▶ base-calling has some probability (p) that we make a mistake.
- ▶ the quality score expresses our *confidence* in a particular base-call; higher quality score, higher confidence
- ▶ one such score for each base of sequencing. i.e. 100 scores for 100 bases of sequencing
- ▶ these are of importance if we want to call SNVs etc.
 - ▶ need to be sure that differences detected from the reference genome are real, and not caused by sequencing error

N?>:<9>>>:=;>>?<>:@?>;==@@@>?=AAA<>=A@?6>4B=<>>.@>?<@;?####

Quality Scores to probabilities

- ▶ look-up the ASCII code for each character
- ▶ subtract the offset to get the Q score
- ▶ convert to a probability using the formula:-

$$p = 10^{-Q/10}$$

Worked Example

for our particular example:

N?>:<9>>>:=;>>?<>:@?>;==@@@>?=AAA<>=A@?6>4B=<>>.@>?<@;?####

it works out as follows:-

	Character	Code	Minus.Offset..33..	Probability
1	N	78	45	0.00003
2	?	63	30	0.00100
3	>	62	29	0.00126
4	:	58	25	0.00316
5	<	60	27	0.00200
6	9	57	24	0.00398
7	>	62	29	0.00126
8	>	62	29	0.00126
9	>	62	29	0.00126
10	:	58	25	0.00316

...

Exercise

- Use the Galaxy tool *Text Manipulation* -> *Select last*

The screenshot shows the Galaxy web interface. On the left is a sidebar with a 'Tools' section containing a search bar and a list of tool categories: 'Send Data', 'Collection Operations', and 'Text Manipulation'. Under 'Text Manipulation', several tools are listed, including 'FASTQ to FASTA converter', 'Regex Find And Replace', 'Concatenate datasets tail-to-head', 'Column Regex Find And Replace', 'Transpose data from a file', 'Add/Remove chr prefix', 'Sort Chromosomal Position', 'Filter Columns', 'File Concatenation', 'Column Select', 'Strip Header', 'Add column', 'Cut columns', 'Merge Columns', 'Convert delimiters', 'Create single Interval', 'Change Case', 'Paste two files', 'Remove beginning', 'Select random lines', 'Select first lines', and 'Select last lines from a dataset'. The last tool is highlighted with a red box.

The main panel displays the configuration for the 'Select last lines from a dataset (Galaxy Version 1.0.0)' tool. It has a 'Select last' input field with the value '10' and a 'from' dropdown menu showing '16: FASTQ Groomer on data 13'. An 'Execute' button is at the bottom of the configuration area.

Below the configuration, the 'What it does' section states: 'This tool outputs specified number of lines from the end of a dataset'.

The 'Example' section shows the input file and the resulting output. The input file is a tab-separated table with 6 columns. The output shows the last two lines of the input file.

Input File:

chr7	57134	57154	D17003_CTCF_R7	356	-
chr7	57247	57267	D17003_CTCF_R4	207	+
chr7	57314	57334	D17003_CTCF_R5	269	+
chr7	57341	57361	D17003_CTCF_R7	375	+
chr7	57457	57477	D17003_CTCF_R3	188	+

Show last two lines of above file. The result is:

chr7	57341	57361	D17003_CTCF_R7	375	+
chr7	57457	57477	D17003_CTCF_R3	188	+

Exercise

- ▶ print the last *12* lines from the file
JoeBlogsBRCAPanel_R2.fastq
- ▶ how many reads are shown in the result?
- ▶ Look at the last read and write down the first five and last five ASCII characters
 - ▶ is the quality greater at the start, or the end of the read?

Exercise

- Use the Galaxy tool *Text Manipulation* -> *Line/Word/Character count*

Tools

search tools

Get Data

Collection Operations

Text Manipulation

- [Transpose](#) rows/columns in a tabular file
- [Reverse columns](#) in a tabular file
- [Datamash](#) (operations on tabular data)
- [Unique](#) occurrences of each record
- [melt](#) collapse combinations of variables: values to single lines
- [cast](#) expand combinations of variables: values to columnar format
- [Regex Replace](#) Regular Expression replacement using the Python re module
- [Compute](#) an expression on every row
- [Add column](#) to an existing dataset
- [Concatenate datasets](#) tail-to-head
- [Cut](#) columns from a table
- [Merge Columns](#) together
- [Convert](#) delimiters to TAB
- [Create single interval](#) as a new dataset
- [Change Case](#) of selected columns
- [Paste](#) two files side by side
- [Remove beginning](#) of a file
- [Select random lines](#) from a file
- [Select first](#) lines from a dataset
- [Select last](#) lines from a dataset
- [Trim](#) leading or trailing characters

Line/Word/Character count of a dataset (Galaxy Version 1.0.0) Options

Text file

2: JoeBlogsBRCAPanel_R2.fastq

Desired values

☒ Select/Unselect all

☒ Line count

☒ Word count

☒ Character count

Include Output header

Yes No

Execute

What it does

This tool outputs counts of specified attributes (lines, words, characters) of a dataset.

Example Output

#Lines	words	characters
7499	41376	624971

Citation

If you use this tool in Galaxy, please cite Blankenberg D, et al. *In preparation*.

Exercise

- ▶ count how many lines are in the file
JoeBlogsBRCAPanel_R2.fastq in total
- ▶ how many reads does this correspond to?

FastQC: Quality Assessment of fastq files

- ▶ FastQC from the Babraham Institute Bioinformatics Core has emerged as the standard tool for performing quality assessment on sequencing reads;
<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- ▶ The manual for FastQC is available online and is very comprehensive
- ▶ A “*traffic light*” system is used to draw your attention to sections of the report that require further investigation.
- ▶ fastqc will not actually *do* anything to your data. If you decide to trim or remove contamination for your samples, you will need to use another tool.
- ▶ it doesn't know what type of sequencing has been performed (WGS, exome, RNA-seq), which can affect interpretation of some of the plots

Example sections of a fastqc report

1. Basic Statistics



Basic Statistics

Measure	Value
Filename	sample.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	9053
Sequences flagged as poor quality	0
Sequence length	36
%GC	50

Figure 10

Example sections of a fastqc report

2. Per-base sequence quality

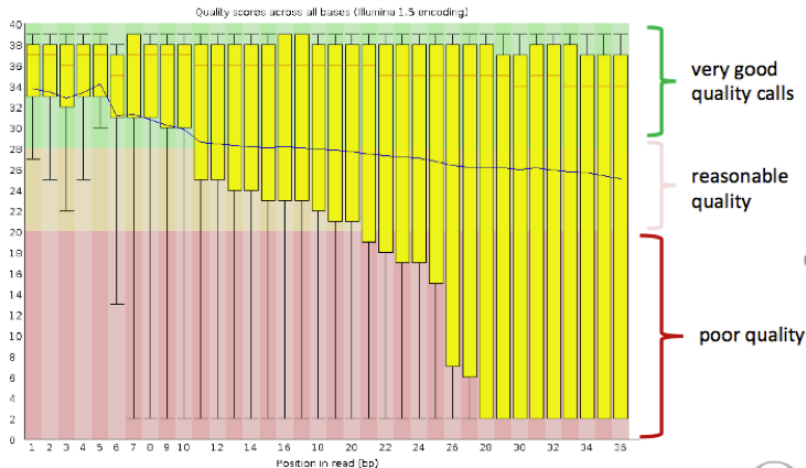


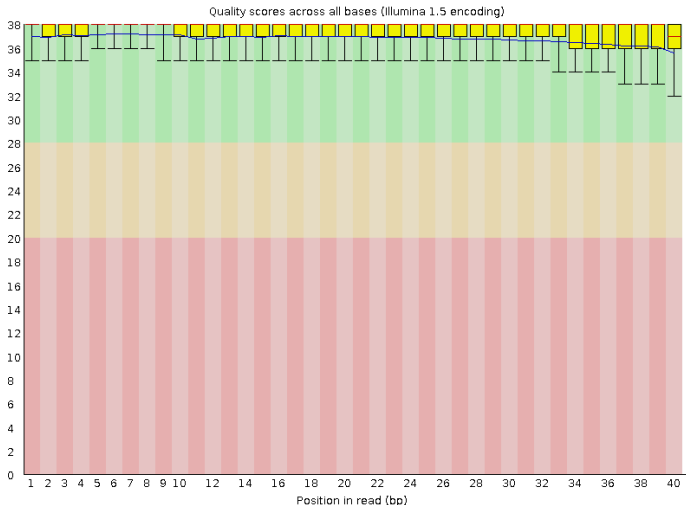
Figure 11

Example sections of a fastqc report

Ideally, the plot should look *something* like following:-



Per base sequence quality

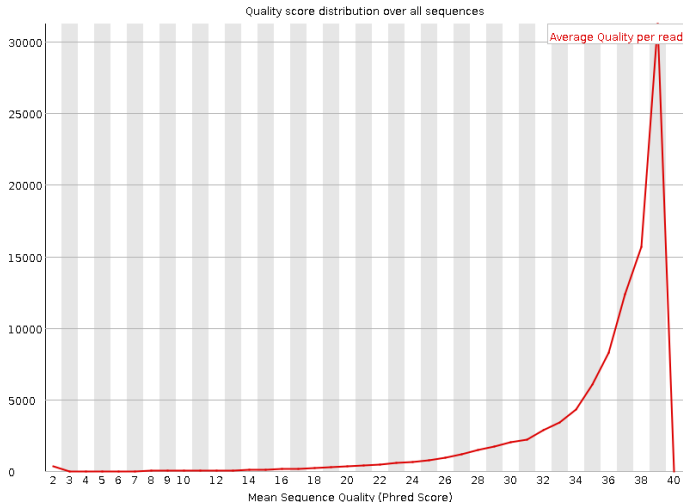


Example sections of a fastqc report

3. Per-sequence quality scores



Per sequence quality scores



Example sections of a fastqc report

[Per-base sequence content]



Per base sequence content

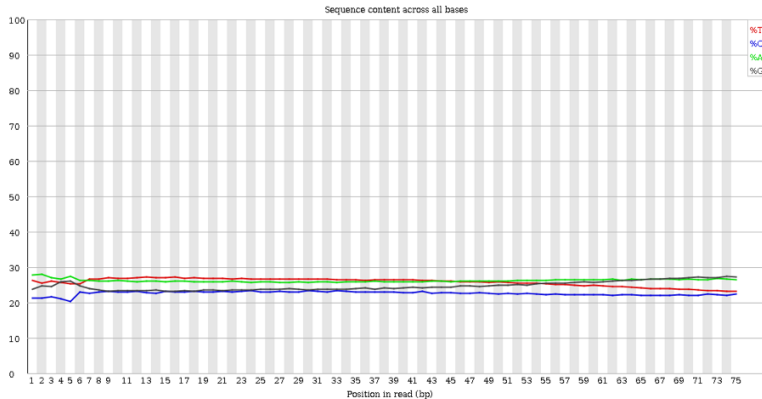


Figure 14

Exercise

- ▶ using the Galaxy Tool *NGS: QC and manipulation* -> *FastQC* Read Quality reports

The screenshot displays the Galaxy web interface. On the left, a sidebar lists various tools, with 'FastQC Read Quality reports' highlighted in a red box. The main panel shows the 'FastQC Read Quality reports (Galaxy Version 0.63)' tool configuration. The 'Short read data from your current history' dropdown is set to '6: JoeBlogsBRCAPanel_R1.fastq'. The 'Contaminant list' is set to 'Nothing selected'. The 'Submodule and Limit specifying file' is also set to 'Nothing selected'. An 'Execute' button is visible at the bottom of the configuration panel. Below the configuration panel, the 'Purpose' section explains that FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It also lists the main functions of FastQC: Import of data from BAM, SAM or FastQ files (any variant); Providing a quick overview to tell you in which areas there may be problems; Summary graphs and tables to quickly assess your data; Export of results to an HTML based permanent report; and Offline operation to allow automated generation of reports without running the interactive application. The 'FastQC' section states that this is a Galaxy wrapper for the external package FastQC, which is documented at [FastQC](#). It also mentions that the contaminants file parameter was borrowed from the independently developed fastqcwrapper contributed to the Galaxy Community Tool Shed by J. Johnson. The 'Inputs and outputs' section states that FastQC is the best place to look for documentation - it's very good. A summary follows below for those in a tearing hurry. It also mentions that the wrapper will accept a Galaxy fastq, sam or bam as the input read file to check. It will also take an optional file containing a list of contaminants information, in the form of a tab-delimited alternative option the tool takes a custom limits.txt file that allows setting the warning thresholds for the different modules and also specifies which modules to include in the output. The tool produces a basic text and a HTML output file that contain all of the results, including the following:

Figure 15

Exercise

- ▶ generate a QC report for the file
JoeBlogsBRCAPanel_R2.fastq
- ▶ look at the basic statistics for the file
 - ▶ does the number of reads agree with your previous answer?
- ▶ is there any evidence for a decrease in quality as the read length increases?
- ▶ See <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/> for descriptions of various sections of report

The SAM / BAM format

- ▶ we don't really spend much time look at *fastq* files
- ▶ most of our time is spent with *aligned* reads
 - ▶ i.e. we have used some software to tell us whereabouts in the genome each read belongs to
 - ▶ we will have a go at this in the practical

The .sam file

- ▶ **Sequence Alignment/Map** (sam)
- ▶ The output from an aligner such as bwa or bowtie
- ▶ Same format regardless of sequencing protocol (i.e. RNA-seq, ChIP-seq, DNA-seq etc)
- ▶ May contain un-mapped reads
- ▶ Official specification can be found online
<http://samtools.github.io/hts-specs/SAMv1.pdf>
- ▶ We normally work on a compressed version called a .bam file.
See later.

The .sam file

Comprises a *tab-delimited* section that describes the alignment of each sequence in detail.

1	2	3	4	5	6	7	8	9	10	11
SRR081708.237649	163	1	10003	6	1567M	=	10041	105		
GACCCTGACCCTAACCCCTGACCCTGACCCTGACCCTGACCCTAACCCCTGACCCTAACCCCTAA S<=====<=>=<?=?=?==@??;?>@@@=?@?@??@??@?>?@<@>@'@=?=??										
=<=>?>?=?Q ZA:Z:<@;0;0;;308;68M;68><@;0;0;;27;;>MD:Z:5A11A5A11A5A11A13 RG:Z:SRR081708 NM:1:6 OQ:Z:GEGFFEGGGDGGGGGGA?										
DCDD:GGGDDGDCFGDFFCCCBEBFDABDD-D:EEEE=D=DDDDC:										

Figure 16

- ▶ 1:- Sequence ID
- ▶ 2:- Sequence quality expressed as a bitwise *flag*
- ▶ 3:- Chromosome that the read aligned to

Fun with flags!

The “*flags*” in the sam file can represent useful QC information

- ▶ Read is unmapped
- ▶ Read is paired / unpaired
- ▶ Read failed QC
- ▶ Read is a PCR duplicate (see later)

SRR081708.237649 163 1 10003 6 1567M = 10041 105
GACCTTGACCCTAACCCAGACCTTGACCCTAACCCAGACCTTGACCCTAACCCATAA S=<====<>?<?=?>==@??;?>000=??@???@??@?>?@<@>@'=??
<=>?>?=Q ZA?Z:<@;>;308;68M;68<@;>0;0;27;;MD:Z:5A11A11A5A11A13 RB:F:SRR081708 EE:E:D:000 QO:Z:GEGFFFEGGDGGDGDDGA?
DCDD:GGGDGCDFGDFGFFCBBFDBADD-D NNM=D=DDD

Figure 17

Derivation

	ReadHasProperty	Binary	MultiplyBy
Paired?	TRUE	1	1
Properly Paired?	TRUE	1	2
Unmapped?	FALSE	0	4
Unmapped Mate?	FALSE	0	8
On Minus Strand?	FALSE	0	16
Mate on Minus Strand?	TRUE	1	32
Is First Read?	FALSE	0	64
Is Second Read?	TRUE	1	128
Is Secondary Alignment?	FALSE	0	256
Is Not Passing QC?	FALSE	0	512
Is Duplicate Read?	FALSE	0	1024

$$1 \times 1 + 1 \times 2 + 0 \times 4 + 0 \times 8 + 0 \times 16 + 1 \times 32 + 0 \times 64 + 1 \times 128 + 0 \times 256 + 0 \times 512 + 0 \times 1024 = 163$$

<https://broadinstitute.github.io/picard/explain-flags.html>

CIGAR string

Compact Idiosyncratic Gapped Alignment Report

- ▶ Value before **M** is number of consecutive mapping bases (can be mismatches)
- ▶ Value before **I** is number of bases inserted relative to reference
- ▶ Value before **D** is number of bases deleted relative to reference
 - ▶ e.g. 142**M**2**I**7**M** 2 bp insertion after 142 bases then 7 aligned bases

The .sam file

```
SRR081708.237649      163          10003    6       1567M   =       10041    105
GACCTCGACCTAACCTGACCCTGACCCTAACCTGACCCTGACCCTAACCTGACCCTAACCTAA S=<=====<?><=?>==???:?@?@???@??@?>?@?<@'=@?=??=
>====>?>ZA:Z:<@;0;308;68M;68M:<@;0;0;27;;MD:Z:5A11A5A11A13RG:Z:SRR081708NM:i:6OQ:Z:GEGFFFGGGDGDGGDDGA?
DCDD:GGDGDCGFDFFFCCBEBFDBBD-D:EEEE=D=DDDDD:
.....
SRR081708.237649      83          1       10041    18       68M   =       10003    -105
CCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCC #####?>@?>. >.>=<B4>6?@A>=<AAA=?>@&==;>?>@?>
>;>::>?>?NZA:Z:<@;0;308;~<@;0;27;1567M;5A11A5A11A5A11A13>MD:Z:68RG:Z:SRR081708NM:i:0OQ:Z:#####?<8@?
@,7:88:165@AA?>?BE@EBBA@>@AB?>B?AAB?B-CCCCACCC
```

Figure 19

- ▶ 7, 8, 9: Alignment information for the paired read (if available)
 - ▶ whether they align to the same chromosome
 - ▶ where the position of the paired read is
 - ▶ how far apart did they map?

The .sam file

1 **2** **3** **4** **5** **6** **7** **8** **9** **11**

SRR081708.237649 163 1 10003 6 1567M = 10041 105

GACCTCGACCTTAACCTGACCCTAACCTGACCCTGACCCTGACCCTGACCCTAACCTTAA S=<=====><?=?==@??;?>@@=??@?????@??@?>7@<-@-'@=?=?

=<=>?>=Q ZA:Z:<6;0;0;308;68M;68<0;0;0;27;?>MD:Z:5A11A5A11A5A11A13 RG:Z:SRR081708 NM:1:6 OQ:Z:GEGFFFGGGDGGGGDGA?

DCDD:GGGDGDCFGDFDDFFCCBEFBADBD-D:EEEE=D=DDDDC|

Figure 20

- ▶ 10:- Sequence
- ▶ 11:- Base Qualities
- ▶ This is the same as the fastq file; so if you have aligned data you can always go back and re-align
- ▶ The file may also have additional (optional) information recorded by the aligner or analysis tool

Sam and Bam

- ▶ sam is a human-readable file
 - ▶ which makes it quite large and unwieldy
- ▶ bam is the compressed binary version
 - ▶ needs special software to interrogate
 - ▶ better way of transferring data
- ▶ they contain *same* data
- ▶ the bam file needs to be *indexed* so we can access it more efficiently

- ▶ *NGS: QC and manipulation -> FASTQ Groomer*

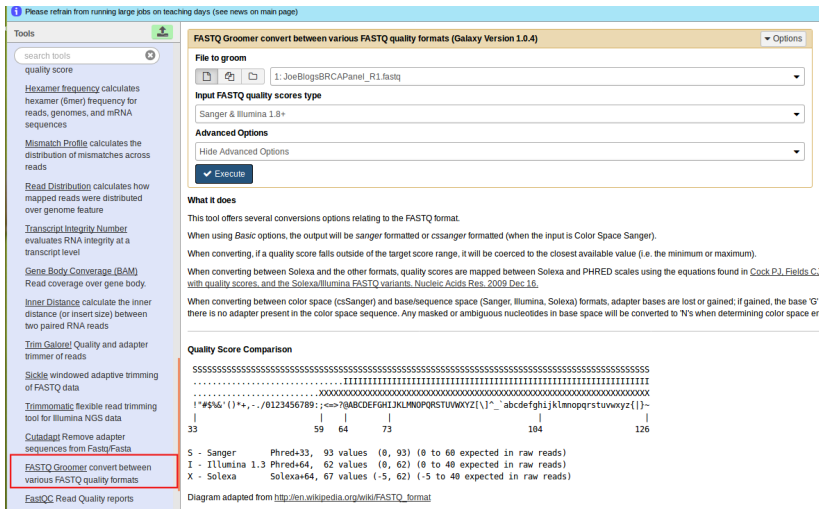


Figure 21

Exercise: Alignment

- ▶ Select file to groom as `JoeBlogsBRCAPanel_R1.fastq`
 - ▶ press **Execute**
- ▶ Repeat with `JoeBlogsBRCAPanel_R2.fastq`

Exercise: Alignment

► NGS: Mapping -> Bowtie2

Please refrain from running large jobs on teaching days (see news on main page)

Tools

search tools

Get Data

Send Data

Collection Operations

Text Manipulation

Filter and Sort

Join, Subtract and Group

NGS

NGS: QC and manipulation

NGS: Mapping

Map with BWA - map short reads (< 100 bp) against reference genome

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome

Bowtie2 - map reads against reference genome

Bowtie2 - map reads against reference genome (Galaxy Version 0.6)

Options

Is this single or paired library

Paired-end

FASTQ file #1

5: FASTQ Groomer on data 1

Must be of datatype "fastqsanger"

FASTQ file #2

6: FASTQ Groomer on data 2

Must be of datatype "fastqsanger"

Write unaligned reads (in fastq format) to separate file(s)

Yes No

--unl--un-conc. This triggers --un parameter for single reads and --un-conc for paired reads

Do you want to set paired-end options?

No

See "Alignment Options" section of Help below for information

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index




Figure 22

Exercise: Alignment

- ▶ In *Is this single-end or Paired-end?* Select **Paired-end**
- ▶ Set *FastQ file #1* and *FastQ file #2* to the two files you created in the previous step
- ▶ Make sure the reference genome is set to **Human (Homo sapiens)(b37):hg19**
- ▶ Press *Execute*
- ▶ Wait!

Exercise: Visualisation of reads




- Download the bam file you have just created, and it's index file

11: Bowtie2 on data 9 and data 8: aligned reads (sorted BAM)   

2.5 MB
format: **bam**, database: **hg19**

22354 reads; of these:
22354 (100.00%) were paired; of these:
212 (0.95%) aligned concordantly 0 times
21841 (97.71%) aligned concordantly exactly 1 time
301 (1.35%) aligned concordantly >1 times

212 pairs aligned concordantly

display at UCSC [main test](#)
display at Ensembl [Current](#)
display with IGV [local](#) [Human hg19](#)
display in IGB [View](#)
display at bam.iobio [bam.iobio.io](#)

Binary bam alignments file

Exercise: Visualisation of reads

- ▶ Load into **IGV**
- ▶ Navigate to the BRCA1 gene and zoom-in to see the reads
- ▶ Can you see any possible mutations?
- ▶ *Hover* over particular reads to get information about the alignment of the read
- ▶ (if you didn't manage to align the data, the file JoeBlogsBRCAPanel_bowtie2.bam can be used)