3000788 Intro to Comp Molec Biol

Lecture 4: Sequencing data processing

Fall 2025

Sira Sriswasdi, PhD

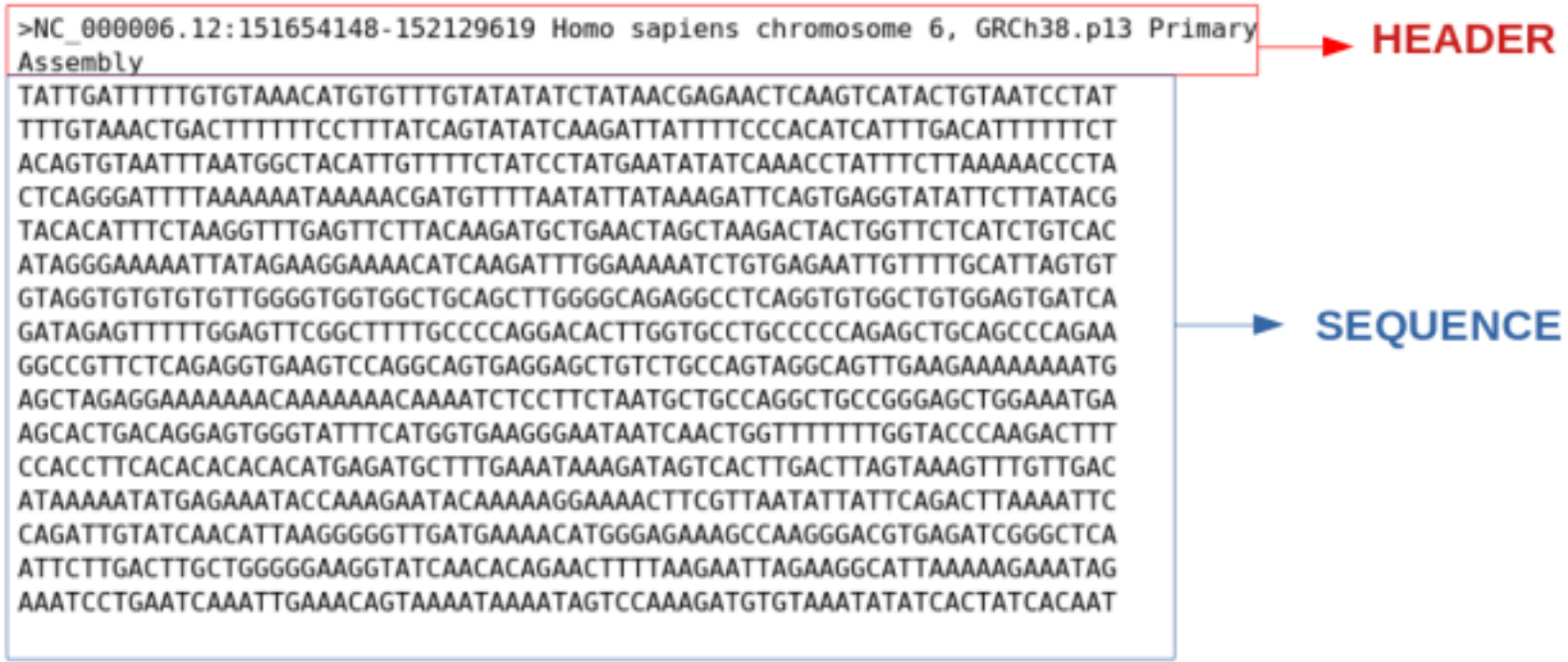
- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Today's agenda

- Sequencing data file formats
- Quality check and processing of sequencing reads
- Read-to-genome alignment

Sequencing data

FASTA format

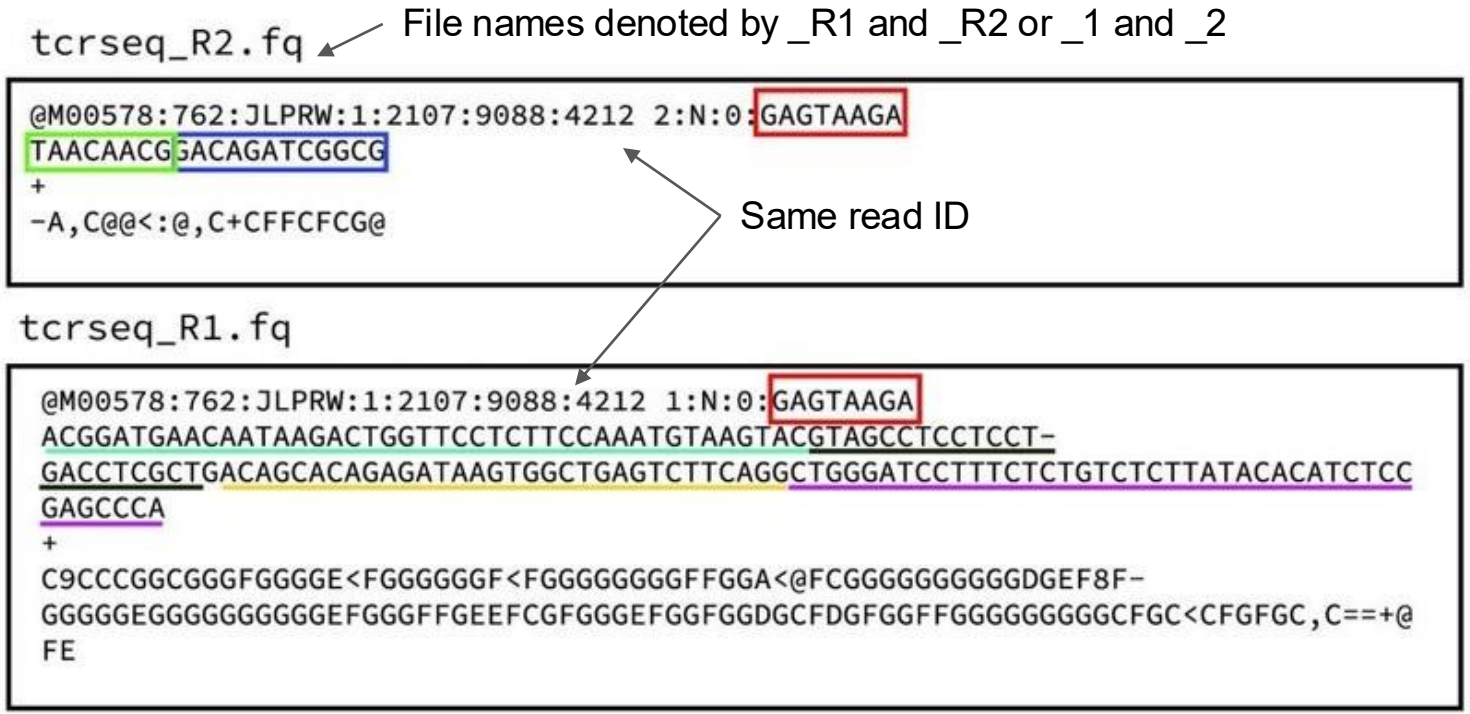


FASTQ format

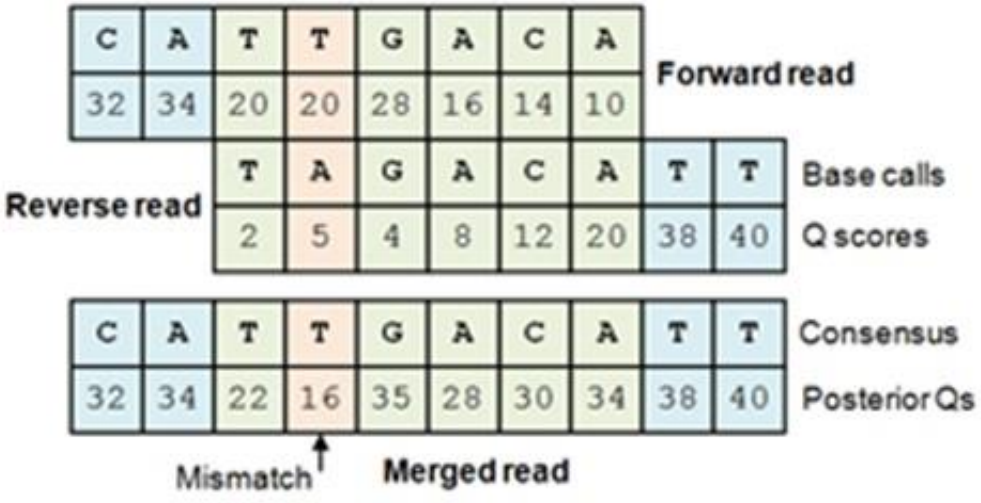
```
1  @ERR000589.41 EAS139_45:5:1:2:111/1
2  CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCCAGGGAACATCTTGTCAT
3  +
4  3IIIIIIIIIIIII>1IIIFF9BG08E00I%IG+&?(4)%00646.C1#&(
5  @ERR000589.42 EAS139_45:5:1:2:1293/1
6  AGTTGTTAAAATCCAAGCCAATTAAGATAGTCTTATCTTTTTAAAAGAAAT
7  +
8  IIIIIGII.AIIII=?I9G-/II=+I=4?761BA2C9I+5A711+&>1$/I
```

- Header: Location of cluster on Illumina's flow cell
- Sequence
- Quality score

FASTQ for paired-end sequencing



Merged FASTQ for paired-end sequencing



https://drive5.com/usearch/manual8.1/merge_pair.html

- Forward & reverse reads must overlap (read length > fragment length)
- **Example**: 300bp paired-end sequencing of 16S rRNA genes

Sequencing quality score (Phred score)

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

https://www.drive5.com/usearch/manual/quality_score.html

- Q score = – 10 x Log_{10} (base call error rate)
- Base call error of 10% → Q score = **+**
- Base call error of 0.0001 → Q score = **I**

Increased error toward the ends of read

```
@ERR000589.41 EAS139_45:5:1:2:111/1
CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCCAGGGAACATCTTGTCAT
+
3IIIIIIIIIIIII>1IIIFF9BG08E00I%IG+&?(4)%00646.C1#&(
```

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

Nanopore FAST5 format

```
HDF5 "/home3/ont/lambda_fc1/downloads/pass/vgb_20170110_FNFAB46402_MN19940_sequencing_run_lambdacontrol_10012017_23602_ch9_read939_strand.fast5" {
DATASET "/Raw/Reads/Read_939/Signal" {
   DATATYPE  H5T_STD_I16LE
   DATASPACE  SIMPLE { ( 142677 ) / ( H5S_UNLIMITED ) }
   DATA {
      1216, 653, 494, 487, 468, 478, 510, 535, 506, 454, 476, 483, 475, 488,
      472, 505, 474, 474, 488, 485, 480, 493, 481, 479, 485, 481, 472, 491, 493,
      480, 480, 487, 477, 500, 484, 488, 486, 493, 458, 480, 491, 487, 477, 489,
      478, 485, 476, 489, 486, 488, 490, 480, 480, 484, 493, 475, 486, 477, 478,
      489, 481, 482, 492, 480, 474, 486, 426, 483, 508, 486, 487, 479, 476, 486,
      473, 485, 487, 484, 456, 485, 484, 466, 466, 483, 484, 484, 474, 480, 498,
      481, 484, 483, 477, 479, 473, 488, 482, 480, 478, 496, 479, 490, 489, 483,
      487, 473, 477, 479, 478, 480, 474, 475, 472, 475, 486, 498, 503, 481, 493,
```

```
GROUP "/Raw/Reads/Read_939" {
   ATTRIBUTE "duration" {
      DATATYPE  H5T_STD_U32LE
      DATASPACE  SCALAR
      DATA {
      (0): 142677
      }
   }
```

```
ATTRIBUTE "read_id" {
   DATATYPE  H5T_STRING {
      STRSIZE 37;
      STRPAD H5T_STR_NULLTERM;
      CSET H5T_CSET_ASCII;
      CTYPE H5T_C_S1;
   }
```

https://bioinformatics.cvr.ac.uk/exploring-the-fast5-format/

- Ion flow rate data through each nanopore, with time stamps
- **FAST5** is an **HDF5** file (a specialized compression for scientific datasets)

Quality check for sequencing data

FastQC tool

Basic Statistics

Measure	Value
Filename	small_rna.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	100
%GC	45

Check number of reads and read length

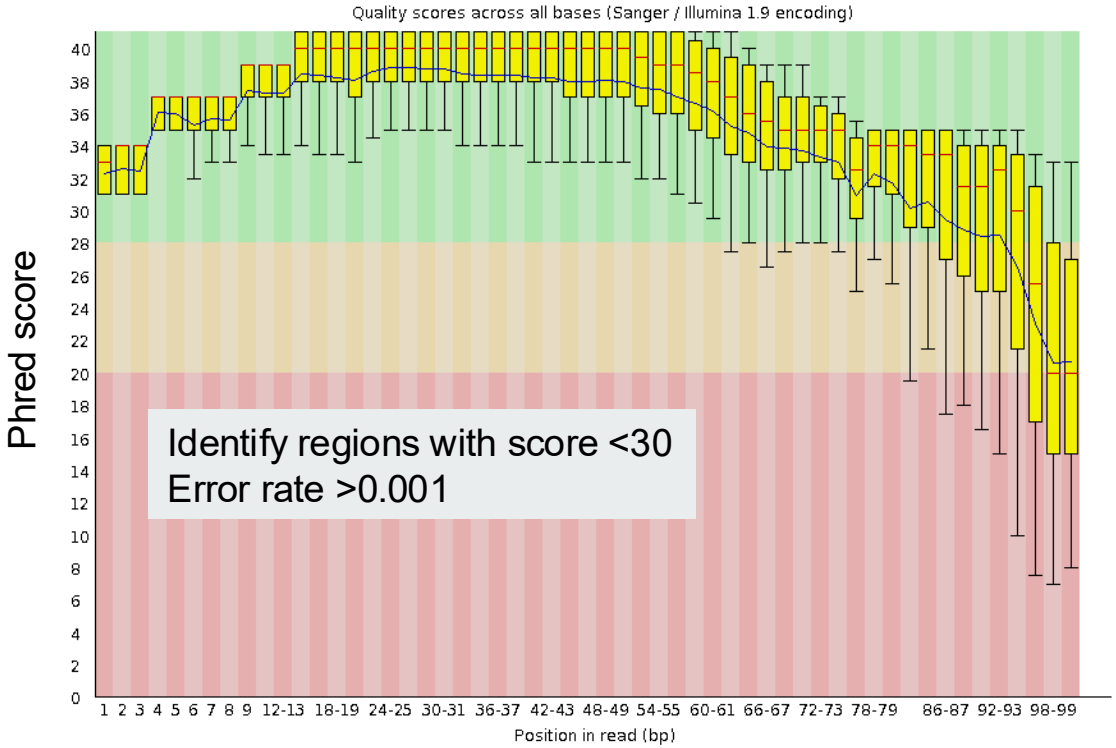
FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Base calling quality

Per base sequence quality



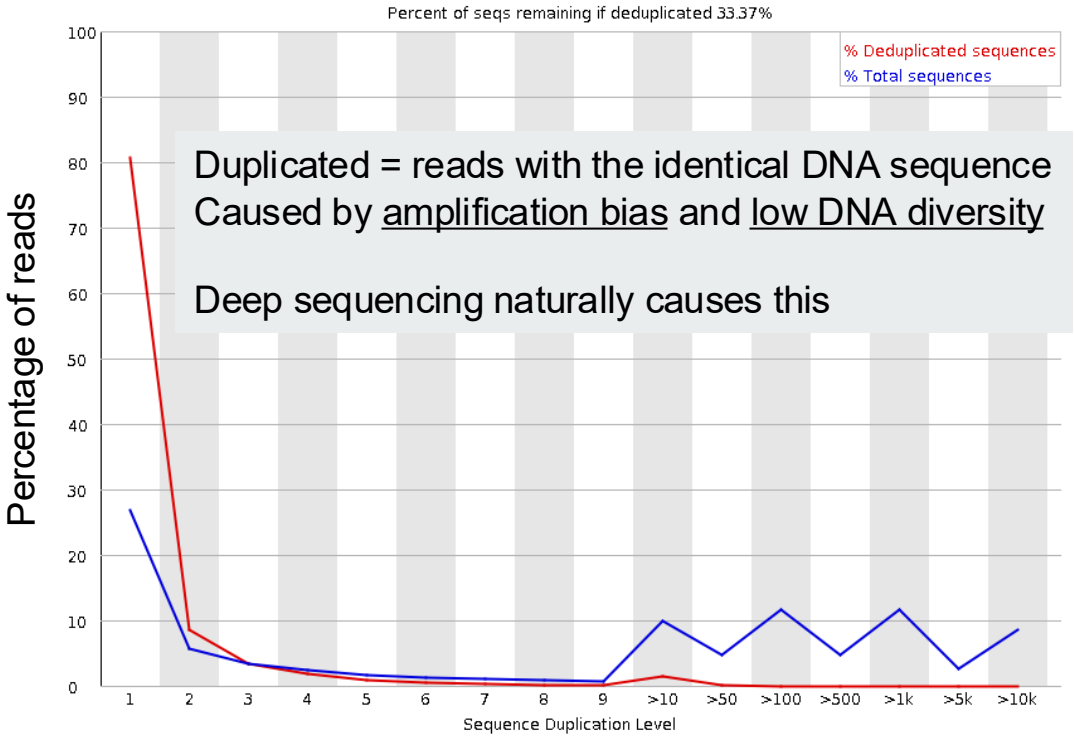
FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Duplicated reads

Sequence Duplication Levels



Duplicated = reads with the identical DNA sequence
Caused by amplification bias and low DNA diversity

Deep sequencing naturally causes this

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Possible adapter read-through

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
TGAGGTAGTAGATTGTATAGTTAGATCGGAAGAGCACACGTCTGAACTCC	10865	4.346	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TAGCTTATCAGACTGATGTTGACAGATCGGAAGAGCACACGTCTGAACTC	10845	4.338	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)
TCTTTGGTTATCTAGCTGTATGAGATCGGAAGAGCACACGTCTGAACTCC	7062	2.8247999999999998	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TCTTTGGTTATCTAGCTGTATGAAGATCGGAAGAGCACACGTCTGAACTC	4056	1.6223999999999998	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)
TGAGGTAGTAGTTTGTGCTGTTAGATCGGAAGAGCACACGTCTGAACTCC	3737	1.4948	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TGAGGTAGTAGTTTGTACAGTTAGATCGGAAGAGCACACGTCTGAACTCC	3549	1.4196	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TGAGGTAGTAGGTTGTATGGTTAGATCGGAAGAGCACACGTCTGAACTCC	2931	1.1724	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
AACCCGTAGATCCGATCTTGTAGATCGGAAGAGCACACGTCTGAACTCCA	1910	0.764	Illumina Multiplexing PCR Primer 2.01 (100% over 29bp)
CGCGACCTCAGATCAGACGTAGATCGGAAGAGCACACGTCTGAACTCCAG	1749	0.6996	Illumina Multiplexing PCR Primer 2.01 (100% over 30bp)
TGAGGTAGTAGGTTGTATAGTTAGATCGGAAGAGCACACGTCTGAACTCC	1647	0.6588	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TCTTTGGTTATCTAGCTGTATAGATCGGAAGAGCACACGTCTGAACTCCA	1622	0.6487999999999999	Illumina Multiplexing PCR Primer 2.01 (100% over 29bp)
TAGCTTATCAGACTGATGTTGATAGATCGGAAGAGCACACGTCTGAACTC	1328	0.5312	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)
TTCAAGTAATCCAGGATAGGCTAGATCGGAAGAGCACACGTCTGAACTCC	1248	0.4992	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
AGCAGCATTGTACAGGGCTATGAAGATCGGAAGAGCACACGTCTGAACTC	1248	0.4992	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)

Check whether reads contain sequencing adapter
Must be removed!

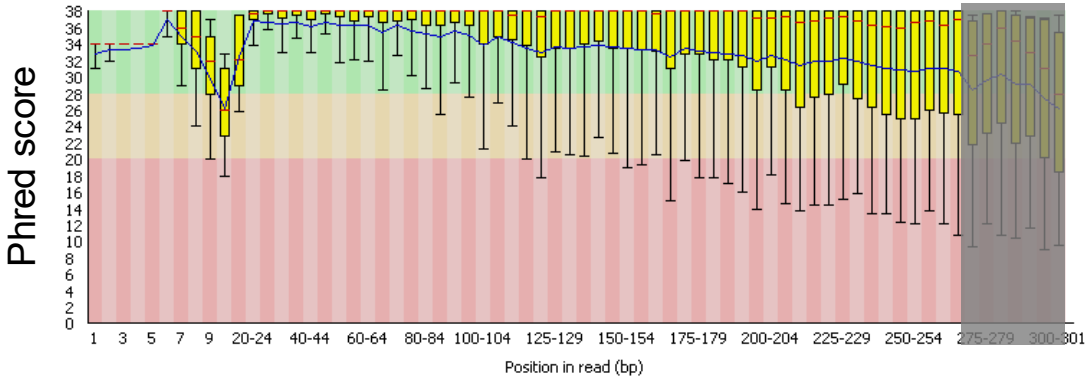
FastQC Report

Summary

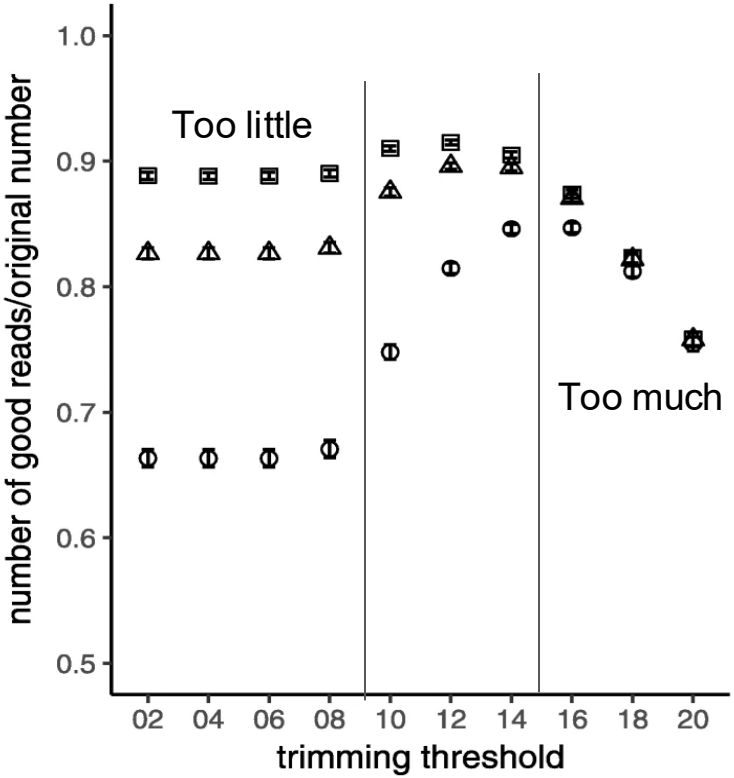
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Read trimming

Quality trimming

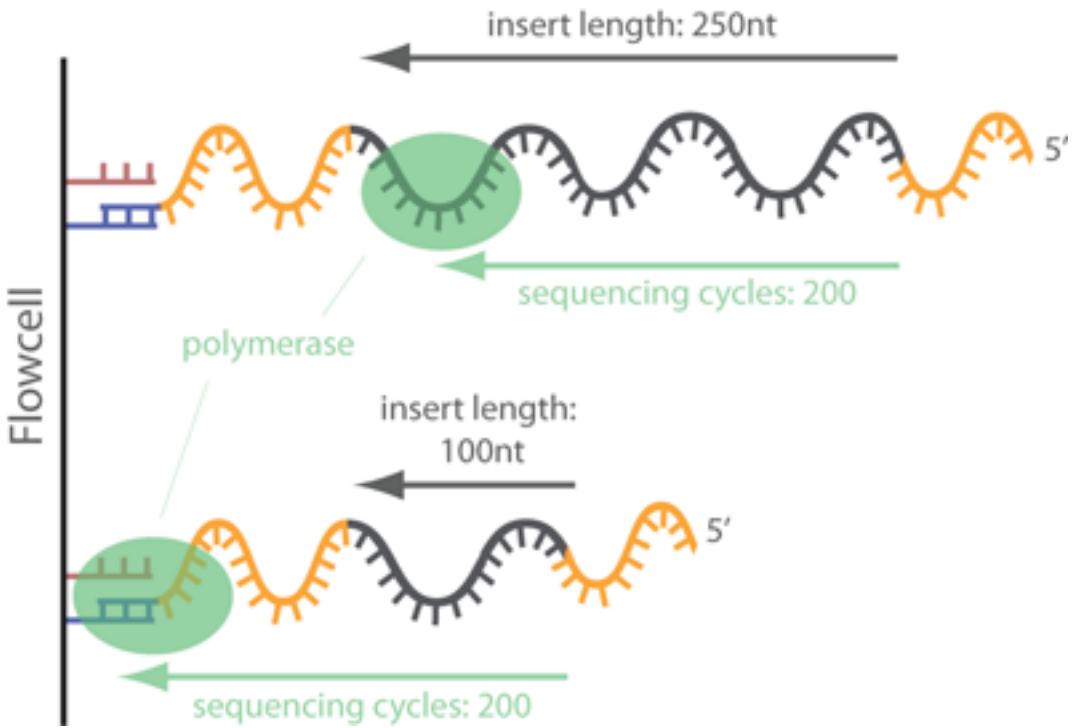


- Remove bases from each end until a minimum quality is reached
  - Or remove a specific number of bases
- Lose some reads but lead to better results



Mohsen, A. et al. BMC Bioinformatics 20:581 (2019)

Adapter trimming



List of known sequencing adapters

main ▾ Trimmomatic / adapters /

TonyBolger Parallel Compression

..

NexteraPE-PE.fa

TruSeq2-PE.fa

TruSeq2-SE.fa

TruSeq3-PE-2.fa

TruSeq3-PE.fa

TruSeq3-SE.fa

https://www.ecseq.com/support/ngs/trimming-adapter-sequences-is-it-necessary

Example of read trimming command

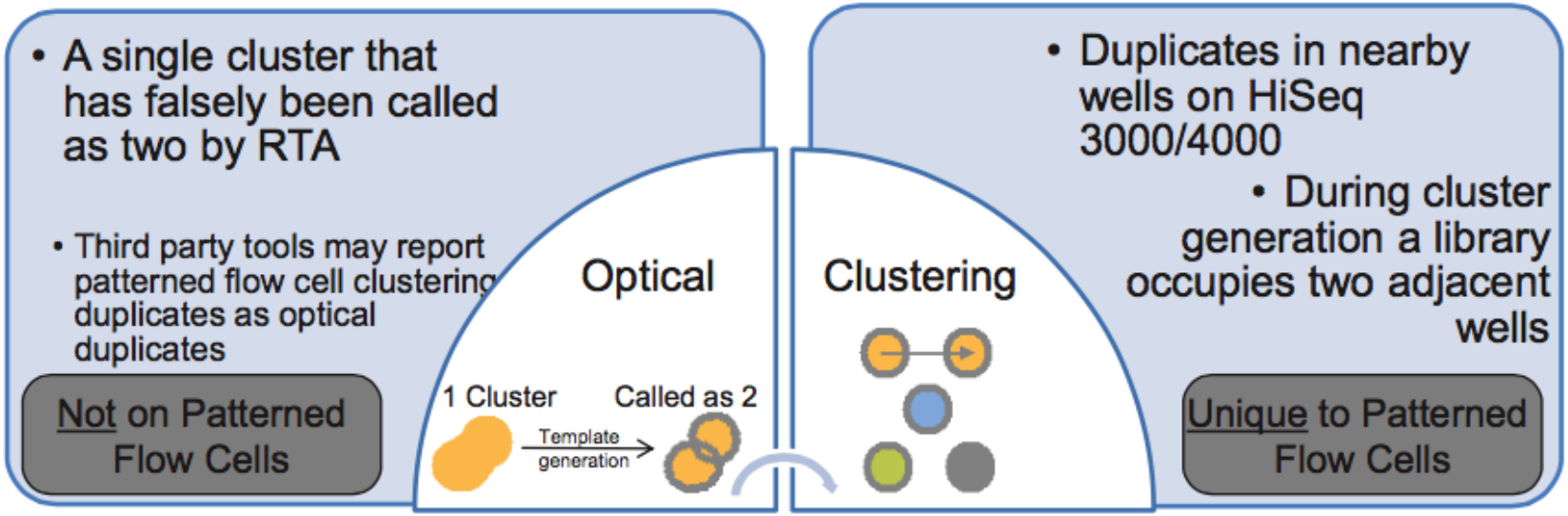
```
trimmomatic PE -threads 4 SRR_1056_1.fastq SRR_1056_2.fastq  \
            SRR_1056_1.trimmed.fastq SRR_1056_1un.trimmed.fastq \
            SRR_1056_2.trimmed.fastq SRR_1056_2un.trimmed.fastq \
            ILLUMINACLIP:SRR_adapters.fa SLIDINGWINDOW:4:20
```

- Use **4 CPU threads**
- Output to trimmed reads to **.trimmed files**
- Output discarded reads to **un.trimmed files**
  - Length became too short after trimming
- Remove adapter sequences listed in **SRR_adapters.fa**
- Check quality score in a sliding window
  - **Average Phred score <= 20** among **4 consecutive nucleotides**

Deduplication

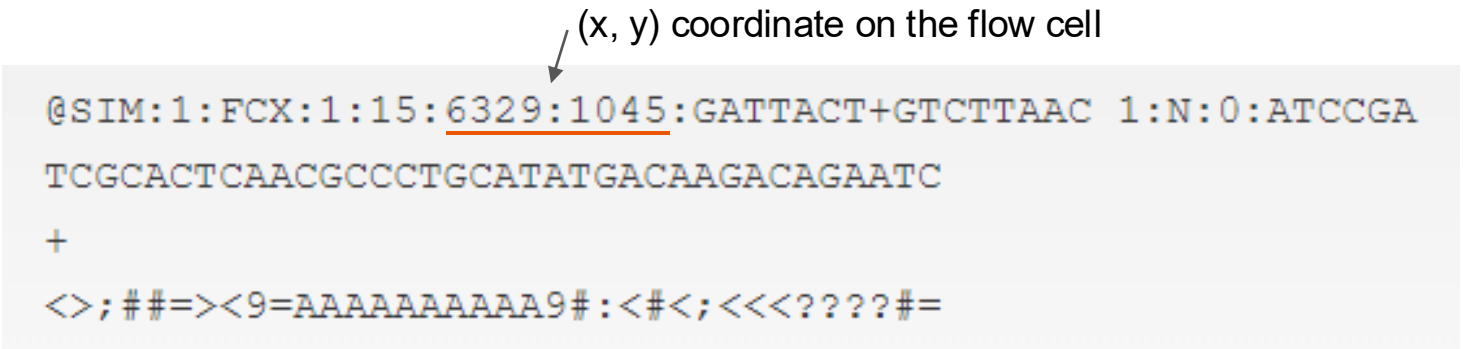
Duplicated reads from technical error

Illumina, 2016



- Same DNA molecule amplified into adjacent clusters in flow cells
- A large cluster was erroneously read as two clusters

Duplicated reads likely came from the same DNA



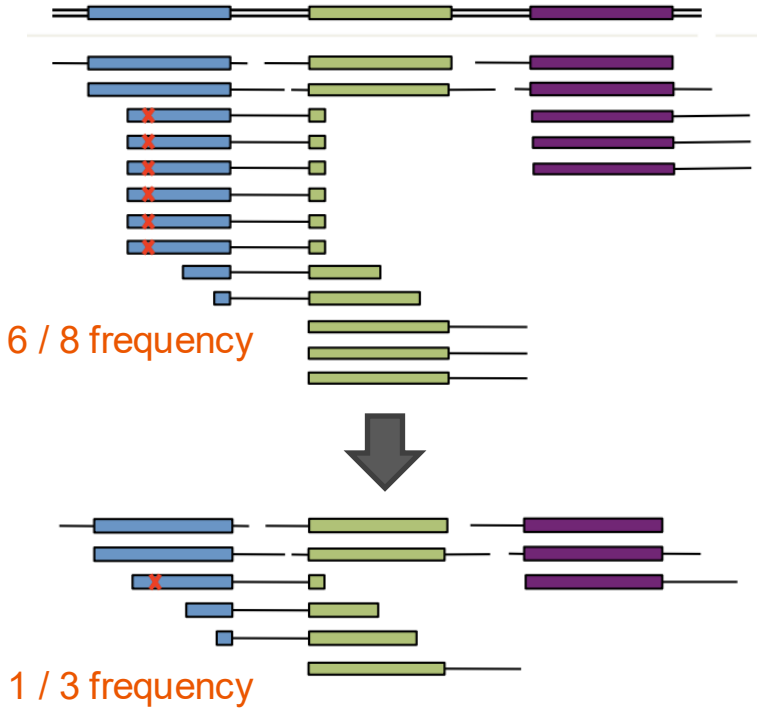
- Reads with identical sequences or mapped positions on the genomes
  - Unlikely to generate multiple identical DNA fragments by chance
  - Retain only one read for each group
- **Illumina**: Nearby cluster coordinates in the header

Problems caused by duplicated reads

- Lead to incorrect frequency estimates
  - Gene expression level
  - Variant allele frequency
- Many tools can de-duplicate reads
  - Perform after alignment
  - No extra parameters required

```
java -jar picard.jar MarkDuplicates \
      I=input.bam \
      O=marked_duplicates.bam \
      M=marked_dup_metrics.txt
```

```
samtools markdup positionsort.bam markdup.bam
```



Images from qcb.ucla.edu

GATK workflow

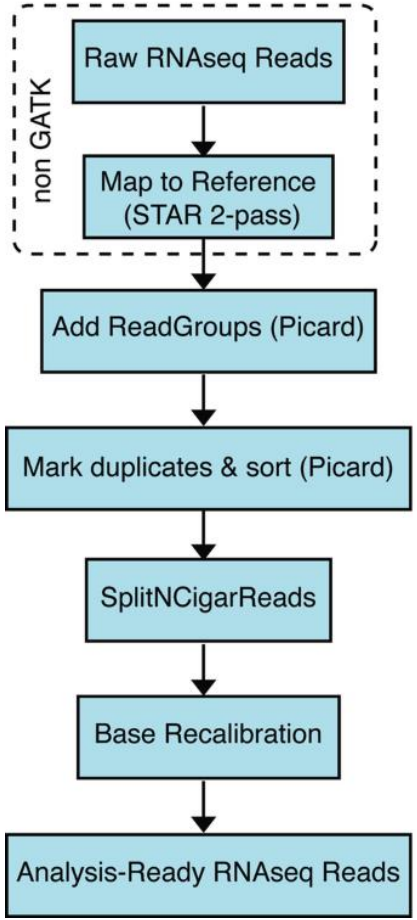
The Genome Analysis Toolkit (GATK)

- Developed by Broad Institute since 2010
- Core software uses Java language, with plugins from R and Python
- Industry standard for variant calling workflow

GATK data processing steps

- Performed after sequence alignment

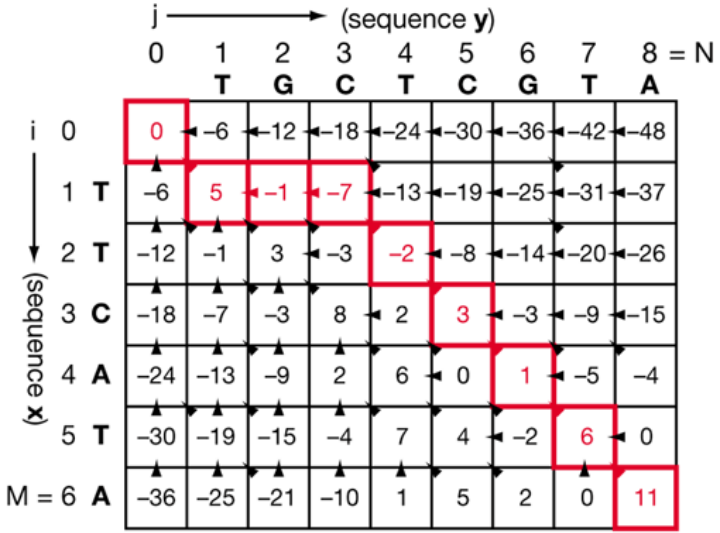
- Key steps:
  - Mark duplicates
  - Split reads based on **N** (unknown base)
    - AACTA**N**CTGAGA → AACTA and CTGAGA
  - Recalibrate quality scores
    - Identify systematic error and apply correction
    - **Example**: 1% more error after reading AAA
    - **Example**: 2% more error after position 120
    - Trained using common variants as truth



Sequence alignment:
Read-to-genome

Dynamic programming alone is not enough

Dynamic programming matrix:



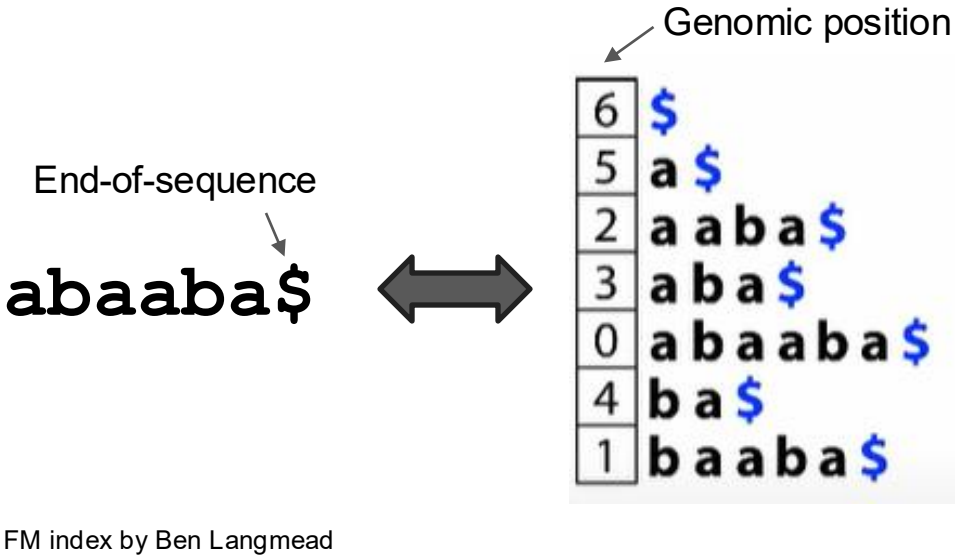
Optimum alignment scores 11:

T	–	–	T	C	A	T	A
T	G	C	T	C	G	T	A
+5	–6	–6	+5	+5	–2	+5	+5

- Aligning a 150 bp read to a human genome would create a 150 by 3 x10^9 table!
- 10 million reads = 10 million searches
- We need a faster strategy

Eddy, S.R. Nature Biotechnology 22:909-10 (2004)

Indexing



FM index by Ben Langmead

- The genome is static → We can preprocess beforehand and reuse
- Indexing create a lookup table like alphabetical order in dictionary
  - Generate all possible short DNA fragments and sort

Suffix array

Reference Sequence

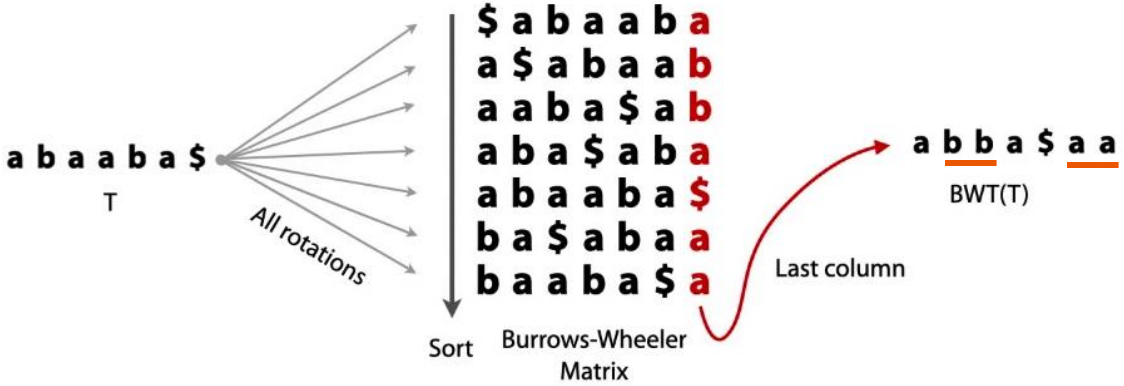
ATTGCAGTCCG

AGTCCG	6
ATTGCAGTCCG	1
CAGTCCG	5
CCG	9
CG	10
G	11
GCAGTCCG	4
GTCCG	7
TCCG	8
TGCAGTCCG	3
TTGCAGTCCG	2

- Suffix = ending part of a string
- Organize suffixes in an easily searchable data structure
- Also record the start positions

Christley, S. et al. BMC Bioinformatics 9:15 (2008)

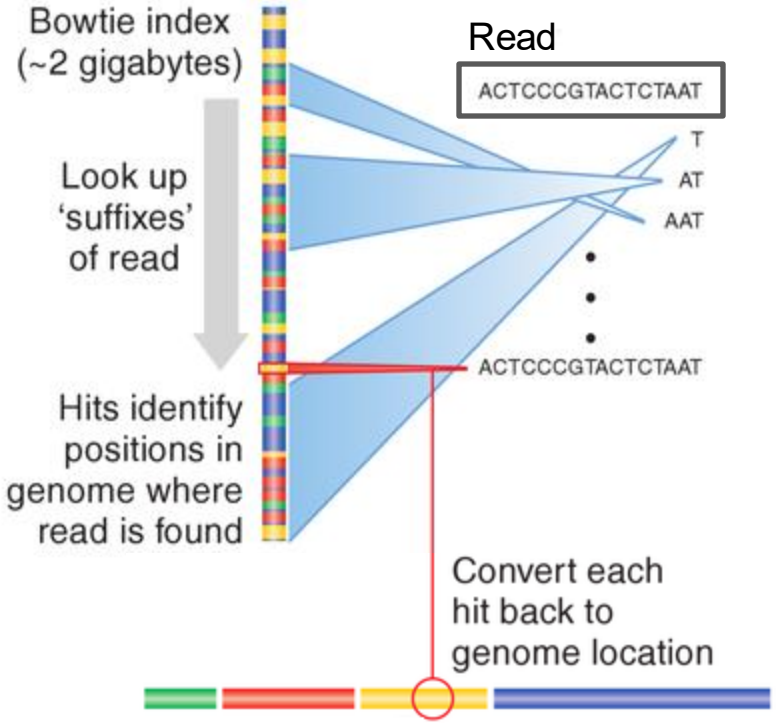
Burrows-Wheeler transform



Burrows, M. and Wheeler, D.J. A block sorting lossless data compression algorithm. 1994

- BWT tends to group the same character consecutively
  - Make the data easier to describe/compress: **ab2a$a2**
- BWT is reversible: can recover the original position

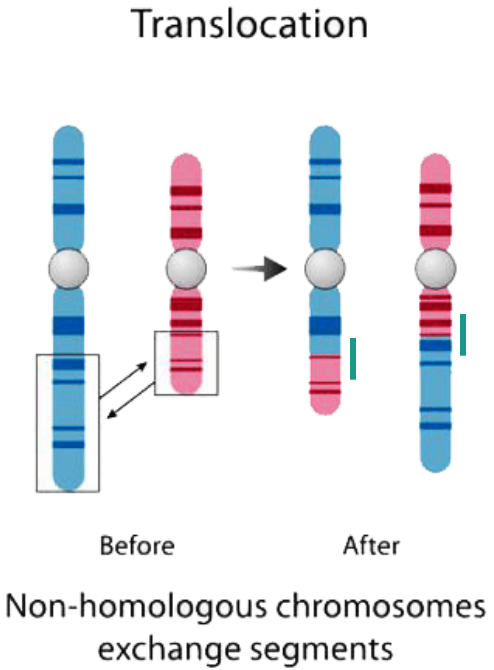
Genome-scale alignment

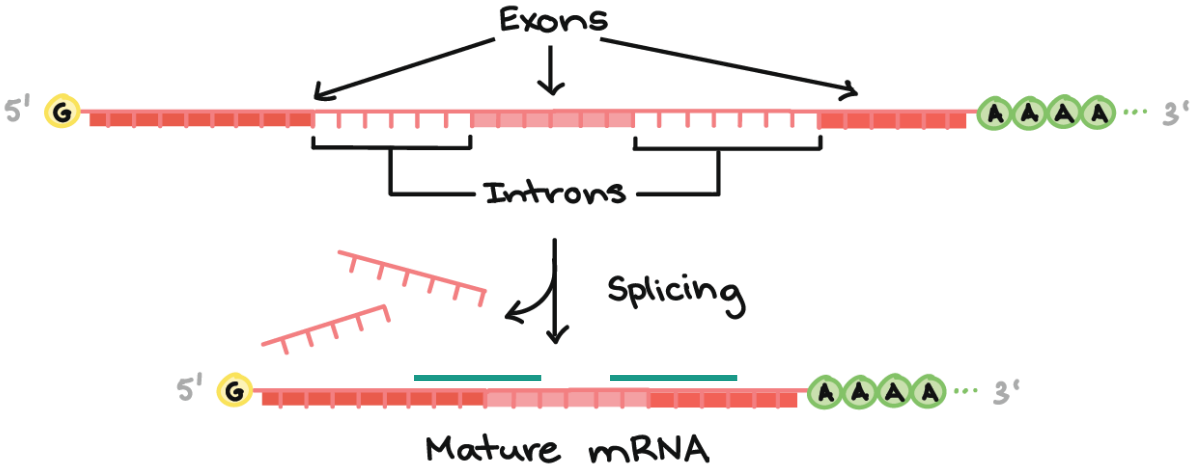


- Use BWT to index the genome
- 20x smaller memory than simple indexing for human genome
- 30x faster search speed

Trapnell, C. and Salzberg, S.L. Nature Biotechnology 27:455-7 (2009)

The need for gapped alignment

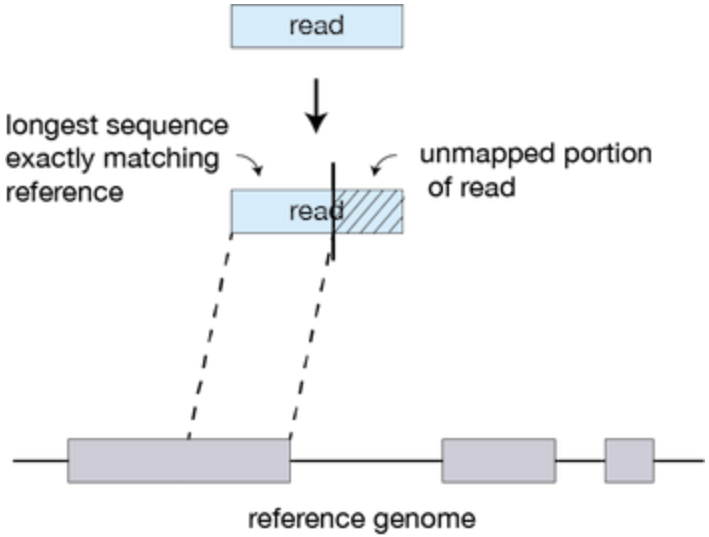




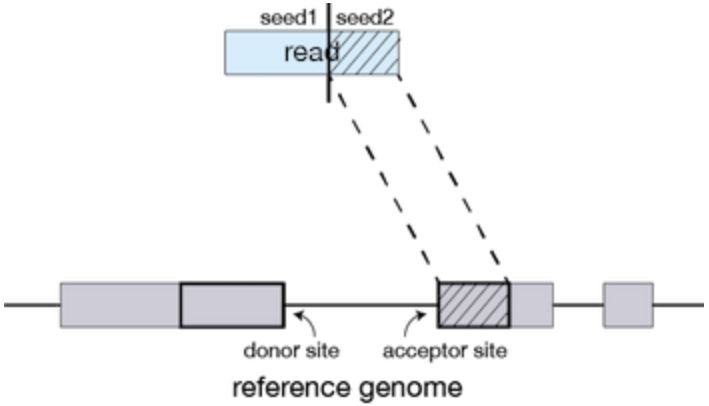
- A read can span a long genomic region due to translocation (DNA) or splicing (RNA)

https://www.paragongenomics.com/targeted-sequencing/rna-fusion-detection/

Gapped alignment



Essential for aligning RNA-seq data (spliced) and identifying translocation/gene fusion



https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

- Split read into two segments
- Align each segment separately → Look for nearby hits

Typical genome alignment commands

- Bowtie2
  - Index a genome database (FASTA file)

    ```
    bowtie-build GRCh38.fasta GRCh38_db
    ```
  - Perform alignment

    ```
    bowtie -x GRCh38_db -1 sample1_R1.fastq -2 sample1_R2.fastq
            -sam --threads 8 sample1.sam
    ```

- BWA
  - Index a genome

    ```
    bwa index GRCh38.fasta
    ```
  - Perform alignment

    ```
    bwa mem sample1_R1.fastq sample1_R2.fastq > sample1.sam
    ```

Sequence alignment results

Sequence Alignment Map (SAM)

Sort Order = by genomic coordinate
SN = reference sequence's name (FASTA header)
LN = reference sequence's length

```
@HD VN:1.6 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
r004     0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC
r003  2064 ref 29 17 6H5M       *  0   0 TAGGC
r001   147 ref 37 30 9M         =  7 -39 CAGCGGCAT
```

- r001 = read name (from sequencing FASTQ)
- ref = reference sequence name (from genomic FASTA)
- 7 = first position on the reference sequence
- 30 = Mapping quality score = $-10 \times \text{Log}_{10}(\text{error})$
- 8M2I4M1D3M = **CIGAR string** = matches, insertion, deletion information

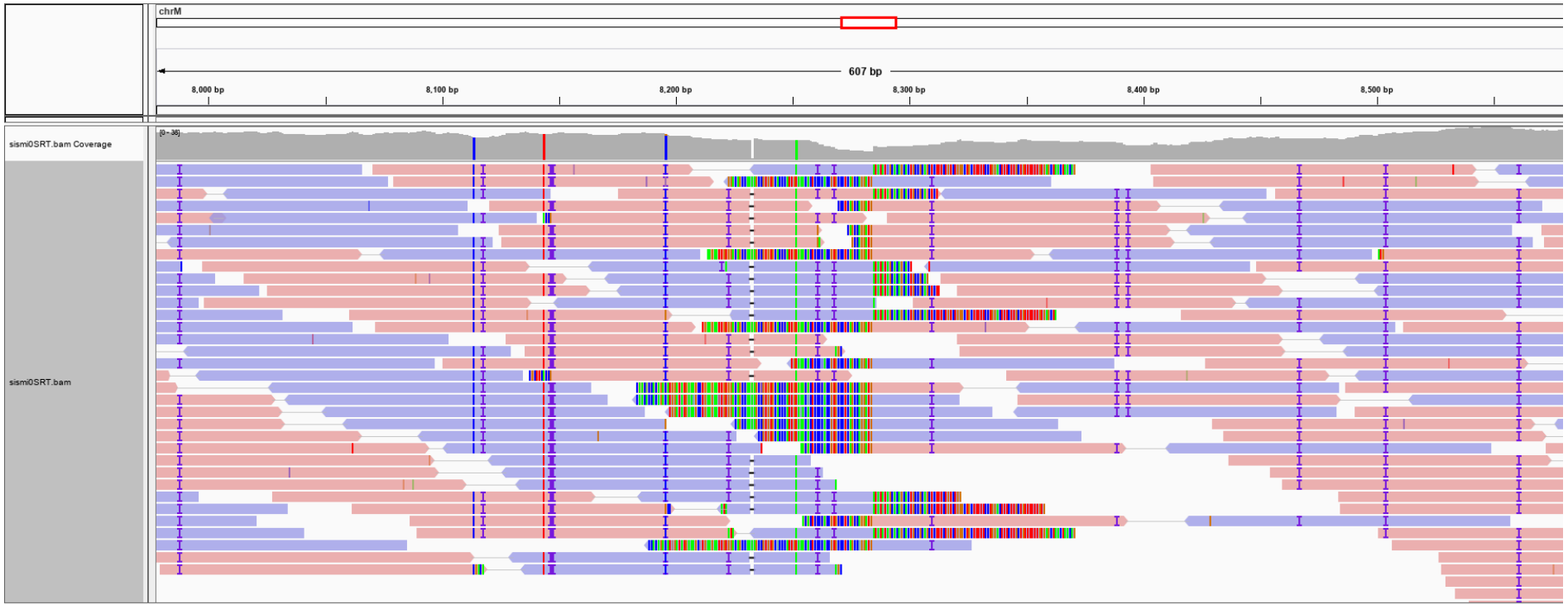
SAM file manipulation

- Sequence alignment results can be sorted, indexed, filtered, and zipped
- BAM is a zipped version of SAM (~40% of the size)

- Sorting and indexing makes alignment results in BAM file easier to be located and analyzed

- Performed with **samtools**

Integrated Genomics Viewer (IGV)



Pileup format

Sequence	Position	Reference Base	Read Count	Read Results	Quality
seq1	272	T	24	,.$....,,.,.,...,,,,.,..^+.	<<<+;<<<<<<<<<<=<;<;7<&
seq1	273	T	23	,.....,,.,.,...,,,.,..A	<<<;<<<<<<<<<3<=<<<;<<+
seq1	274	T	23	,.$....,,.,.,...,,,.,...	7<7;<;<<<<<<<<<=<;<;<<6
seq1	275	A	23	,$....,,.,.,...,,,.,...^l.	<+;9*<<<<<<<<<=<<:;<<<<
seq1	276	G	22	...T,,.,.,...,,,.,....	33;+<<7=7<<7<&<<1;<<6<
seq1	277	T	22	....,,.,.,.C.,,,.,..G.	+7<;<<<<<<<&<=<<:;<<&<
seq1	278	G	23	....,,.,.,...,,,.,....^k.	%38*<<;<7<<7<=<<<;<<<<<
seq1	279	C	23	A..T,,.,.,...,,,.,.....	75&<<<<<<<<<=<<<9<<:<<<

Image from wikipedia

- Focus on each base pair position
  - Summarize match, mismatch, and gap from multiple reads

Any question?

- See you next time