RNA Sequencing workflow

CANBIO

Arnaud Muller

November-2019

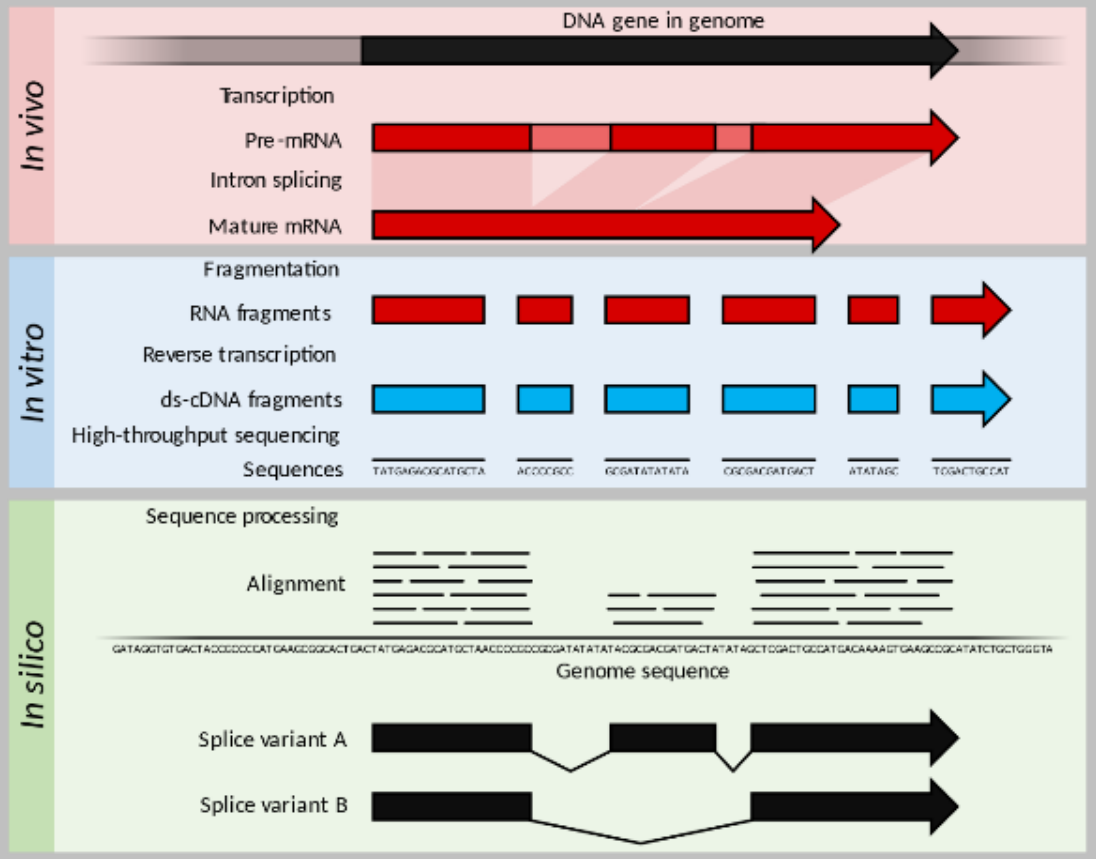
LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

LUGEN
NEXT GENERATION SEQUENCING

Biology

Biochemistry

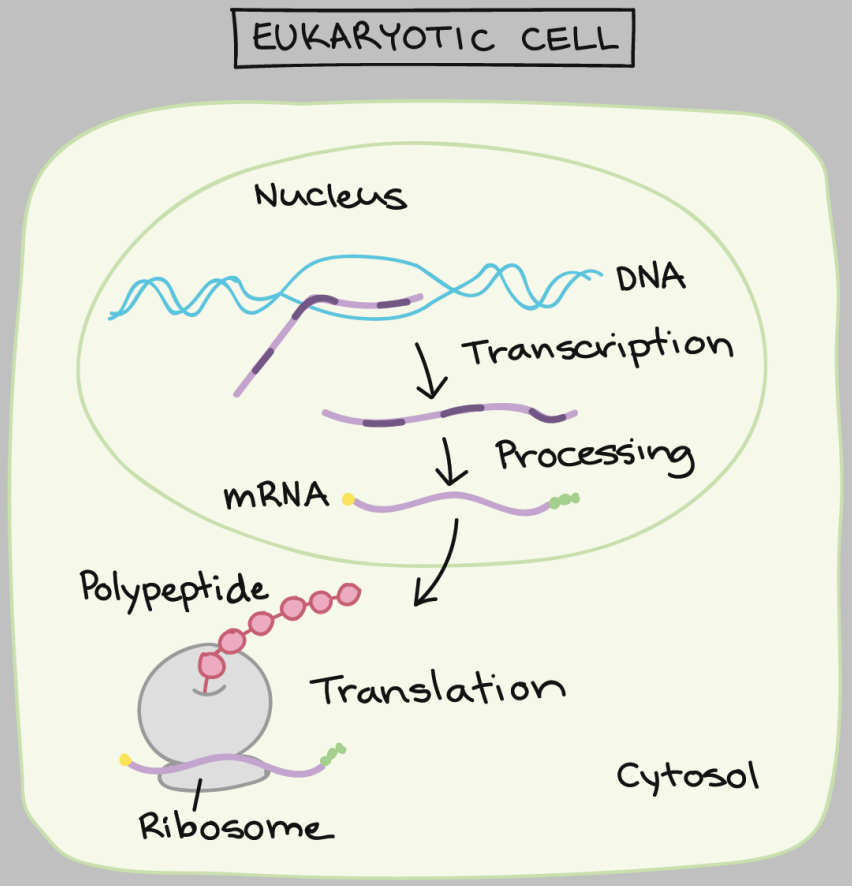
Bioinformatics



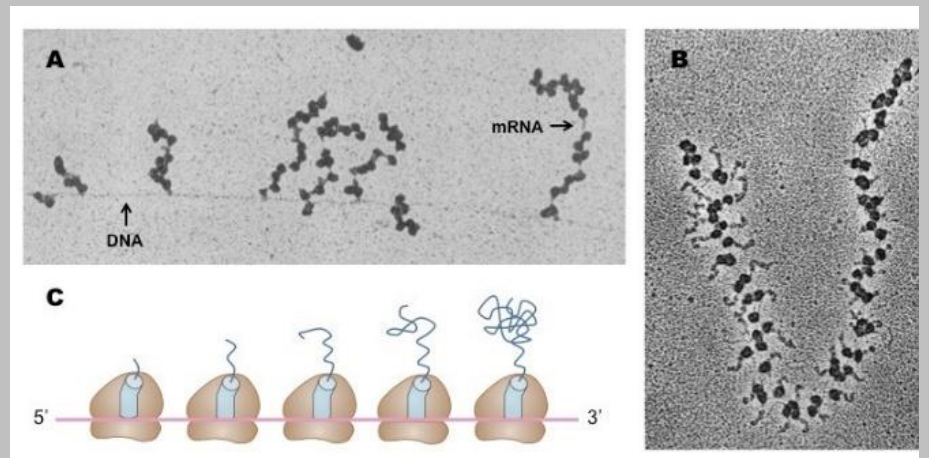
source: wikipedia

in vivo

mRNA processing

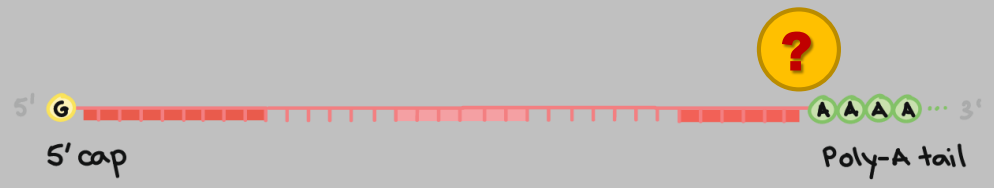


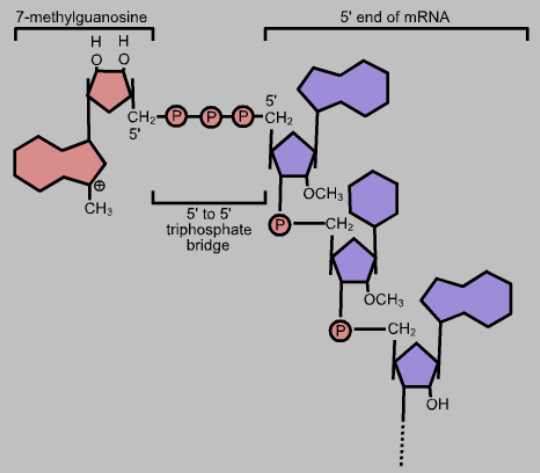
source: khanacademy



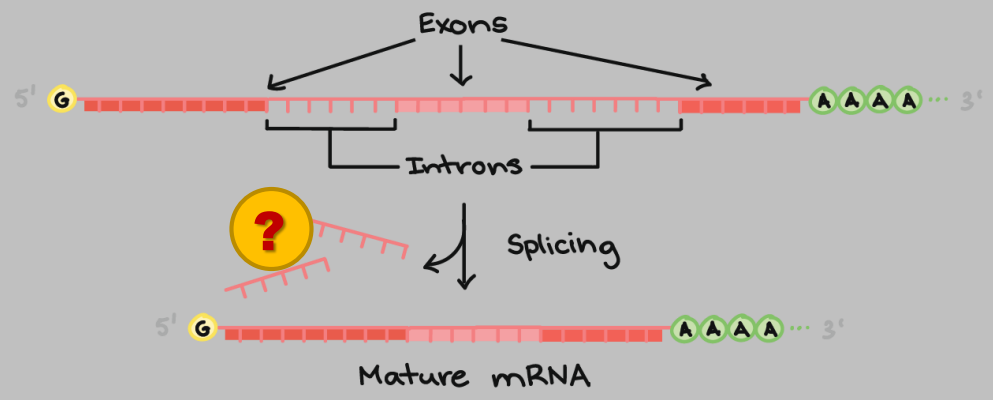
A: Polysome formation on RNA strands being transcribed from a DNA template (prokaryote)
B: Polysome formation on a eukaryotic mRNA transcript (arrow indicates 5'-end of transcript)
C: Diagrammatic representation of polysomes (polypeptide chain gets longer as ribosome moves 5' →

mRNA processing





- 5’ capping: addition of a modified *guanine* 5’-end of the pre-mRNA
- 3’ poly-A tail: a polyadenylation signal (AAUAAA) is recognized, followed by a cleavage and polyA tail synthesis (200 ntd)
- Splicing:



source: khanacademy

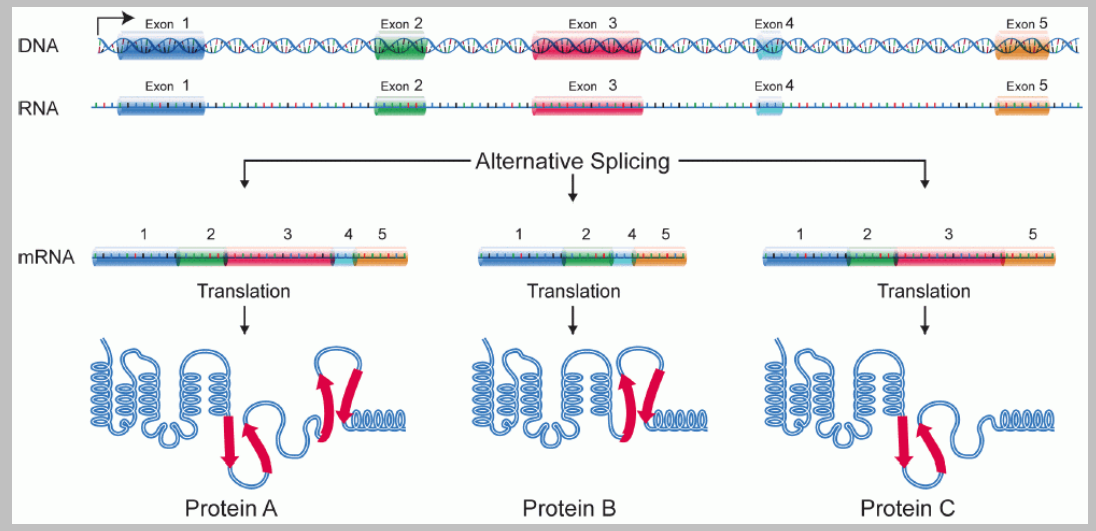


Image credit: "DNA, alternative splicing," by the National Human Genome Research Institute.

THEDOGRAMAPQANANDAYAPTQMTETHEHAT

Splicing exercice

THEDOGRAMAPQANANDAYAPTQMTETHEHAT

Are you able to visually splice that sequence?

Splicing exercice

THEDOGRAMAPQANANDAYAPTQMTETHEHAT

THEDOGRAMAPQANANDAYAPTQMTETHEHAT

Here are the introns.

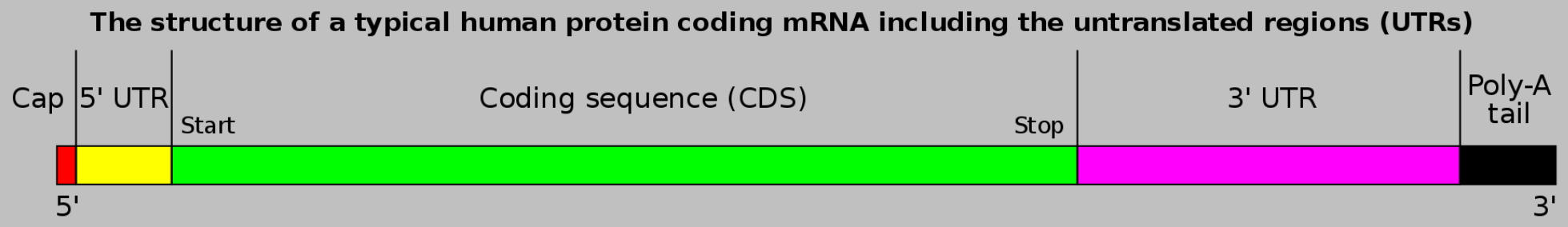
Splicing exercice

THEDOGRAMAPQANANDAYAPTQMTETHEHAT

THEDOGRAMAPQANANDAYAPTQMTETHEHAT

THE DOG R------AN AND A-------TE THE HAT

mature mRNA structure

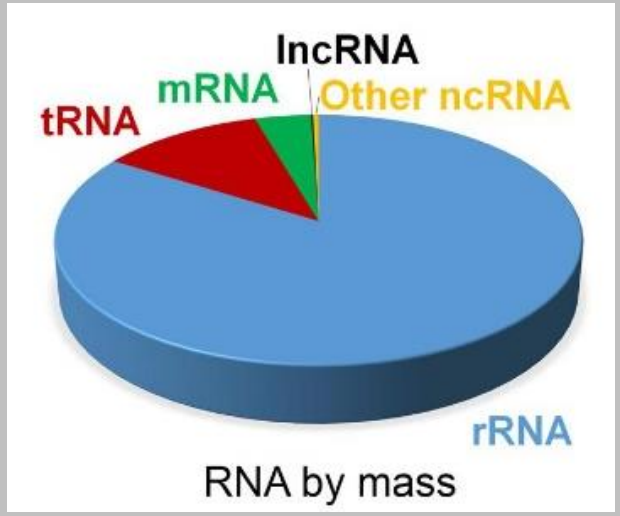


This dogmatic view is INcomplete.

RNA biotypes abundance

Type	Percent of total RNA by mass	Molecules per cell	Average size (kb)	Total weight picograms/cell	Notes	Reference
rRNAs	80 to 90	3–10 × 10^6 (ribosomes)	6.9	10 to 30		Blobel and Potter (1967), Wolf and Schlessinger (1977), Duncan and Hershey (1983)
tRNA	10 to 15	3–10 × 10^7	<0.1	1.5 to 5	About 10 tRNA molecules /ribosome	Waldron and Lacroute (1975)
mRNA	3 to 7	3–10 × 10^5	1.7	0.25 to 0.9		Hastie and Bishop (1976), Carter et al. (2005)
hnRNA (pre-mRNA)	0.06 to 0.2	1–10 × 10^3	10*	0.004 to 0.03	Estimated at 2–4% of mRNA by weight	Mortazavi et al. (2008), Menet et al. (2012)
Circular RNA	0.002 to 0.03	3–20 × 10^3	~0.5	0.0007 to 0.005	Estimated at 0.1–0.2% of mRNA**	Salzman et al. (2012), Guo et al. (2014)
snRNA	0.02 to 0.3	1–5 × 10^5	0.1–0.2	0.008 to 0.04		Kiss and Filipowicz (1992), Castle et al. (2010)
snoRNA	0.04 to 0.2	2–3 × 10^5	0.2	0.02 to 0.03		Kiss and Filipowicz (1992), Cooper (2000), Castle et al. (2010)
miRNA	0.003 to 0.02	1–3 × 10^5	0.02	0.001 to 0.003	About 10^5 molecules per 10 pg total RNA	Bissels et al. (2009)
7SL	0.01 to 0.2	3–20 × 10^4	0.3	0.005 to 0.03	About 1–2 SRP molecules/100 ribosomes	Raue et al. (2007), Castle et al. (2010)
Xist	0.0003 to 0.02	0.1–2 × 10^3	2.8	0.0001 to 0.003		Buzin et al. (1994), Castle et al. (2010)
Other lncRNA	0.03 to 0.2	3–50 × 10^3	1	0.002 to 0.03	Estimated at 1–4% of mRNA by weight	Mortazavi et al. (2008), Ramsköld et al. (2009), Menet et al. (2012)

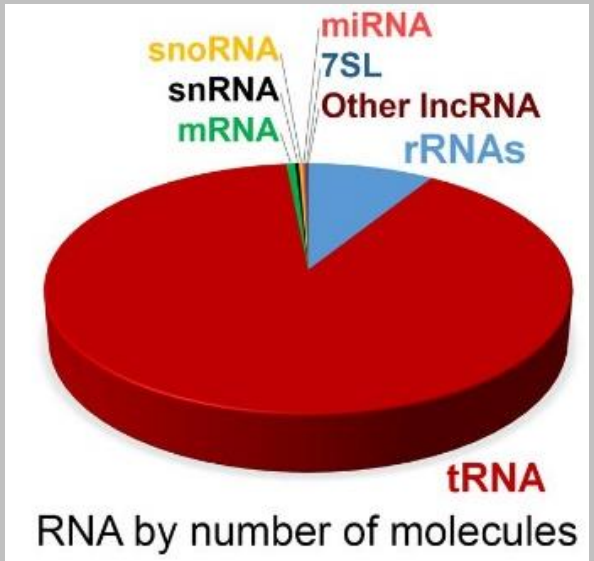
*The size for the average unspliced pre-mRNA is 17 kb; however, most pre-mRNAs are partially spliced at any given time, and the average size of hnRNA is estimated at 10 kb (Salditt-Georgieff et al., 1976).
**Based on the finding that 1–2% of all mRNA species generate circular RNA, which is present at 10% of the level of the parental mRNA.



Alexander F. Palazzo and Eliza S. Lee, Front. Genet., 2015

- ribosomal RNA represents up to 90% of the total RNA mass, the second most represented biotype is
- tRNA, followed by
- mRNA (3% to 7%) and
- pre-mRNA (<0.2%).
- … only later comes the lncRNA, microRNA, circRNA …

Types/abundance of RNA-Seq



Alexander F. Palazzo and Eliza S. Lee, Front. Genet., 2015

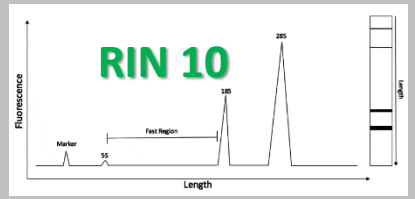
Number of molecules per cell:

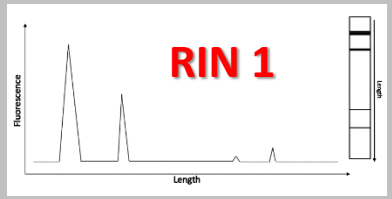
- tRNA: 100E06
- rRNA: 10E06
- mRNA: 1E06
- microRNA: 0.1E06

in vitro

Library preparation

Prepare a complementary DNA (cDNA) library ready for sequencing, with the following main steps:

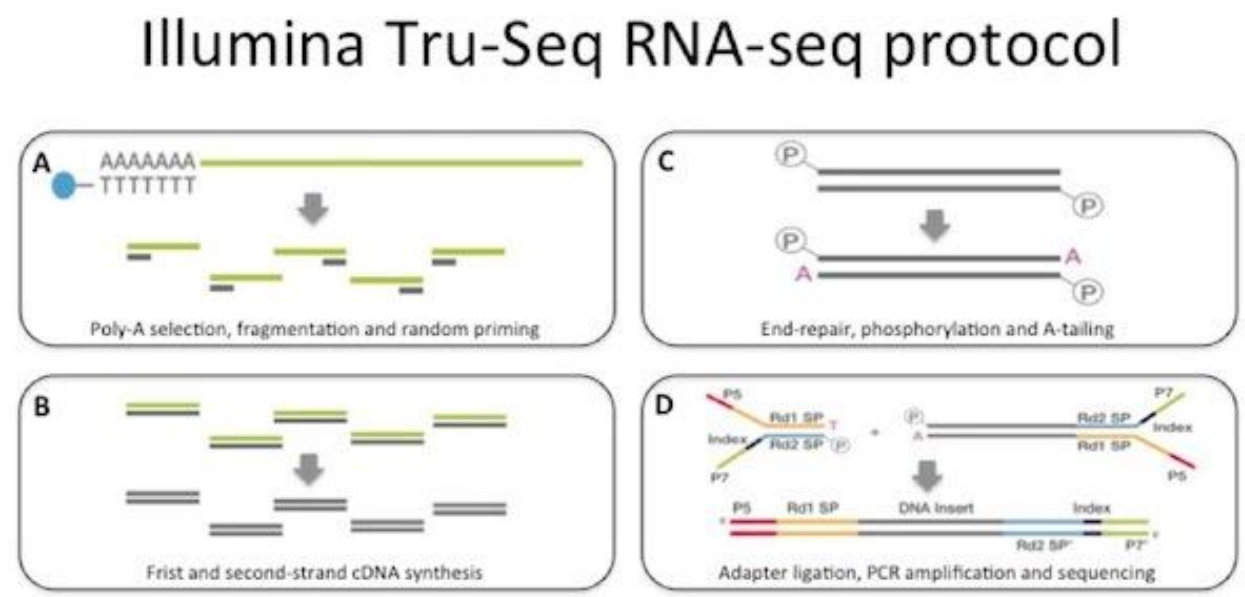




- RNA extraction (RNA Integrity Number)

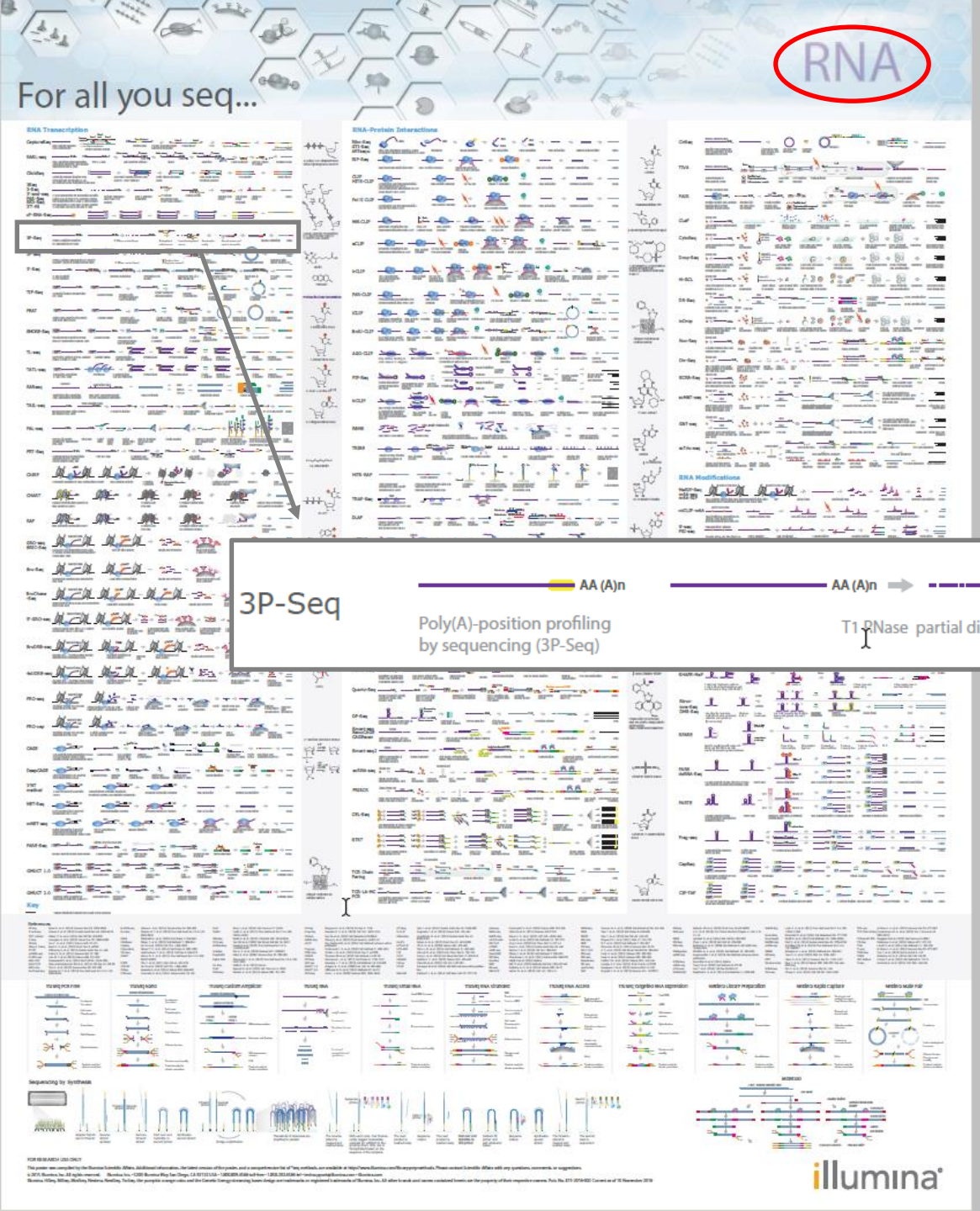
- RNA selection or depletion:
  - polyA selection
  - ribosomal RNA depletion
  - targeted RNA capture
  - small RNA
  - … many library preparation methods …

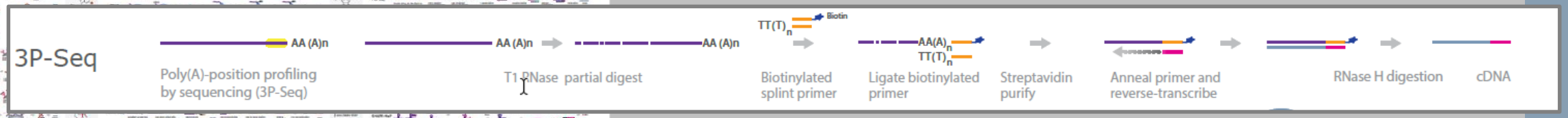
- Fragmentation
- double strand cDNA synthesis (random priming)
- Addition of adapters, indexes and sequencing primers
- Amplification



Library prep begins from 100ng-1ug of Total RNA which is poly-A selected (A) with magnetic beads. Double-stranded cDNA (B) is phosphorylated and A-tailed (C) ready for adapter ligation. The library is PCR amplified (D) ready for clustering and sequencing.

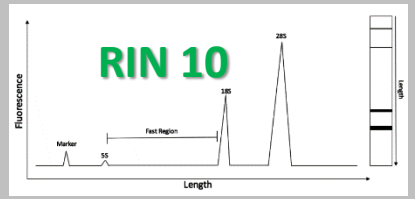
Library preparation

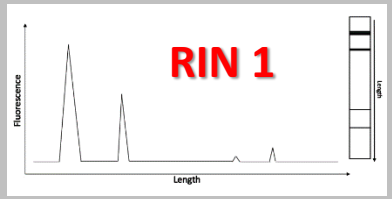




Library preparation

Prepare a complementary DNA (cDNA) library ready for sequencing, with the following main steps:

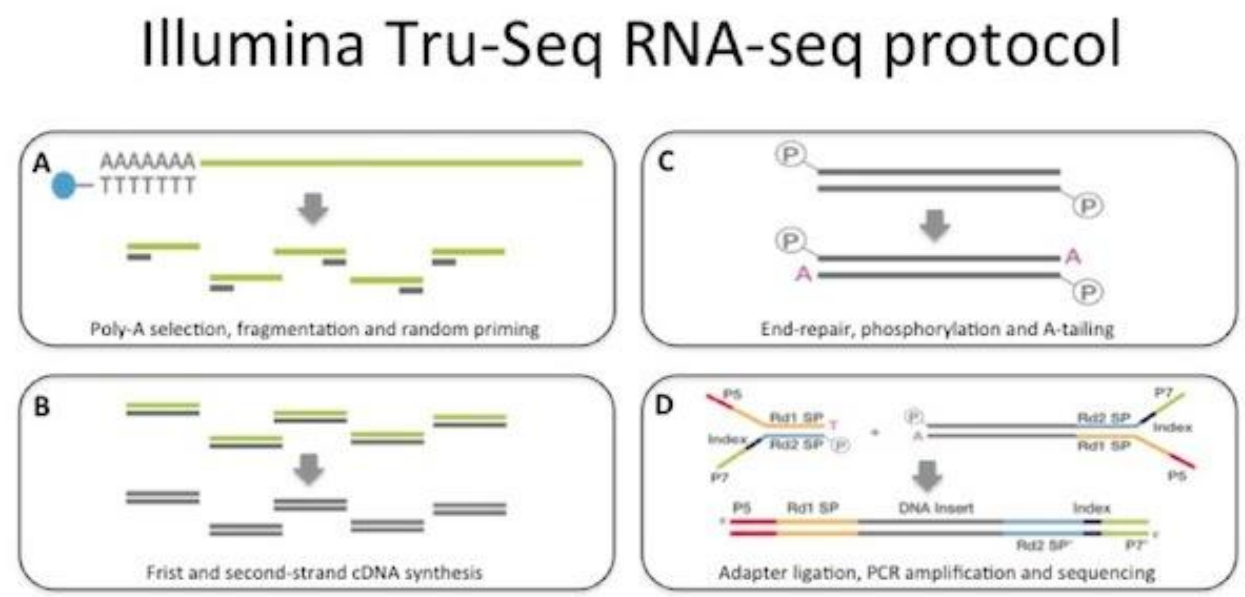




- RNA extraction (RNA Integrity Number)

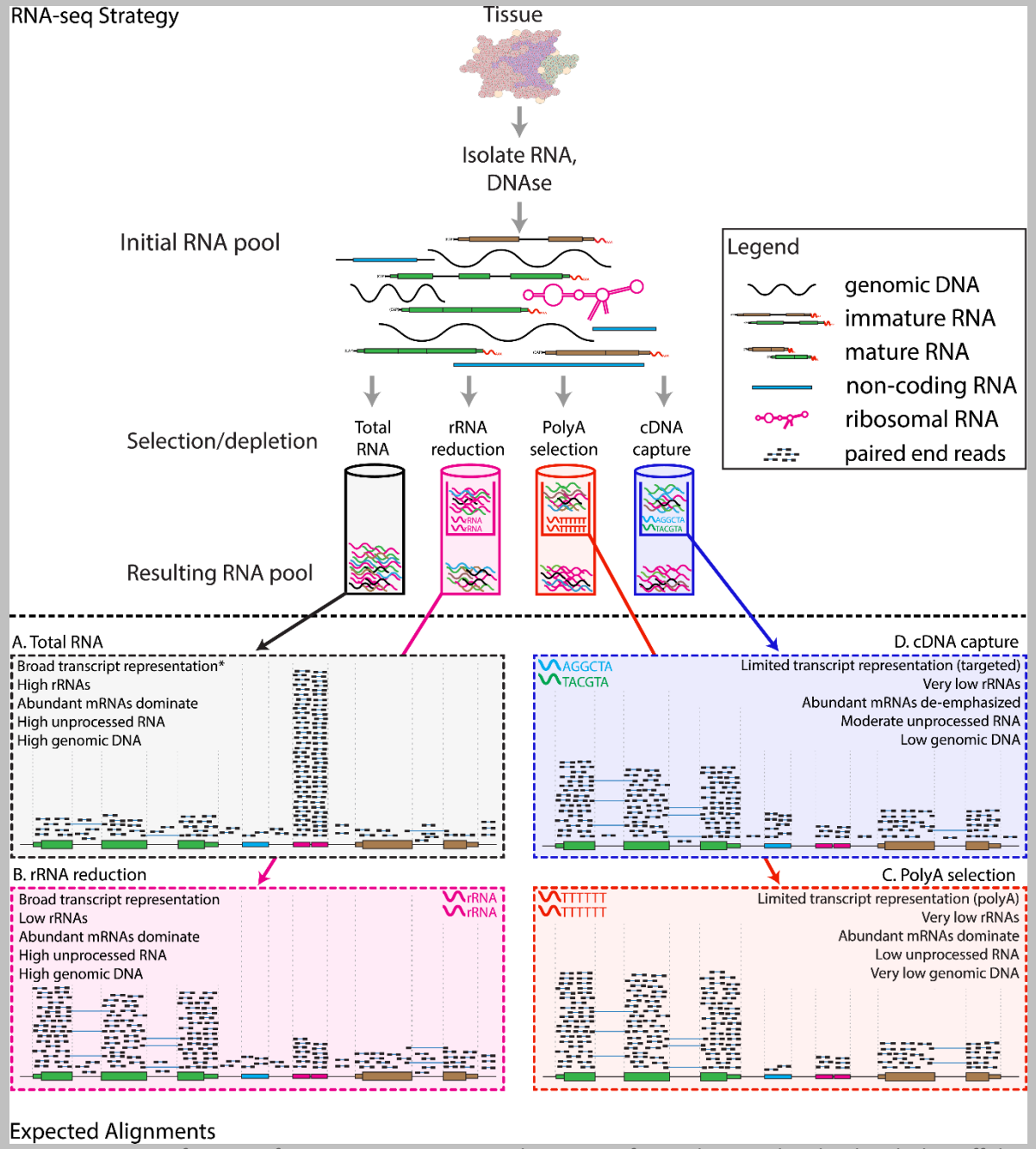
- RNA selection or depletion:
  - polyA selection
  - ribosomal RNA depletion
  - targeted RNA capture
  - small RNA
  - … many library preparation methods …

- Fragmentation
- double strand cDNA synthesis (random priming)
- Addition of adapters, indexes and sequencing primers
- Amplification



Library prep begins from 100ng-1ug of Total RNA which is poly-A selected (A) with magnetic beads. Double-stranded cDNA (B) is phosphorylated and A-tailed (C) ready for adapter ligation. The library is PCR amplified (D) ready for clustering and sequencing.

Library preparation

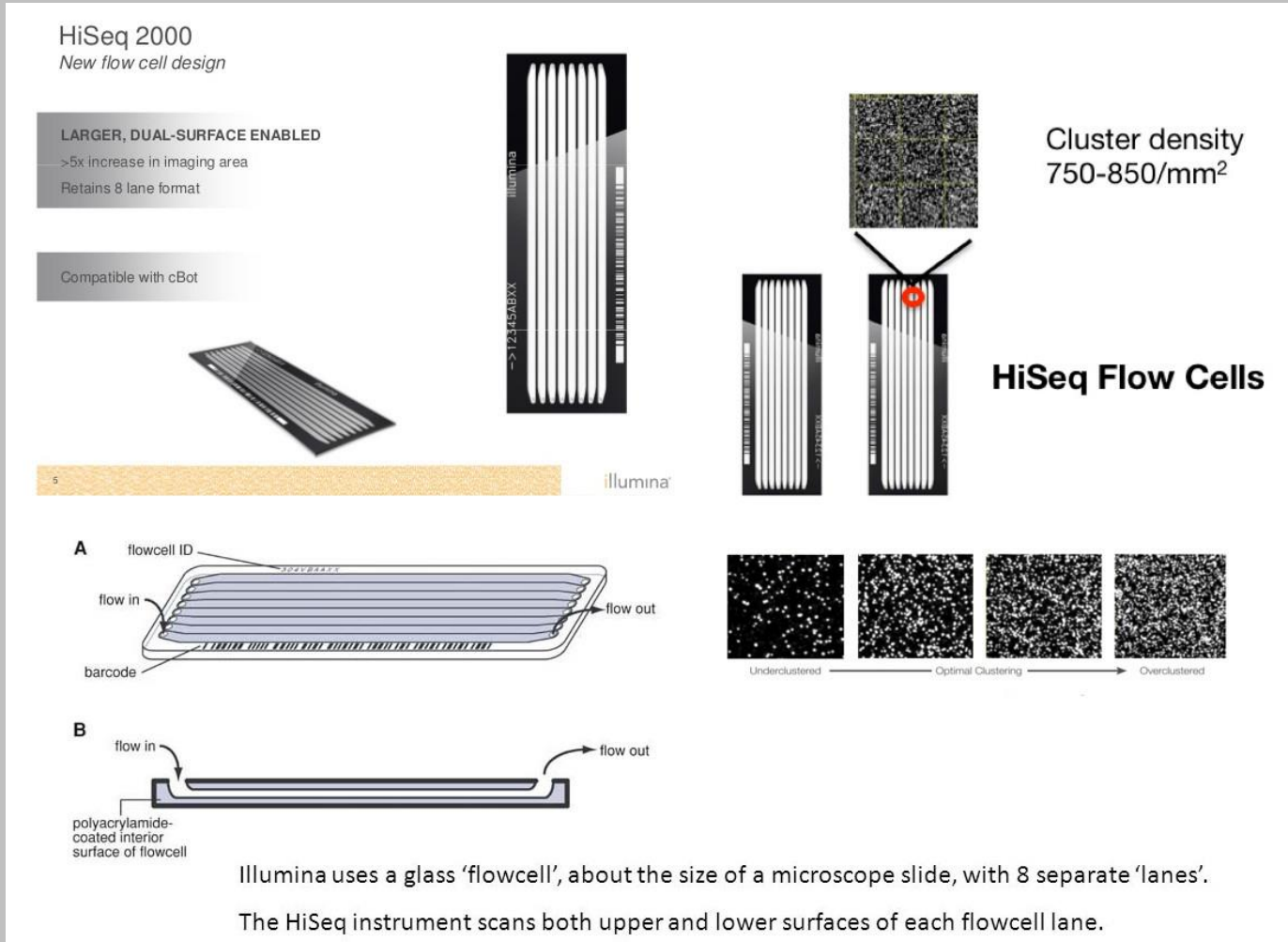


Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud Malachi Griffith

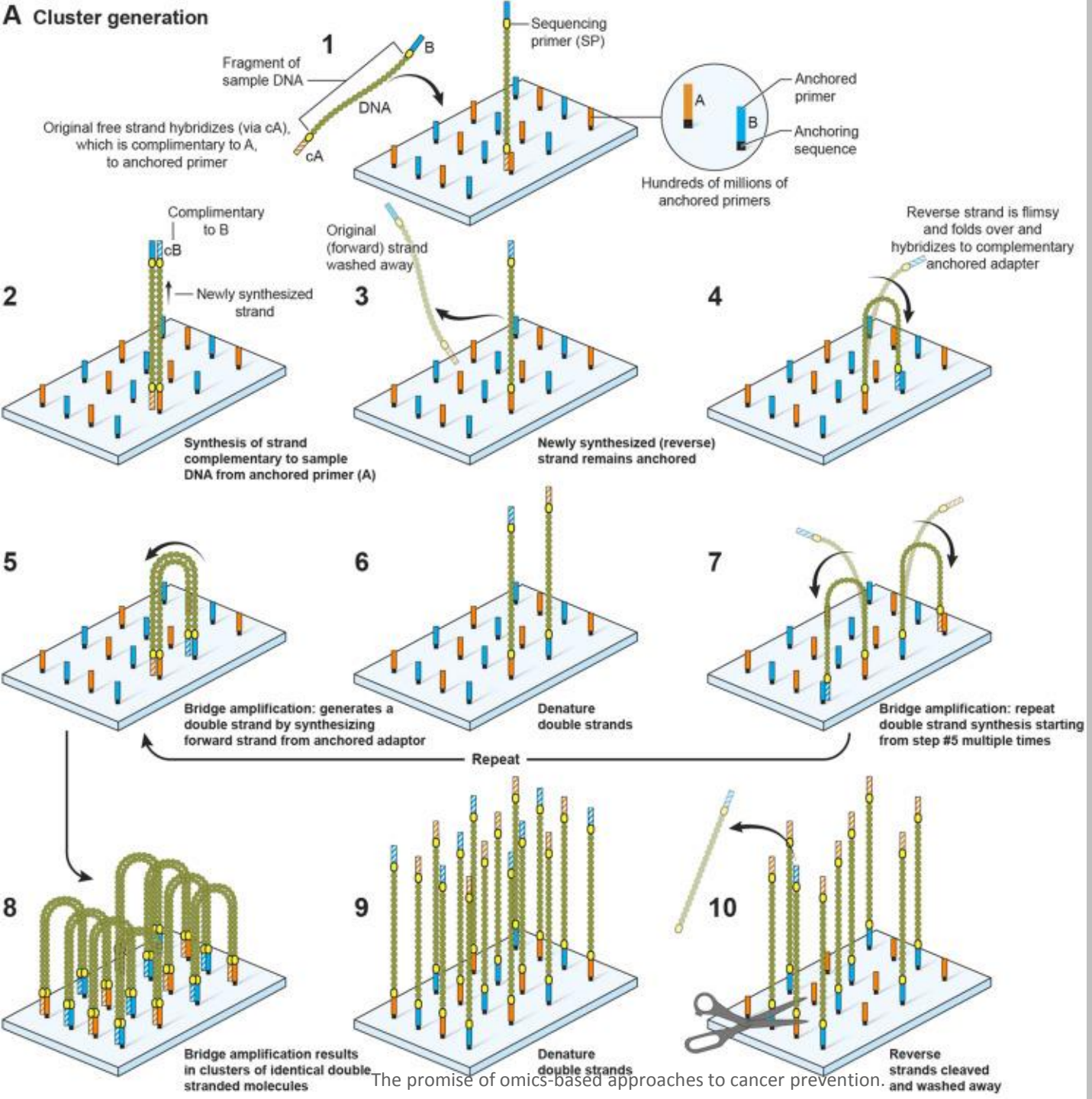
Library preparation

Since the library is prepared, it's now ready for sequencing:

- cluster generation
- Sequencing By Synthesis



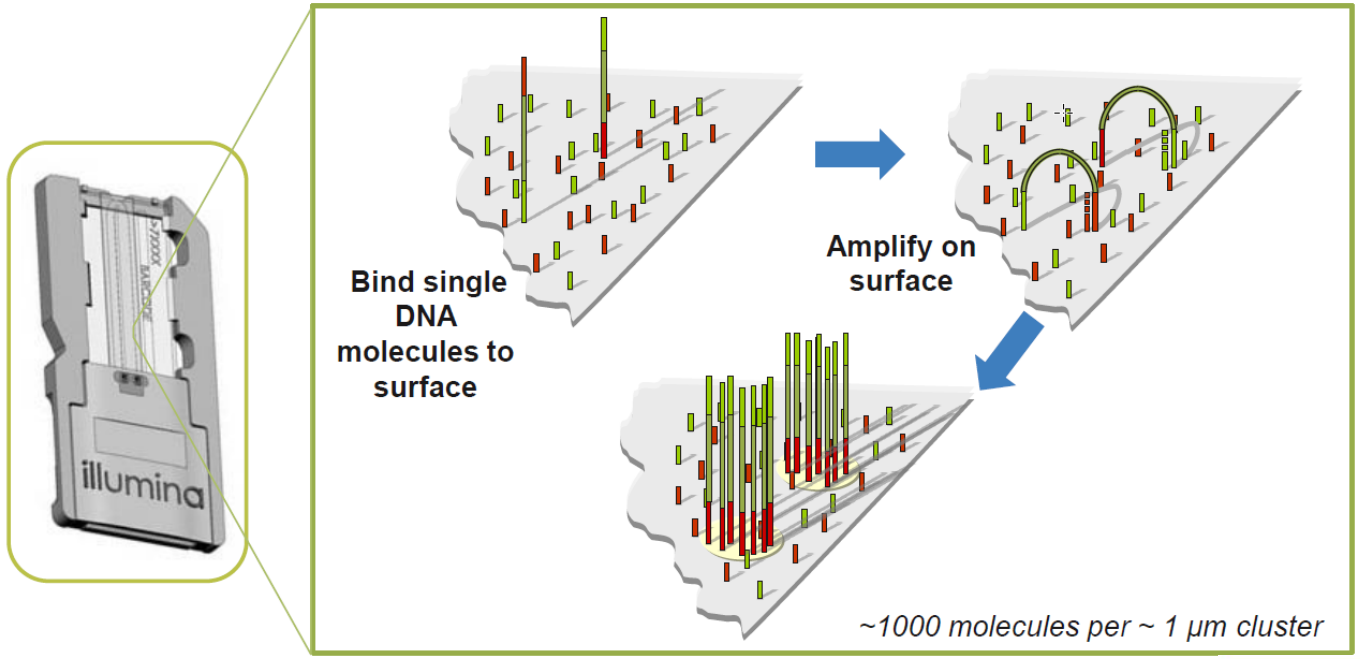
Cluster generation



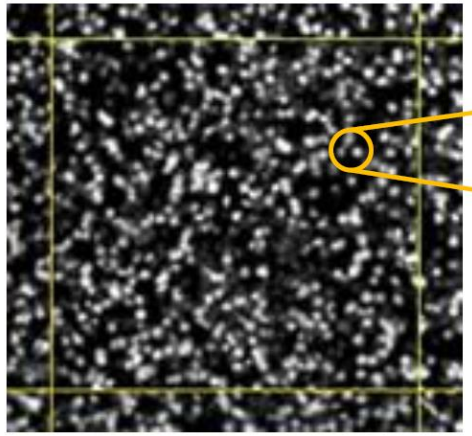
- "Bridge amplification"

- Cluster: clones of fragments from the input library

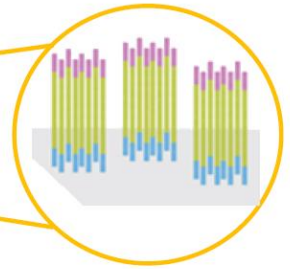
Cluster generation



Clusters are bright spots on an image

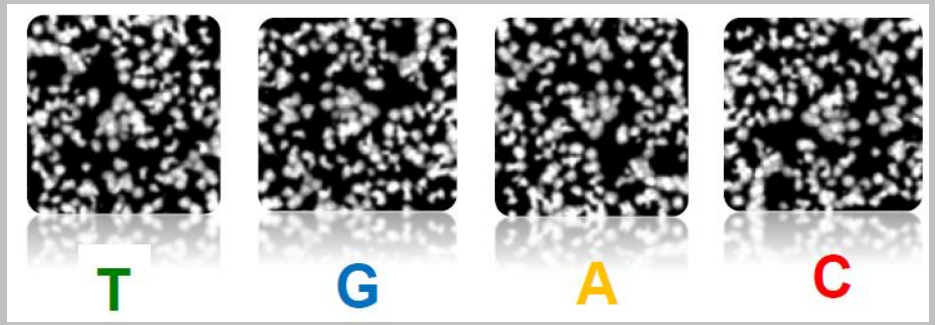


Each cluster represents thousands of copies of the same DNA strand in a 1–2 micron spot



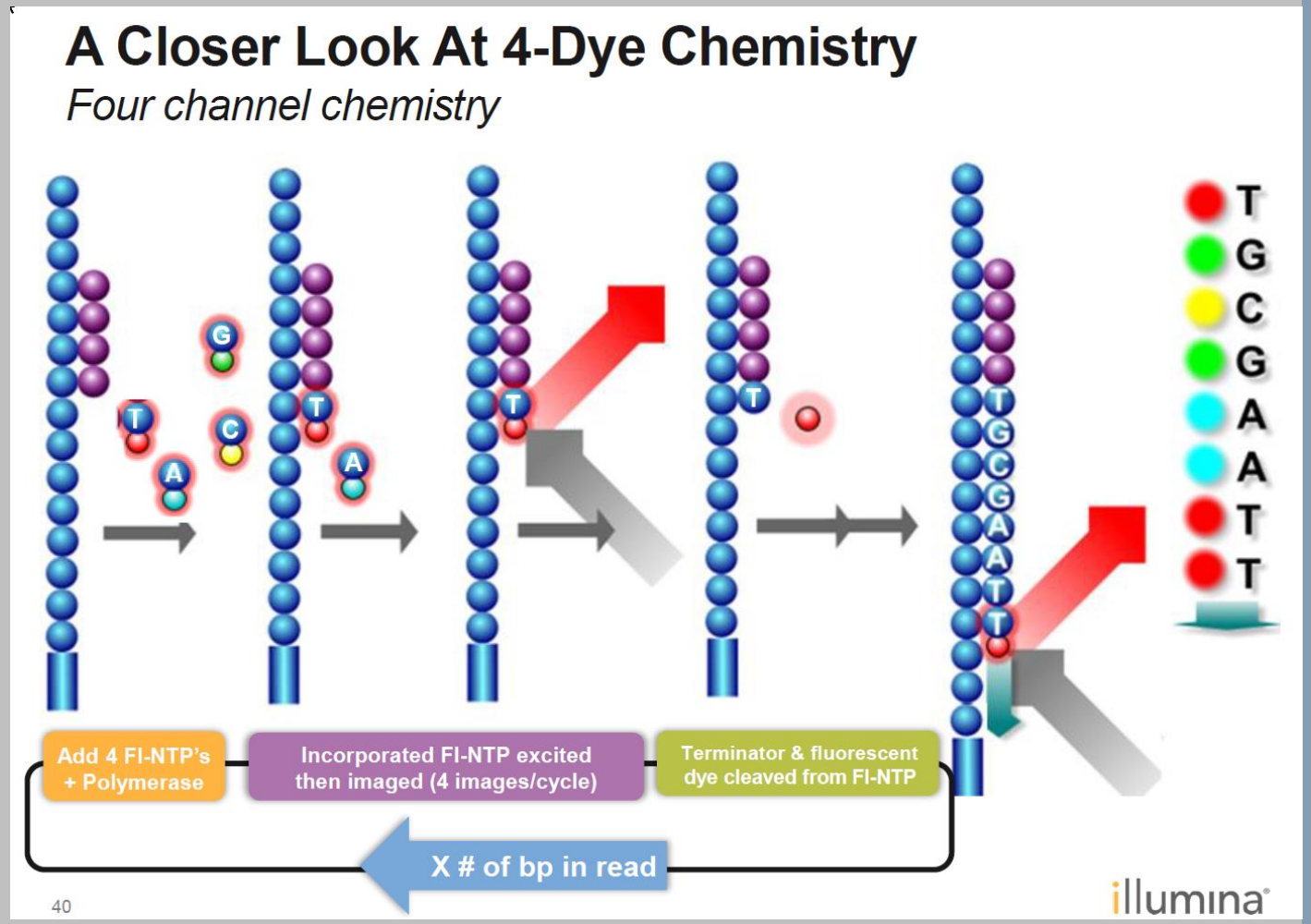
Sequencing By Synthesis

1. Addition of A C G T which contains a fluorophore and a terminator + polymerase
2. Integration of the nucleoside
3. Laser excitation + image capture (4 times!)
4. Removal of fluorescent dye and terminator



These 4 steps are repeated as many times as the read length.

-> computational/storage cost: made on the instrument, on the fly
-> biochemical cost
-> microfluidics
-> optics
-> time: 10-30 h

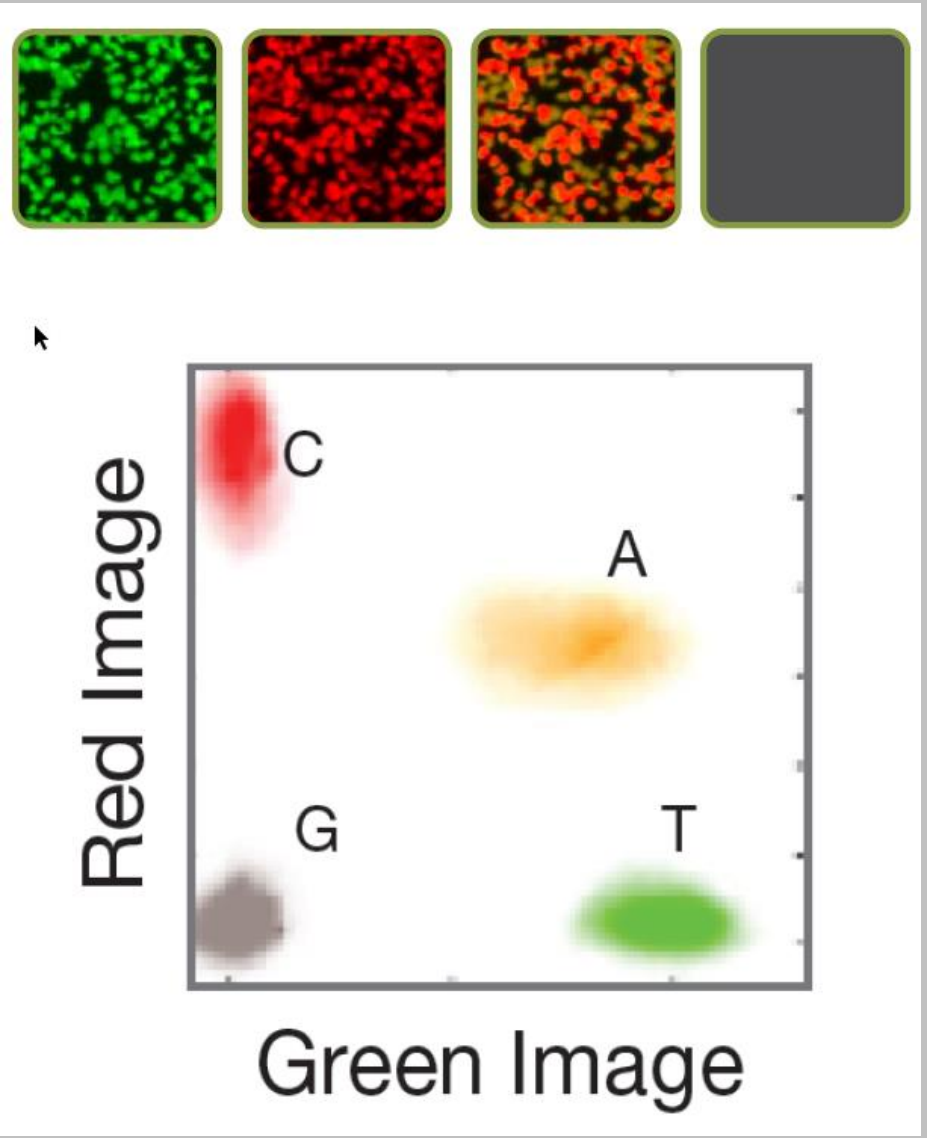


Sequencing By Synthesis

Two channel SBS uses 2 images

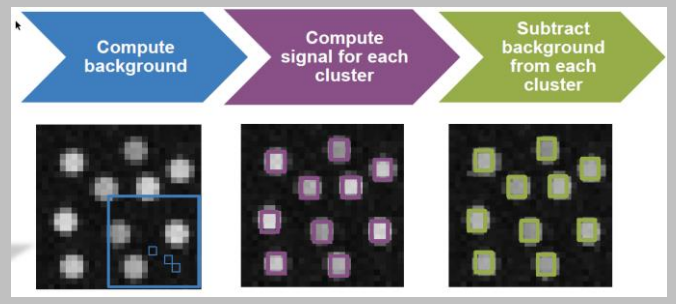
- Builds template over 5 cycles
- Clusters appearing in green only are T
- Clusters appearing in red only are C
- Clusters appearing in both images are A
- Clusters not present in either green nor red are G
- Cluster intensities are plotted and bases are called accordingly

-> time cost reduced by 50%



Base call

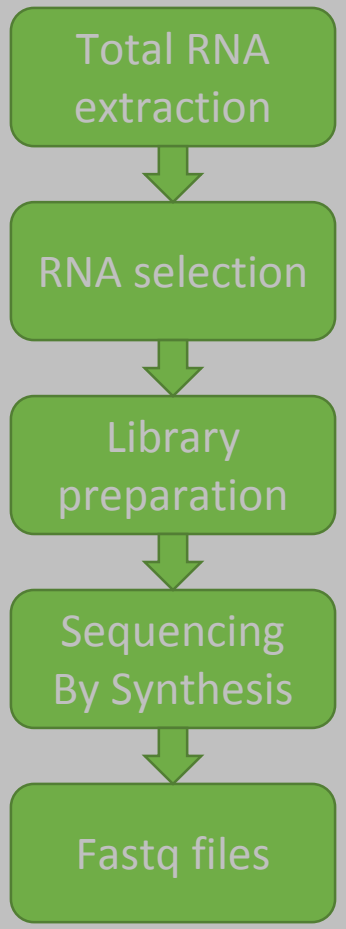
- The Sequencing By Synthesis steps generates BCL files.
- Binary Base Call (BCL) files contains the base and the confidence in the call.
- This file is generated on the fly.
- The bcl generation workflow is the following:
  - Template generation (cluster map)
  - extract intensities
  - intensity norm
  - phasing estimate
  - base call filtering
  - quality score (phred score): Q40 ⇔ Error probability < 0.0001
    - quality score is a combination of metrics, such as: intensity, S/N ratio, phasing..
- Because of the size of the optical sensor, the flow cell is split into multiple tiles, generating several bcl files.
- After the experiment, bcl file is converted into fastq file.



Base quality score

Q-score	Base Call Accuracy	Probability of Incorrect Based Call	ASCII Quality Score
Q10	90%	0.1	+
Q20	99%	0.01	5
Q30	99.9%	0.001	?
Q40	99.99%	0.0001	|

Wrap up


Total RNA extraction
RNA selection
Library preparation
Sequencing By Synthesis
Fastq files

in silico

Fastq file

coordinates of the cluster

Y: cluster did not pass the filter

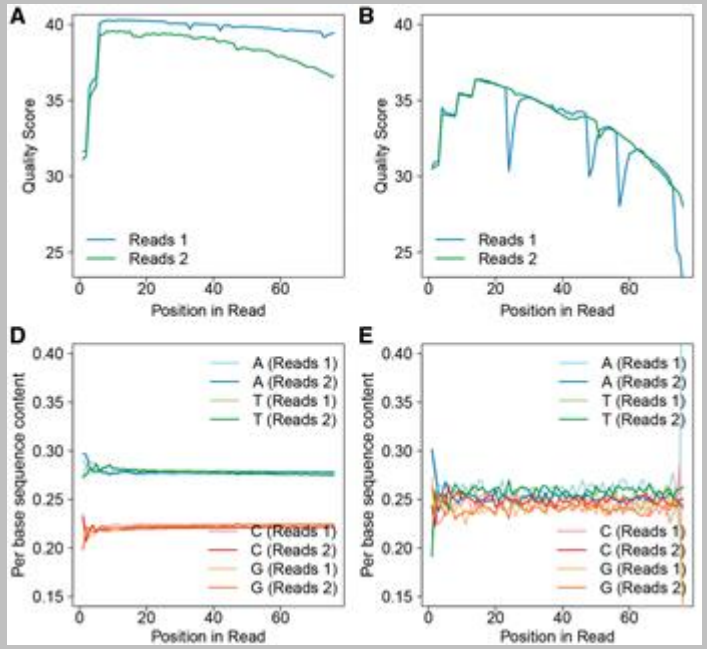
@<instrument>:<run number>:<flowcellID>:<lane>:<tile>: <x-pos>:<y-pos> <read> : <is filtered> : <control number> : <sample number>

unique ID

```
$ gunzip -c 2T210_S1_R1_001.fastq.gz|head -4
@NB551409:68:H7CN7BGXB:1:11101:16989:1046 1:N:0:ATCCACTG+ACGCACCT
CTCTATACCANTGGTCCAATGGGCTTAAAAAAGAGCAAATATTACCAAATGGATATGCTCTGAAGTTGTCGTTAAT
+
AAAAAEEEEE#EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

- Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description (like a FASTA title line).
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character, this is historical and mysterious.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

Quality Controls



Quanhu Sheng et al., 2017, *Briefings in Functional Genomics*

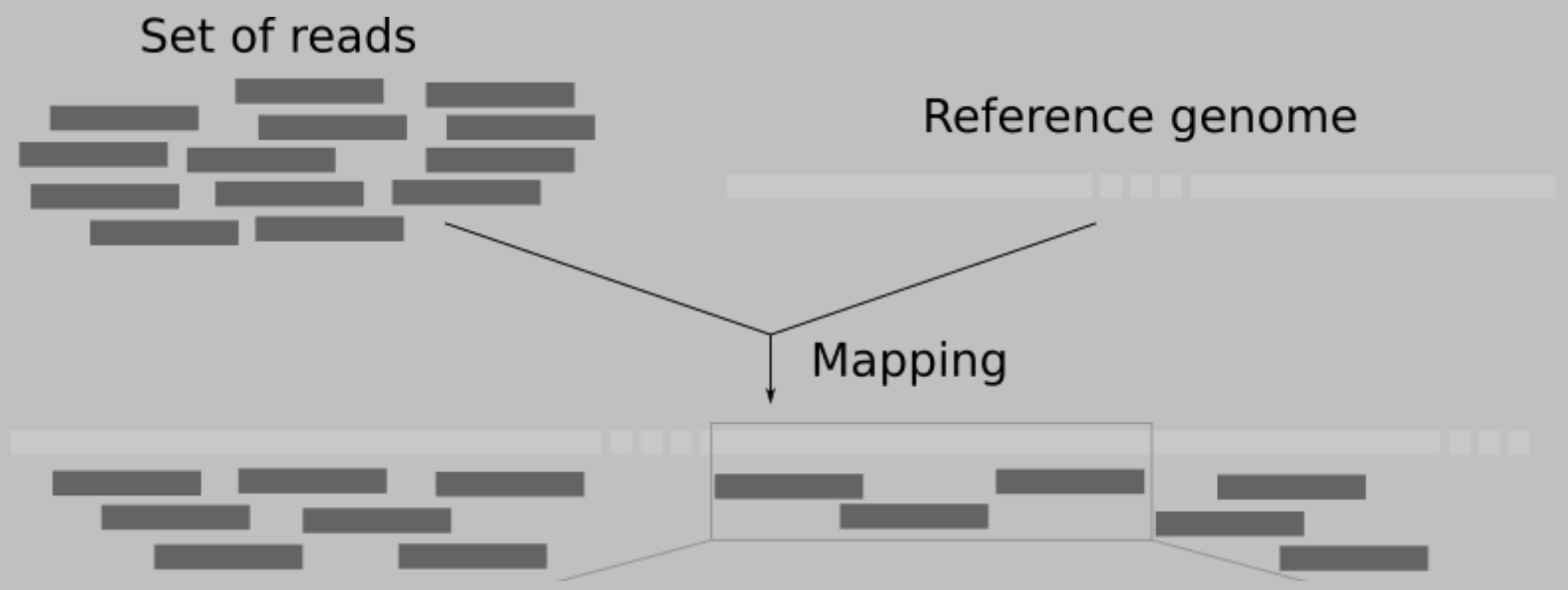
(A) Example of a long RNA-seq sample with expected base quality score. Read 2 tends to have a slightly lower median base score than read 1, but it is not usually a quality concern.

(**B**) Example of a long RNA-seq sample with potential base quality problem, as denoted by the sudden drops of median base quality in read 2 of pair-end read sequencing.

(**D**) Example of a long RNA-seq sample with expected nucleotide distribution, as denoted by the stable nucleotide distribution across the samples.

(**E**) Example of a long RNA-seq sample with a potential nucleotide distribution issue, as denoted by the unstable distribution across the cycles.

Mapping - the ingredients

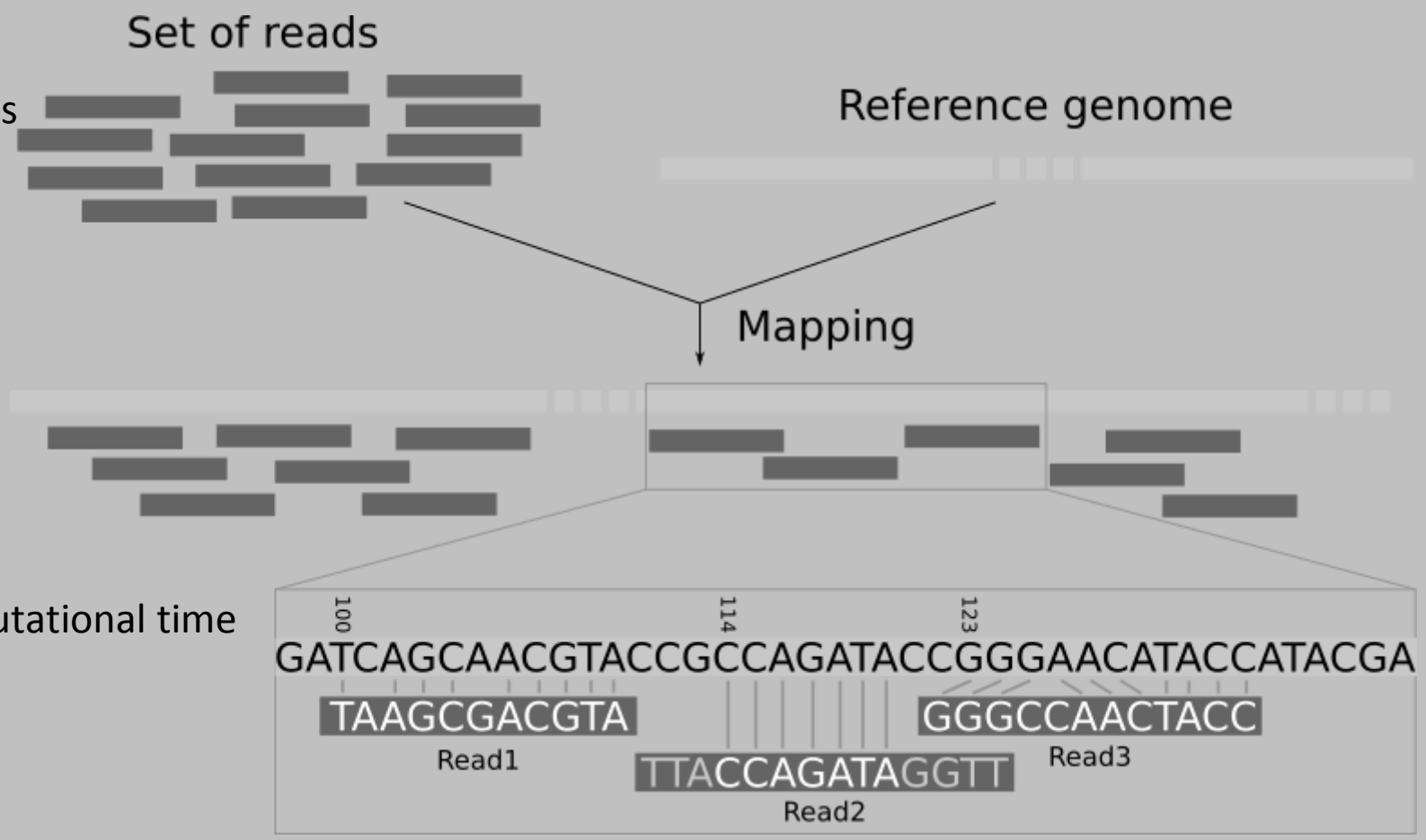


Mapping - the challenge

approx. 50 millions reads

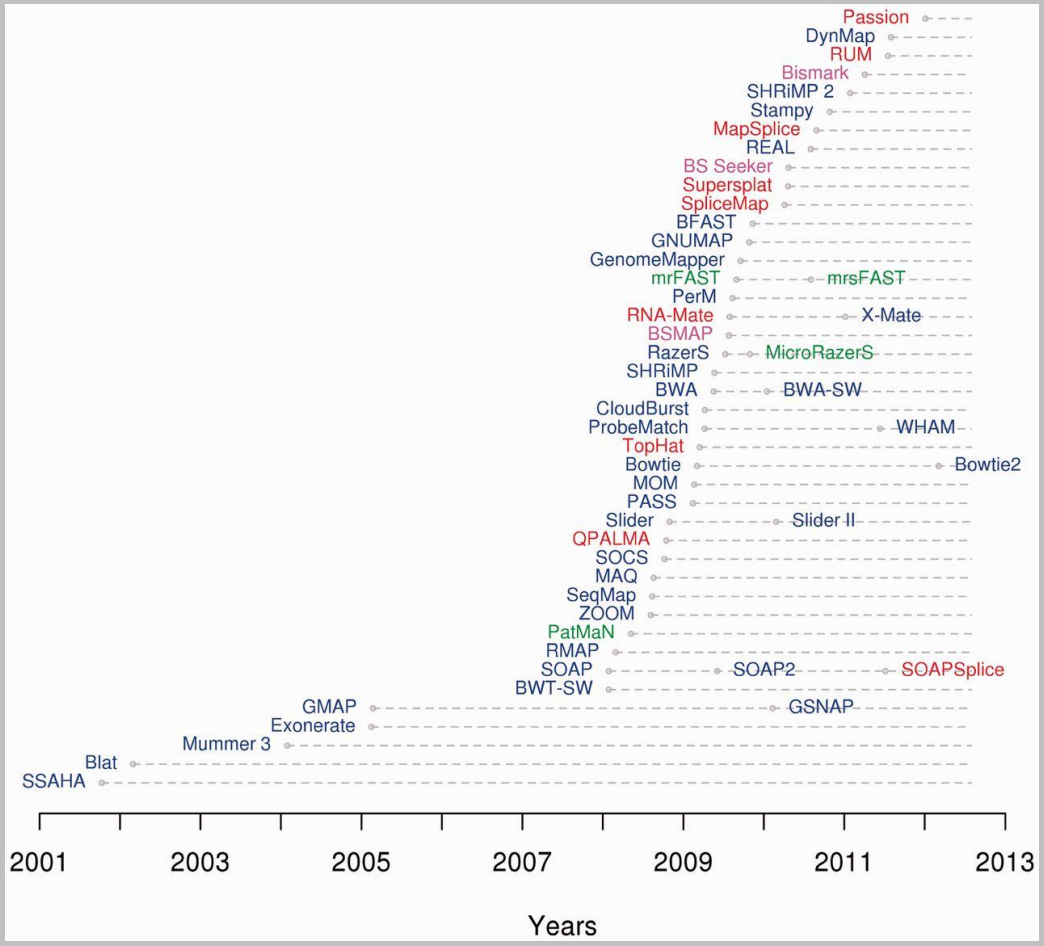
3 billions bp

reasonable computational time



Mapping tools

Mappers timeline (since 2001). DNA mappers are plotted in blue, RNA mappers in red, miRNA mappers in green.



Bioinformatics, Volume 28, Issue 24, December 2012, Pages 3169–3177

Alignment scoring

Sequence One : GGCTGG

Sequence Two : GAGG

G	G	C	T	G	G
G	A	-	-	G	G
10	-5	-5	-1	10	10
10	5	0	-1	9	19

Your cumulative score

- **Reward** for a match (e.g. +10), **penalty** for a mismatch (e.g. -5)
- **Penalty** for gaps
  - *Linear:* every gap same penalty (e.g. -5)
  - *Affine:* gap open vs gap extend (e.g. -5 and -1)
- Different tools use different scoring values (and give different results)

Alignment scoring

- **Reference:** **AAA CAGTGA GAA**
- **Observed:** **AAA TCTCT GAA**

Alignment:

```
AAA-CAGTGAGAA
|||-|--|::|||
AAATC—TCTGAA

AAACAGTGAGAA
|||-::|::|||
AAA-TCTCTGAA

AAACAGTGAGAA
|||:-:|::|||
AAAT-CTCTGAA

AAACAGTCA-----GAA
|||-----------|||
AAA------TCTCTGAA
```

Alignment scoring

- **Reference:** **AAA CAGTGA GAA**
- **Observed:** **AAA TCTCT GAA**

Alignment:

```
AAA-CAGTGAGAA
|||-|--|::|||
AAATC—TCTGAA
```

Novoalign

```
AAACAGTGAGAA
|||-::|::|||
AAA-TCTCTGAA
```

Ssaha2

```
AAACAGTGAGAA
|||:-:|::|||
AAAT-CTCTGAA
```

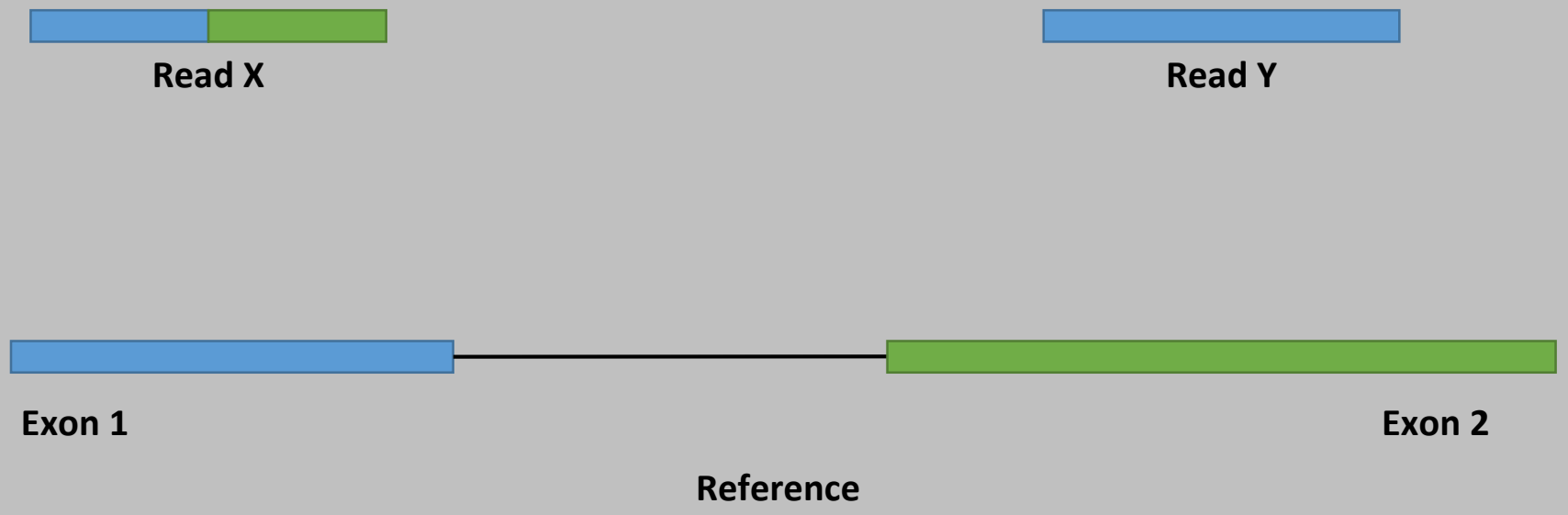
BWA

```
AAACAGTCA-----GAA
|||-----------|||
AAA------TCTCTGAA
```

Complete Genomics

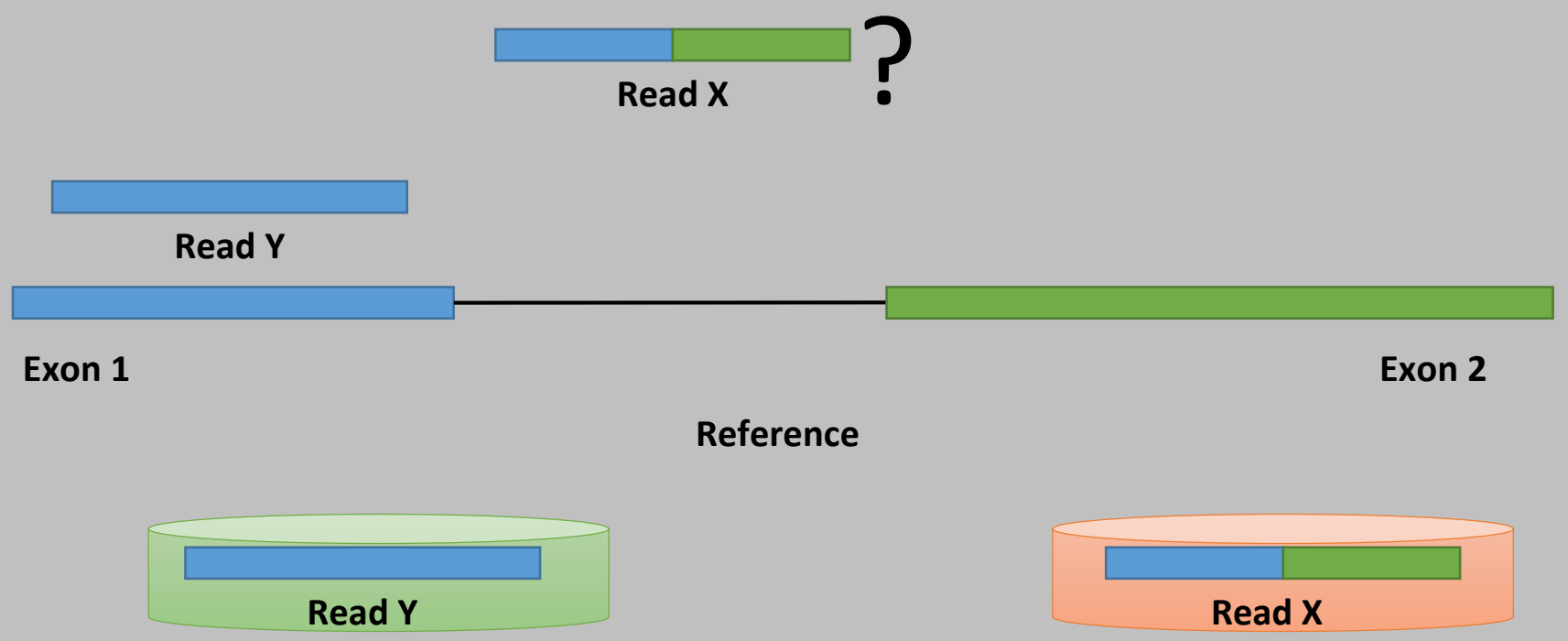
Mapping - bowtie

Bowtie / TopHat



Mapping - bowtie

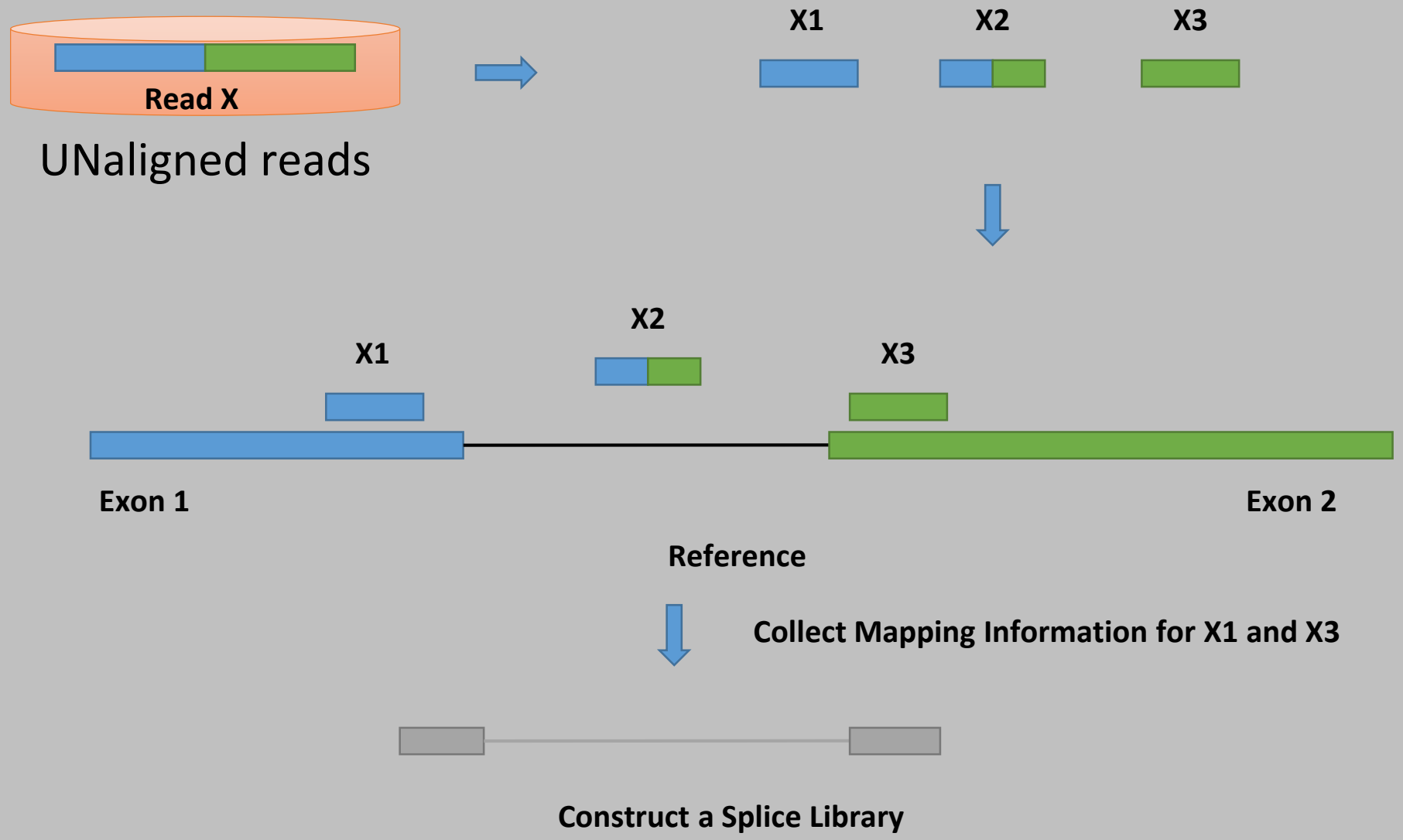
Bowtie / TopHat



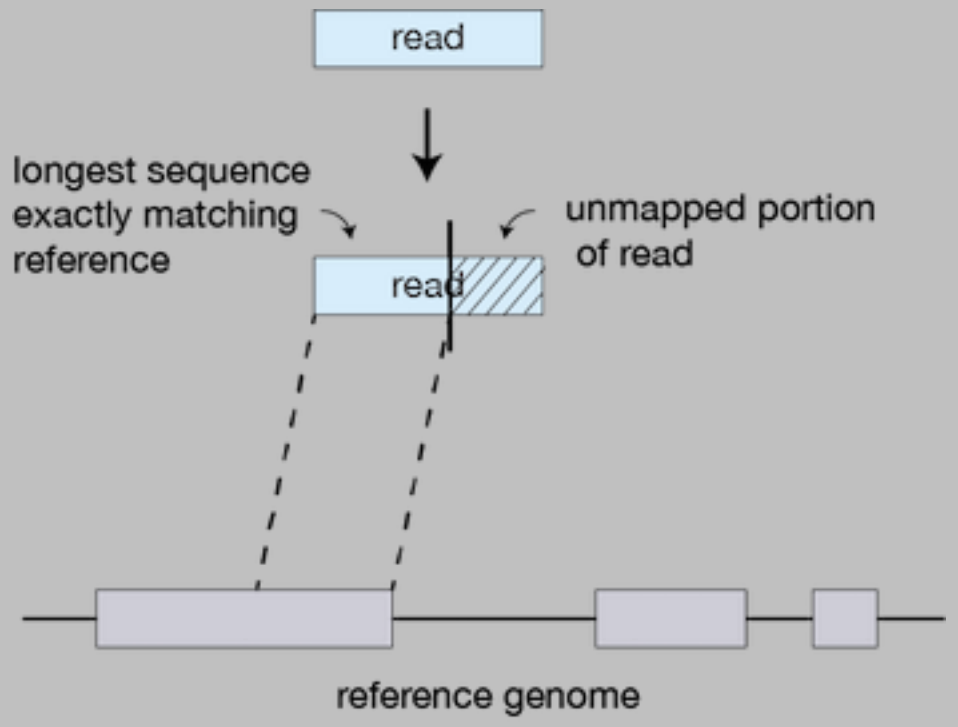
Aligned reads

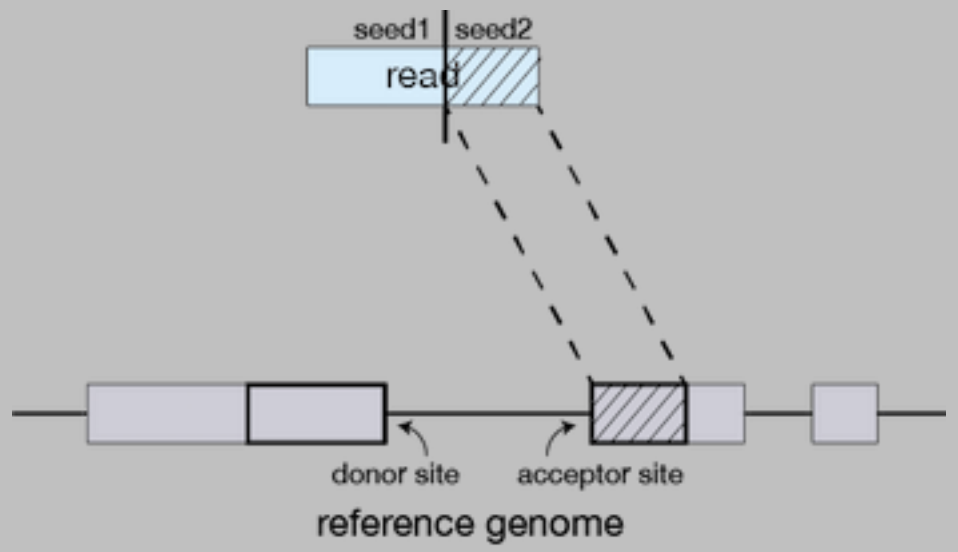
UNaligned reads

Mapping - bowtie



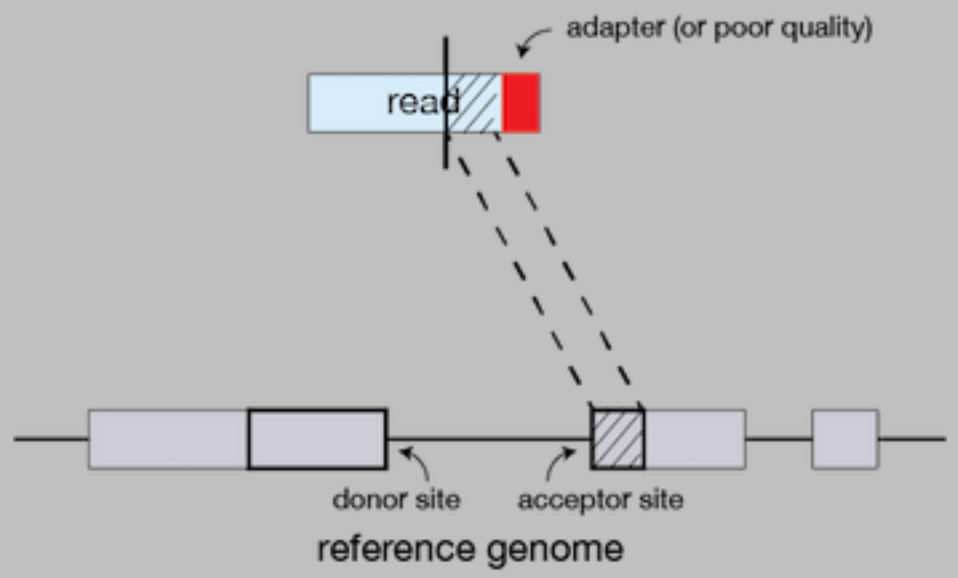
Mapping - STAR





1. Find the longest perfect match seed
2. Independent search for the unmapped portion

Mapping - STAR



If extension does not give a good alignment, then the poor quality or adapter sequence (or other contaminating sequence) will be **soft clipped**.

The separate seeds are **stitched** together to create a complete read by first clustering the seeds together based on proximity to a set of 'anchor' seeds, or seeds that are not multi-mapping.

Then the seeds are stitched together based on the best alignment for the read (scoring based on mismatches, indels, gaps, etc.).

Genome versions

- Completion of the *Human Genome Project* in 2003.
- Continuation by the *Genome Reference Consortium* (GRC):
  - NCBI
  - Welcome Trust Sanger Institute
  - European Bioinformatics Institute
  - Genome Institute at Washington University

- Now: closing gaps, centromers/telomers representation, error correction

	Release Name	UCSC Version	Release Date
	NCBI Build 34	hg16	July 2003
	NCBI Build 35	hg17	May 2004
Human	NCBI Build 36.1	hg18	March 2006
	GRCh37	hg19	February 2009
	GRCh38	hg38	December 2013

Genome sequence file

```
head  ~/Tools/Genome/Human/hg19/chrM.fa
>chrM
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCAT
TTGGTATTTTCGTCTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCTG
GAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCATT
CTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACCTACTA
AAGTG
GTCTG
```


```
head  ~/Tools/Genome/Human/hg19/chrM.fa
>chrM
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCAT
TTGGTATTTTCGTCTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCTG
GAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCATT
CTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACCTACTA
AAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATAACAATTGAAT
GTCTGCACAGCCGCTTTCCACACAGACATCATAACAAAAAATTTCCACCA
```

Genome sequence file

```
$ wc -l `ls  |egrep 'chr[0-9XYM]+.fa'`
   2710696 chr10.fa
   2700132 chr11.fa
   2677039 chr12.fa
   2303399 chr13.fa
   2146992 chr14.fa
   2050629 chr15.fa
   1807097 chr16.fa
   1623906 chr17.fa
   1561546 chr18.fa
   1182581 chr19.fa
   4985014 chr1.fa
   1260512 chr20.fa
    962599 chr21.fa
   1026093 chr22.fa
   4863989 chr2.fa
   3960450 chr3.fa
   3823087 chr4.fa
   3618307 chr5.fa
   3422303 chr6.fa
   3182775 chr7.fa
   2927282 chr8.fa
   2824270 chr9.fa
       333 chrM.fa
   3105413 chrX.fa
   1187473 chrY.fa
  61913917 total
```

=> leading to 3 095 695 850 nucleotides.

Sequence Alignment Map - SAM

```
$ more S1AR005ACAGTG.sam
@HD     VN:1.0  SO:unsorted
@SQ     SN:1    LN:248956422
@SQ     SN:2    LN:242193529
@PG     ID:bowtie2      PN:bowtie2      VN:2.3.4.1     CL:"/mnt/pcpnfs/homedirs/luxgen/Tools/Bowtie/bowtie2-
2.3.4.1-linux-x86_64/bowtie2-align-l --wrapper basic-0 -p20 -t -x
/mnt/pcpnfs/homedirs/luxgen/Tools/Genome/ensembl/Homo_sapiens_GRCh38 --met-file S1AR005
ACAGTG.metrics --passthrough -1 ../fastq/S1AR005ACAGTG_S1_R1_001.fastq.gz -2
../fastq/S1AR005ACAGTG_S1_R2_001.fastq.gz"

NB551409:37:H5FHLBGXB:1:11101:3778:1083 83      16      177285  1       75M     =       177055  -305
CCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCNNCNGAGTTCACCCNTGCGGTGCACGCCTCCCG
EEE/AEEEEEEEEEEEEEAEEEEEAEEEEEEEEEEEEE##E#EAEEEEEEE#EEEEEEEEEEEEAAAAA  AS:i:-9  XS:i:-9 XN:i:0  XM:i:5
XO:i:0  XG:i:0  NM:i:5  MD:Z:42C0G1C10C17T0     YS:i:-3 YT:Z:CP

NB551409:37:H5FHLBGXB:1:11101:3778:1083 163     16      177055  1       76M     =       177285  305
GGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGCCCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTTC
AA/AA/E/EEEEA/EEEEEEEEEEEEEEAEEE/EEEEEEEEEEEEEAEEEEEEEEEE6E<EEE/EE/EEEE AS:i:-3  XS:i:-3 XN:i:0  XM:i:1
XO:i:0  XG:i:0  NM:i:1  MD:Z:35A40      YS:i:-9 YT:Z:CP

NB551409:37:H5FHLBGXB:1:11101:8321:1083 99      10      1051875 42      76M     =       1051950 150
AGGAAAACTTTTTTTGANACAGAGTCTCNTNNTTTCGCCCAGGCTGAAGTGCAGTGGTGCGATCTTGGCTCACTG
AAAAAEEEEEEEEAEEE#EEEEEEEEE#E##EEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEE<AEEE AS:i:-4  XN:i:0  XM:i:4  XO:i:0
XG:i:0  NM:i:4  MD:Z:18G10G1T0C43       YS:i:0  YT:Z:CP
```

Sequence Alignment Map - SAM

```
$ more S1AR005ACAGTG.sam
@HD     VN:1.0  SO:unsorted
@SQ     SN:1    LN:248956422
@SQ     SN:2    LN:242193529
@PG     ID:bowtie2      PN:bowtie2      VN:2.3.4.1      CL:"/mnt/pcpnfs/homedirs/luxgen/Tools/Bowtie/bowtie2-
2.3.4.1-linux-x86_64/bowtie2-align-l --wrapper basic-0 -p20 -t -x
/mnt/pcpnfs/homedirs/luxgen/Tools/Genome/ensembl/Homo_sapiens_GRCh38 --met-file S1AR005
ACAGTG.metrics --passthrough -1 ../fastq/S1AR005ACAGTG_S1_R1_001.fastq.gz -2
../fastq/S1AR005ACAGTG_S1_R2_001.fastq.gz"
```

Header:

- version and sorting status
- reference sequence name (chromosome) + length
- program used for read processing

More information can be added: https://samtools.github.io/hts-specs/SAMv1.pdf

Sequence Alignment Map - SAM

```
$ more S1AR005ACAGTG.sam

NB551409:37:H5FHLBGXB:1:11101:3778:1083 83      16      177285  1       75M     =       177055  -305
CCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCNNCNGAGTTCACCCNTGCGGTGCACGCCTCCCG
EEE/AEEEEEEEEEEEEEEAEEEEEAEEEEEEEEEEEEE##E#EAEEEEEEE#EEEEEEEEEEEEAAAAA  AS:i:-9  XS:i:-9 XN:i:0  XM:i:5
XO:i:0  XG:i:0  NM:i:5  MD:Z:42C0G1C10C17T0     YS:i:-3 YT:Z:CP

NB551409:37:H5FHLBGXB:1:11101:3778:1083 163     16      177055  1       76M     =       177285  305
GGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGCCCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTTC
AA/AA/E/EEEEA/EEEEEEEEEEEEEEEAEEE/EEEEEEEEEEEEEAEEEEEEEEEE6E<EEE/EE/EEEE AS:i:-3  XS:i:-3 XN:i:0  XM:i:1
XO:i:0  XG:i:0  NM:i:1  MD:Z:35A40      YS:i:-9 YT:Z:CP

NB551409:37:H5FHLBGXB:1:11101:8321:1083 99      10      1051875 42      76M     =       1051950 150
AGGAAAACTTTTTTTGANACAGAGTCTCNTNNTTTCGCCCAGGCTGAAGTGCAGTGGTGCGATCTTGGCTCACTG
AAAAAEEEEEEEEAEEE#EEEEEEEEE#E##EEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEE<AEEE AS:i:-4  XN:i:0  XM:i:4  XO:i:0
XG:i:0  NM:i:4  MD:Z:18G10G1T0C43       YS:i:0  YT:Z:CP
```

Read ID

@<instrument>:<run number>:<flowcellID>:<lane>:<tile>: <x-pos>:<y-pos>

Position

CIGAR

read sequence

read quality score

Counting

- The purpose of the counting is to estimate the number of reads (counts) associated with each feature of interest (gene, exon, transcript).
- Input required: BAM/SAM files + General Transfer Format (**GTF**) file.
- Output generated: a gene expression matrix with genes as rows and samples as column.
- Top popular tools *htseq-count* (more cited, same developers as *DESeq*) and *featureCounts* (a way faster, Unix and R package, a bit more liberal).

General Transfer Format - GTF

```
$ head ../../ensembl/Homo_sapiens.GRCh38.93.gtf|cut -f 1-8
#!genome-build GRCh38.p12
#!genome-version GRCh38
#!genome-date 2013-12
#!genome-build-accession NCBI:GCA_000001405.27
#!genebuild-last-updated 2018-01
1       havana  gene    11869   14409   .       +       .
1       havana  transcript      11869   14409   .       +       .
1       havana  exon    11869   12227   .       +       .
1       havana  exon    12613   12721   .       +       .
1       havana  exon    13221   14409   .       +       .
```

chr id

gene feature

end location

strand

source

start location

score

reading frame

attributes

```
$ head -8 ../../ensembl/Homo_sapiens.GRCh38.93.gtf|cut -f 9
#!genome-build GRCh38.p12
#!genome-version GRCh38
#!genome-date 2013-12
#!genome-build-accession NCBI:GCA_000001405.27
#!genebuild-last-updated 2018-01
gene_id "ENSG00000223972"; gene_version "5"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene";
gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype
"transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-202"; transcript_source "havana"; transcript_biotype "processed_transcript"; tag "basic";
transcript_support_level "1";
gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; exon_number "1"; gene_name "DDX11L1"; gene_source "havana";
```

Counting

- The purpose of the counting is to estimate the number of reads (counts) associated with each feature of interest (gene, exon, transcript).
- Input required: BAM/SAM files + General Transfer Format (GTF) file.
- Output generated: a gene expression matrix with genes as rows and samples as column.
- Top popular tools *htseq-count* (more cited, same developers as *DESeq*) and *featureCounts* (a way faster, Unix and R package, a bit more liberal).

aligned read:

start: 113217600 end: 113217650

GTF

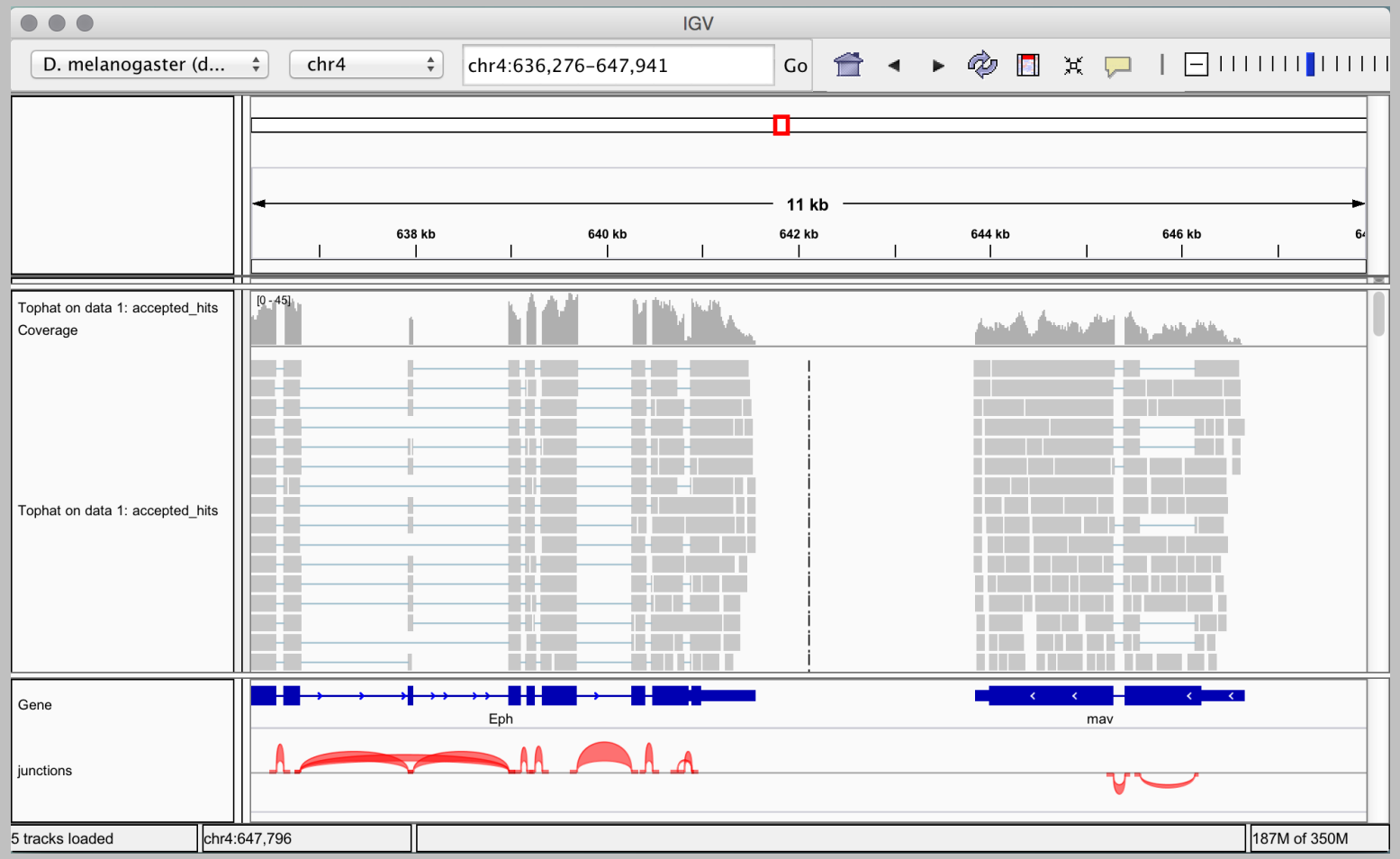
```
chr1 unknown  exon        113217048  113217252   .   +   .   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  exon        113217048  113217351   .   +   .   gene_id "MOV10";p_id "P5535";transcript_id "NM_020963"
chr1 unknown  exon        113217470  113217671   .   +   .   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  CDS         113217535  113217671   .   +   0   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  start_codon 113217535  113217537   .   +       gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
```

feature type

feature

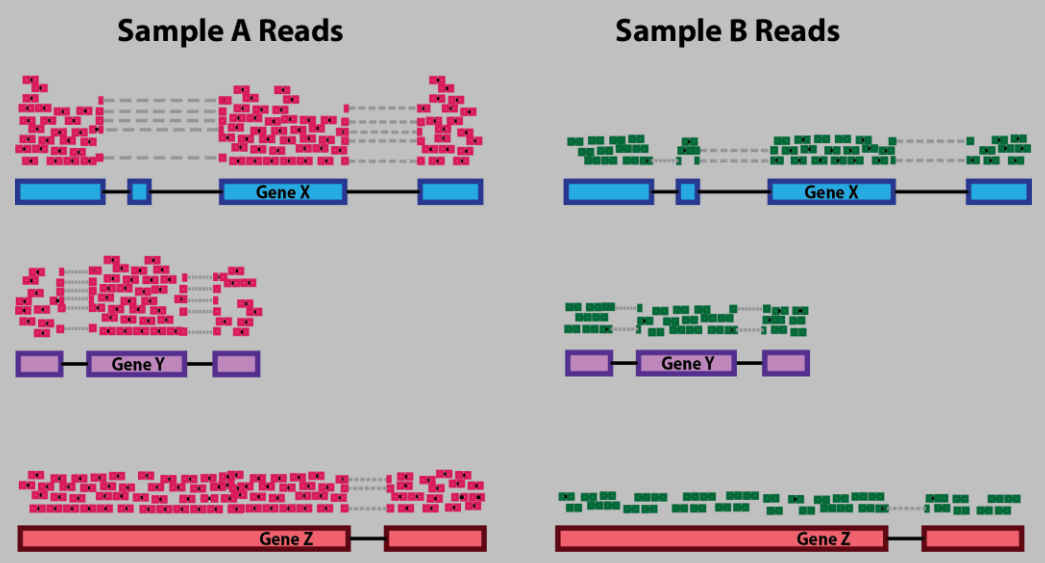
Visualisation

- Input: BAM file (alignment output, indexed) + genome sequences + genome annotation.



Counting issue: let's consider a gene A with 100 counts, a gene B with 100 counts as well but gene A is TWICE longer than gene B…
"Are these 2 genes expressed at the same level?"

Normalisation



2 biases to correct:

- library size or sequencing depth
- gene length

Gene name	Rep1 Count	Rep2 Count	Rep3 Count
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1 kb)	5	8	15
D (10 kb)	0	0	1

Normalisation with TPM

Transcript Per Million

Gene name	Rep1 Count	Rep2 Count	Rep3 Count
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1 kb)	5	8	15
D (10 kb)	0	0	1

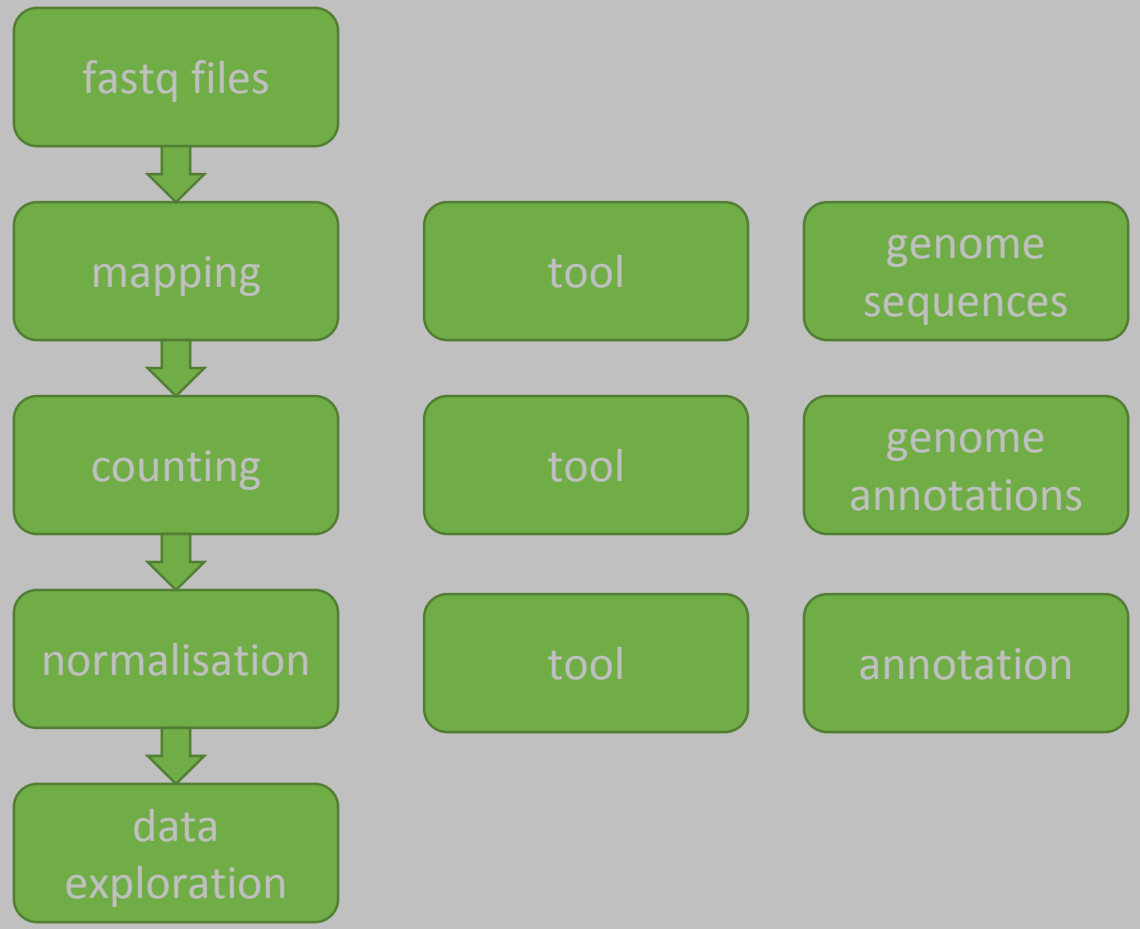
Step 1:
Normalisation in **gene length** (divide by gene length).

Gene name	Rep1 RPK	Rep2 RPK	Rep3 RPK
A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1 kb)	5	8	15
D (10 kb)	0	0	0.1
Total RPK:	15	20.25	45.1
Tens of RPK:	1.5	2.025	4.51

Step 2:
Normalisation in sequencing depth (divide by **scaling factor**).

Gene name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1 kb)	3.33	3.95	3.326
D (10 kb)	0	0	0.02

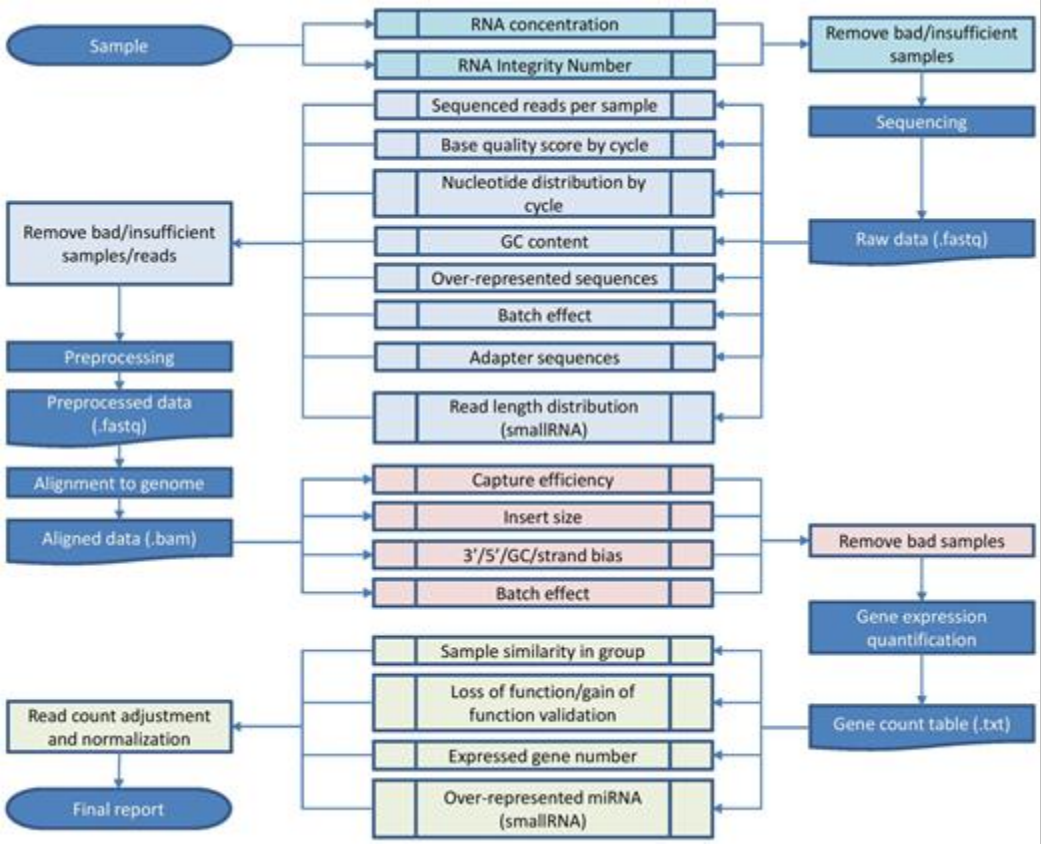
Wrap up


fastq files
mapping
counting
normalisation
data exploration
tool
tool
tool
genome sequences
genome annotations
annotation

Thank you for your attention

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Quality Controls



Third gen sequencing

Principle of nanopore and single-molecule real-time (SMRT) sequencing