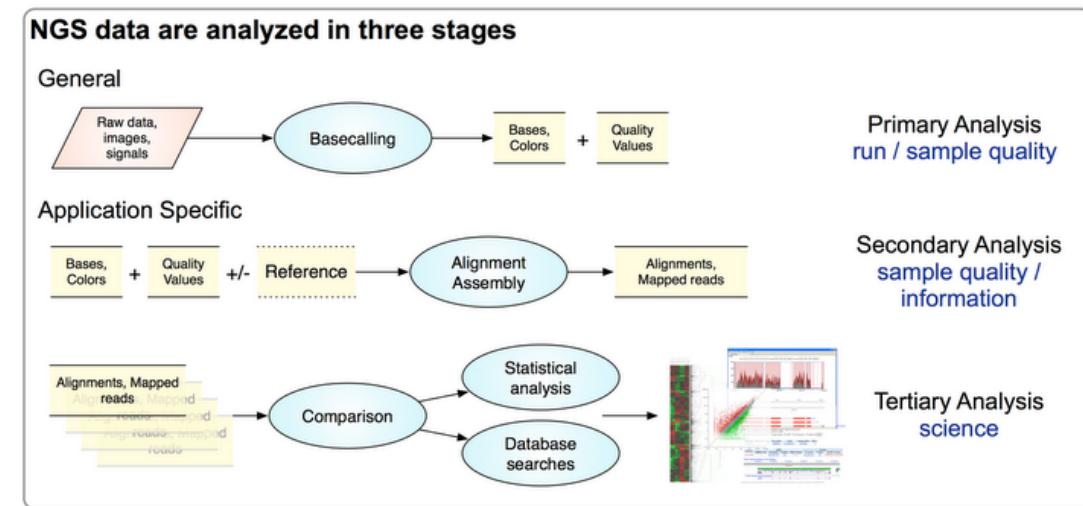


# BIOINFORMATICS COURSE

## Data formats

# Data formats

- There are many different types of file formats depending on
  - Type of information they contain
    - Raw Sequence files
    - Co-ordinate files
    - Parameter files
    - Annotation files
    - Metadata files
  - Sequencing platform
  - Analysis stage
  - Data source



# Data formats

- Formats are designed to hold sequence data and other information about sequence
- All Sequence formats are ASCII text containing sequence ID, Quality Scores, Annotation details, comments, and other descriptions about sequence
- FastA format (everybody knows about it)
  - Header line starts with “>” followed by a sequence ID
  - Sequence (string of nt).

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGYVLPWGQMSFWGATVITNLFSAI PYIGTNLV
EWIWGGFSVDKATLNRF FAFHFILEPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQVEYPYTIIGQM ASILYFSIILAFLPIAGX
IENY
```

# Data formats

## Fastq format

- Output of most actual sequencing platforms for raw data
- There is no standard file extension for a FASTQ file, but .fq and .fastq, are commonly used.
- FASTQ files are uncompressed and quite large because they contain the following information for every single sequencing read:
  1. @ followed by the read ID and possibly information about the sequencing run
  2. sequenced bases
  3. + (perhaps followed by the read ID again, or some other description)
  4. quality scores for each base of the sequence (ASCII-translated Phred scores)

```
@Seq description
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTT
+
! '' * ((((*++)) %%%++) (%%%%) . 1***-+*'') ) **55CCF>>>>CCCCCCCC65
```

# Data formats

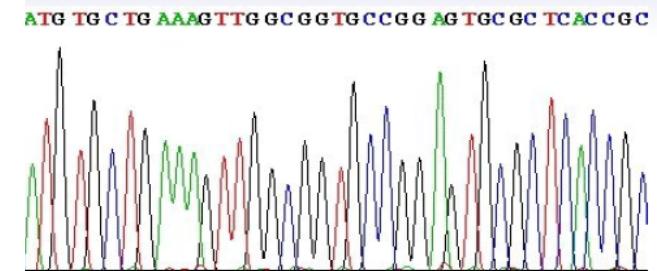
## Fastq format

### Phred Scores

- Sequencing systems assign quality scores to each peak, that represents the error probability that an individual base call is incorrect.
- Phred scores provide  $\log_{10}$ -transformed error probability values:

If  $p$  is probability that the base call is wrong the Phred score is

$$Q = -10 \cdot \log_{10} p$$



PHRED Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

- The base calling (A, T, G or C) is performed based on Phred scores.
- Ambiguous positions with Phred scores  $\leq 20$  are labeled with N.
- To assign each base a unique score identifier (instead of numbers of varying character length), Phred scores are typically represented as ASCII characters.

# Data formats

## Fastq format

- Different sequencing platforms may use different ASCII ranges for Phred encoding
- In Illumina 1.8, the quality scores have basically returned to the use of the Sanger format (Phred+33).

Description	ASCII characters		Quality score	
	Range	Offset	Type	Range
Solexa/early Illumina (1.0)	59 to 126 (; to `)	64	Solexa	-5 to 62
Illumina 1.3+	64 to 126 (@ to `)	64	Phred	0 to 62
Sanger standard/Illumina 1.8+	33 to 126 (! to `)	33	Phred	0 to 93

Base call quality scores are represented with the Phred range. Different Illumina (formerly Solexa) versions used different scores and ASCII offsets. Starting with Illumina format 1.8, the score now represents the standard Sanger/Phred format that is also used by other sequencing platforms and the sequencing archives.

# Data formats

# BAM/SAM formats

- The Sequence Alignment/Map (SAM) format is a generic nucleotide alignment format that describes the alignment of sequencing reads to a reference.
  - SAM files typically contain a short header section with information about the genomic loci of each read and a very long alignment section where each row represents a single read alignment. For each read, there are 11 mandatory fields that always appear in the same order:

<QNAME> <FLAG> <RNAME> <POS> <MAPQ> <CIGAR> <MRNM> <MPOS> <ISIZE> <SEQ>  
<QUAL>

- BAM is a dedicated binary format including the compressed SAM. It enables fast access to data without having to “unzip” the whole file. For the typically large data, BAM is currently the most recommended and most “standard” format.

```
@HD VN:1.0
@SQ SN:chr20 LN:62435964
@RG ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_28833_29006_6945 99 chr20 28833 20 10M1D25M = 28993 195 \
    AGCTTAGCTAGCTACCTATCTGGTCTGGCCG <<<<<<<<<<<<:<9/,&,22;;<<< \
        NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M      = 28701 -168 \
    ACCTATATCTTGGCCTTGGCCGATGC GG CTTGCA <<<<;<<<7;:<<<6; <<<<<<<<<7<<< \
        MF:i:18 RG:Z:L2
```

# Data formats

## BAM/SAM formats

### Mandatory Alignment Section Fields

Position	Field	Description
1	QNAME	Query template (or read) name
2	FLAG	Information about read mapping (see next section)
3	RNAME	Reference sequence name. This should match a @SQ line in the header.
4	POS	1-based leftmost mapping position of the first matching base. Set as 0 for an unmapped read without coordinate.
5	MAPQ	Mapping quality of the alignment. Based on base qualities of the mapped read.
6	CIGAR	Detailed information about the alignment (see relevant section).
7	RNEXT	Used for paired end reads. Reference sequence name of the next read. Set to “=” if the next segment has the same name.
8	PNEXT	Used for paired end reads. Position of the next read.
9	TLEN	Observed template length. Used for paired end reads and is defined by the length of the reference aligned to.
10	SEQ	The sequence of the aligned read.
11	QUAL	ASCII of base quality plus 33 (same as the quality string in the Sanger FASTQ format).
12	OPT	Optional fields (see relevant section).

# Data formats

## BAM/SAM formats

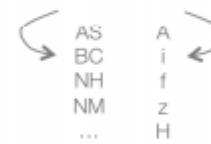
@HD	VN:	
@SQ	SN:	LN:
@RG	ID:	SM:
@PG	ID:	
@CO		

(theoretically) optional  
 HEADER SECTION  
 general information about the file

1	2	3	4	5	6	7	8	9	10	11	>11
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT

Paired read?  
 Unmapped?  
 Mapped to rev.  
 strand?  
 1<sup>st</sup> in pair?  
 2<sup>nd</sup> in pair?  
 Failed QC?  
 ...

M (mis)match  
 I insertion  
 D deletion  
 N skipped  
 S soft clipped  
 H hard clipped  
 P padding

<TAG>:<TYPE>:<VALUE>  

 AS  
 BC  
 NH  
 NM  
 ...

ALIGNMENT  
SECTION  
1 line per locus

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT

# Data formats

## Formats for genome annotations (BED, GFF)

- One line per genomic feature
- The BED format is the simplest way to store annotation tracks. It has three required fields (chromosome, start, end) and up to 9 optional fields (name, score, strand, thickStart, thickEnd, itemRgb, blockCount, blockSizes, blockStarts).

```
# 6-column BED file defining transcript loci
chr1 66999824 67210768 NM_032291 0 +
chr1 33546713 33586132 NM_052998 0 +
chr1 25071759 25170815 NM_013943 0 +
chr1 48998526 50489626 NM_032785 0 -
```

- The General Feature Format (GFF) has nine required fields; the first three fields form the basic name, start, end tuple that allows for the identification of the location in respect to the reference genome.

```
##gff-version 3
ctg123 . operon      1300 15000  .  +  .  ID=operon001;Name=superOperon
ctg123 . mRNA        1300 9000   .  +  .  ID=mRNA0001;Parent=operon001;Name=sonichedgedehog
ctg123 . exon         1300 1500   .  +  .  Parent=mRNA0001
ctg123 . exon         1050 1500   .  +  .  Parent=mRNA0001
ctg123 . exon         3000 3902   .  +  .  Parent=mRNA0001
ctg123 . exon         5000 5500   .  +  .  Parent=mRNA0001
ctg123 . exon         7000 8000   .  +  .  Parent=mRNA0001
```

# Data formats

## Formats for genome annotations (BED, GFF)

1. **reference sequence:** coordinate system of the annotation (e.g., "Chr1")
2. **source:** describes how the annotation was derived (e.g., the name of the annotation software)
3. **method:** annotation type (e.g., gene)
4. **start position:** 1-based integer, always less than or equal to the stop position
5. **stop position:** for zero-length features, such as insertion sites, start equals end and the implied site is to the right of the indicated base
6. **score:** e.g., sequence identity
7. **strand:** "+" for the forward strand, "-" for the reverse strand, or "." for annotations that are not stranded
8. **phase:** codon phase for annotations linked to proteins; 0, 1, or 2, indicating the frame, or the number of bases that should be removed from the beginning of this feature to reach the first base of the next codon
9. **group:** contains the class and ID of an annotation which is the logical parent of the current one ("feature is composed of")

# Data formats

So, in summary...

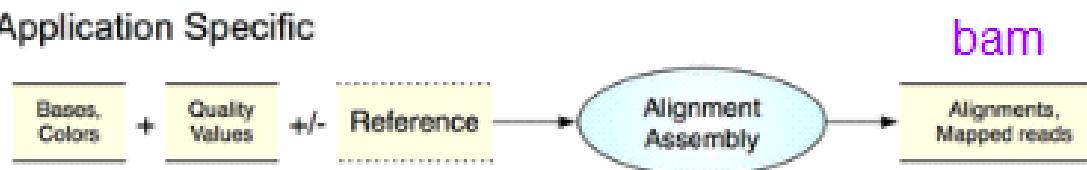
## NGS data are analyzed in three stages

### General



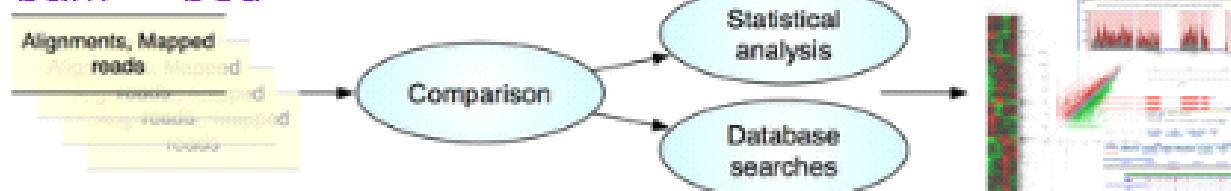
Primary Analysis  
run / sample quality

### Application Specific



Secondary Analysis  
sample quality / information

### bam → bed



Tertiary Analysis  
science

- For base-call data, FASTQ (Sanger, Phred)
- For read alignments, SAM/BAM/MAQ format
- For annotation results, GFF or BED format

# BIOINFORMATICS COURSE

## Quality control and Preprocessing of NGS data

# Quality Control

## FastQC tool

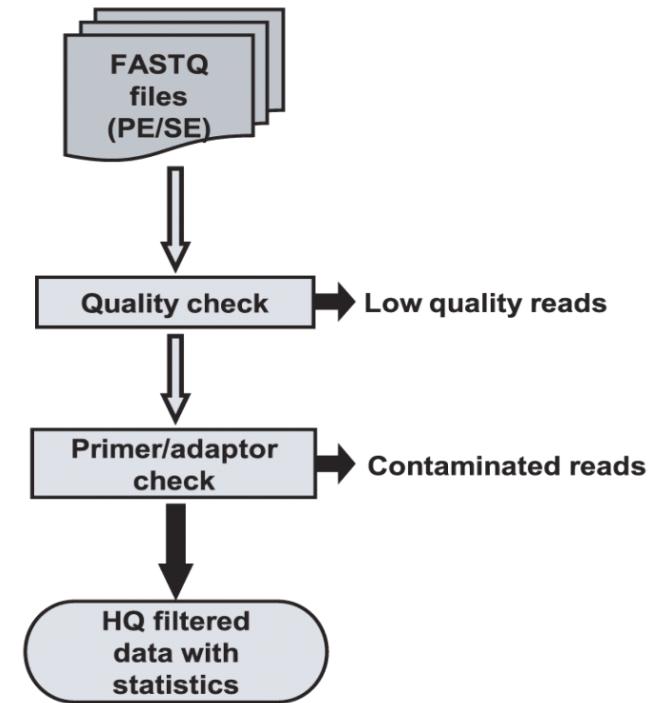
<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

- Basic statistics
- Quality- Per base position
- Per Sequence Quality Distribution
- Nucleotide content per position
- Per sequence GC distribution
- Per base GC distribution
- Per base N content
- Length Distribution
- Overrepresented/ duplicated sequences
- K-mer content

# Quality Control

QC analysis of sequence data is extremely important for meaningful downstream analysis

- To analyze problems in quality scores/ statistics of sequencing data
- To check whether further analysis with sequence is possible
- To remove redundancy (filtering)
- To remove low quality reads from analysis
- To remove adapter contamination



Highly efficient and fast processing tools are required to handle large volume of datasets

# Quality Control

## FastQC tool

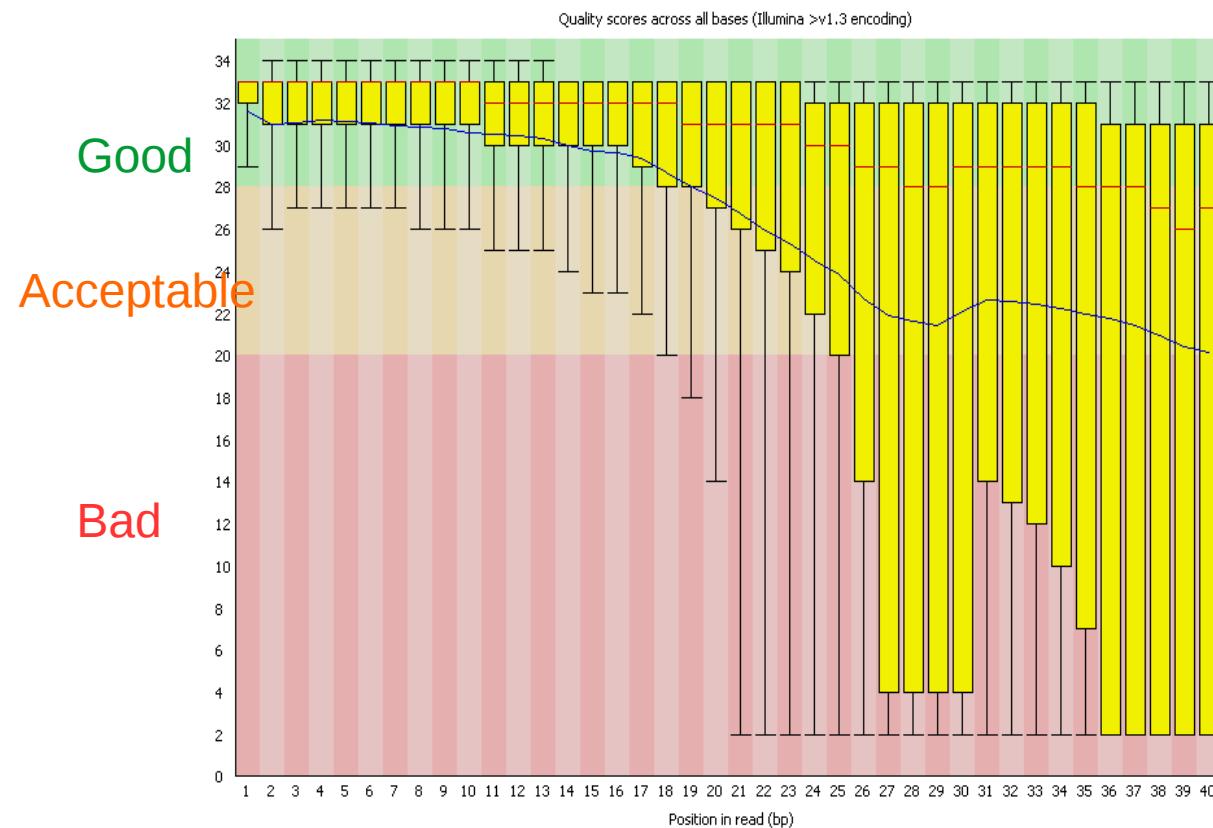
<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

- Basic statistics
- Quality- Per base position
- Per Sequence Quality Distribution
- Nucleotide content per position
- Per sequence GC distribution
- Per base GC distribution
- Per base N content
- Length Distribution
- Overrepresented/ duplicated sequences
- K-mer content

# Quality Control

## FastQC

### Per base sequence quality (Boxplot)

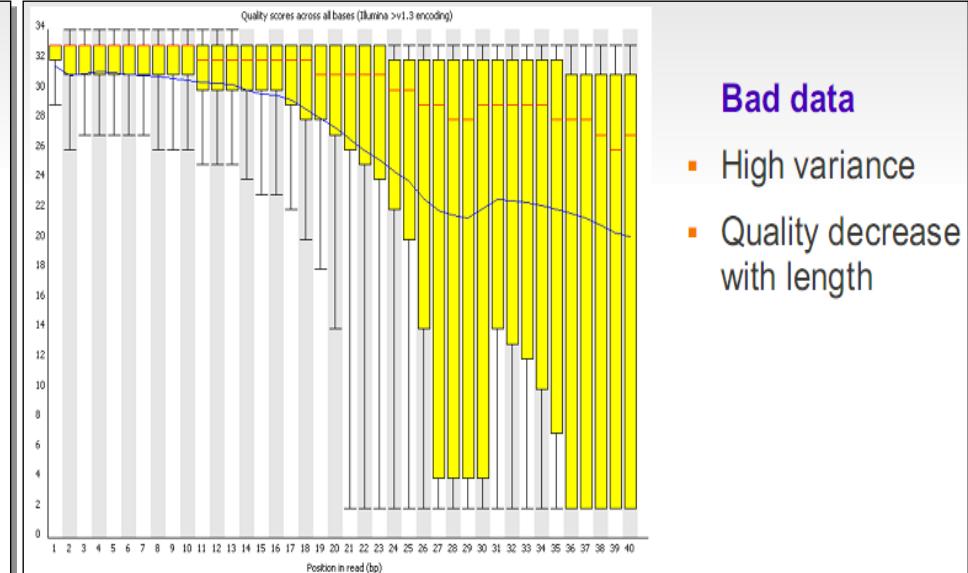
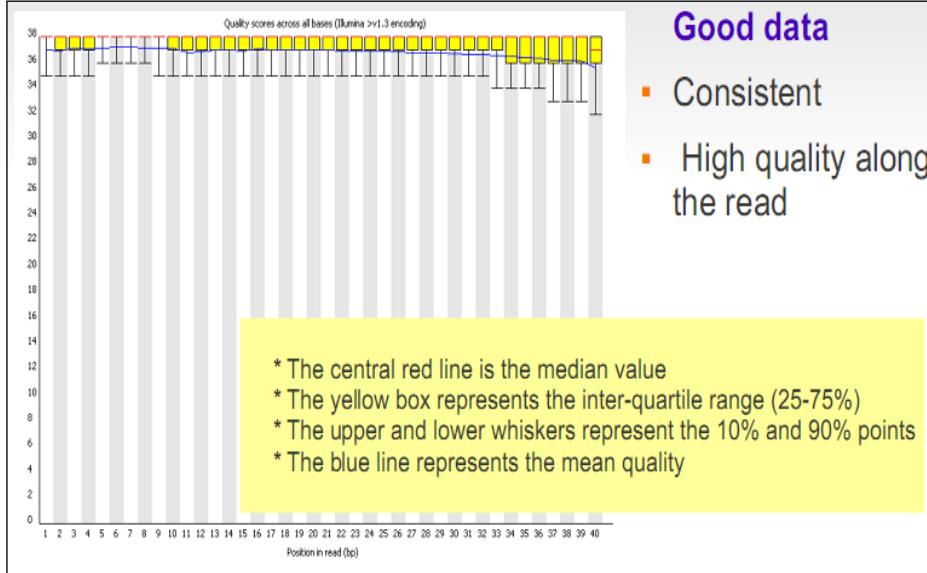


shows an overview of the range of quality values across all bases at each position in the FastQ file

# Quality Control

## FastQC

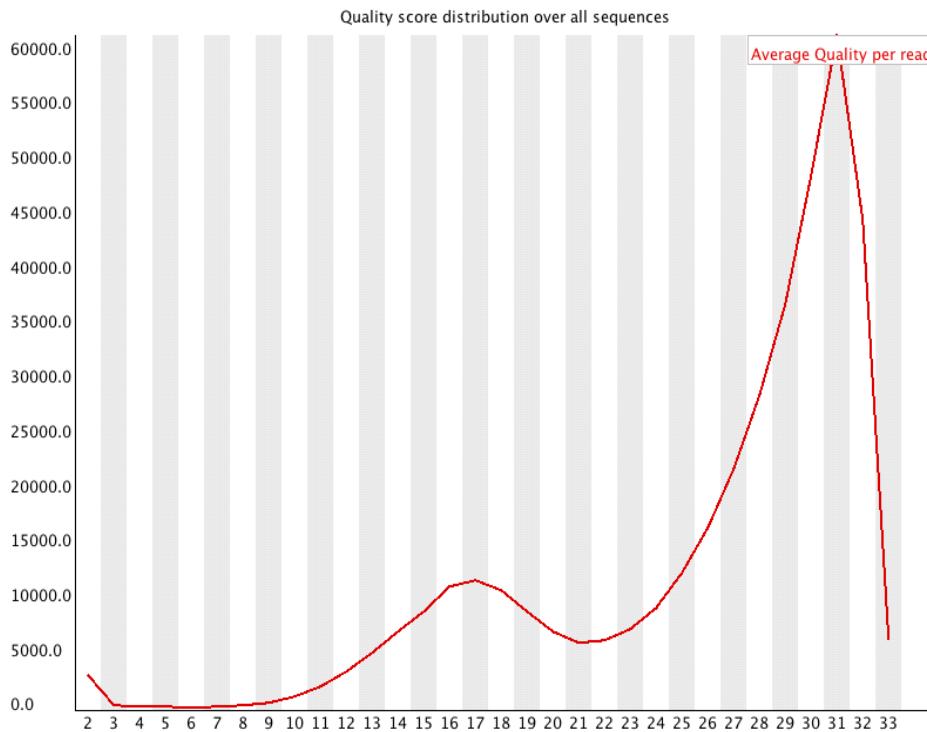
### Per base sequence quality (Boxplot)



# Quality Control

## FastQC

### Per sequence quality scores

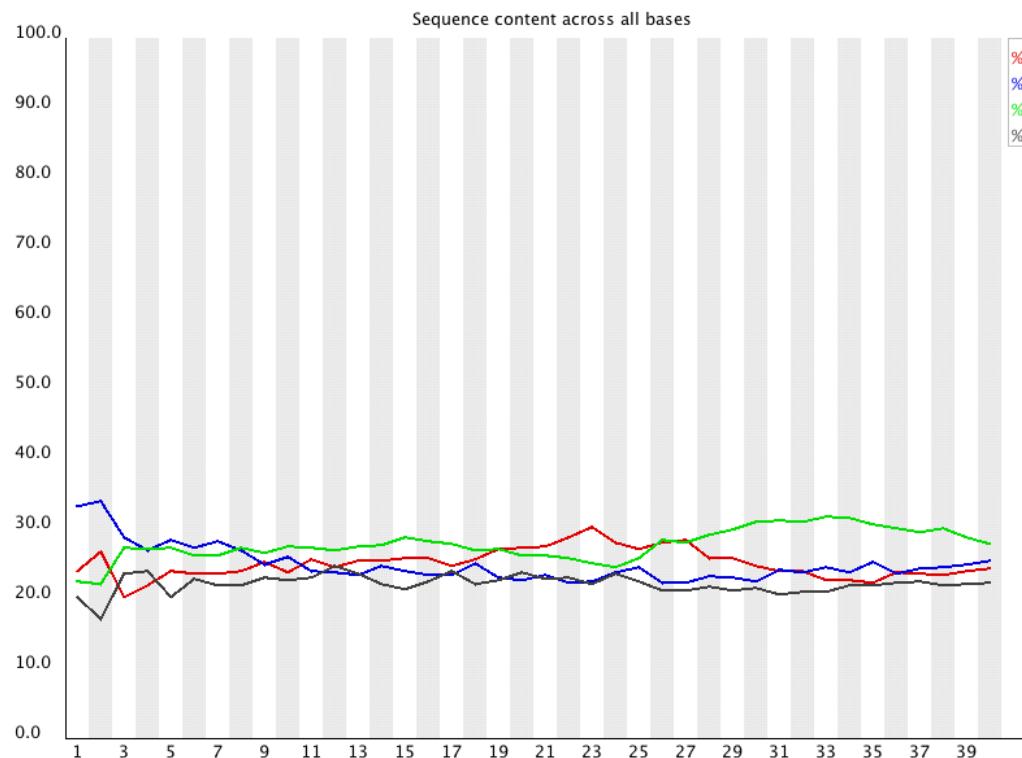


allows you to see if a subset of your sequences have universally low quality values.

# Quality Control

## FastQC

### Per base sequence content



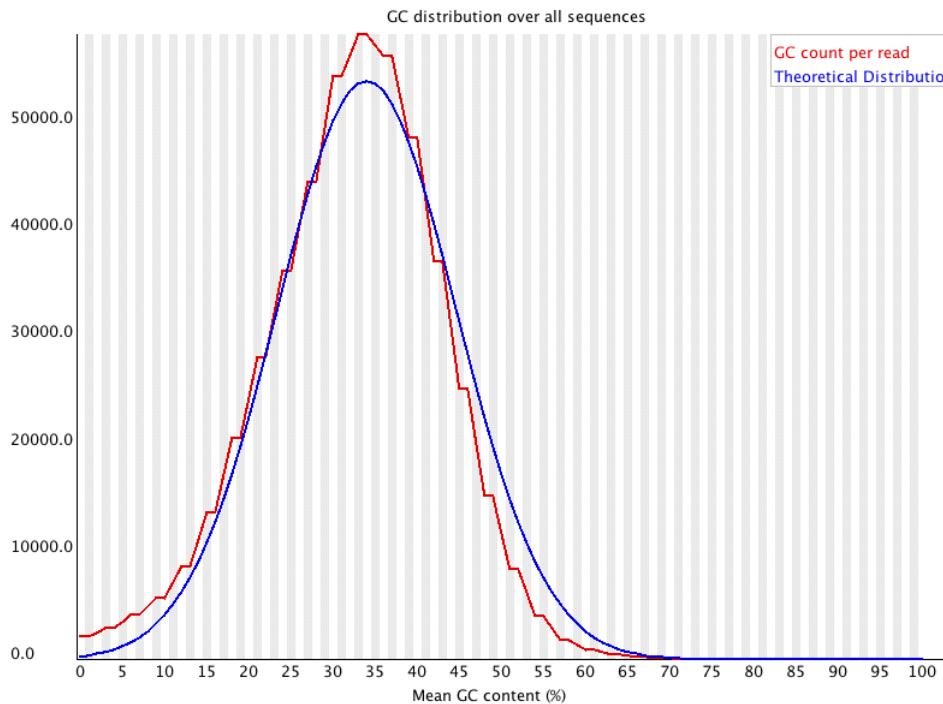
Per Base Sequence Content plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called

- In a random library you would expect that there would be little to no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other.

# Quality Control

## FastQC

### Per sequence GC content



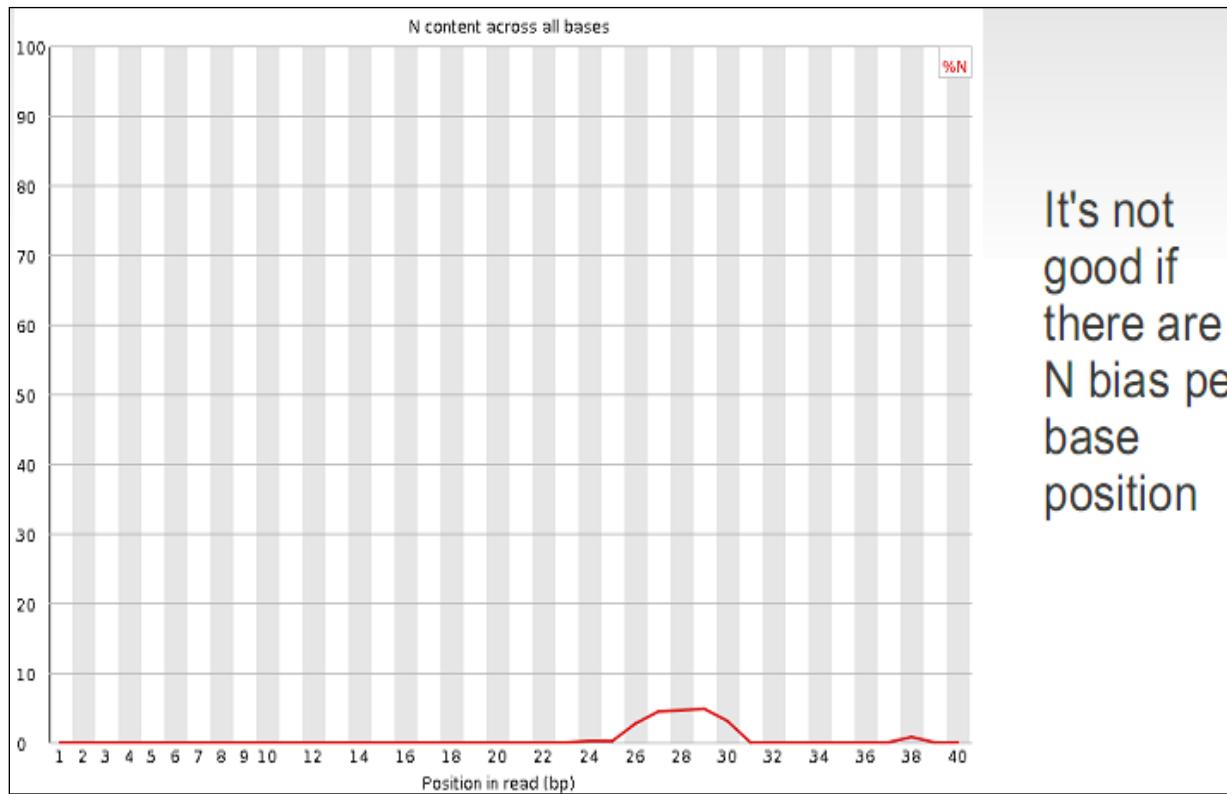
measures the GC content across the whole length of each sequence in a file and compares it to a modelled normal distribution of GC content

- you would expect to see a roughly normal distribution of GC content
- An unusually shaped distribution could indicate a contaminated library or some other kinds of biased subset.

# Quality Control

## FastQC

### Per base N content



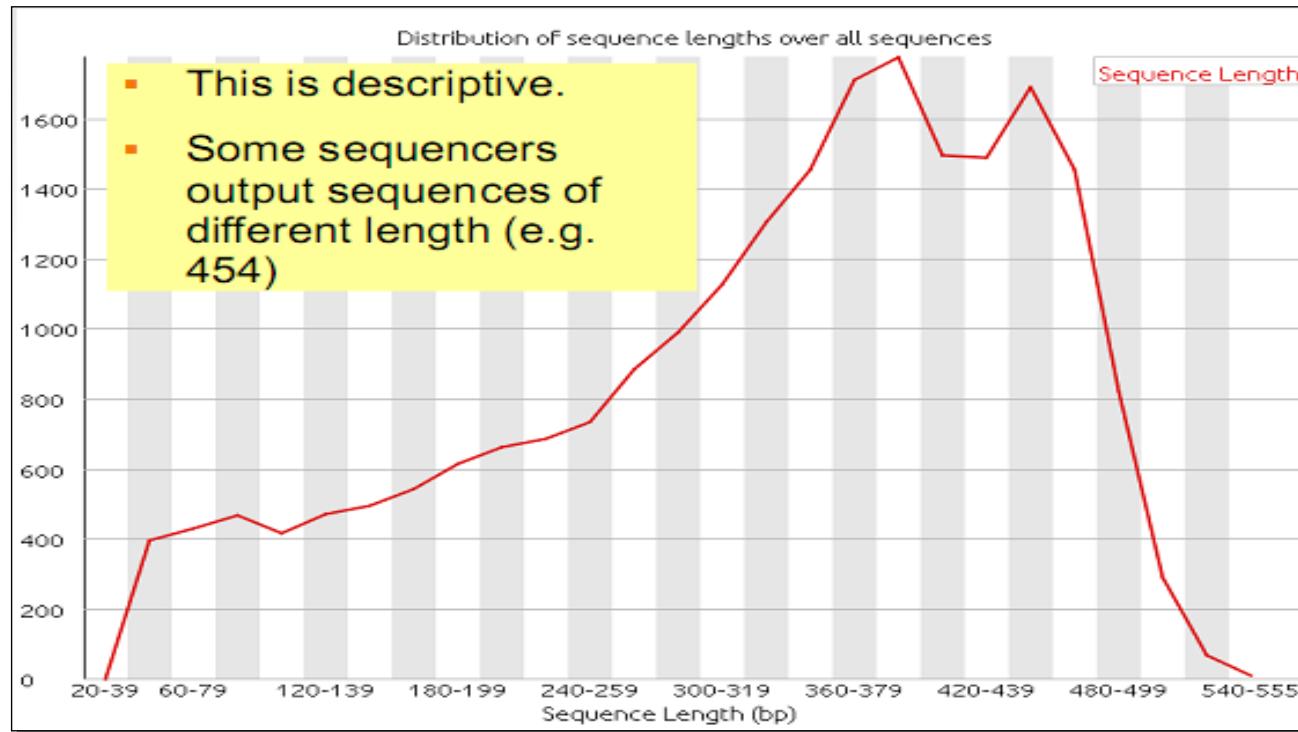
It's not good if there are N bias per base position

If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base call. It plots out the percentage of base calls at each position for which an N was called.

# Quality Control

## FastQC

### Sequence length distribution

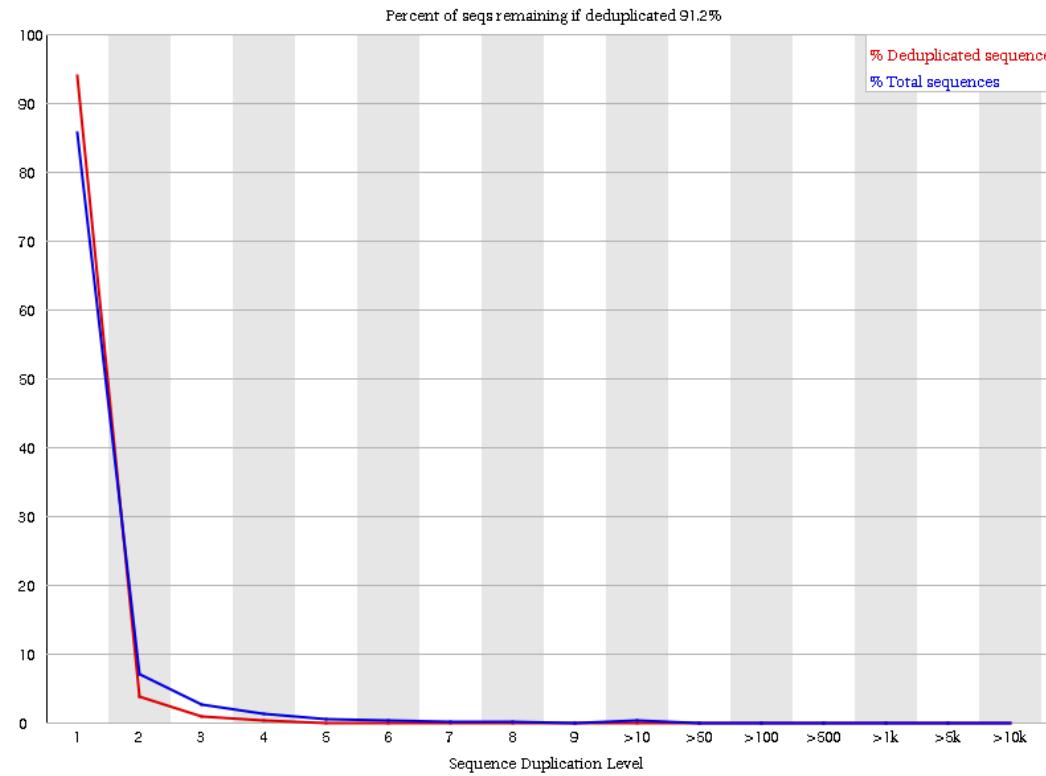


In many cases it will produce a simple graph showing a peak only at one size, but for variable length FASTQ files, it will show the relative amounts of each different size of sequence fragment.

# Quality Control

## FastQC

### Sequence duplication level



Counts the degree of duplication for every sequence. Too many duplicate regions in the sequence may indicate contamination or technical problems

# Quality Control

## FastQC

### Overrepresented sequences

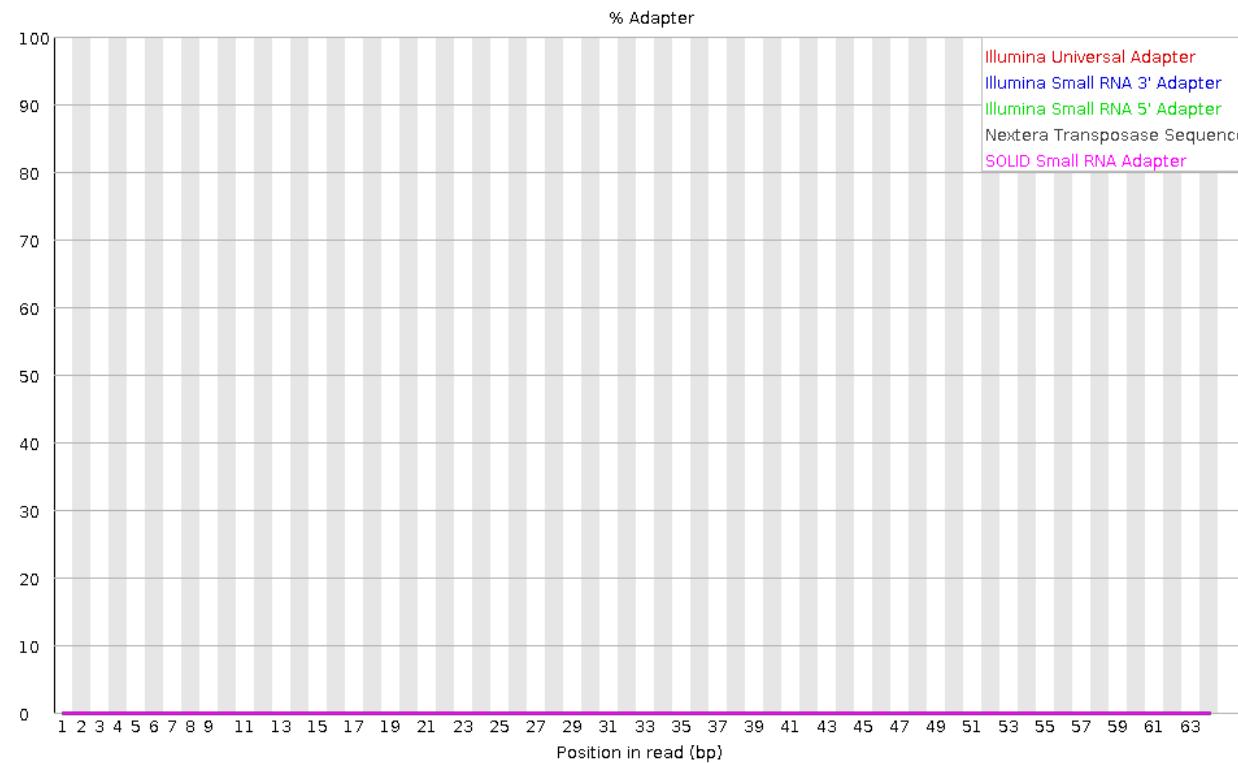
Sequence	Count	Percentage	Possible Source
AAGATCCGAGTCGTCCGGAAATCATTGCCGTGTTCTCACAGTTATTAA	432	0.43585733743631133	No Hit
AGATCCGAGTCGTCCGGAAATCATTGCCGTGTTCTCACAGTTATTAAAC	335	0.33799122231750994	No Hit
TGGCAGAACAGTAGAGCAGAAGAAGAAGCGGACCTTCCGCAAGTTCACCTAC	250	0.25223225546082834	No Hit
CAGAAGTAGAGCAGAAGAAGAAGCGGACCTTCCGCAAGTTCACCTACCGC	237	0.23911617817686526	No Hit
GTAGAGCAGAAGAAGAAGCGGACCTTCCGCAAGTTCACCTACCGCGGCGT	223	0.22499117187105888	No Hit
AAGAAATCTGACCCGGTCGTCTCGTACCGCGAGACGGTCAGTGAAGAGTC	204	0.2058215204560359	No Hit
AAGTAGAGCAGAAGAAGAAGCGGACCTTCCGCAAGTTCACCTACCGCGGC	151	0.1523482822983403	No Hit
CACCTGGAGATCTGCCTGAAGGACCTGGAGGAGGACCACGCCCTGCATCCC	147	0.14831256621096706	No Hit
TCTGCCTGAAGGACCTGGAGGAGGACCACGCCCTGCATCCCCATCAAGAAA	146	0.14730363718912376	No Hit

Lists all of the sequence which make up more than 0.1% of the total. Finding that a single sequence is very overrepresented in the set either means that is highly biologically significant, or that the library is contaminated. For each overrepresented sequence it will look for matches in a database of common contaminants.

# Quality Control

## FastQC

### Adapter content



Does a generic analysis of all the Kmers in the library to find those that don't have even coverage through the length of the reads.

# Quality Control

## FastQC

- Good (Illumina) quality data:

[http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good\\_sequence\\_short\\_fastqc.html](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html)

- Bad (Illumina) quality data:

[http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad\\_sequence\\_fastqc.html](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html)

# Preprocessing raw data

# Preprocessing raw data

Removing technical artifacts

Adjusting biases

etc

# Preprocessing raw data

Based on the information provided by the QC graphs, the sequences may be treated to reduce bias in downstream analysis by:

- Filtering sequences
  - with low mean quality score
  - too short
  - with too many ambiguous (N) bases
  - based on their GC content
- Cutting/Trimming/masking sequences
  - from low quality score regions
  - beginning/end of sequence
  - removing adapters, primers

But, removing duplicate reads is not advised since high expressed genes can have genuine duplicate reads that are not due to the PCR amplification step.

# Preprocessing raw data

After preprocessing, run QC again.

If too bad quality, it may be advisable to rerun the experiment...

# Introduction to Galaxy

*A web-based genome analysis platform*

# Galaxy

- An open, web-based platform integrating many popular tools and resources for intensive biomedical research.
- **What can be done?**
  - Obtain data from many data sources like UCSC Table Browser, Biomart, WormBase, or your own data
  - Prepare data for further analysis by rearranging or cutting data columns, filtering data and many other options
  - Analyze data by finding overlapping regions, determining statistics, preprocessing NGS data and much more
  - Share data and workflows

<https://usegalaxy.org/>

# Galaxy Interface

The Galaxy page is divided into three panels:

Tools for uploading,  
processing and  
analysis

Viewing panel  
(menus, data, results)

Login

[Register](#)

History of analysis  
steps and datasets

The screenshot illustrates the Galaxy web interface with three main panels:

- Tools Panel (Left):** A sidebar containing a search bar and a list of tool categories, including "Get Data", "Send Data", "Lift-Over", "Collection Operations", "Text Manipulation", "Datamash", "Convert Formats", "Filter and Sort", "Join, Subtract and Group", "Fetch Alignments/Sequences", "NGS: QC and manipulation", "NGS: DeepTools", "NGS: Mapping", "NGS: RNA Analysis", "NGS: SAMtools", "NGS: BamTools", "NGS: Picard", "NGS: VCF Manipulation", "NGS: Peak Calling", "NGS: Variant Analysis", "NGS: RNA Structure", "NGS: Du Novo", "NGS: Gemini", "NGS: Assembly", "NGS: Chromosome Conformation", "NGS: Mothur", "Operate on Genomic Intervals", "Statistics", "Graph/Display Data", "Phenotype Association", "BEDTools", "Genome Diversity", "EMBOSS", "Regional Variation", "FASTA manipulation", "Multiple Alignments", and "Metagenomic Analysis".
- Viewing Panel (Center):** The main content area featuring:
  - A banner: "Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#). You can install your own Galaxy by following the [tutorial](#) and choose from thousands of tools from the [Tool Shed](#).
  - A "Running Your Own Understanding how Galaxy works" section with a "Main" button.
  - A "Tweets" section showing tweets from @galaxyproject and @denbiOffice.
  - A "History" section showing a message: "This history is empty. You can load your own data or get data from an external source".
- History Panel (Right):** A panel titled "History" showing a search bar and a message: "This history is empty. You can load your own data or get data from an external source".

Logos for Penn State, Johns Hopkins University, Oregon Health & Science University, TACC, and CYVERSE are visible at the bottom of the central panel.

# Galaxy Interface



The screenshot shows the Galaxy web interface with a sidebar on the left containing a list of tools categorized under 'Tools'. A red box highlights the 'Get Data' section, which includes links to 'Upload File', 'UCSC Main table browser', 'UCSC Archaea table browser', 'Get Microbial Data', 'BioMart Central server', 'GrameneMart Central server', 'Flymine server', 'EuPathDB server', 'EncodeDB at NHGRI', and 'EpiGRAPH server'. Below this, other sections like 'Send Data', 'ENCODE Tools', 'Lift-Over', 'Text Manipulation', 'Convert Formats', 'FASTA manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Graph/Display Data', 'Regional Variation', 'Multiple regression', 'Evolution', 'Metagenomic analyses', and 'EMBOSS' are listed. At the bottom, there are sections for 'NGS TOOLBOX BETA', 'NGS: QC and manipulation', 'NGS: Mapping', and 'NGS: SAM Tools'.

## Tools for data analysis

### Get Data

- From databases (UCSC Table Browser, ...)
- From uploaded files
- From urls

### Text manipulation

### Filter and Sort

### Operate on Genomic Intervals

### FASTA manipulation

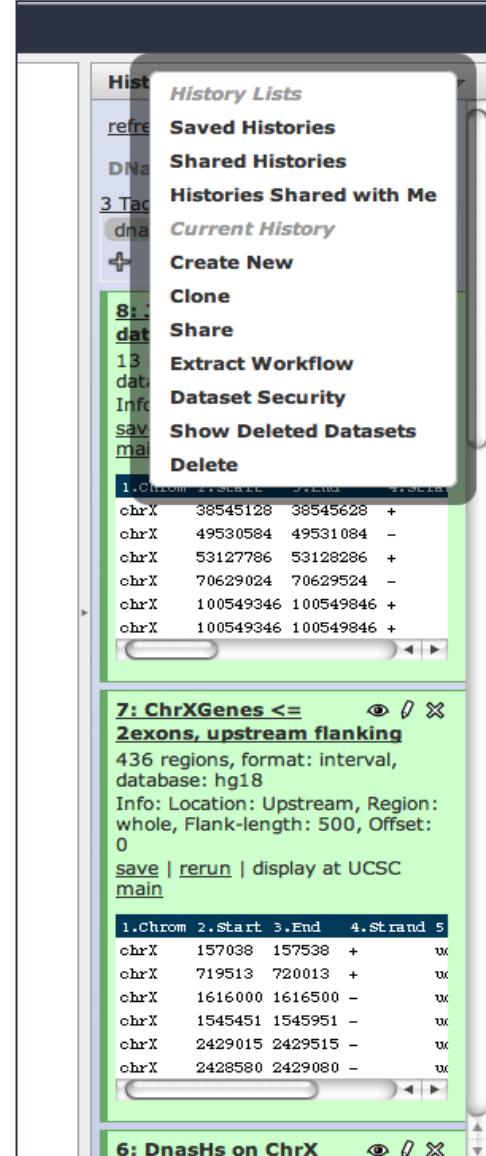
### NGS analysis

- QC
- Fastq file pre-processing
- Read Alignment / Mapping
- SAM tools

# Galaxy Interface

## Histories

List saved histories and shared histories.  
Work on Current History, create new, clone, share, create workflow, set permissions, show deleted datasets or delete history.



The screenshot shows the Galaxy web interface. On the left, a sidebar titled "Histories" lists "Saved Histories" and "Shared Histories". A context menu is open over a history named "8: ChrXGenes <= 2exons, upstream flanking". The menu options include "History Lists", "Saved Histories", "Shared Histories", "Histories Shared with Me", "Current History", "Create New", "Clone", "Share", "Extract Workflow", "Dataset Security", "Show Deleted Datasets", and "Delete".

**8: ChrXGenes <= 2exons, upstream flanking**

436 regions, format: interval, database: hg18

Info: Location: Upstream, Region: whole, Flank-length: 500, Offset: 0

[save](#) | [rerun](#) | [display at UCSC main](#)

1.Chrom	2.Start	3.End	4.Strand	5
chrX	157038	157538	+	ux
chrX	719513	720013	+	ux
chrX	1616000	1616500	-	ux
chrX	1545451	1545951	-	ux
chrX	2429015	2429515	-	ux
chrX	2428580	2429080	-	ux

**7: ChrXGenes <= 2exons, upstream flanking**

436 regions, format: interval, database: hg18

Info: Location: Upstream, Region: whole, Flank-length: 500, Offset: 0

[save](#) | [rerun](#) | [display at UCSC main](#)

1.Chrom	2.Start	3.End	4.Strand	5
chrX	157038	157538	+	ux
chrX	719513	720013	+	ux
chrX	1616000	1616500	-	ux
chrX	1545451	1545951	-	ux
chrX	2429015	2429515	-	ux
chrX	2428580	2429080	-	ux

**6: DnasHs on ChrX**

# Galaxy Interface

## Workflows

Galaxy Analyze Data Workflow Visualize Shared Data Help User Using 2%

Galaxy will be down for six hours beginning at 2:30 PM UTC, Tuesday, November 20 for filesystem maintenance.

Tools

search tools

Inputs

Get Data

Send Data

Lift-Over

Collection Operations

Text Manipulation

Datamash

Convert Formats

Filter and Sort

Join, Subtract and Group

Fetch Alignments/Sequences

NGS: QC and manipulation

NGS: DeepTools

NGS: Mapping

NGS: RNA Analysis

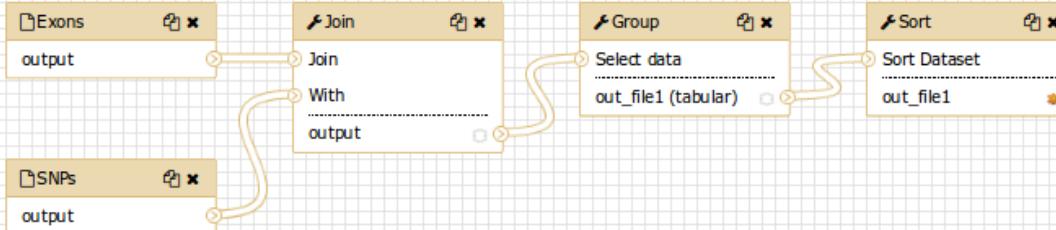
NGS: SAMtools

NGS: BamTools

NGS: Picard

NGS: VCF Manipulation

Workflow Canvas | Coding Exon SNPs



Details

**Edit Workflow Attributes**

**Name:** Coding Exon SNPs

**Version:** Version 1, 5 steps (active)

**Tags:**  Apply tags to make it easy to search for and find items with the same tag.

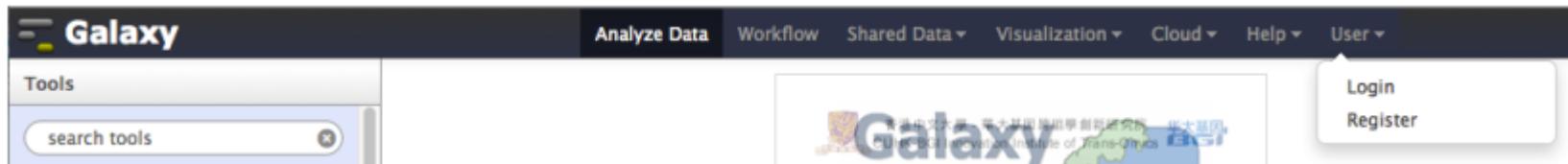
**Annotation / Notes:**  
*Describe or add notes to workflow*  
 Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.

Workflows with all the analysis steps, allows user to repeat analysis using different datasets

# Practicum

## Register for a Galaxy account

This will only take a moment, and will allow all the work that you do to persist between sessions and allow you to name, save, share, and publish Galaxy histories, workflows, datasets and pages. It allows you to store up to 250GB of data on this public server.



# Practicum

## Importing data into Galaxy

1. From database queries (eg. UCSC): obtain a BED-formatted dataset of all RefSeq genes from platypus.

Get Data > UCSC Main – Table Browser tool  
 Set genome, RefSeq Genes, and BED output format (send to Galaxy)

Genomes    Genome Browser    Tools    Mirrors    Downloads    My Data    Help    About Us

**Table Browser**

This program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve the browser for a description of the controls in this form, and the User's Guide for general information and sample queries. For more biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Download

de: Mammal    genome: Platypus    assembly: Feb. 2007 (ASM227v2/ornAna2)

up: Genes and Gene Predictions    track: RefSeq Genes    add custom tracks    track hubs

le: refGene    describe table schema

ion:  genome  position chrX5:870777-1056769    lookup    define regions

ntifiers (names/accessions): paste list    upload list

er: create

intersection: create

relation: create

put format: **BED - browser extensible data**     Send output to  Galaxy  GREAT  GenomeSpace

put file: (leave blank to keep output in browser)

type returned:  plain text  gzip compressed

et output    summary/statistics

**Output refGene as BED**

Include [custom track header](#):  
 name= tb\_refGene  
 description= table browser query on refGene  
 visibility= pack  
 url=

Create one BED record per:  
 Whole Gene  
 Upstream by 200 bases  
 Exons plus 0 bases at each end  
 Introns plus 0 bases at each end  
 5' UTR Exons  
 Coding Exons  
 3' UTR Exons  
 Downstream by 200 bases

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream

Send query to Galaxy  
 Cancel

# Practicum

## Importing data into Galaxy

### 2. From a File on your computer / FTP file:

Get Data > Upload File

**Download from web or upload from disk**

Regular    Composite    Collection    Rule-based

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
New File	72 b	fastqsang...	----- Additional Sp...		0%

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.  
[http://chagall.med.cornell.edu/galaxy/rnaseq/GM12878\\_rnaseq1.fastqsanger](http://chagall.med.cornell.edu/galaxy/rnaseq/GM12878_rnaseq1.fastqsanger)

Type (set all):

Genome (set all):

# Practicum

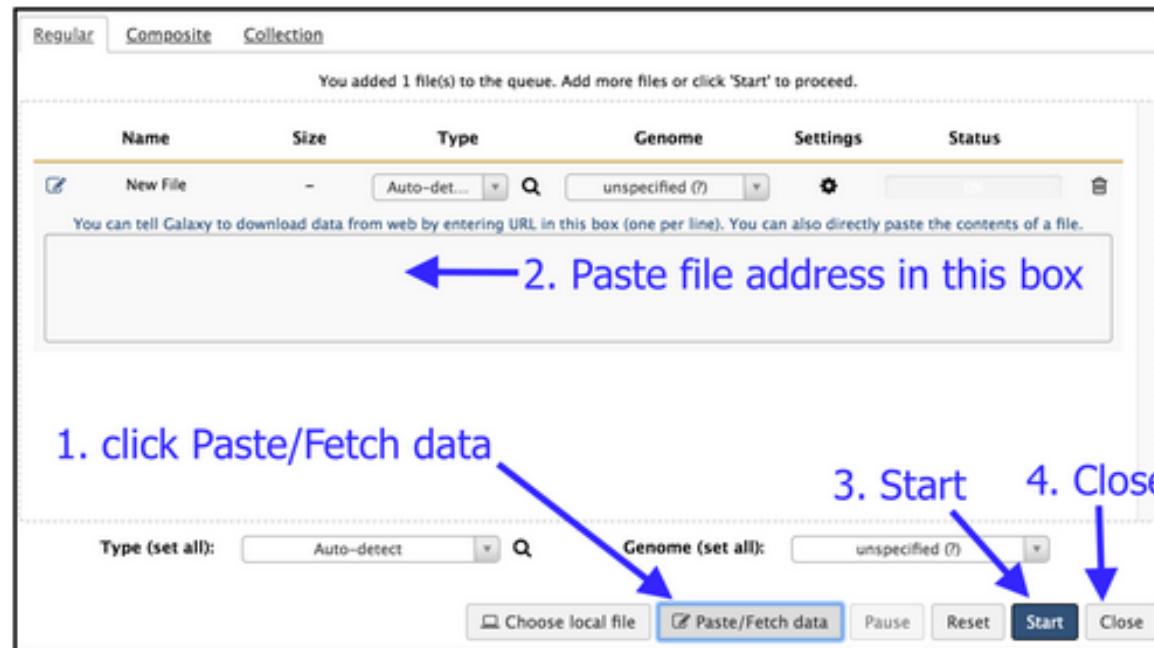
## Importing data into Galaxy

### 3. From a website:

Get Data > Upload File

Copy this URL into the text-entry box:

url: [https://zenodo.org/record/582600/files/mutant\\_R1.fastq](https://zenodo.org/record/582600/files/mutant_R1.fastq)



# Practicum

## Importing data into Galaxy

### 4. From the Sequence Read Archive (EBI/NCBI): eg. SRR925743

NCBI SRA Tools > Extract reads in FASTQ format

Extract reads in FASTQ/A format from NCBI SRA. (Galaxy Version 2.6.2)

**select input type**

SRR accession

**SRR accession**

Must start with SRR, DRR or ERR, e.g. SRR925743 , ERR343809

**select output format**

fastq

Advanced Options

Execute



This tool extracts reads from SRA archives using fastq-dump. The fastq-dump program is developed at NCBI, and is available at <http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software>.

NB: Single-end or pair-end collections may be empty if given SRRs LibraryLayout contains only either SINGLE or PAIRED results. Browse the NCBI SRA for SRR accessions at <http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=studies>.

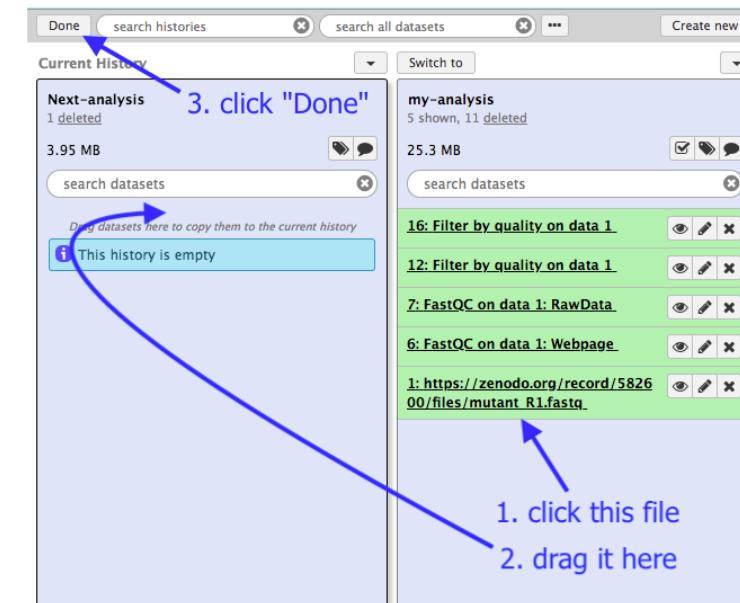
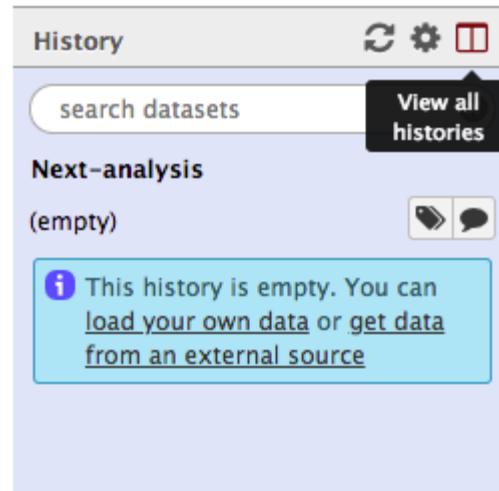
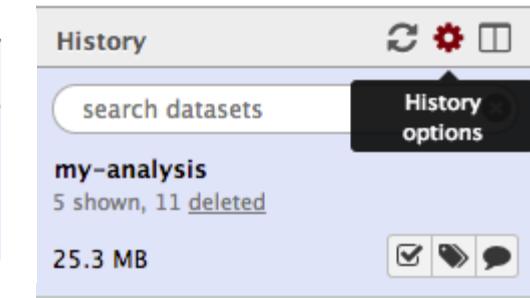
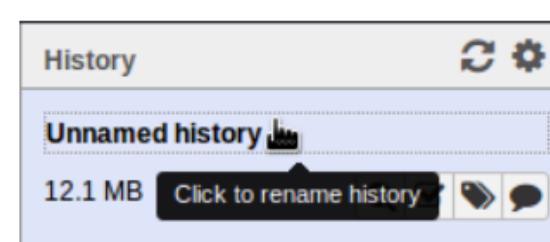
Galaxy tool wrapper originally written by Matt Shirley (mdshw5 at gmail.com).

Wrapper modified by Philio Mabon ( philio.mabon at phac-aspc.ca ).

# Practicum

## Managing histories

- Name your current history
- Create new history and rename it (eg. “pract1”)
- Manage datasets and histories:
  - View all histories
  - Drag files between histories (**new history must be set to current**)



# Practicum

## Visualizing the dataset

- You can view file content clicking the eye icon in history.
- The mutant\_R1.fastq file contains DNA sequencing reads from a bacteria, in FASTQ format:

```
@mutant-no_snps.gff-24960/1                               read 1 sequence
AATGTTGTCACTGGATTCAAATGACATTAACTAAATTCTAATTATTCAATGAATCGAACTAGTAGCAGAAATGCAATGAG
+
5??A9?BBBDDDBEDDBFF+FGHHIIHHHEIHHIIHIIAHDHIIHIG#IIHIFHHHFGIII*IHHHIHFIIHGIC1
@mutant-no_snps.gff-24958/1
CAAAGTCGTTGGTCATATAAAAAACCGCGTACAGTCAACTATAGATAACAATCAAGATAAAACTCATGCACAGATTG
+
?A????@?DDDABDE9FGGGFGICFHIIIBGHIIIGICHHIFH=IHAFIHIHHHHIFCIIEIHAIFGIHIDDIHE
@mutant-no_snps.gff-24956/1
TATAAATTCAACTTGCAACAGAACCATCTAATCTTCAACAAACTGGCCCGTTGTTGAACTACTCTTTAATAAA
+
?????BBADD5DDDDDGFGCFEECFBBICIII,IIHIICHIIIFHHHHHIIIIIIIAHHIHHH5FHDHHHH
```

The screenshot shows a bioinformatics analysis interface. On the left, a 'History' panel lists datasets: 'my-analysis' (1 shown, 3.95 MB). Below it, a specific dataset is detailed: '1: https://zenodo.org/reCORD/582600/files/mutant\_R1.fastq'. This entry includes a preview section with an eye icon, a pencil icon for editing, and a delete icon. A 'View data' button is also present. The main area of the interface displays the content of the 'mutant\_R1.fastq' file, showing DNA sequencing reads in FASTQ format. The first few lines of the sequence are visible, including the header '@mutant-no\_snps.gff-24960/1', the sequence itself (AATGTTGTCACTGGATTCAAATGACATTAACTAAATTCTAATTATTCAATGAATCGAACTAGTAGCAGAAATGCAATGAG), and the quality scores (5??A9?BBBDDDBEDDBFF+FGHHIIHHHEIHHIIHIIAHDHIIHIG#IIHIFHHHFGIII\*IHHHIHFIIHGIC1).

# Practicum

## Update dataset attributes

The attributes of the dataset can be updated by clicking the pencil icon next to the dataset.

- Set a name (eg. “pract1”)
- Check the format of your Fastq dataset (Data Type)\*
- Add notes (eg. “url”)
- You can also associate a genome to the dataset if you know the information

**Edit attributes**

**Name**

**Info**

**Annotation**  
  
Add an annotation or notes to a dataset; annotations are available when a history is viewed.

**Database/Build**

**\* Warning!**: different data formats may have quality scores in different ranges. When changing between data formats, use the tool “NGS: QC and manipulate > FASTQ Groomer”. This will ensure that quality scores are correctly rescaled between formats.

Description Galaxy format name	ASCII characters		Quality score	
	Range	Offset	Type	Range
Sanger standard/Illumina 1.7+ fastqsanger	33 to 126	33	PHRED	0 to 93
Solexa/early Illumina fastqsolexa	59 to 126	64	Solexa	-5 to 62
Illumina 1.3+ fastqillumina	64 to 126	64	PHRED	0 to 62

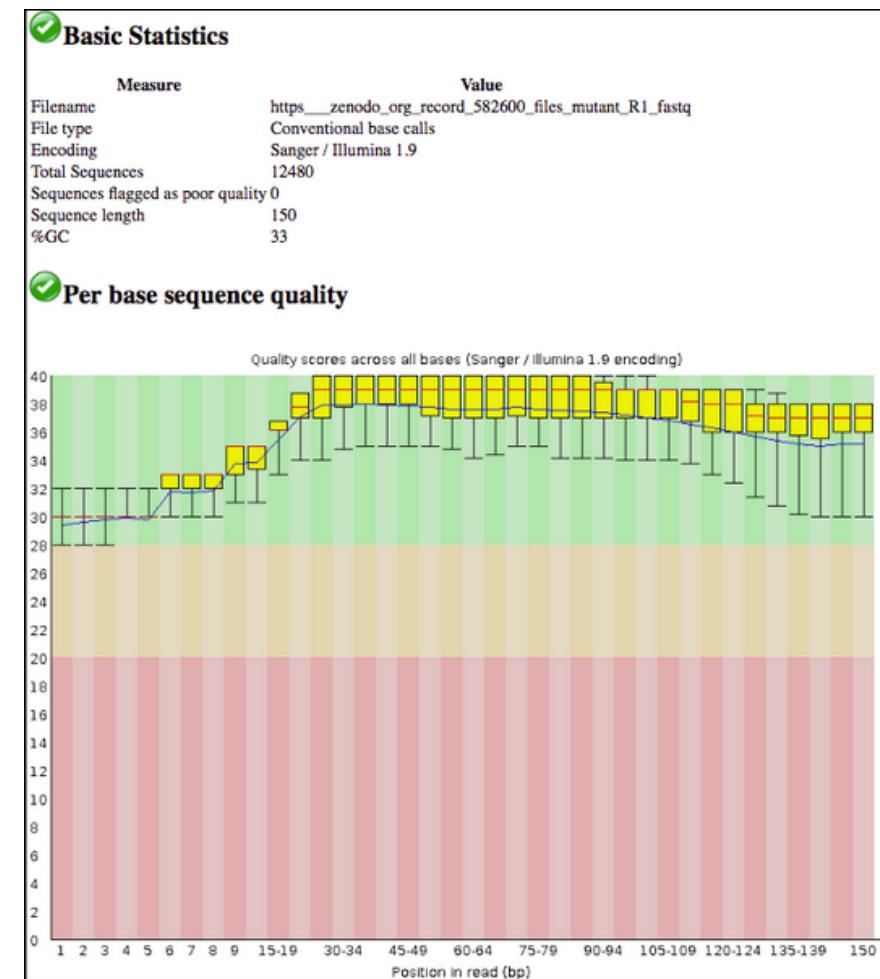
Most tools that accept FASTQ data expect it to be in a specific FASTQ version: **.fastqsanger**

# Practicum

## Quality Control of Raw Data

To do FastQC in this dataset run the tool “NGS: QC and manipulation” > “FastQC”

- Look at the output file called FastQC on data 1: Webpage (click on eye icon).
- What was the length of the reads in the input FASTQ file?
- Do these reads have higher quality scores in the centre or at the ends?



# Practicum

## Preprocessing of Raw Data

Filter out lower-quality reads from our FASTQ file using the tool NGS: QC > Filter by quality with the following parameters:

- Quality cutoff value: 35
  - %bases in sequence with  $\geq$ quality cutoff: 80
- 
- How many reads have been discarded?

History

search datasets

my-analysis

4 shown, 8 deleted

24.99 MB

12: Filter by quality on data 1

3.4 MB

format: fastqsanger, database: ?

Quality cut-off: 35  
Minimum percentage: 80  
Input: 12480 reads.  
Output: 10694 reads.  
discarded 1786 (14%) low-quality reads.

@mutant-no\_snps.gff-24960/1  
AATGTTGTCACTGGATTCAAATGACATTAAATCTA  
+  
5??A9?BBBDDDBEDDBFF+FGHHIIHHHEIHHIIIA  
@mutant-no\_snps.gff-24958/1

7: FastQC on data 1: Raw Data

6: FastQC on data 1: Web page

1: https://zenodo.org/record/582600/files/mutant\_R1.fastq

# Practicum

## Create workflow from history

- From history options: Export workflow

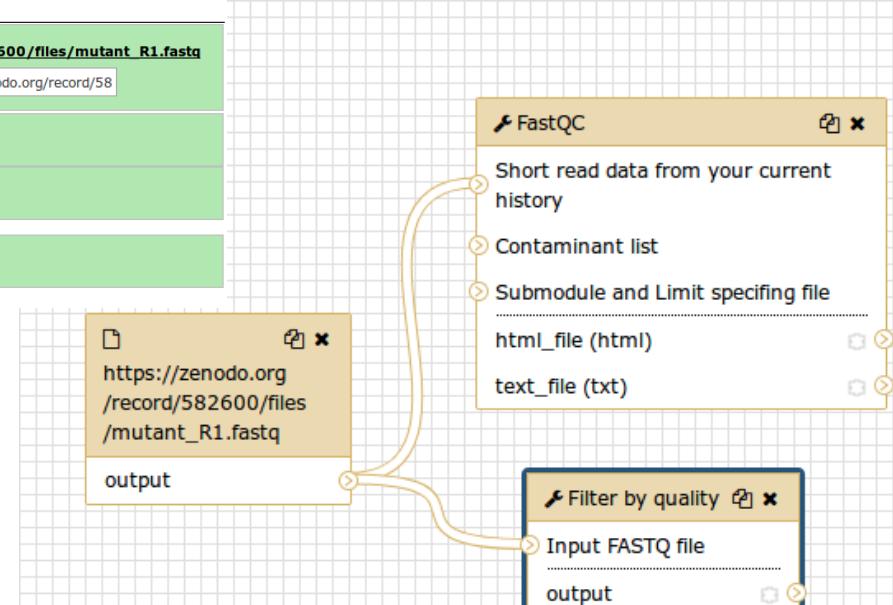
The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

**Workflow name**  
Workflow constructed from history 'prova'

**Create Workflow** **Check all** **Uncheck all**

Tool	History Items created
Upload File <small>This tool cannot be used in workflows</small>	1 <a href="https://zenodo.org/record/582600/files/mutant_R1.fastq">https://zenodo.org/record/582600/files/mutant_R1.fastq</a> <input checked="" type="checkbox"/> Treat as input dataset <a href="https://zenodo.org/record/582600">https://zenodo.org/record/582600</a>
FastQC <input checked="" type="checkbox"/> Include "FastQC" in workflow	2 FastQC on data 1: Webpage
Filter by quality <input checked="" type="checkbox"/> Include "Filter by quality" in workflow	3 FastQC on data 1: RawData 4 Filter by quality on data 1



**FastQC**  
Short read data from your current history  
Contaminant list  
Submodule and Limit specifying file

**html\_file (html)**  
**text\_file (txt)**

**Filter by quality**  
Input FASTQ file  
Output

```

graph LR
    A[https://zenodo.org/record/582600/files/mutant_R1.fastq] --> B[FastQC]
    A --> C[Filter by quality]
    B --> D[html_file (html)]
    B --> E[text_file (txt)]
    C --> F[Input FASTQ file]
    F --> G[Output]
  
```

# Practicum

## Your turn!

1. Create a new history and name it “pract2”
2. Upload “raw\_child-ds-1.fq” and “raw\_child-ds-2.fq” files from your USB
3. Update the attributes of the two datasets
  - Change the names of the datasets to “sample-f” and “sample-r”, respectively.
  - Check data type is set to “fastqsanger”
  - Associate the dataset with the human hg38 genome in the Database/Build field.
4. Run a quality control on each dataset using the FastQC tool.
  - What can we say about the quality of these datasets?

# Practicum

## Your turn!

5. Trim the reads in each dataset using the tool “NGS: QC and manipulation > FastQ Trimmer” tool. Set the parameters:

- Determine from the boxplot and FastQC figures where the quality of the reads begins to drop off sharply.
- Calculate how many bases have to be trimmed from the end and use that number as the Offset from 3' end.

6. Using the tool “NGS: QC and manipulation > Filter by quality”, filter out all sequences with  $\geq 80\%$  bases that have a quality less than 20 (use as input the trimmed sequences).

- How many sequences do you have left in your dataset?

7. Re-run FastQC on the quality-controlled data, and inspect the new FastQC report

- Has the sequence quality been improved?

8. Convert your analysis history into a workflow

# Practicum

## Solution

Galaxy

Analyze Data Workflow Visualize Shared Data Help User Using 2%

Tools search tools

quality formats

Filter FASTQ reads by quality score and length

Combine FASTA and QUAL into FASTQ

Trim Galore! Quality and adapter trimmer of reads

Select high quality segments

Build base quality distribution

Collapse sequences

FASTQ Trimmer by column

FASTQ Quality Trimmer by sliding window

Convert SOLiD output to fastq

Compute quality statistics for SOLiD data

Draw quality score boxplot for SOLiD data

FASTQ Trimmer by column (Galaxy Version 1.0.0)

Options

FASTQ File

2: raw\_child-ds-1.fq  
1: raw\_child-ds-2.fq

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Define Base Offsets as

Absolute Values

Use Absolute for fixed length reads (Illumina, SOLiD)  
Use Percentage for variable length reads (Roche/454)

Offset from 5' end

0

Values start at 0, increasing from the left

Offset from 3' end

41

Values start at 0, increasing from the right

Keep reads with zero length

Yes No

Execute

This tool allows you to trim the ends of reads.

History

search datasets

prova2

6 shown, 9 deleted

5.71 GB

15: FastQC on data 2: RawData

14: FastQC on data 2: Webpage

5: FastQC on data 1: RawData

4: FastQC on data 1: Webpage

2: raw\_child-ds-1.fq

1: raw\_child-ds-2.fq

# Practicum

## MultiQC tool

- MultiQC tool allows to summarize multiple QC reports at once, though FastQC needs to be run on each dataset first.

The screenshot shows the MultiQC tool's configuration interface. At the top, it asks "Which tool was used generate logs?" with "FastQC" selected. Below that, under "FastQC output", it asks "Type of FastQC output?" with "Raw data" selected. A note states: "↓ MultiQC does not accept the HTML report generated by FastQC, only the Raw Data". A list of FastQC outputs is shown, with the first item highlighted: "14: FastQC on data 4: RawData". An arrow points from this highlighted item to a callout box on the right. The callout box contains the text: "Select FASTQC\_RawData outputs for the datasets to be included". Other items in the list include: 13: FastQC on data 4: Webpage, 12: FastQC on data 3: RawData, 11: FastQC on data 3: Webpage, 10: FastQC on data 2: RawData, 9: FastQC on data 2: Webpage, 8: FastQC on data 1: RawData, 7: FastQC on data 1: Webpage, and 4: [https://zenodo.ora/record/583613/files/sample2-rfq.oz\\_\(as.fq.gz\)](https://zenodo.ora/record/583613/files/sample2-rfq.oz_(as.fq.gz)). At the bottom, there are buttons for "Report title" (set to "multiQC"), "Custom comment" (empty), "Output the multiQC log file?" (radio buttons for "Yes" and "No" with "No" selected), and "Execute".

Select *FASTQC\_RawData* outputs for the datasets to be included

# Practicum

## MultiQC tool

General Statistics				
Sample Name	% Dups	% GC	Length	M Seqs
sample1-f	53.2%	44%	233 bp	0.0
sample1-r	55.1%	44%	233 bp	0.0
sample2-f	66.3%	44%	251 bp	0.1
sample2-r	65.7%	44%	251 bp	0.1

### FastQC

FastQC is a quality control tool for high throughput sequence data, written by Simon Andrews at the Babraham Institute in Cambridge.

#### Sequence Quality Histograms

1 1 2

The mean quality value across each base position in the read. See the [FastQC help](#).

Y-Limits:  on

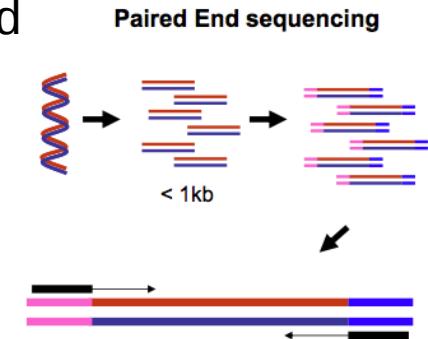


# Practicum

## Processing paired-end data

**Paired-end data:** a single physical piece of DNA/RNA is sequenced from two ends and so generates two reads. These can be represented as separate files (two fastq files with first and second reads) or a single file where reads for each end are interleaved.

- The datasets we used (child1 and child2) are paired-end data, the first representing the forward (f) and the second the reverse (r)



## Two single files

File 1

@M02286:19:00000000-AA549:1:1101:12677:1273 1:N:0:23  
CCTACGGTGGCAGTCAGTGAGGAATTGGTCAATGGACCGGAAGTCT  
+  
ABC8C,:@F:CE8,B-,C,-6-9-C,CE9-CC--C-<-C++,,+;CE  
@M02286:19:00000000-AA549:1:1101:15048:1299 1:N:0:23  
CCTACGGTGGCTGCACTGAGGAATTGGACAATGGTCGGAAGACT  
+  
ABC@CC77CFCEG;F9<F89<9--C,CE,--C-6C-,CE:++7:,CF

File 2

@M02286:19:00000000-AA549:1:1101:12677:1273 2:N:0:23  
CACTACCCGTATCTAACCTGTTGATACCGCACCTCGAGCTTA  
+  
--8A,CCE+, , ;,<CE@,CFD,,C,CFF+@+@CCEF,,B+C,  
@M02286:19:00000000-AA549:1:1101:15048:1299 2:N:0:23  
CACTACCGGGTATCTAACCTGTTCGCTCCCCACGCTTCGTCCATC  
+  
-6AC,EE@::CF7CFF<<FFGGDFFF,@FGGGG?F7FEGGGDEFF>FF

## Interleaved file

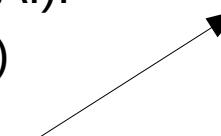
# Practicum

## Alignment/Mapping of paired-end data

- Mapping of NGS reads against reference sequences is one of the key steps of the analysis.
- Map datasets uploaded before using “NGS: Mapping > Map with BWA-MEM” tool against hg38 *canonical* version\* of the human genome (your datasets are *mitochondrial* DNA!).
- Set paired data (child1 (f) / child2 (r))

\*hg38 versions:

- hg38 contains all chromosomes as well as all unplaced contigs.
- hg38 canonical does not contain unplaced sequences and only consists of chromosomes 1 through 22, X, Y, and mitochondrial DNA.
- hg38 canonical female contains everything from the canonical set with the exception of chromosome Y.



Map with BWA-MEM – map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.7.15.1)

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

Use a built-in genome index

Built-ins were indexed using default options. See ‘Indexes’ section of help below

Using reference genome

Human (Homo sapiens) (b38): hg38

Select genome from the list

Single or Paired-end reads

Paired

Select between paired and single end data

Select first set of reads

4: sample2-r.fq.gz

Specify dataset with forward reads

Select second set of reads

4: sample2-r.fq.gz

Specify dataset with reverse reads

Set read groups information?

Set read groups (SAM/BAM specification)

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

- Note that as output we get a unique BAM file

# Practicum

## Your turn!

- When preprocessing paired-end data coming from separate files, this must be taken into account so that reads are preprocessed “in pairs”
  - After the trimming and filtering that we did before, how many reads did you get for each dataset as final output?
  - What happens if you try to map these datasets as paired-end data?
- Do the preprocessing again starting from the original raw data sets, but this time taking into account that they are paired-end data [child1=sample-f (forward), child2=sample-r (reverse)]. To accomplish this task use the tool “NGS: QC and manipulation > Cutadapt” (combines trimming and quality control). Set the following parameters:
  - Paired-end
  - File 1: child1/sample-f (forward)
  - File 2: child2/sample-r (reverse)
  - Filter options: minimum length=20
  - Read modification options: Quality cutoff=20
  - Output options: Report=yes
- How many datasets do we get? How many bp in total do we get?
- Perform the alignment with the output of Cutadapt

# Practicum

## Which coding exon has the highest number of single nucleotide polymorphisms (SNPs) on human chromosome 22?

First retrieve all coding exons and SNPs in chromosome 22

- Create a new history. Call it “SNPs in coding exons”.
- Next retrieve all coding exons in chromosome 22 from UCSC – Table browser:
  1. Use [Get Data](#) → [UCSC Main](#) tool.
  2. Change the position to search to chr22.
  3. Click the get Output button.
  4. On the output format screen, change it to one BED record per coding exon.
  5. Click Send query to Galaxy.
  6. Rename this exon dataset to “Exons\_22”.
- Now retrieve the SNP data, again from UCSC.
  1. Use [Get Data](#) → [UCSC Main](#) tool.
  2. This time, change the group to Variation and Repeats.
  3. Change the position to chr22 again.
  4. On the output format page, the one BED record per gene should again be selected. In this case, “gene” is actually “SNP” (or “feature”).
  5. Click the Send query to Galaxy button.
  6. Rename this dataset to “SNPs\_22”.

Genomes    Genome Browser    Tools    Mirrors    Downloads    My Data    Help    About Us

## Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and this application see [Using the Table Browser](#) for a description of the controls in this form, and the [User's Guide](#) for general information. You may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation, you may want to use [GenomeSpace](#) for use with diverse computational tools. Refer to the [Credits](#) page for the list of contributors and usage information. Data can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

**clade:** Mammal    **genome:** Human    **assembly:** Dec. 2013 (GRCh38/hg38)

**group:** Genes and Gene Predictions    **track:** NCBI RefSeq    [add custom tracks](#)    [track hubs](#)

**table:** UCSC RefSeq (refGene)    [describe table schema](#)

**region:**  genome  position chr22    [lookup](#)    [define regions](#)

**identifiers (names/acccessions):** [paste list](#)    [upload list](#)

**filter:** [create](#)

**subtrack merge:** [create](#)

**intersection:** [create](#)

**correlation:** [create](#)

**output format:** BED - browser extensible data     Send output to  Galaxy  GREAT  GenomeSpace

**output file:**  (leave blank to keep output in browser)

**file type returned:**  plain text  gzip compressed

[get output](#)    [summary/statistics](#)

To reset **all** user cart settings (including custom tracks), [click here](#).

# Practicum

**Which coding exon has the highest number of single nucleotide polymorphisms (SNPs) on human chromosome 22?**

- Now we need to associate the SNPs with their exons. Use the database join command to associate with an exon any SNP whose given genomic interval overlaps that of any exon by at least one base:
  - Open the [Operate on Genomics Intervals](#) → Join tool.
  - Use the exons dataset as the first dataset and the SNPs dataset as the second dataset.
  - Click Execute.
  - View the results

Note that the first 6 columns correspond to exons and the next 6 to SNPs positions

# Practicum

**Which coding exon has the highest number of single nucleotide polymorphisms (SNPs) on human chromosome 22?**

- Next, we want to count the number of SNPs associated with each exon:
  1. Open the **Join, Subtract, and Group** → **Group** tool.
  2. Change the Group by Column to column 4 (i.e., the exon name).
  3. Add a new operation.
  4. Change the type of the new operation to Count and the column to col 4. This groups the dataset on the exon name, and count how many exons of the same name are in each unique group.

The resulting dataset will have two columns: one is a list of all the distinct exon names, and the other is the number of times that exon name was associated with a SNP.

# Practicum

**Which coding exon has the highest number of single nucleotide polymorphisms (SNPs) on human chromosome 22?**

- We can now rearrange this dataset such that the exons with the highest number of SNPs are at the top.
  1. Open the Filter and Sort → Sort tool.
  2. Sort the dataset by column 2 (SNP count), in descending order (i.e., highest first).

Which exon has the highest number of SNPs?

- Finally, we want to select the five exons with the highest numbers of SNPs.
  1. Open the Text Manipulation → Select First tool.
  2. Make sure the sorted dataset is selected.
  3. Change the number of lines to select to 5.
  4. Click Execute.

1	2
NM_001136213_cds_0_0_chr22_15690078_f	40
NM_006071_cds_0_0_chr22_46256561_r	31
NM_001005239_cds_0_0_chr22_15528159_f	24
NM_138433_cds_0_0_chr22_50546244_f	24
NM_173566_cds_5_0_chr22_31712083_r	15

# References

## Bibliography

- Goecks J, Nekrutenko A, Taylor J; Galaxy Team. *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences*. Genome Biol. 2010

## Links and Tutorials

- Data formats  
<https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/>
- FASTQC  
<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Galaxy workshop  
<http://chagall.med.cornell.edu/galaxy/GalaxyWorkshopNotes.pdf>
- Importing data, data formats, QC and mapping  
<https://galaxyproject.github.io/training-material/topics/introduction/tutorials/galaxy-intro/ngs-data-managment/tutorial.html>
- A complete RNA-seq analysis using Galaxy  
<https://galaxyproject.github.io/training-material/topics/transcriptomics/tutorials/ref-based/tutorial.html>
- And a lot more in  
<https://galaxyproject.github.io/training-material/>