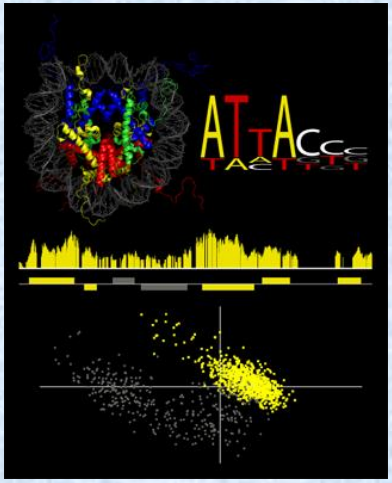
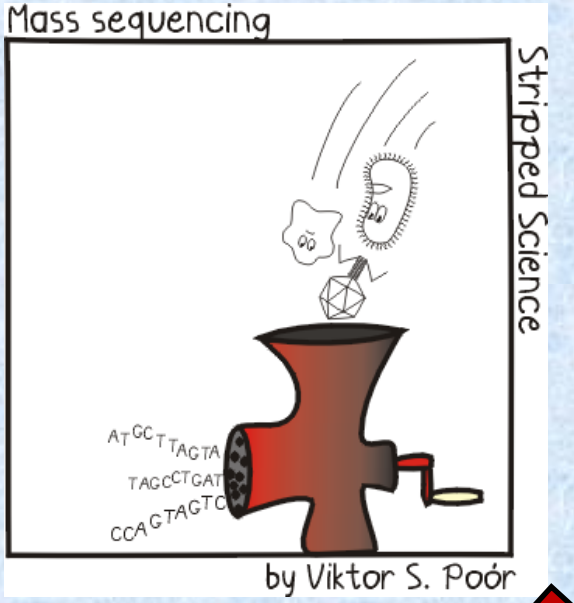
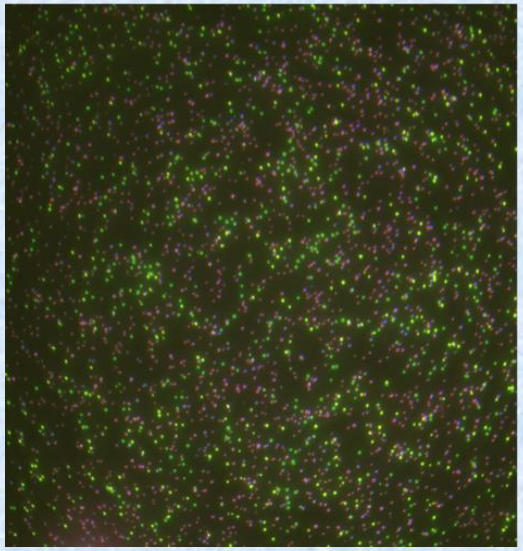
MAPPING OR ALIGNMENT OF SHORT READS

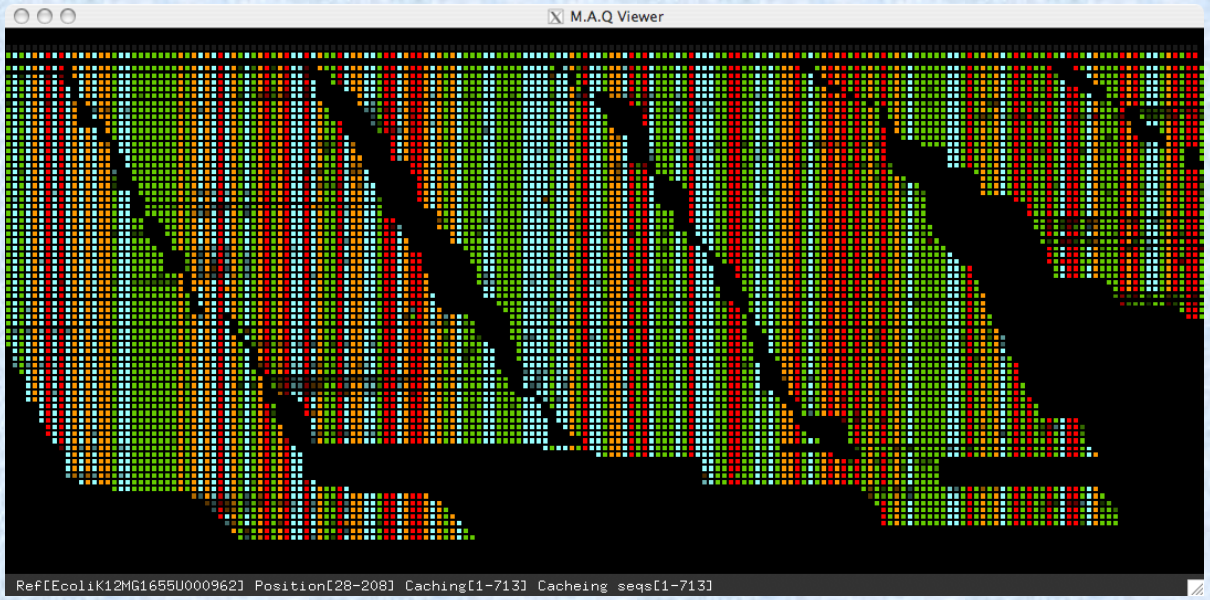


Ester Feldmesser

Introduction to Bioinformatics 2019



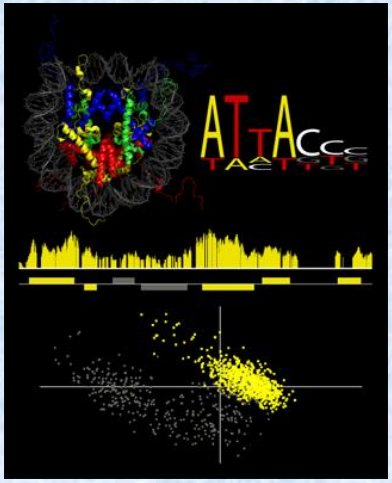




https://www.wiki.ed.ac.uk/download/attachments/36021339/maqscreenshot.png

Outline

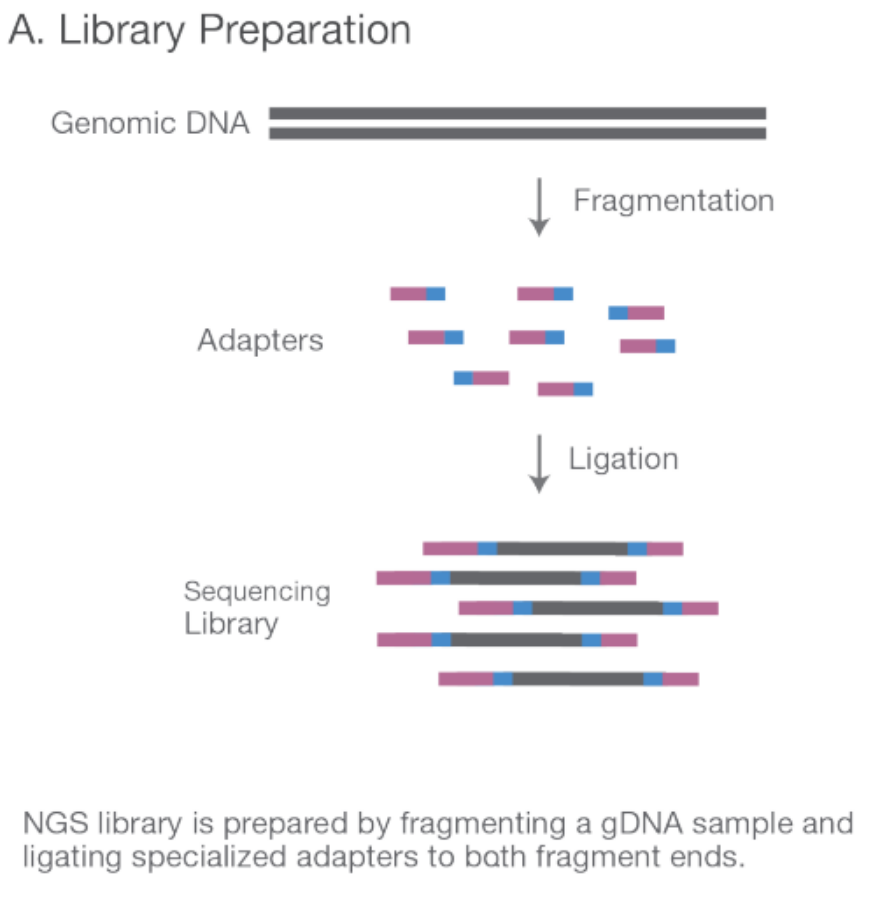
- Introduction
- Local and global alignments
- Blast
- Needs of NGS
- Short reads alignment
  - Burrows Wheeler transform
- Genome structure
- DNA alignment
- Transcriptome alignment

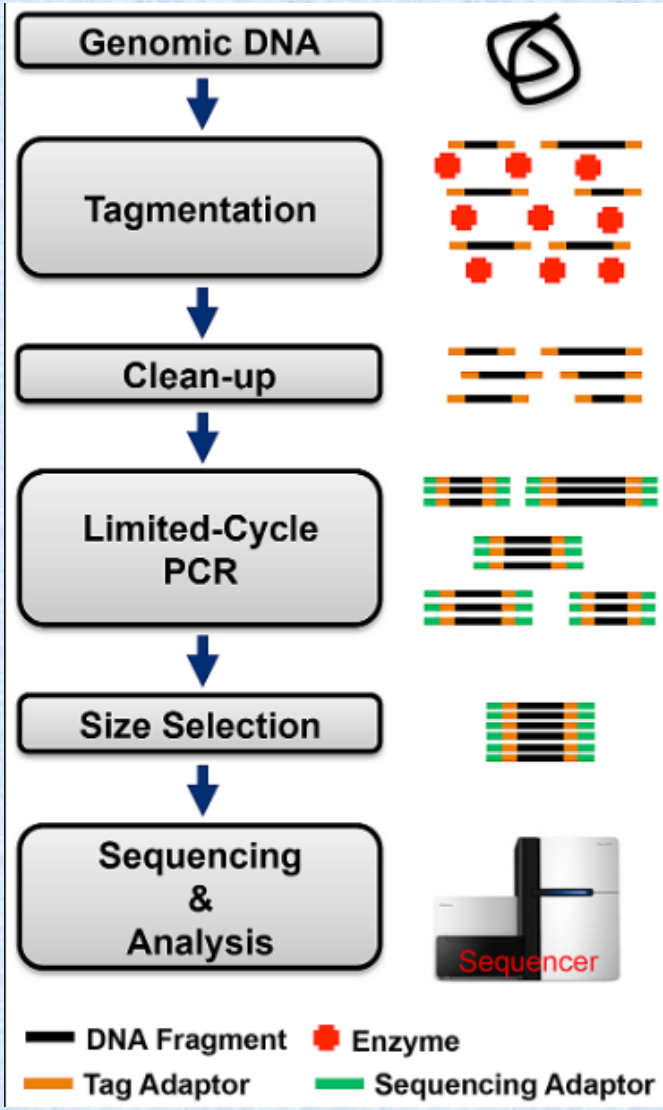


Illumina sequencing

- Prepare library
- Amplify
  - Why?
- Sequence
  - How?
- Analyze
  - What do we need?

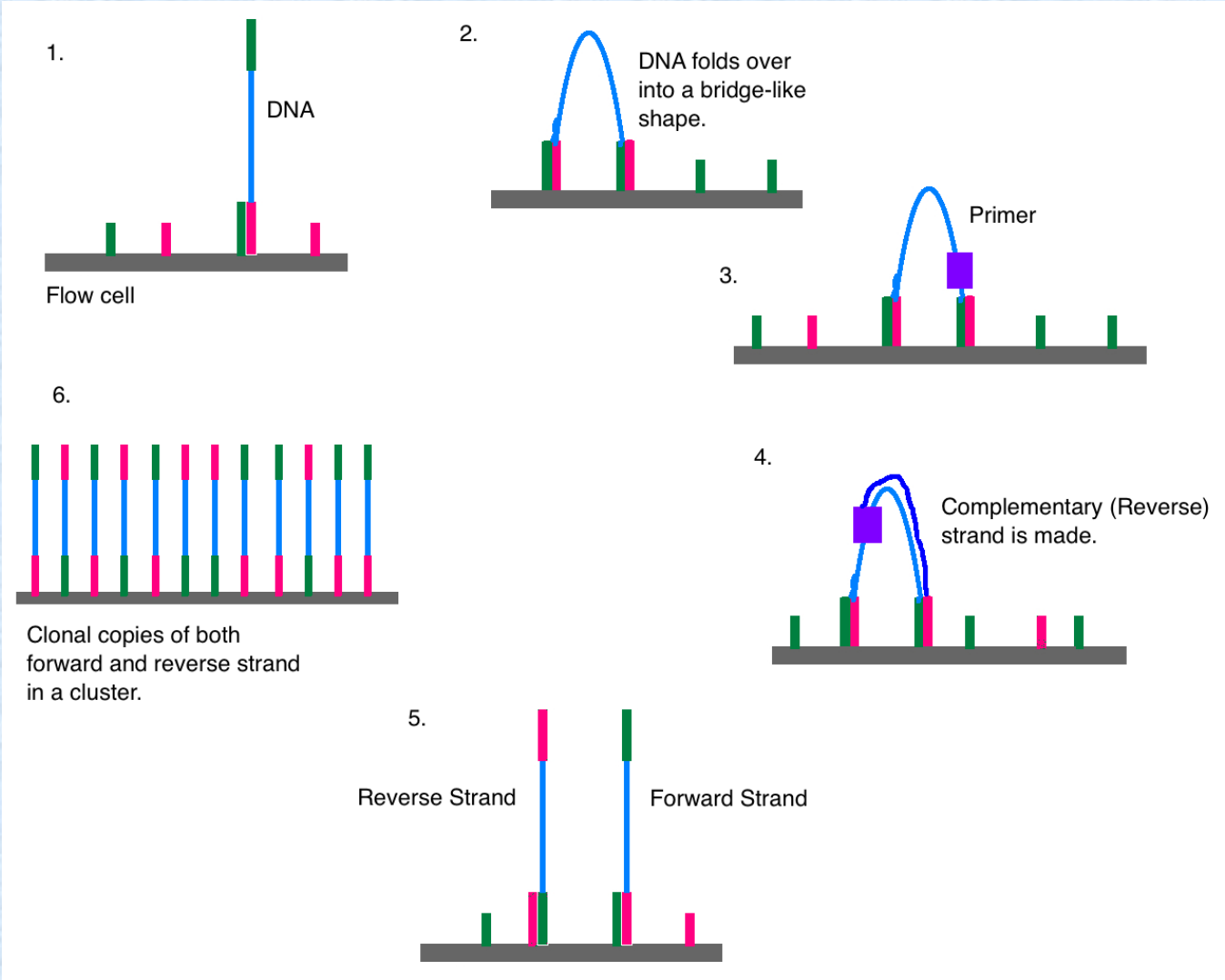
Library preparation





https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf,
https://www.researchgate.net/figure/NexteraR-protocol-for-preparing-DNA-libraries-compatible-with-Illumina-sequencers-Key_fig6_252321920

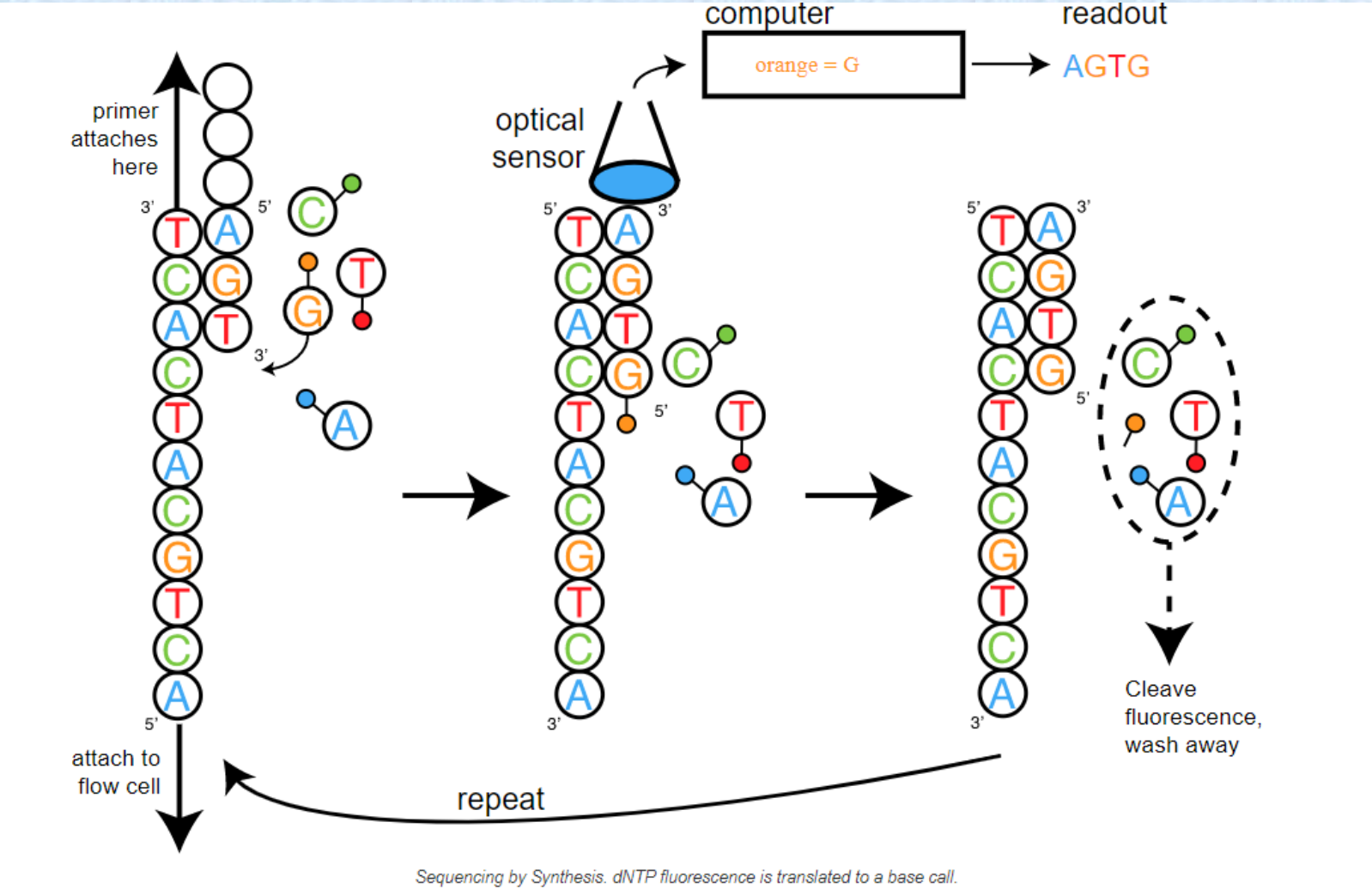
Amplify



1. The DNA attaches to the flow cell via complementary sequences.
2. The strand bends over and attaches to a second oligo forming a bridge.
3. A polymerase synthesizes the reverse strand. The two strands release and straighten.
4. Each forms a new bridge (bridge amplification).
5. The result is a cluster of DNA forward and reverse strands clones.

https://en.wikipedia.org/wiki/Illumina_dye_sequencing

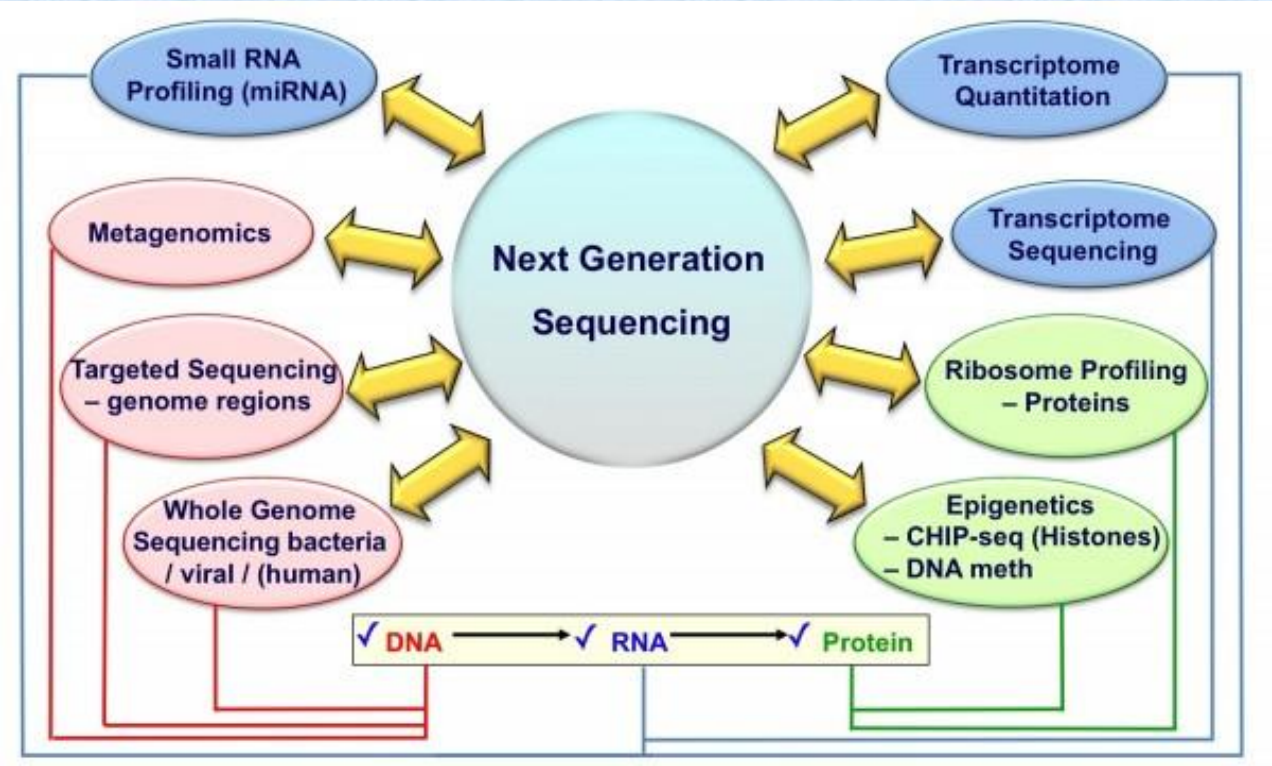
Sequence



Sequencing by Synthesis. dNTP fluorescence is translated to a base call.

https://binf.snipcademy.com/lessons/ngs-techniques/illumina-solexa

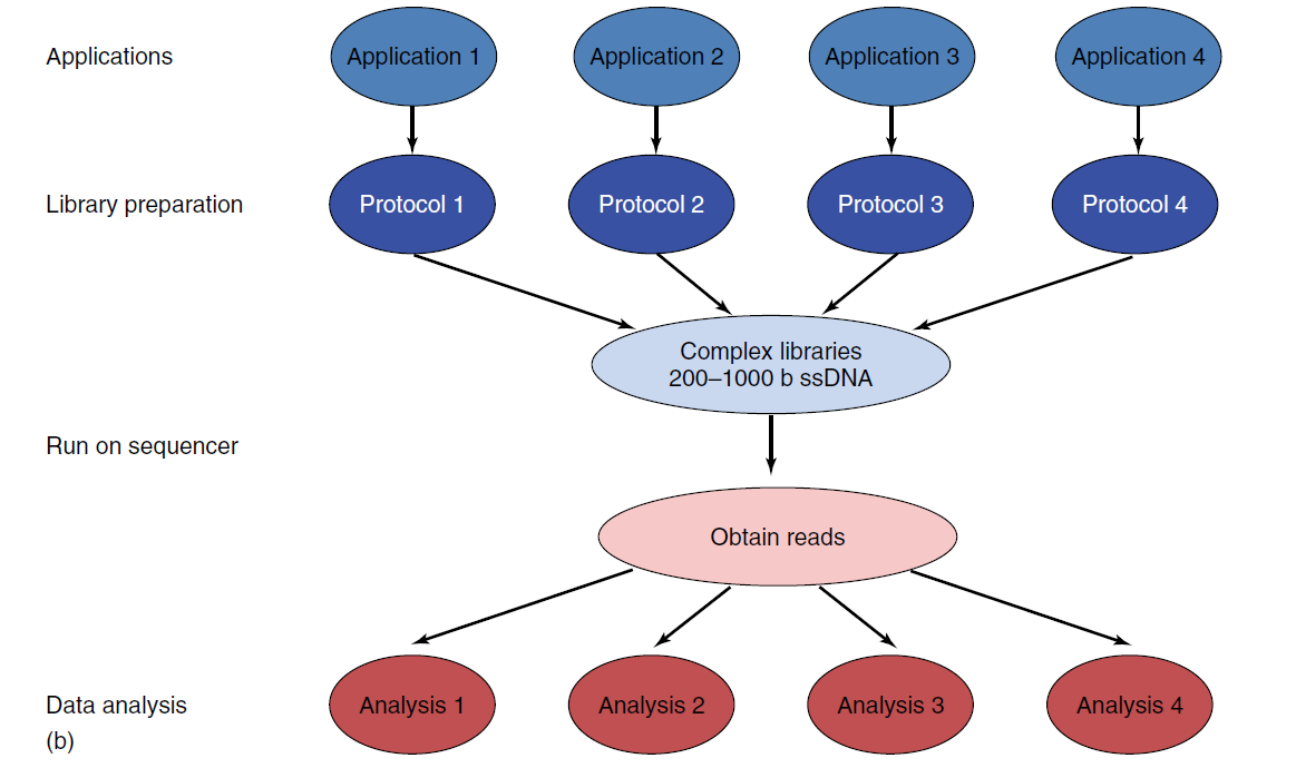
Analyze: applications



Applications of Next Generation Sequencing Technologies. Listed are some of the different applications possible on a single NGS sequencing machine which span across the central dogma of molecular biology ('DNA makes RNA makes protein')

https://www.europeanpharmaceuticalreview.com/article/10409/dna-sequencing-technologies-and-emerging-applications-in-drug-discovery/

Protocols and analyses



Each application is processed in a different way but all protocols result in short fragments ready to be sequenced. Once reads are obtained, downstream computational analyses are once again application specific.

Horn-Saban, S. and Amann-Zalcenstein, D. 2011. Frontiers in DNA Sequencing: the (R)Evolution of Sequencing Technologies. Encyclopedia of Analytical Chemistry.

First step in analysis

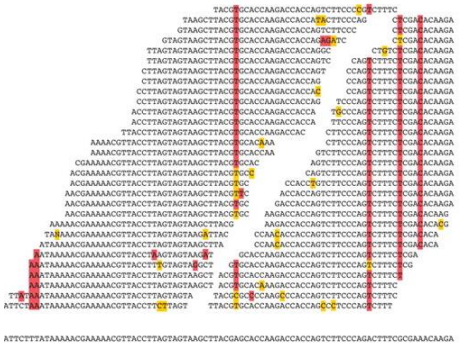
What is the first bioinformatics step to do after quality control?

- For DNA
- For RNA
- Model or non-model organism?

First step in analysis

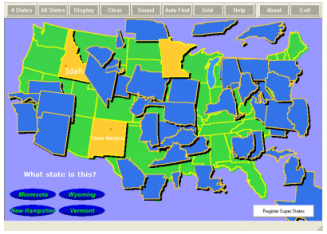
READS

Sequencing

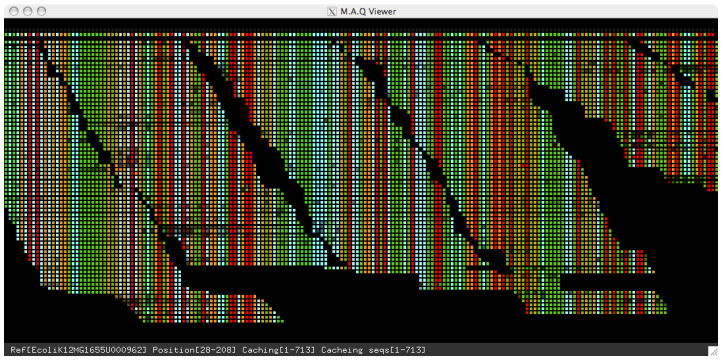


Assembly

Mapping
Alignment



Viewing



Sequence alignment

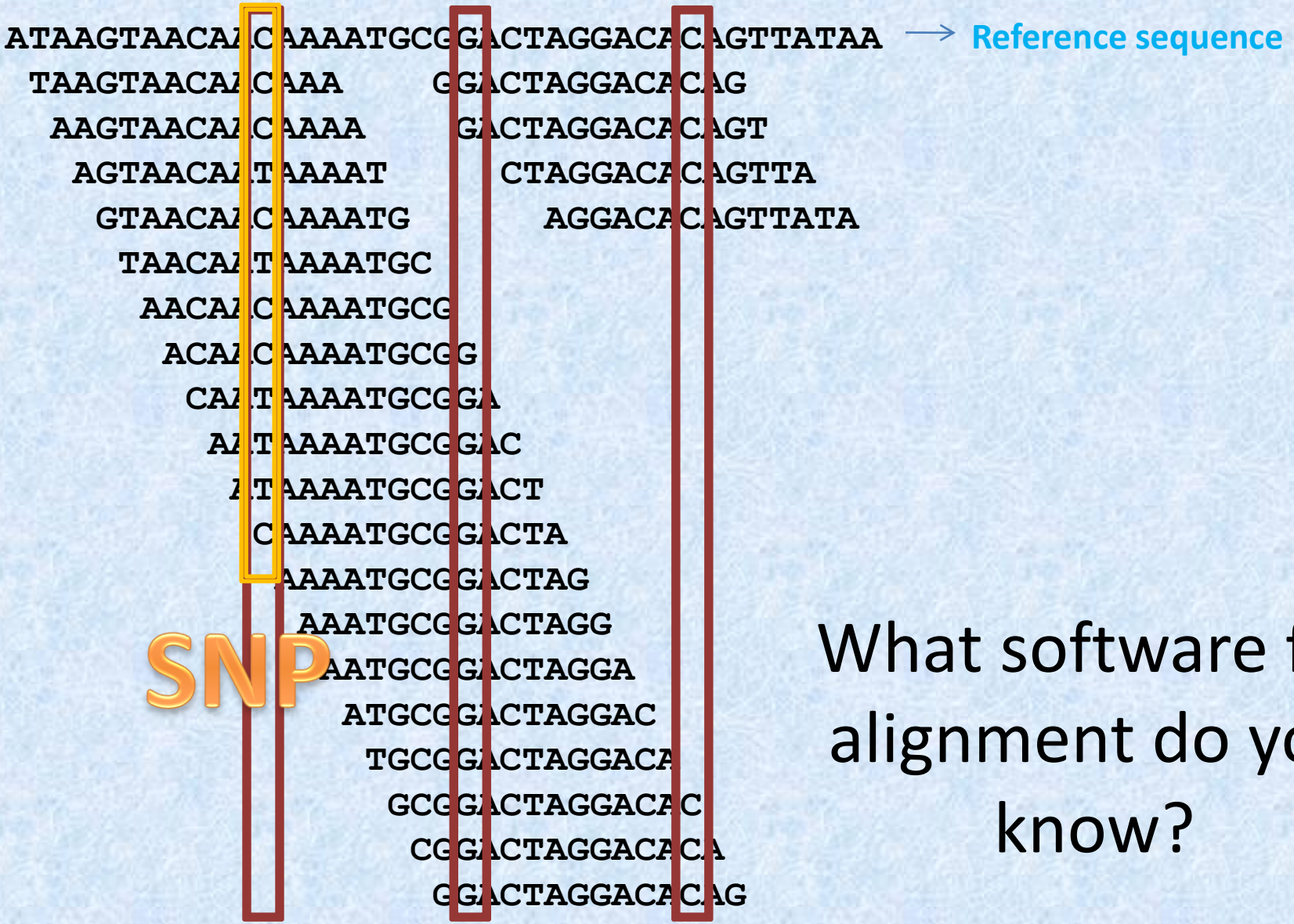
A **sequence alignment** is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix.
Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns.

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

A sequence alignment, produced by ClustalW, of two human zinc finger proteins

Alignment coverage



What software for alignment do you know?

DNA Coverage

Coverage (or **depth**) in DNA sequencing is the number of unique reads that include a given nucleotide in the reconstructed sequence

Depth is commonly a term used for genome or exome sequencing and means the number of reads covering each position.

But that is for RNA-seq totally pointless since the coverage pattern is so uneven due to differences in expression.

Which aligners do you know?

What type of alignment are they looking for?

Partial or for the full length of the sequence

Alignment type

- Global
  - the whole of each sequence is included, end to end

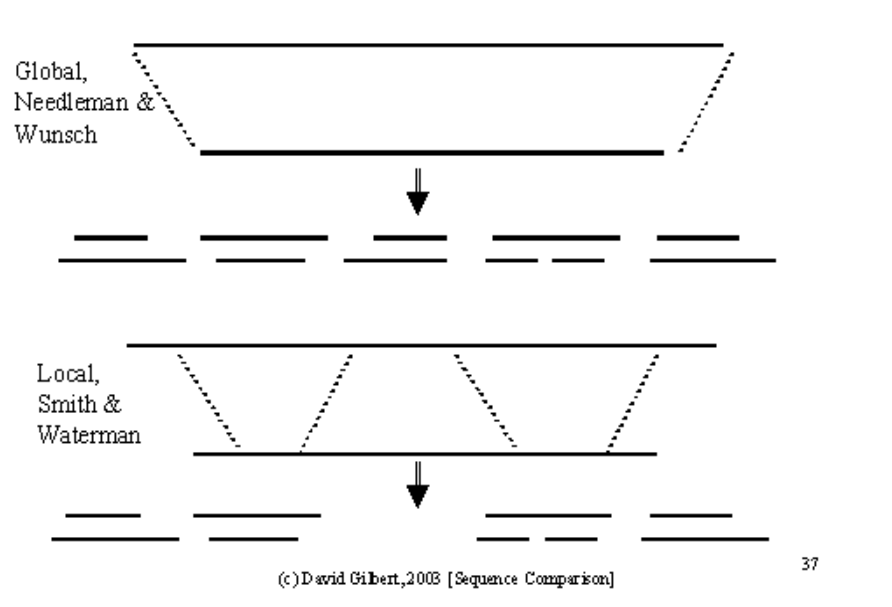
```
FTFTALILLAVAV
F--TAL-LLA-AV
```

- Local
  - only the best matching parts of each sequence

```
 FTALILL-AVAV
 FTAL-LL
```

Local and global alignments

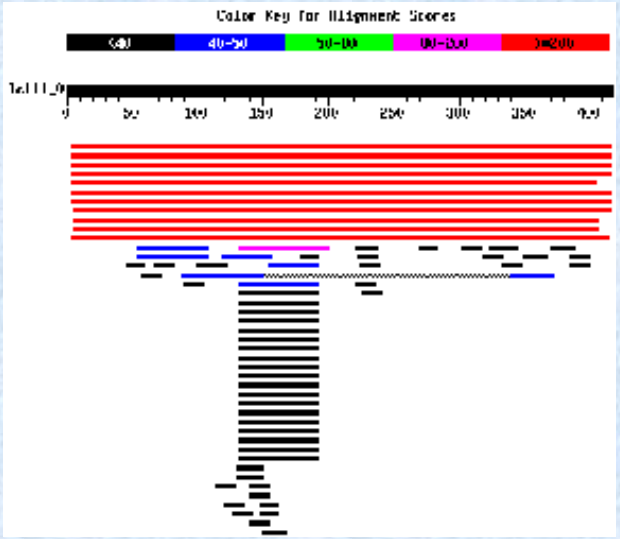
- **Global alignments** attempt to align every residue in every sequence
- Useful when the sequences in the query set are similar and of roughly equal size
- The FASTA software performs global alignments



- **Local alignments** looks for regions of similarity or similar sequence motifs within their larger sequence context
- Useful for dissimilar sequences
- BLAST performs local alignment

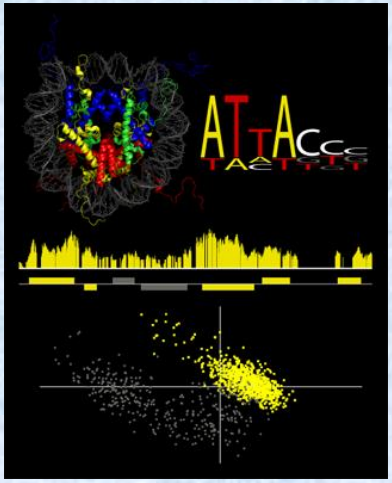
Blast (Basic Local Alignment Search Tool)

- Locates **short matches** (common words)
- Searches for high scoring sequence alignments between these matches in the database using a heuristic approach that approximates the Smith-Waterman algorithm
- BLAST algorithm uses a heuristic approach that is less accurate than the Smith-Waterman but over 50 times faster.
- In typical usage, the query sequence is much smaller than the database (query ~1000 nucleotides and the database ~ several billion nucleotides.



Outline

- Introduction
- Local and global alignments
- Blast
- Needs of NGS
- Short reads alignment
  - Burrows Wheeler transform
- Genome structure
- DNA alignment
- Transcriptome alignment



New needs for NGS

- Many of the next-generation sequencing projects begin with a known, or so-called ‘reference’, genome. This process is known as aligning or ‘mapping’ the read to the reference.
- There are many programs that perform both spliced and unspliced alignment for the older Sanger-style capillary reads.
- These programs neither scale up to the much greater volumes of data produced by short read sequencers nor scale down to the short read lengths.
- Aligning the reads from ChIP-Seq or RNA-Seq experiments can take hundreds or thousands of central processing unit (CPU) hours using conventional software tools such as BLAST or BLAT.

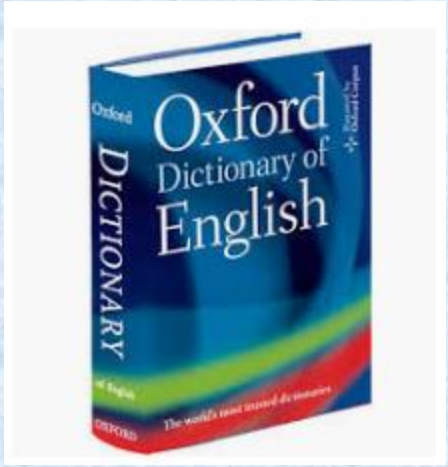
Nature Biotechnology **27**, 455 - 457 (2009)

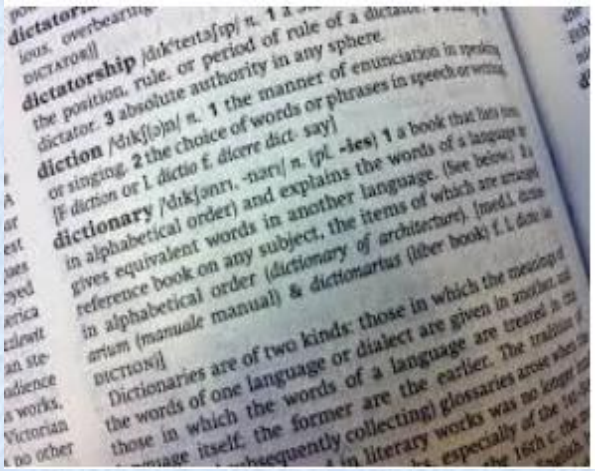
The problem

- 50 -100 million reads (sequences)
- Read length between 50 and 150 base pairs
- We would like to know if these reads are in the human genome
- The human genome is ~ 3 x 10^9 base pairs

A dictionary

The genome is our dictionary and the reads are the words we are looking for





designdictatorshi
pdictiondictionar
ydisagreeparent

parentdisagreedi
ctatorshipdiction
dictionarydesign

Searching …

- Searching every single read by text comparison, will take years
- Even using blat in a parallel computer will take months
- What do you suggest to improve this?
- Do we need to allow for mismatches?

Comparison between generations of alignment programs

Blast	New programs
Alignment of a protein to large databases	Alignment of DNA reads within the genome sequenced
Alignment of one or a few sequences	Alignment of millions of reads
Usually the sequence is long	Sequence reads are short
To find some degree of homology	To find almost perfect matches (up to two mismatches, short indels)
Permits gaps	Take into account splice junctions
Algorithms not fully optimized	Algorithms optimized for speed and memory

Short read aligners: Differences

Alignment tools differ in

- speed
- suitability for use on computer clusters
- memory requirements
- accuracy
    - Is a good match always found?
    - What is the maximum number of allowed mismatches?
- ease of use

http://marray.economia.unimi.it/2009/material/lectures/L5_HTS_Overview.pdf

Short read aligners: Differences

Alignment tools also differ in whether they can

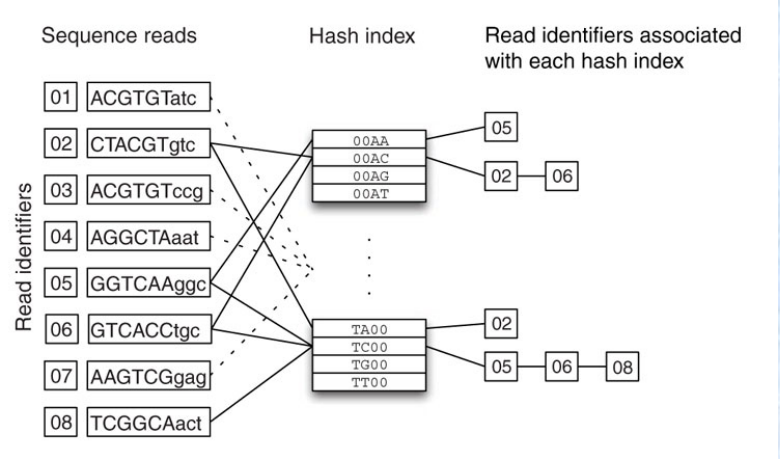
- make use of base-call quality scores
- estimate alignment quality
- detect small indels
- report multiple matches
- work with longer than normal reads
- match in colour space (for SOLiD systems)
- align data from methylation experiments
- deal with splice junctions

http://marray.economia.unimi.it/2009/material/lectures/L5_HTS_Overview.pdf

Short read alignment: Algorithms

Short-read aligners use indexing to speed up mapping

- use spaced seed indexing
  - hash seed words from the reference
  - hash seed words from the reads



- use the Burrows-Wheeler transform (BWT)

BWT seems to be the winning idea (very fast, sufficiently accurate), and is used by the newest tools (Bowtie, SOAPv2, BWA, Hisat).

http://marray.economia.unimi.it/2009/material/lectures/L5_HTS_Overview.pdf

Burrows–Wheeler transform

- **Burrows–Wheeler transform** (**BWT**, also called **block-sorting compression**), is an algorithm used in data compression techniques.
- It was invented by Michael Burrows and David Wheeler in 1994 while working at DEC Systems Research Center in Palo Alto, California.
- When a character string is transformed by the BWT, none of its characters change value.
- The transformation permutes the order of the characters.
- If the original string had several substrings that occurred often, then the transformed string will have several places where a single character is repeated multiple times in a row.
- This is useful for compression.

http://en.wikipedia.org/wiki/Burrows–Wheeler_transform

Example

The transform is done by sorting all rotations of the text, then taking the last column. For example, the text "^BANANA@" is transformed into "BNN^AA@A" through these steps (the red @ character indicates the 'EOF' pointer):

Transformation			
Input	All Rotations	Sort the Rows	Output
^BANANA@	^BANANA@ @^BANANA A@^BANAN NA@^BANA ANA@^BAN NANA@^BA ANANA@^B BANANA@^	ANANA@^**B** ANA@^BA**N** A@^BANA**N** BANANA@**^** NANA@^B**A** NA@^BAN**A** ^BANANA**@** @^BANAN**A**	BNN^AA@A

BWT not only generates a more easily encoded string
It is reversible, allowing the original data to be re-generated

http://en.wikipedia.org/wiki/Burrows–Wheeler_transform

Getting back the original string

Inverse Transformation	
Input	
BNN^AA@A	
Add 1	Sort 1
B	A
N	A
N	A
^	B
A	N
A	N
@	^
A	@

- The last column sorted by characters return the first column
- The first and last columns together give you all *pairs* of successive characters , where pairs are taken cyclically so that the last and first character form a pair.
- Sorting the list of pairs gives the first and second columns.

http://en.wikipedia.org/wiki/Burrows–Wheeler_transform

Getting back the original string 2

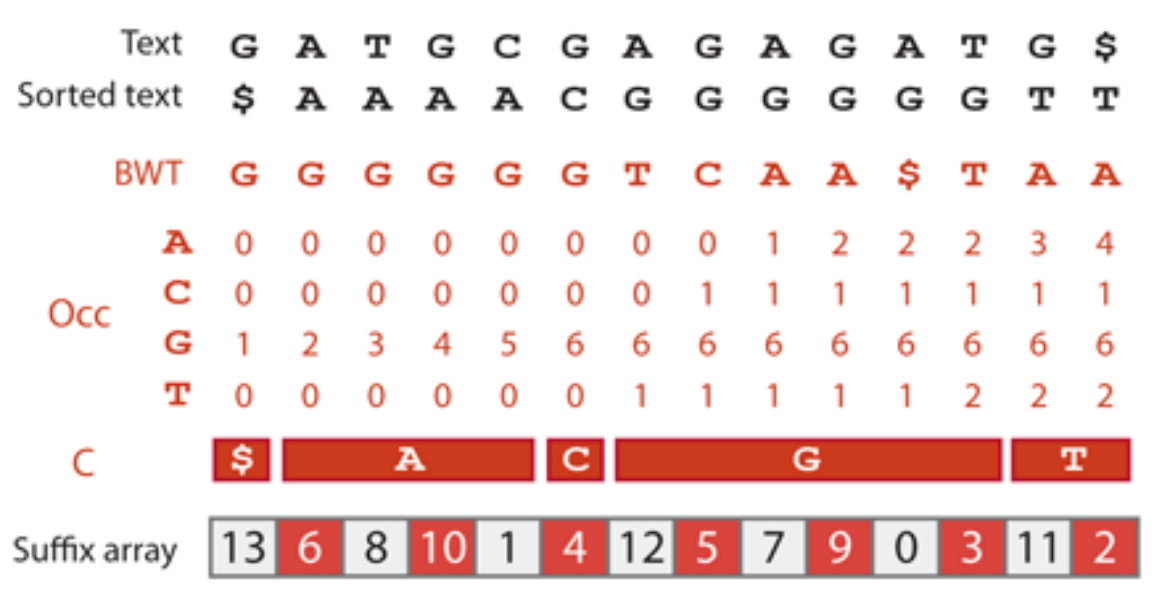
Inverse Transformation							
Input							
BNN^AA@A							
Add 5	Sort 5	Add 6	Sort 6	Add 7	Sort 7	Add 8	Sort 8
BANAN	ANANA	BANANA	ANANA@	BANANA@	ANANA@^	BANANA@^	ANANA@^B
NANA@	ANA@^	NANA@^	ANA@^B	NANA@^B	ANA@^BA	NANA@^BA	ANA@^BAN
NA@^B	A@^BA	NA@^BA	A@^BAN	NA@^BAN	A@^BANA	NA@^BANA	A@^BANAN
^BANA	BANAN	^BANAN	BANANA	^BANANA	BANANA@	^BANANA@	BANANA@^
ANANA	NANA@	ANANA@	NANA@^	ANANA@^	NANA@^B	ANANA@^B	NANA@^BA
ANA@^	NA@^B	ANA@^B	NA@^BA	ANA@^BA	NA@^BAN	ANA@^BAN	NA@^BANA
@^BAN	^BANA	@^BANA	^BANAN	@^BANAN	^BANANA	@^BANANA	^BANANA@
A@^BA	@^BAN	A@^BAN	@^BANA	A@^BANA	@^BANAN	A@^BANAN	@^BANANA
Output							
^BANANA@							

- Continuing in this manner, entire list can be reconstructed.
- The row with the "end of file" character at the end is the original text.

http://en.wikipedia.org/wiki/Burrows–Wheeler_transform

Index the whole genome

Additional requirements:

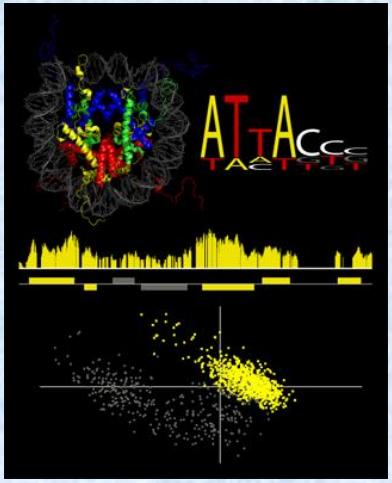


1. Sorted text
2. Suffix array, the position of the base in the original sequence (starting the count from 0)
3. Occurrence (Occ), the cumulative numbers of occurrences of each base in the BW transformed sequence
4. C, stores the position of the first occurrence of each character in the sorted text

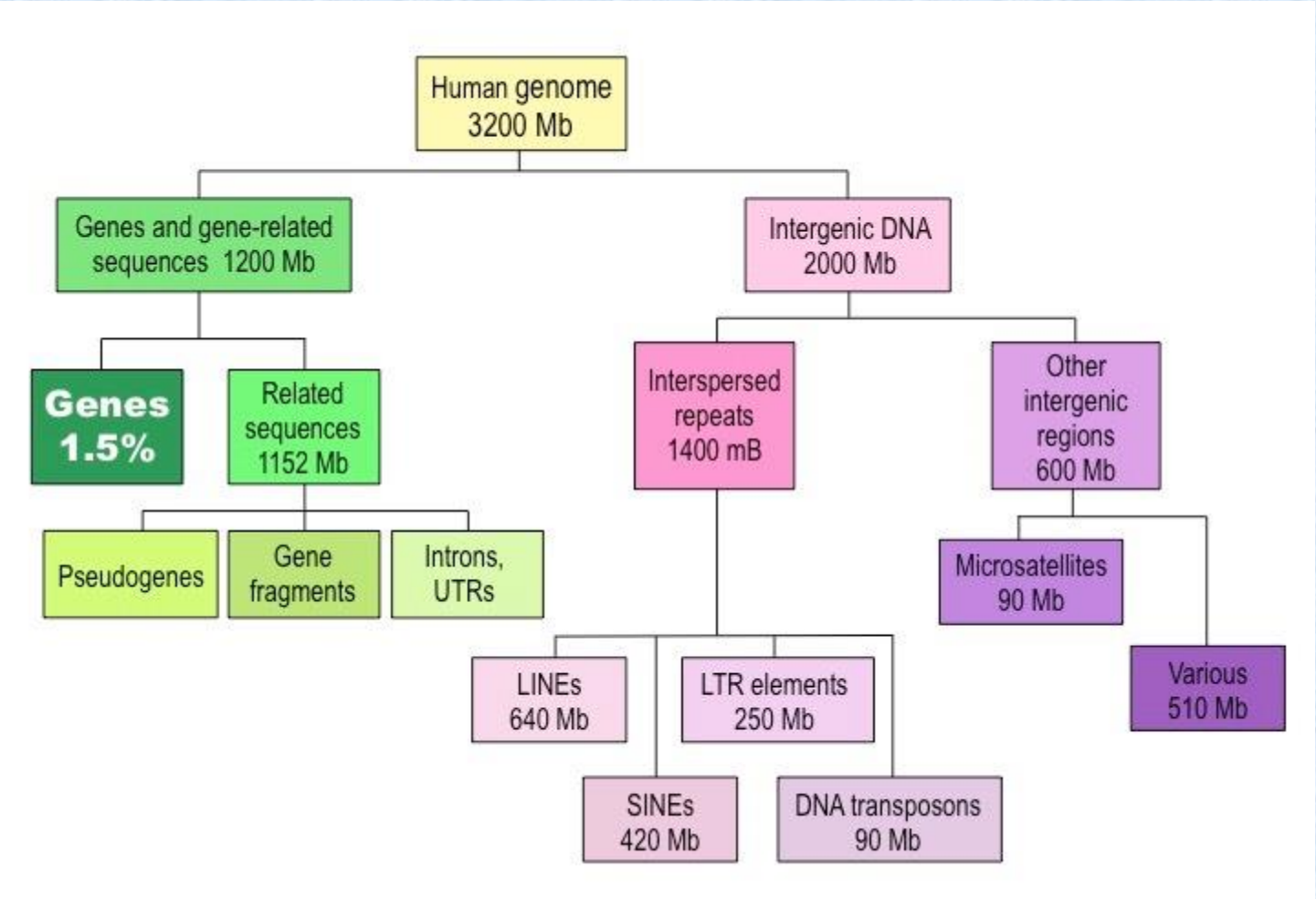
http://blog.thegrandlocus.com/2016/07/a-tutorial-on-burrows-wheeler-indexing-methods

Outline

- Introduction
- Local and global alignments
- Blast
- Needs of NGS
- Short reads alignment
  - Burrows Wheeler transform
- Genome structure
- DNA alignment
- Transcriptome alignment

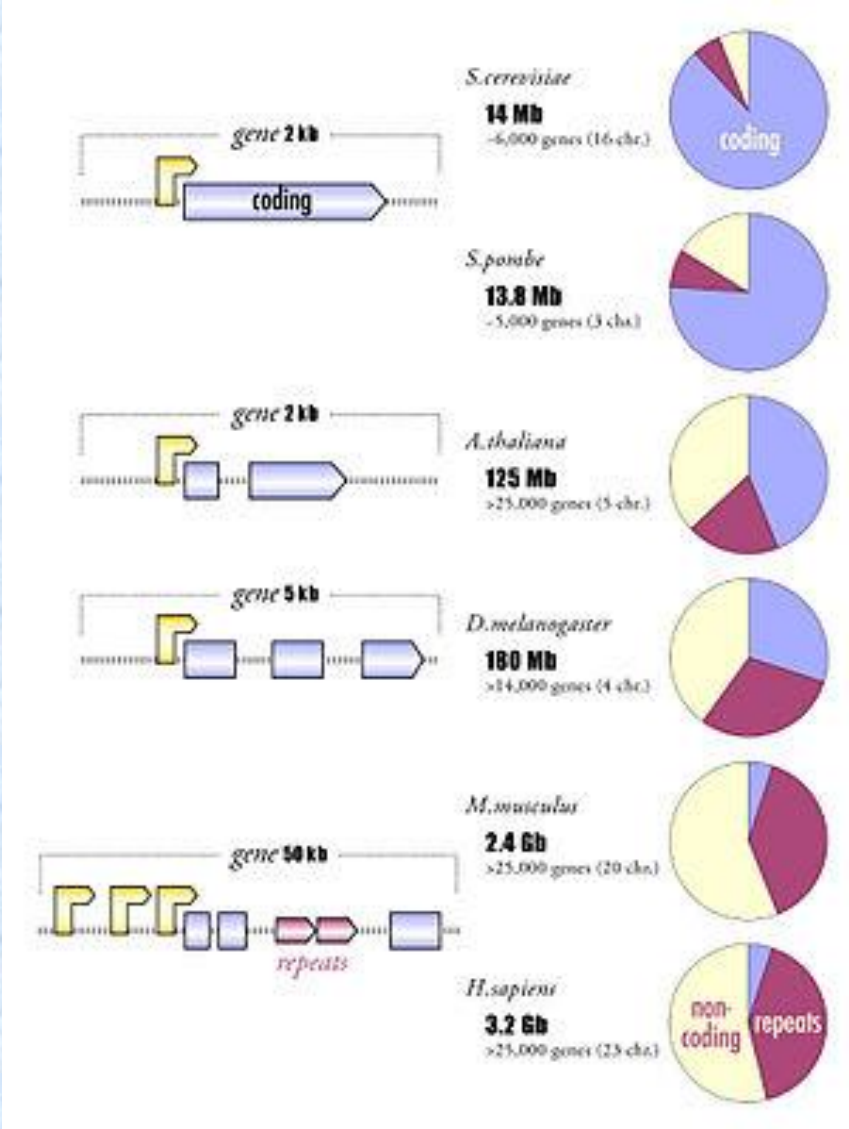


Genome Structure



https://ib.bioninja.com.au/standard-level/topic-3-genetics/31-genes/genome-structure.html

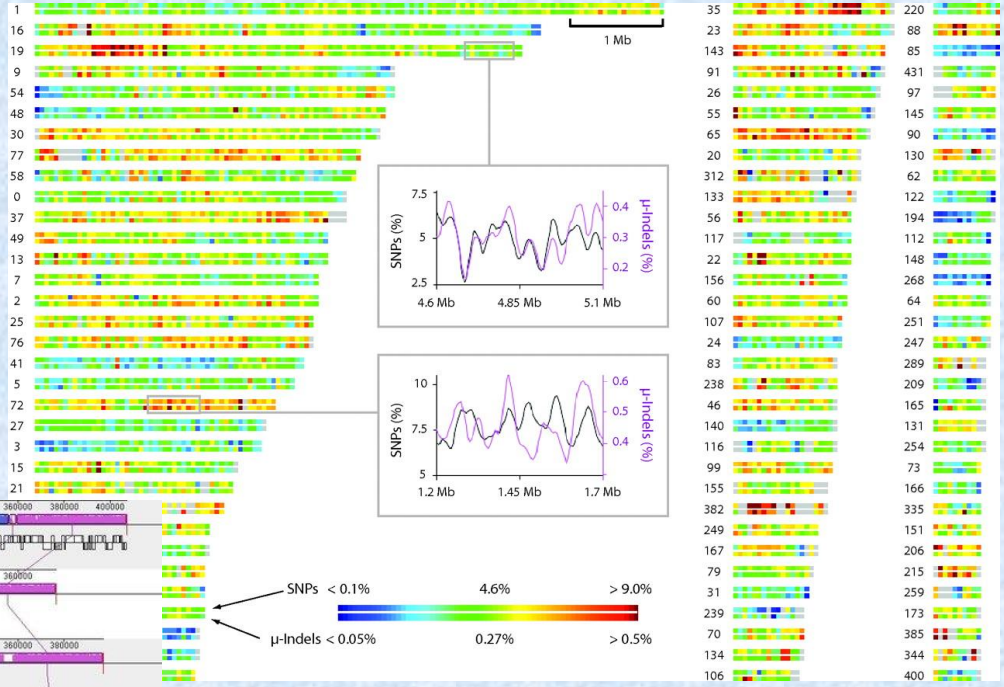
Composition of six major model genomes

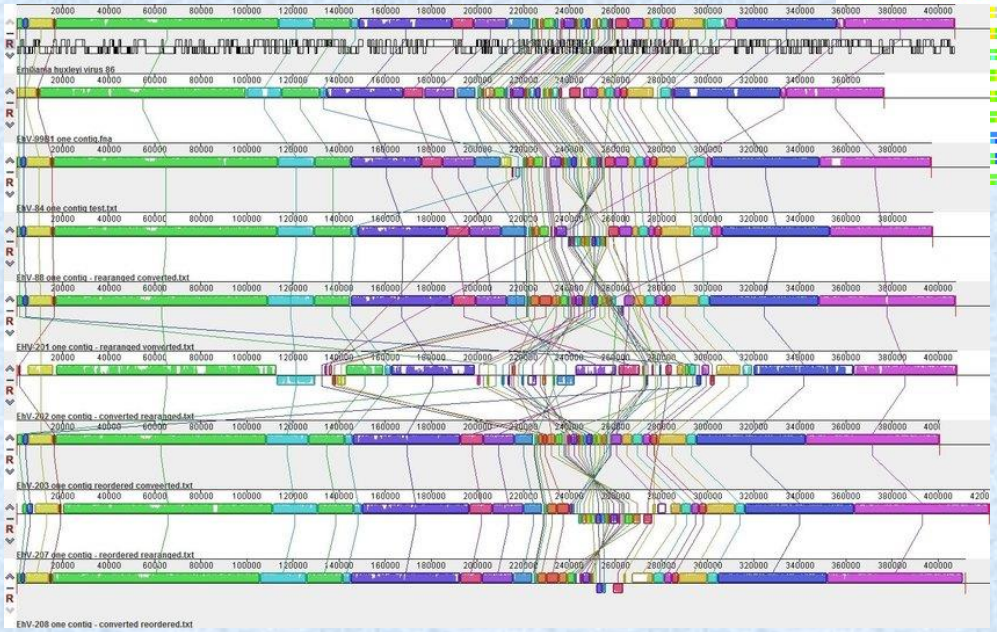


The increase in genome size correlates with the vast expansion of noncoding (i.e., intronic, intergenic, and interspersed repeat sequences) and repeat DNA(e.g., satellite, LINEs, Short interspersed nuclear element (SINEs), DNA (Alu sequence), in red) sequences in more complex muiticellular organisms.

https://en.wikipedia.org/wiki/Genomic_organization

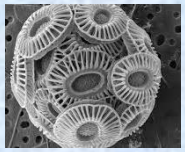
Genomic variations

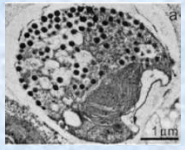




Ciona savignyi

EhV



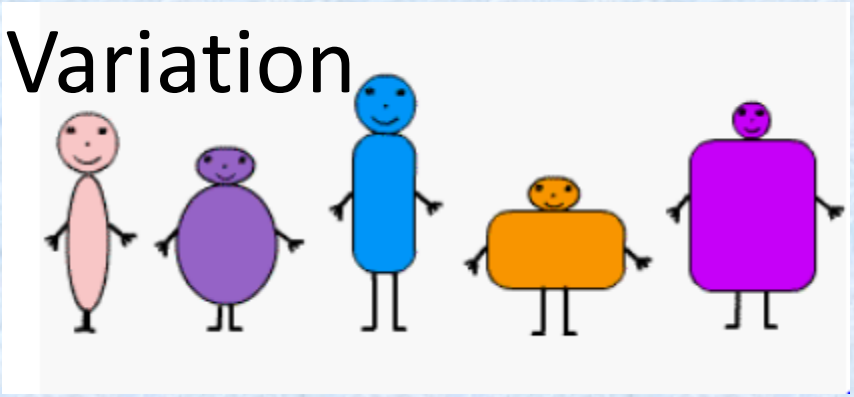


http://eprints.nottingham.ac.uk/13776/, https://www.pnas.org/content/104/13/5698

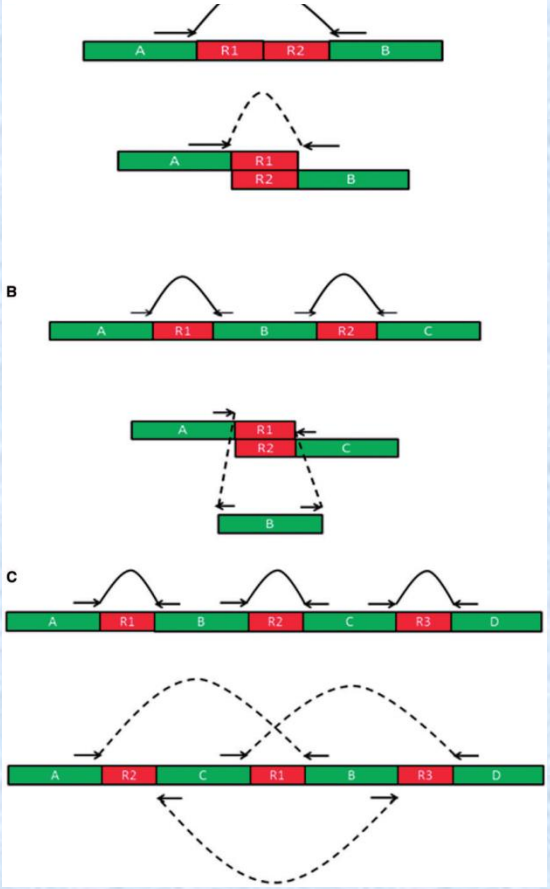
Issues mapping reads onto a genome

Size

Variation

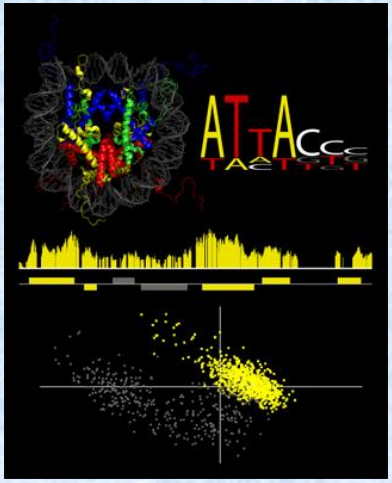


Errrors in the genome



Outline

- Introduction
- Local and global alignments
- Blast
- Needs of NGS
- Short reads alignment
  - Burrows Wheeler transform
- Genome structure
- DNA alignment
- Transcriptome alignment



Bowtie algorithm (DNA) without indexing

The places to look for the read

```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

Bowtie algorithm

The places to look for the read

```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

Bowtie algorithm

The places to look for the read

```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

Bowtie algorithm

The places to look for the read

```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

Bowtie algorithm

The places to look for the read

```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

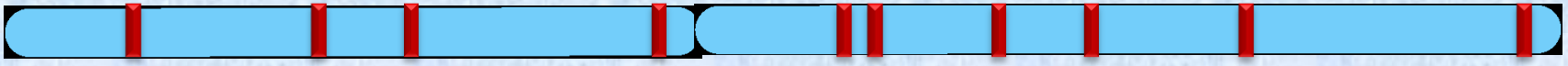
Bowtie algorithm

The places to look for the read

```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

Bowtie algorithm

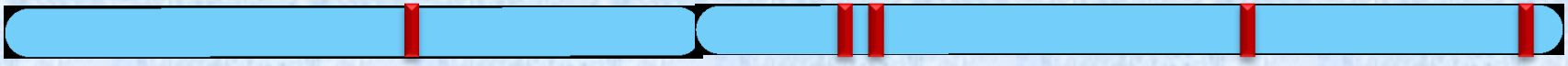


The places to look for the read

```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

Bowtie algorithm

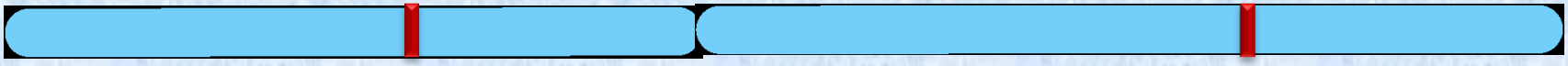


The places to look for the read

```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

Bowtie algorithm



The places to look for the read

```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

Bowtie algorithm

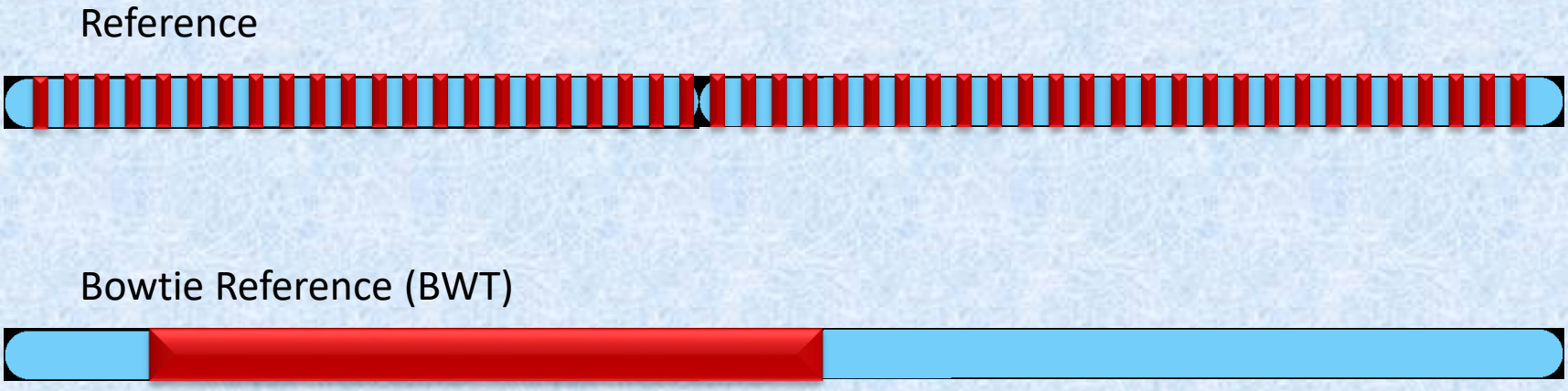
Reference

Bowtie Reference (BWT)

```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

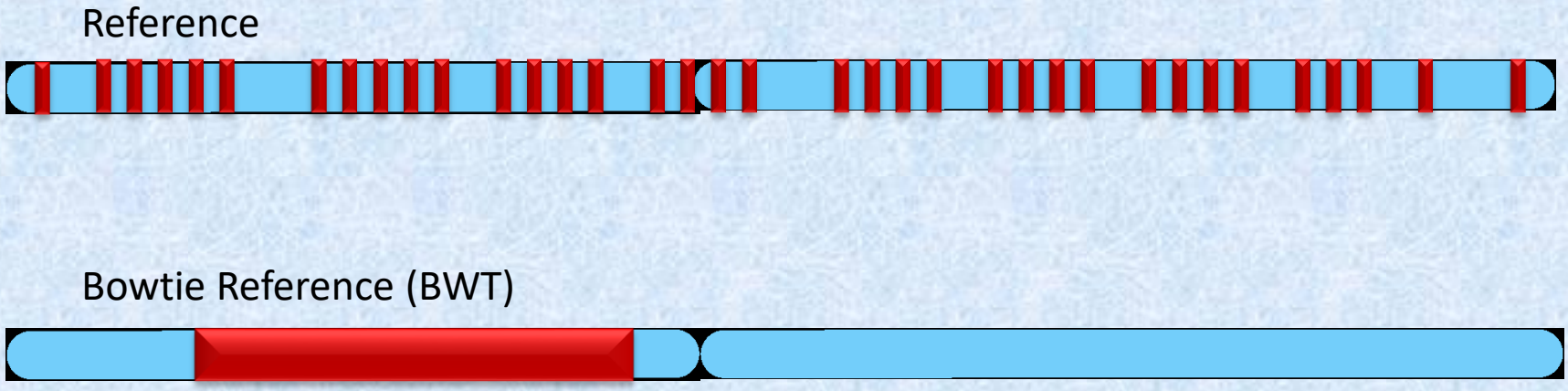
Bowtie algorithm with indexing




```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

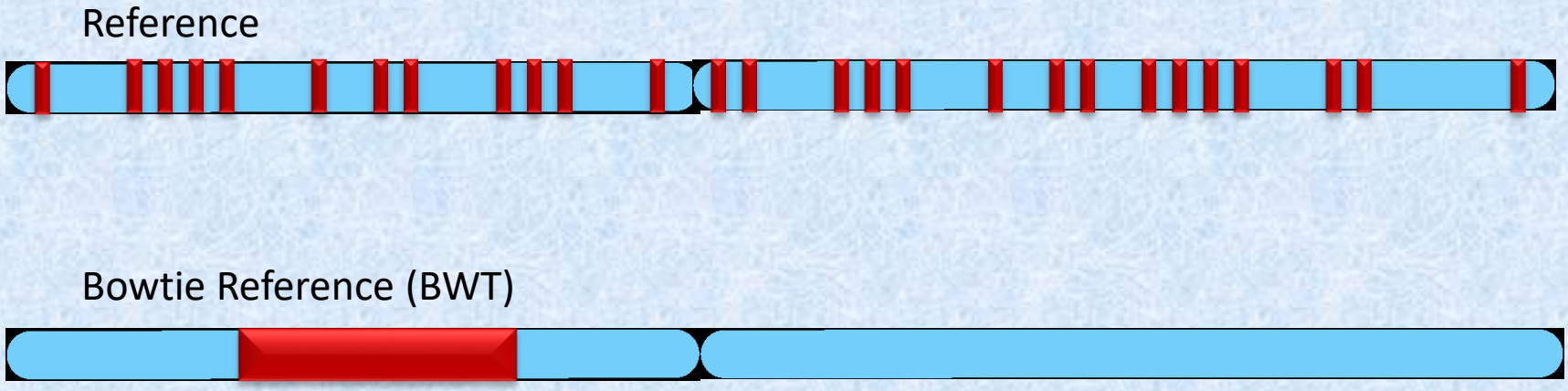
Bowtie algorithm




```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

Bowtie algorithm

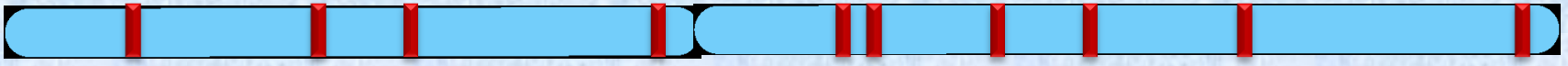



```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

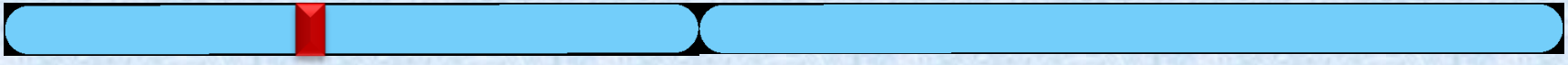
http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

Bowtie algorithm

Reference



Bowtie Reference (BWT)

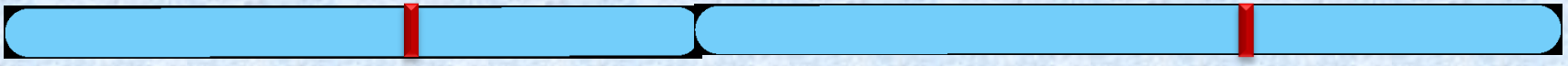


```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

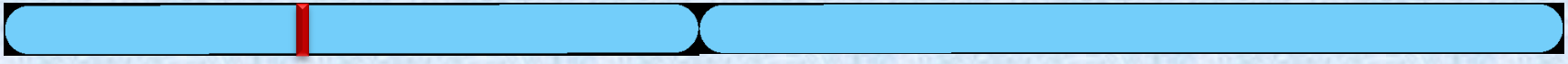
http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

Bowtie algorithm

Reference



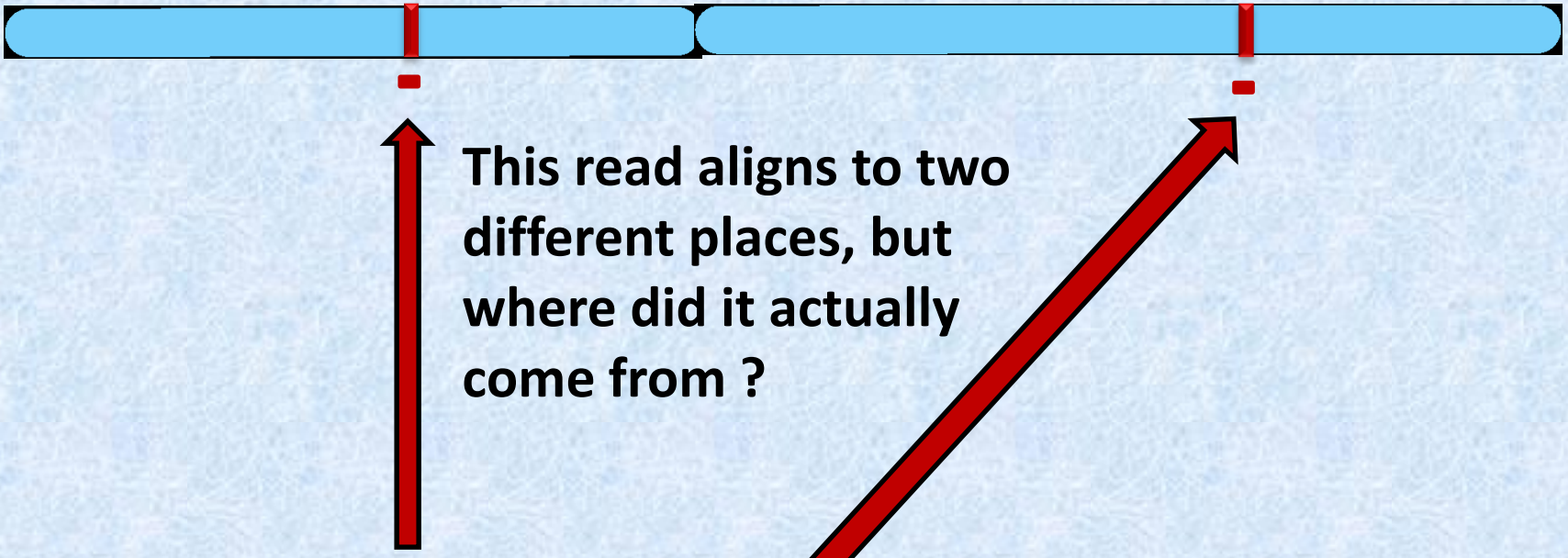
Bowtie Reference (BWT)



```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

Picking the right alignment



This read aligns to two different places, but where did it actually come from ?

```
@HWI-EAS412_4:1:1:1376:380
AATGATACGGCGACCACCGAGATCTA
```

http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

Popular DNA aligners

For DNA
Bowtie2
BWA
BWA-MEM

Why do we need different aligners for DNA and RNA?

We need to allow for gaps because of the exon – intron gene structure

Would you prefer to have local or global alignments?

Depending on the application and read length

Popular DNA aligners

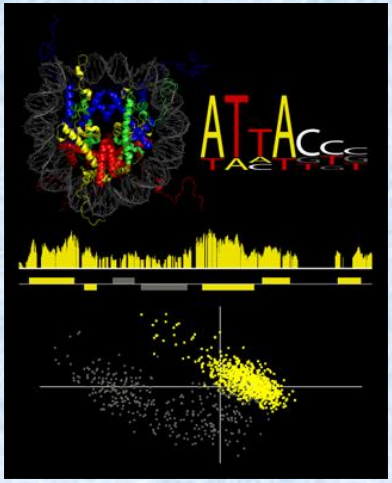
For DNA
Bowtie2
BWA
BWA-MEM

DNA Aligners

- They can use global or local alignments
- BWA-MEM uses both global and local alignments
- Bowtie2 and both BWA allow indels
- BWA-MEM:
  - Fitted for aligning sequence reads or long query sequences against a large reference genome
  - It automatically chooses between local and end-to-end alignments, supports paired-end reads and performs chimeric alignment
  - It is permissive to long gaps up to tens of bp for 100bp reads, or up to several hundred bp (tunable) for contig alignment

Outline

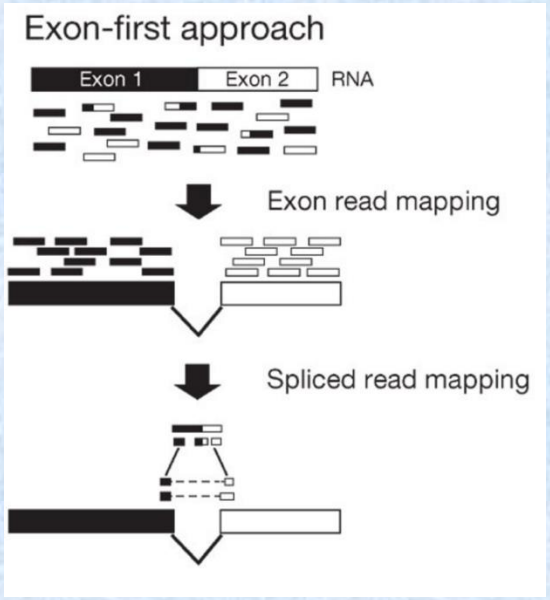
- Introduction
- Local and global alignments
- Blast
- Needs of NGS
- Short reads alignment
  - Burrows Wheeler transform
- Genome structure
- DNA alignment
- Transcriptome alignment



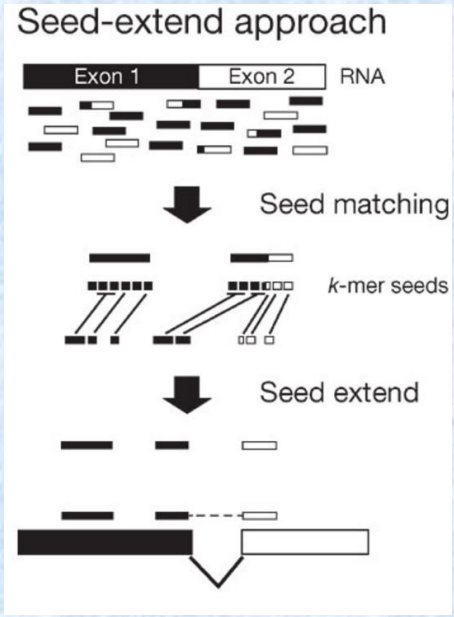
RNA alignment

- Align the reads to the transcriptome
- Align the reads to the genome
- When to use each of them?
- Local or global alignments

Transcriptome alignment



- Map full, unspliced reads (exonic reads).
- Remaining reads are divided into smaller pieces and mapped to the genome.
- An extension process extends mapped pieces to find candidate splice sites to support a spliced alignment.



- Store a map of all small words (*k*-mers) in an efficient lookup data structure.
- Each read is divided into *k*-mers, which are mapped to the genome via the lookup structure.
- Mapped *k*-mers are extended into larger alignments, which may include gaps flanked by splice sites.

http://www.nature.com/nmeth/journal/v8/n6/fig_tab/nmeth.1613_F1.html

Tophat

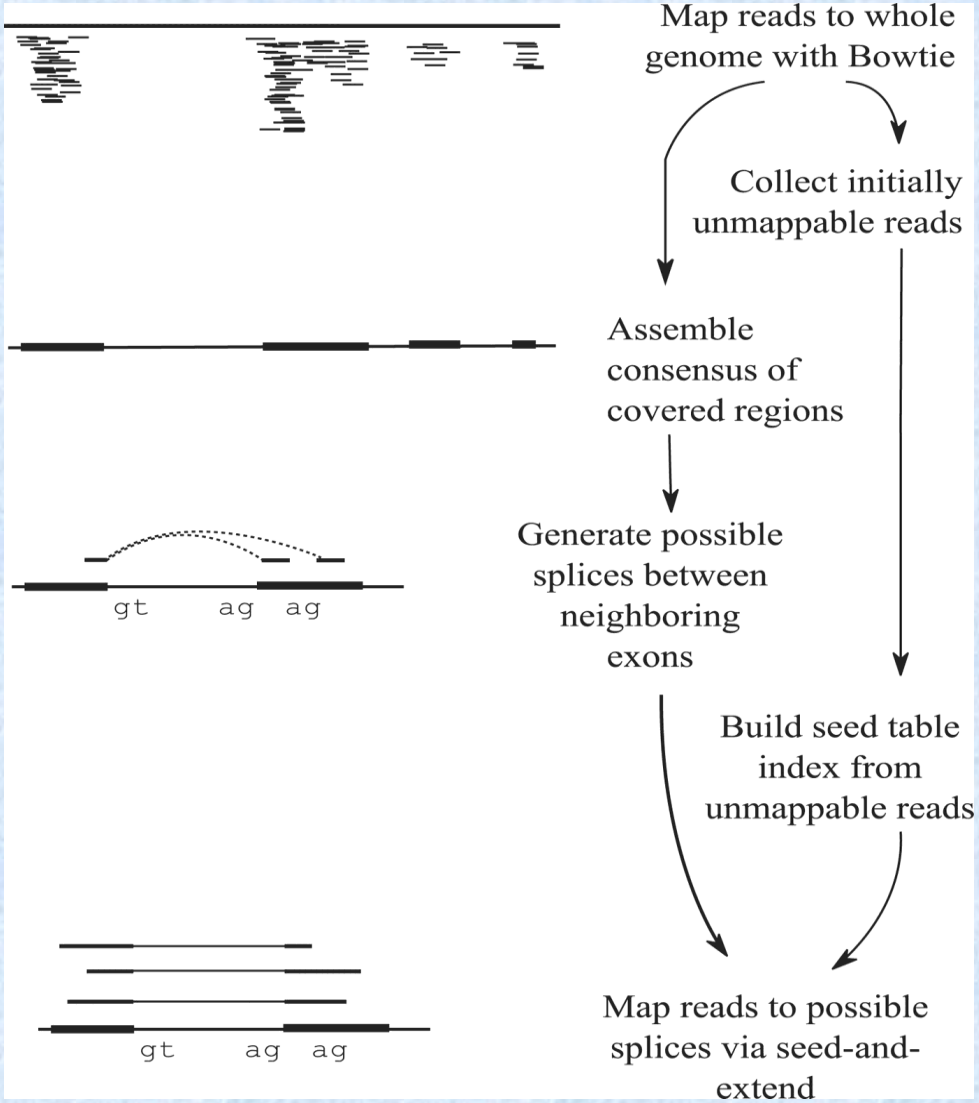
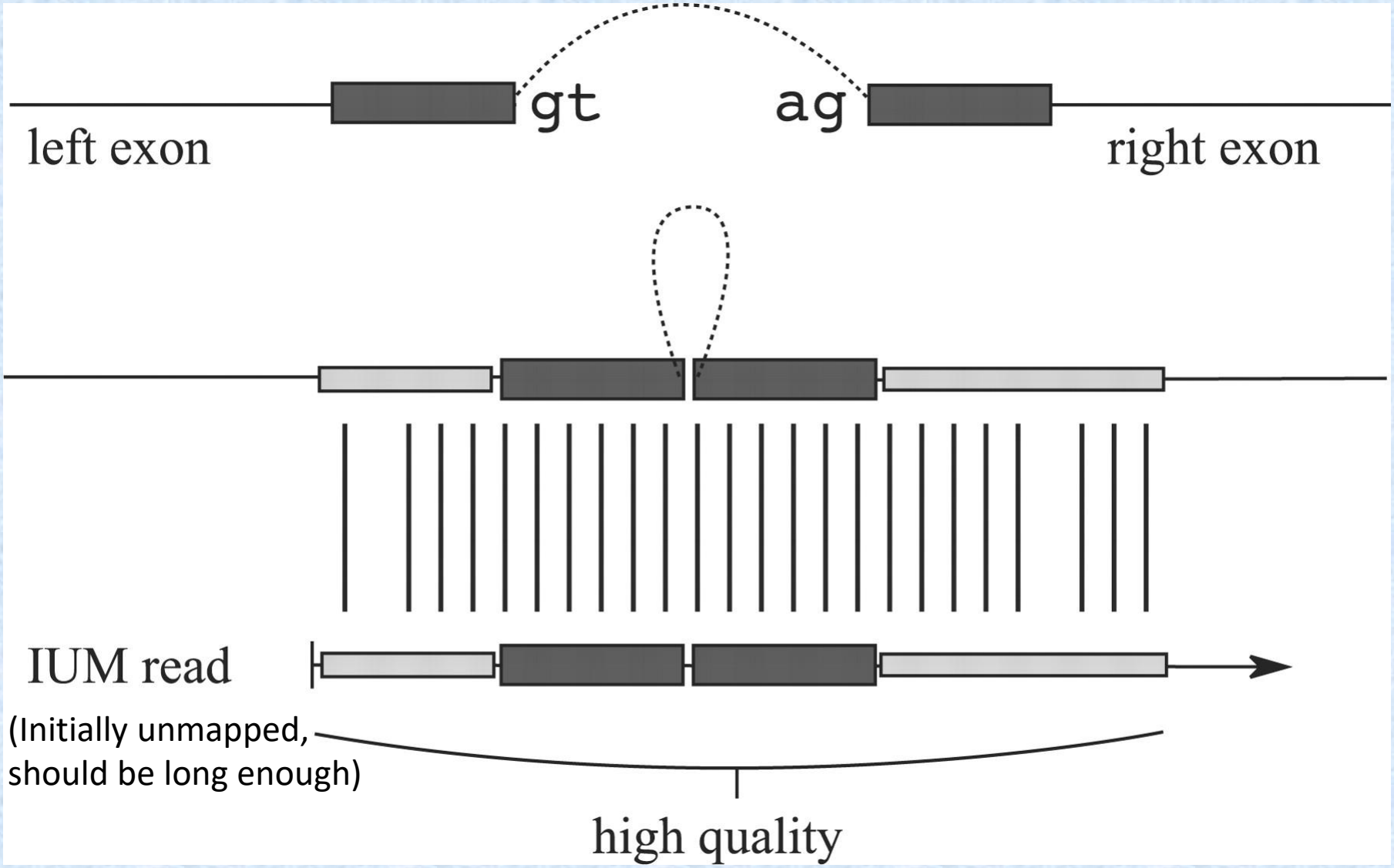


Fig. 1. The TopHat pipeline. RNA-Seq reads are mapped against the whole reference genome, and those reads that do not ...

Bioinformatics, Volume 25, Issue 9, 1 May 2009, Pages 1105–1111, https://doi.org/10.1093/bioinformatics/btp120

The seed and extend alignment



Star

- Does not use BWT
- STAR search is implemented through uncompressed suffix arrays (index)
- Requires huge amount of memory
- Very fast and accurate

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3530905/

Star

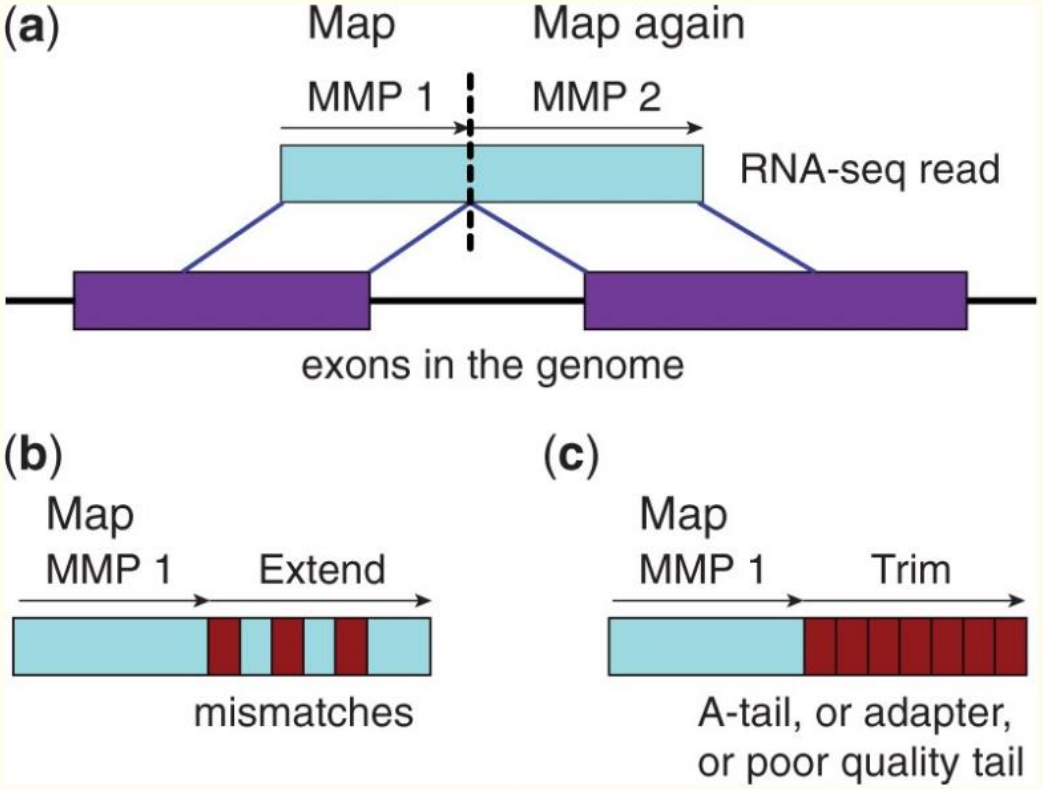


Fig. 1.

Schematic representation of the Maximum Mappable Prefix search in the STAR algorithm for detecting (**a**) splice junctions, (**b**) mismatches and (**c**) tails

- Seed search step
  Find Maximal Mappable Prefix (*MMP*)

  The splice junctions are detected in a single alignment pass

- Clustering/ stitching/scoring step

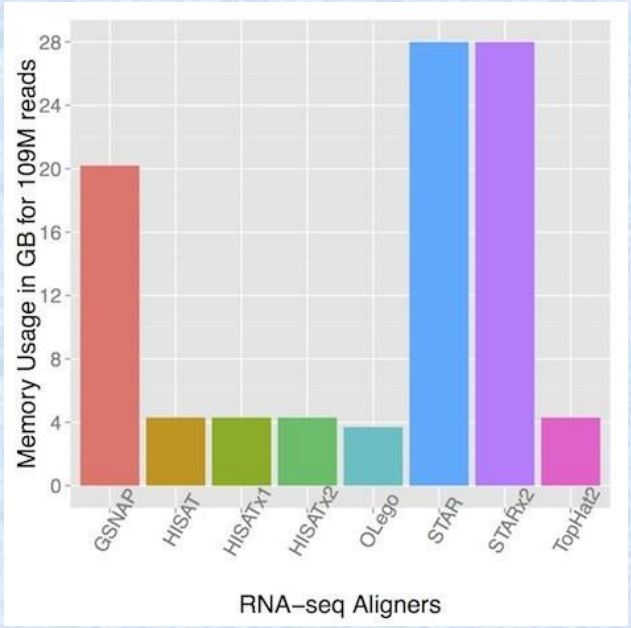
  The seeds are clustered together by proximity

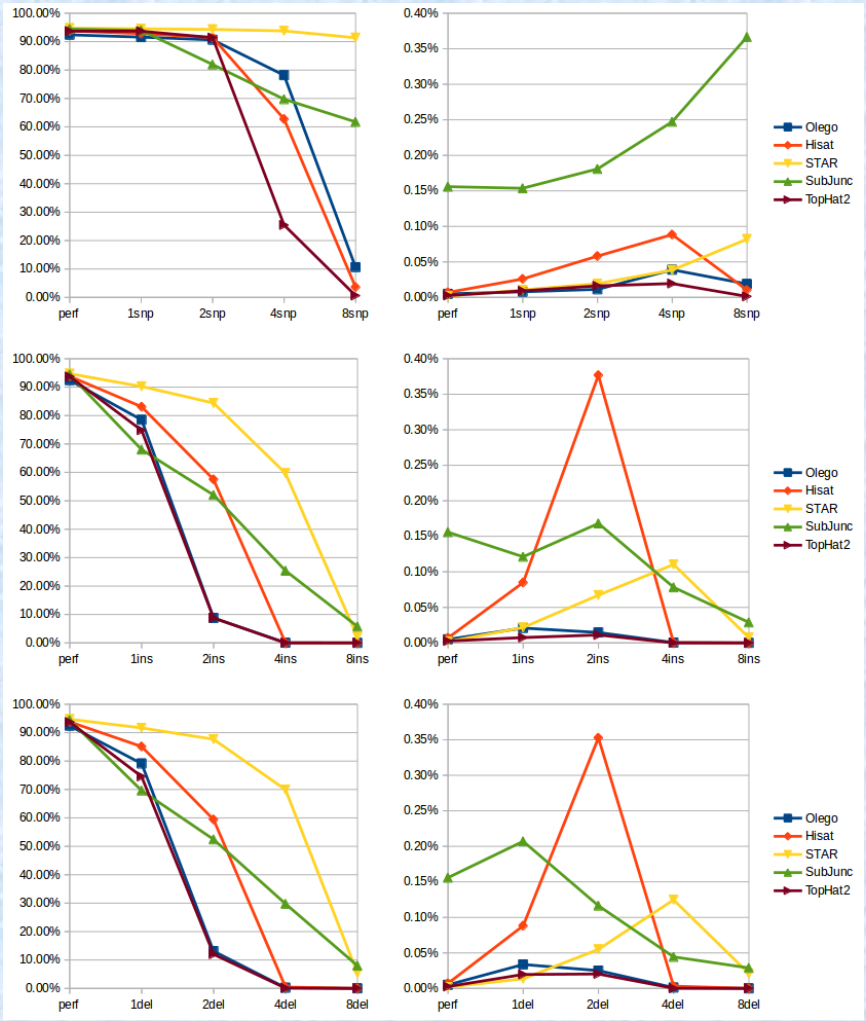
Hisat

- Employs two different types of indexes:
  - (i) a global index that represents the entire genome and
  - (ii) numerous small indexes for regions that collectively cover the genome, where each index represents 64,000 bp.
- A combined aproach of anchor and extension

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4655817/

Comparison of RNA aligners





Accuracy of mapping mutated 100bp reads. Left hand side graphs show the correct mapping rates and right hand side shows incorrect mapping rates. Top panels show effect of single nucleotide changes, middle panels show effect of single base insertions and lower panels show single base deletions.

http://nextgenseek.com/2015/03/hisat-a-fast-and-memory-lean-rna-seq-aligner/
http://genomespot.blogspot.com/2015/03/hisat-vs-star-vs-tophat2-vs-olego-vs.html

Popular aligners

For RNA
TopHat2
Star
hisat2

All of them can perform local and global alignment

Would you prefer to have local or global alignments?

Depending on the application and read length

Popular RNA aligners

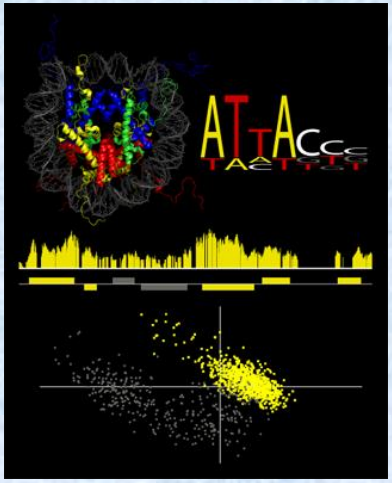
For RNA	
TopHat2	Exon-first approach
Star	Seed searching step (Maximum Mappable Prefix) and clustering/stitching/scoring step.
Hisat2	Double index and combined approach of anchor and extension

RNA Aligners

- **Star features**
  - Outperformed other aligners by a factor of >50 in mapping speed in 2012
  - Detects canonical junctions *de novo*
  - Can discover non-canonical splices and chimeric (fusion) transcripts
  - Can deal with long reads (up to full transcripts)
- **Hisat2 features**
  - The fastest today
  - The lowest memory requirement
  - High sensitivity and accuracy of splice site detection

Outline

- Introduction
- Local and global alignments
- Blast
- Needs of NGS
- Short reads alignment
  - Burrows Wheeler transform
- Genome structure
- DNA alignment
- Transcriptome alignment



QUESTIONS?

Thank you