RNA-SEQ

Francesco Russo

CNR Naples

14 January 2013

Outline

Outline

- DNA, Sequencing, Sanger Sequencing, Next Generation Sequencing.

Outline

- DNA, Sequencing, Sanger Sequencing, Next Generation Sequencing.
- Gene, Transcription, RNA, Gene Expression, Isoform Expression,

Outline

- DNA, Sequencing, Sanger Sequencing, Next Generation Sequencing.
- Gene, Transcription, RNA, Gene Expression, Isoform Expression,
- RNA-SEQ,

Outline

- DNA, Sequencing, Sanger Sequencing, Next Generation Sequencing.
- Gene, Transcription, RNA, Gene Expression, Isoform Expression,
- RNA-SEQ,
- Alignment, Gene Expression Quantification, Coverage,

Outline

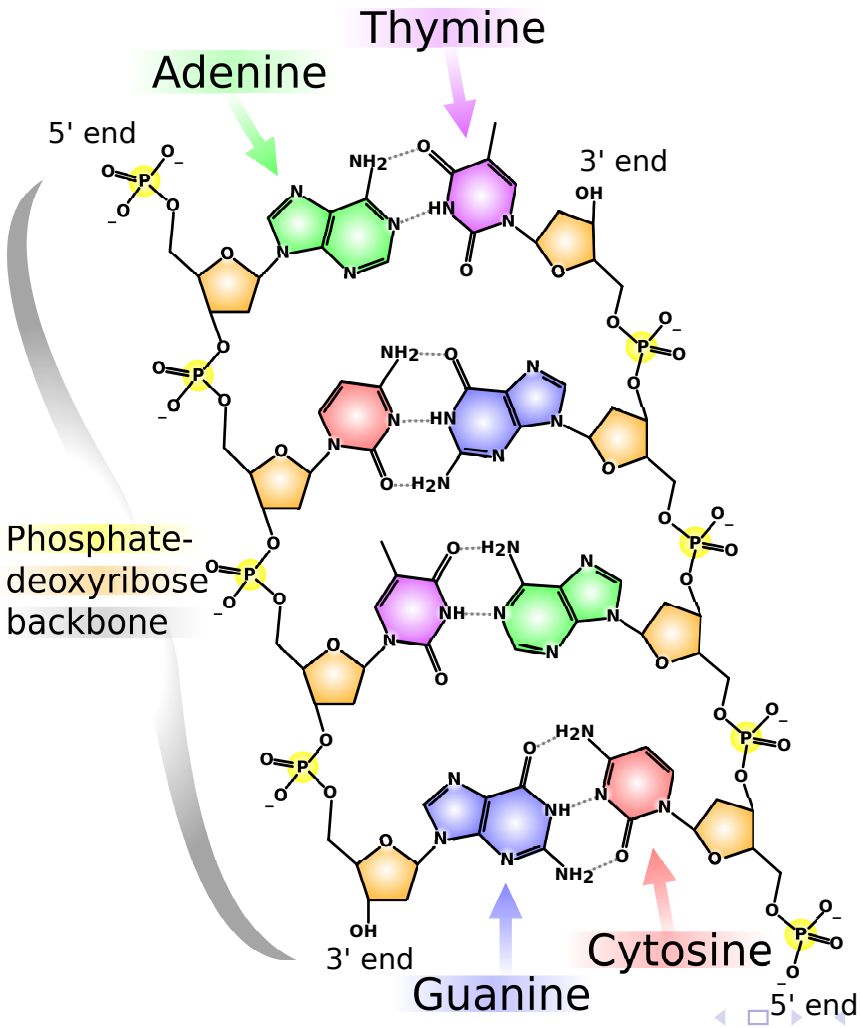
- DNA, Sequencing, Sanger Sequencing, Next Generation Sequencing.
- Gene, Transcription, RNA, Gene Expression, Isoform Expression,
- RNA-SEQ,
- Alignment, Gene Expression Quantification, Coverage,
- Computational Challenges, Questions of Interest,

Outline

- DNA, Sequencing, Sanger Sequencing, Next Generation Sequencing.
- Gene, Transcription, RNA, Gene Expression, Isoform Expression,
- RNA-SEQ,
- Alignment, Gene Expression Quantification, Coverage,
- Computational Challenges, Questions of Interest,
- Some Open Problems.

What is DNA?

DNA (Deoxy Ribonucleic Acid) is a informational molecule made from repeating units called *nucleotides*: Adenine, Guanine, Cytosine, Thymine.



What is Sequencing?

What is Sequencing?

Sequencing is the process of determining the precise order of nucleotides within a DNA molecule.

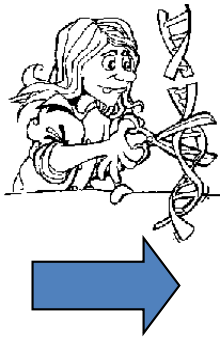
What is Sequencing?

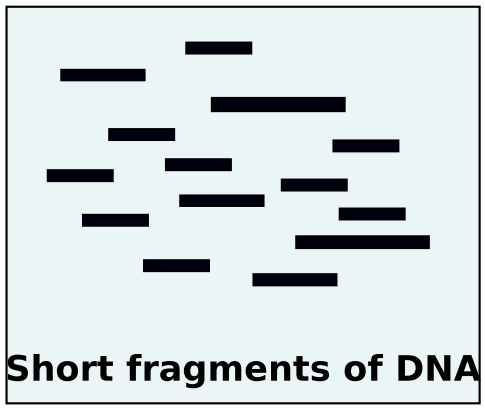
Sequencing is the process of determining the precise order of nucleotides within a DNA molecule.

It can be used to determine the sequence of individual genes, chromosomes or entire genomes.

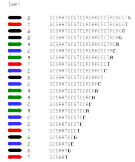
Sequencing process

Genome





Short fragments of DNA



AC..GC TG..GT TC..CC CG..CA TT..TC AC..GC GA..GC TG..AC CT..TG GT..GC AC..GC AC..GC AA..GC AT..AT TT..CC

Short DNA sequences

ACGTGGTAA CGTATACAC TAGGCCATA GTAATGGCG CACCCTTAG TGGCGTATA CATA...

ACGTGGTAATGGCGTATACACCCTTAGGCCATA

ACGTGACCGGTACTGGTAACGTACA
CCTACGTGACCGGTACTGGTAACGT
ACGCCTACGTGACCGGTACTGGTAA
CGTATACACGTGACCGGTACTGGTA
ACGTACACCTACGTGACCGGTACTG
GTAACGTACGCCTACGTGACCGGTA
CTGGTAACGTATACCTCT...

Sequenced genome

Sanger sequencing

Sanger sequencing provides high quality and it is the most used method.

Sanger sequencing

Sanger sequencing provides high quality and it is the most used method. This method makes use of special enzymes to synthesize fragments of DNA that terminate when a selected base appears in the sequence of DNA being read.

Sanger sequencing

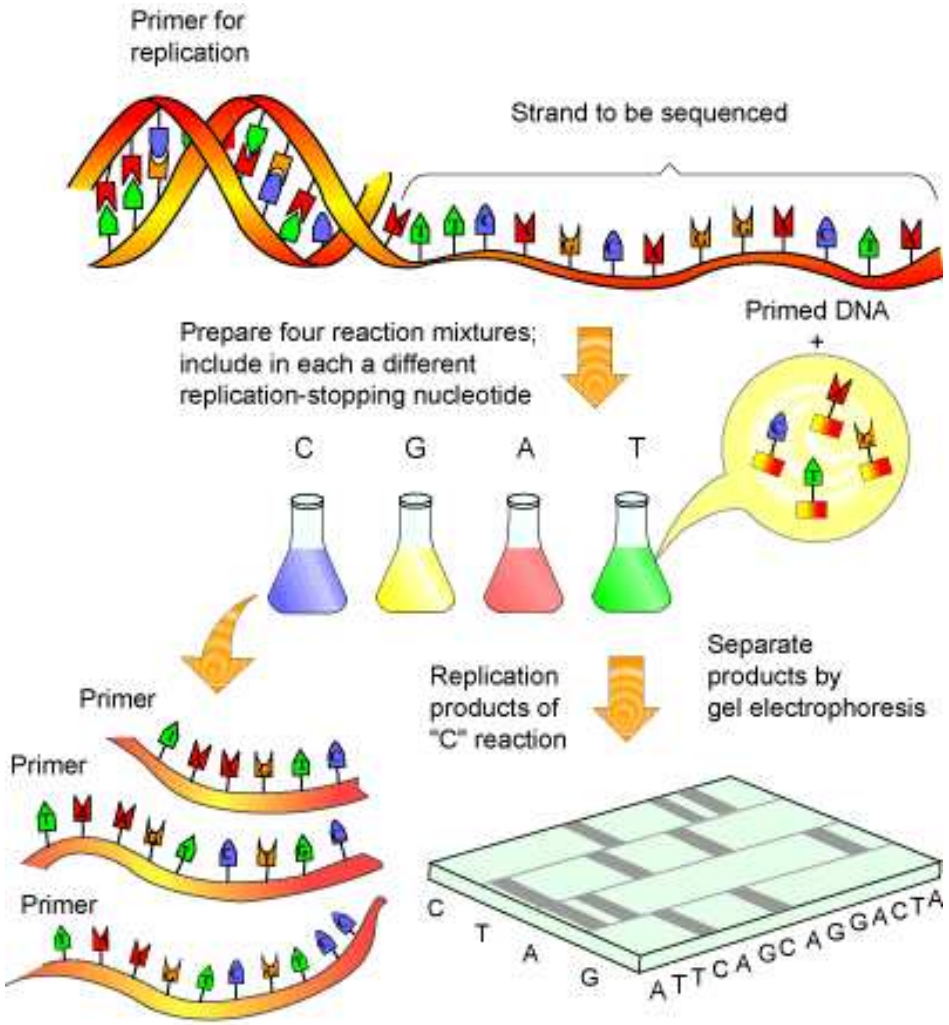
Sanger sequencing provides high quality and it is the most used method. This method makes use of special enzymes to synthesize fragments of DNA that terminate when a selected base appears in the sequence of DNA being read.
These fragments are then sorted according to size by placing them in a slab of polymeric gel and applying an electric field. This technique is called *electrophoresis*.

Sanger sequencing

Sanger sequencing provides high quality and it is the most used method.
This method makes use of special enzymes to synthesize fragments of DNA that terminate when a selected base appears in the sequence of DNA being read.
These fragments are then sorted according to size by placing them in a slab of polymeric gel and applying an electric field. This technique is called *electrophoresis*.
Because of DNA's negative charge, the fragments move across the gel toward the positive electrode. The shorter the fragment, the faster it moves.

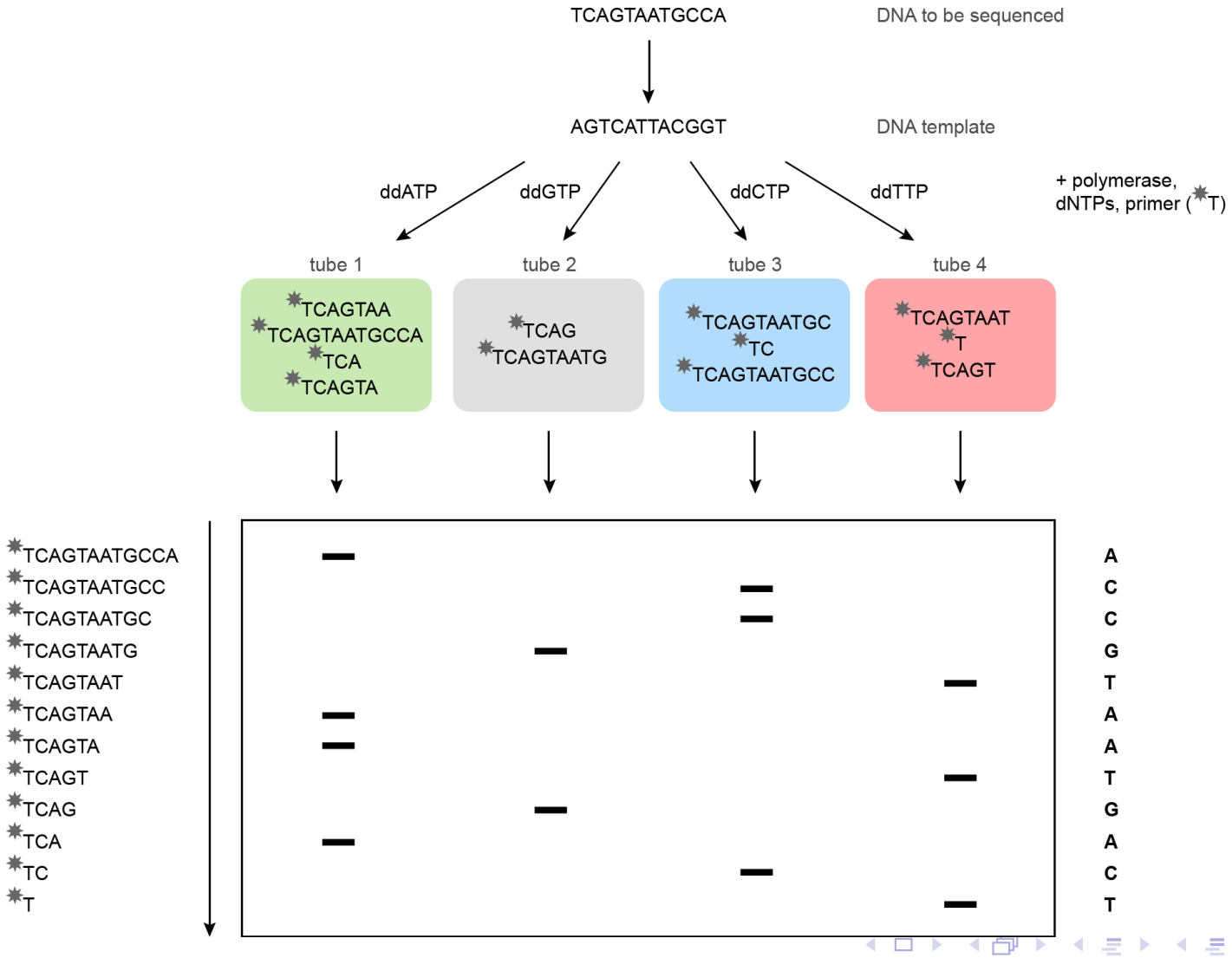
Sanger sequencing

Shortest fragments make the furthest progress. By considering the results of our race, we can reconstruct the nucleotide sequence.



Sanger sequencing

Each tube contains different length fragments.



The output Sanger sequencing machines are FASTA files

FASTA format is a text-based format for representing nucleotide sequences, in which nucleotides are represented using single-letter codes.

```
>gi|9626685|ref|NC_001477.1| Dengue virus type 1, complete genome
AGTTGTTAGTCTACGTGGACCGACAAGAACAGTTTCGAATCGGAAGCTTGCTTAACGTAGTTCTAACAGT
TTTTTATTAGAGAGCAGATCTCTGATGAACAACCAACGGAAAAAGACGGGTCGACCGTCTTTCAATATGC
TGAAACGCGCGAGAAACCGCGTGTCAACTGTTTCACAGTTGGCGAAGAGATTCTCAAAAGGATTGCTTTC
AGGCCAAGGACCCATGAAATTGGTGATGGCTTTTATAGCATTCCTAAGATTTCTAGCCATACCTCCAACA
GCAGGAATTTTGGCTAGATGGGGCTCATTCAAGAAGAATGGAGCGATCAAAGTGTTACGGGGTTTCAAGA
AAGAAATCTCAAACATGTTGAACATAATGAACAGGAGGAAAAGATCTGTGACCATGCTCCTCATGCTGCT
GCCCACAGCCCTGGCGTTCCATCTGACCACCCGAGGGGGAGAGCCGCACATGATAGTTAGCAAGCAGGAA
AGAGGAAAATCACTTTTGTTTAAGACCTCTGCAGGTGTCAACATGTGCACCCTTATTGCAATGGATTTGG
GAGAGTTATGTGAGGACACAATGACCTACAAATGCCCCCGGATCACTGAGACGGAACCAGATGACGTTGA
CTGTTGGTGCAATGCCACGGAGACATGGGTGACCTATGGAACATGTTCTCAAACTGGTGAACACCGACGA
GACAAACGTTCCGTCGCACTGGCACCACACGTAGGGCTTGGTCTAGAAACAAGAACCGAAACGTGGATGT
CCTCTGAAGGCGCTTGGAAACAAATACAAAAAGTGGAGACCTGGGCTCTGAGACACCCAGGATTCACGGT
GATAGCCCTTTTTCTAGCACATGCCATAGGAACATCCATCACCCAGAAAGGGATCATTTTTATTTTGCTG
ATGCTGGTAACTCCATCCATGGCCATGCGGTGCGTGGGAATAGGCAACAGAGACTTCGTGGAAGGACTGT
CAGGAGCTACGTGGGTGGATGTGGTACTGGAGCATGGAAGTTGCGTCACTACCATGGCAAAAGACAAACC
AACACTGGACATTGAACTCTTGAAGACGGAGGTCACAAACCCTGCCGTCCTGCGCAAACTGTGCATTGAA
GCTAAAATATCAAACACCACCACCGATTCGAGATGTCCAACACAAGGAGAAGCCACGCTGGTGGAAGAAC
AGGACACGAACTTTGTGTGTCGACGAACGTTCGTGGACAGAGGCTGGGGCAATGGTTGTGGGCTATTCGG
AAAAGGTAGCTTAATAACGTGTGCTAAGTTTAAGTGTGTGACAAAACTGGAAGGAAAGATAGTCCAATAT
GAAAACTTAAAATATTCAGTGATAGTCACCGTACACACTGGAGACCAGCACCAAGTTGGAAATGAGACCA
CAGAACATGGAACAACTGCAACCATAACACCTCAAGCTCCCACGTCGGAAATACAGCTGACAGACTACGG
AGCTCTAACATTGGATTGTTCACCTAGAACAGGGCTAGACTTTAATGAGATGGTGTTGTTGACAATGAAA
```

The output Sanger sequencing machines are FASTA files

FASTA format is a text-based format for representing nucleotide sequences, in which nucleotides are represented using single-letter codes.

```
>gi|9626685|ref|NC_001477.1| Dengue virus type 1, complete genome
AGTTGTTAGTCTACGTGGACCGACAAGAACAGTTTCGAATCGGAAGCTTGCTTAACGTAGTTCTAACAGT
TTTTTATTAGAGAGCAGATCTCTGATGAACAACCAACGGAAAAAGACGGGTCGACCGTCTTTCAATATGC
TGAAACGCGCGAGAAACCGCGTGTCAACTGTTTCACAGTTGGCGAAGAGATTCTCAAAAGGATTGCTTTC
AGGCCAAGGACCCATGAAATTGGTGATGGCTTTTATAGCATTCCTAAGATTTCTAGCCATACCTCCAACA
GCAGGAATTTTGGCTAGATGGGGCTCATTCAAGAAGAATGGAGCGATCAAAGTGTTACGGGGTTTCAAGA
AAGAAATCTCAAACATGTTGAACATAATGAACAGGAGGAAAAGATCTGTGACCATGCTCCTCATGCTGCT
GCCCACAGCCCTGGCGTTCCATCTGACCACCCGAGGGGGAGAGCCGCACATGATAGTTAGCAAGCAGGAA
AGAGGAAAATCACTTTTGTTTAAGACCTCTGCAGGTGTCAACATGTGCACCCTTATTGCAATGGATTTGG
GAGAGTTATGTGAGGACACAATGACCTACAAATGCCCCCGGATCACTGAGACGGAACCAGATGACGTTGA
CTGTTGGTGCAATGCCACGGAGACATGGGTGACCTATGGAACATGTTCTCAAACTGGTGAACACCGACGA
GACAAACGTTCCGTCGCACTGGCACCACACGTAGGGCTTGGTCTAGAAACAAGAACCGAAACGTGGATGT
CCTCTGAAGGCGCTTGGAAACAAATACAAAAAGTGGAGACCTGGGCTCTGAGACACCCAGGATTCACGGT
GATAGCCCTTTTTCTAGCACATGCCATAGGAACATCCATCACCCAGAAAGGGATCATTTTTATTTTGCTG
ATGCTGGTAACTCCATCCATGGCCATGCGGTGCGTGGGAATAGGCAACAGAGACTTCGTGGAAGGACTGT
CAGGAGCTACGTGGGTGGATGTGGTACTGGAGCATGGAAGTTGCGTCACTACCATGGCAAAAGACAAACC
AACACTGGACATTGAACTCTTGAAGACGGAGGTCACAAACCCTGCCGTCCTGCGCAAACTGTGCATTGAA
GCTAAAATATCAAACACCACCACCGATTCGAGATGTCCAACACAAGGAGAAGCCACGCTGGTGGAAGAAC
AGGACACGAACTTTGTGTGTCGACGAACGTTCGTGGACAGAGGCTGGGGCAATGGTTGTGGGCTATTCGG
AAAAGGTAGCTTAATAACGTGTGCTAAGTTTAAGTGTGTGACAAAACTGGAAGGAAAGATAGTCCAATAT
GAAAACTTAAAATATTCAGTGATAGTCACCGTACACACTGGAGACCAGCACCAAGTTGGAAATGAGACCA
CAGAACATGGAACAACTGCAACCATAACACCTCAAGCTCCCACGTCGGAAATACAGCTGACAGACTACGG
AGCTCTAACATTGGATTGTTCACCTAGAACAGGGCTAGACTTTAATGAGATGGTGTTGTTGACAATGAAA
```

FASTA file are easy to manipulate by using scripting languages such as: Python, AWK, and Perl.

Sanger sequencing limitations

Sanger sequencing limitations

- Main limitation: the size of DNA fragments that can be read in this way is about 10^3 base pairs.

Sanger sequencing limitations

- Main limitation: the size of DNA fragments that can be read in this way is about 10^3 base pairs.
- Main problem: most genomes are enormous (e.g $3 \cdot 10^9$ base pairs in case of human genome, for each strand of DNA). Therefore, it is impossible for DNA to be sequenced directly.

Sanger sequencing limitations

- Main limitation: the size of DNA fragments that can be read in this way is about 10^3 base pairs.
- Main problem: most genomes are enormous (e.g $3 \cdot 10^9$ base pairs in case of human genome, for each strand of DNA). Therefore, it is impossible for DNA to be sequenced directly.
- This is called *Large-Scale Sequencing*.

Sanger sequencing limitations

- Main limitation: the size of DNA fragments that can be read in this way is about 10^3 base pairs.
- Main problem: most genomes are enormous (e.g $3 \cdot 10^9$ base pairs in case of human genome, for each strand of DNA). Therefore, it is impossible for DNA to be sequenced directly.
- This is called *Large-Scale Sequencing*.
- A very good video about Sanger sequencing can be found here:

http://www.youtube.com/watch?v=bEFLBf5WEtc

What is Next Generation Sequencing?

What is Next Generation Sequencing?

The *Human Genome Project* began in 1990 and finished in 2003. It required the collaboration of hundreds of researchers from 20 institutions in 6 countries. The estimated cost is about 270.000.000 dollars.

What is Next Generation Sequencing?

The *Human Genome Project* began in 1990 and finished in 2003. It required the collaboration of hundreds of researchers from 20 institutions in 6 countries. The estimated cost is about 270.000.000 dollars.
With the increase of computer technology and with the decrease sequencer costs it has been possible to produce more and more reads and to achieve the sequencing task in a shorter and shorter time.

What is Next Generation Sequencing?

The *Human Genome Project* began in 1990 and finished in 2003. It required the collaboration of hundreds of researchers from 20 institutions in 6 countries. The estimated cost is about 270.000.000 dollars.
With the increase of computer technology and with the decrease sequencer costs it has been possible to produce more and more reads and to achieve the sequencing task in a shorter and shorter time.
"Nowadays", thanks to NGS methodologies sequencers are capable to sequence the whole human genome in less that one week at a cost of less than 50.000 dollars.

What is Next Generation Sequencing?

The *Human Genome Project* began in 1990 and finished in 2003. It required the collaboration of hundreds of researchers from 20 institutions in 6 countries. The estimated cost is about 270.000.000 dollars.
With the increase of computer technology and with the decrease sequencer costs it has been possible to produce more and more reads and to achieve the sequencing task in a shorter and shorter time.
"Nowadays", thanks to NGS methodologies sequencers are capable to sequence the whole human genome in less that one week at a cost of less than 50.000 dollars.
Cost and time are decreasing at each new release, while speed, accuracy and resolution are improving dramatically.

What is Next Generation Sequencing?

What is Next Generation Sequencing?

"NGS extends the sequencing processes across millions of reactions in a massively parallel fashion, rather than being limited to few DNA fragments.

What is Next Generation Sequencing?

"NGS extends the sequencing processes across millions of reactions in a massively parallel fashion, rather than being limited to few DNA fragments. This advance enables rapid sequencing of large stretches of DNA,

What is Next Generation Sequencing?

"NGS extends the sequencing processes across millions of reactions in a massively parallel fashion, rather than being limited to few DNA fragments. This advance enables rapid sequencing of large stretches of DNA, The latest instruments are capable of producing hundreds of GigaBytes of data in a single sequencing run."

- An Introduction to Next-Generation Sequencing Technology, from illumina.com

What is Next Generation Sequencing?

"NGS extends the sequencing processes across millions of reactions in a massively parallel fashion, rather than being limited to few DNA fragments. This advance enables rapid sequencing of large stretches of DNA, The latest instruments are capable of producing hundreds of GigaBytes of data in a single sequencing run."

- An Introduction to Next-Generation Sequencing Technology, from illumina.com

NGS sequencing methodologies are able to generate hundreds of millions of short *reads* (e.g. 30 bps - 100 bps) with their quality values.

Data produced after a typical NGS process

FASTQ format is a text-based format for storing both a biological sequence and its corresponding quality scores.

```
@HWI-1KL111:53:D176AACXX:1:1101:1221:1861 1:N:0:CGATGT
NTGCTTCCTCAAAAAATTACAAAAATGCCAGTGGAGTTGTGAACTCTTCACCTCGAAGTCATAGCGCCACAAATG
+
#4=DDFFFHHHGHJIIIJJIJJJJJHHIJJJFHIHGDGHIJJIJJJGJJJJJIJI<FHHHIGGIJHHFFFEDEEC
@HWI-1KL111:53:D176AACXX:1:1101:1197:1873 1:N:0:CGATGT
NTGATGCTTTCAACTGCATACTTAAAGCTTGCTCCAGTTTTGTCCATCTTGTTTAAAAAACAGATTCGAGGAAAA
+
#1:DBDDDFHBDFIEGHIIHHIDHGHGHAHII>C@GIIIIIEDHGGHEGF@BDH:BGHH6:==3@@##########
@HWI-1KL111:53:D176AACXX:1:1101:1228:1956 1:N:0:CGATGT
NTGAAGGGCTCTCCTTCCTGCACCAAGCTGGGAGGTTTCGGGCTCCTCTGGGGCATTCAATGCCTGCAAGCAAG
+
#1=DDFDDHHHHHJJJJJJIIJJJIJIGHHIGE@GHGHIIIIIJIJIIIFHEGIE<EDFFFFFFECEEECDD=CC
@HWI-1KL111:53:D176AACXX:1:1101:1371:1853 1:N:0:CGATGT
NGCCGGCGGACCGAAGAACGCAGGAAGGGGGCCGGGGGGACCCGCCCCCGGCCGGCCGCAGCCATGAACTCCAAC
+
#1=DDFFFHDFHHIGIIJIGIGHJHHHGB<>B3=BDDDD0<?B@BBBBD8B5<8B8BBBDDBB<CCACCADD:?>
@HWI-1KL111:53:D176AACXX:1:1101:1492:1889 1:N:0:CGATGT
NTATTTTTAGTAGAGACTGGATTTCTCCATGTTGGTCAGACAGGTCTCGAACTCCTGACCTCAGGTGATCTGCCT
+
#1=DDFFFFHHHFIIJIJJJGIJJJJJJBIIIEHIBGIIJJJIJBGIIJIDGGIJIJJIHJIEII7@GIIJJJIC
```

However...

However...

... as the quantities of sequence data increase exponentially thanks to NGS methodologies, ...

However...

... as the quantities of sequence data increase exponentially thanks to NGS methodologies, ...

... the analysis of such a large data (usually hundreds of GigaBytes) became the major "bottom-neck" of experimental investigations.

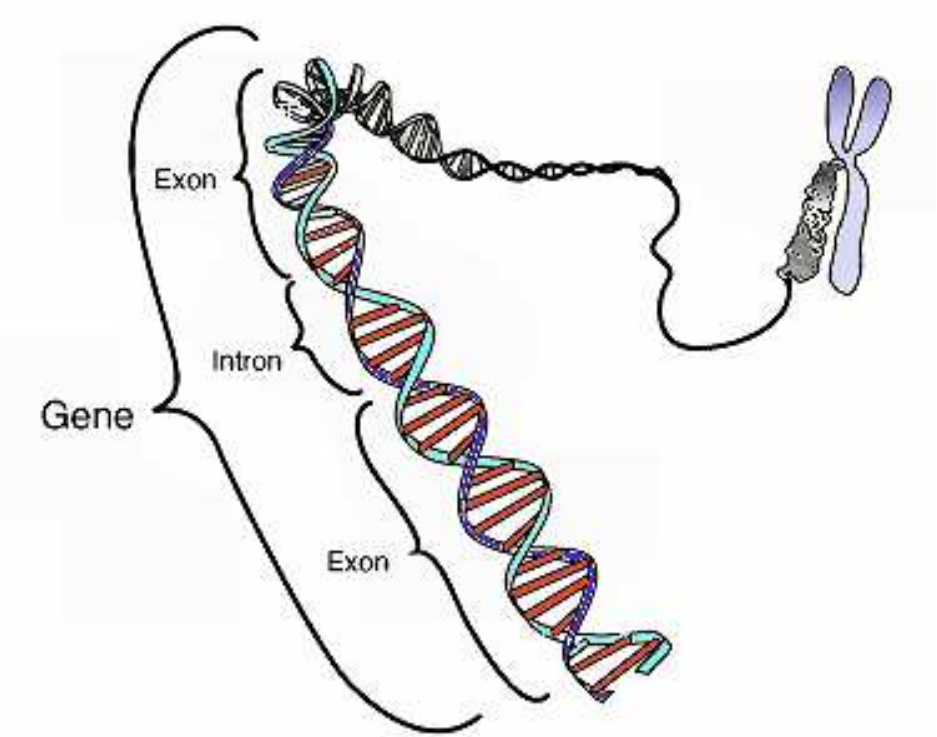
What are Genes?

What are Genes?

- A *Gene* is a stretch of DNA that codes for a specific product. Usually, a gene codes for proteins.

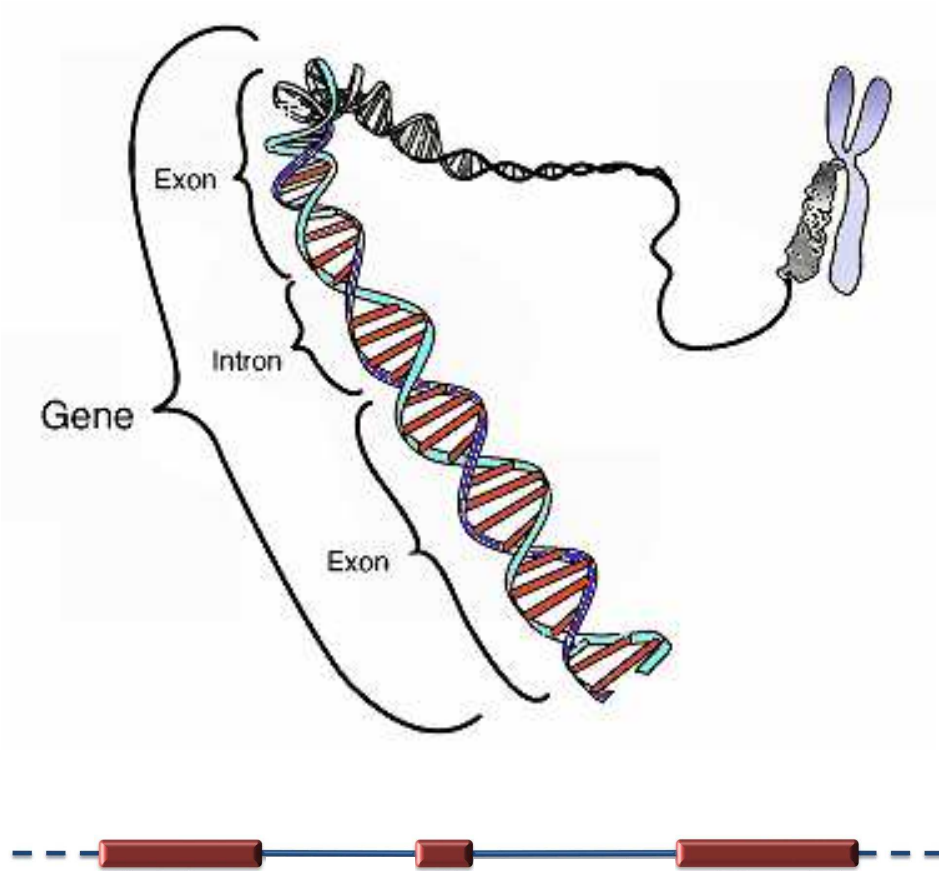
What are Genes?

- A *Gene* is a stretch of DNA that codes for a specific product. Usually, a gene codes for proteins.



What are Genes?

- A *Gene* is a stretch of DNA that codes for a specific product. Usually, a gene codes for proteins.



Transcription: pre-mRNA, mRNA and RNA

Transcription: pre-mRNA, mRNA and RNA

- RNA (Ribonucleic Acid) is a single-stranded nucleic acid.

Transcription: pre-mRNA, mRNA and RNA

- RNA (Ribonucleic Acid) is a single-stranded nucleic acid.
- *Transcription* is the production of RNA copies of the DNA.

Transcription: pre-mRNA, mRNA and RNA

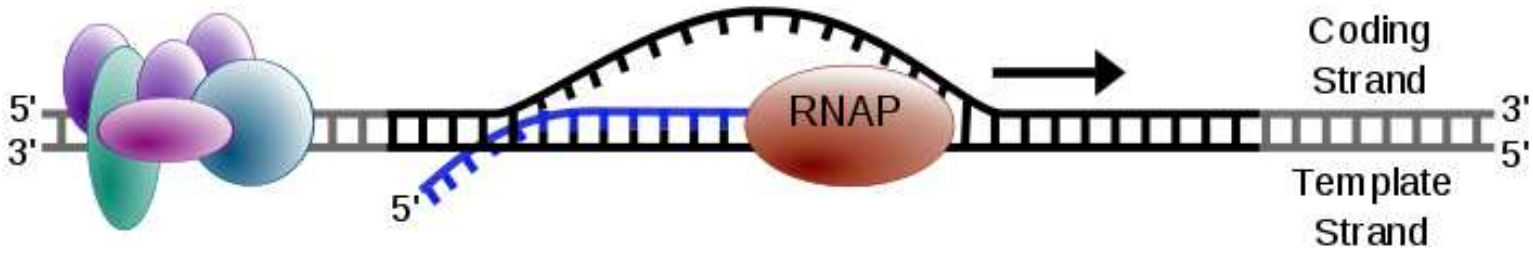
- RNA (Ribonucleic Acid) is a single-stranded nucleic acid.
- *Transcription* is the production of RNA copies of the DNA.
- Transcription is performed by a special enzyme called *RNA-polymerase*.

Transcription: pre-mRNA, mRNA and RNA

- RNA (Ribonucleic Acid) is a single-stranded nucleic acid.
- *Transcription* is the production of RNA copies of the DNA.
- Transcription is performed by a special enzyme called *RNA-polymerase*.
- Transcription of genes forms a primary transcript of messenger RNA, called *pre-mRNA*.

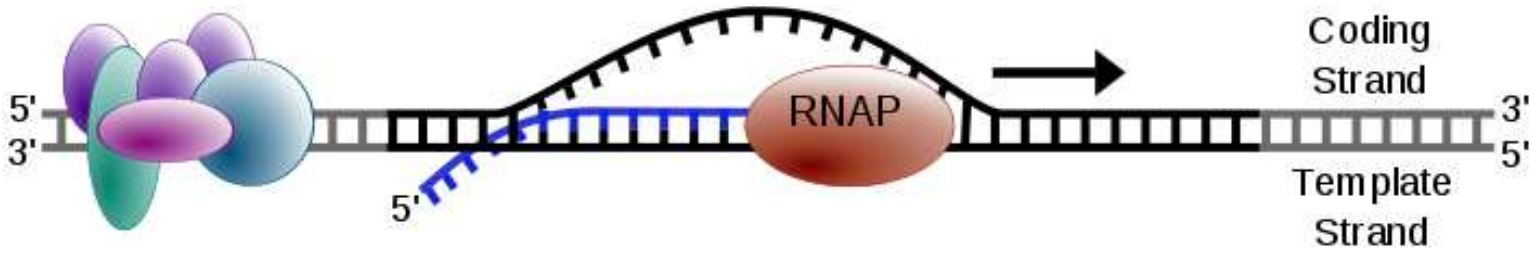
Transcription: pre-mRNA, mRNA and RNA

- RNA (Ribonucleic Acid) is a single-stranded nucleic acid.
- *Transcription* is the production of RNA copies of the DNA.
- Transcription is performed by a special enzyme called *RNA-polymerase*.
- Transcription of genes forms a primary transcript of messenger RNA, called *pre-mRNA*.



Transcription: pre-mRNA, mRNA and RNA

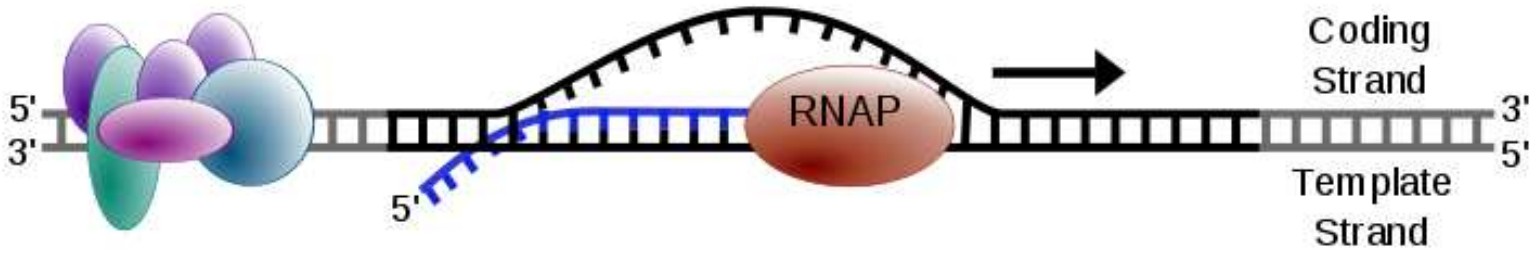
- RNA (Ribonucleic Acid) is a single-stranded nucleic acid.
- *Transcription* is the production of RNA copies of the DNA.
- Transcription is performed by a special enzyme called *RNA-polymerase*.
- Transcription of genes forms a primary transcript of messenger RNA, called *pre-mRNA*.



Pre-mRNA first has to undergo a series of modifications to become a mature *mRNA*.

Transcription: pre-mRNA, mRNA and RNA

- RNA (Ribonucleic Acid) is a single-stranded nucleic acid.
- *Transcription* is the production of RNA copies of the DNA.
- Transcription is performed by a special enzyme called *RNA-polymerase*.
- Transcription of genes forms a primary transcript of messenger RNA, called *pre-mRNA*.



Pre-mRNA first has to undergo a series of modifications to become a mature *mRNA*.
Differently from DNA, RNA is very perishable.

Splicing

One of those pre-mRNA modifications is called *Splicing* (it only takes place in *eukaryotic* cells).

Splicing

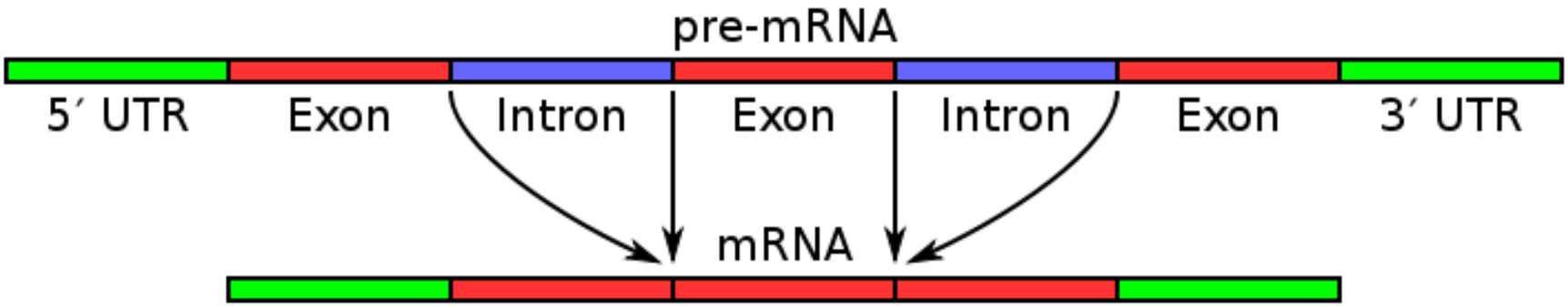
One of those pre-mRNA modifications is called *Splicing* (it only takes place in *eukaryotic* cells).
Splicing is a modification in which introns are removed and exons are joined.

Splicing

One of those pre-mRNA modifications is called *Splicing* (it only takes place in *eukaryotic* cells).
Splicing is a modification in which introns are removed and exons are joined.
It takes place after or concurrently with transcription process.

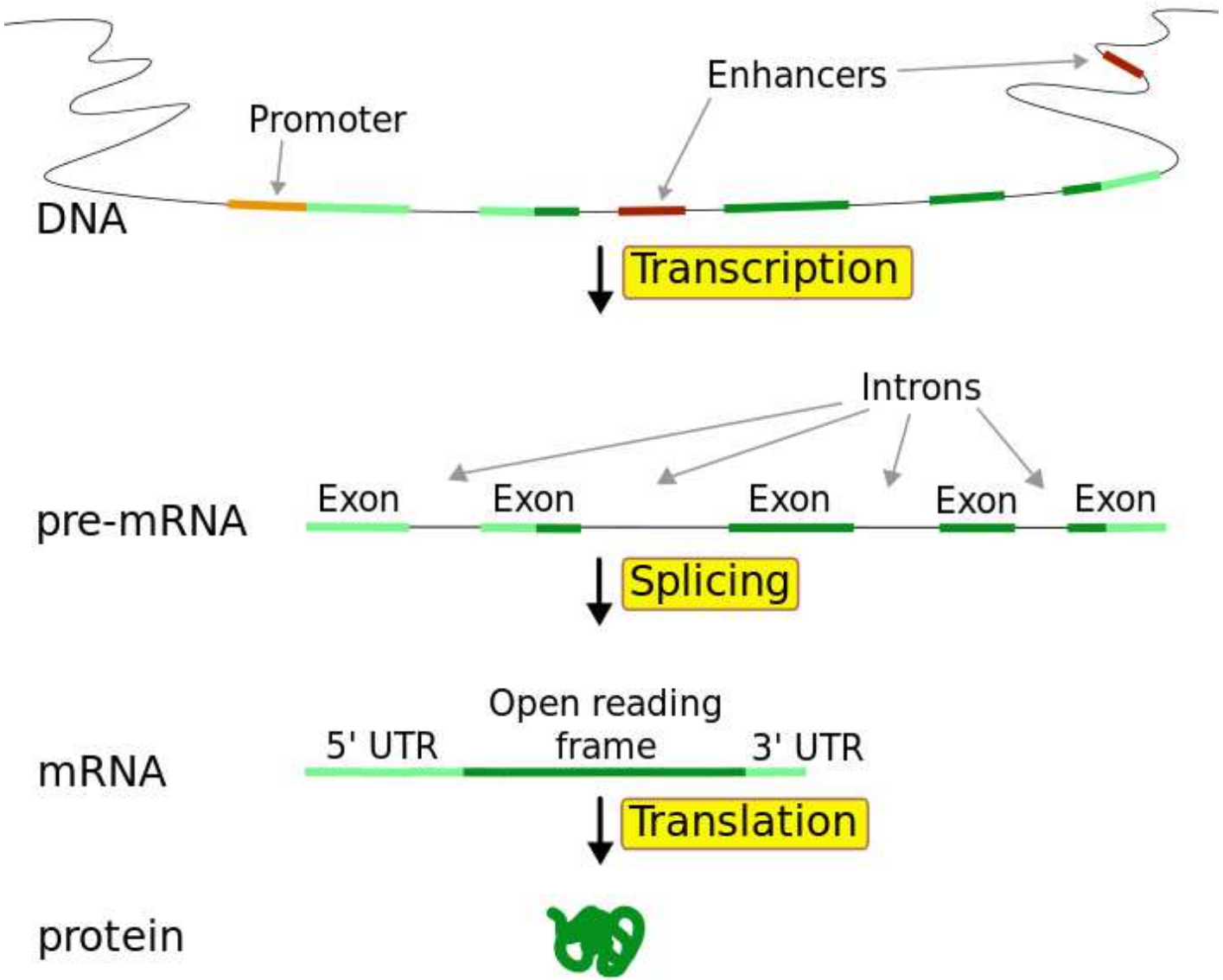
Splicing

One of those pre-mRNA modifications is called *Splicing* (it only takes place in *eukaryotic* cells).
Splicing is a modification in which introns are removed and exons are joined.
It takes place after or concurrently with transcription process.

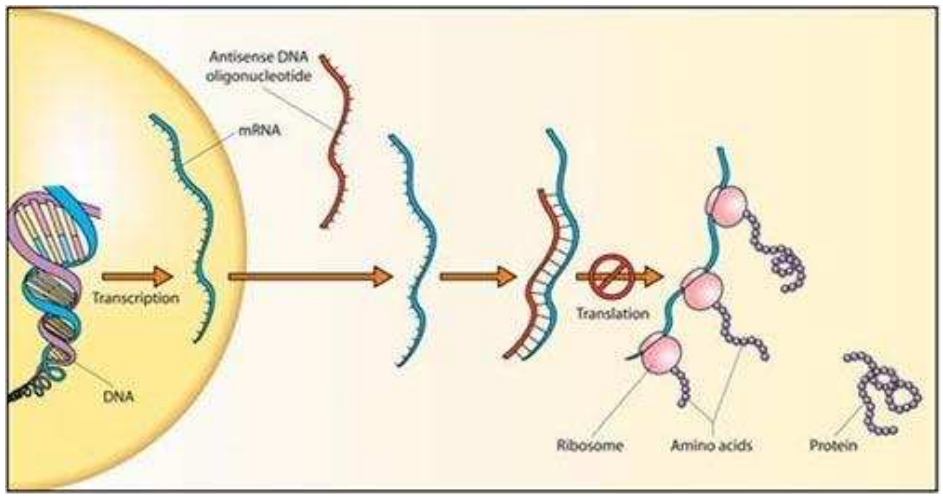


UTR stands for Untranslated Region.

Protein production



Protein production

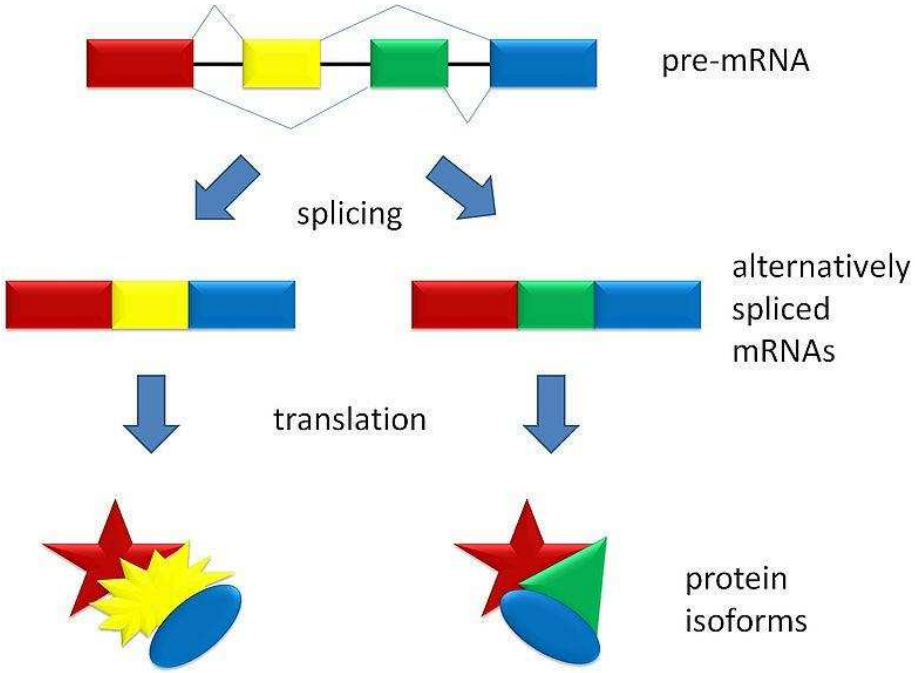


Alternative Splicing

- Another pre-mRNA modification is *Alternative Splicing* that creates series of different transcripts originating from a single gene.

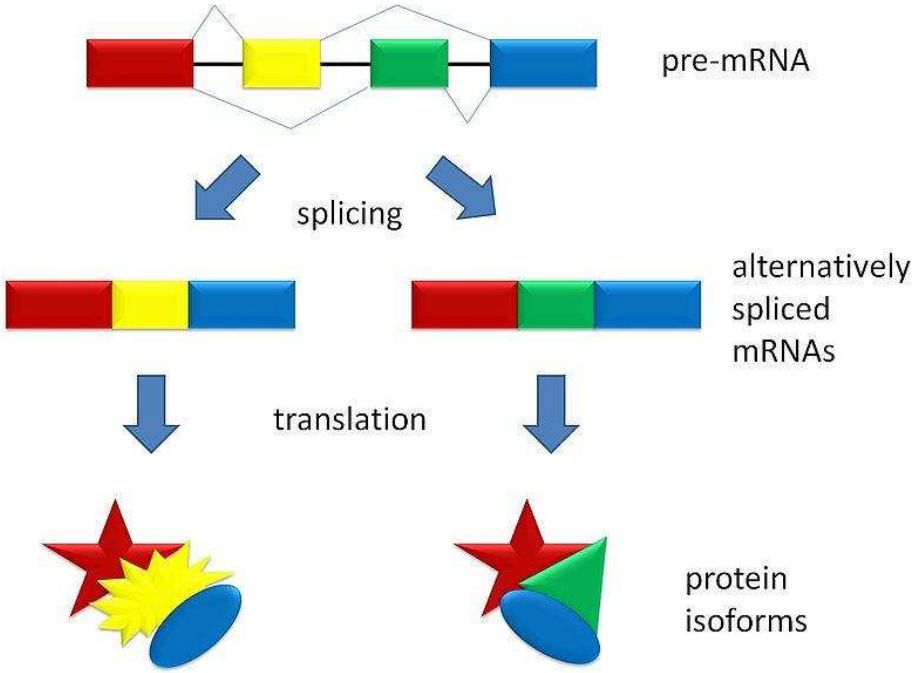
Alternative Splicing

- Another pre-mRNA modification is *Alternative Splicing* that creates series of different transcripts originating from a single gene.



Alternative Splicing

- Another pre-mRNA modification is *Alternative Splicing* that creates series of different transcripts originating from a single gene.



- Therefore, Alternative Splicing extends the complexity of gene expression.

Gene Expression

Gene Expression is the process by which information stored by a gene is used to build some products.

Gene Expression

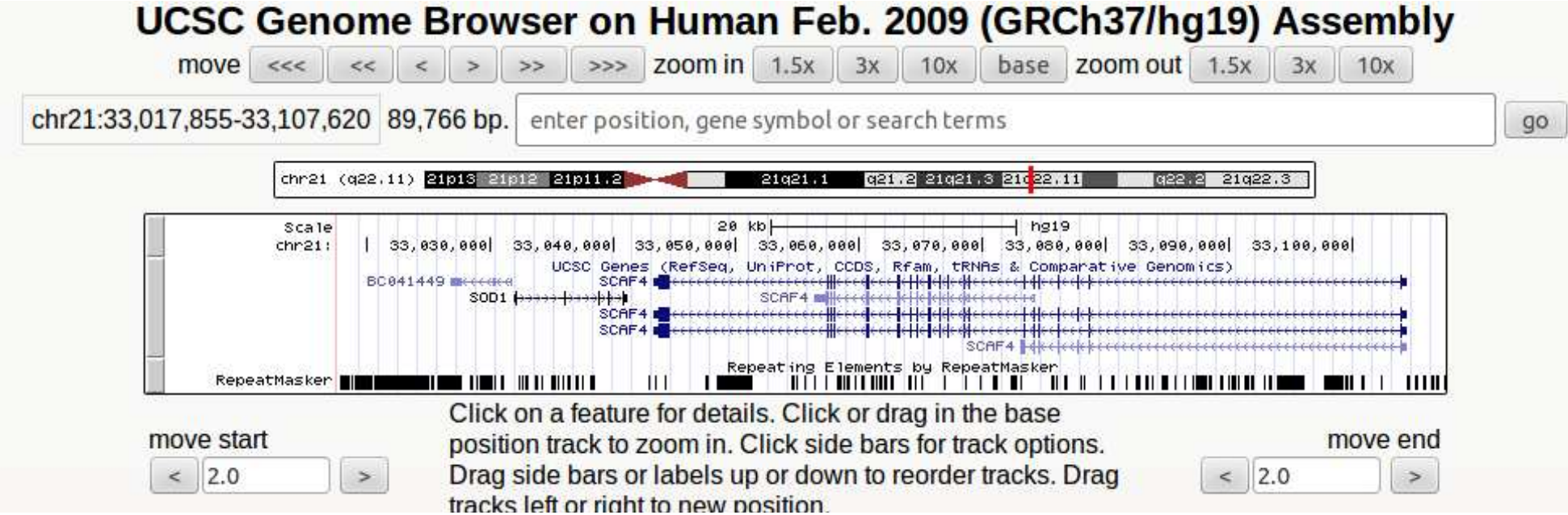
Gene Expression is the process by which information stored by a gene is used to build some products.
These products are often proteins, but they might also be functional RNA.

Gene Expression

Gene Expression is the process by which information stored by a gene is used to build some products.
These products are often proteins, but they might also be functional RNA.
Therefore, Alternative Splicing extends the amount of protein production.

Genome Browser

A way to display and navigate across annotated genes is using a free software called: Genome Browser.



GTF

The Gene Transfer Format is a file format used to hold information about gene structure.

```
chr1    protein_coding  exon    34554   35174   .       -       .       gene_id "ENSG00000237613"; transcript_id "ENST00000417324";
chr1    protein_coding  stop_codon      35138   35140   .       -       0       gene_id "ENSG00000237613"; transcript_id "ENST00000417324";
chr1    protein_coding  CDS     35141   35174   .       -       1       gene_id "ENSG00000237613"; transcript_id "ENST00000417324";
chr1    processed_transcript    exon    35245   35481   .       -       .       gene_id "ENSG00000237613"; transcript_id "ENST00000461467";
chr1    protein_coding  CDS     35277   35481   .       -       2       gene_id "ENSG00000237613"; transcript_id "ENST00000417324";
chr1    protein_coding  exon    35277   35481   .       -       .       gene_id "ENSG00000237613"; transcript_id "ENST00000417324";
chr1    processed_transcript    exon    35721   36073   .       -       .       gene_id "ENSG00000237613"; transcript_id "ENST00000461467";
chr1    protein_coding  CDS     35721   35736   .       -       0       gene_id "ENSG00000237613"; transcript_id "ENST00000417324";
chr1    protein_coding  exon    35721   36081   .       -       .       gene_id "ENSG00000237613"; transcript_id "ENST00000417324";
chr1    protein_coding  start_codon     35734   35736   .       -       0       gene_id "ENSG00000237613"; transcript_id "ENST00000417324";
chr1    protein_coding  CDS     69091   70005   .       +       0       gene_id "ENSG00000186092"; transcript_id "ENST00000335137";
chr1    protein_coding  exon    69091   70008   .       +       .       gene_id "ENSG00000186092"; transcript_id "ENST00000335137";
chr1    protein_coding  start_codon     69091   69093   .       +       0       gene_id "ENSG00000186092"; transcript_id "ENST00000335137";
chr1    protein_coding  stop_codon      70006   70008   .       +       0       gene_id "ENSG00000186092"; transcript_id "ENST00000335137";
chr1    protein_coding  exon    367640  368634  .       +       .       gene_id "ENSG00000235249"; transcript_id "ENST00000426406";
chr1    protein_coding  CDS     367659  368594  .       +       0       gene_id "ENSG00000235249"; transcript_id "ENST00000426406";
chr1    protein_coding  start_codon     367659  367661  .       +       0       gene_id "ENSG00000235249"; transcript_id "ENST00000426406";
chr1    protein_coding  stop_codon      368595  368597  .       +       0       gene_id "ENSG00000235249"; transcript_id "ENST00000426406";
chr1    protein_coding  exon    621059  622053  .       -       .       gene_id "ENSG00000185097"; transcript_id "ENST00000332831";
chr1    protein_coding  stop_codon      621096  621098  .       -       0       gene_id "ENSG00000185097"; transcript_id "ENST00000332831";
chr1    protein_coding  CDS     621099  622034  .       -       0       gene_id "ENSG00000185097"; transcript_id "ENST00000332831";
chr1    protein_coding  start_codon     622032  622034  .       -       0       gene_id "ENSG00000185097"; transcript_id "ENST00000332831";
chr1    protein_coding  exon    860260  860328  .       +       .       gene_id "ENSG00000187634"; transcript_id "ENST00000420190";
chr1    protein_coding  exon    860530  860569  .       +       .       gene_id "ENSG00000187634"; transcript_id "ENST00000437963";
chr1    protein_coding  exon    861118  861180  .       +       .       gene_id "ENSG00000187634"; transcript_id "ENST00000342066";
chr1    protein_coding  exon    861302  861393  .       +       .       gene_id "ENSG00000187634"; transcript_id "ENST00000342066";
chr1    protein_coding  exon    861302  861393  .       +       .       gene_id "ENSG00000187634"; transcript_id "ENST00000420190";
chr1    protein_coding  exon    861302  861393  .       +       .       gene_id "ENSG00000187634"; transcript_id "ENST00000437963";
chr1    protein_coding  CDS     861322  861393  .       +       0       gene_id "ENSG00000187634"; transcript_id "ENST00000342066";
chr1    protein_coding  CDS     861322  861393  .       +       0       gene_id "ENSG00000187634"; transcript_id "ENST00000420190";
chr1    protein_coding  CDS     861322  861393  .       +       0       gene_id "ENSG00000187634"; transcript_id "ENST00000437963";
```

RNA-SEQ

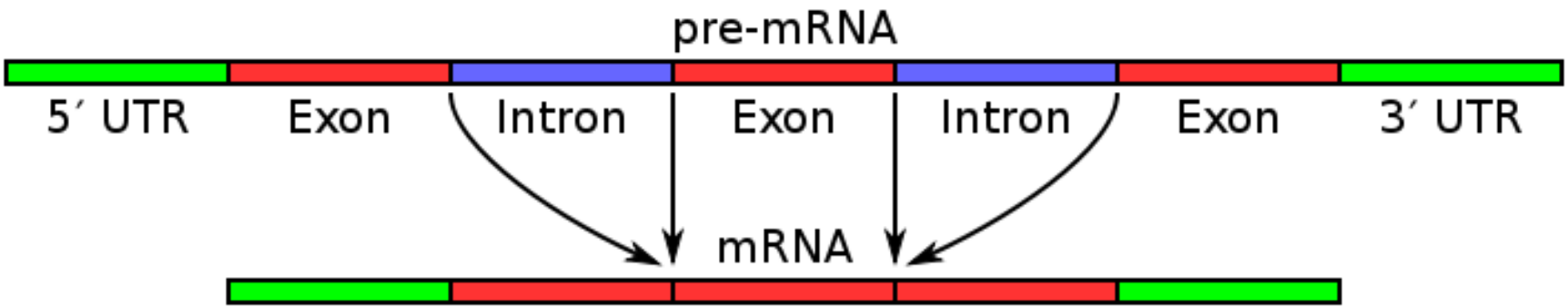
RNA-SEQ refers to the technologies used to sequence *cDNA* in order to get information about the corresponding RNA.

RNA-SEQ

RNA-SEQ refers to the technologies used to sequence *cDNA* in order to get information about the corresponding RNA.
RNA-SEQ is sequencing fragments of the transcriptome.

RNA-SEQ

RNA-SEQ refers to the technologies used to sequence *cDNA* in order to get information about the corresponding RNA.
RNA-SEQ is sequencing fragments of the transcriptome.



Introns are not sequenced.

Alignment

- After the sequencing process, we can use some tools (e.g.: Bowtie) that perform a particular task, called Alignment.

Alignment

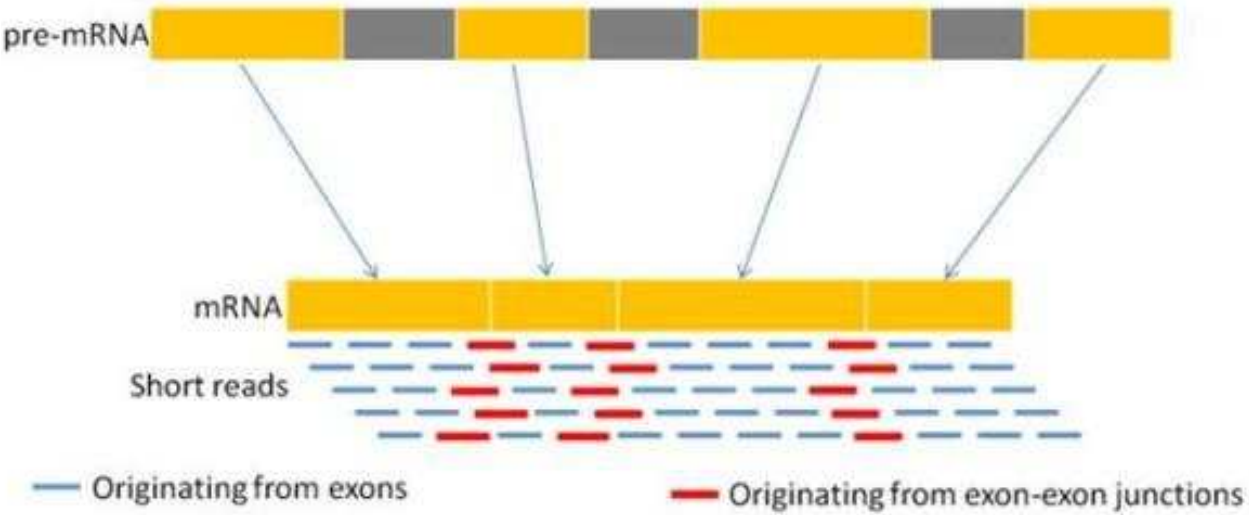
- After the sequencing process, we can use some tools (e.g.: Bowtie) that perform a particular task, called Alignment.
- Alignment is the task of finding a unique location where a short read is identical to the reference.

Alignment

- After the sequencing process, we can use some tools (e.g.: Bowtie) that perform a particular task, called Alignment.
- Alignment is the task of finding a unique location where a short read is identical to the reference.
- Alignment is a useful task to study gene expression.

Alignment

- After the sequencing process, we can use some tools (e.g.: Bowtie) that perform a particular task, called Alignment.
- Alignment is the task of finding a unique location where a short read is identical to the reference.
- Alignment is a useful task to study gene expression.
- RNA alignment, also called Transcript alignment, is particularly challenging due to splicing.



Alignment to a reference genome

Given that the transcriptome is built from the genome, the most commonly used approach is to use the genome itself as the reference.

Alignment to a reference genome

Given that the transcriptome is built from the genome, the most commonly used approach is to use the genome itself as the reference. After the Alignment process we get three types of reads:

Alignment to a reference genome

Given that the transcriptome is built from the genome, the most commonly used approach is to use the genome itself as the reference. After the Alignment process we get three types of reads:

- Uniquely mapped reads,

Alignment to a reference genome

Given that the transcriptome is built from the genome, the most commonly used approach is to use the genome itself as the reference. After the Alignment process we get three types of reads:

- Uniquely mapped reads,
- Multiple mapped reads,

Alignment to a reference genome

Given that the transcriptome is built from the genome, the most commonly used approach is to use the genome itself as the reference. After the Alignment process we get three types of reads:

- Uniquely mapped reads,
- Multiple mapped reads,
- Unmapped reads.

Alignment to a reference genome

Given that the transcriptome is built from the genome, the most commonly used approach is to use the genome itself as the reference. After the Alignment process we get three types of reads:

- Uniquely mapped reads,
- Multiple mapped reads,
- Unmapped reads.

Aligner tools can also make use of paired-end reads.

Alignment to a reference genome

Given that the transcriptome is built from the genome, the most commonly used approach is to use the genome itself as the reference.
After the Alignment process we get three types of reads:

- Uniquely mapped reads,
- Multiple mapped reads,
- Unmapped reads.

Aligner tools can also make use of paired-end reads.
A *Paired-end read* is a read that has been sequenced from each end of the fragment (usually called *R1* and *R2*).

Alignment to a reference genome

Given that the transcriptome is built from the genome, the most commonly used approach is to use the genome itself as the reference.
After the Alignment process we get three types of reads:

- Uniquely mapped reads,
- Multiple mapped reads,
- Unmapped reads.

Aligner tools can also make use of paired-end reads.
A *Paired-end read* is a read that has been sequenced from each end of the fragment (usually called *R1* and *R2*).
Paired-end reads are used to increase the mapping accuracy and to reduce the number of multiple mapped reads.

Alignment to a reference genome

- The reference is never a perfect representation of the actual source.

Alignment to a reference genome

- The reference is never a perfect representation of the actual source.
- Many different *biological variations* might be present such as: SNPs (Single-Nucleotide Polymorphism), indels, structural variations and rearrangements.

Alignment to a reference genome

- The reference is never a perfect representation of the actual source.
- Many different *biological variations* might be present such as: SNPs (Single-Nucleotide Polymorphism), indels, structural variations and rearrangements.
- Reads can be affected by sequencing errors.

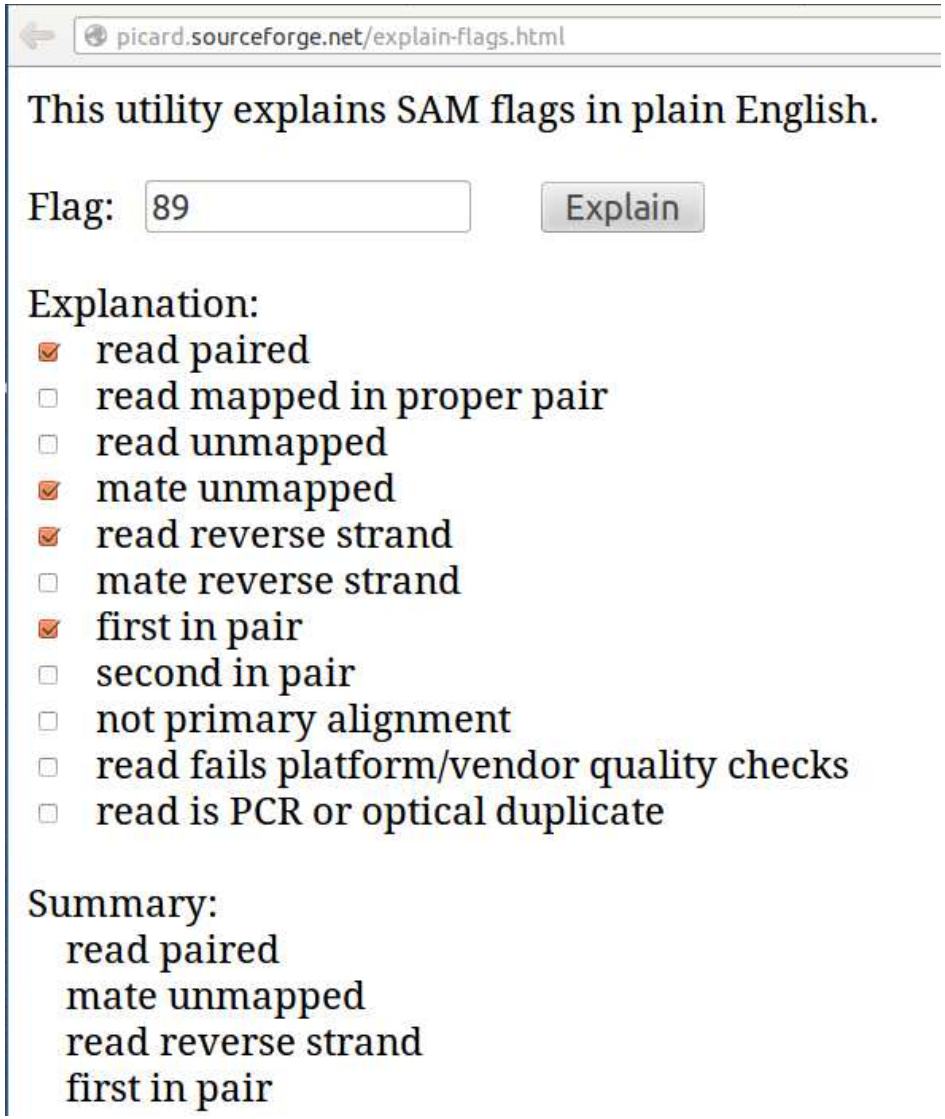
Alignment to a reference genome

- The reference is never a perfect representation of the actual source.
- Many different *biological variations* might be present such as: SNPs (Single-Nucleotide Polymorphism), indels, structural variations and rearrangements.
- Reads can be affected by sequencing errors.
- Therefore, the real task is to find the location where each short read best matches the reference, while allowing for errors and structural variations.

Alignment output: SAM format

```
HWI-1KL111:53:D176AACXX:1:1314:7404:89056       89      chr1    10535   3       75M     *       0       0
GTACCACCGAAATCTGTGCAGAGGAGAACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAAC     BAB@<2BBBBBABBBBABBBBABA@?
<7BBBBB=>@EFBIIFBFECCFGD?AEFIFIIFEIIFFDFFBDA=4+1=       AS:i:-5 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1
MD:Z:25C49      YT:Z:UU NH:i:2  CC:Z:chr15      CP:i:102520475  HI:i:0

HWI-1KL111:53:D176AACXX:1:2105:1895:21247       419     chr1    11650   3       75M     =       11718   143
TGGATTTTTGCCAGTCTAACAGGTGAAGCCCTGGAGATTCTTATTAGTGATTTGGGCTGGGGCCTGGCCATGTNN     ?=?BDDDAF@<DFGFDBGIHIJGEFH<<FE
F3:CGIA9??4?9B*?EGHGIGHIBA@@E;;EECA@?###        AS:i:-2 XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:73G0T0
YT:Z:UU NH:i:2  CC:Z:chr15      CP:i:102519446  HI:i:0

HWI-1KL111:53:D176AACXX:1:2210:10686:50410      329     chr1    11672   1       75M     *       0       0
GTGAAGCCCTGGAGATTCTTATTAGTGATTTGGGCTGGGGCCTGGCCATGTGTATTTTTTTAAATTTCCACTGAT
CC@FFFFFHHHHHJHJIIJJIJJJJHIIJJJJJJJJJJJJJJJIJJJHIJGIGGIJJJJJHHFFFFFFFEEEEEE     AS:i:0  XN:i:0  XM:i:0  XO:i:0
XG:i:0  NM:i:0  MD:Z:75 YT:Z:UU NH:i:3  CC:Z:chr15      CP:i:102519424  HI:i:0

HWI-1KL111:53:D176AACXX:1:1106:4169:92324       419     chr1    11685   3       75M     =       11790   180
GATTCTTATTAGTGATTTGGGCTGGGGCCTGGCCATGTGTATTTTTTTAAATTTCCACTGATGATTTTGCTGCAT     B?
@DBDDDHHBDAFGIIJJIEHJJJE<GGHGGBH@?D?DGHFGEGD@AHGIHGGGDGIIICEH:?EHFBDF@###      AS:i:0  XN:i:0  XM:i:0  XO:i:0
XG:i:0  NM:i:0  MD:Z:75 YT:Z:UU NH:i:2  CC:Z:chr15      CP:i:102519411  HI:i:0

HWI-1KL111:53:D176AACXX:1:1303:20062:7739       419     chr1    11708   3       75M     =       11771   138
GGGGCCTGGCCATGTGTATTTTTTTAAATTTCCACTGATGATTTTGCTGCATGGCCGGTGTTGAGAATGACTGCN
@<BBFFFDHFHB;ECGDGGHGIGGGEFBFHIGGGHII<FH@FEGHGBGGDFFHJICHGHHEFB@C;A@#######     AS:i:-1 XN:i:0  XM:i:1  XO:i:0
XG:i:0  NM:i:1  MD:Z:74G0       YT:Z:UU NH:i:2  CC:Z:chr15      CP:i:102519388  HI:i:0

HWI-1KL111:53:D176AACXX:1:2105:1895:21247       339     chr1    11718   3       75M     =       11650   -143
CATGTGTATTTTTTTAAATTTCCACTGATGATTTTGCTGCATGGCCGGTGTTGAGAATGACTGCGCAAATTTGCC     CCCCCCC?
9<@?;=EHHDHCEE@CIGIHCCDHFB2HDGAFDD<G@F?FCC@DHAFHBGGAC8:AFHFDDDDD@@<     AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0
NM:i:0  MD:Z:75 YT:Z:UU NH:i:2  CC:Z:chr15      CP:i:102519378  HI:i:0
```

Flags


picard.sourceforge.net/explain-flags.html
This utility explains SAM flags in plain English.
Flag: 89
Explain
Explanation:
read paired
read mapped in proper pair
read unmapped
mate unmapped
read reverse strand
mate reverse strand
first in pair
second in pair
not primary alignment
read fails platform/vendor quality checks
read is PCR or optical duplicate
Summary:
read paired
mate unmapped
read reverse strand
first in pair

Coverage as a measure of Expression

Coverage as a measure of Expression

- *Coverage* describes the number of reads that align to a known sequence (reference).

Coverage as a measure of Expression

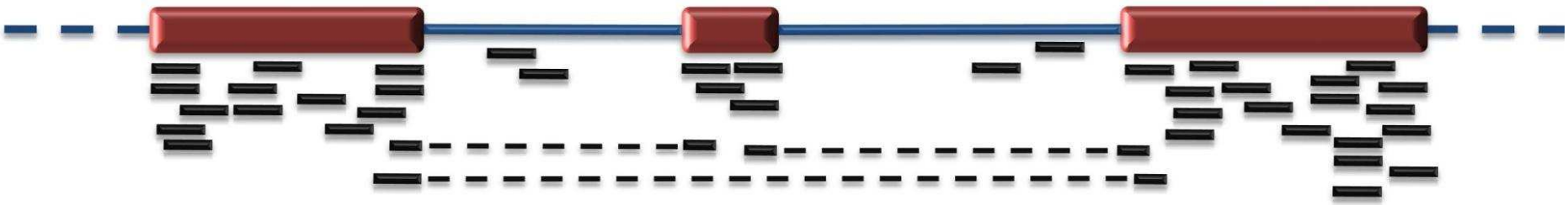
- *Coverage* describes the number of reads that align to a known sequence (reference).
- *Coverage* is used as measure of either gene or isoform expression.

Coverage as a measure of Expression

- *Coverage* describes the number of reads that align to a known sequence (reference).
- *Coverage* is used as measure of either gene or isoform expression.
- Expression is usually measured in Reads Per Kilobase of transcript per Million mapped reads (RPKM).

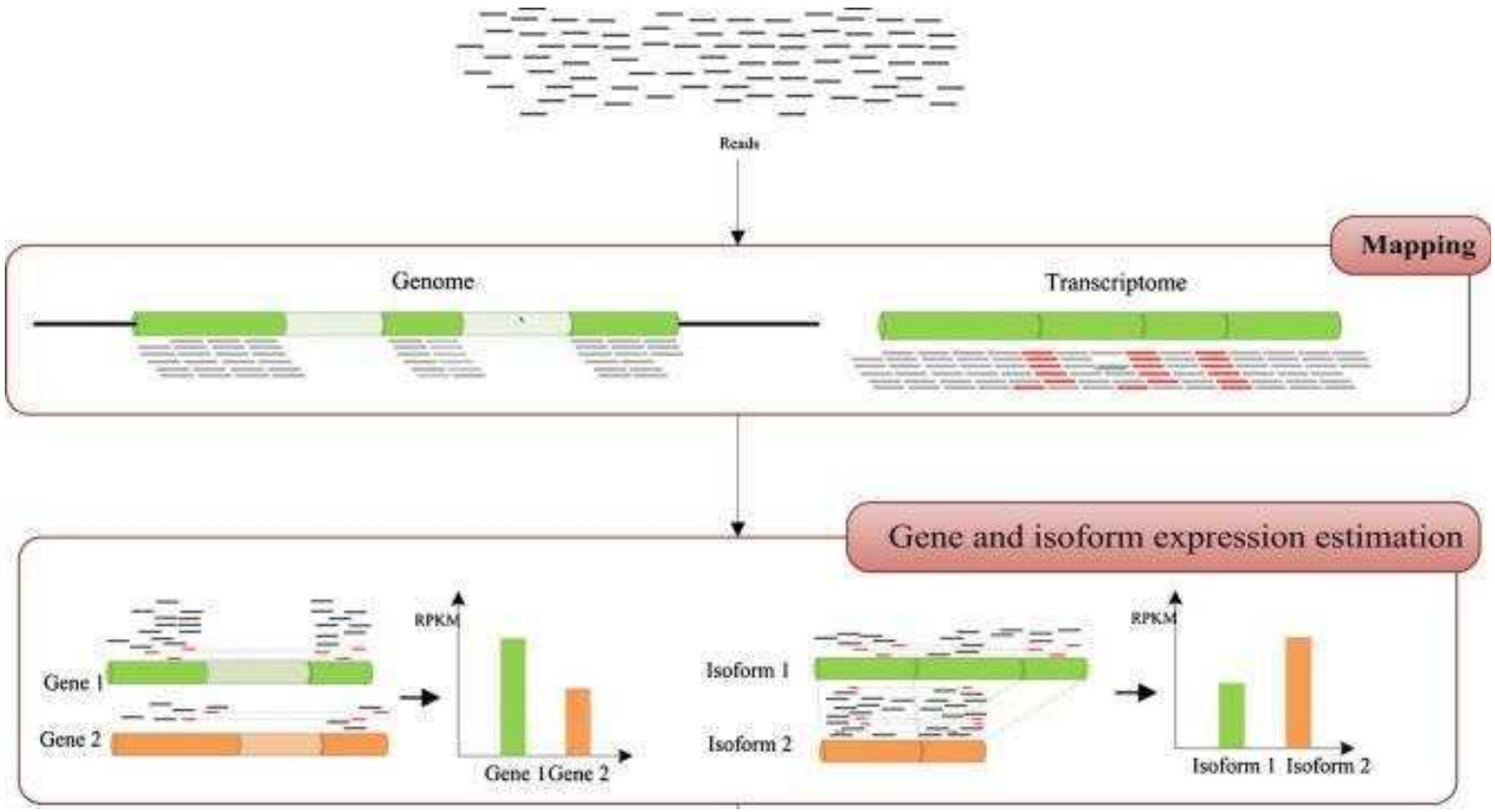
Coverage as a measure of Expression

- *Coverage* describes the number of reads that align to a known sequence (reference).
- *Coverage* is used as measure of either gene or isoform expression.
- Expression is usually measured in Reads Per Kilobase of transcript per Million mapped reads (RPKM).



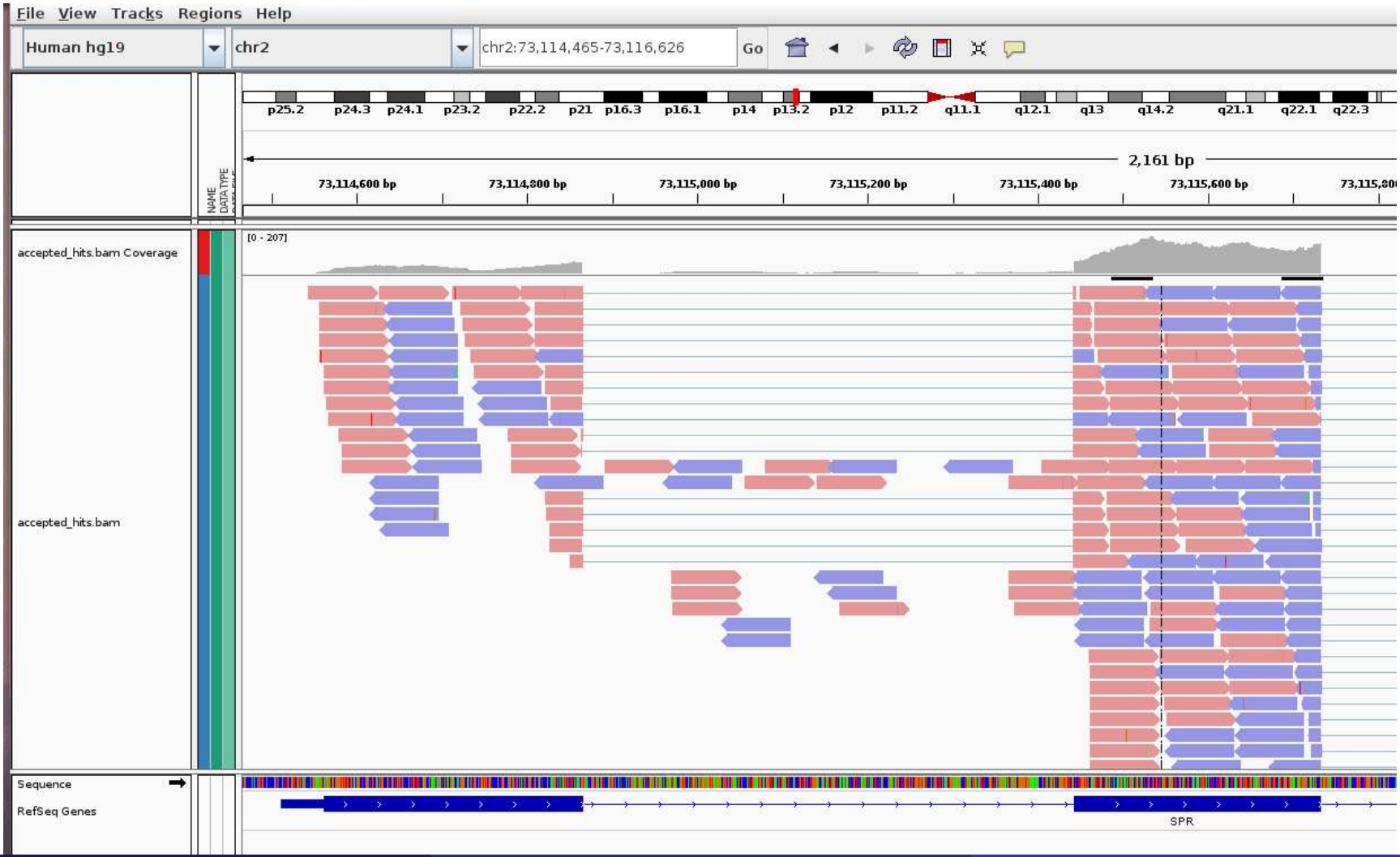
A pipeline to study gene expression and isoform expression

One of the most commonly used workflows is to map reads with a tool like *Tophat* and then use a tool like *HTSeq* to count the number of reads overlapping a gene.

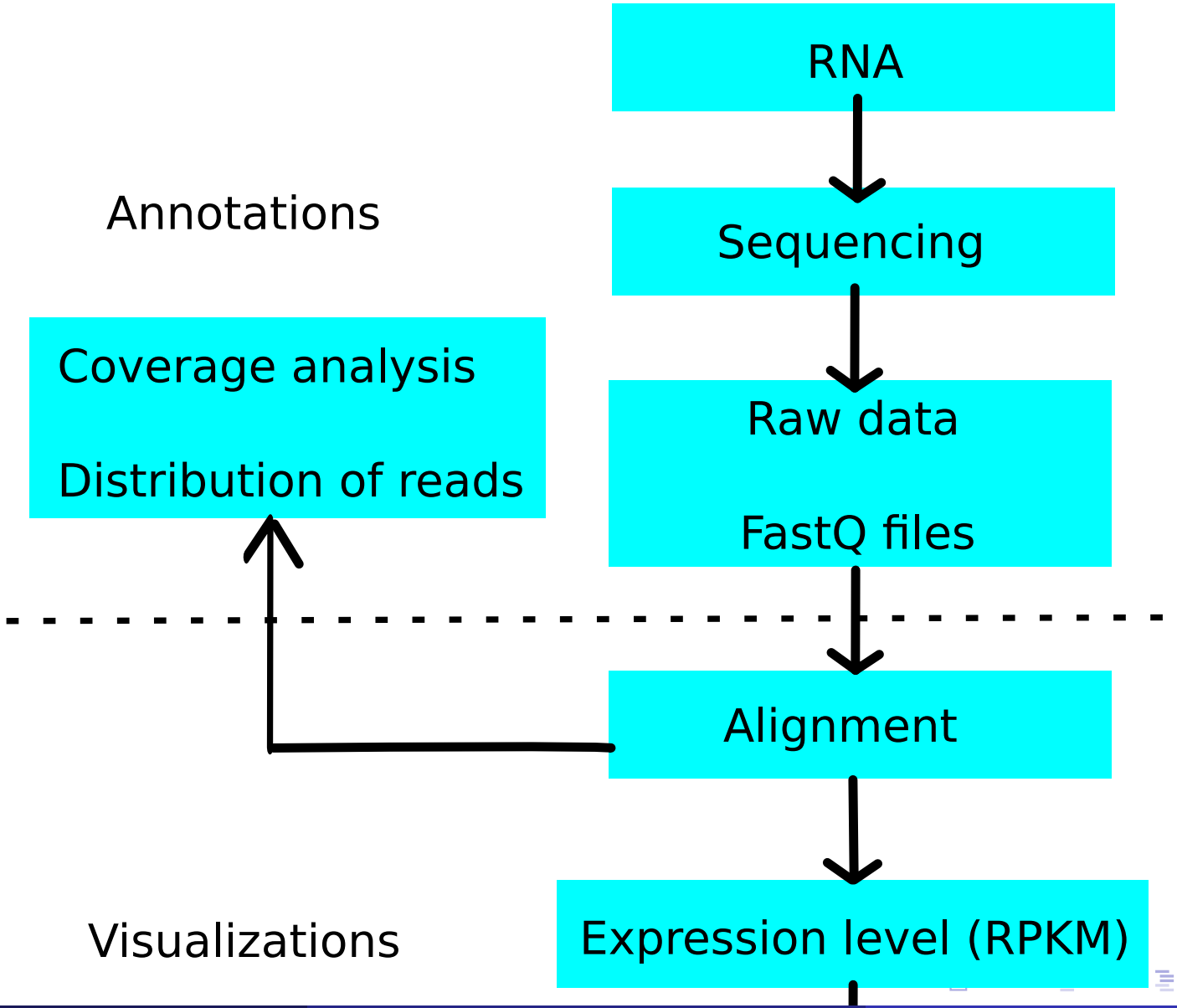


Integrated Genome Viewer

Some reads fall within junctions. First half falls on one exon and the second half on the other exon.



Overview of RNA-SEQ analysis workflow



Common goals of RNA-Seq analysis

RNA-SEQ provides information on:

Common goals of RNA-Seq analysis

RNA-SEQ provides information on:

- differential expression of genes,

Common goals of RNA-Seq analysis

RNA-SEQ provides information on:

- differential expression of genes,
- differently spliced transcripts,

Common goals of RNA-Seq analysis

RNA-SEQ provides information on:

- differential expression of genes,
- differently spliced transcripts,
- non-coding RNAs,

Common goals of RNA-Seq analysis

RNA-SEQ provides information on:

- differential expression of genes,
- differently spliced transcripts,
- non-coding RNAs,
- post-transcriptional mutations,

Common goals of RNA-Seq analysis

RNA-SEQ provides information on:

- differential expression of genes,
- differently spliced transcripts,
- non-coding RNAs,
- post-transcriptional mutations,
- gene fusions.

Computational Challenges

The computational challenges fall into three main categories:

Computational Challenges

The computational challenges fall into three main categories:

- building fast and reliable read aligners,

Computational Challenges

The computational challenges fall into three main categories:

- building fast and reliable read aligners,
- transcriptome reconstruction (genome-guided, genome-independent),

Computational Challenges

The computational challenges fall into three main categories:

- building fast and reliable read aligners,
- transcriptome reconstruction (genome-guided, genome-independent),
- expression quantification (gene or isoform quantification, differential expression).

Questions of Interest

Questions of Interest

RNA-SEQ experiments allow us to answer several questions about sequenced transcripts.

Questions of Interest

RNA-SEQ experiments allow us to answer several questions about sequenced transcripts.

- Expression levels (from counts over exons and junctions)

Questions of Interest

RNA-SEQ experiments allow us to answer several questions about sequenced transcripts.

- Expression levels (from counts over exons and junctions)
- Structure (e.g.: isoforms from the alignments)

Questions of Interest

RNA-SEQ experiments allow us to answer several questions about sequenced transcripts.

- Expression levels (from counts over exons and junctions)
- Structure (e.g.: isoforms from the alignments)
- Variants (e.g.: *Single-Nucleotide Polymorphism*, *indels*)

Transcriptome Analysis

Transcriptome investigations are usually performed for the following tasks:

Transcriptome Analysis

Transcriptome investigations are usually performed for the following tasks:

- Characterize the full genome of an organism.

Transcriptome Analysis

Transcriptome investigations are usually performed for the following tasks:

- Characterize the full genome of an organism.
- Discover unknown genes (possibly also non-coding ones).

Transcriptome Analysis

Transcriptome investigations are usually performed for the following tasks:

- Characterize the full genome of an organism.
- Discover unknown genes (possibly also non-coding ones).
- Compare genes across organisms.

Transcriptome Analysis

Transcriptome investigations are usually performed for the following tasks:

- Characterize the full genome of an organism.
- Discover unknown genes (possibly also non-coding ones).
- Compare genes across organisms.
- Characterize transcript isoforms.

Transcriptome Analysis

Transcriptome investigations are usually performed for the following tasks:

- Characterize the full genome of an organism.
- Discover unknown genes (possibly also non-coding ones).
- Compare genes across organisms.
- Characterize transcript isoforms.
- Discover new alternative isoforms.

Transcriptome Analysis

Transcriptome investigations are usually performed for the following tasks:

- Characterize the full genome of an organism.
- Discover unknown genes (possibly also non-coding ones).
- Compare genes across organisms.
- Characterize transcript isoforms.
- Discover new alternative isoforms.
- Monitor transcriptome changes across tissues or in response to environmental changes.

Transcriptome Analysis

Transcriptome investigations are usually performed for the following tasks:

- Characterize the full genome of an organism.
- Discover unknown genes (possibly also non-coding ones).
- Compare genes across organisms.
- Characterize transcript isoforms.
- Discover new alternative isoforms.
- Monitor transcriptome changes across tissues or in response to environmental changes.
- Identify significant gene expression changes across different environmental conditions.

Transcriptome Analysis

Transcriptome investigations are usually performed for the following tasks:

- Characterize the full genome of an organism.
- Discover unknown genes (possibly also non-coding ones).
- Compare genes across organisms.
- Characterize transcript isoforms.
- Discover new alternative isoforms.
- Monitor transcriptome changes across tissues or in response to environmental changes.
- Identify significant gene expression changes across different environmental conditions.
- Study what is encoded in a genome and how is it processed.

Transcriptome Analysis

Transcriptome investigations are also used for the identification and quantification of transcripts, such as:

Transcriptome Analysis

Transcriptome investigations are also used for the identification and quantification of transcripts, such as:

- Annotated regions stored in classical databases (e.g.: RefSeq, UCSC, Ensembl).

Transcriptome Analysis

Transcriptome investigations are also used for the identification and quantification of transcripts, such as:

- Annotated regions stored in classical databases (e.g.: RefSeq, UCSC, Ensembl).
- Boundary annotations (e.g.: exon boundaries, splicing sites, UTR's).

Transcriptome Analysis

Transcriptome investigations are also used for the identification and quantification of transcripts, such as:

- Annotated regions stored in classical databases (e.g.: RefSeq, UCSC, Ensembl).
- Boundary annotations (e.g.: exon boundaries, splicing sites, UTR's).
- Novel transcribed regions (e.g.: new exons, new isoforms).

Transcriptome Analysis

Transcriptome Analysis

- Since the first step of any RNA-SEQ data analysis is usually the alignment of the reads, all subsequent results and inferences are based on the aligned reads.

Transcriptome Analysis

- Since the first step of any RNA-SEQ data analysis is usually the alignment of the reads, all subsequent results and inferences are based on the aligned reads.
- Alignment is an inferential process.

Transcriptome Analysis

- Since the first step of any RNA-SEQ data analysis is usually the alignment of the reads, all subsequent results and inferences are based on the aligned reads.
- Alignment is an inferential process.
- Therefore, uncertainty should be taken into account.

Transcriptome Analysis

- Since the first step of any RNA-SEQ data analysis is usually the alignment of the reads, all subsequent results and inferences are based on the aligned reads.
- Alignment is an inferential process.
- Therefore, uncertainty should be taken into account.
- Reads do not have the same quality and the mapping can produce single mapped reads, multiple mapped reads and unmapped ones.

Transcriptome Analysis

- Since the first step of any RNA-SEQ data analysis is usually the alignment of the reads, all subsequent results and inferences are based on the aligned reads.
- Alignment is an inferential process.
- Therefore, uncertainty should be taken into account.
- Reads do not have the same quality and the mapping can produce single mapped reads, multiple mapped reads and unmapped ones.

Therefore:

Transcriptome Analysis

- Since the first step of any RNA-SEQ data analysis is usually the alignment of the reads, all subsequent results and inferences are based on the aligned reads.
- Alignment is an inferential process.
- Therefore, uncertainty should be taken into account.
- Reads do not have the same quality and the mapping can produce single mapped reads, multiple mapped reads and unmapped ones.

Therefore:

- How can we take into account all this information?

Transcriptome Analysis

- Since the first step of any RNA-SEQ data analysis is usually the alignment of the reads, all subsequent results and inferences are based on the aligned reads.
- Alignment is an inferential process.
- Therefore, uncertainty should be taken into account.
- Reads do not have the same quality and the mapping can produce single mapped reads, multiple mapped reads and unmapped ones.

Therefore:

- How can we take into account all this information?
- How can we can handle all this information?

Open research topics in Statistics for RNA-SEQ

Statistical analysis usually makes use of Poisson and Negative Binomial distributions.

Open research topics in Statistics for RNA-SEQ

Statistical analysis usually makes use of Poisson and Negative Binomial distributions. These distributions do not take into account such uncertainty about reference and reads.

Open research topics in Statistics for RNA-SEQ

Statistical analysis usually makes use of Poisson and Negative Binomial distributions. These distributions do not take into account such uncertainty about reference and reads.

Novel Probabilistic models for RNA-SEQ are still needed.

De novo transcriptome assembly

De novo transcriptome assembly

- Since a genome contains the sum of all introns and exons that might be present in a transcript, spliced variants that do not align continuously along the genome might be neglected as actual protein isoforms.

De novo transcriptome assembly

- Since a genome contains the sum of all introns and exons that might be present in a transcript, spliced variants that do not align continuously along the genome might be neglected as actual protein isoforms.
- *De novo transcriptome assembly* is the method of creating a transcriptome without the aid of a reference genome.

Questions about Isoform Expression

Questions about Isoform Expression

- Is an isoform expressed?

Questions about Isoform Expression

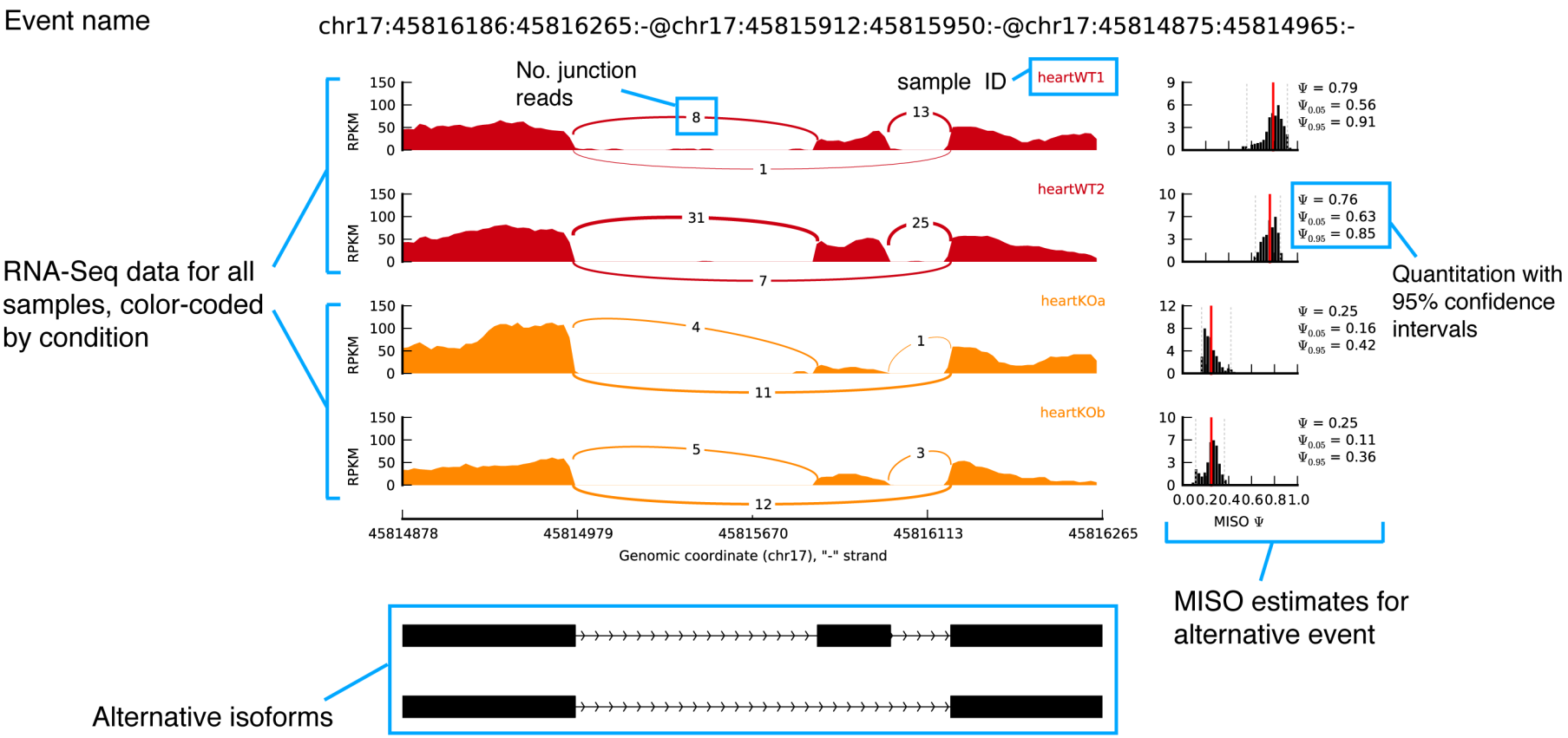
- Is an isoform expressed?
- What is the expression level of a particular isoform, and how can we compare it to that of other isoforms?

Questions about Isoform Expression

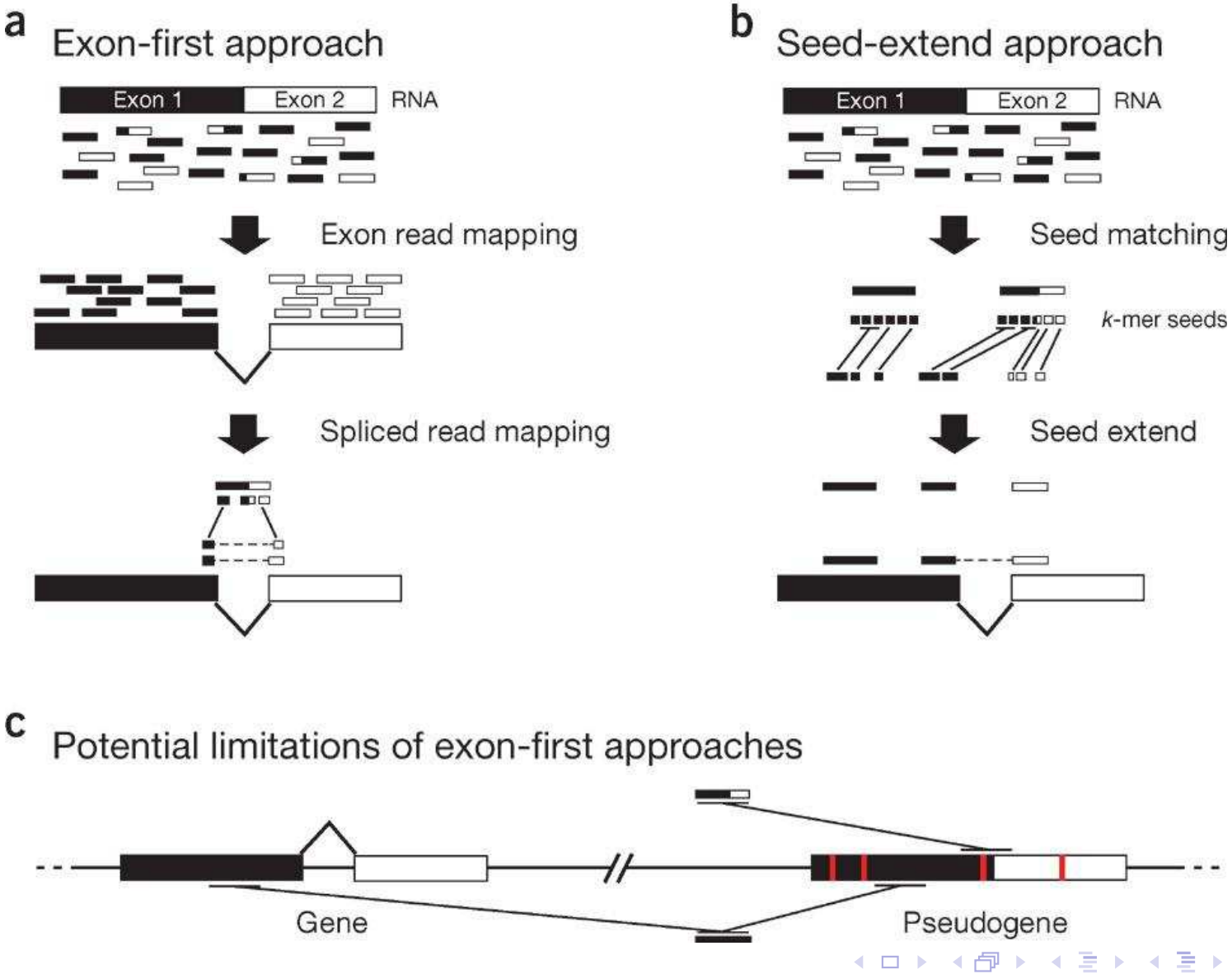
- Is an isoform expressed?
- What is the expression level of a particular isoform, and how can we compare it to that of other isoforms?
- Is the balance in isoform expression different across samples?

Questions about Isoform Expression

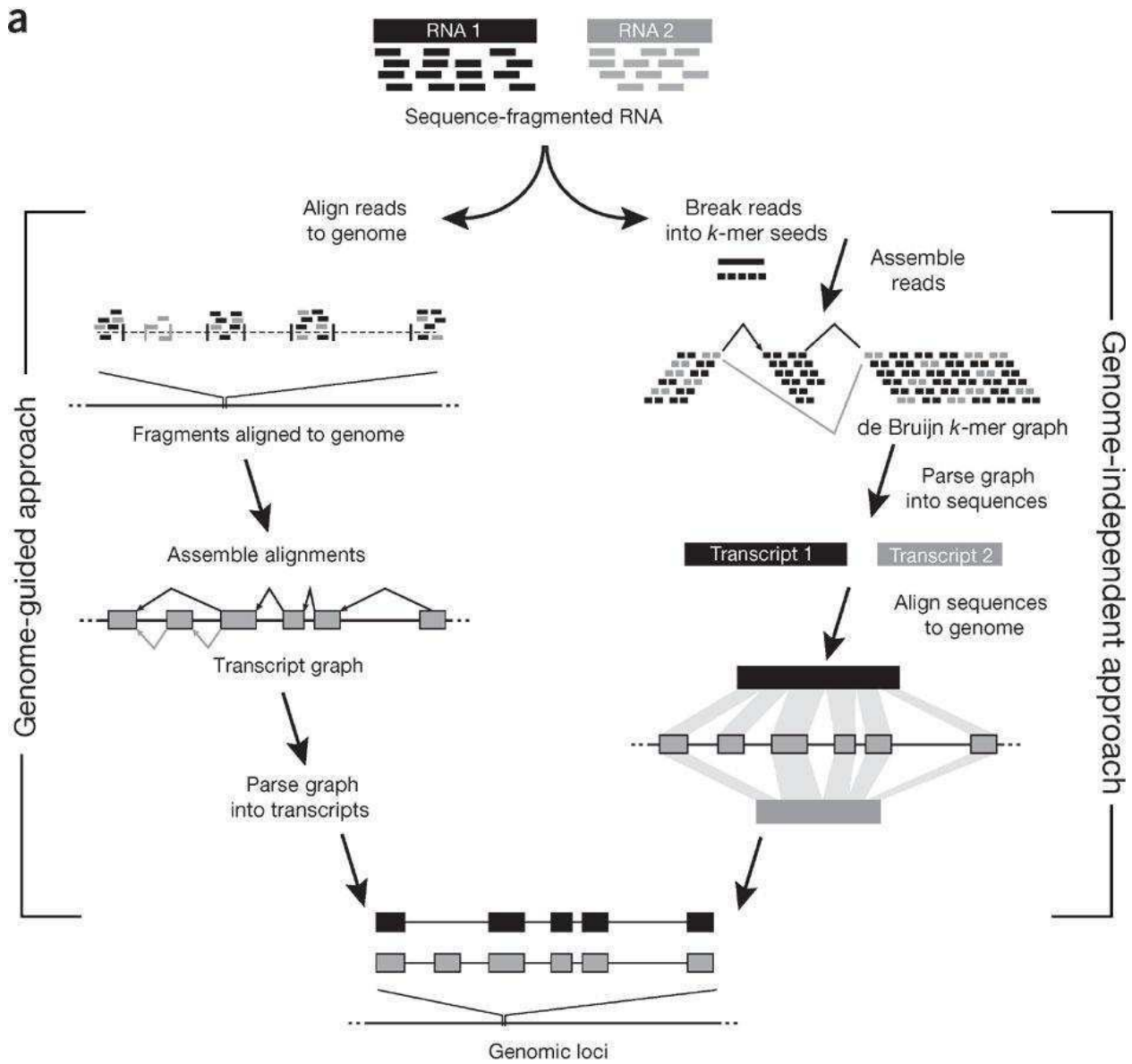
- Is an isoform expressed?
- What is the expression level of a particular isoform, and how can we compare it to that of other isoforms?
- Is the balance in isoform expression different across samples?



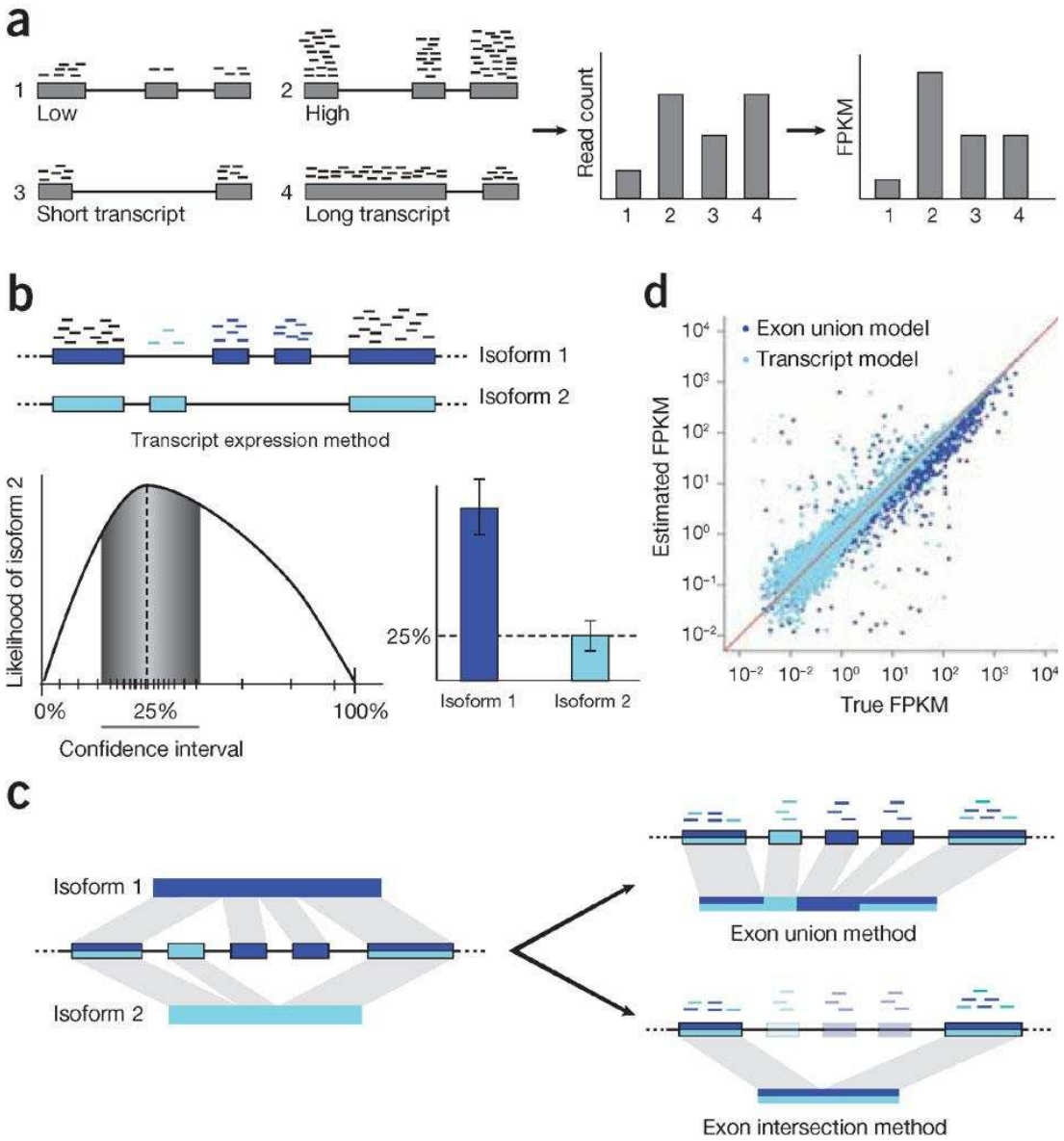
Alignment strategies of RNA-SEQ reads to the genome



Transcriptome reconstruction methods



Gene espression quantification with RNA-SEQ



Graph theory to locally assembly genes/transcripts

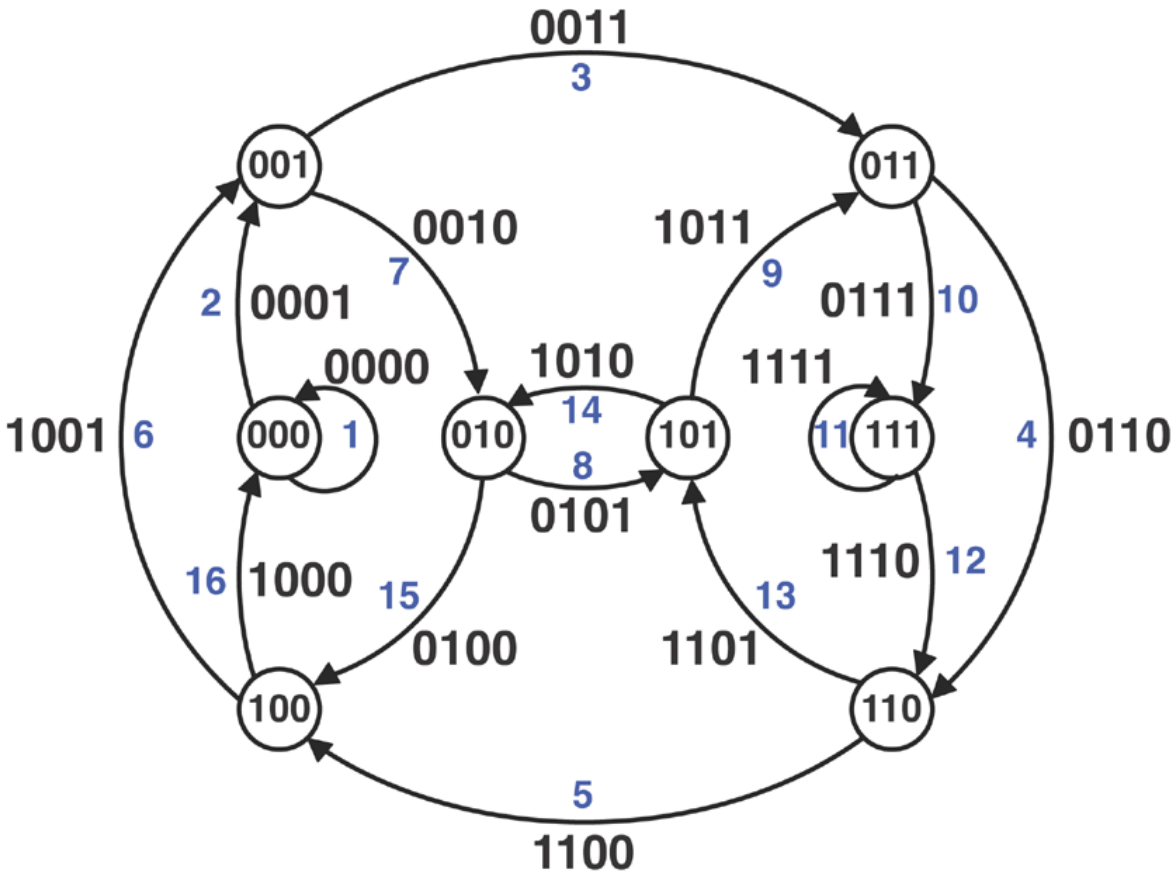
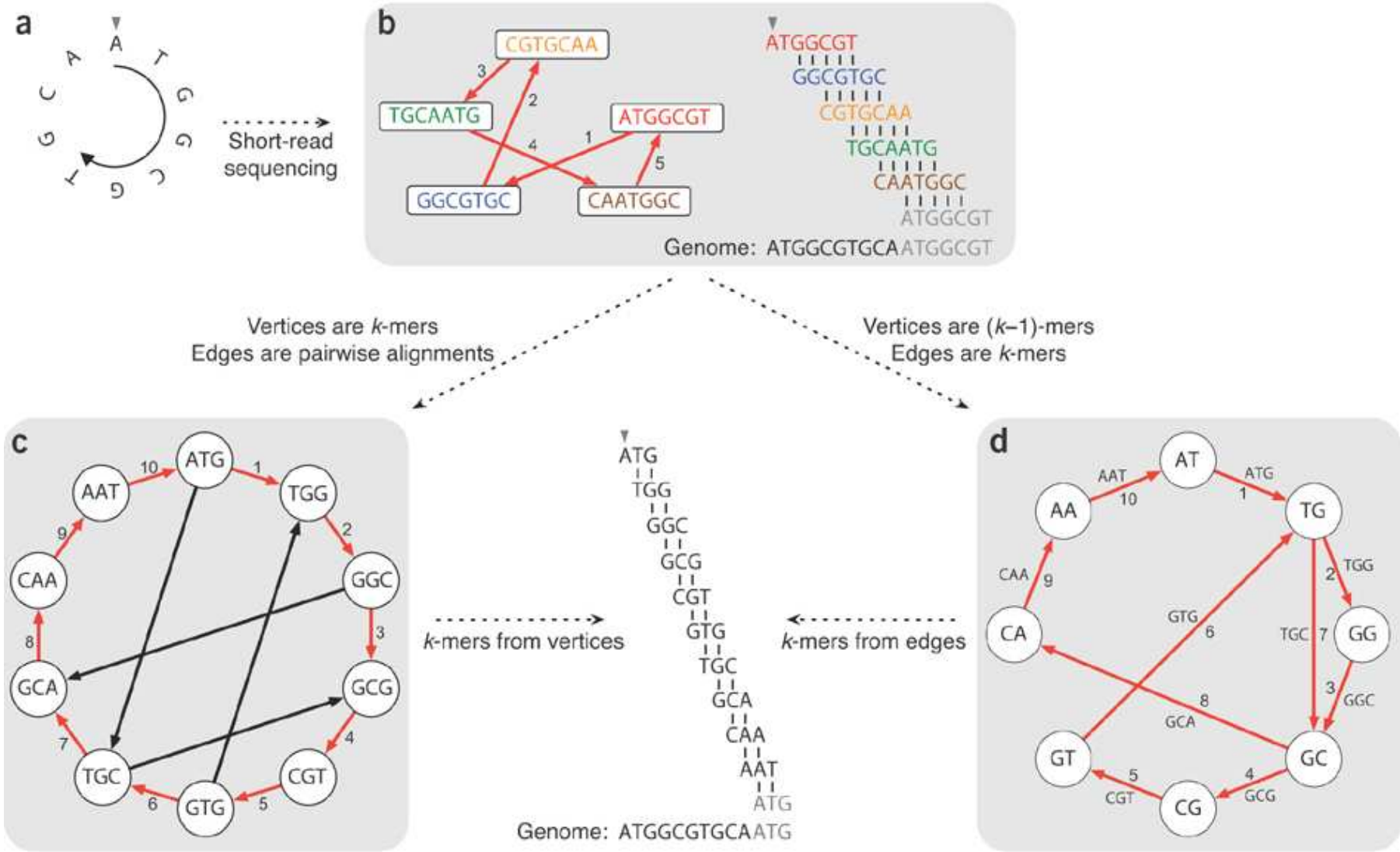
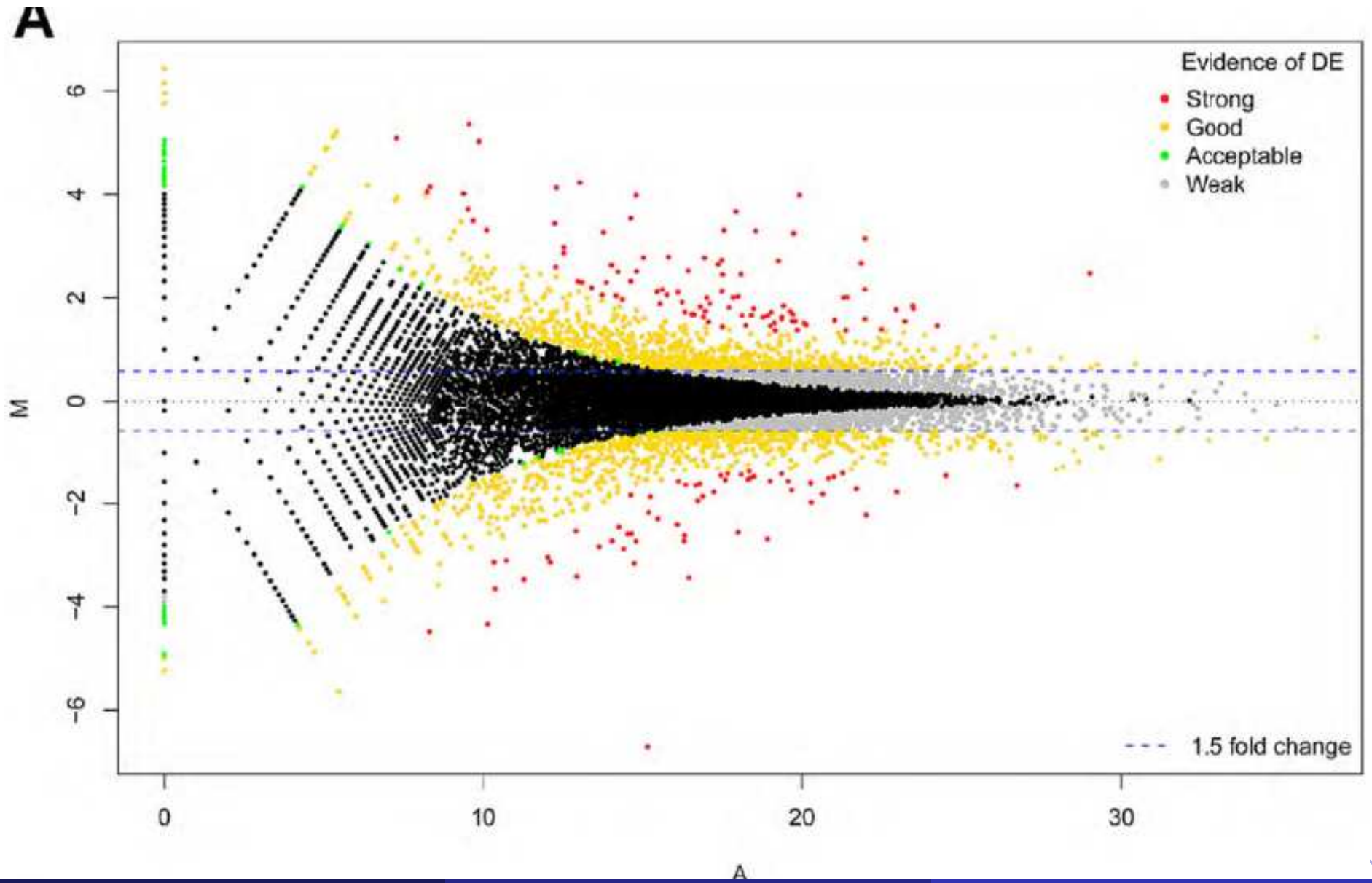


Figure:

Two strategies for genome assembly: from Hamiltonian cycles to Eulerian cycles



Differential Expression



THANK YOU !