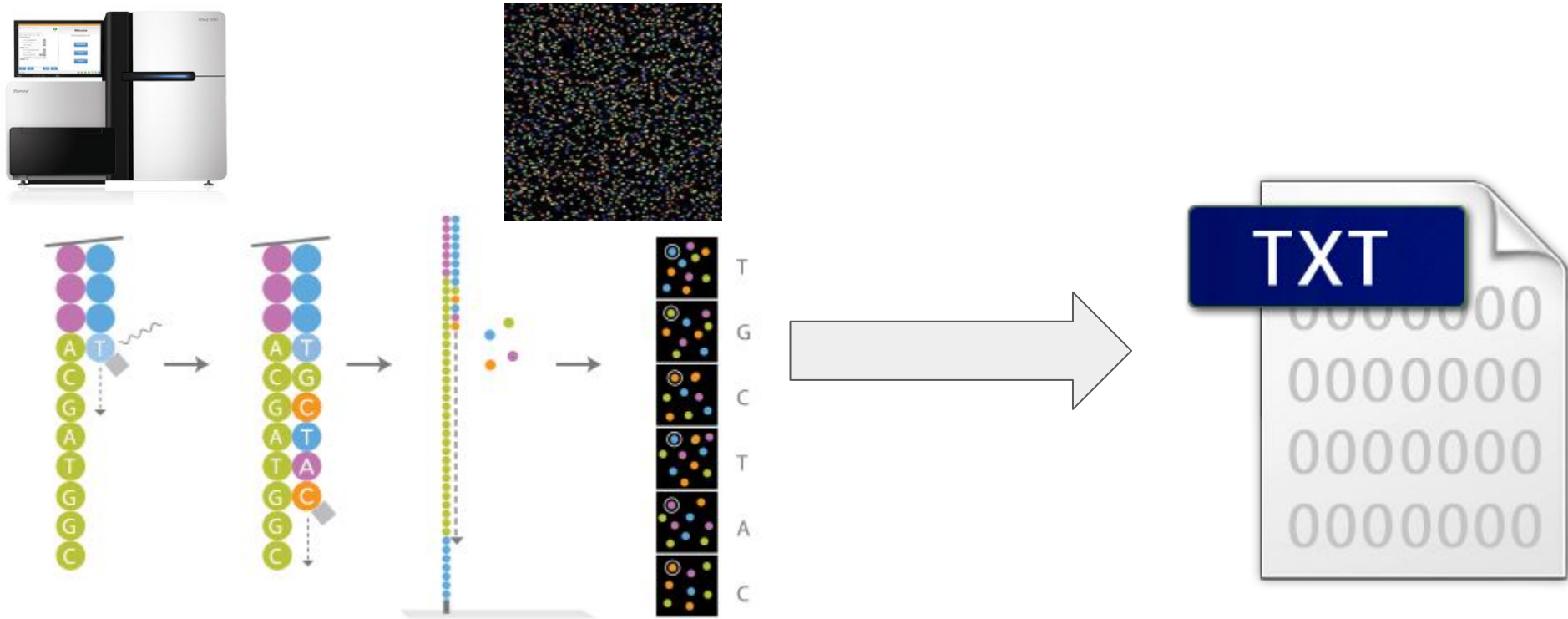


RNA-seq

From sequencing to counts

Sequence Data: Raw Reads

Demultiplexing: from image to fastq

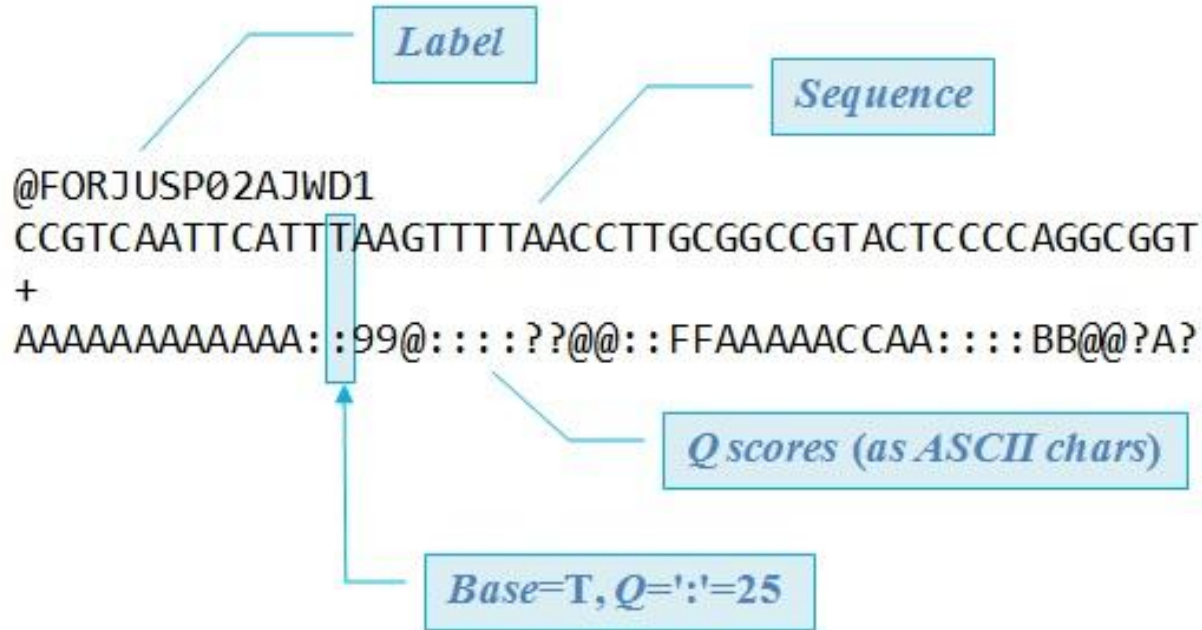


Fastq

Fastq is a text based format containing the reads (sequences) that came from the machine.

Often this file is several MB to GB in size, therefore it is gzipped (.fastq.gz).

Fastq: File format



Fastq: The label

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139:136	Machine and run ID
FC706VJ	Flowcell ID
2	Lane
2104	Tile number in the lane
15343:197393	Coordinate of the cluster within the tile
1	Member of the pair (1 or 2)
Y:18	Past filter information of the Illumina software
ATCACG	The Illumina index/barcode

Fastq: Scores



S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

Fastqc: Quality control of fastq files

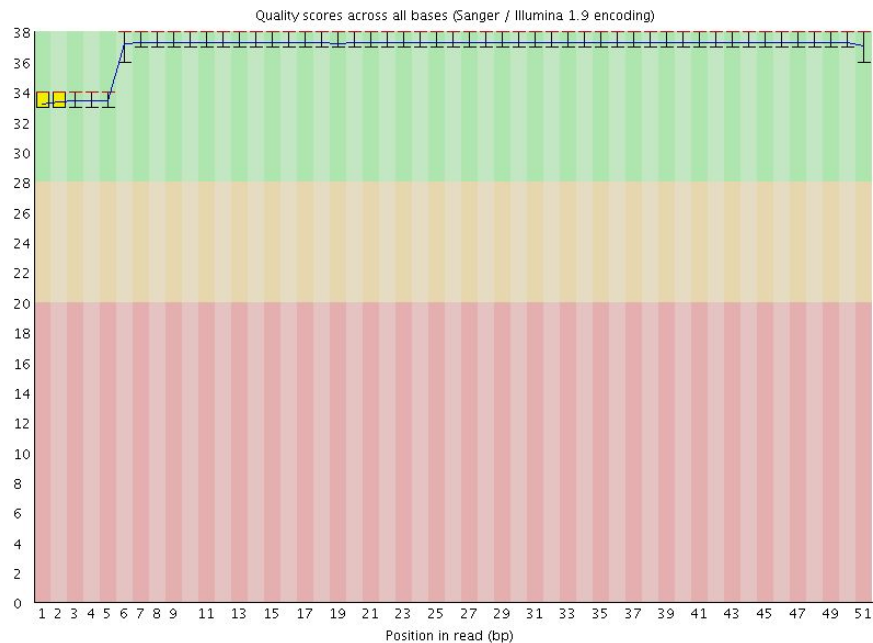
Some important stats for RNA-seq:

- Number of reads
- Base quality
- Length of reads

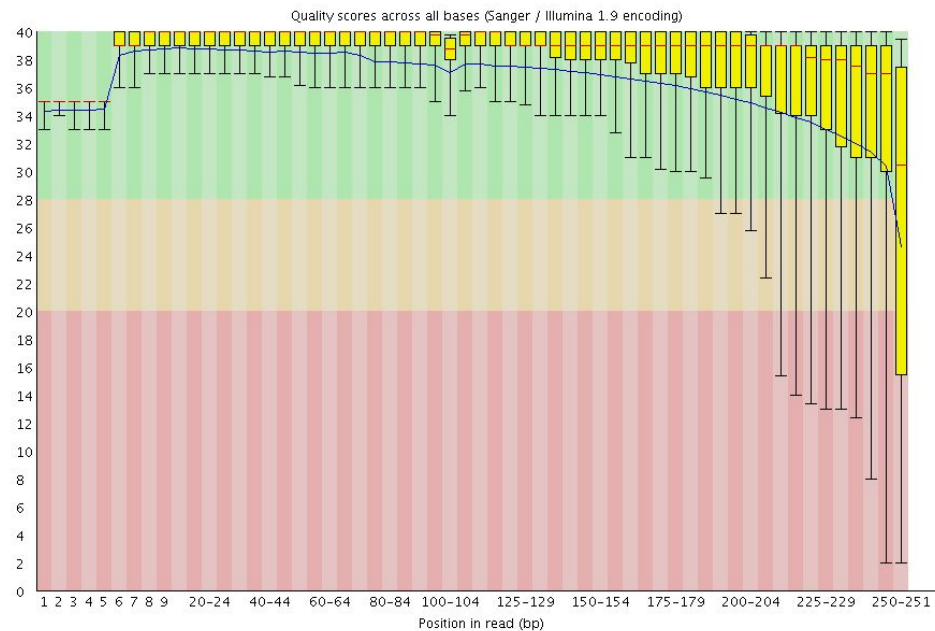
Open Fastqc

Fastqc: Base Quality

✓ Per base sequence quality

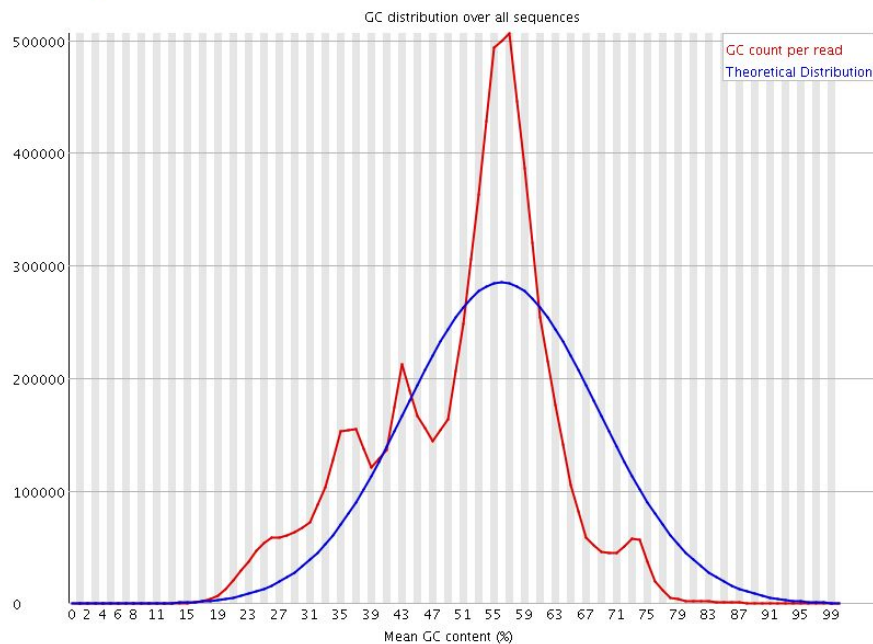


✓ Per base sequence quality

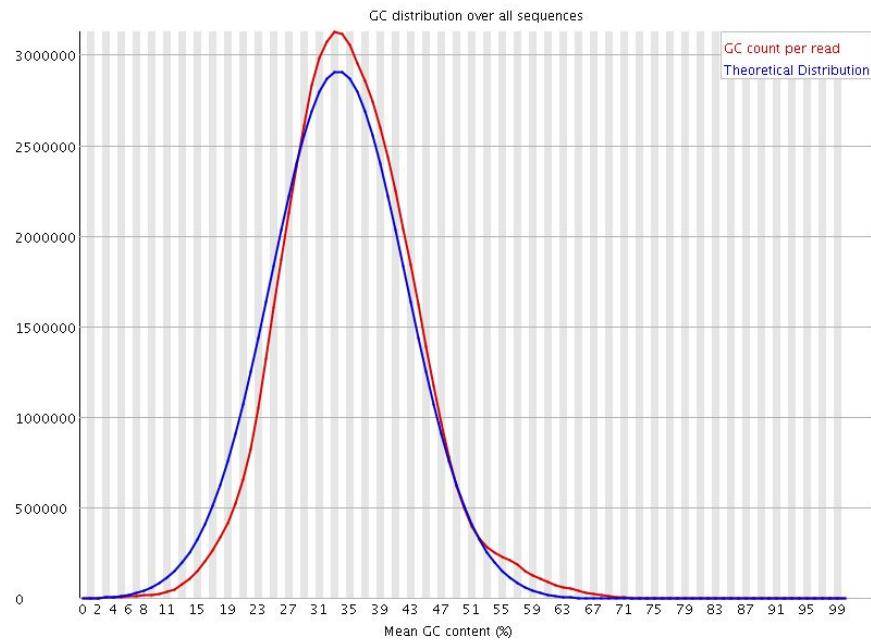


Fastqc: GC content

❌ Per sequence GC content

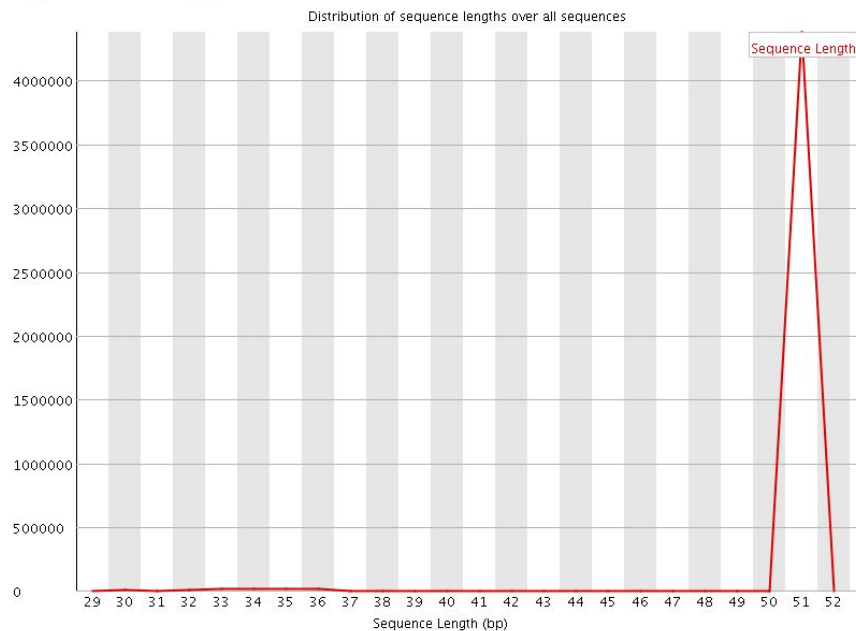


✅ Per sequence GC content

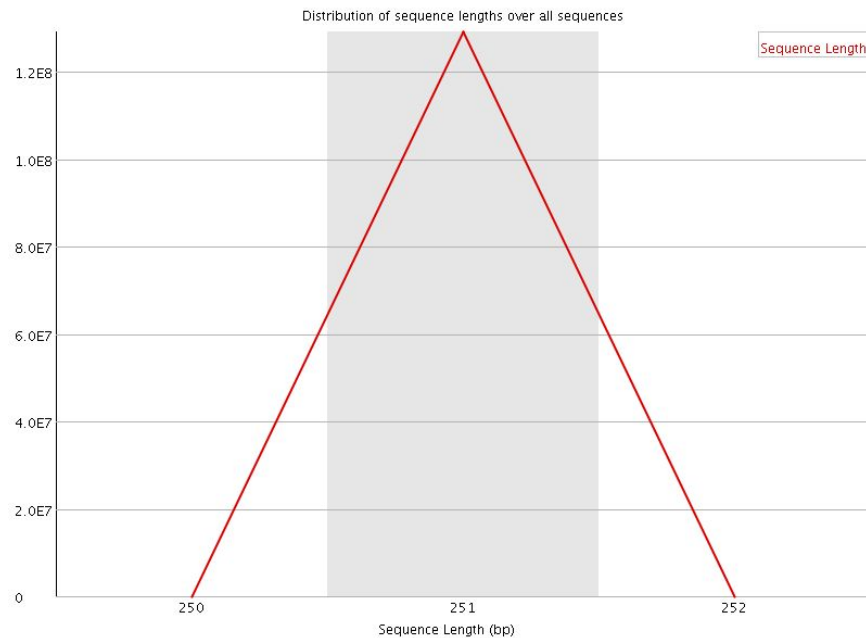


Fastqc: Read length

❗ Sequence Length Distribution

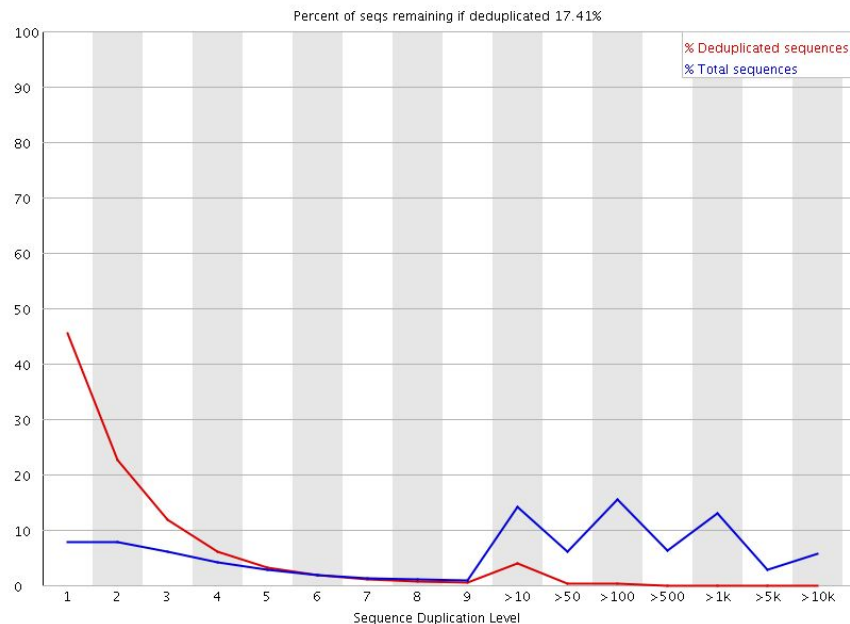


✅ Sequence Length Distribution

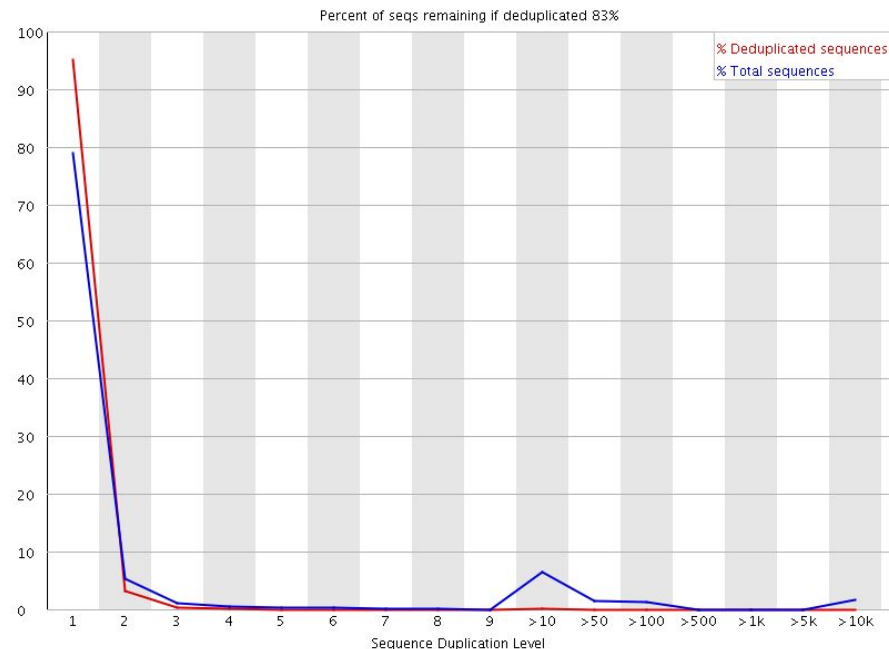


Fastqc: duplication levels

✗ Sequence Duplication Levels

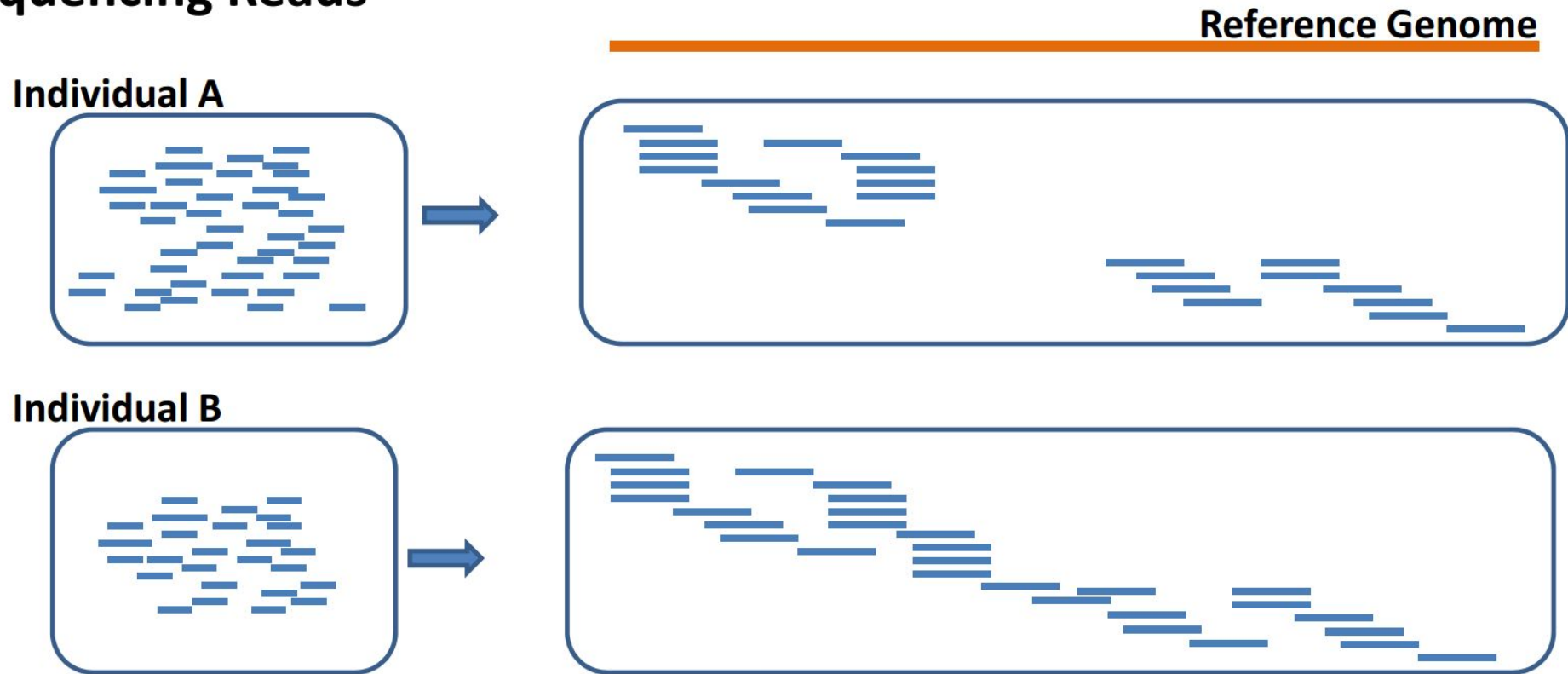


✓ Sequence Duplication Levels



Finding the Location of the Reads in the Reference

Mapping Sequencing Reads



Mapping: Where to start?

Needed information:

1. The dataset in fastq format (1 or 2 files, depends on single reads or paired end data)
2. A reference
 - a. The reference is species dependent.
 - b. It can have any quality (from 1k+ contigs, to chromosome level)
 - c. It can be the genome or the transcriptome
 - d. The version is tracked using a certain build. Make sure the build you use has an annotation!

Mapping

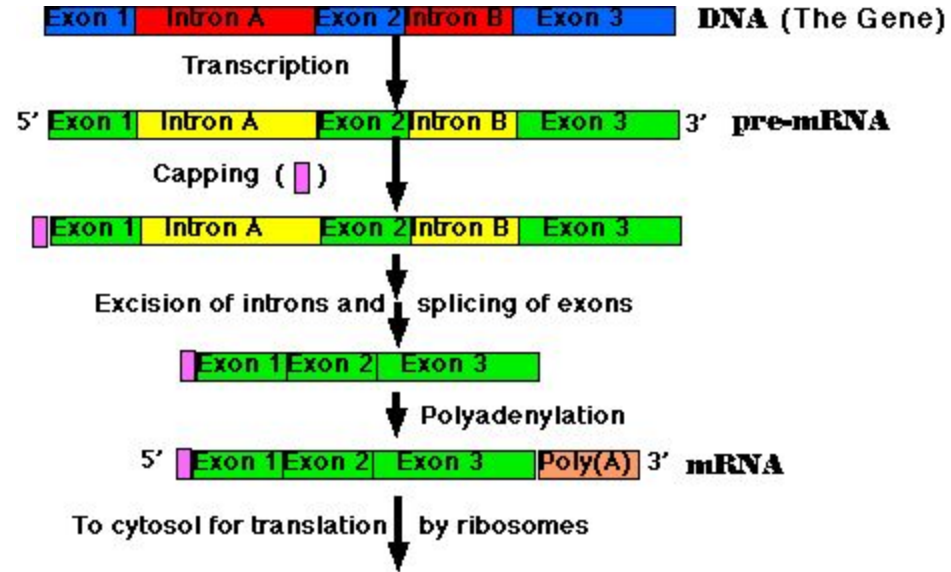
Possible References:

- **Transcriptome:**

Created from mRNA, does only include exons.
All positions in the reference are possible to be in the expressed dataset

- **Genome:**

Created from the DNA, does include introns, and other 'junk'
Only a small part of the reference will be in the expressed dataset



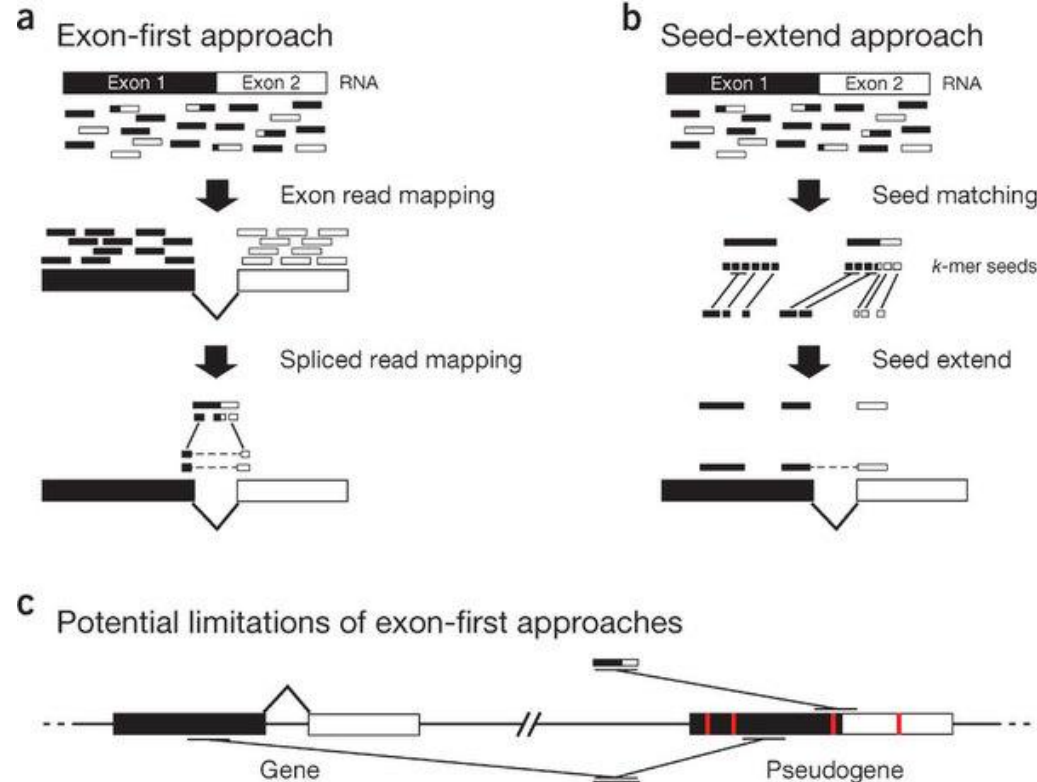
Split Read Mapping

Exon-first approach

1. map read to transcriptome
2. if mapped, map to genome
3. if read was mapped on a known splicing location, splice.

Seed-extend approach

1. Find seed (k-mer) in genome
2. Extend seed to maximum length
3. Repeat for multiple seed in read
4. Combine the found positions



Mapping: Output format: SAM and BAM format

- SAM
Sequence Alignment/Map format
Header starts with @
Mapped Reads are tab-delimited lines
1-based system (includes SAM, VCF, GFF, GTF)
- BAM
Binary Alignment/Map format
Is the binary form of the SAM format, so reduces storage
Contains exact the same information as the SAM format
0-based system (includes BAM, BED)

Mapping: Output format: SAM and BAM format

>gene1:3-53 l:49 nM:i:0	0	chr0	22	255	50M	*	0	0	TTTGTTTCATCGTATTTTCTACAGTCGGGTAGCAAAGTATAACTGGATT	AAAAA5AFFFFFFFFF@FFFFFFFFF7FFFFFFFFFFFFFFFFF7FFFFFFF	NH:i:1	HI:i:1	AS:
>gene1:4-54 l:47 nM:i:1	0	chr0	23	255	50M	*	0	0	TTGTTTCATCGTATTTTCTACAGTCGAGTAGCAAAGTATAACTGGATT	AAA@AAAAFFFFFFFFFFFFFFFFF@FFFFFFFFFFFFFFFFF7FFFFFFF	NH:i:1	HI:i:1	AS:
>gene1:6-56 l:49 nM:i:0	0	chr0	25	255	50M	*	0	0	GTTTCATCGTATTTTCTACAGTCGGGTAGCAAAGTATAACTGGATTAA	AAAAAAA>FFFFFF>FFFFFFFFFFFFFF<FFFF+FFFFFFFFF@:FFF	NH:i:1	HI:i:1	AS:
>gene1:8-58 l:49 nM:i:0	0	chr0	27	255	50M	*	0	0	TCATCGGTATTTTCTACAGTCGGGTAGCAAAGTATAACTGGATTAA	AAAAAAAF8FFFFF8+F<FFFFFFFFFFFFFFFFF5FFFFFFFFFFFFFFF	NH:i:1	HI:i:1	AS:
>gene1:14-64 l:49 nM:i:0	0	chr0	33	255	50M	*	0	0	GTATTTTCTACAGTCGGGTAGCAAAGTATAACTGGATTAAATTAGAAA	AA6AAAA8FF	NH:i:1	HI:i:1	AS:
>gene1:17-67 l:48 nM:i:0	0	chr0	37	255	1549M	*	0	0	ATTTTCTACAGTCGGGTAGCAAAGTATAACTGGATTAAATTAGAAAAA	AAAAAAA86FFFFFFFFFFFFFFFFFFFFFFFFF@FFFFFFFFFFFFFFFFF	NH:i:1	HI:i:1	AS:
>gene1:19-69 l:49 nM:i:0	0	chr0	38	255	50M	*	0	0	TTTCTACAGTCGGGTAGCAAAGTATAACTGGATTAAATTAGAAAAA	AAAAAAAF8FFFFFFFFF>FFFFF5F<FF@FFFFFFFFF@FFFFFFFF>	NH:i:1	HI:i:1	AS:
>gene1:19-69 l:47 nM:i:1	0	chr0	38	255	50M	*	0	0	TTTCTACAGTCGGATAGCAAAGTATAACTGGATTAAATTAGAAAAA	AAAAAA7AFF5FFFFFF:FFFFFFFFF@FFFFFFFFFFFFFFFFFFFFF	NH:i:1	HI:i:1	AS:
>gene1:26-76 l:49 nM:i:0	0	chr0	45	255	50M	*	0	0	AGTCGGGTAGCAAAGTATAACTGGATTAAATTAGAAAAAATACAGGTG	A+AAAAAF8FFFFFFFFF+FFFFFFFFFFFFFFFFF8FFFFFFFFFFFFF	NH:i:1	HI:i:1	AS:
>gene1:26-76 l:49 nM:i:0	0	chr0	45	255	50M	*	0	0	AGTCGGGTAGCAAAGTATAACTGGATTAAATTAGAAAAAATACAGGTG	AA>AAAAAF7FFFF@FFFFFFFFFFFFFFFFF8FFFFFFFFF>FF>FF	NH:i:1	HI:i:1	AS:
>gene1:28-78 l:49 nM:i:0	0	chr0	47	255	50M	*	0	0	TCGGGTAGCAAAGTATAACTGGATTAAATTAGAAAAAATACAGGTGTG	8AA@AAAAFFFFFFFFF:FFF7FFFFF5F<FFF<F:5FFFFFF<@FFFF	NH:i:1	HI:i:1	AS:
>gene1:33-83 l:49 nM:i:0	0	chr0	52	255	50M	*	0	0	TAGCAAAGTATAACTGGATTAAATTAGAAAAAATACAGGTGTTGTTTC	A6AAAAAF8FFF7FFFFFFFFFFFF:FFFFFFFFFFFFF+FFFFFFFFF	NH:i:1	HI:i:1	AS:
>gene1:33-83 l:49 nM:i:0	0	chr0	52	255	50M	*	0	0	TAGCAAAGTATAACTGGATTAAATTAGAAAAAATACAGGTGTTGTTTC	:AAAAAF8FFFF5FFF<>FFFFFFFFFFFFFFFFFFFFFFFFFFFFF>	NH:i:1	HI:i:1	AS:
>gene1:39-89 l:47 nM:i:1	0	chr0	58	255	50M	*	0	0	AGTATAACTGGATTAAATTAGAAAAAATACAGGTGTTGATTCTAATTA	AA:AAAAAF8FF@FFFFFFFF8FFFFF+FFFFF>FFFFFFFFFFFFF	NH:i:1	HI:i:1	AS:
>gene1:45-95 l:49 nM:i:0	0	chr0	64	255	50M	*	0	0	ACTGGATTAAATTAGAAAAAATACAGGTGTTGGTTCTAATTAGTCGGC	AAA<A7<FFFFFFFFF8FFF@FFFFFFFFFFFFFFFFFFFFF:FFFF	NH:i:1	HI:i:1	AS:
>gene1:45-95 l:49 nM:i:0	0	chr0	64	255	50M	*	0	0	ACTGGATTAAATTAGAAAAAATACAGGTGTTGGTTCTAATTAGTCGGC	A>AAAAAF8FFFFFFFFFFFFFFFFFFFFFFFFFFFFF6FF7FFFFFFF	NH:i:1	HI:i:1	AS:
>gene1:57-107 l:49 nM:i:0	0	chr0	76	255	50M	*	0	0	TTAGAAAAAATACAGGTGTTGGTTCTAATTAGTCGGCGTACGCCGTTA	AAAAAAAF8FFFFFFFFFFFFFFFFFFFF7F5F6FFFF>FFFFFFFF6FF	NH:i:1	HI:i:1	AS:
>gene1:60-110 l:49 nM:i:0	0	chr0	79	255	50M	*	0	0	GAAAAAATACAGGTGTTGGTTCTAATTAGTCGGCGTACGCCGTTACAT	AAAAAAA:FFFF+FF6FFFF+FF+FFFFFFFFFFFFFFFFFFFFFFFFF	NH:i:1	HI:i:1	AS:
>gene1:66-116 l:49 nM:i:0	0	chr0	85	255	50M	*	0	0	ATACAGGTGTTGGTTCTAATTAGTCGGCGTACGCCGTTACATTATTCG	AAAAAAAF8FFFF8FFFFF<FFFF8FFFF8FFFFFFFFFFFFF8F	NH:i:1	HI:i:1	AS:
>gene1:67-117 l:49 nM:i:0	0	chr0	86	255	50M	*	0	0	TACAGGTGTTGGTTCTAATTAGTCGGCGTACGCCGTTACATTATTCG	AAAAAAAF8FFF>FFFFFFFFF5:6FFFFFFFFFFFFFFFFFFFFF	NH:i:1	HI:i:1	AS:
>gene1:67-117 l:47 nM:i:1	0	chr0	86	255	48M2S	*	0	0	TACAGGTGTTGGTTCTAATTAGTCGGCGTACGCCGTTACATTATTCAT	AAAAA:AFF8FFFFF8FFFFF>FFFFFFFF:FF+FF@FF5F	NH:i:1	HI:i:1	AS:
>gene1:68-118 l:47 nM:i:1	0	chr0	87	255	50M	*	0	0	ACAGGTGATGTTTCTAATTAGTCGGCGTACGCCGTTACATTATTCGTG	AAAAAAAF8FFF>FFFFFFFF>F7FFFFFFFF6FFFFF:FF@FFFFF	NH:i:1	HI:i:1	AS:
>gene1:70-120 l:48 nM:i:0	0	chr0	89	255	49M1S	*	0	0	AGGTGTTGTTTCTAATTAGTCGGCGTACGCCGTTACATTATTCGTGTA	AAAA5AAAF:FFFFFFFFF8FFFF8FFFFFFFFFFFF5FFFFF	NH:i:1	HI:i:1	AS:

Mapping: Output format: SAM and BAM format

```
>gene1:3-53      0      chr0      22      255      50M      *      0      0      TTTGTTTCATGCGTATTTTCTACAGTCGGGTAGCAAAGTATAACTGGATT
AAAAAA5AFFFFFFFFF@FFFFFFFFF7FFFFFFFFFFFFF7FFFFFFF      NH:i:1      HI:i:1      AS:i:49      nM:i:0
```

>gene1:3-53	Name of the sequence
0	The flag (containing information about mapped/unmapped, forward/reverse, paired, info of the paired read)
chr0	The chromosome
22	Start position
255	Mapping quality
50M	CIGAR string (info about the alignment)
*	Chromosome of the paired read (no pair here)
0	Start position of the paired read
0	Template length (calculated from the start from the first read, to the end of the second read)
TTTTT....	Sequence
AAAAA	Quality

Mapping: view in IGV

- Open IGV
- Load reference file
Genomes > Load Genome from File
- Index the bam file, if no index available (.bai)
Tools > Run IGVtools > Select Index
- Open 2 bam files
File > Load from File
- Load gtf file
File > Load from File

Mapping: view in IGV

- Adjust view:
 - change read view
Right mouse button on Track > Collapsed/Expanded
 - sort per read strand
Right mouse button on Track > Group alignments by > read strand
 - show soft clipped
View > Preferences > Alignment > Show soft-clipped bases
 - Remove low quality reads
View > Preferences > Alignment > Mapping quality threshold

- How to search for a position/gene

The position track

- How to take screenshot

File > Save Image

Counting the Reads per Feature

Counting: Where to start?

Needed information:

1. The mapped data
=> BAM file
2. Information about the annotation.
Where are the genes?

=> GTF file

GTF files are linked with the used species, version and build. These files also has there own versions (often used are the Ensembl versions)

Counting: GTF file

```
1 transcribed_unprocessed_pseudogene   gene      11869 14409 . + . gene_id "ENSG00000223972"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene";
1 processed_transcript                  transcript 11869 14409 . + . gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; gene_name "DDX11L1"; gene_source "havana"; transcript_biotype "processed_transcript";
```

seqname	Name of the chromosome/contig
source	The database or program that generated this feature
feature	Feature type: e.g. gene, transcript, exon, ...
start	Start position of the feature (1-based)
end	End position of the feature
score	A floating point value (. when not available)
strand	Forward (+) or reverse (-)
frame	0, 1 or 2. Indicates the start of a codon
attribute	Semicolon separated tag-value pairs, like gene, gene_id, gene_name, ...

Counting

- **Simple**

Combine exons, analyse at gene level

Simple, powerful, sometimes inaccurate

HTSeq-count *STANDARD*

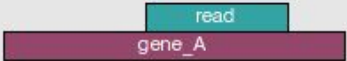

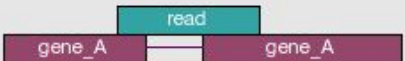


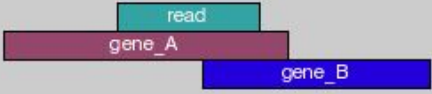
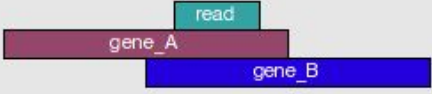
- **Complex**

Quantitate transcripts and merge to gene level

Potentially cleaner, more powerful signal

High degree of uncertainty

Cuffquant of Cufflinks 2.2.1

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Counting

Example of a file (first lines):

```
A4GN1 0
AA06 0
AAAS 0
AACS 2
AACSP1 0
AADAC 0
AADACL2 0
AADACL2-AS1 0
AADACL3 0
AADACL4 0
AADACP1 0
AADAT 0
AAED1 0
AAGAB 3
AAK1 0
AAMDC 0
AAMP 4
AANAT 0
AAR2 0
AARD 0
AARS 0
AARS2 0
AARSD1 0
AARSP1 0
AASDH 0
AASDHPPT 0
```

Example of a file (last lines):

```
snoZ6 0
snosnR66 0
uc_338 0
yR211F11.2 0
_no_feature 357
_ambiguous 5007
_too_low_aQual 47490
_not_aligned 10470
_alignment_not_unique 0
```

_no_feature	Reads map to location outside known genes
_ambiguous	Reads map to an overlap of multiple genes
_too_low_aQual	Reads whom the mapping quality read is to low
_not_aligned	Reads that are not mapped
_alignment_not_unique	Reads that mapped to multiple locations

These output count files can be opened with any text editor or spreadsheet.

Beware if opened with a spreadsheet, the gene names can be automatically changed:

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren and Assam El-Osta 

Genome Biology 2016 17:177 | DOI: 10.1186/s13059-016-1044-7 | © The Author(s). 2016

Published: 23 August 2016

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.