RNA-seq Bioinformatics

Vincent Lacroix
LBBE

Outline

- What is RNA-seq ?
- What is seq ?
- What is transcription and splicing ?
- What is the RNA-seq protocol ?
- What new bioinformatics challenges ?

RNA-seq

- RNA-seq: application of deep sequencing technologies to RNA
- Purpose: Identification and quantification of all RNAs present in a sample
- RNA-seq Vs microarrays:
  - better resolution (1nt),
  - better quantification,
  - no a priori knowledge of the sequences,
  - more expensive ?

A short history of sequencing

- 1951: first protein sequence (Insulin, Sanger)
- 1953: Discovery of the double helix structure (Franklin, Watson & Crick)
- 1970: Global alignment of 2 sequences (Needleman & Wunsch)
- 1977: Development of techniques for large scale DNA sequencing (Sanger, Maxam, Gilbert)
- 1984: Creation of the first sequence databanks, ACNUC, PIR, EMBL, GenBank

A short history of sequencing

- 1984: First genome sequenced: EBV virus (170 Kb)
- 1990: Initiation of the human genome sequencing project
- 1996: First eukaryotic genome sequenced (S. cerevisaie, 12Mb)
- 2001: First mammalian genome (Homo Sapiens, 3Gb)
- 2005: New sequencing protocols: Solexa, 454, SOLiD
- Today:
  - Genbank: 106 533 156 756 bases in 108 431 692 sequences
  - Over 2000 complete genomes, still counting

Deep sequencing (or Next-Generation Sequencing)

- Characteristics:
  - Millions of reads
  - Short length
  - Larger error rates

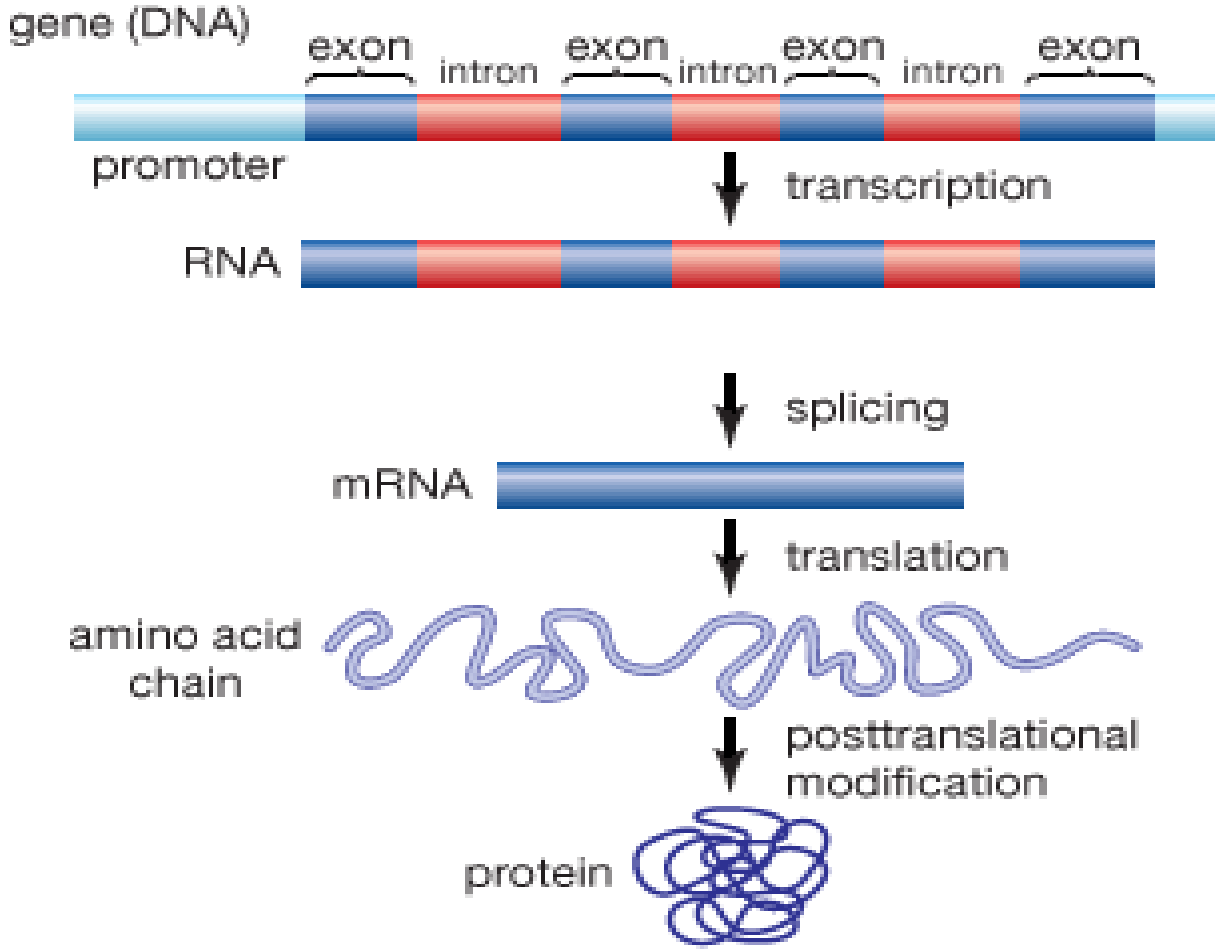
- Main platforms:
  - Illumina (100M, 100nt, 0.1%)
  - Pacbio (1M, 10000nt, 10%)
  - Oxford Nanopore (1M, 10000nt, 10 %)

Applications of NGS

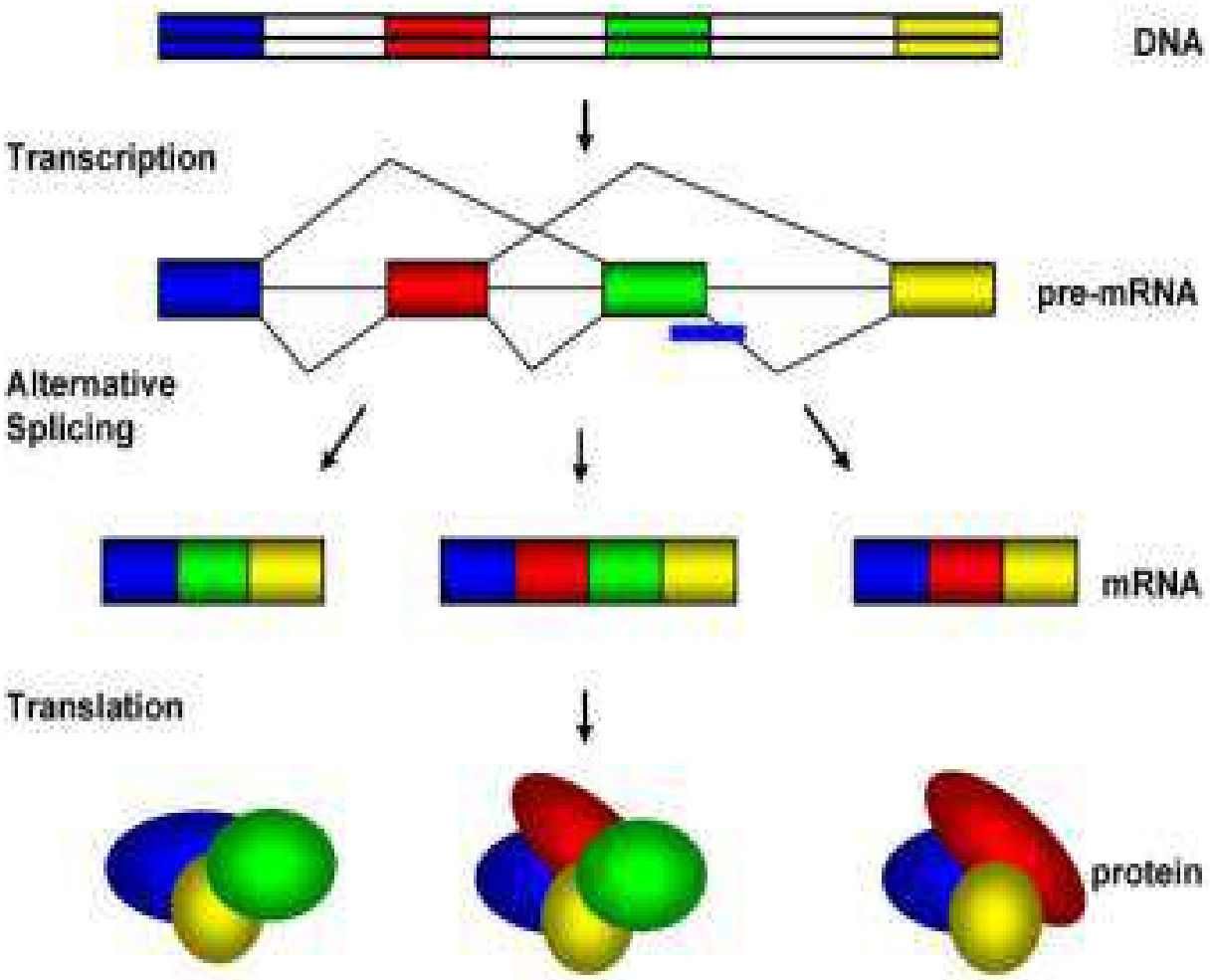
- De novo genome sequencing
- Genome resequencing (polymorphism: SNPs, rearrangements)
- Metagenomics
- Transcriptome sequencing :RNA-seq (genome (re-)annotation, alternative splicing)
- Protein-DNA interaction : Chip-seq
- Protein-RNA interaction : CLIP-seq
- DNA-DNA interaction : Hi-C

Transcription and Splicing

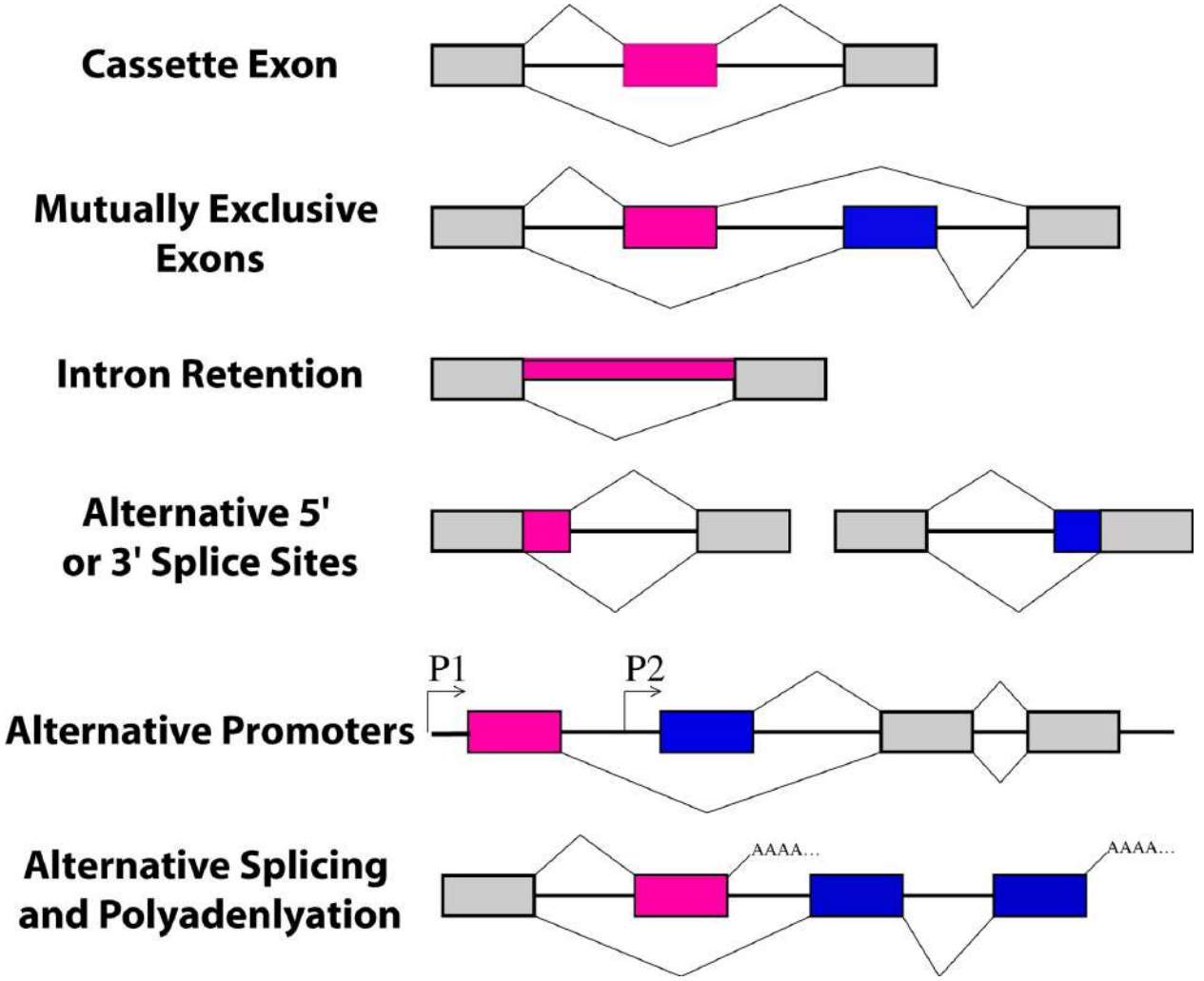
Transcription and Splicing



Alternative splicing (strict)



Alternative splicing (general)



Alternative splicing (definition)

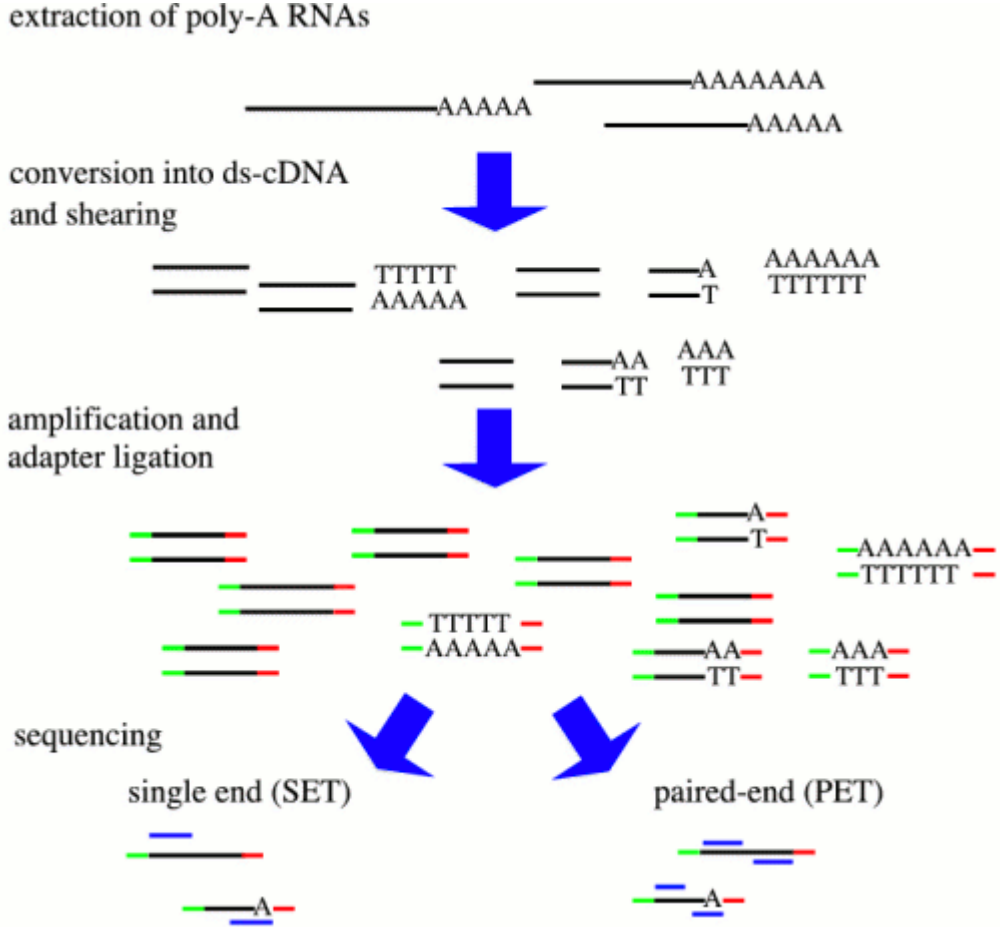
- Alternative splicing is a process by which the exons of a pre-mRNA produced by transcription of a gene are reconnected in multiple ways during RNA splicing.
- Alternative transcription start (alternative promoters)
- Alternative transcription end (alternative polyAdenylation)

RNA-seq and splicing

- RNA-seq enables to characterise all the variations in
  - transcription initiation,
  - transcription termination (polyadenylation)
  - splicing

RNA-seq protocol(s)

RNAseq protocol



Bioinformatics of RNA-seq

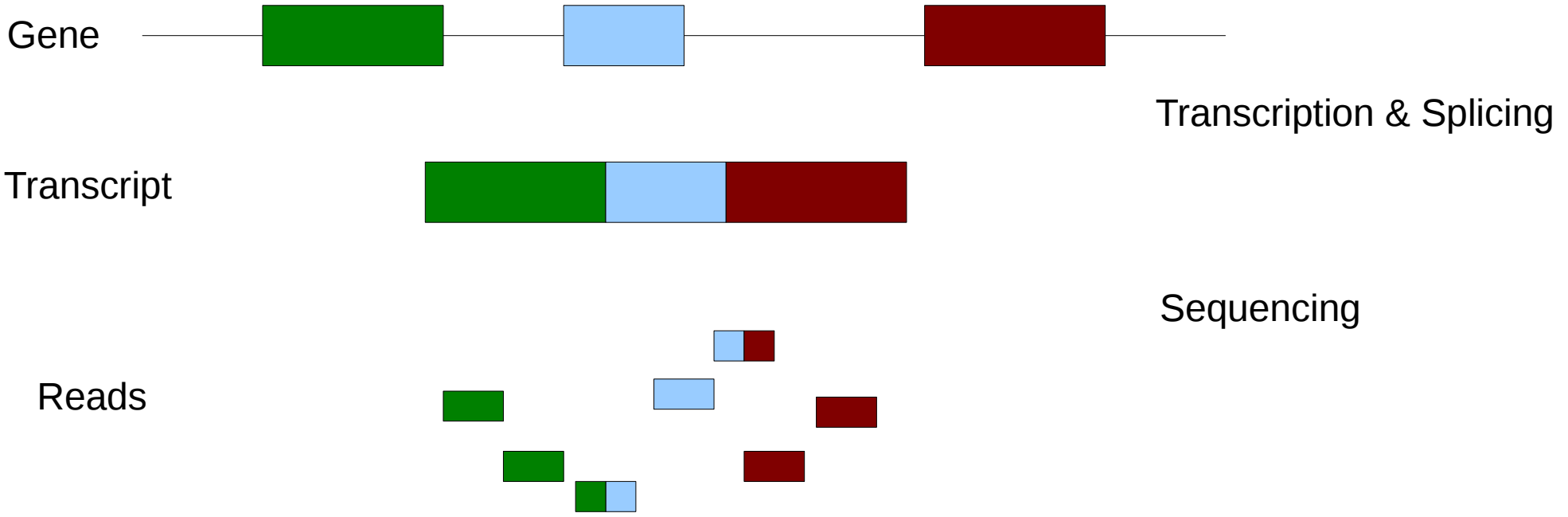
RNA-seq Bioinformatics

- Step 1 : Mapping of RNA-seq reads to a reference genome
- Step 2 : Identify and quantify exons, exon junctions, loci and transcripts
- Step 3 : Call differentially expressed genes, differentially spliced genes

Mapping

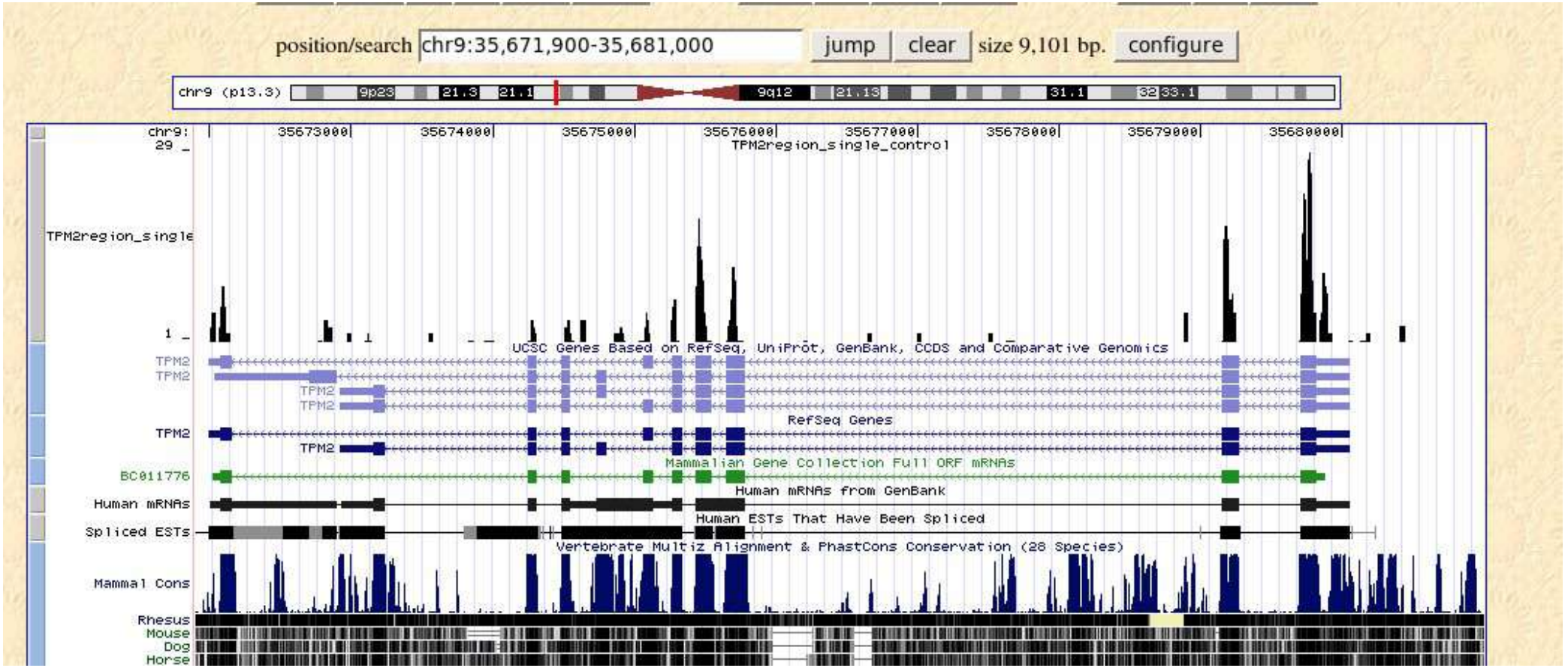
- Two main algorithmic approaches
  - Preprocessing of the reads
    - build an index (trie, hash table) of the reads
    - scan the genome once
  - Preprocessing of the genome
    - Build a index (suffix tree, suffix array, bwt) of the genome
    - Query each read
- Most short read mappers (bwa, bowtie, gem) rely on indexing the genome with a burrows-wheeler transform

Exonic reads and Junction Reads



Exonic read : the whole read in included in one exon
Junction read : the read spans more than one exon

Mapping reads to the genome



- Only exonic reads can be mapped

Junction reads

- Junction reads will not be mapped to the genome
- Two situations can be distinguished :
  - The genome is annotated
    - The reads can be mapped to known exon junctions, or to a reference transcriptome
  - The genome is not annotated
    - The reads have to be split-mapped

Mapping to known exon junctions

- Construct a set of exon junctions of half length L-MM-2
  - L: read length
  - MM: number of mismatches allowed
- Rationale
  - If a read maps to the junction, it will have at least 2 matched bases on each side of the junction
- Example
  - L=36, MM=2 => half length=32
  - The exon junction will contain 32 bases of the upstream exon and 32 bases of the downstream exon

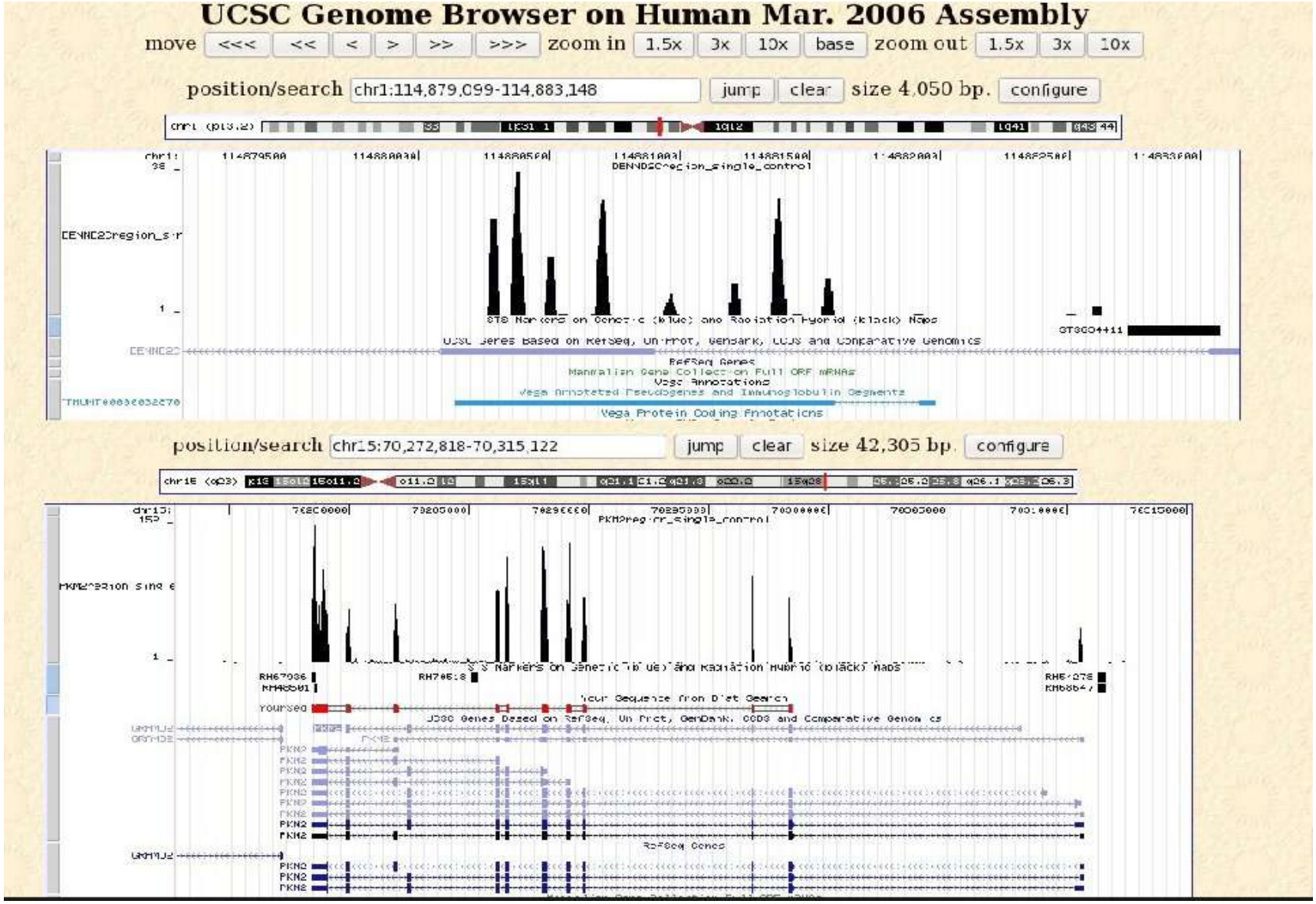
Mapping to known exon junctions

- A read may map both to the genome and to an exon junction, hence it is important to map to the genome AND the exon junctions at the same time

The case of processed pseudogenes

- A processed pseudogene originates from the inclusion of a processed transcript in the genome. Being non functional, the pseudogene starts accumulating mutations.
- In an RNAseq experiment, if the parent transcript is expressed, it will generate reads from its exons, and from its exon junctions
- Mapping the reads to the genome, exonic reads will map to the correct location, but junction reads will map to the processed pseudogene

The case of processed pseudogenes



The case of processed pseudogenes

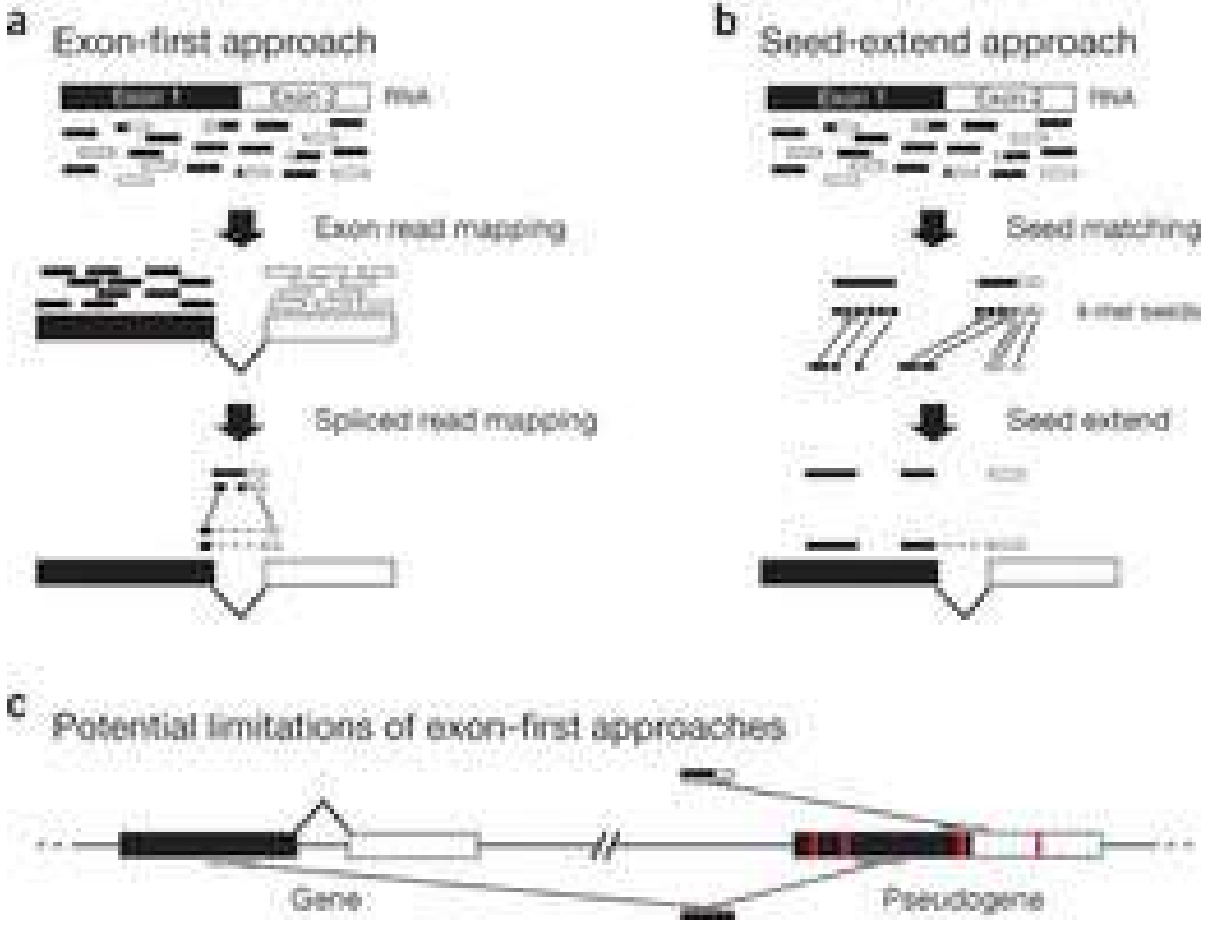
- The worse solution is to map the reads first to the genome, then to exon junctions. This leads to map all junction reads of the parent gene to the pseudogene. Beware ! This is the default behaviour of exon-first mappers like TopHat.
- The not-so-bad solution is to map the reads first to annotated exon junctions, and then to the genome. This is the behaviour of most mappers IF YOU PROVIDE THE GTF OPTION.
- The ideal solution would be to map the reads both to the genome and to exon junctions. This enables to deal with the situation where both the parent gene and the pseudogene are expressed. This happens and is probably underestimated.

Identifying new exon junctions

- Split-map reads to the genome
- Use a priori knowledge of the splice sites (GT-AG

- Available software :
  - STAR
  - TopHat
  - HiSat2

Splice alignment



Mapping to the transcriptome

- Enables to identify known exons and known exon junctions
- Issue of multi mapping in the case of alternative splicing
- Assume that the annotation is complete

Mapping summary

- Mapping to a reference genome will enable to identify known and new exons
- Mapping to a reference transcriptome will enable to identify known exons and known exon junctions
- Identifying new exon junctions is challenging

Quantifying exons, loci and transcripts

- Quantifying exons, junctions and loci is easy
- Quantifying transcripts is much harder

Quantifying exons, junctions and loci

- Number of reads
  - Issue : Long genes have higher expression values
- Number of reads per kb
  - Issue : Different experiments cannot be compared
- Number of reads per kb per experiment size
- RPKM : read per kb per million mapped read
- FPKM : fragment per kb per million mapped read

Issue of paralog genes

- For a family of paralog genes, the same read may map to multiple locations.
- Ignoring reads that map to multiple locations will result in an under-estimation of the expression level of each gene
- Keeping all locations for each read will result in over-estimating the expression level of each gene
- Randomly chosing one location for each read implicitely assumes that all genes are equally expressed. THIS IS THE SOLUTION IMPLEMENTED IN MOST MAPPERS.
- One solution : restrict the analysis to the mapable genome.
- Issue : this depends on the read size
- Alternative solution : use dowstream software which can handle multi-mapping reads (ex : mmquant)

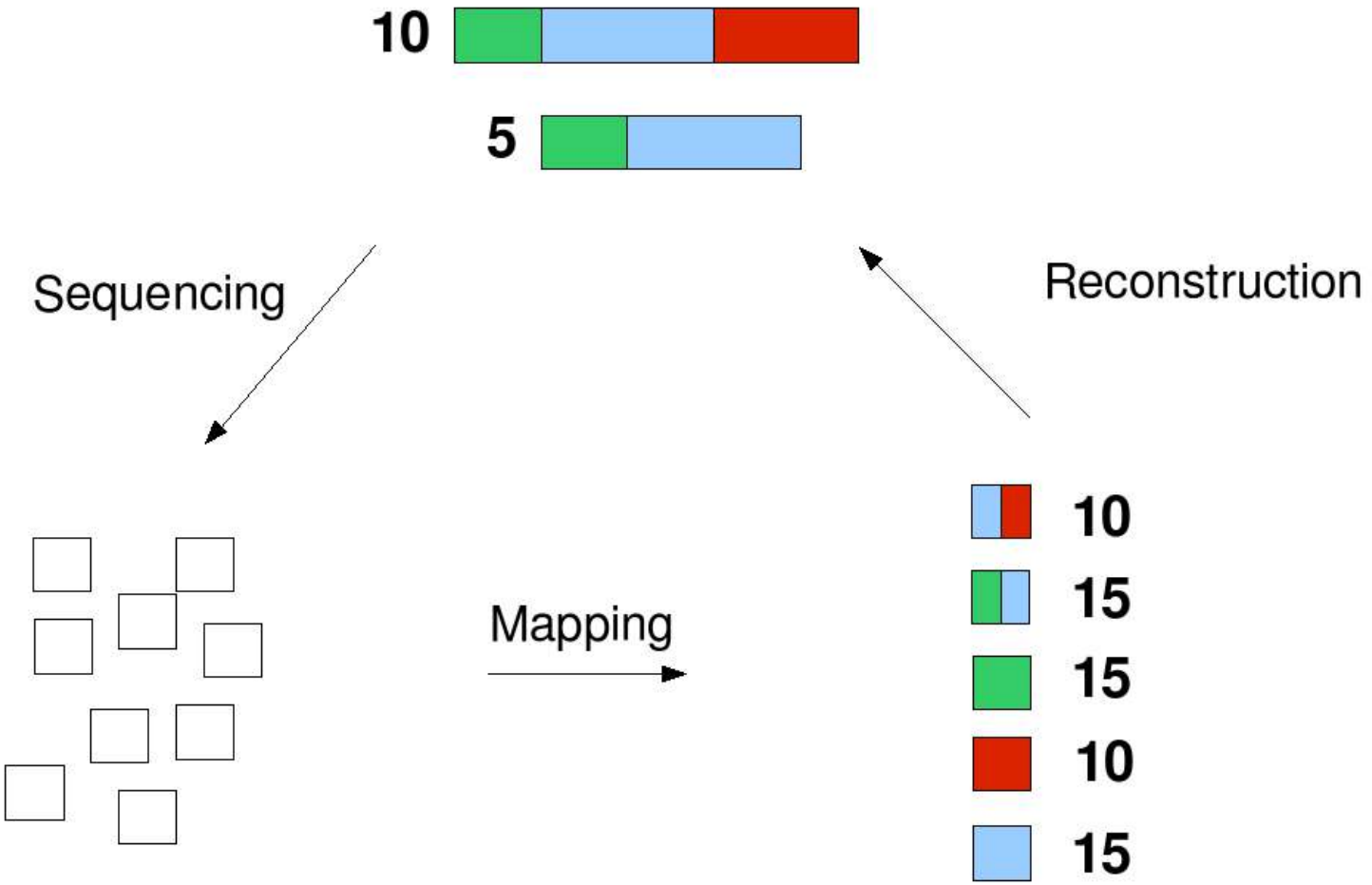
Transcript quantification

- Once exons and exon junctions have been identifed and quantified, how to identify and quantify the full length transcripts ?

Transcript reconstruction

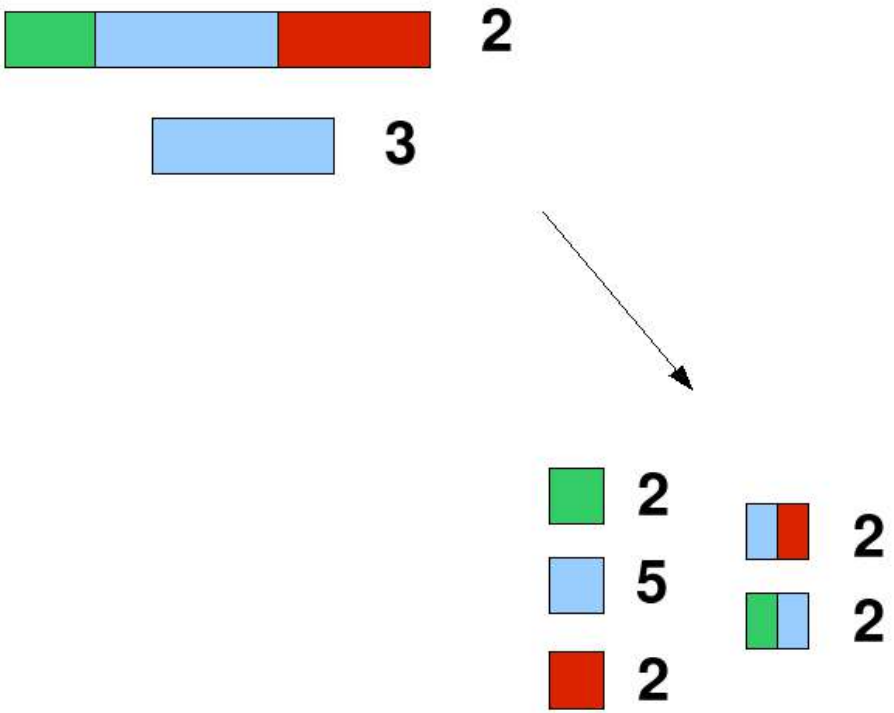
- Is it theoretically possible to assign a unique expression value to each transcript ?
- Is there always a unique solution to the problem of transcriptome reconstruction ?
- Assumptions:
  - Large coverage
  - Reads are uniformly sampled from 5'-3'
  - Mapping is solved

Overview of the problem

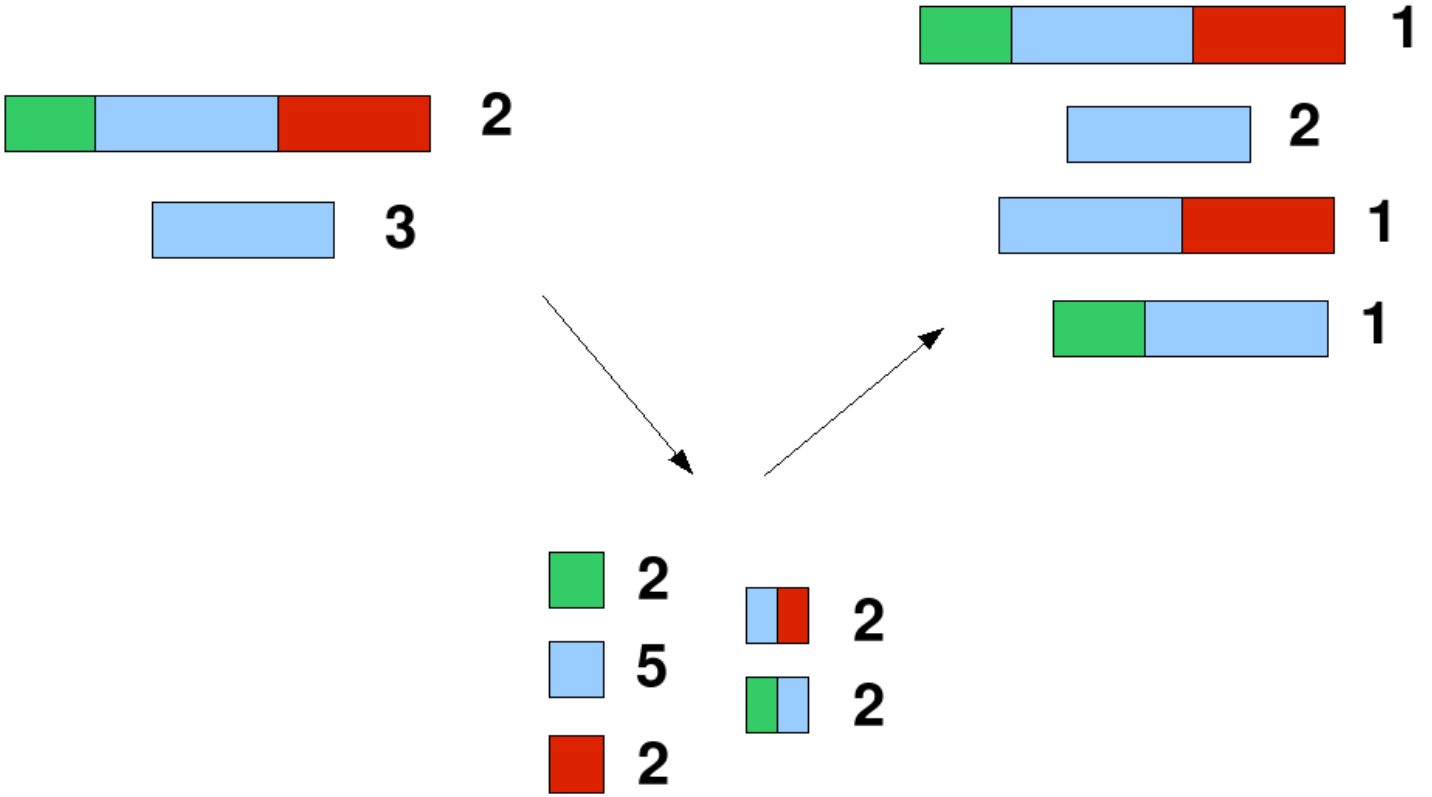


Example

Example



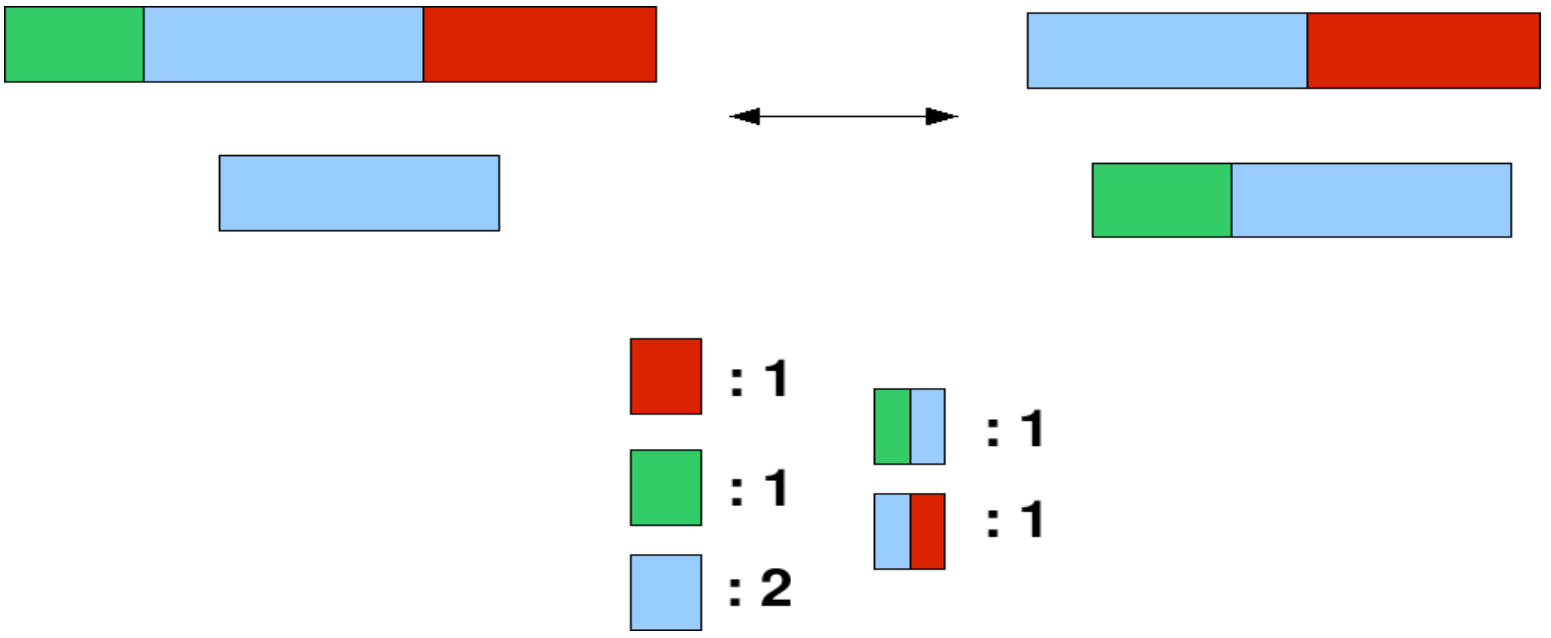
Example



Two transcriptomes may have the same signature

Interchangeable sets

- Two disjoint sets of variants S1 and S2 are interchangeable if each exon and each exon junction has the same number of occurrences in each set



Impact on real data

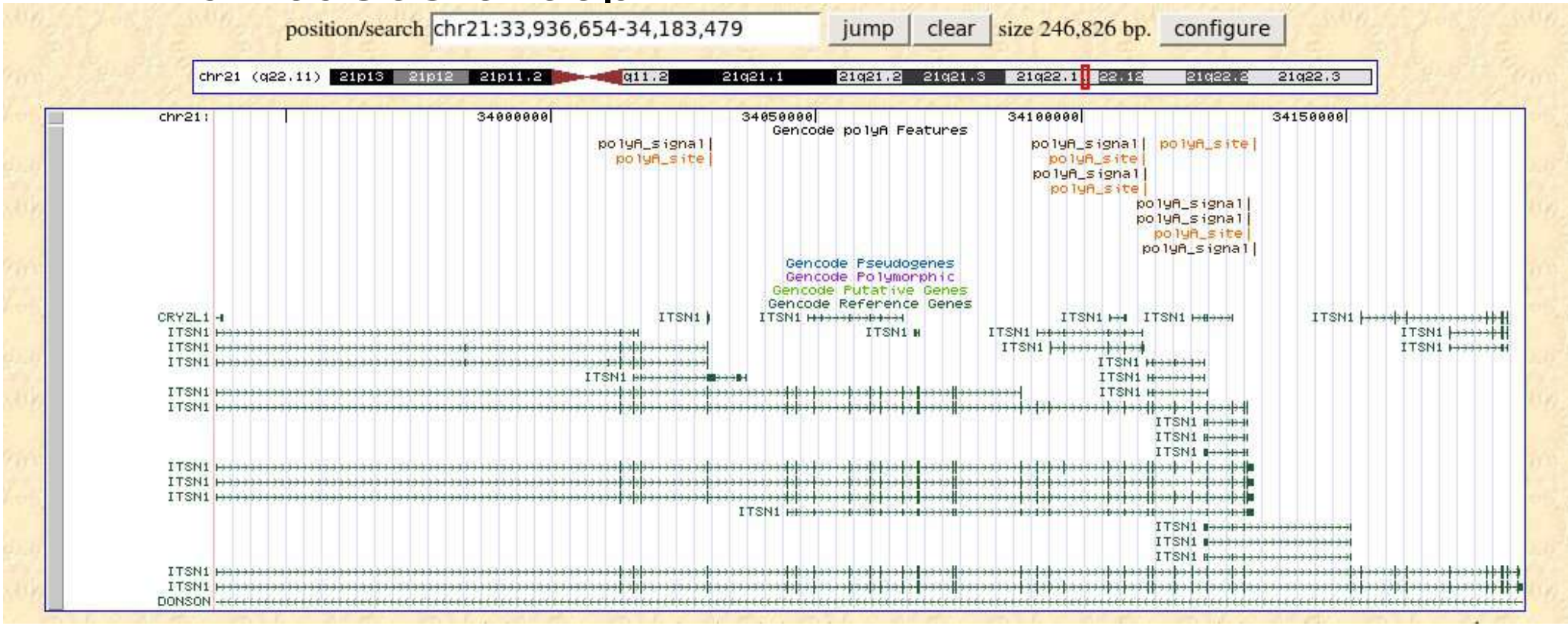
- Gencode dataset: a set of extensively annotated transcripts (681 loci, 2981 variants, 1% of human genome)
- Assume that Gencode is the real transcriptome, how many transcripts would you recover ?

Number of variants	2	3	4	5	6	7	8	9	10	11	12	13	14	≥ 15
Total	110	141	141	124	144	133	144	171	110	154	168	130	70	895
Correctly predicted	8	7	4	1	0	0	2	2	0	0	1	1	0	4

- Very few transcript abundances can be determined

Better situation 1

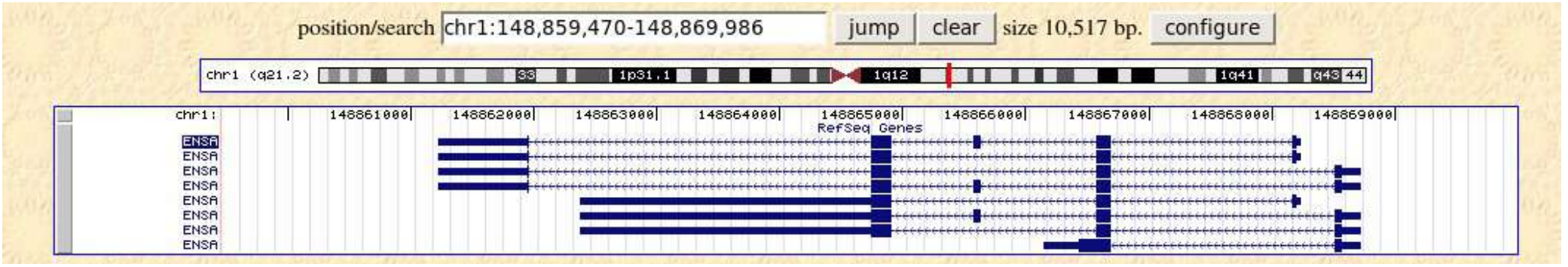
- If we know which transcripts are present in the sample, then assigning abundances is doable in all cases except 1



Better situation 2

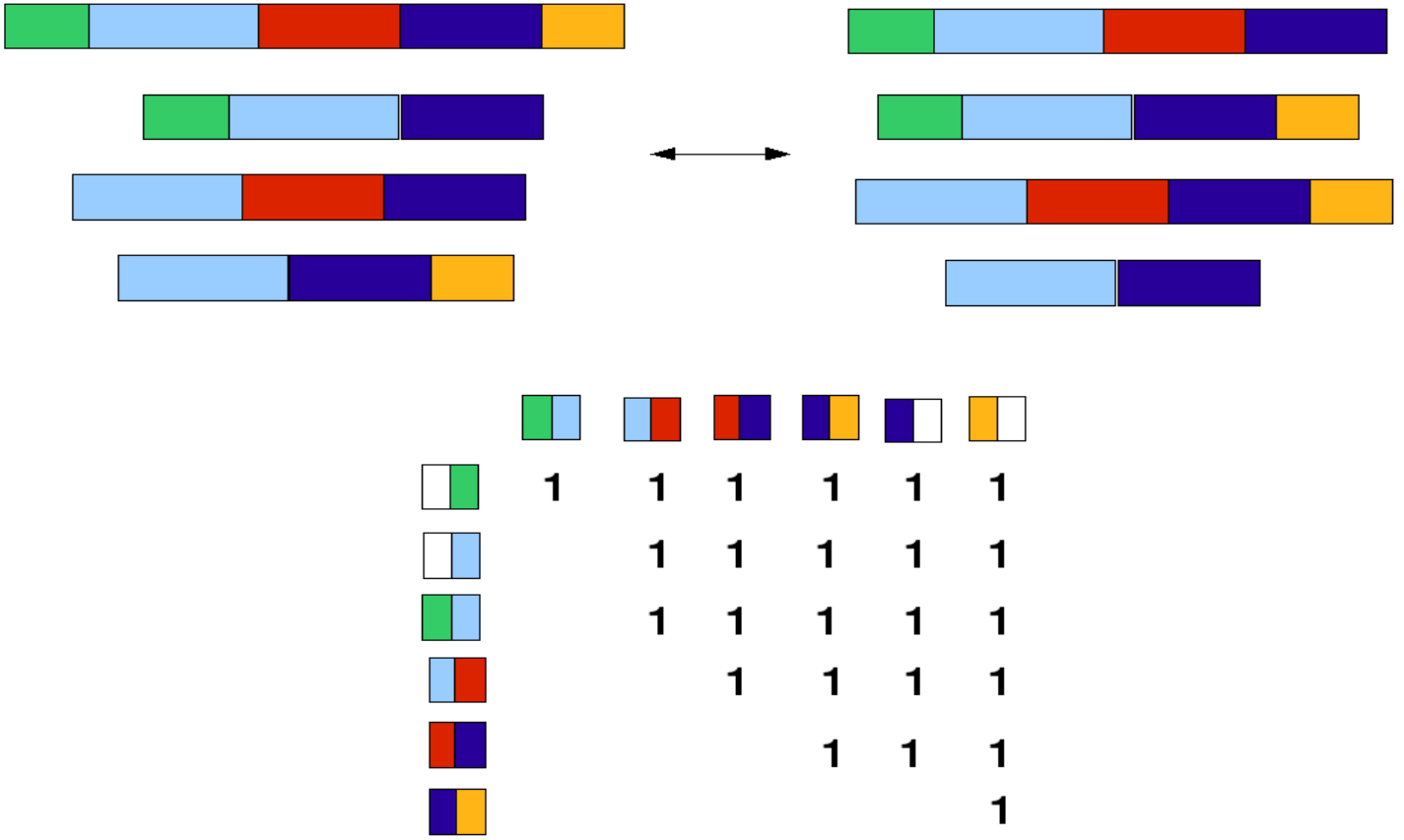
- If we have paired-end reads, then both identifying transcripts and quantifying their abundance should be doable for most loci

- Counter-example:



Interchangeable sets

- The same concept can be defined for pairs of block junctions



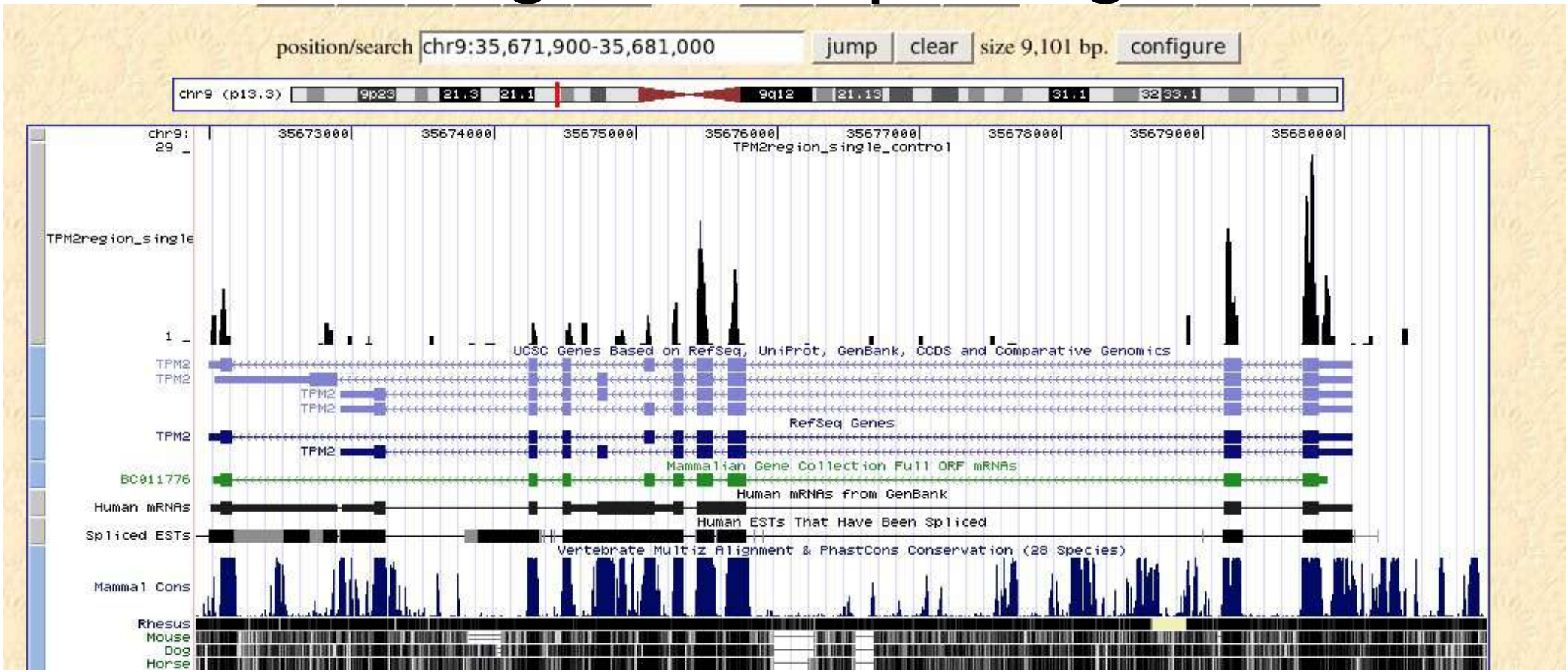
Transcript reconstruction

- Transcriptome reconstruction from short (paired) reads may not be unique, even with only 3 (5) exons
- In practice, single reads should be sufficient to quantify known transcripts, not to discover transcripts
- Paired-end reads should be sufficient to discover and quantify transcripts
- In general, interchangeable sets can be identified and removed a priori

Transcript quantification

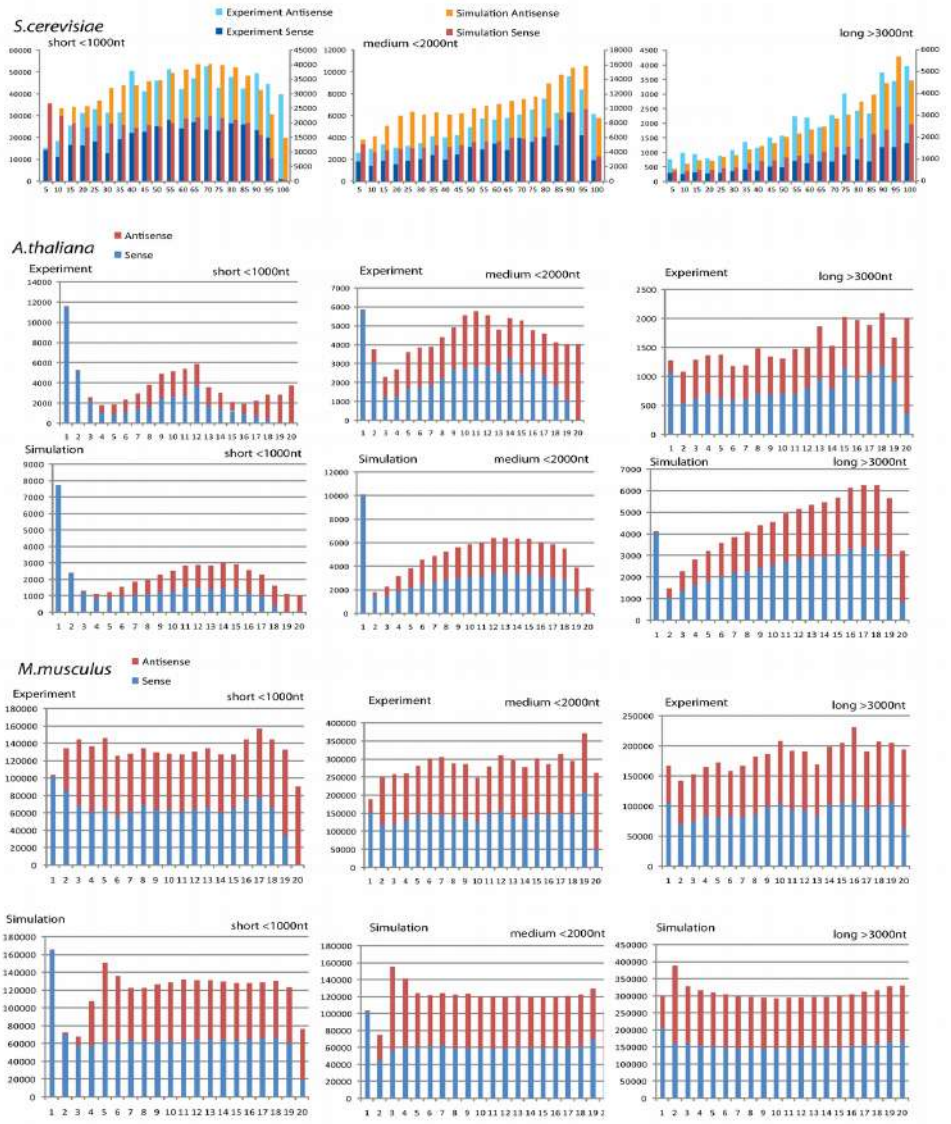
- In practice, the assumptions are not met:
  - Read coverage is not uniform across transcript length (technical reasons depending on the RNA-seq protocol)
  - Read coverage is not uniform across genes (some are very expressed, some are not)

Non uniform distribution of reads along transcript length



More reads at the 5'end, because of random priming during RT.
Assigning abundances to all annotated spliceforms is not trivial.
Read distribution has to be modelled.
http://flux.sammeth.net

Simulating the RNA-seq protocol(s) to understand/model the biases



http://sammeth.net/confluence/display/SIM/Home

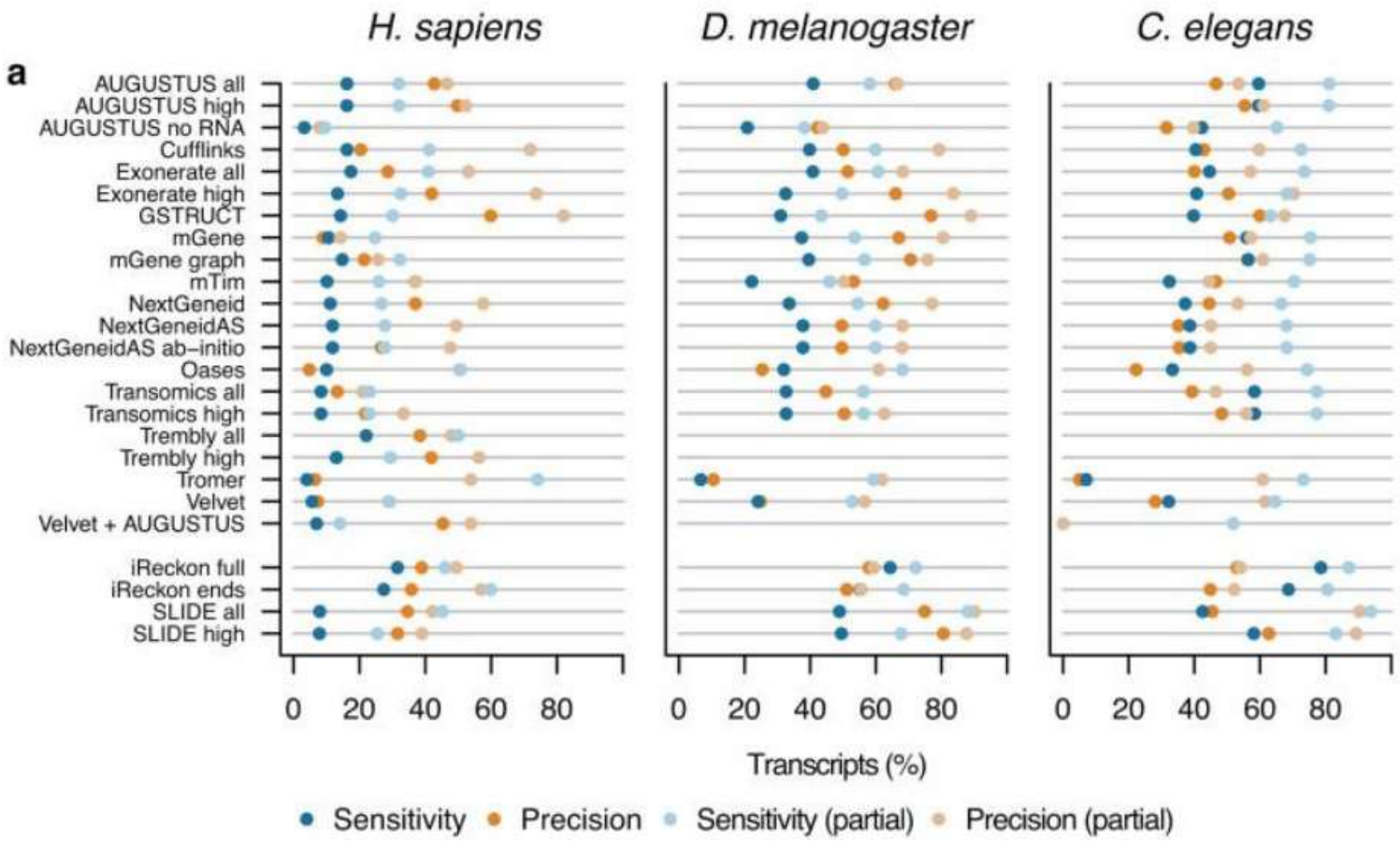
Dynamic range of gene expression

- Many transcripts are present in one copy per cell, few transcripts are present in 10^5 copies
- RNA-seq will identify the most abundant transcripts

Transcript quantification

- The RGASP competition:
  - Assess the status of computational methods to map human RNAseq data, assemble them into transcripts and quantify the abundance of the transcripts in particular datasets

Rgasp results

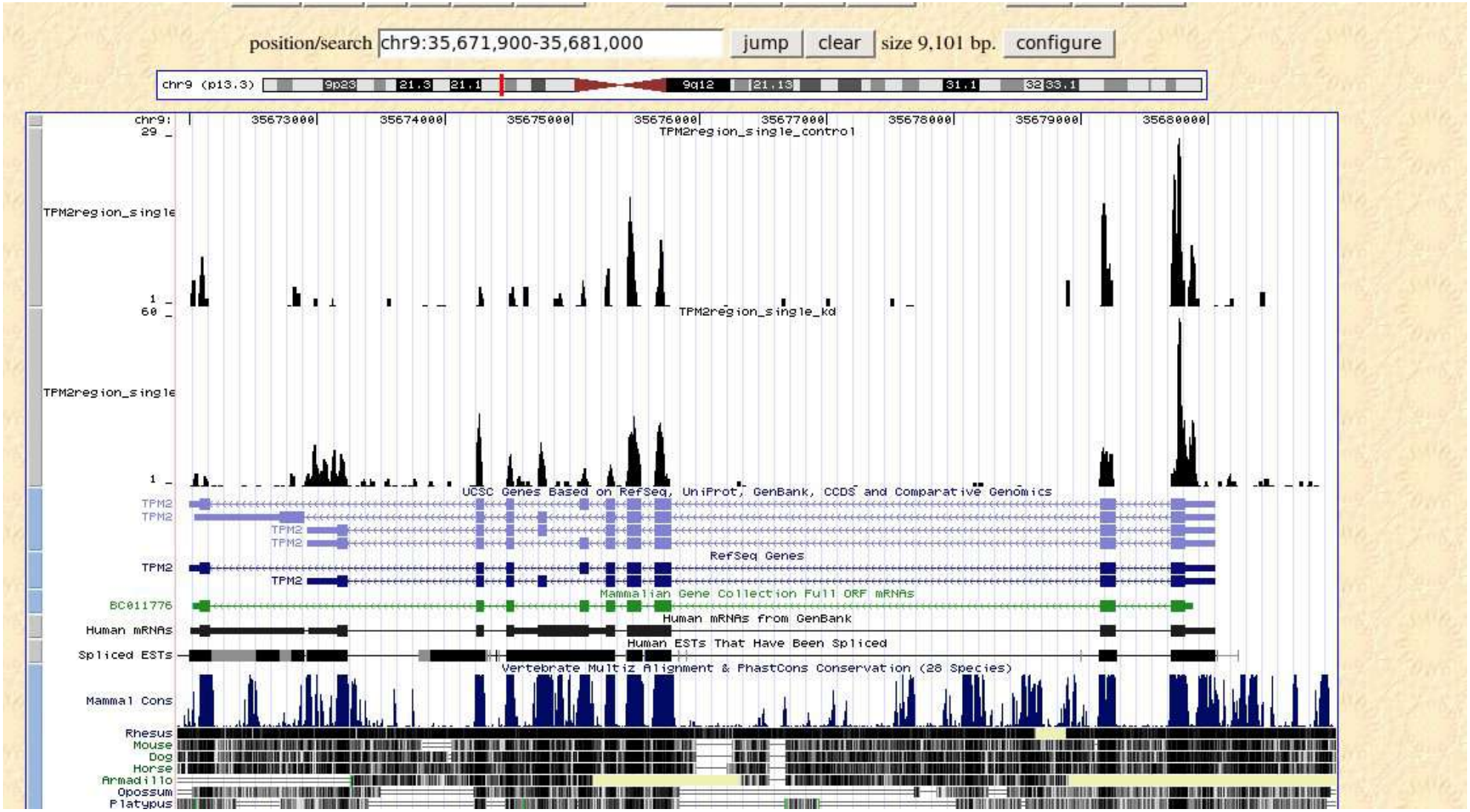


Steijger et al, Assessment of transcript reconstruction methods from RNAseq, Nature Methods, 20

Detecting differentially spliced genes

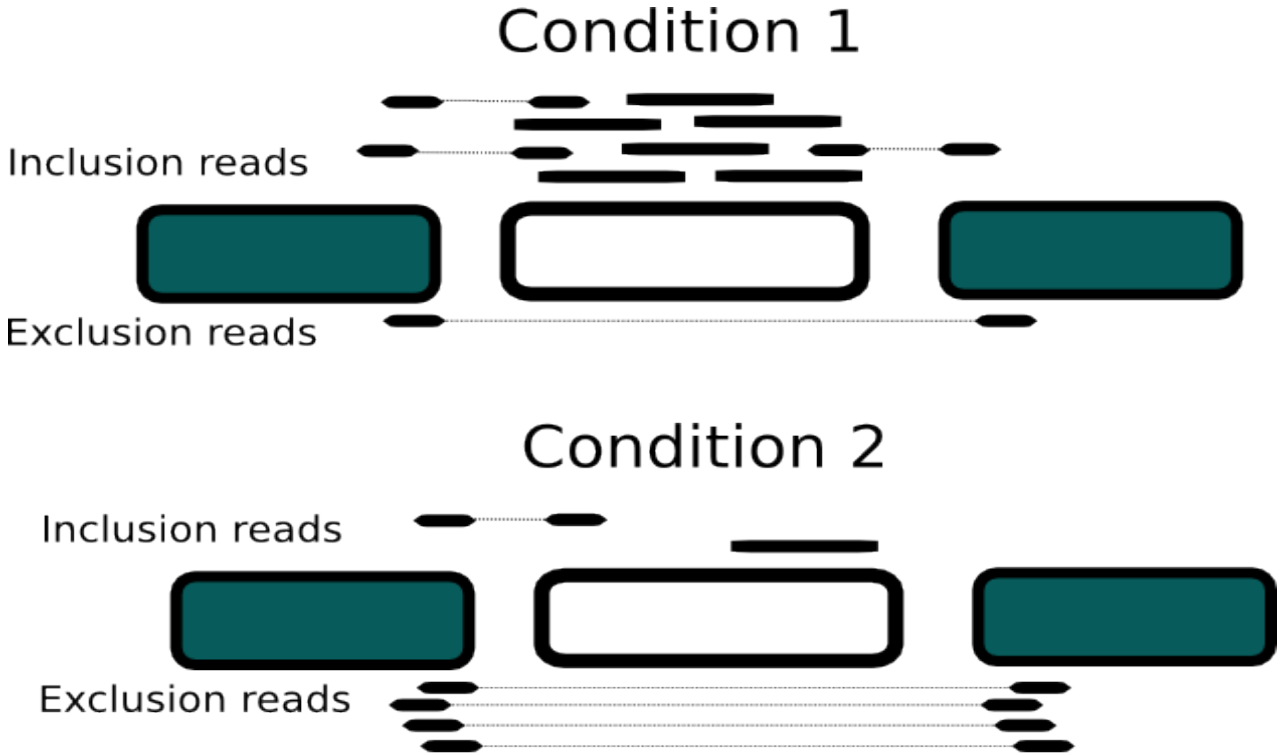
- When we work with two different conditions, we may assume that the distribution of reads along transcripts is equally biased in both conditions
- Modelling precisely this distribution is therefore less crucial when comparing conditions
- Note that if the bias in the distribution is position dependent, then it is important that the transcripts compared in the two conditions have similar length.

Read distribution, two conditions



We can still detect differences in the signal, and call them differences in splicing
No need to fully understand read distribution... in most cases

A test for differential splicing



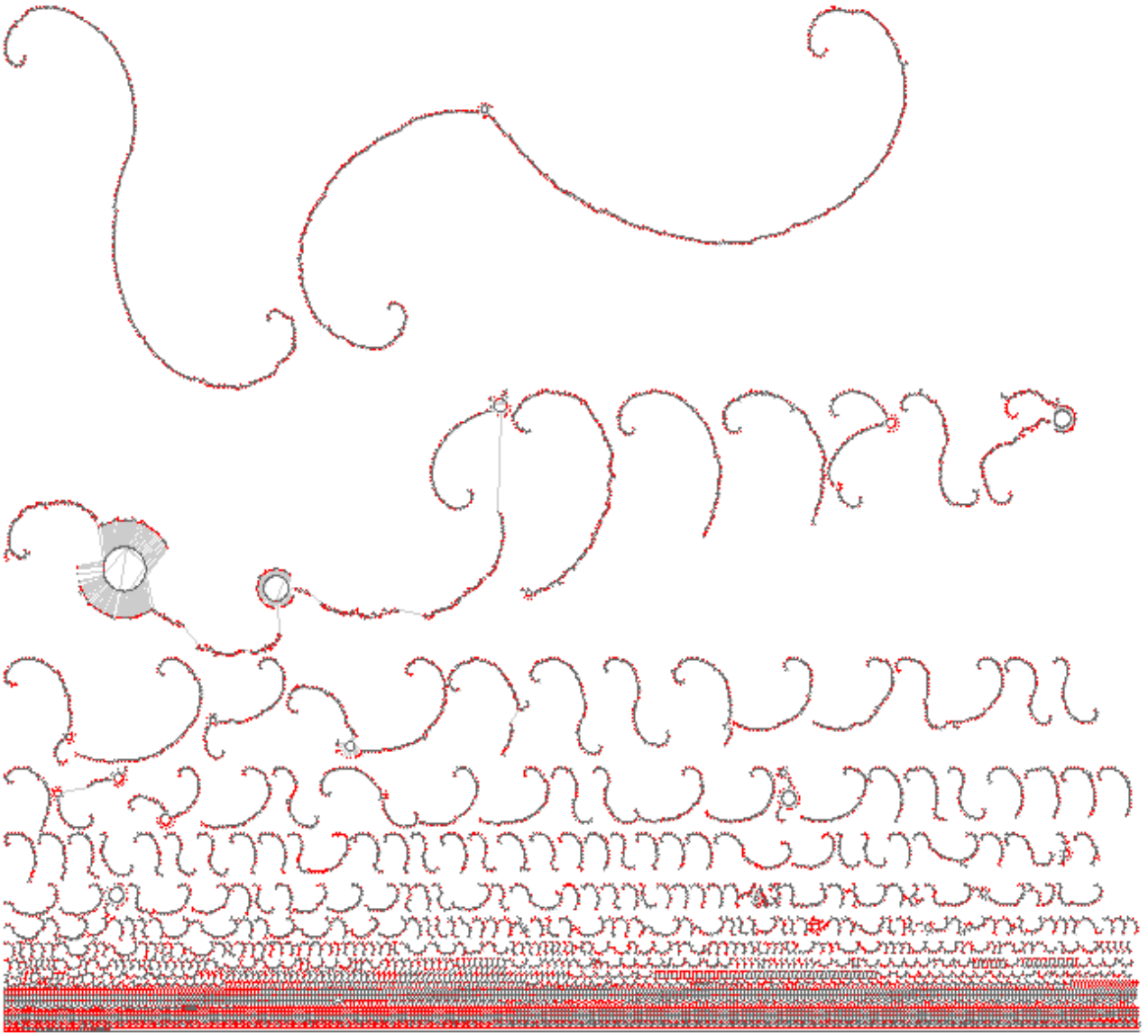
	Condition 1	Condition 2
Inclusion	9	2
Exclusion	1	4

p=0.036

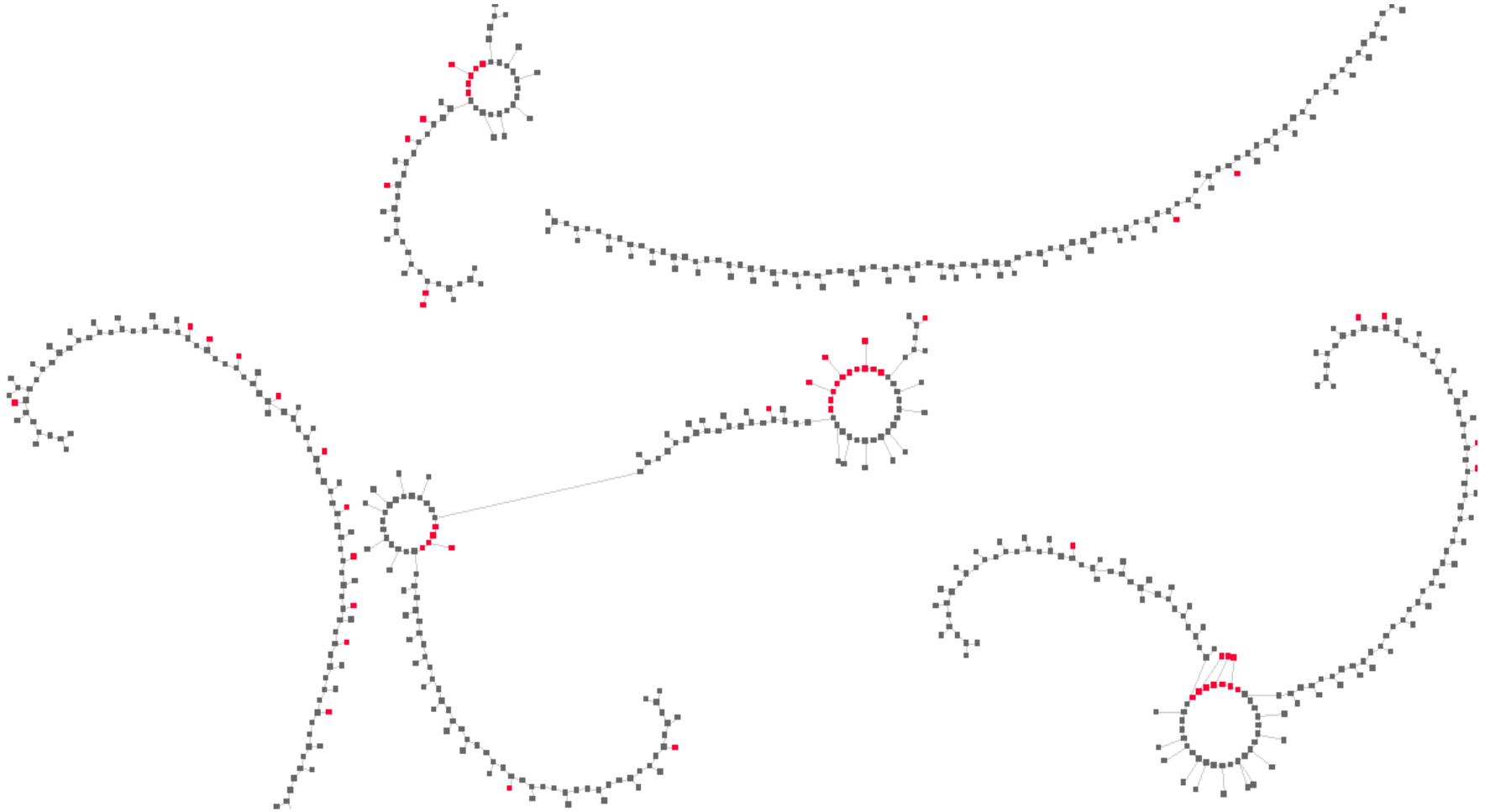
RNA-seq for non model organisms

- What can we do without a reference genome:
  - Map the reads to a related genome, allowing more mismatches
    - Issues: many reads will not map
    - Divergence differs between genes
  - De novo transcriptome assembly
    - Available software are : Trinity, Oases, Trans-abyss
    - Issue : As in most assemblers, heuristics are applied to linearize the graph, which leads to underestimation of SNPs, indels and alternative splicing
    - Solution : Use also a local assembler : KisSplice

Example of DBG built from RNA-seq data

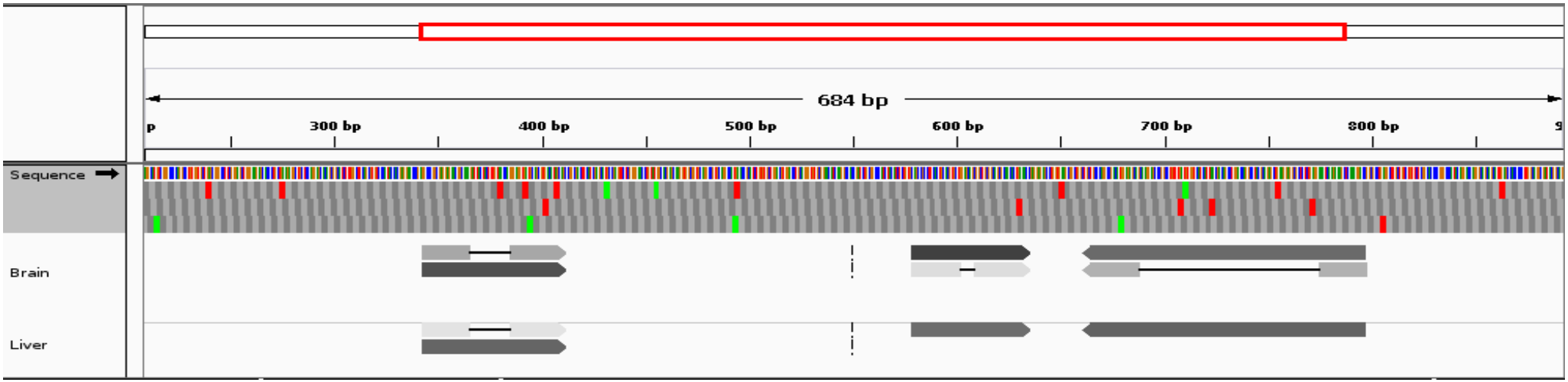


Polymorphism in RNA-seq data



Global Vs Local Transcriptome Assembly

- Global Transcriptome Assembly
  - The goal is to produce full length transcripts, at the expense of underestimating polymorphism
- Local Transcriptome Assembly
  - The goal is to review all polymorphisms, but it does not produce full length transcripts
- Combination of both



Softwares

- Global
  - Trinity
    - Trinityrnaseq.sourceforge.net
  - Oases
    - http://www.ebi.ac.uk/~zerbino/oases/
- Local
  - KisSplice
    - http://kissplice.prabi.fr

Transcriptome Assembly

- Open questions :
  - Should we use transcriptome assembly even when a reference genome is available ?
  - How to efficiently deal with complex splicing events (more than 2 transcripts involved)
  - How to differentiate AS and indels ?

Summary

- Many problems are still open :
  - Identify new exon junctions
  - Quantify full-length transcripts
  - Call differentially spliced genes

RNAseq original articles

- Wang et al (2008): Alternative isoform regulation in human tissue transcriptomes. Nature
- Nagalakshimi et al (2008): The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science
- Mortazavi et al (2008): Mapping and quantifying mammalian transcriptomes by RNA-seq. Nature Methods
- Sultan et al (2008): A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science