NGS Analysis and Galaxy

Kathryn Crouch
kathryn.crouch@glasgow.ac.uk

Introduction

University of Glasgow

Who am I?

- PhD in comparative immunology 2005
- Worked in industry (big pharma and small biotech)
- MSc bioinformatics 2013
- Core bioinformatician WCIP 2013 - 2021
- Create online resources for public use

VEuPathDB
Eukaryotic Pathogen, Vector & Host Informatics Resources

MicrobiomeDB
A Microbiome Resource

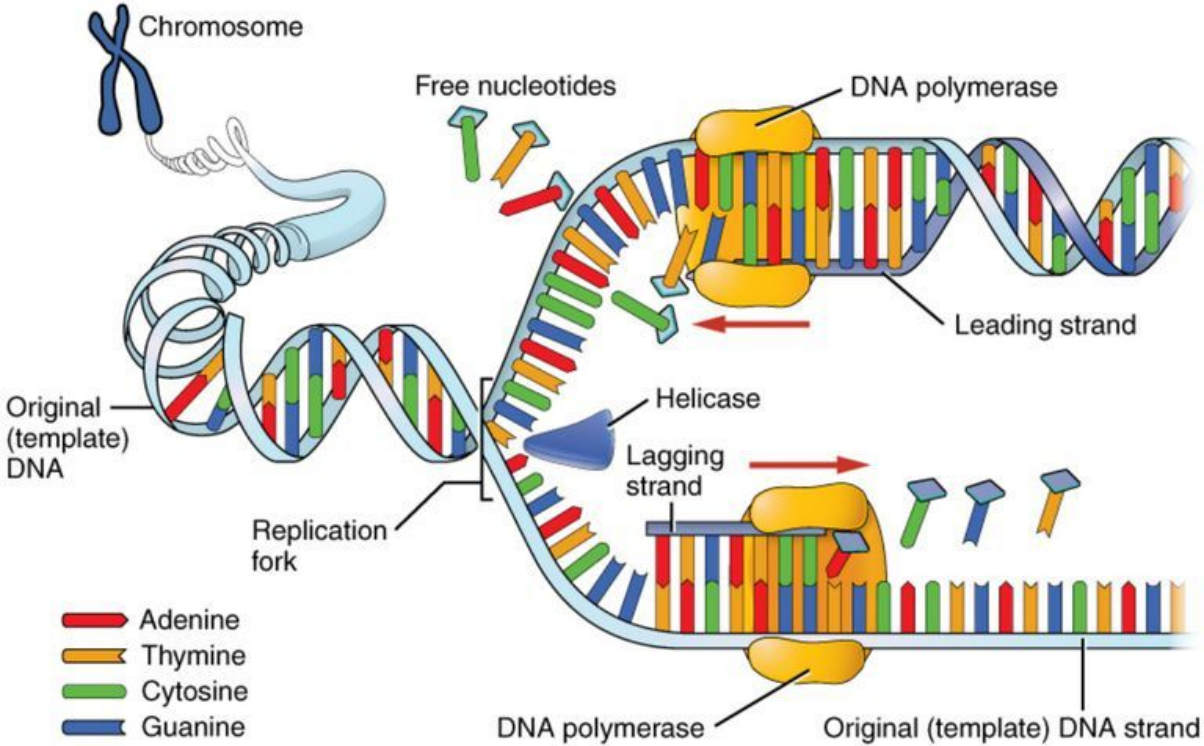
COMPANION

Outline

- Background
  - How does sequencing work?
  - What does the data look like?
- Pre-processing for Analysis
  - QC and trimming
  - Aligning to a reference genome
- Whole-genome sequencing
  - Calling SNPs
  - CNVs
- RNA-seq and differential expression
- Using Galaxy

How Does Sequencing Work?

- Carry out replication under controlled conditions
- Artificially slow down the reaction to see the order in which bases are incorporated



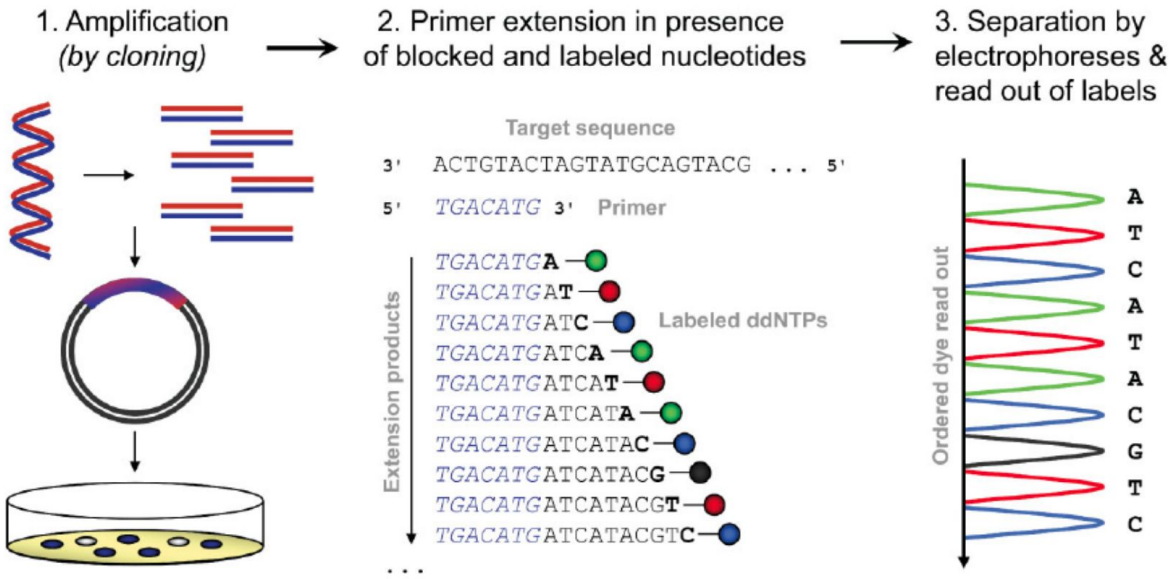
How Does Sequencing Work?

Sanger Sequencing (1975)

Uses modified nucleotides (ddNTPs) that cannot be extended

Each ddNTP is labelled with a different dye so you can see the order in which they are incorporated

Long reads and low error rate, but low throughput



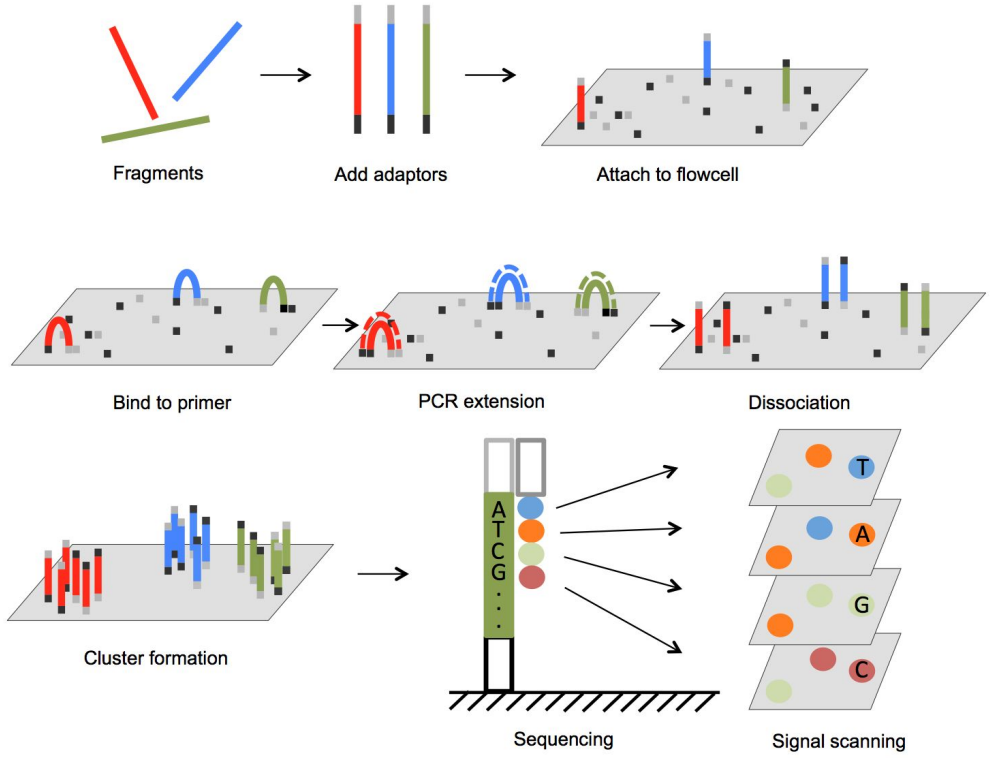
How Does Sequencing Work?

Illumina Sequencing (2005)

DNA fragments are adaptor-ligated and attached to a flow cell

PCR is carried out in situ to form clusters

Sequencing can be carried out on millions of clusters simultaneously



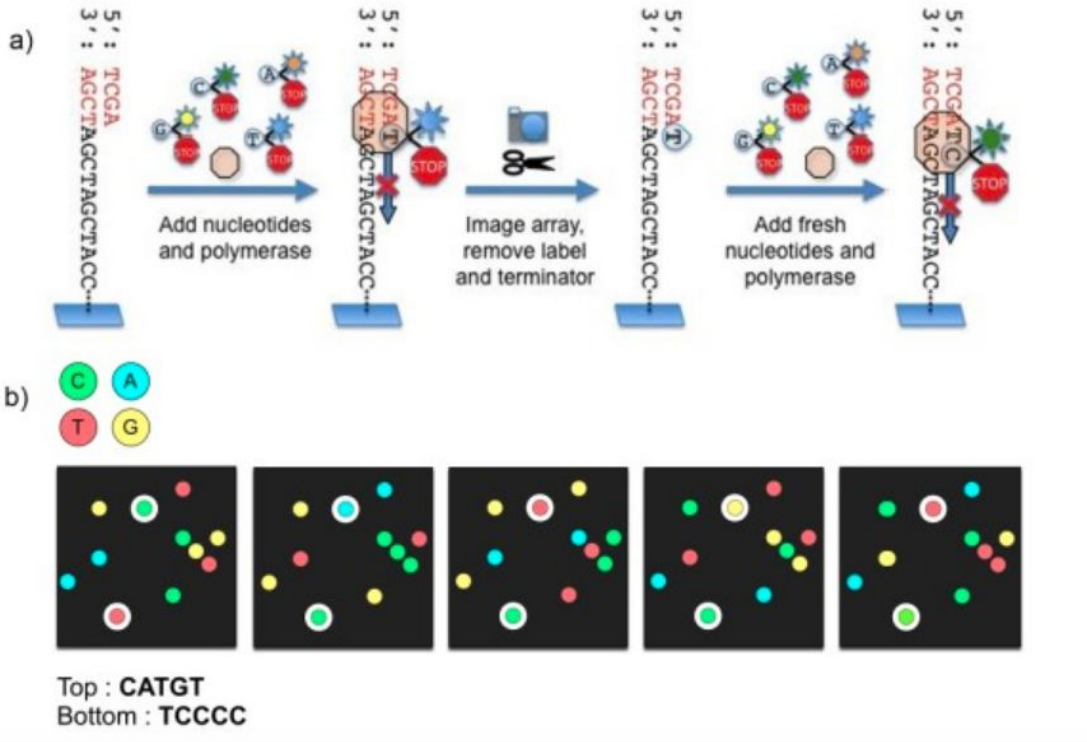
How Does Sequencing Work?

Illumina Sequencing

Blocked and labelled nucleotides are added

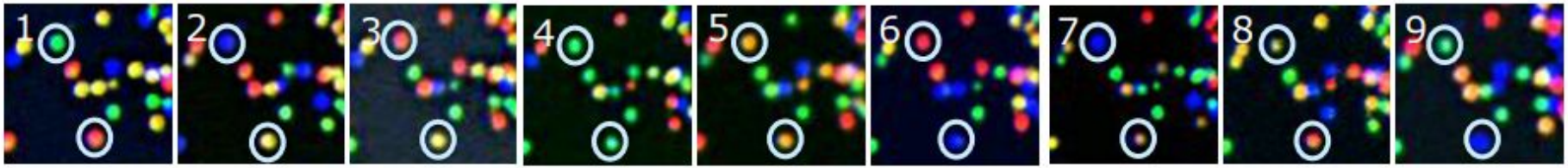
1 nucleotide is incorporated and an image is taken of the array

Label and block are removed and cycle repeats



How Does Sequencing Work?

T G C T A C G A T



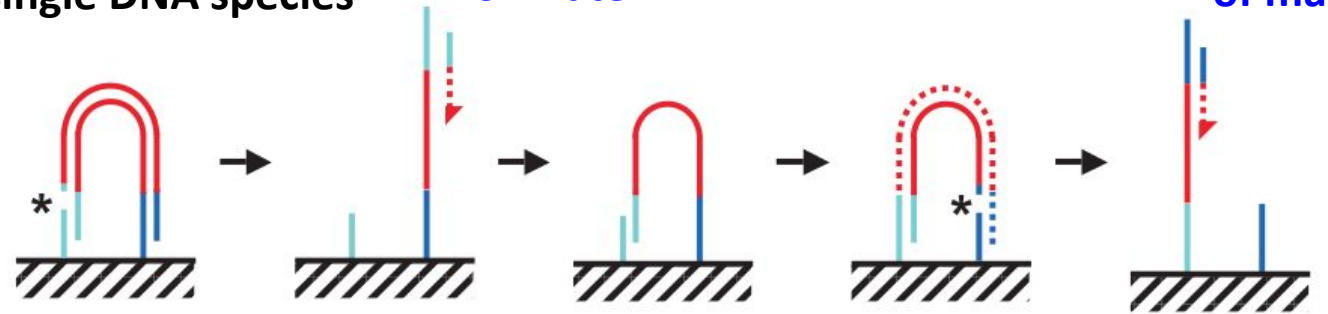
C A A T A G A C G

How Does Sequencing Work?

Denaturing
& washing
➔ single DNA species

Sequencing of mate 1

Sequencing of mate 2



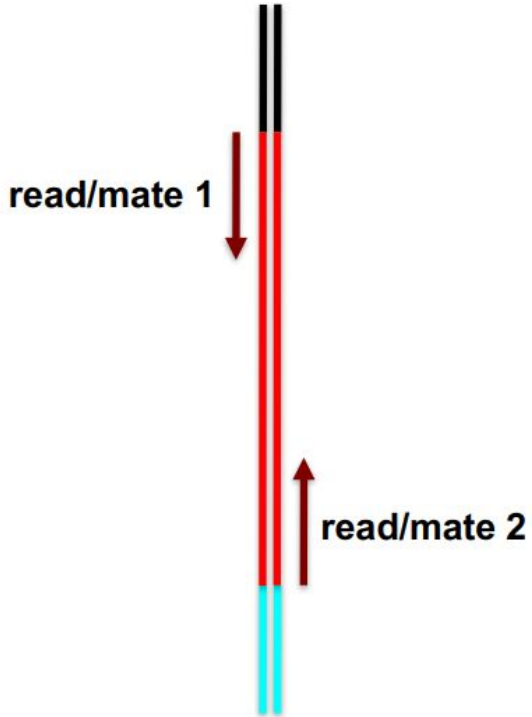
Adaptor-specific DNA-cleavage

Re-synthesis of complementary DNA strand & adaptor-specific DNA-cleavage

Paired End Illumina Sequencing

A short read is sequenced from each end of each fragment

Blue and black are adaptors

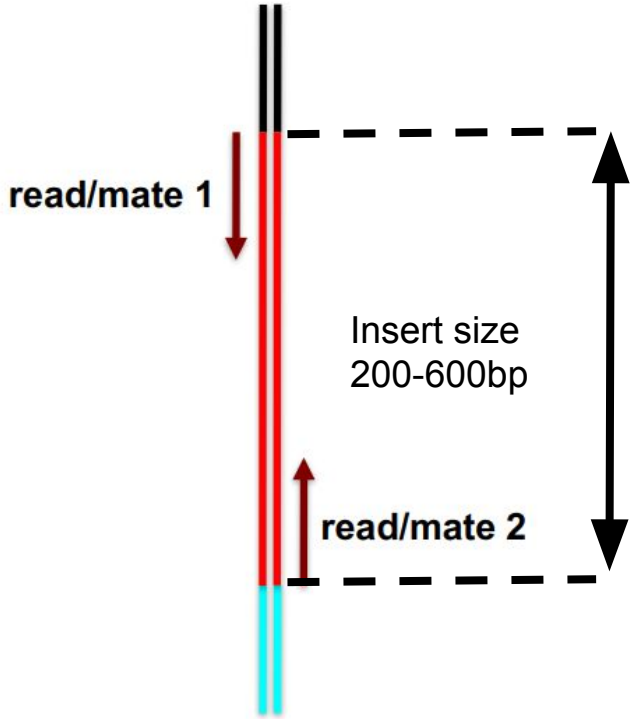


Paired End Sequencing

A short read is sequenced from each end of each fragment

Blue and black are adaptors

Insert or fragment size is the distance between adaptors



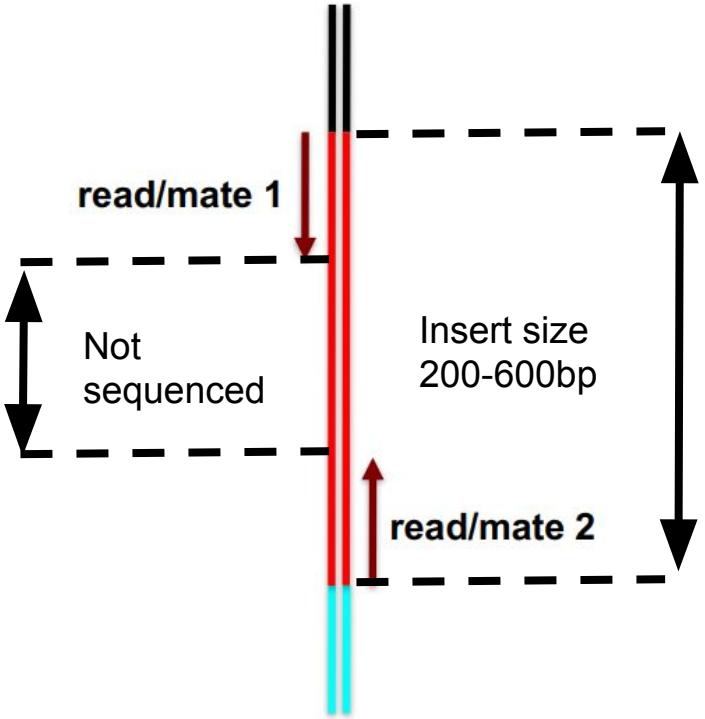
Paired End Sequencing

A short read is sequenced from each end of each fragment

Blue and black are adaptors

Insert or fragment size is the distance between adaptors

The region between the reads is not sequenced - it is covered by other fragments



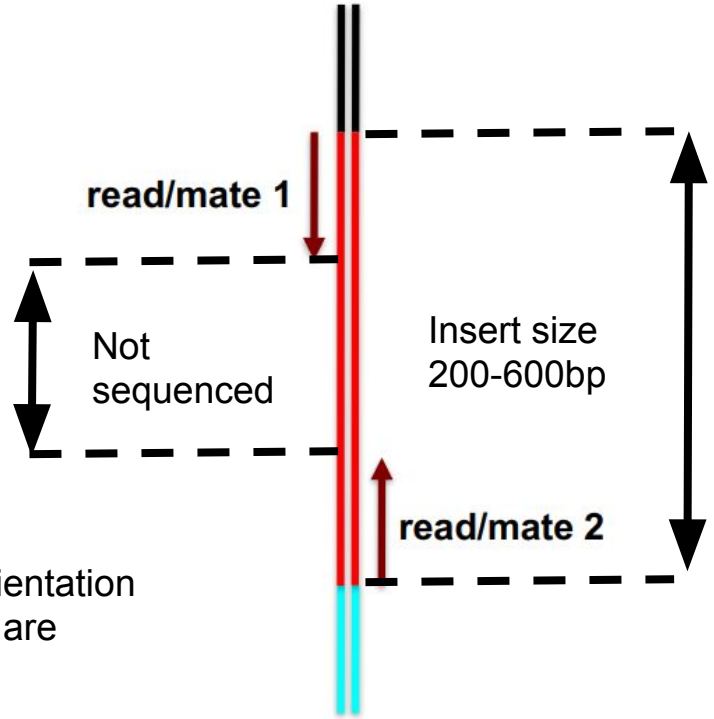
Paired End Sequencing

A short read is sequenced from each end of each fragment

Blue and black are adaptors

Insert or fragment size is the distance between adaptors

The region between the reads is not sequenced - it is covered by other fragments



Reads that map in the correct orientation and the expected distance apart are “concordant” or “proper pairs”

Concordant alignments are prioritised

File Formats: FASTQ

```
@NS500205:127:HW272BGXX:1:11101:7788:1040 1:N:0:TCTCGCGC+AGGCTATA
```

For paired-end reads you will have two files

4 lines per read

- Line 1 is a unique header (this will be shared between the pairs)

File Formats: FASTQ

```
@NS500205:127:HW272BGXX:1:11101:7788:1040 1:N:0:TCTCGCGC+AGGCTATA
CCTTGNTCCGTCATATTTTTTAGCATTGCAATGACGCTAAGTCCCGATTGACGCGCACGTGCTCACCCGGTTTCC
```

For paired-end reads you will have two files

4 lines per read

- Line 1 is a unique header (this will be shared between the pairs)
- Line 2 is the sequence of the read

File Formats: FASTQ

```
@NS500205:127:HW272BGXX:1:11101:7788:1040 1:N:0:TCTCGCGC+AGGCTATA
CCTTGNTCCGTCATATTTTTTAGCATTGCAATGACGCTAAGTCCCGATTGACGCGCACGTGCTCACCCGGTTTCC
+
AAAAA#EEEEEEEEEEEEEEEEEEEEAEEEEEAEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEAEEEEEEEEEA
```

For paired-end reads you will have two files

4 lines per read

- Line 1 is a unique header (this will be shared between the pairs)
- Line 2 is the sequence of the read
- Line 4 is the quality for each base
    - Quality is encoded using ASCII
    - http://www.asciitable.com/

Phred quality scores are logarithmically linked to error probabilities

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 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Alignment-based Analysis

Quality Control

↓

Read Trimming

↓

Alignment

Analysis of NGS Sequencing

Quality Control

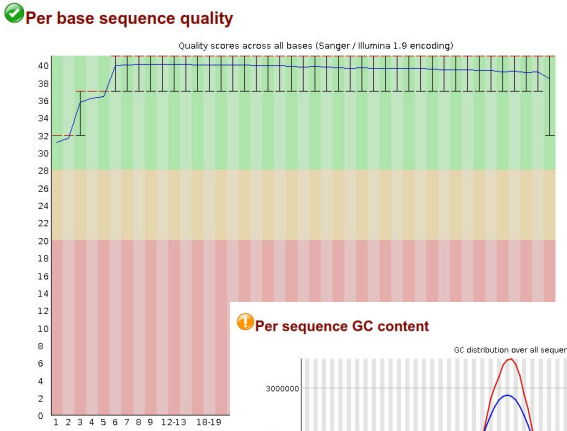
↓

Read Trimming

↓

Alignment

- FASTQC https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
  - Overall sequencing quality
  - GC content
  - N content
  - Read length distribution
  - Over-represented sequences
  - Adaptor content
- Output is an html file that can be opened in a web browser



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
ACAAGTGTGTAACATTAATTTGCAAGTTTGCAACGCTGTTCTTTAGTGTT	70896	0.12562741276052788	No Hit

Analysis of NGS Sequencing

Quality Control

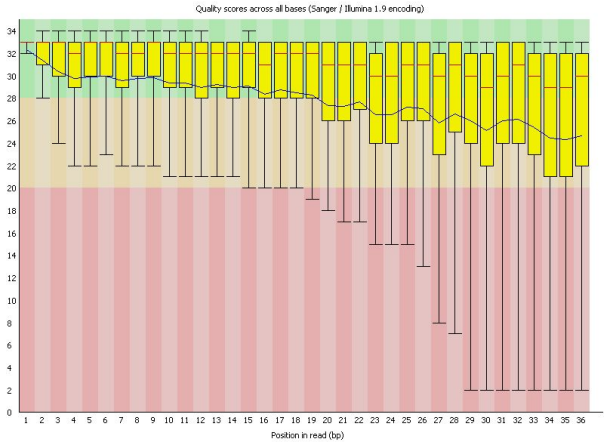
↓

Read Trimming

↓

Alignment

- Trimmomatic https://github.com/usadellab/Trimmomatic
- Sequence quality tends to decrease towards the 3’ end of reads - these can affect mapping
- Sickle will:
  - Remove poor quality reads from the 3’ end of each read
  - Check for reads that are too short and discard them
  - Check that all reads still have a pair and discard those that don’t
- Some trimming tools can also remove adaptors



Analysis of NGS Sequencing

Quality Control

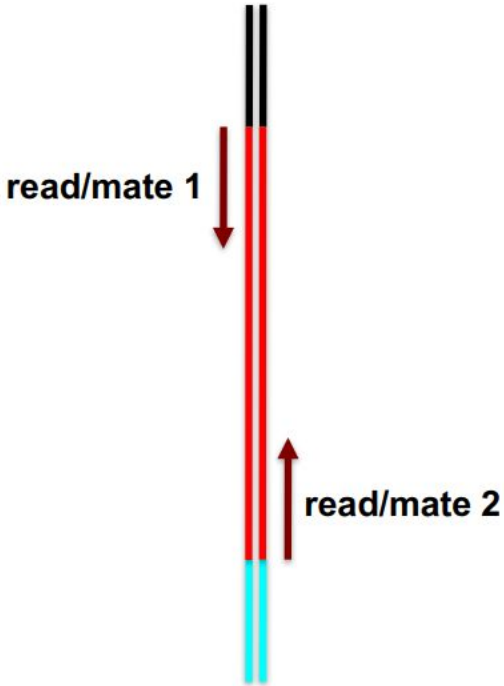
↓

Read Trimming

↓

Alignment

- Trimmomatic https://github.com/usadellab/Trimmomatic
- Sequence quality tends to decrease towards the 3’ end of reads - these can affect mapping
- Sickle will:
  - Remove poor quality reads from the 3’ end of each read
  - Check for reads that are too short and discard them
  - Check that all reads still have a pair and discard those that don’t
- Some trimming tools can also remove adaptors



Analysis of NGS Sequencing

Quality Control

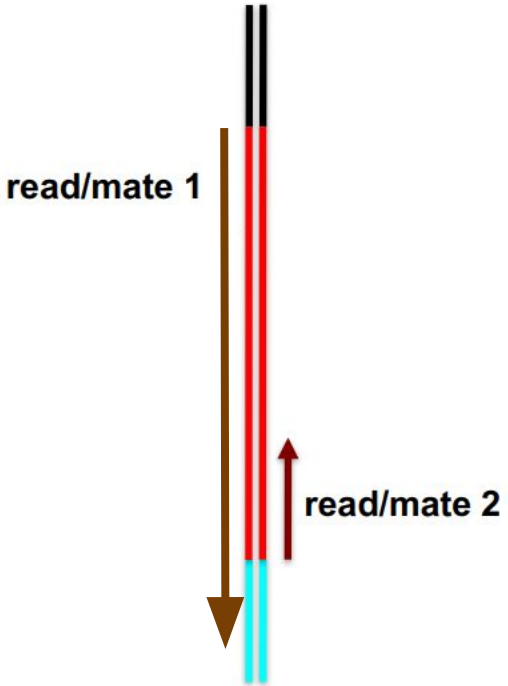
↓

Read Trimming

↓

Alignment

- Trimmomatic https://github.com/usadellab/Trimmomatic
- Sequence quality tends to decrease towards the 3’ end of reads - these can affect mapping
- Sickle will:
  - Remove poor quality reads from the 3’ end of each read
  - Check for reads that are too short and discard them
  - Check that all reads still have a pair and discard those that don’t
- Some trimming tools can also remove adaptors



Analysis of NGS Sequencing

Quality Control

↓

Read Trimming

↓

Alignment

What is an alignment?

Two sequences:

```
ATTGAAAGCTA
GAAATGAAAAGG
```

How would you align one to the other?

Analysis of NGS Sequencing

Quality Control

↓

Read Trimming

↓

Alignment

What is an alignment?

Two sequences:

```
ATTGAAAGCTA
GAAATGAAAAGG
```

How would you align one to the other?

```
--ATTGAAA-GCTA
  | ||||| |
GAAATGAAAAGG
```

```
ATTGAAA-GCTA---
   |||| |   |
---GAAATGAAAAGG
```

Which one is better??

Analysis of NGS Sequencing

Quality Control

↓

Read Trimming

↓

Alignment

What is an alignment?

Two sequences:

```
ATTGAAAGCTA
GAAATGAAAAGG
```

How would you align one to the other?

```
--ATTGAAA-GCTA
  | ||||| |
GAAATGAAAAGG--

ATTGAAA-GCTA---
   |||| |   |
---GAAATGAAAAGG
```

Which one is better??

Alignment scoring:

- 1 for a match
- -1 for a mismatch
- -2 for a gap

Now alignment 1 scores -4 and alignment 2 scores -10 so we would choose alignment 1

Alignment algorithms are designed to align data in a reasonable time on a standard computer

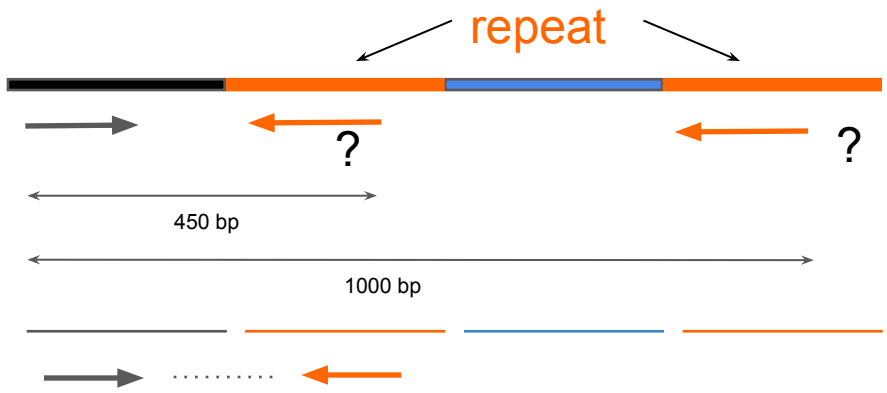
Heuristic - not exhaustive, but “good enough”

Improving Alignments with Paired End Reads

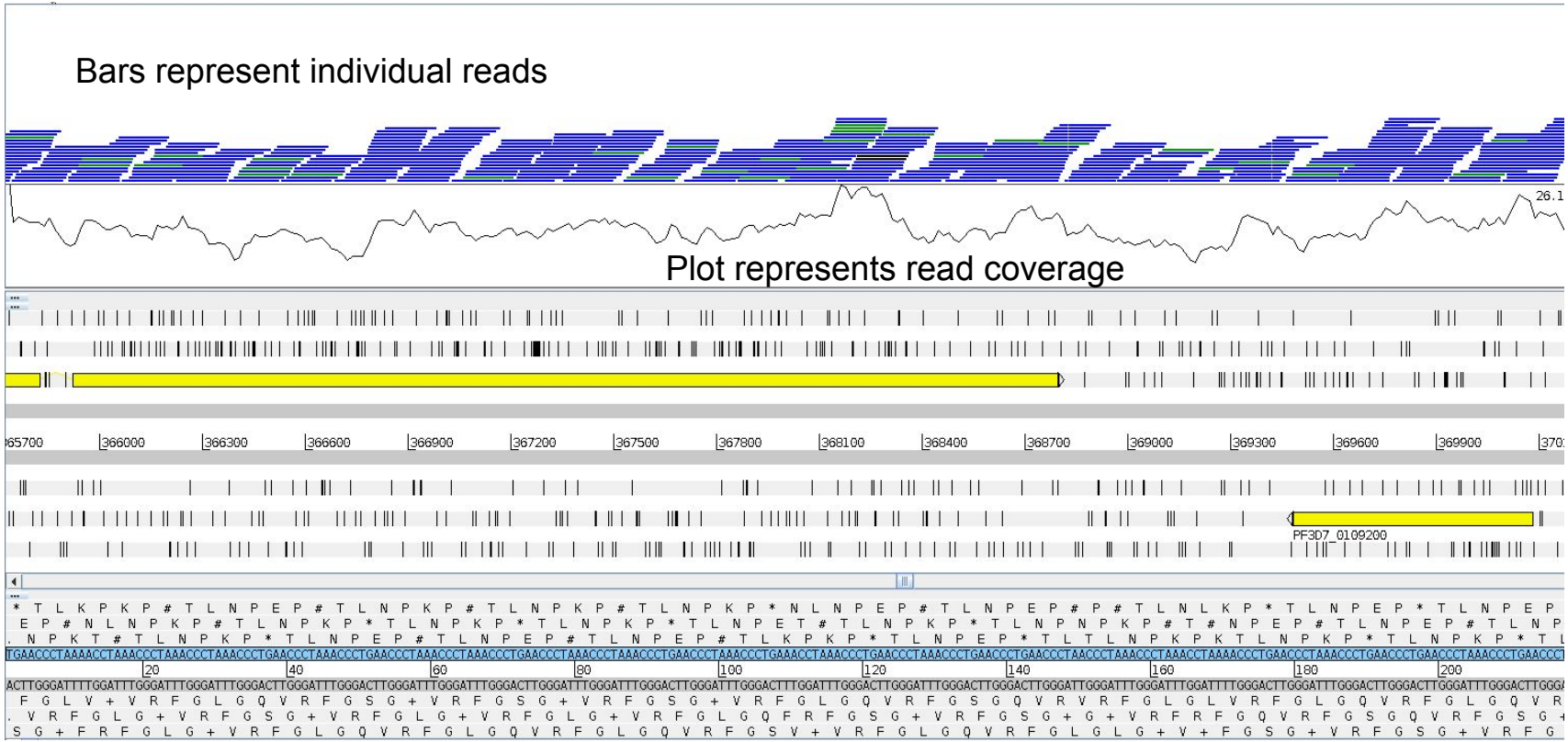
Paired end reads can resolve alignments in repetitive regions

A read that could map in multiple locations due to repetitions in the genome can be located accurately by inference from the position of its pair

In this case, we know that the insert size is ~400 bp, so we can infer that the first alignment is more likely to be correct.



Visualising Alignments (Artemis)



Aligning RNA-seq Data

Quality Control

↓

Read Trimming

↓

Alignment

Spliced mRNA

Aligning RNA-seq Data

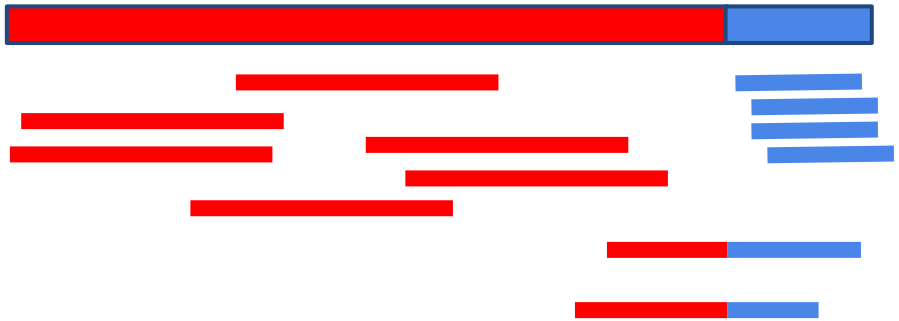
Quality Control

↓

Read Trimming

↓

Alignment



Spliced mRNA

Aligning RNA-seq Data

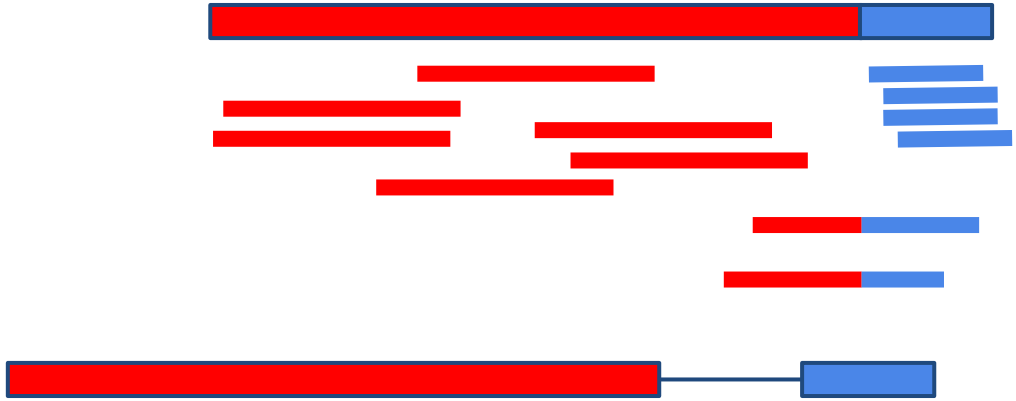
Quality Control

↓

Read Trimming

↓

Alignment



Spliced mRNA

Reference genome (genomic sequence with introns)

Aligning RNA-seq Data

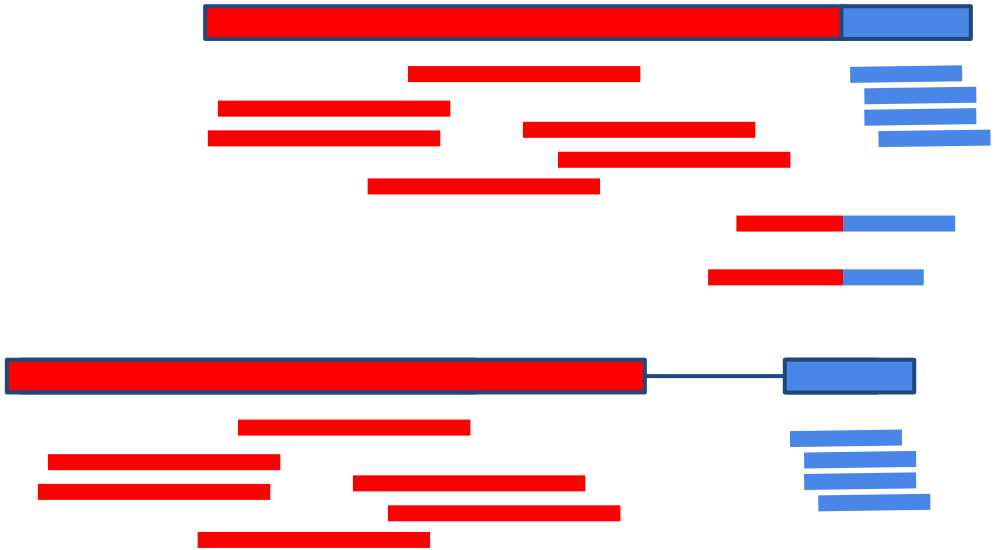
Quality Control

↓

Read Trimming

↓

Alignment



Spliced mRNA

Reference genome

Aligning RNA-seq Data

Quality Control

↓

Read Trimming

↓

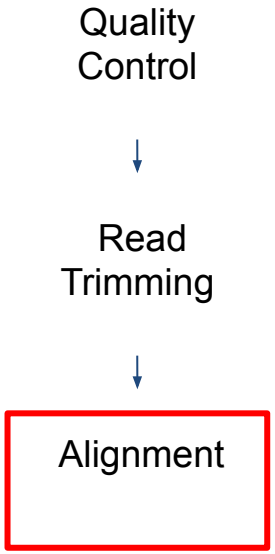
Alignment

Spliced mRNA

Reference genome

Intron-spanning reads

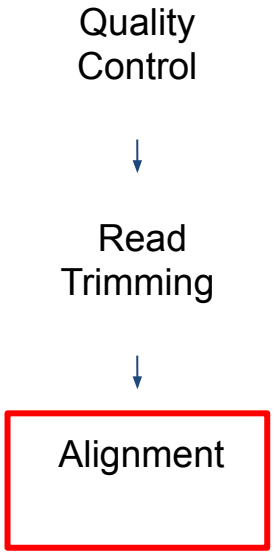
Alignment Tools



Mapping Tools for DNA-seq data

- BWA
  - https://github.com/lh3/bwa
- Bowtie2
  - https://bowtie-bio.sourceforge.net/bowtie2/index.shtml

Alignment Tools



Mapping Tools for RNA-seq data

- Must be capable of aligning intron-spanning reads
- Hisat2
  - Fast, sacrifices sensitivity
  - http://daehwankimlab.github.io/hisat2/
- STAR
  - Very sensitive, but slow
  - https://github.com/alexdobin/STAR

Whole Genome Sequencing (DNA-seq, WGS)

What Can We Discover From Aligned Reads?

- Where and how is our sample different from the reference?
  - Discovery of SNVs and Indels

- Coverage
  - Discovery of copy number variations
  - Discovery of regions of high variability

Finding SNVs

Quality Control

↓

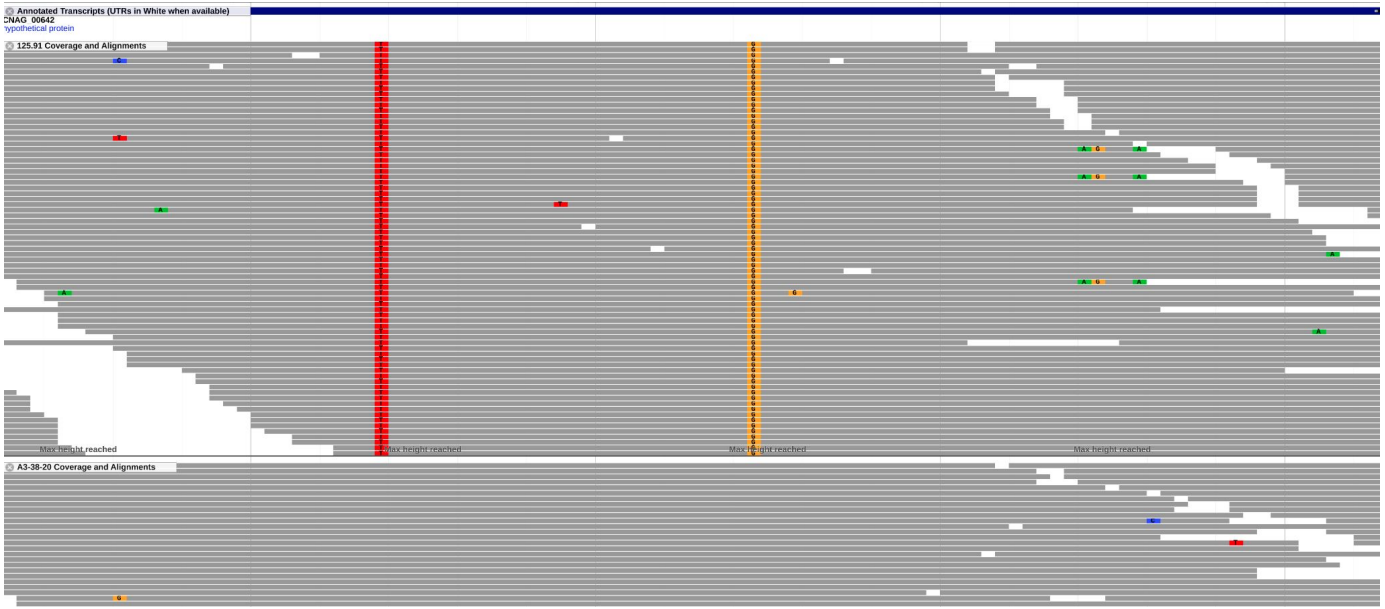
Read Trimming

↓

Alignment

↓

SNP Calling



Finding Variants

If we load alignments into a genome viewer, we can see variants

How do we find them globally? How do we assess them?

Analysis of NGS Sequencing

Quality Control

↓

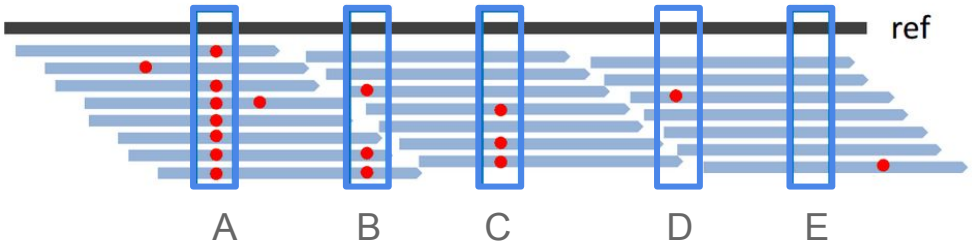
Read Trimming

↓

Alignment

↓

SNP Calling



Blue lines are reads aligned against a reference (black). Red dots indicate individual bases where a base in a read differs from the reference.

A: Most reads differ from the reference -> homozygous SNP
B and C: Roughly 50% of reads differ from the reference -> potential heterozygous SNP
D: Only one base differs from the reference -> probably a sequencing error
E: All bases the same as the reference

Analysis of NGS Sequencing

Quality Control

↓

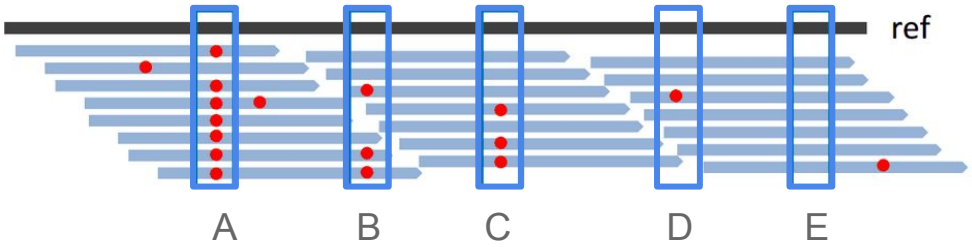
Read Trimming

↓

Alignment

↓

SNP Calling



Things to think about:

- What happens if your sample is not a clone?
- What happens if your sequencing depth is low?

Analysis of NGS Sequencing

Quality Control

↓

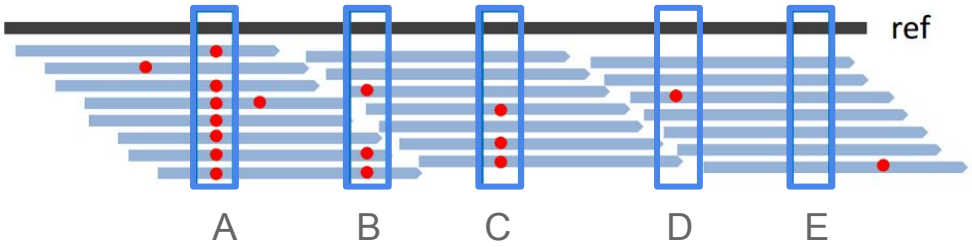
Read Trimming

↓

Alignment

↓

SNP Calling

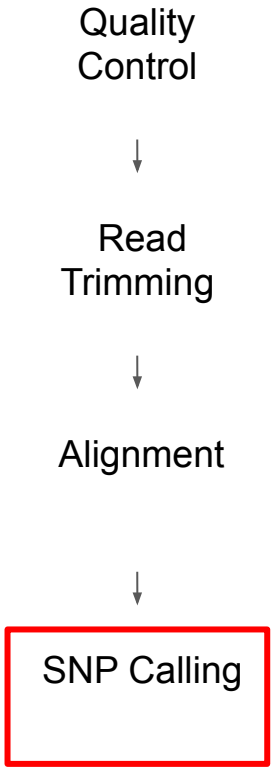


Freebayes https://github.com/freebayes/freebayes

Automated tool to call SNPs

You may also come across other tools including GATK and BCFTools.

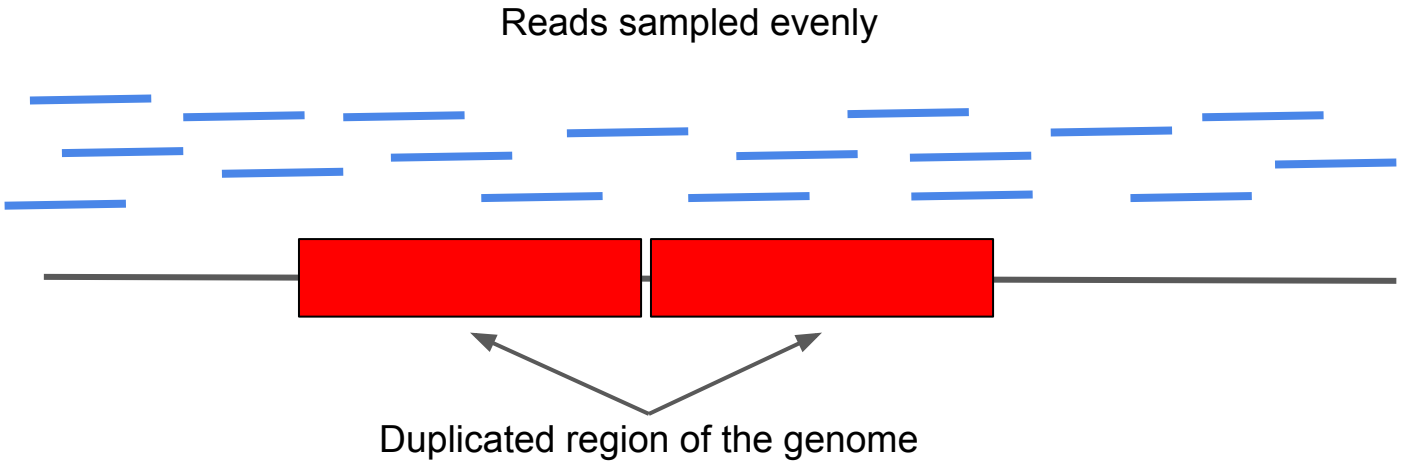
What Else Can We Find Out?



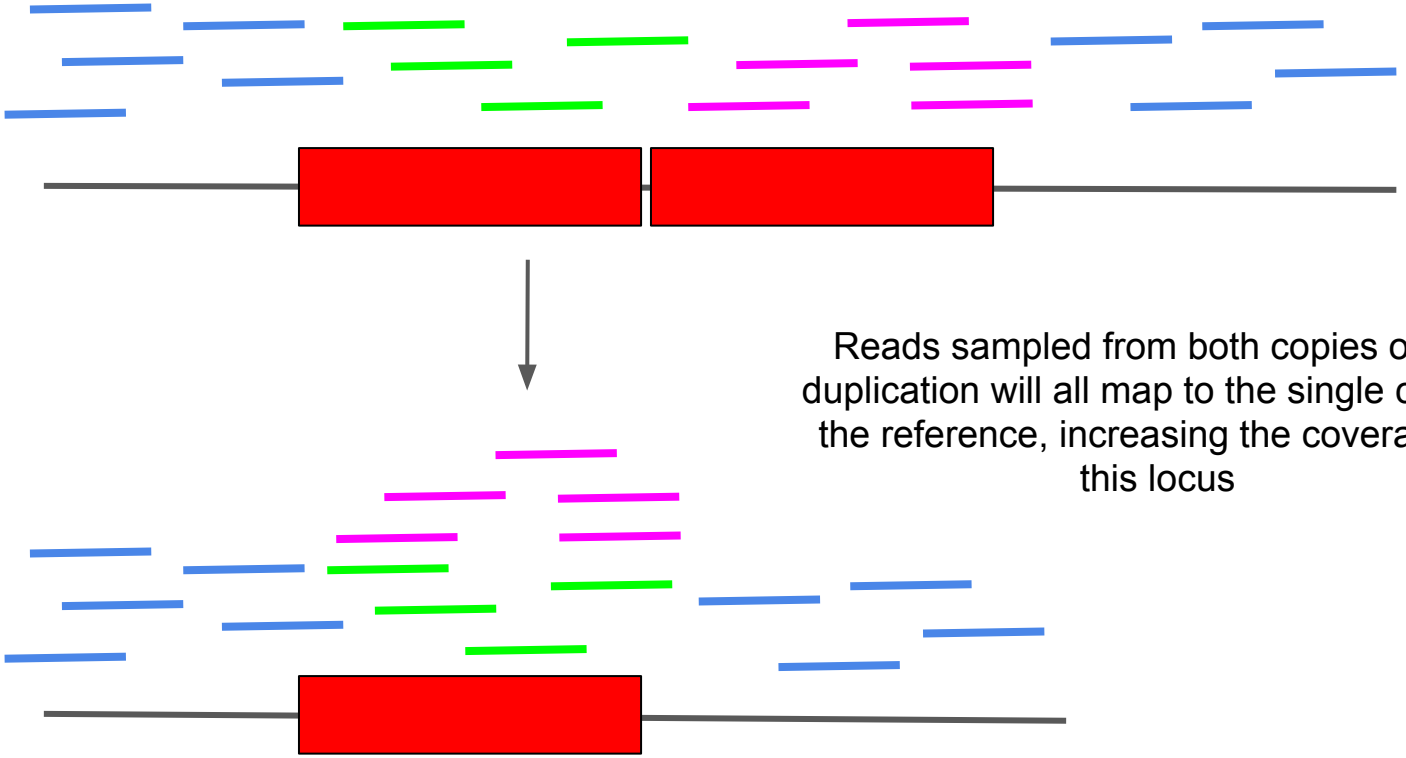
Coverage

- Expect coverage to be even across the genome
- In reality, we see variation associated with:
  - GC content
  - Repetitive or highly variable regions
  - Large scale insertions and deletions
- Note that doing alignments and examining coverage is the basis of RNAseq and ChIPseq analysis too!

Coverage and Copy Number Variations



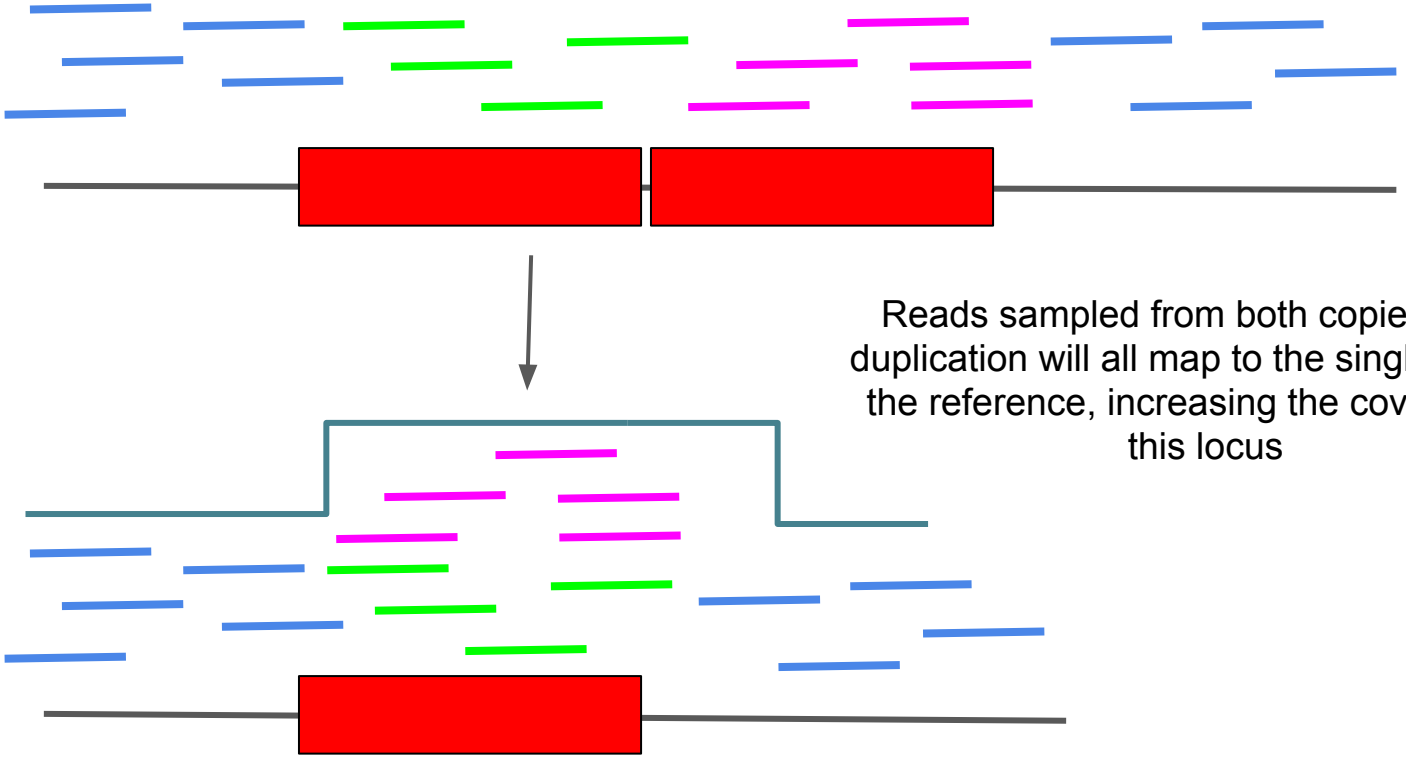
Copy Number Variations



Reads sampled from both copies of the duplication will all map to the single copy in the reference, increasing the coverage at this locus

Reference genome without duplication

Copy Number Variations



Reads sampled from both copies of the duplication will all map to the single copy in the reference, increasing the coverage at this locus

Reference genome without duplication

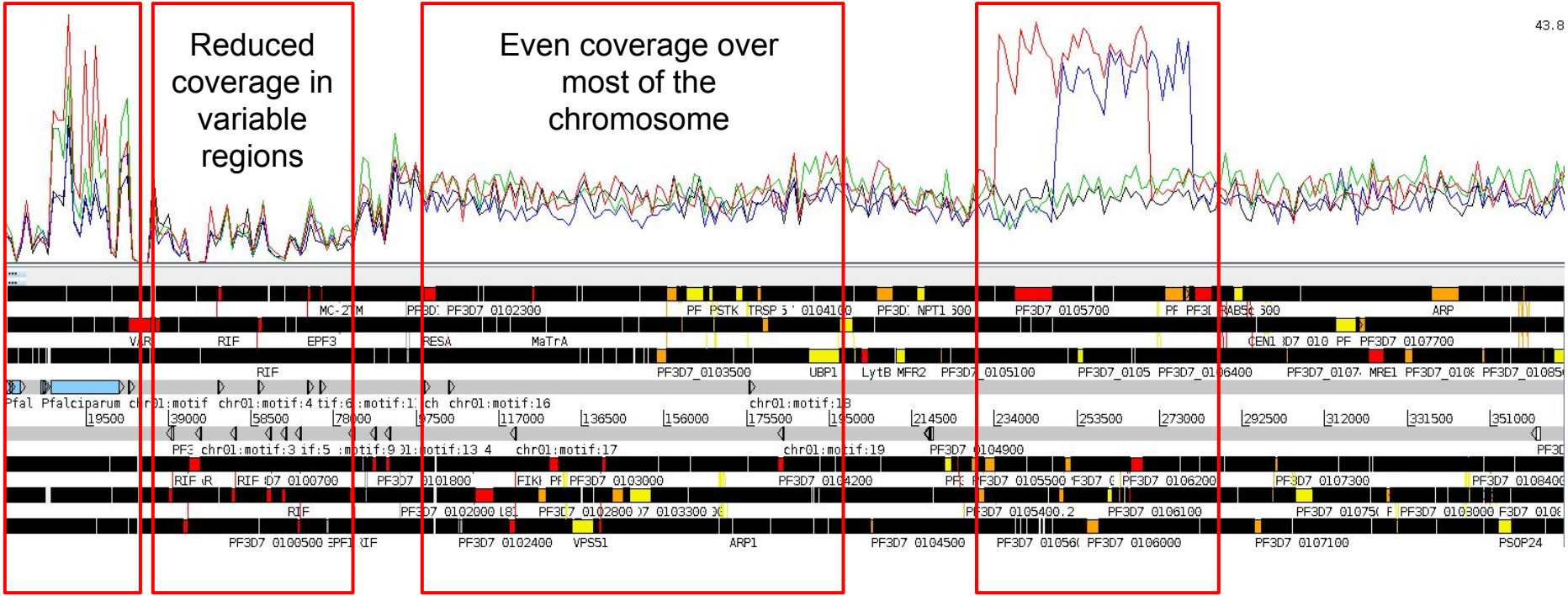
Global Coverage in Artemis

Variable coverage in repetitive regions

Reduced coverage in variable regions

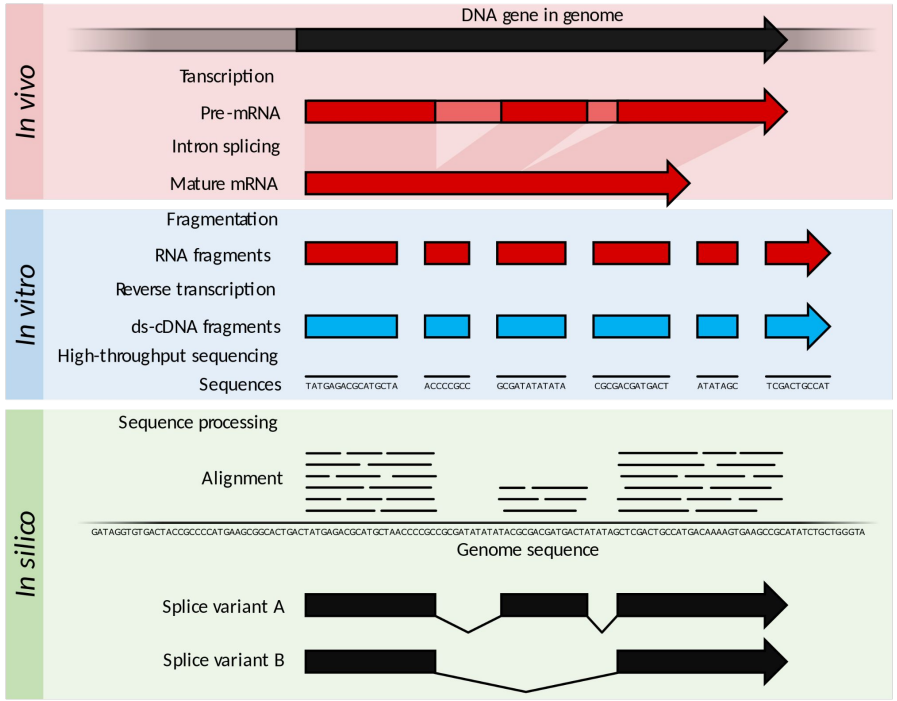
Even coverage over most of the chromosome

CNV (triplication) in two samples

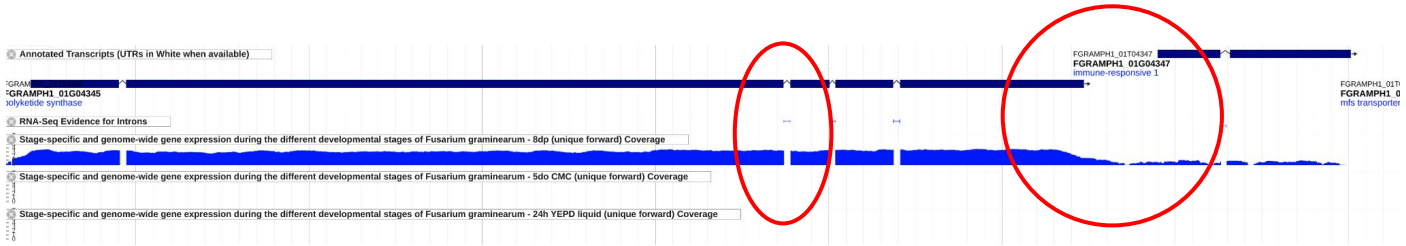


Transcript Sequencing (RNA-seq)

Transcriptome sequencing



What Can We Learn from RNA-seq?



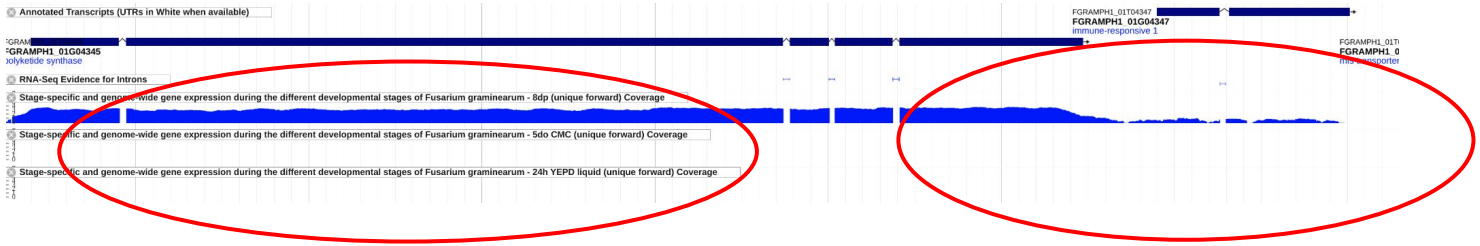
Gene Model Prediction

Alignment of RNA-seq reads to a genomic reference can help us to predict and confirm gene model structure

- Introns can be predicted based on coverage and on individual reads that cross splice junctions
- UTRs can be predicted based on coverage
- Differential splicing can also be predicted from coverage

https://tinyurl.com/mven959f

What Can We Learn from RNA-seq?



Differential Expression

Depth of coverage can help us learn about transcript abundance

- Differential transcript abundance can be observed both within and between samples

https://tinyurl.com/mven959f

Quantifying Expression

Quality Control

↓

Trimming

↓

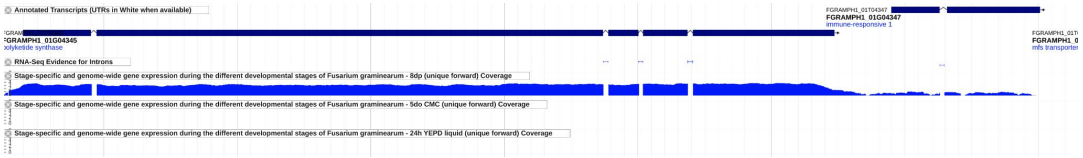
Read Alignment

↓

Read Counting

↓

Differential Expression



Quantifying Expression

We've seen that we can see expression differences in a genome browser

Looking at plots like this is great for one gene, but it is too much to look at every gene individually and is not statistically robust

To examine transcript expression globally and perform robust statistics, we must count how many reads map to each gene.

Quantifying Expression

Quality Control

↓

Trimming

↓

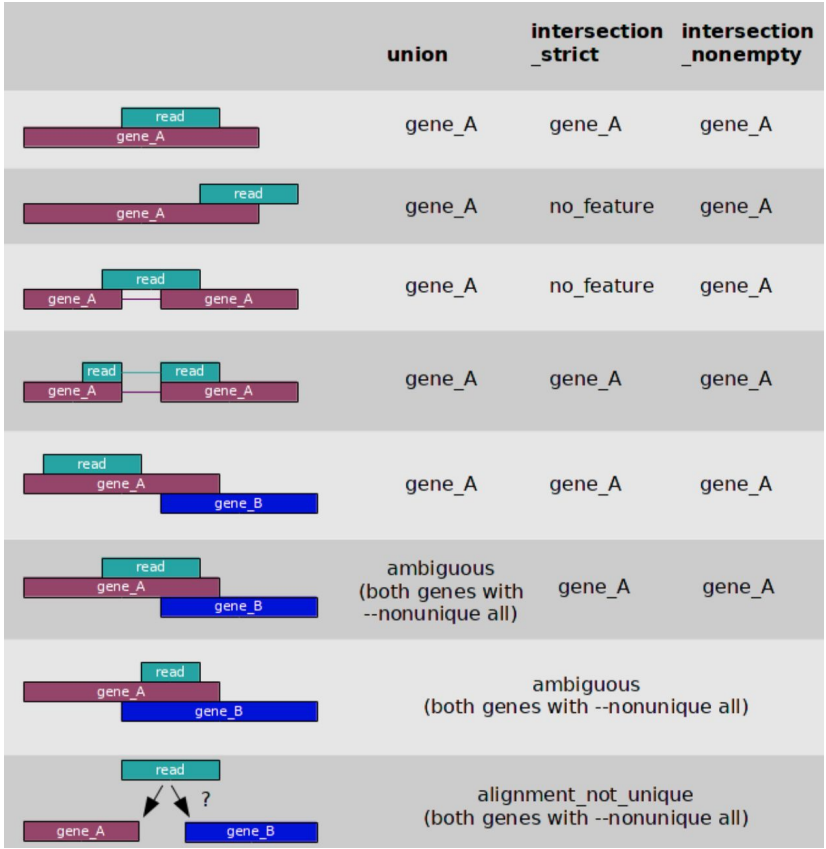
Read Alignment

↓

Read Counting

↓

Differential Expression



Read Counting Tools:

htseq-count:
https://htseq.readthedocs.io/en/release_0.11.1/count.html

FeatureCounts:
https://subread.sourceforge.net/featureCounts.html

Kallisto:
https://pachterlab.github.io/kallisto/

Normalisation

Quality Control

↓

Trimming

↓

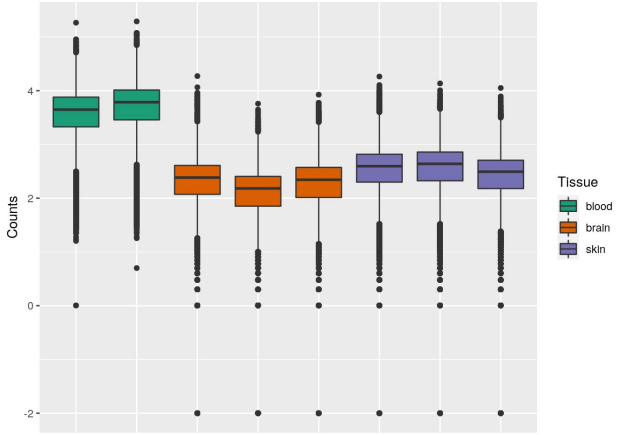
Read Alignment

↓

Read Counting

↓

Differential Expression

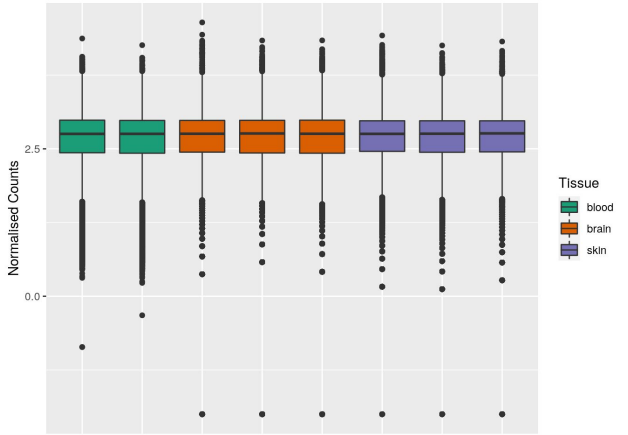


Raw count data

Each box represents is one sample and shows the distribution of read counts for each gene

Normalised count data

After normalising, count distributions are aligned so individual genes can be compared



With Quantitative Data We Can…

Quality Control

↓

Trimming

↓

Read Alignment

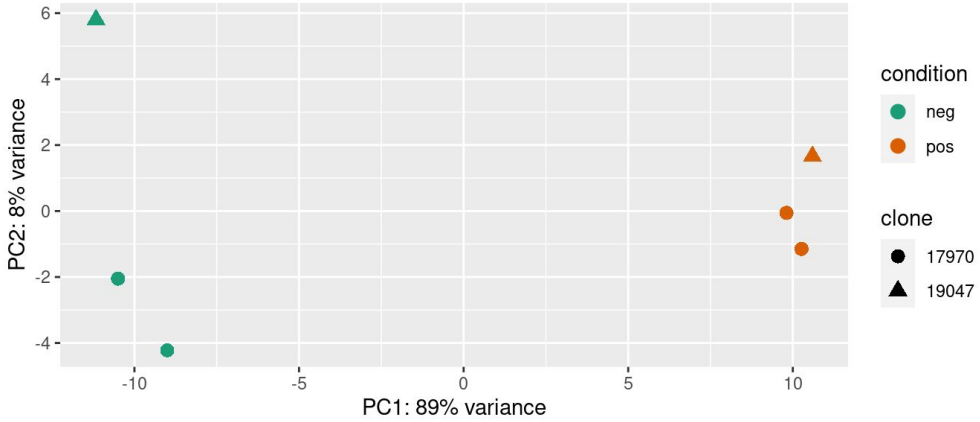
↓

Read Counting

↓

Differential Expression

- Explore our data



With Quantitative Data We Can…

Quality Control

↓

Trimming

↓

Read Alignment

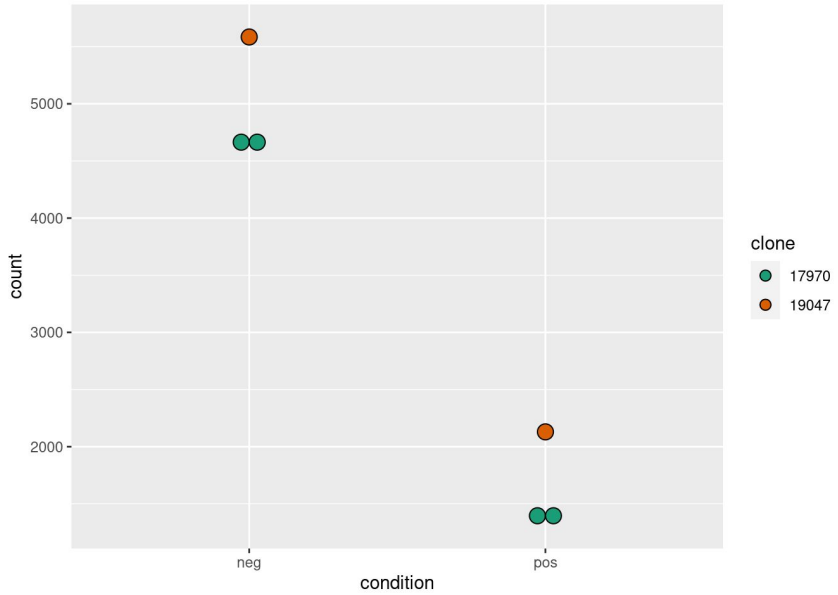
↓

Read Counting

↓

Differential Expression

- Explore our dataset
- Look at expression for individual genes



With Quantitative Data We Can…

Quality Control

↓

Trimming

↓

Read Alignment

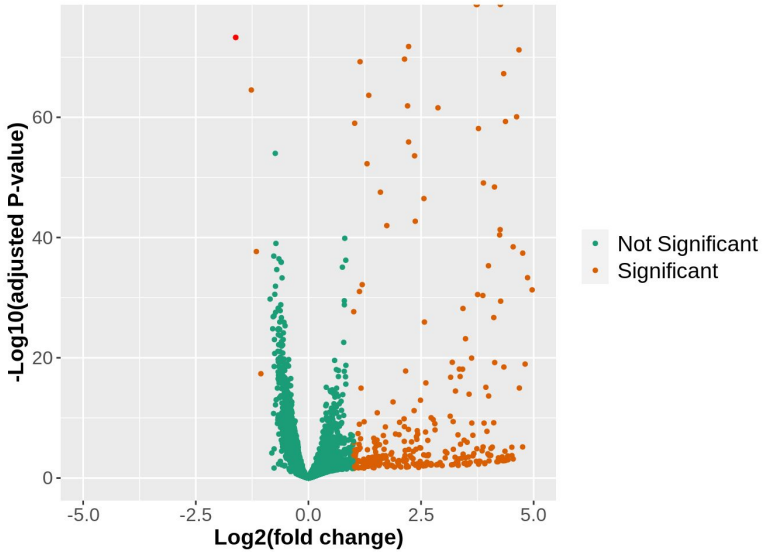
↓

Read Counting

↓

Differential Expression

- Explore our dataset
- Look at expression for individual genes
- Do pairwise statistical tests (differential expression)



With Quantitative Data We Can…

Quality Control

↓

Trimming

↓

Read Alignment

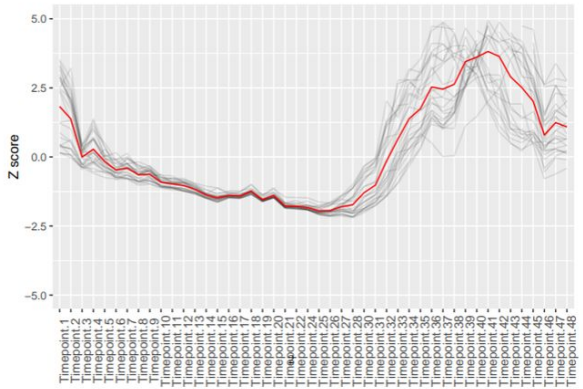
↓

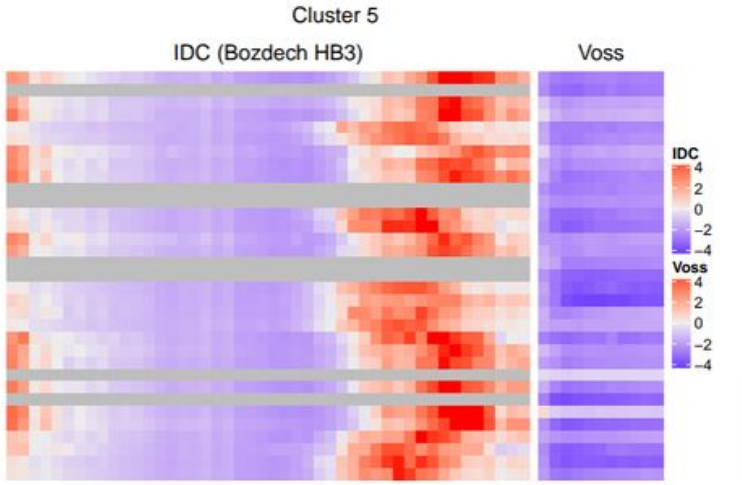
Read Counting

↓

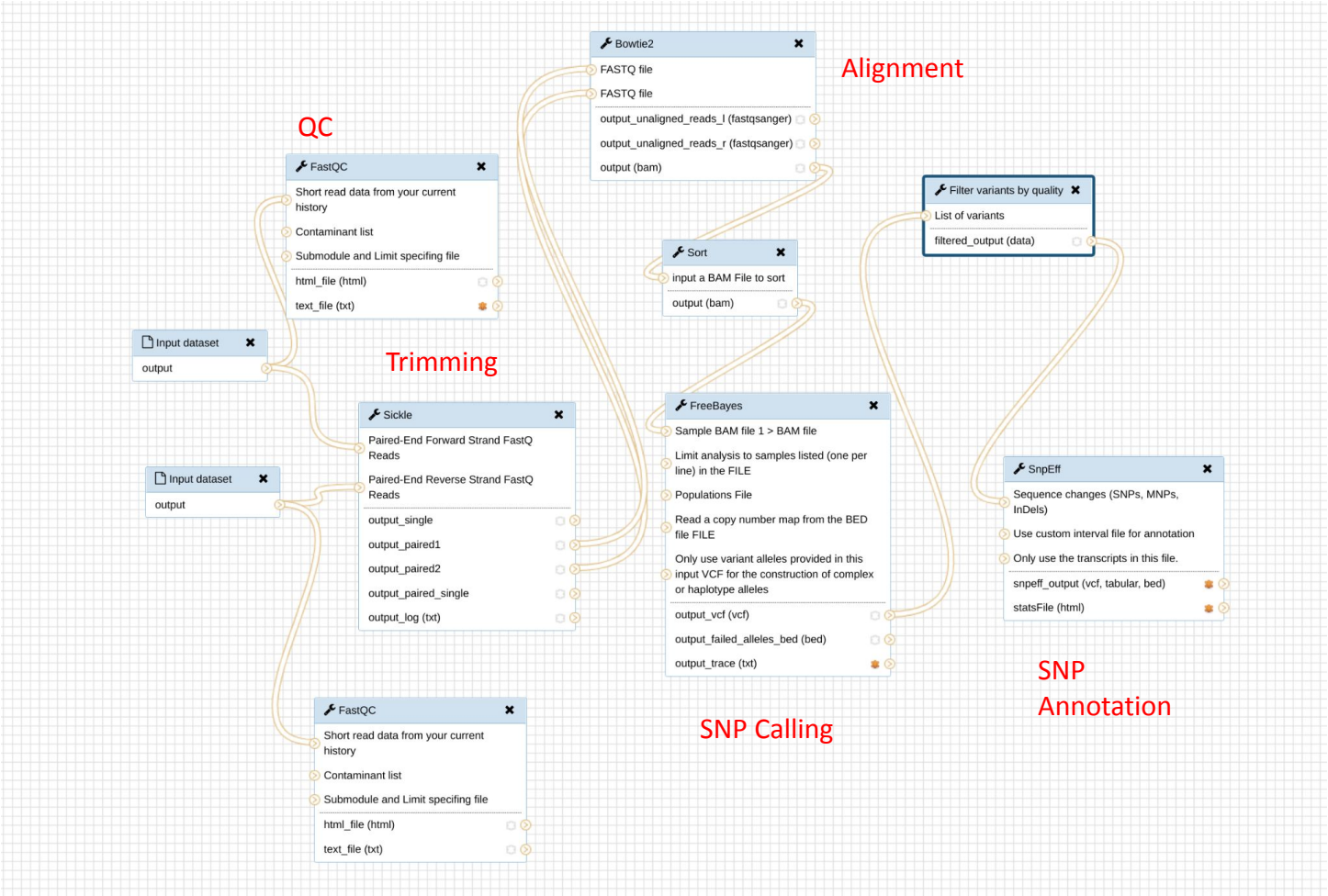
Differential Expression

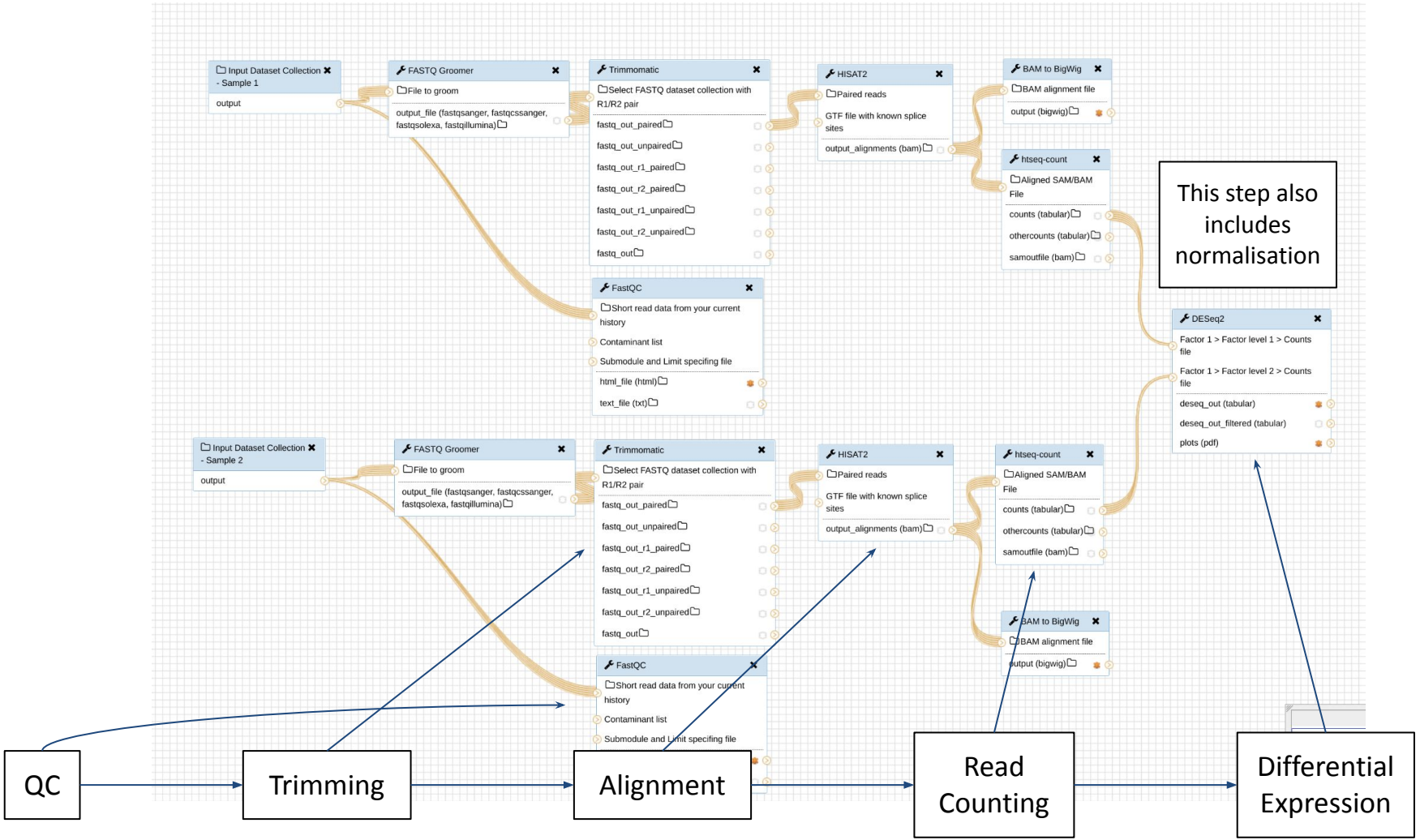
- Explore our dataset
- Look at expression for individual genes
- Do pairwise statistical tests (differential expression)
- Do advanced analysis (clustering, coexpression, etc.)





Galaxy


Bowtie2
FASTQ file
FASTQ file
output_unaligned_reads_l (fastqsanger)
output_unaligned_reads_r (fastqsanger)
output (bam)
Alignment
QC
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
Filter variants by quality
List of variants
filtered_output (data)
Sort
input a BAM File to sort
output (bam)
Input dataset
output
Trimming
Sickle
Paired-End Forward Strand FastQ Reads
Paired-End Reverse Strand FastQ Reads
output_single
output_paired1
output_paired2
output_paired_single
output_log (txt)
FreeBayes
Sample BAM file 1 > BAM file
Limit analysis to samples listed (one per line) in the FILE
Populations File
Read a copy number map from the BED file FILE
Only use variant alleles provided in this input VCF for the construction of complex or haplotype alleles
output_vcf (vcf)
output_failed_alleles_bed (bed)
output_trace (txt)
SnpEff
Sequence changes (SNPs, MNPs, InDels)
Use custom interval file for annotation
Only use the transcripts in this file.
snpeff_output (vcf, tabular, bed)
statsFile (html)
Input dataset
output
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
SNP Calling
SNP Annotation


Input Dataset Collection - Sample 1
output
FASTQ Groomer
File to groom
output_file (fastqsanger, fastqcssanger, fastqsolexa, fastqillumina)
Trimmomatic
Select FASTQ dataset collection with R1/R2 pair
fastq_out_paired
fastq_out_unpaired
fastq_out_r1_paired
fastq_out_r2_paired
fastq_out_r1_unpaired
fastq_out_r2_unpaired
fastq_out
HISAT2
Paired reads
GTF file with known splice sites
output_alignments (bam)
BAM to BigWig
BAM alignment file
output (bigwig)
htseq-count
Aligned SAM/BAM File
counts (tabular)
othercounts (tabular)
samoutfile (bam)
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
This step also includes normalisation
DESeq2
Factor 1 > Factor level 1 > Counts file
Factor 1 > Factor level 2 > Counts file
deseq_out (tabular)
deseq_out_filtered (tabular)
plots (pdf)
Input Dataset Collection - Sample 2
output
FASTQ Groomer
File to groom
output_file (fastqsanger, fastqcssanger, fastqsolexa, fastqillumina)
Trimmomatic
Select FASTQ dataset collection with R1/R2 pair
fastq_out_paired
fastq_out_unpaired
fastq_out_r1_paired
fastq_out_r2_paired
fastq_out_r1_unpaired
fastq_out_r2_unpaired
fastq_out
HISAT2
Paired reads
GTF file with known splice sites
output_alignments (bam)
htseq-count
Aligned SAM/BAM File
counts (tabular)
othercounts (tabular)
samoutfile (bam)
BAM to BigWig
BAM alignment file
output (bigwig)
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
QC
Trimming
Alignment
Read Counting
Differential Expression