INSTITUTE OF PLANT SCIENCES

Sant'Anna
School of Advanced Studies – Pisa

Advanced Genomics

Next-Generation sequencing: principles of alignment and variant calling

Some considerations before we start…

We will take into consideration Illumina paired-end sequencing

Currently is the leading technology in characterization of genomic diversity, and not only…

We will go through some **reference-based methods** for generating and processing population-scale **NG resequencing** data, producing high-quality variants

Schumtz et al. 2014,
Nature Genetics

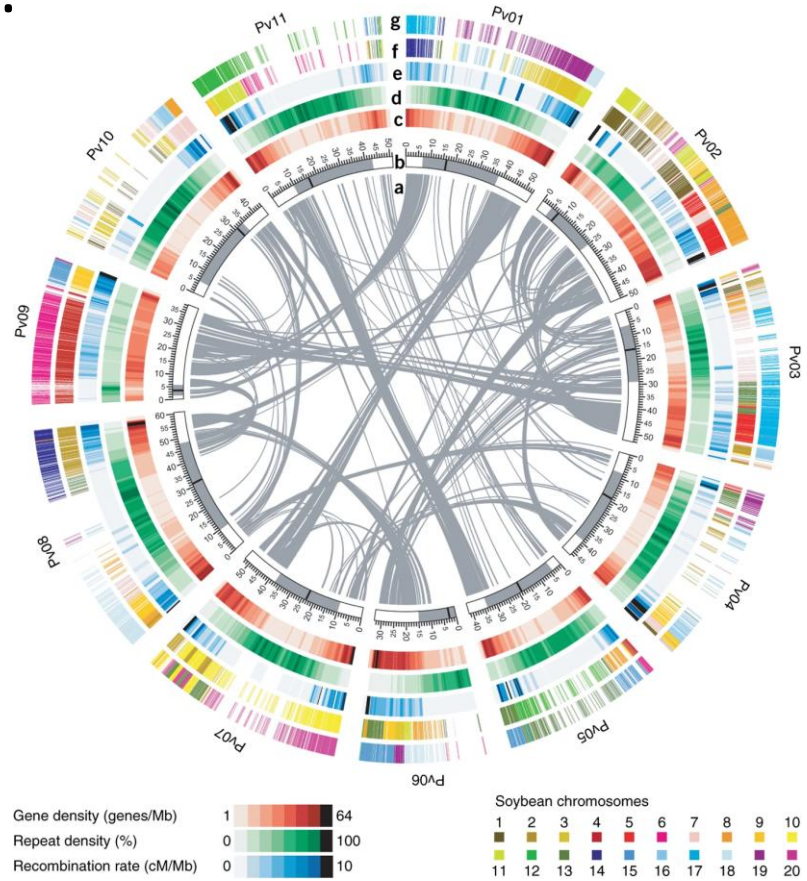
What is reference genome?

A **genome** is the full complement of DNA characterizing an individual organism and is inherent to all living beings

A **reference genome** is a tool used for research:

Some general features:

- Representative of the species
- Used alone for interspecific comparisons
- Used as a reference against which population-level resequencing data can be aligned
- Reference genomes act as catalogues of gene coding sequences and other functional components

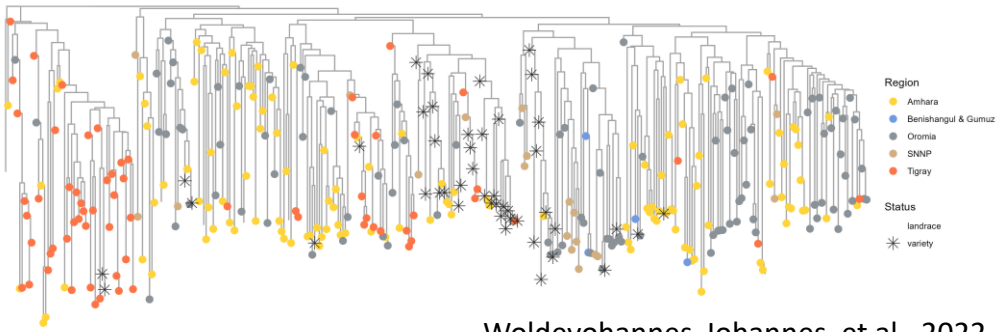


Genome resequencing

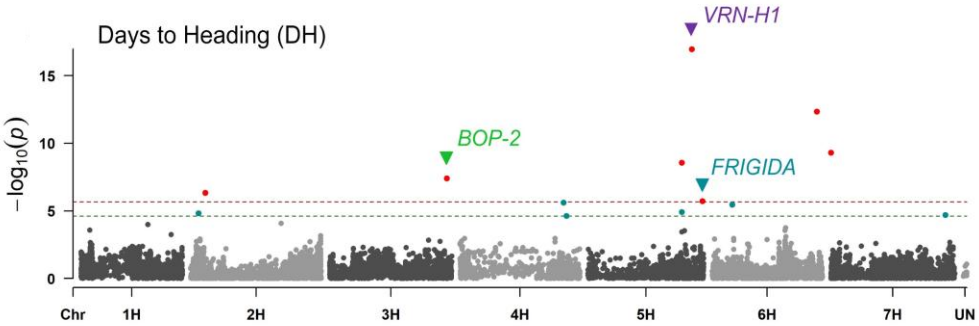
With NGS we observed a shift towards resequencing

Within species (or within population) genetic diversity

- Natural variation
- Demographic history
- Relation between phenotype and genotype



Woldeyohannes, Iohannes, et al., 2022



Caproni et al., 2023

Resequencing using NGS… in a nutshell

ADVANTAGES

↑ throughput

↓ per-base cost

~~BOTTLENECKS~~

↑ per-base error rate
complexity

↓ sequence length

COMPUTATIONAL CHALLEGES

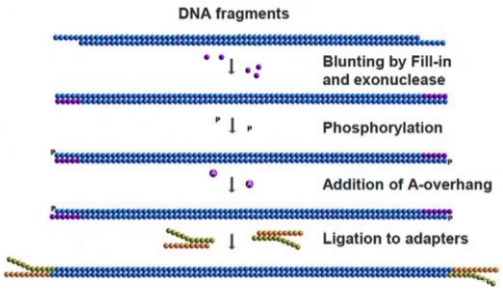
SOS

→ Multitude of bioinformatics algorithms help us managing these challenges

1. Data generation: **NGS**

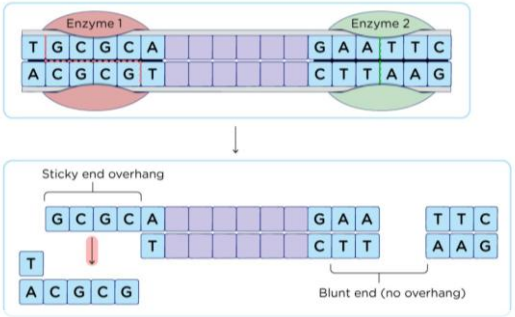
The DNA or the RNA of each organism took long journey to get to your PC…

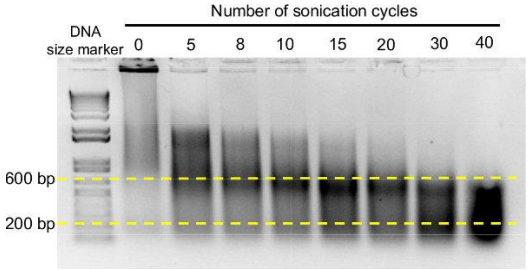
1. DNA extraction



3. Library Preparation

FASTQ files





2. DNA fragmentation

4. Sequencing

DNA fragmentation

Most NGS protocols start with preparation of libraries by **shearing** the DNA

Randomly

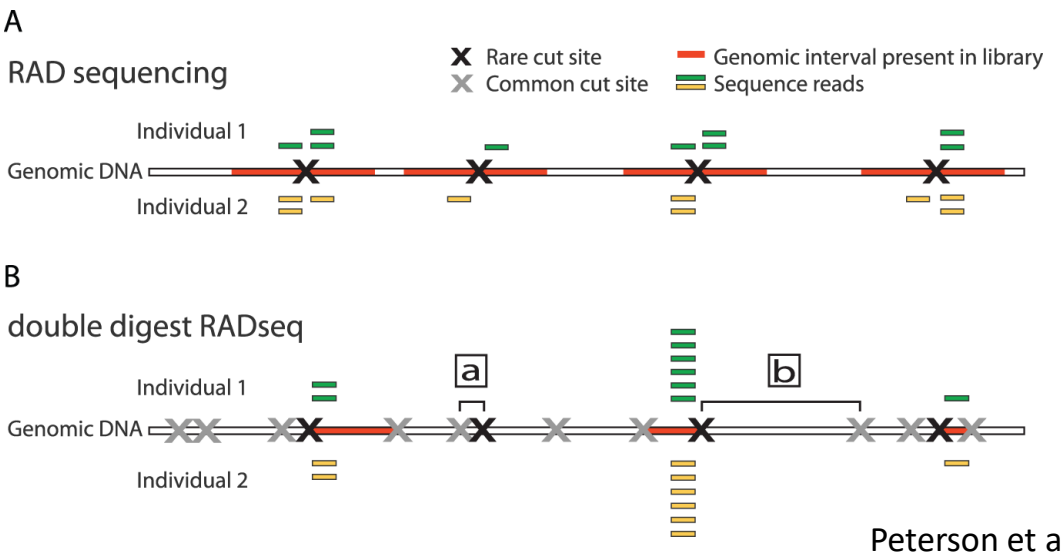
Physical

- Acoustic shearing
- Sonication
- Hydrodynamic

Chemical

- Metal cations (divalent cations like magnesium or zinc + heat – RNA)

Systematically (e.g RAD and ddRAD)



Peterson et al., 2014

ddRAD-seq: pros and cons

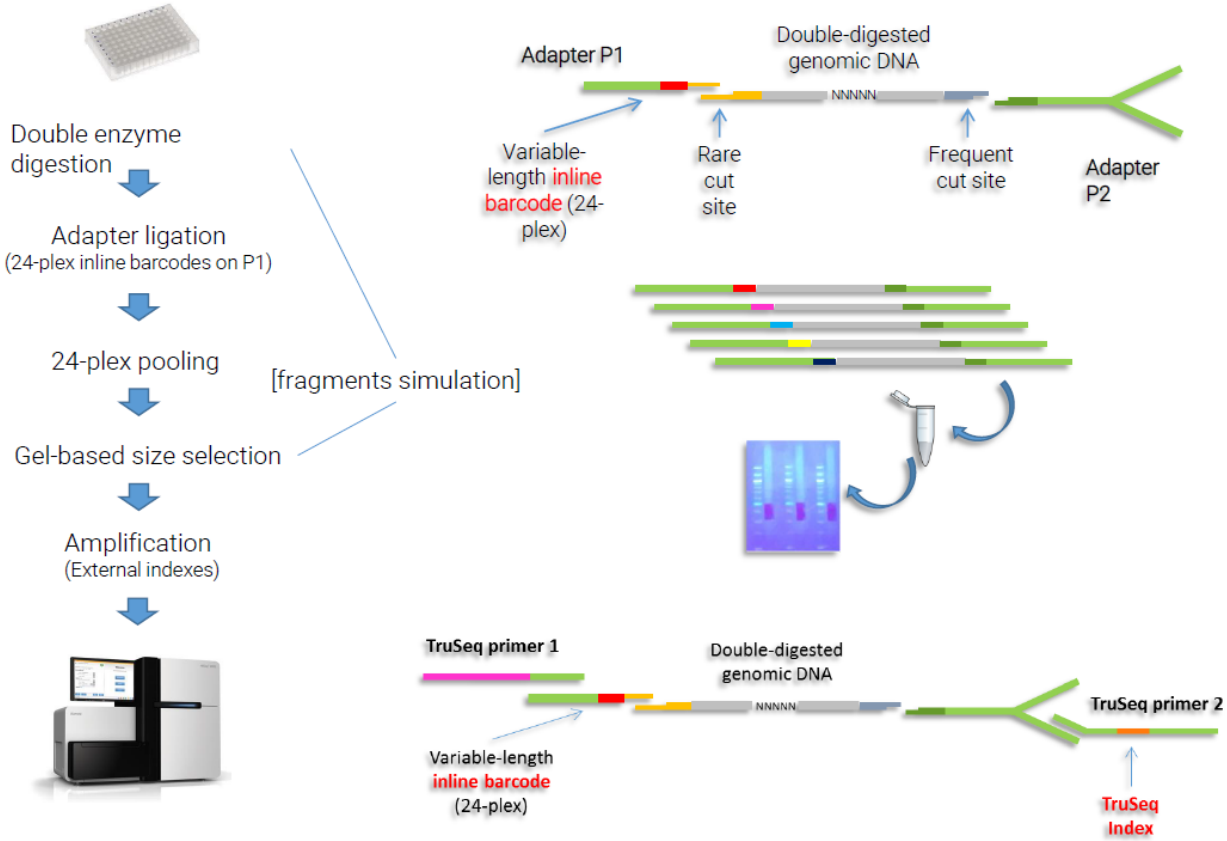
Pros:

- Relatively inexpensive, compared to whole-genome sequencing
- The degree of genome coverage can be adjusted by selecting various restriction enzymes

Cons:

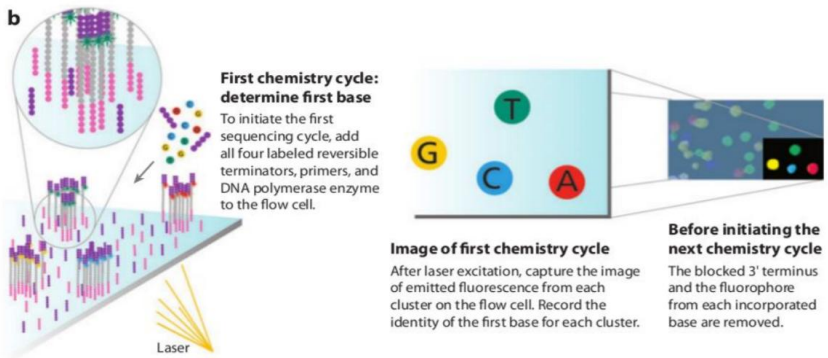
- There can be gaps in the genome coverage
- Requires high-quality DNA

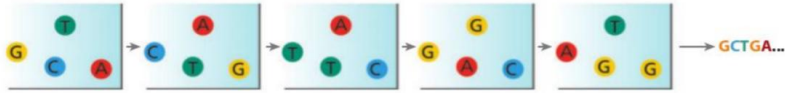
Platform specific adapters + multiplexing



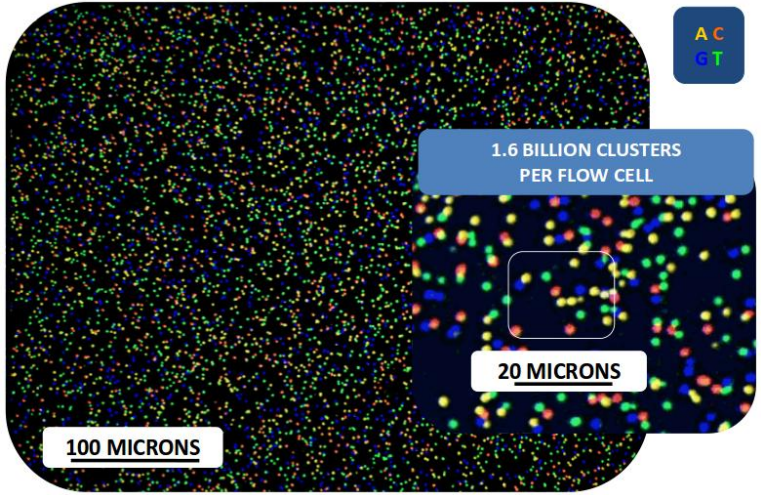
Illumina in brief

- A polymerase reaction is used to extend fragments from free primers
- Nt are labeled with different colours
- When a Nt is incorporated, it emits the corresponding light color
- A laser captures the light emission and maps it to a cluster





Sequence read over multiple chemistry cycles



Keep in mind about Illumina sequencing

1. The signal obtained from each cluster is the sequence of a single, *single-stranded DNA fragment*
2. In paired-end sequencing, sequence from both ends of the fragment (but not necessarily the middle) is obtained.
3. The technique lends itself to sequencing millions of “anonymous” chunks of DNA.
4. The indexes, or “barcodes” allow DNA from multiple individuals to be sequenced in a single run.
5. This is how most high-throughput resequencing is done today.

The FASTQ format

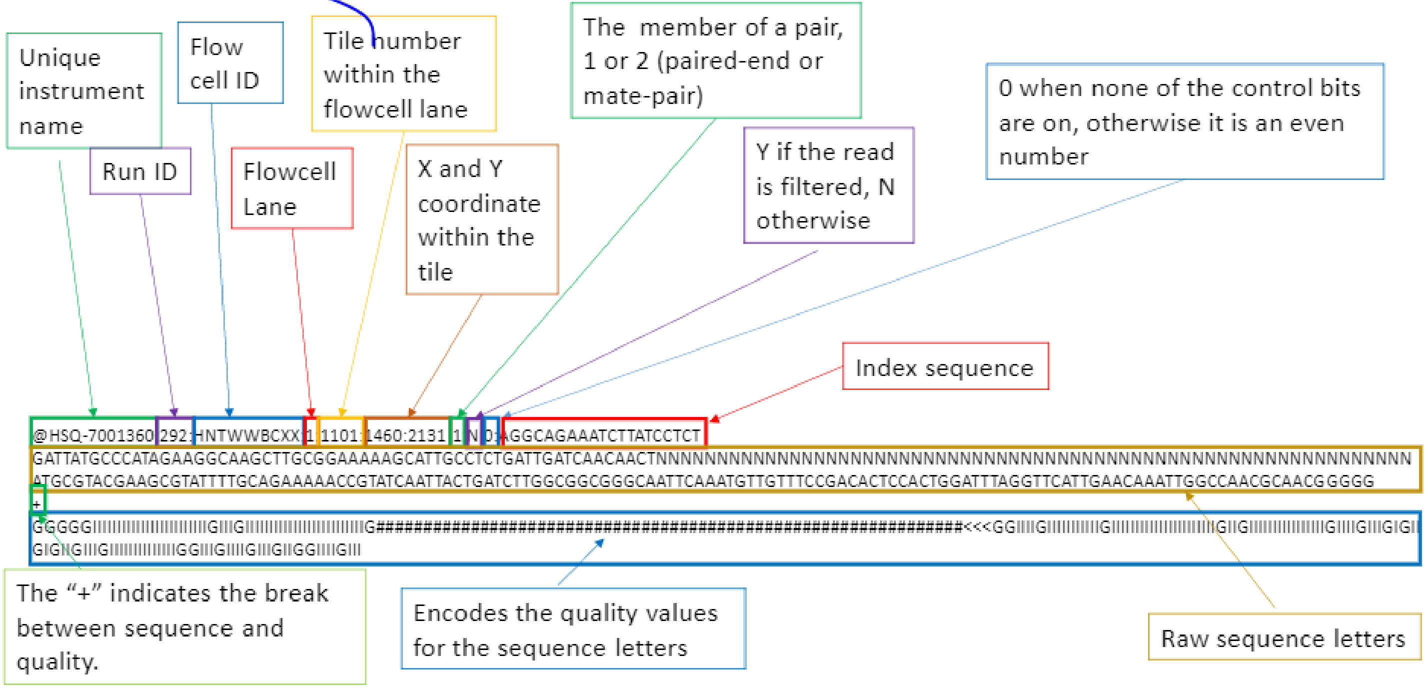
Light signals are **translated into base-calls associated with ASCII-encoded PHRED-like quality scores** (statistical measures of call certainty)

$$Q_{PHRED} = - \log_{10} p(error)$$

Ewing and Green 1998

*Depending on the platform information can be obtain as single-end or pair-end

FASTQ File Format Analysis



Unique instrument name

Run ID

Flow cell ID

Flowcell Lane

Tile number within the flowcell lane

X and Y coordinate within the tile

The member of a pair, 1 or 2 (paired-end or mate-pair)

Y if the read is filtered, N otherwise

0 when none of the control bits are on, otherwise it is an even number

Index sequence

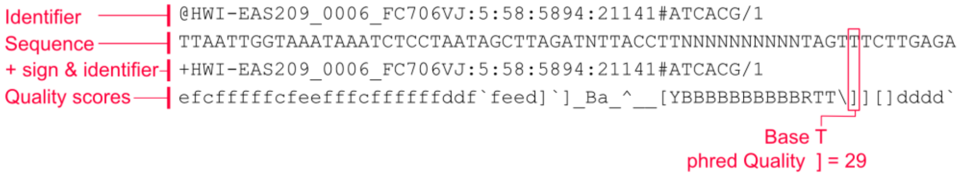
```
@HSQ-7001360:292:HNTWWBCXX:1:1101:1460:2131 1:N:0:AGGCAGAAATCTTATCCTCT
GATTATGCCCATAGAAGGCAAGCTTGCGGAAAAAGCATTGCCTCTGATTGATCAACAACTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
ATGCGTACGAAGCGTATTTTGCAGAAAAACCGTATCAATTACTGATCTTGGCGGCGGGCAATTCAAATGTTGTTTCCGACACTCCACTGGATTTAGGTTCATTGAACAAATTGGCCAACGCAACGGGGG
+
GGGGGIIIIIIIIIIIIIIIIIIIIGIIIGIIIIIIIIIIIIIIIIIIIIIIG######################################################################<<<GGIIIIGIIIIIIIIIGIIIIIIIIIIIIIIIIIIGIIGIIIIIIIIIIIIIIGIIIIGIIIGIGII
GIGIIGIIIGIIIIIIIIIIIIIIGGIIIGIIIIGIIIGIIGGIIIIGIII
```

The "+" indicates the break between sequence and quality.

Encodes the quality values for the sequence letters

Raw sequence letters

In summary

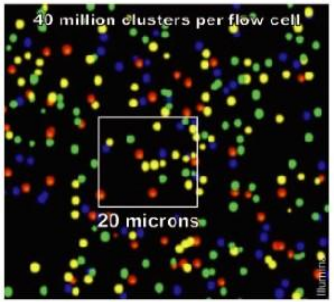


```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGA
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd`
```

Sequencing errors

Specific biases

- Library preparation -> **contamination**
- Amplification -> **errors of polymerase *in vitro***
- Sequencing -> **long-homopolymers** (or high content in GC**)**



Common biases across (Illumina) platforms

- **Signal intensity** decreases towards the end of each read
- Incomplete read extension or termination -> **desynchronization of clusters**
- Particles -> **chemical crystallization** (or others)

2. Quality control and data processing

We want to make sure that samples have been sequenced correctly, with minimal contamination and sufficient coverage

SUMMARY STATISTICS

- Nucleotide quality scores distributions
- Sequence characteristics (including GC content)

FastQC - fast Quality Control

http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc

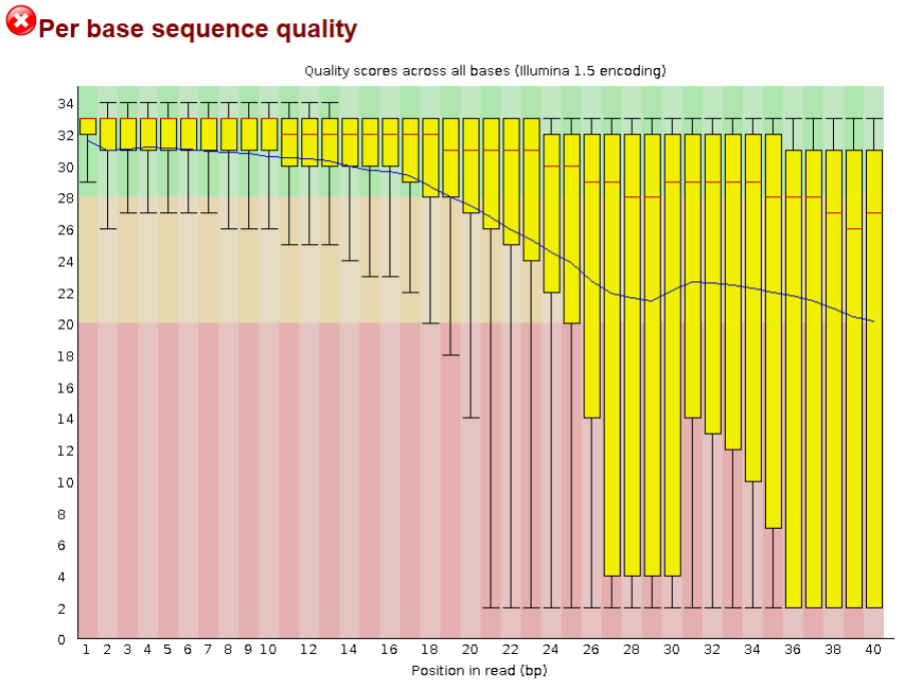
Issue 1: low quality data

Good seq data = most reads have high PHRED-like score

When sequences with universally low-quality scores are present these should be removed from subsequent analyses

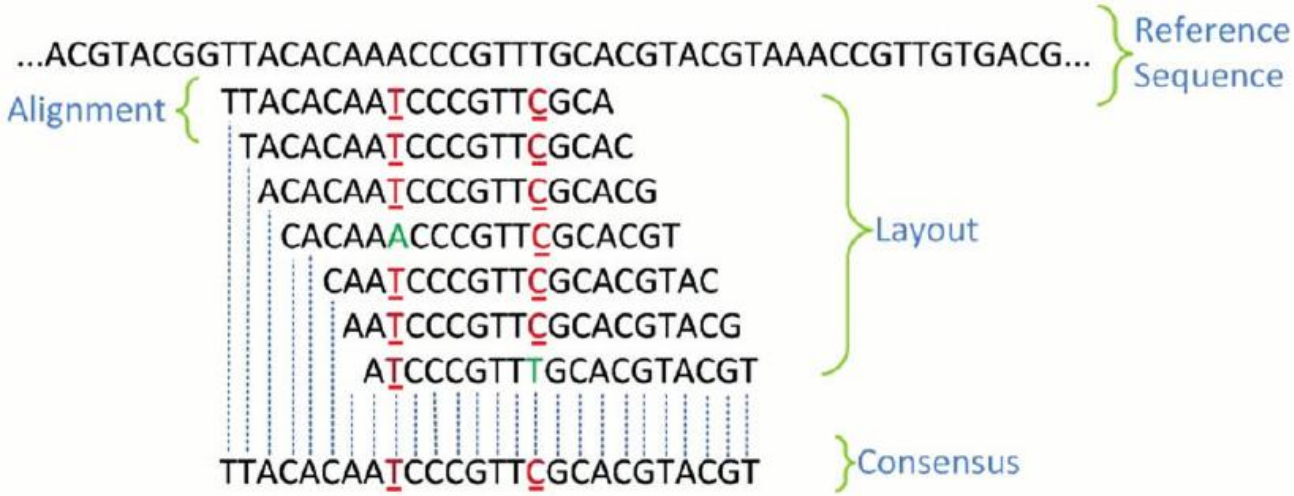
Causes:

- Bubbles
- Spot-specific signal noise
- Problems with readout (*e.g.* edge of the flow cell)



Managing erroneous and/or missing calls

- Modify **low-frequency patterns** and superimposing **high-frequency consensus sequences**



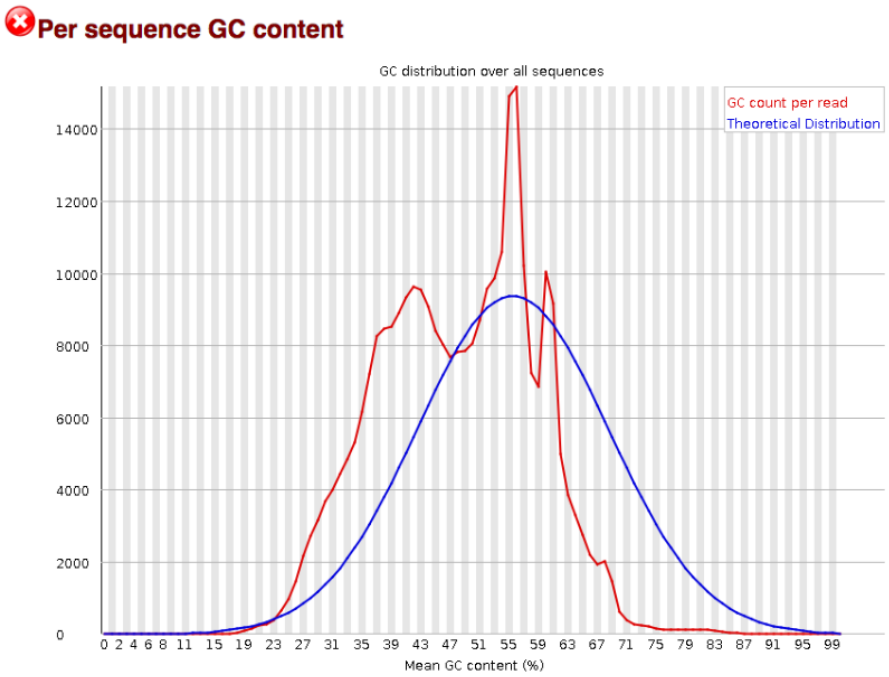
Issue 2: presence of adapters of exogenous sequences

Read longer than the target length

partial/complete sequencing of adapters/primers:

- Issues in downstream alignment (False positive variants)
- When this occurs in 5’ is highly problematic (most aligners require high similarity in 5’)
  - Even more problematic with projects using short reads (ancient DNA of forensic samples)

Per-base nucleotide proportion along reads and GC content

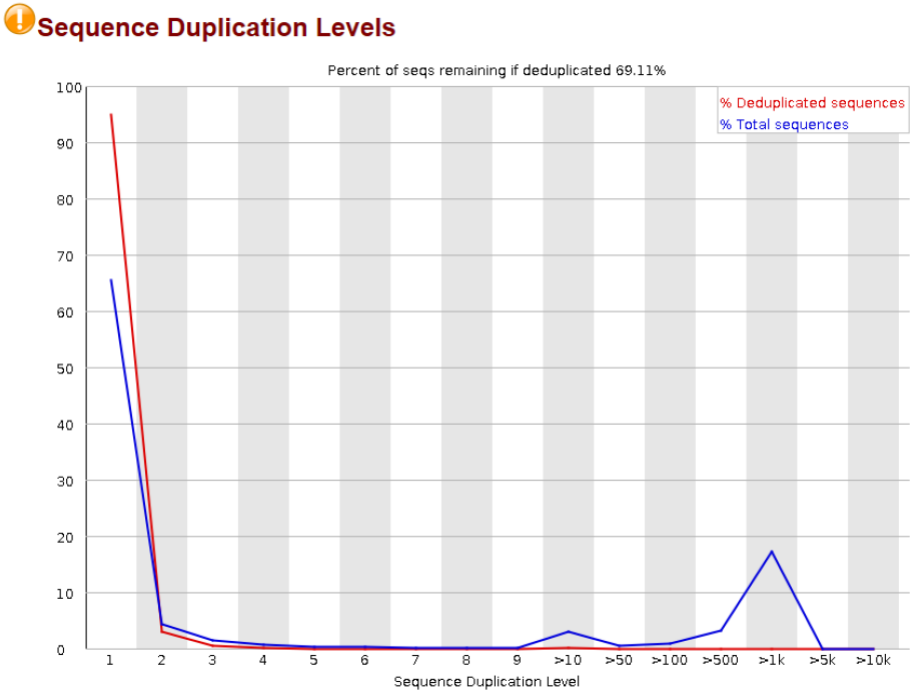


Issue 3: enrichment bias

Low levels of sequence duplication

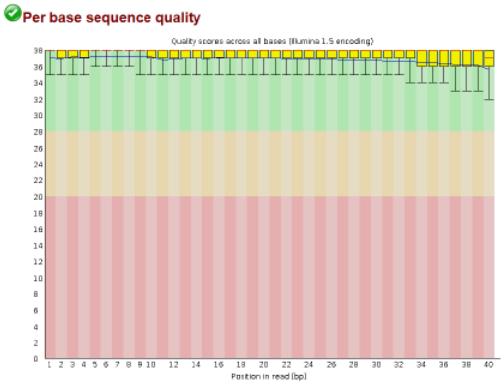
- Suggest that a given library has been sequenced with high coverage.
- In contrast high levels of sequence duplication arise from:
  - Technical artefacts (*e.g.* PCR overamplification)
  - Biological duplication

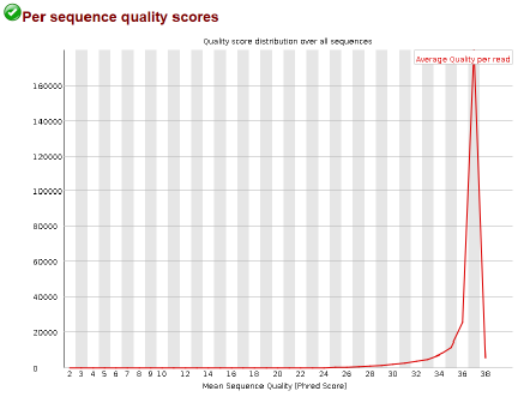
Keep in mind: Moderate levels of seq duplication (10-50 copies) are expected in approaches like ddRAD

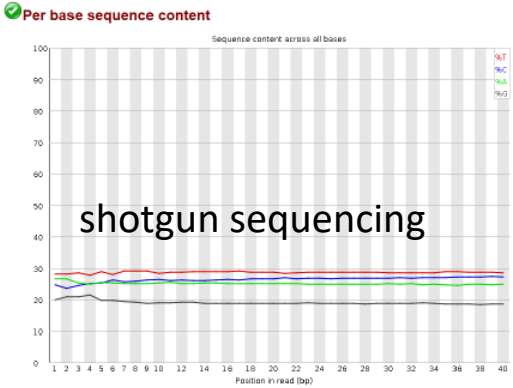


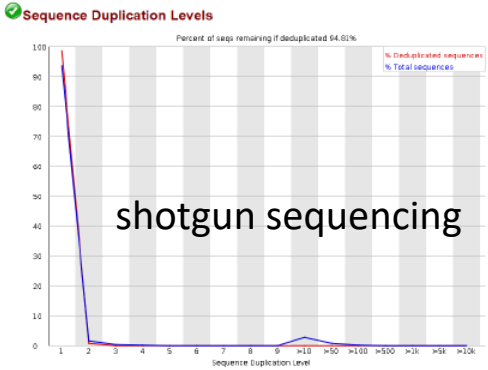
True for shotgun sequencing

Example of **GOOD** quality reads

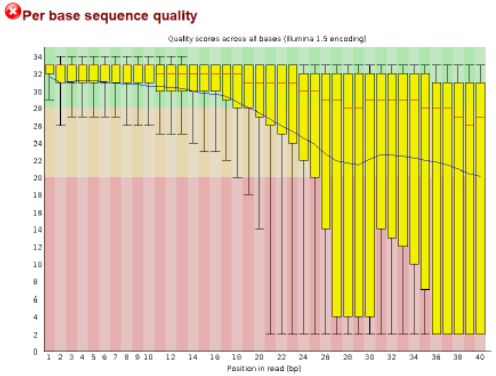


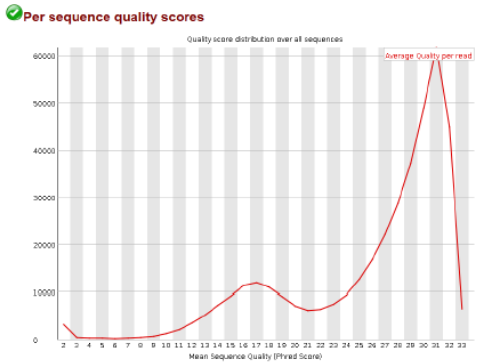


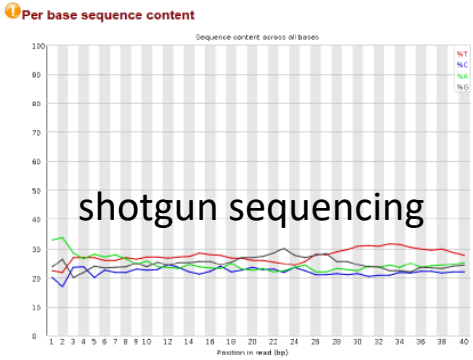


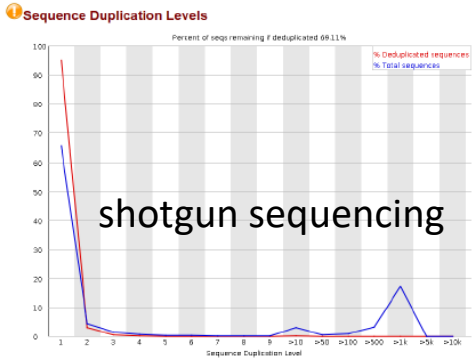


Example of **BAD** quality reads







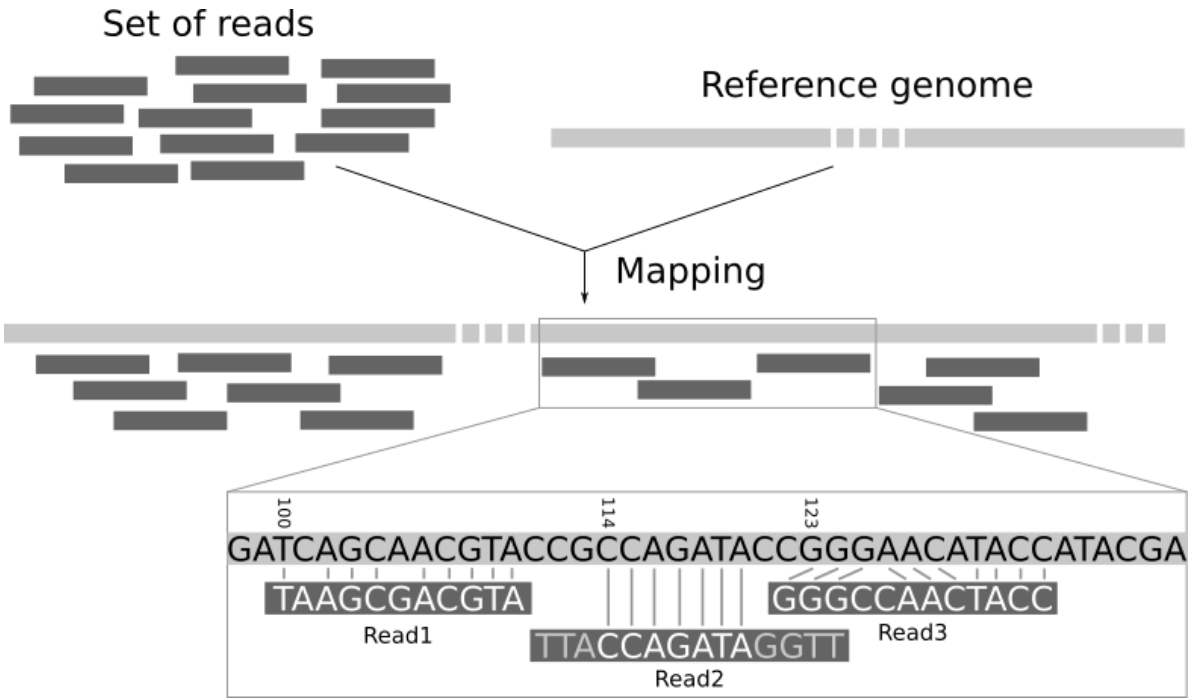


Popular tools for preprocessing

Pfeifer 2017, Heredity

Software	*Ability of the software to:*			*Reference*
	Handle multiple adapter sequences	*Trim low-quality bases*	*Demultiplex barcodes*	
AdapterRemoval	–	+	–	Lindgreen (2012)
AlienTrimmer	+	+	–	Criscuolo and Brisse (2013)
Btrim	–	+	+	Kong (2011)
CONDETRI	–	+	–	Smeds and Künstner (2011)
Cutadapt	+	+	–	NA
EA-Utils	–	+	+	NA
ERNE-FILTER	–	+	–	Del Fabbro *et al.* (2013)
FASTX-Toolkit	–	–	+	NA
Flexbar	+	+	+	Dodt *et al.* (2012)
Kraken	+	+	+	Davis *et al.* (2013)
NGSQC	+	+	–	Dai *et al.* (2010)
ngsShoRT	+	+	–	Chen *et al.* (2014)
PEAT	+	–	–	Li *et al.* (2015)
PRINSEQ	–	+	–	Schmieder and Edwards (2011b)
QC-Chain	+	+	–	Zhou *et al.* (2013)
QcReads	+	+	–	Ma *et al.* (2013)
Reaper	+	+	+	NA
SeqTrim	+	+	–	Falgueras *et al.* (2010)
Sickle	–	+	–	NA
Skewer	+	+	+	Jiang *et al.* (2014)
TrimGalore!	–	+	–	NA
Trimmomatic	+	+	–	Bolger *et al.* (2014)

3. Alignment (aka mapping)



Credits: Galaxy training Wolff et al. 2022

3. Alignment (aka mapping)

This is the step in which any NGS study is based upon

State-of-art aligners consider several factors to circumvent possible issues:

- **Gapped alignments** (prevent false positives deriving from 'ungapped' alignment)
- **Integration of quality scores** (deal with low-quality regions)

The product of alignment is stored in SAM/BAM files

- usually manipulated with or Picard or SAMtools (Li et al. 2009) -

SAM files

A

```
Coor      12345678901234  5678901234567890123456789012345
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1         TTAGATAAAGGATA*CTG
+r002          aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                     CAGCGGCAT
```

Example of alignment: Read r001/1 and r001/2 constitute a read pair; r003 is a chimeric read; r004 represents a split alignment (there is a gap).

B

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M * 0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       * 0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    * 0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       * 0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         = 7 -39 CAGCGGCAT         * NM:i:1
```

Header section

QUAL (read quality; * meaning such information is not available)

Alignment section

QNAME (query template name, aka. read ID)

FLAG (indicates alignment information about the read, e.g. paired, aligned, etc.)

RNAME (reference sequence name, e.g. chromosome /transcript id)

POS (1-based position)

MAPQ (mapping quality)

CIGAR (summary of alignment, e.g. insertion, deletion)

RNEXT (reference sequence name of the primary alignment of the NEXT read; for paired-end sequencing, NEXT read is the paired read; corresponding to the RNAME column)

PNEXT (Position of the primary alignment of the NEXT read in the template; corresponding to the POS column)

TLEN (the number of bases covered by the reads from the same fragment. In this particular case, it's 45 - 7 + 1 = 39 as highlighted in Panel A). Sign: plus for leftmost read, and minus for rightmost read

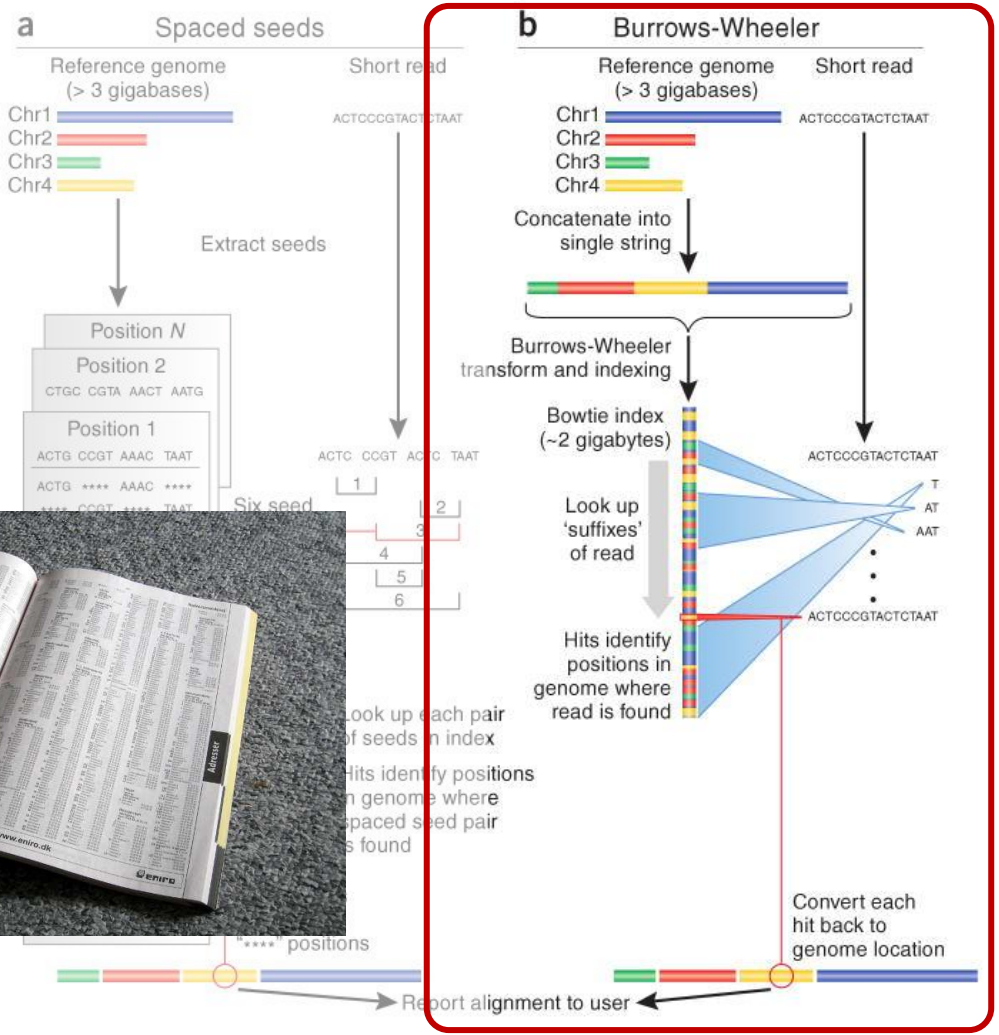
SEQ (read sequence)

Optional fields in the format of TAG:TYPE:VALUE

SAM stands for **Sequence Alignment/Map format**; **BAM** is its corresponding binary version (compressed – that is the actual format used to store alignment results)

The Burrows-Wheeler transform, store a memory-efficient representation of the reference genome.

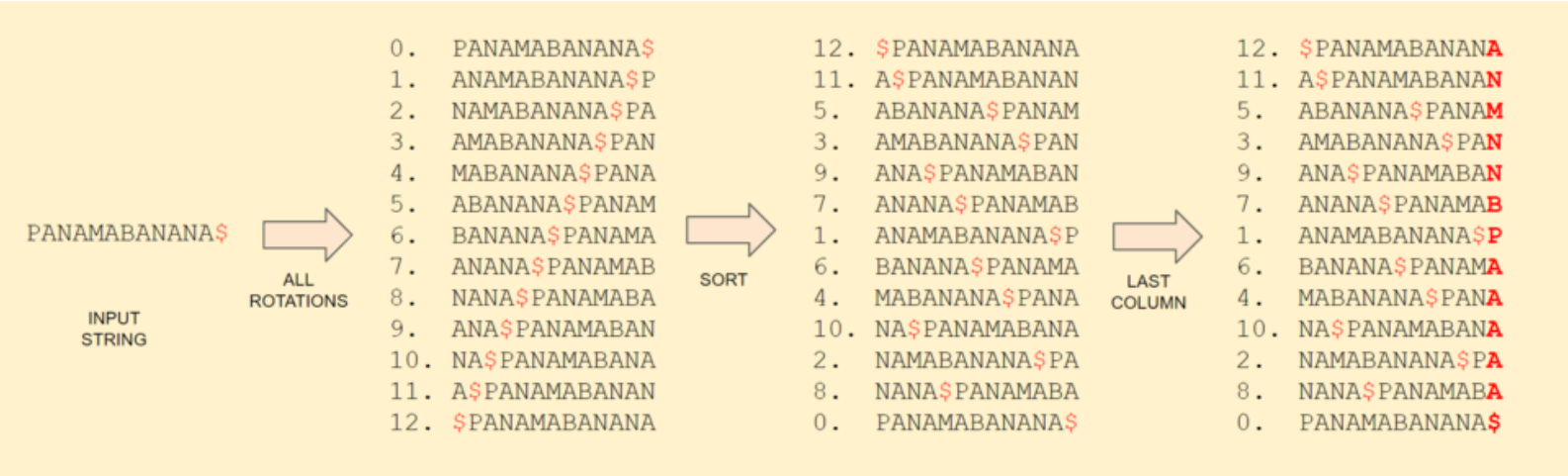
- Reads are aligned character by character from right to left against the transformed string.
- With each new character, the algorithm updates an interval (indicated by blue 'bear transformed string.
- When all characters in t been processed, alignm represented by any pos the interval.



Trapnell and Salzberg, 2009, Nature Biotech

The **Burrows-Wheeler algorithm**

How it works… PANAMABANANA$

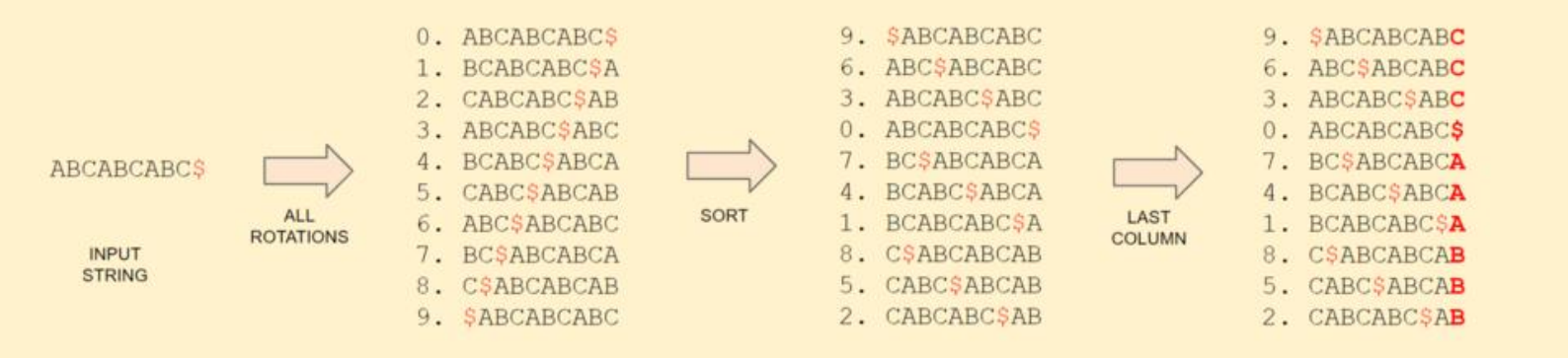


The final transformed string we get is “ANMNNBPAAAAA$”

Which we can store as “**ANM2NBP5A$**”

The **Burrows-Wheeler algorithm**

Good compression



“3C$3A3B”

If there are substrings which occur often in the, then there will be more characters together. This turns out to be great for compressing strings with repeats, like the DNA sequences, where we have only 4 characters (A, C, G, T) and a lot of repeated patterns.

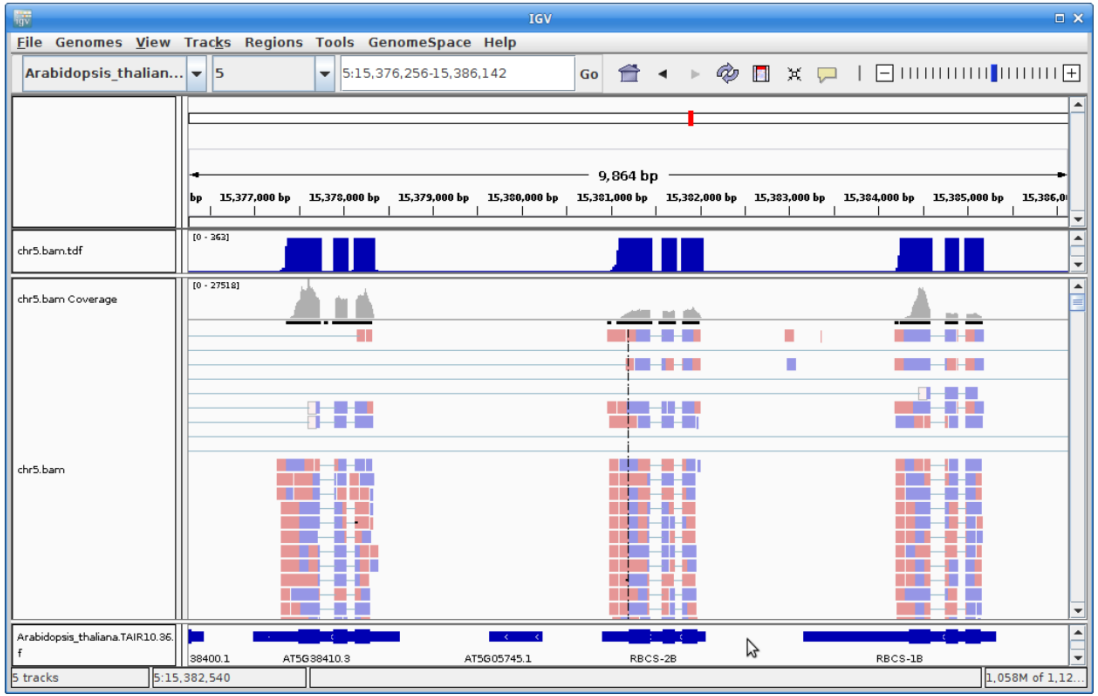
Open-source/binary NGS aligners

Software	*Sequencing platform*	*Ability to perform gapped alignment*	*Quality awareness*	*Ability to align PE reads*	*Reference*
BFAST	I,4	+	–	+	Homer *et al.* (2009)
Bowtie	I,4,Sa	–	+	+	Langmead *et al.* (2009)
Bowtie 2	I,4,Ion	+	+	+	Langmead and Salzberg (2012)
BWA	I,4,Sa	+	+	+	Li and Durbin (2009)
CloudBurst	non-specific	+	–	–	Schatz (2009)
GSNAP	I,4,Sa,Ion	+	–	+	Wu and Nacu (2010)
MAQ	I	–	+	+	Li *et al.* (2008)
MOSAIK	I,4,Sa,Ion	+	+	+	NA
mrFAST	I	–	+	+	Alkan *et al.* (2009)
mrsFAST	I	–	+	+	Hach *et al.* (2010)
NextGenMap	I,4,Ion	+	–	+	Sedlazeck *et al.* (2013)
PASS	I,4	+	+	+	Campagna *et al.* (2009)
RazerS	I,4	+	–	+	Weese *et al.* (2009)
segemehl	I,4,Sa,Ion	+	–	+	Hoffmann *et al.* (2009)
SHRiMP	I,4	+	–	+	Rumble *et al.* (2009)
SHRiMP 2	I,4	–	+	+	David *et al.* (2011)
SOAP2	I	+	–	+	Li *et al.* (2009b)
Stampy	I	+	+	+	Lunter and Goodson (2011)

Pfeifer 2017, Heredity

Visualizing alignments

- BamViewer (Carver et al., 2010)
- Gap5 (Bonfield and Whitwham 2010)
- **Integrative Genomics Viewer** (IGV, Broad Institute, Robinson et al., 2011, example in Figure)
- MapViewer (Boa et al., 2009)
- Tablet (Milne et al., 2009)
- SAMtools (Text Aligner Viewer, Li et al., 2009)



4. Alignment post-processing

Before variant calling, we need to **detect** and **correct** spurious alignments as to minimize artifacts

Potential issues at this stage to consider are:

1. Real unidentified insertions or deletions may cause spurious variant calls -> **Local alignment**
2. Some library preparation approaches may induce artificial coverage -> **not true for approaches like ddRAD**
3. Row base quality scores may not reflect true quality -> **recalibration based on position**

5. Variant calling and filtering

Definitions:

- **VARIANTS**: positions in which at least one individual differs from the reference
- **GENOTYPING**: individual alleles at all variant sites are estimated
- **FILTERING**: removal of false positives from initial variant data to improve specificity

The VCF format

Example

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0|1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1|0:77	1/1:95
1	100		T	<DEL>	.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Phased data (G and C above are on the same chromosome)

Deletion

SNP

Large SV

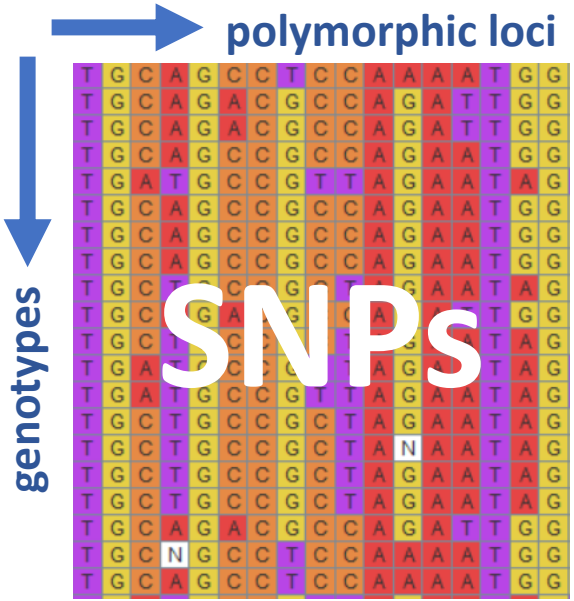
Insertion

Other event

Single-nucleotide polymorphism

DEFINITION: a germline substitution of a single nucleotide at a specific position in the genome and is present in a sufficiently large fraction of the population (1% or more).

Reference	ATTCGCTCAGATTACAAACTACTTA
Ind 1	ATTCGCTCAGATTACAAACTACTTA
Ind 2	ATTCGCTCAGATTACAAACTACTTA
Ind 3	ATTCGCACAGATTACAAACTACTTA
Ind 4	ATTCGCTCAGATTACAAACTACTTA
Ind 5	ATTCGCACAGATTACAAACTACTTA
Ind 6	ATTCGCTCAGATTACAAACTACTTA
Ind 7	ATTCGCTCAGATTACAAACTACTTA
Ind 8	ATTCGCACAGATTACAAACTACTTA
Ind 9	ATTCGCTCAGATTACAAACTACTTA



Recommended read

Heredity (2017) 118, 111–124

www.nature.com/hdy

REVIEW

From next-generation resequencing reads to a high-quality variant data set

SP Pfeifer[1,2,3]

Large-scale whole Genome resequencing projects

Article

The sequences of 150,119 genomes in the UK Biobank

https://doi.org/10.1038/s41586-022-04965-x

Received: 5 November 2021

Accepted: 10 June 2022

Published online: 20 July 2022

Open access

Check for updates

Bjarni V. Halldorsson[1,2]✉, Hannes P. Eggertsson[1], Kristjan H. S. Moore[1], Hannes Hauswedell[1], Ogmundur Eiriksson[1], Magnus O. Ulfarsson[1,3], Gunnar Palsson[1], Marteinn T. Hardarson[1,2], Asmundur Oddsson[1], Brynjar O. Jensson[1], Snaedis Kristmundsdottir[1,2], Brynja D. Sigurpalsdottir[1,2], Olafur A. Stefansson[1], Doruk Beyter[1], Guillaume Holley[1], Vinicius Tragante[1], Arnaldur Gylfason[1], Pall I. Olason[1], Florian Zink[1], Margret Asgeirsdottir[1], Sverrir T. Sverrisson[1], Brynjar Sigurdsson[1], Sigurjon A. Gudjonsson[1], Gunnar T. Sigurdsson[1], Gisli H. Halldorsson[1], Gardar Sveinbjornsson[1], Kristjan Norland[1], Unnur Styrkarsdottir[1], Droplaug N. Magnusdottir[1], Steinunn Snorradottir[1], Kari Kristinsson[1], Emilia Sobech[1], Helgi Jonsson[4,5], Arni J. Geirsson[4], Isleifur Olafsson[4], Palmi Jonsson[4,5], Ole Birger Pedersen[6], Christian Erikstrup[7,8], Søren Brunak[9], Sisse Rye Ostrowski[10,11], DBDS Genetic Consortium*, Gudmar Thorleifsson[1], Frosti Jonsson[1], Pall Melsted[1,3], Ingileif Jonsdottir[1,5], Thorunn Rafnar[1], Hilma Holm[1], Hreinn Stefansson[1], Jona Saemundsdottir[1], Daniel F. Gudbjartsson[1,3], Olafur T. Magnusson[1], Gisli Masson[1], Unnur Thorsteinsdottir[1,5], Agnar Helgason[1,12], Hakon Jonsson[1], Patrick Sulem[1] & Kari Stefansson[1]✉