Gene Fusion Tools

Chimeric Transcripts

- Chimeric RNA encoded by
  - a fused gene resulting by the fusion at DNA level of two different genes
  - two different genes by subsequent trans-splicing

  (see "Course introduction & Molecular Biology" slides for definitions and details)

- Certain fusion transcripts are commonly expressed by cancer cells

- Finding the exact point of fusion (namely *breakpoint*) helps in the better characterization of the disease

- RNA-Seq reads and in particular *paired end reads* are helpful in detecting chimeric transcripts.

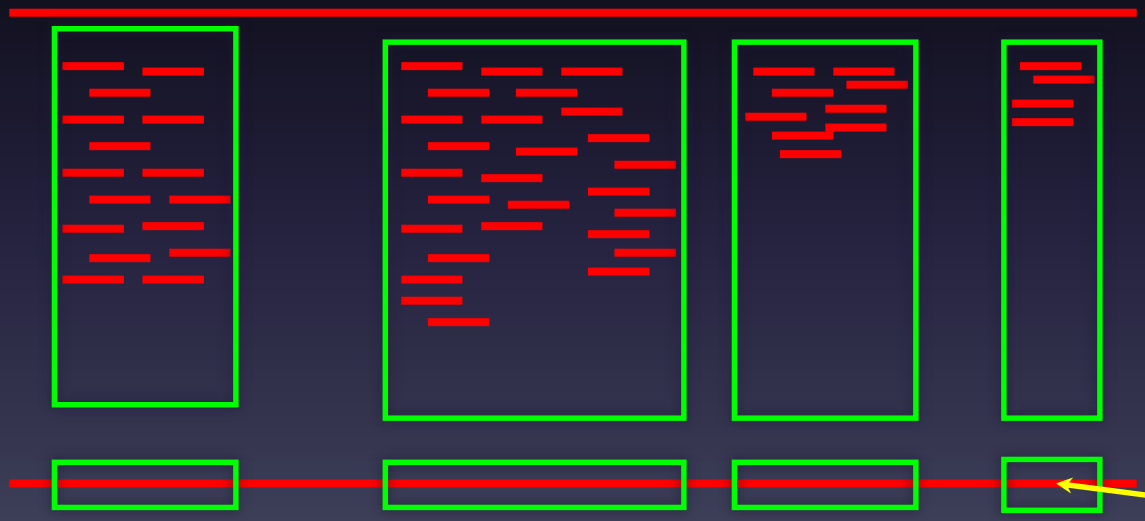
  Remember that reads in paired ends mode are reads sequenced just on the two extremities of RNA subsequence for about one hundred bases. The two extremities of the read are named *mates* (in particular, *mate 1* and *mate 2*). The sequence between the two mates is not known.

Computational Issues

- Alternative splicing and chimerism makes analysis of RNA expression or alteration much more complex
- The problem is to map the reads on the correct location on the genome avoiding mismatches and, vice versa, multiple matching
- The identification of alternative splicing and chimerism it is very hard because splicing and fusion breakpoints are either unknown or predictable.

Splicing-aware Alignment

- Tools such as TopHat or SplitSeq takes alternative splicing into account, by identifying exon regions and mapping reads across putative junctions (up to a certain intron size)



TopHat assembles the mapped reads in consensus sequences

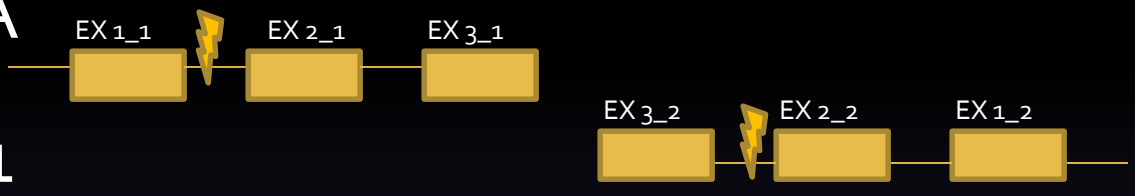
TopHat extracts the sequences for the resulting islands of contiguous sequence inferring them to be putative exons and aligns on them.

Computational Issues

- Alternative splicing and chimerism makes analysis of RNA expression or alteration much more complex
- The problem is to map the reads on the correct location on the genome avoiding mismatches and, vice versa, multiple matching
- The identification of alternative splicing and chimerism it is very hard because splicing and fusion points are not known
- In particular, fusion breakpoints could be not-canonical, i.e. genes can be broken inside exons, keeping subsequences of them and loosing the others. Moreover, the genes involved in the fusion can be, often, located in different chromosomes and thus far away from each others.

Example of gene fusion in mRNA

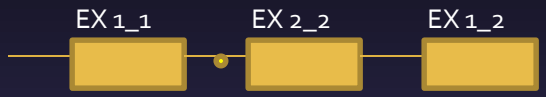
- DNA



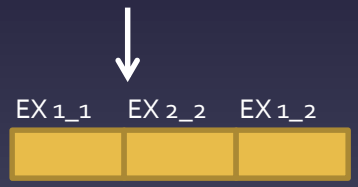
Gene 1

Gene 2

- Fused DNA



- mRNA

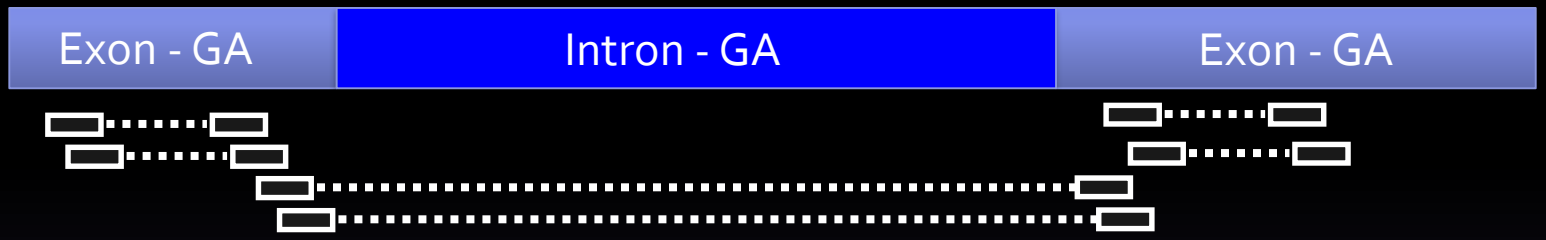


Common case: on the mRNA the breakpoints are at the exon's boundaries whereas on the DNA they are within the introns.

However, fusion may not happen at exon boundaries
=> Non-canonical junctions must be considered

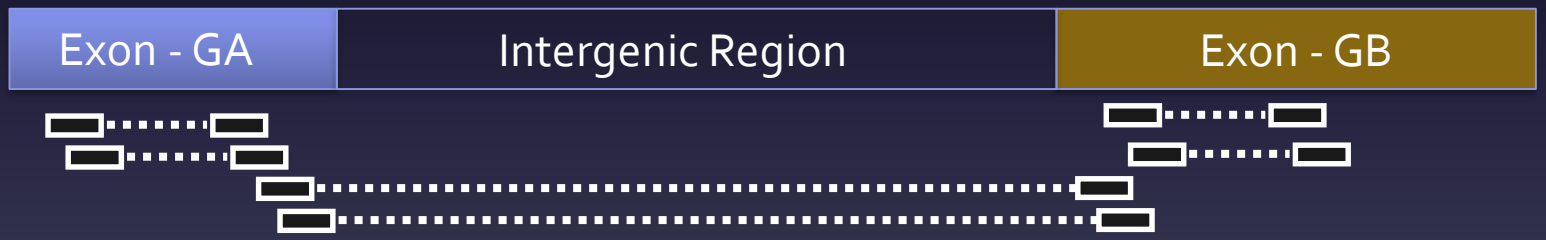
Mapping the Reads on the Genome

Concordant Reads



Splicing Event

Discordant Reads



Gene Fusion

Intergenic Regions are much more wide than intron gaps. Moreover, differently from splicing, searching for a fusion requires the computation of all the combinations along the genome

WGS vs RNA-Seq

The ability of an approach to identify fusions from NGS data relies on the types of sequencing data it aims to work on as well as its computational strategies to process the data.

WGS, RNA-Seq are the major NGS technologies for fusion gene detection

WGS (Whole Genome Sequencing): It is a laboratory process that determines the complete DNA sequence of an organism's genome at a single time. It provides the most comprehensive and unbiased characterization of genomic alterations in genomes, especially cancer genomes. Using WGS technology, a variety of fusion genes have been discovered, some of which are believed important for the growth of certain cancer cells. One drawback of WGS, however, is that it requires a great amount of sequencing and intensive computational analysis. Finally, the significance of a fusion gene discovered using WGS relies on its effects on expression and on whether it produces fusion transcripts.

RNA-Seq (RNA Sequencing): It only sequences the regions of the genome that are transcribed and spliced into mature mRNA, which is 2% of the entire genome. Another advantage that makes RNA-Seq ideal for the discovery of expressed fusion genes is that it allows for detection of multiple alternative splice variants resulting from a fusion event. These distinct features of RNA-Seq, together with its low cost and quick turnaround time, make RNA-Seq very popular in fusion gene studies. However, one main limitation of RNA-Seq is that it cannot detect fusion events involving non transcribed regions.

Mapping First vs Assembly First

The computational strategies for fusion gene detection can be then grouped in two different categories:

- **Mapping-First Approach**: Reads are first aligned to reference DNA/RNA sequences and then fusion breakpoints found from the resulting alignment patterns. Compared to the assembly-first approach, the mapping first approach is faster and has dominated the field of NGS-based gene fusion studies.

- **Assembly-First Approach:** Initially reads that overlap are assembled. The long reads assembled, also called contigues, are then mapped to reference sequences for structure alteration identification.
If the algorithm assembles short reads directly without mapping them to the references, then it is called de novo assembly: It does not need a reference genome/transcriptome for fusion detection but the assembly of short sequences is too time-consuming and too error prone.

Reference Sequences

The detection of Structural Variations that may result in gene fusions imposes the aligment of the reads to a reference genome sequence.

NCBI build 37/36

UCSC hg19/38

BUT

Fusion genes involving novel sequences not represented in the reference will be missed!

RNA-Seq data can be mapped also to a transcriptome library so that the genes involved in each fusion can be identified **BUT** only candidates involving annotated exons are in this manner considered and fusion genes with novel exons cannot be detected.

Single vs Paired-End Reads

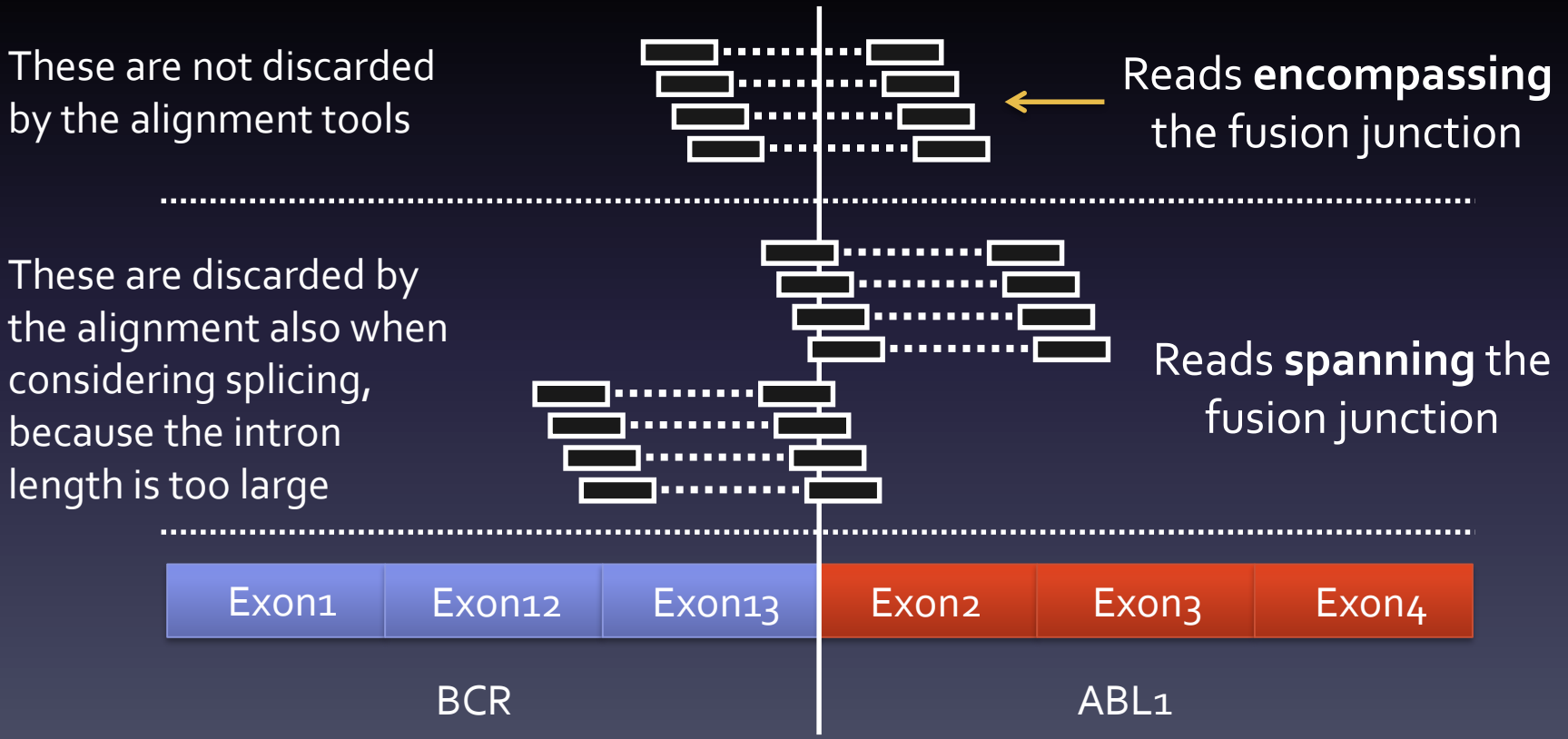
Single-end reads were used at the beginning to detect fusion genes:

Paired-end reads (obtained by sequencing both the ends of a RNA fragment) are nowadays widely used to detect gene fusions:

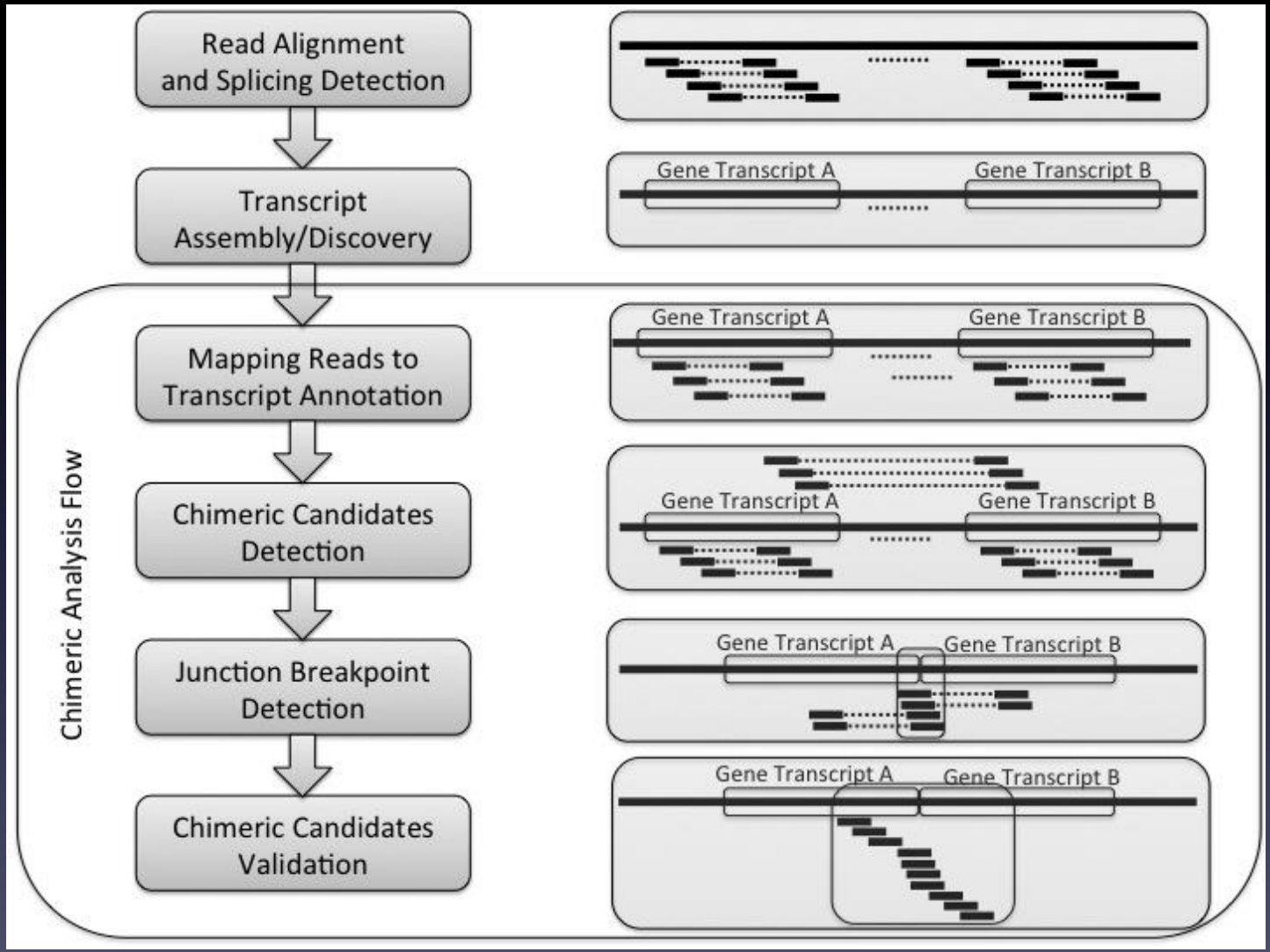
- **_Encompassing_ read (named also Split reads): a read that contains a fusion, one mate on the first gene, the other mate on the second gene.**
- **_Spanning_ reads: a pair of reads that passes through a fusion junction with one of the two mates.**
- **Discordant Mapping: the two mates are aligned to different genes.**

Chimeric Transcript Detection

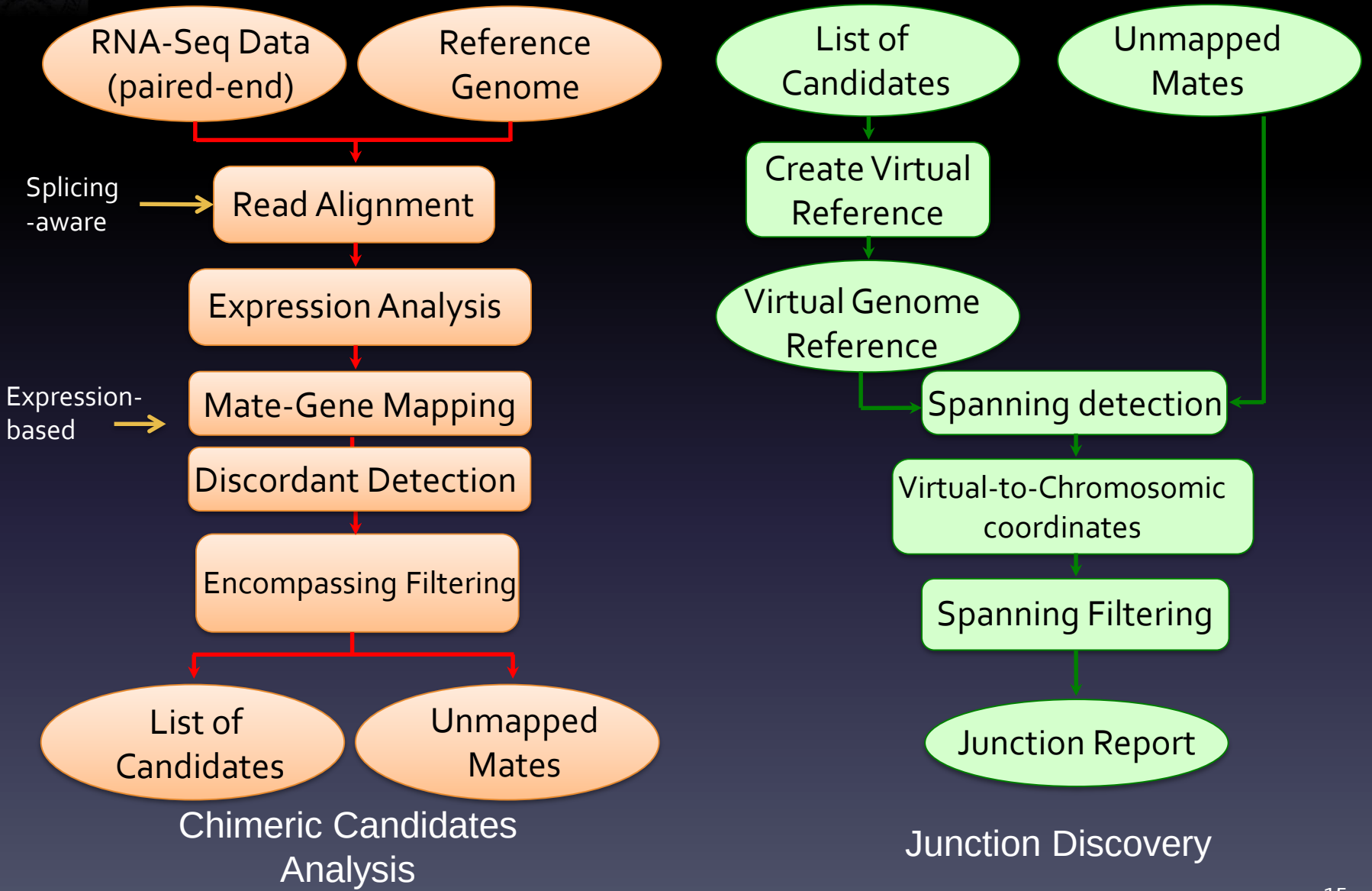
Approach proposed by Maher et al., 2009



Analysis Flow

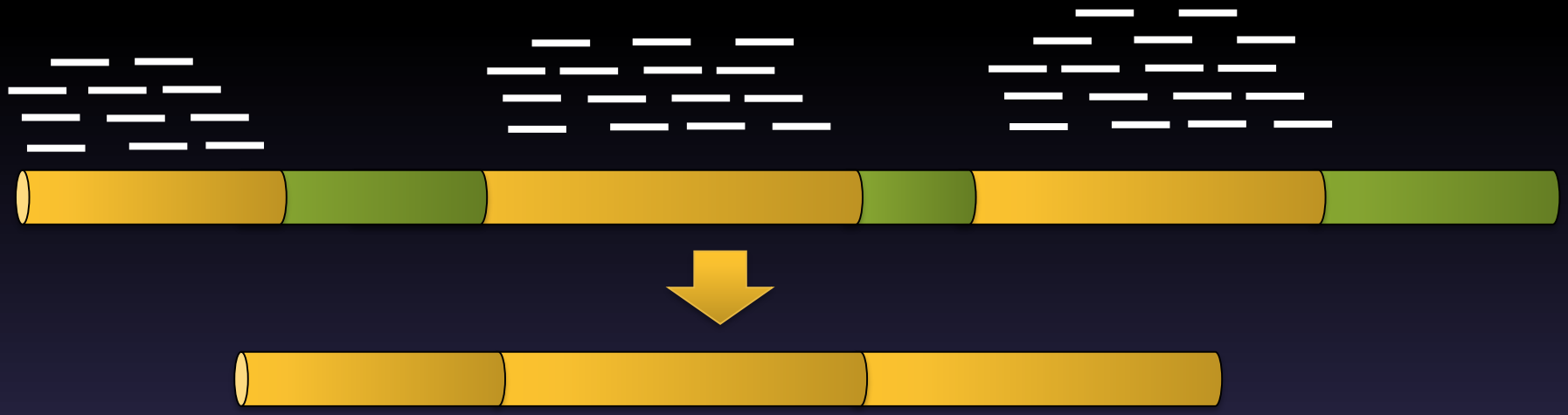


Chimeric Detection Flow



Initial Expression Analysis

- Cufflinks allows to assembly the transcriptome from the experimental sample

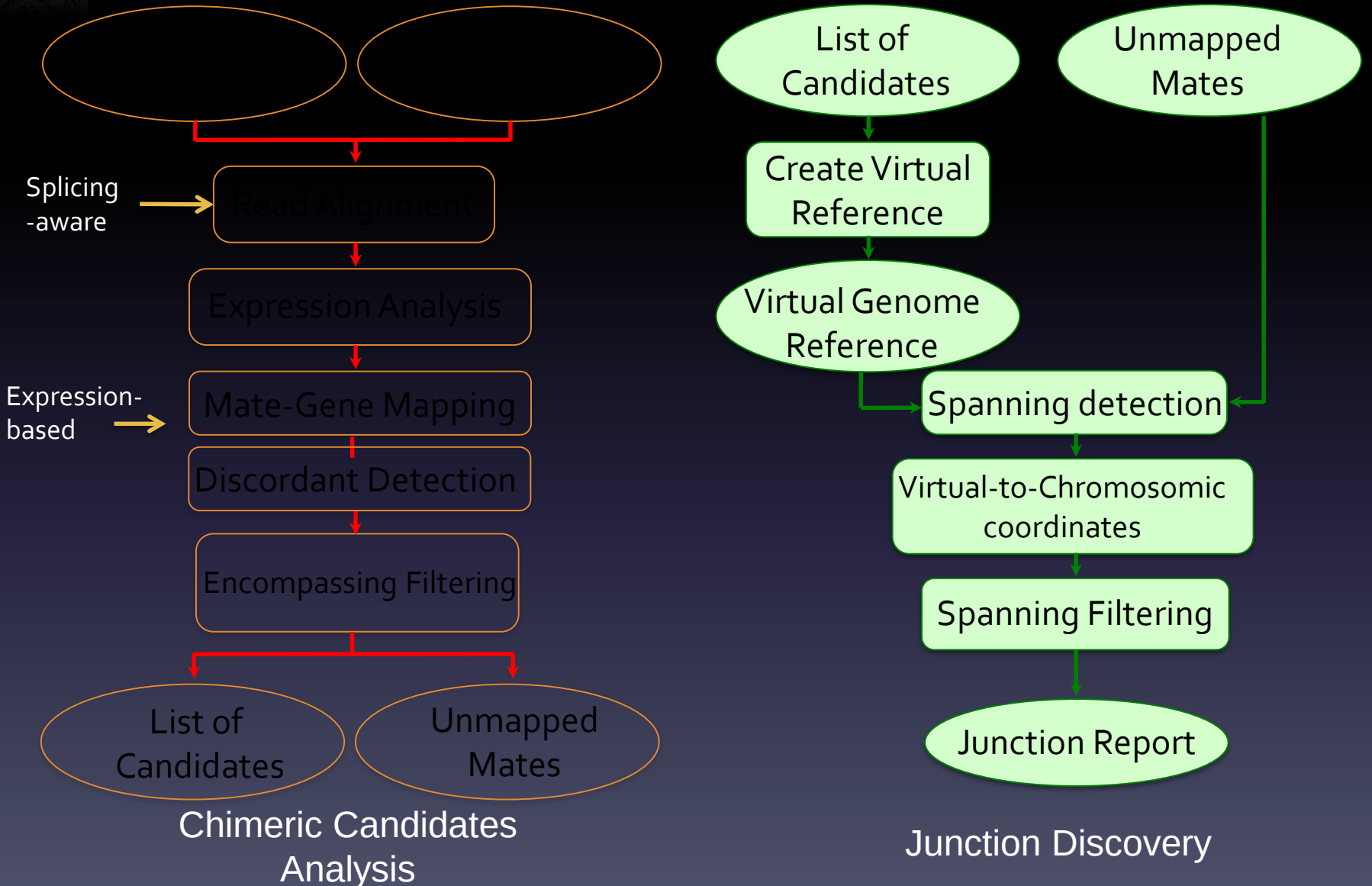


- The resulting GTF annotation transcriptome contains the set of novel and known expressed genes
  - Reduced ambiguities => only the expressed transcripts are considered

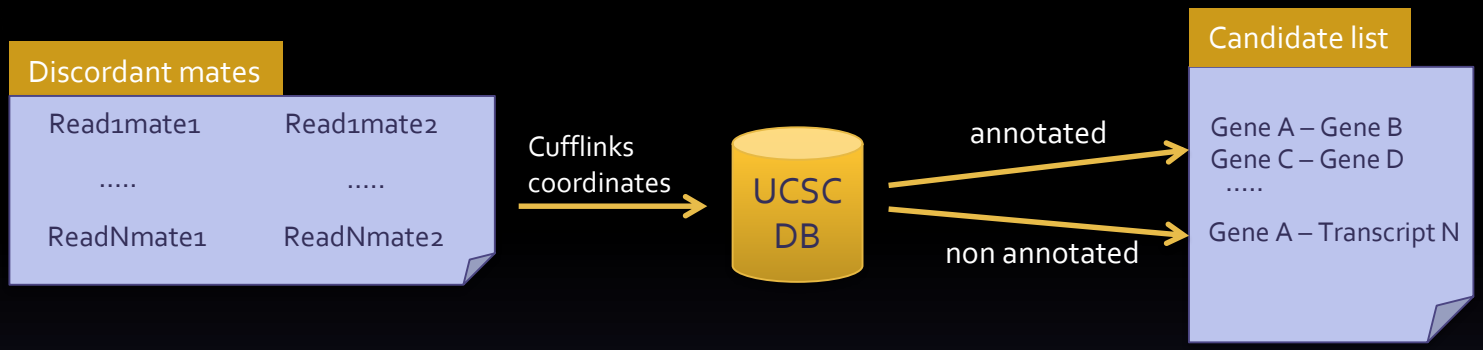
Expression-Based Mapping

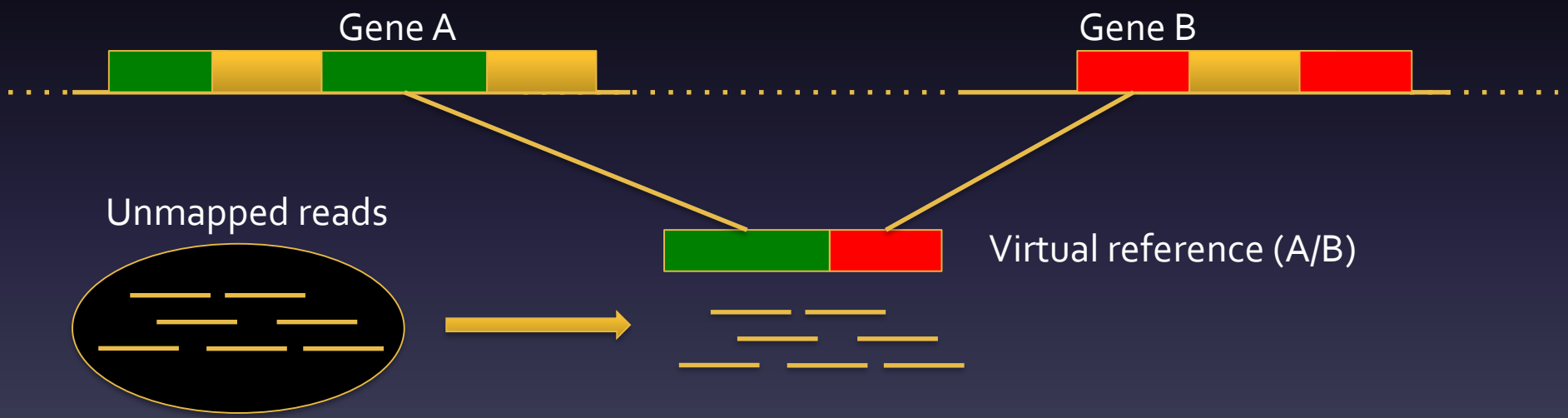
- Advantages
  - Gives emphasis to mostly expressed transcripts
  - Reduce multiple alignment problems because it reduces alignment alternatives
  - Improve accuracy by including non-annotated transcripts
- Drawbacks
  - May disregard fusions involving poorly expressed transcripts
  - Requires high coverage to be effective

Chimeric Detection Flow



Virtual Reference

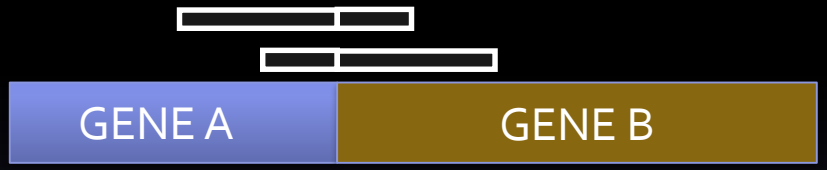




- Alignment using unmapped reads is done on each virtual reference

Junction Breakpoint

On the Transcriptome

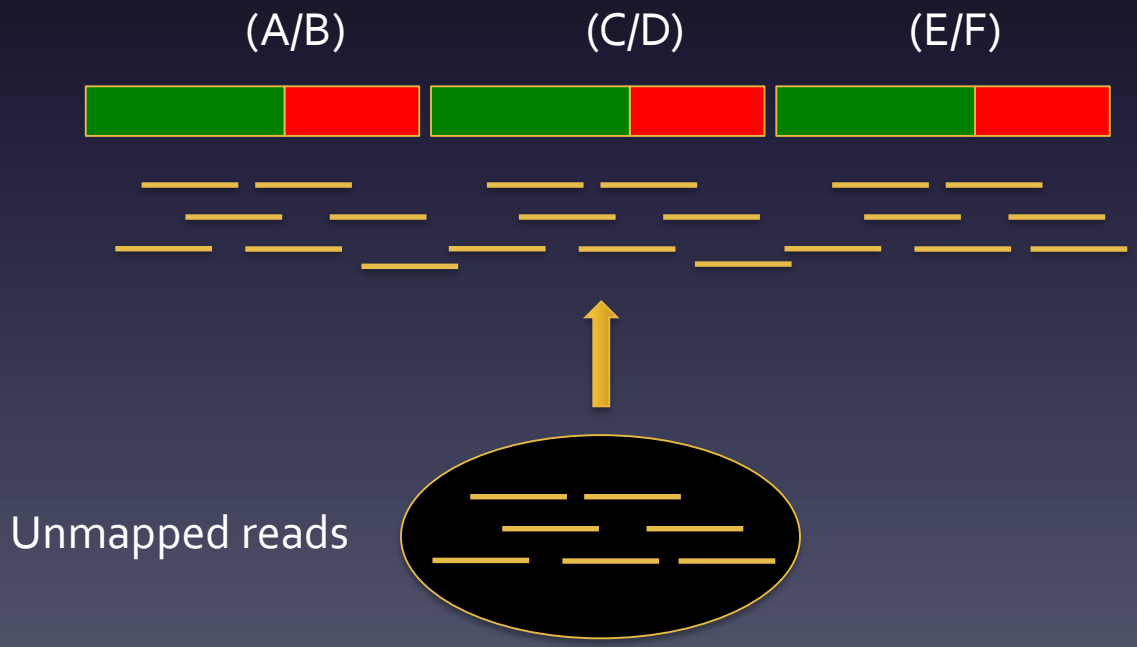


On the Genome

For the identification of the final candidates, we consider spanning reads, reads with one mate mapping on the junction and the other mate mapping on one of the two genes

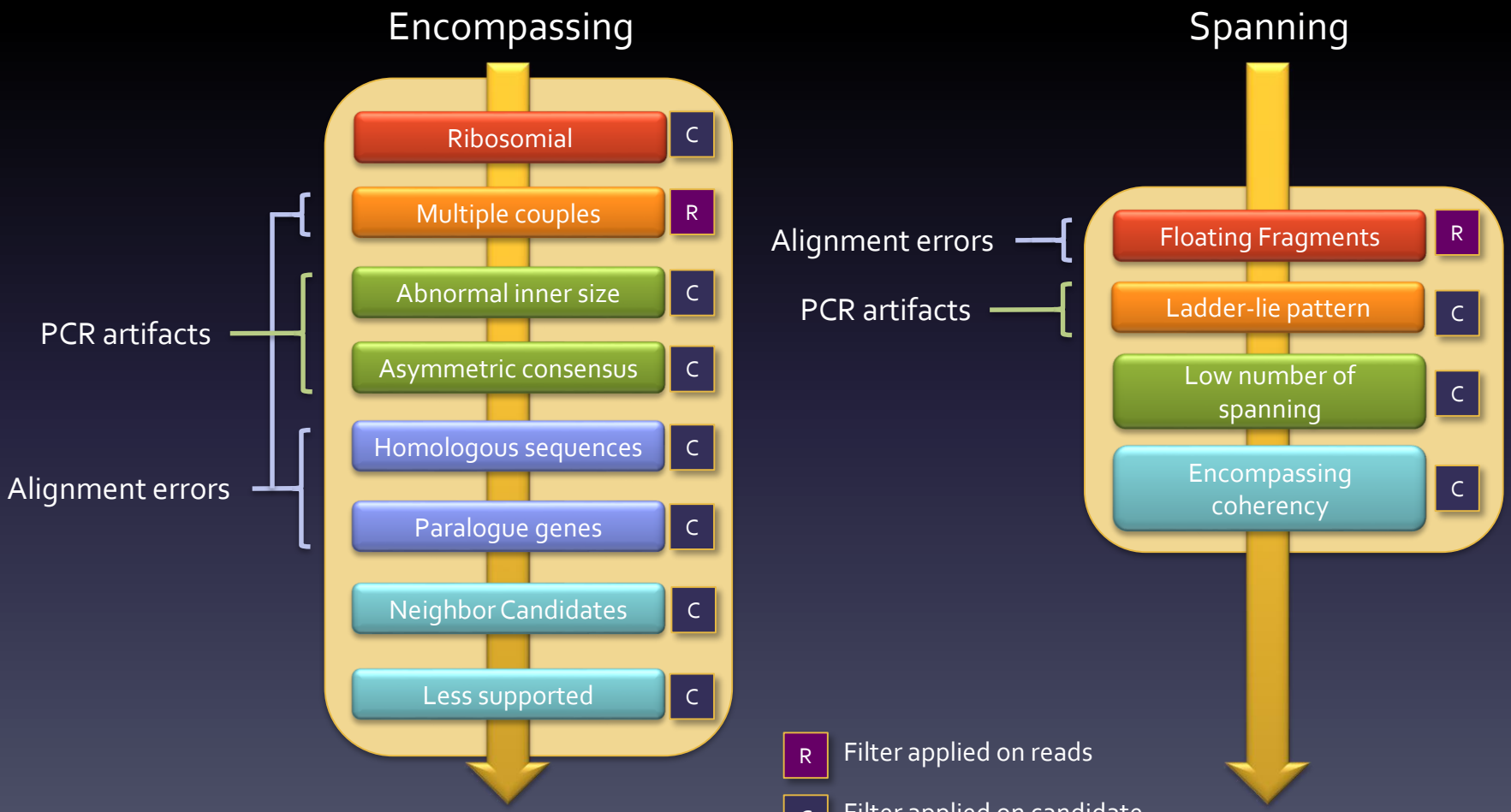
Junction Detection

- Splicing-aware tool is used for alignment in this phase
- A chain of virtual references is created and passed to aligner to reduce computation time
- Coordinates remapping from virtual to chromosomal is needed



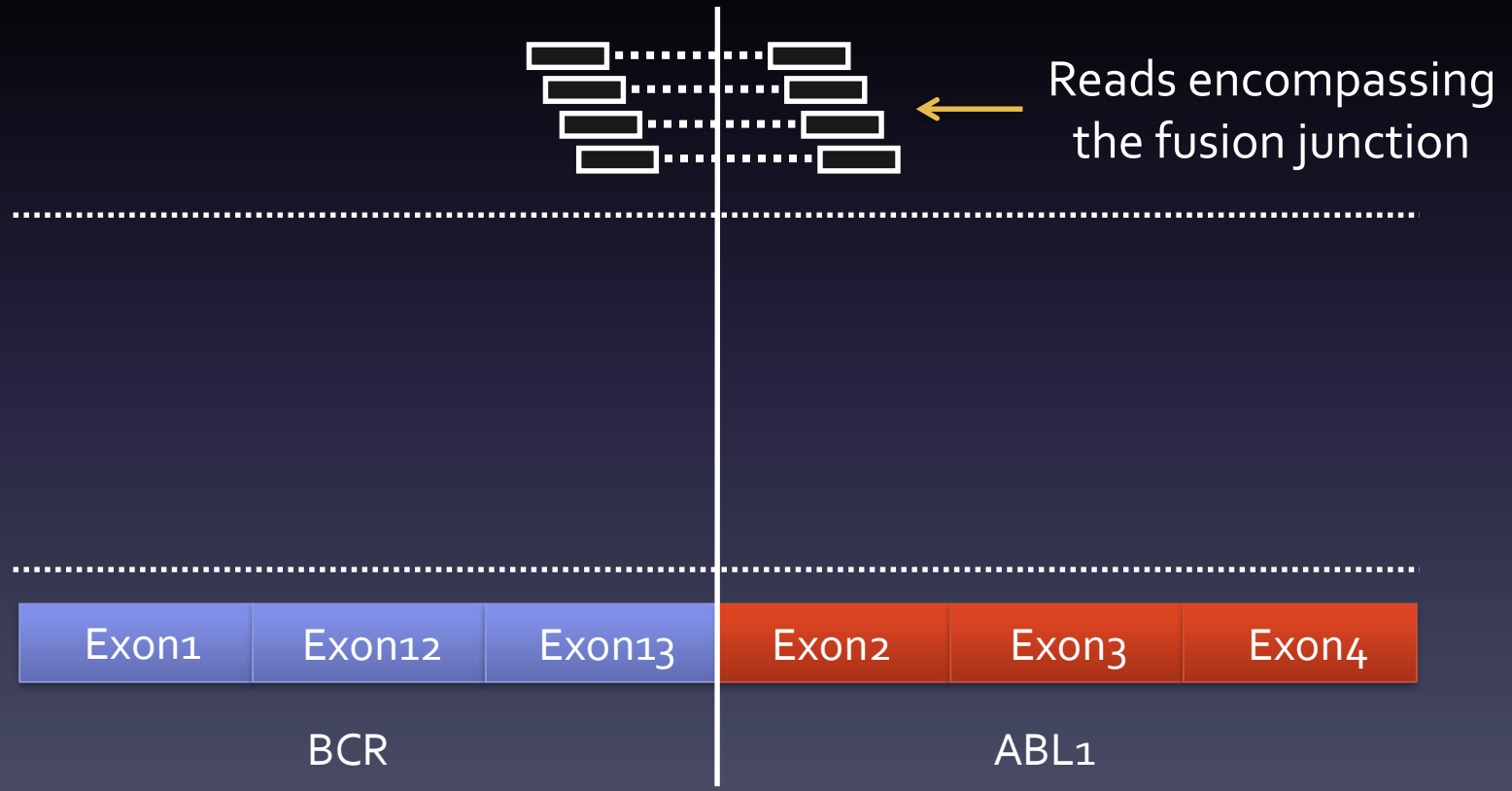
Filtering Scheme

- Filter are needed to remove false positives



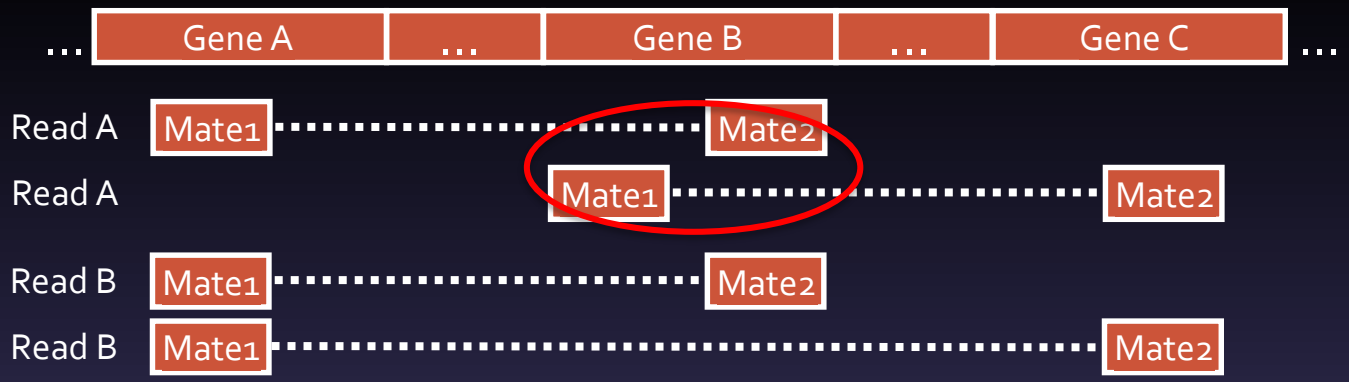
Filtering

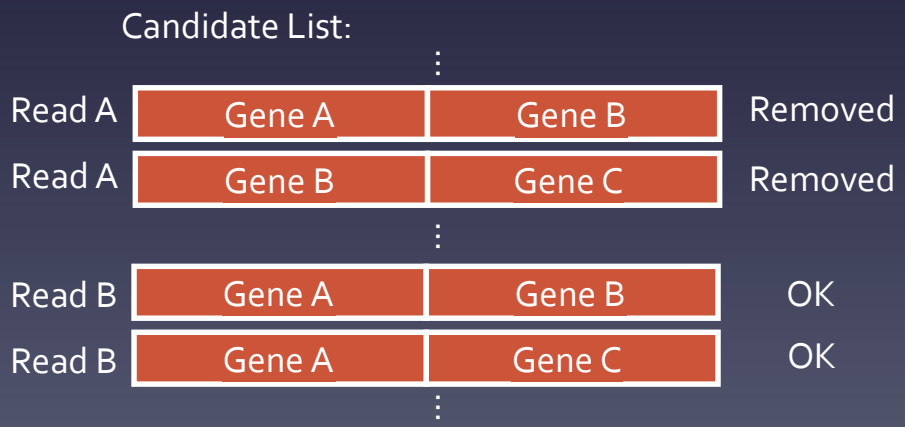
- Encompassing



Filter on Encompassing Reads

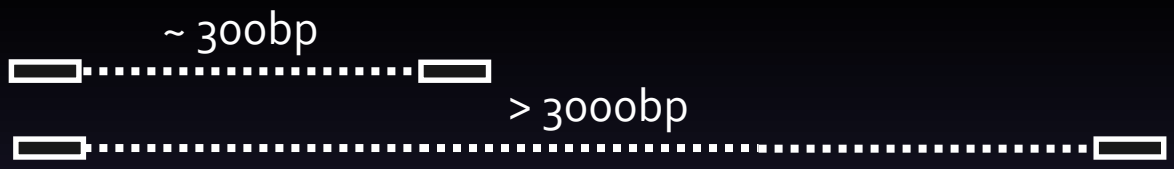
- Selection based on multiple mismatches
  - Mate 1 and Mate2 identify two couples of candidates





Filters on Encompassing Candidates

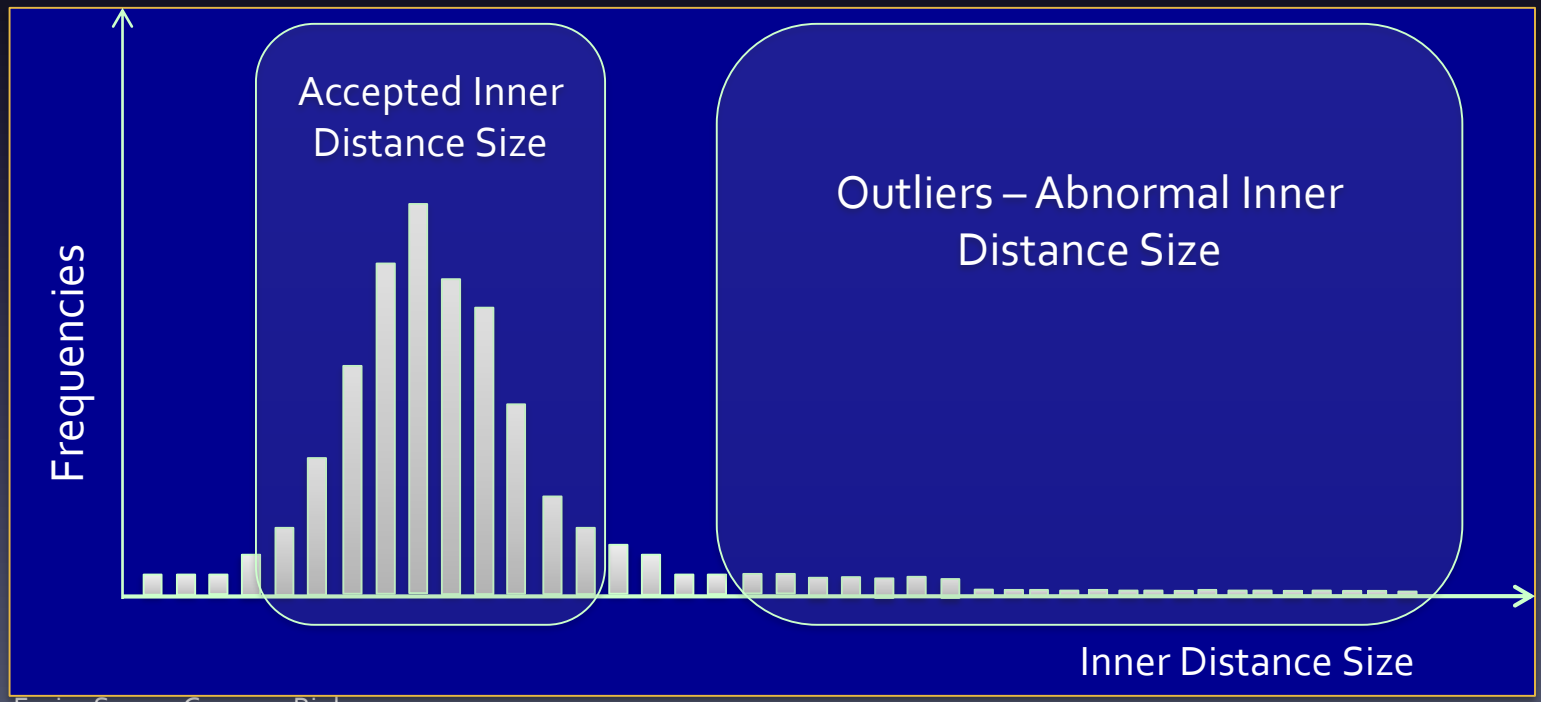
- Remove candidates supported by encompassing on very large fragments



- Due to PCR artifacts, some fragments may have an abnormal size, thus reads have an abnormal inner-size (gap)
- Encompassing with inner size >> threshold are discarded
- Need to:
  - Compute the distribution of fragment lengths (average + stdev)
  - Compute the inner size of the encompassings supporting each candidate
  - Problem: The inner distance is unknown since we don't know the breakpoint yet

Filters on Encompassing

- The inner size (gap between the two mates) must be lower than a threshold determined by inner size distribution analysis (*)
  - Minimum inner distance is compared against the threshold
  - Minimum inner distance is computed by looking at mapping pattern



(*) Sboner et al, FusionSeq ..., Genome Biology, 2010

Computing the Inner Size

On the Genome

The intergenic region length is not known a priori

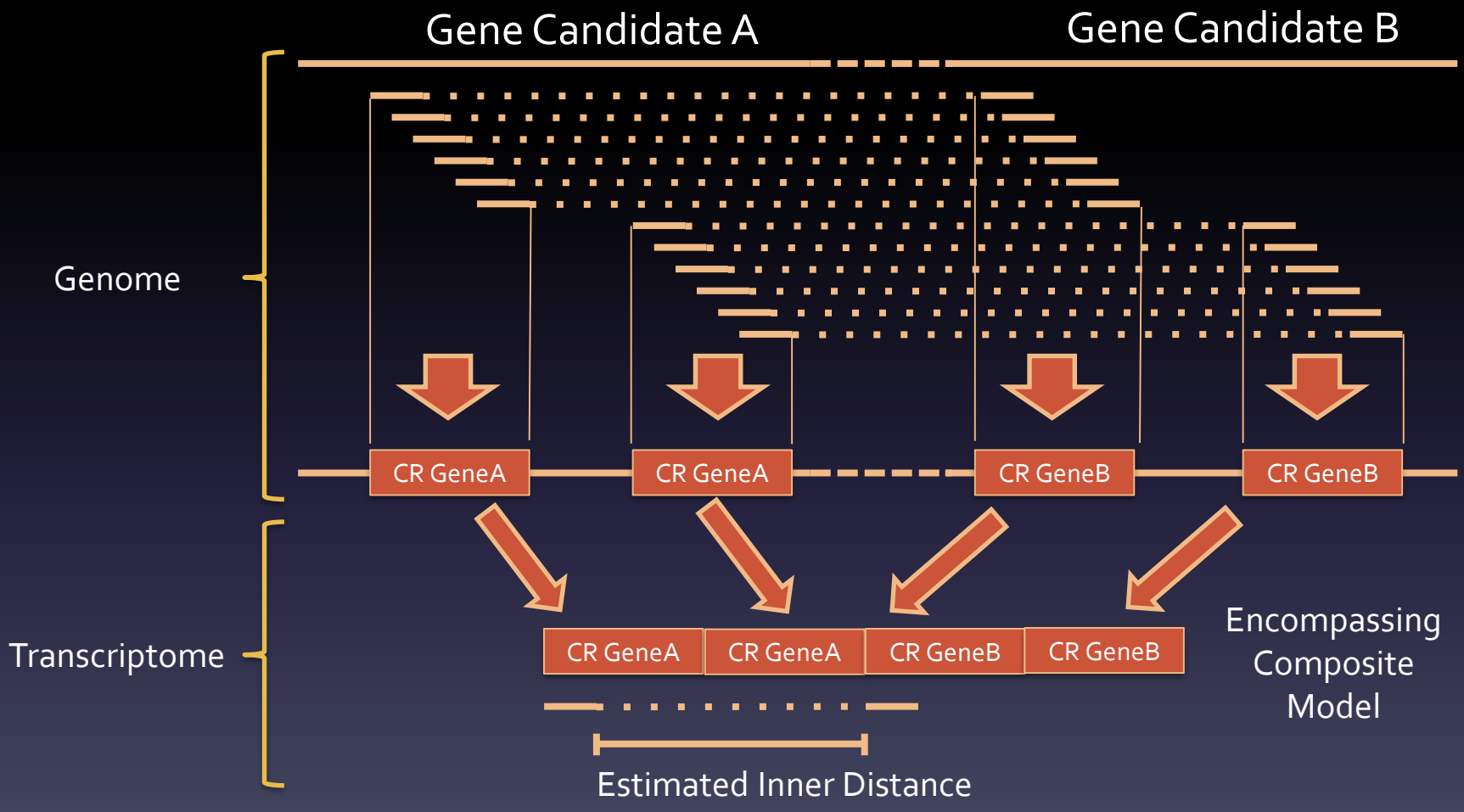
and it is necessary to assembly the fusion transcript

On the Transcriptome

The Fused transcript can be only estimated

because the breakpoint is unknown in this phase

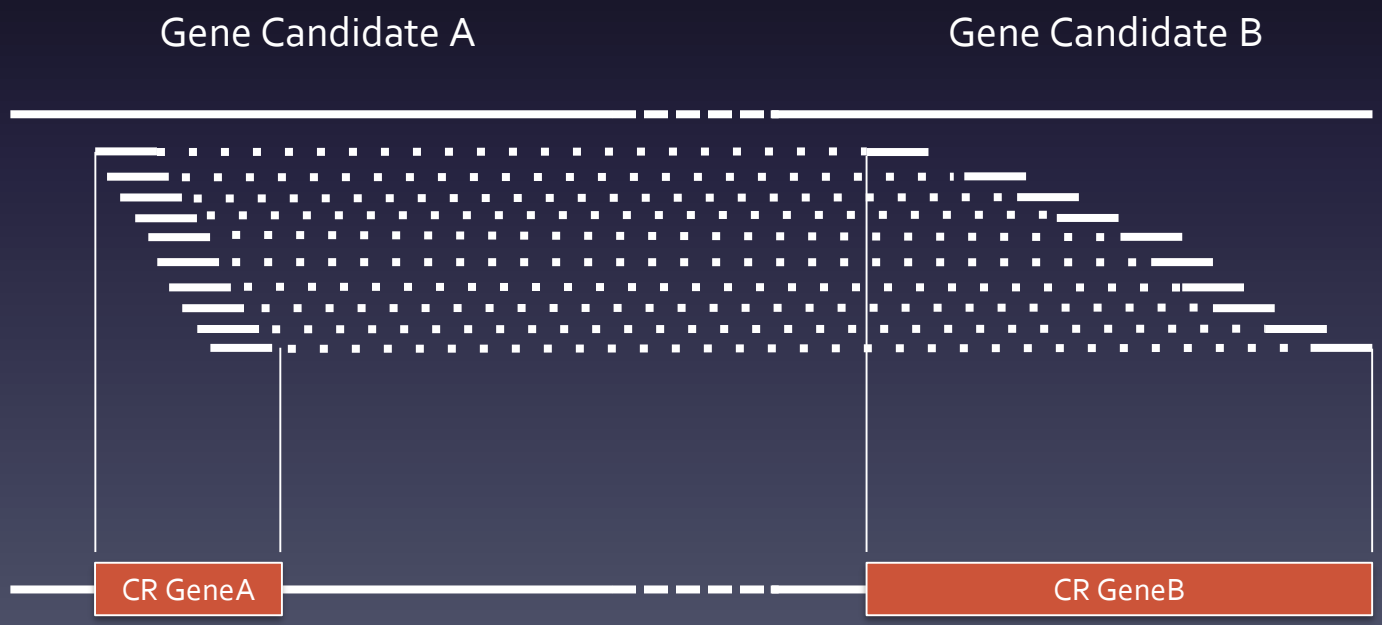
Computing Inner Distance



(the minimum expressed region in between the two genes)

Filters on Encompassing Reads

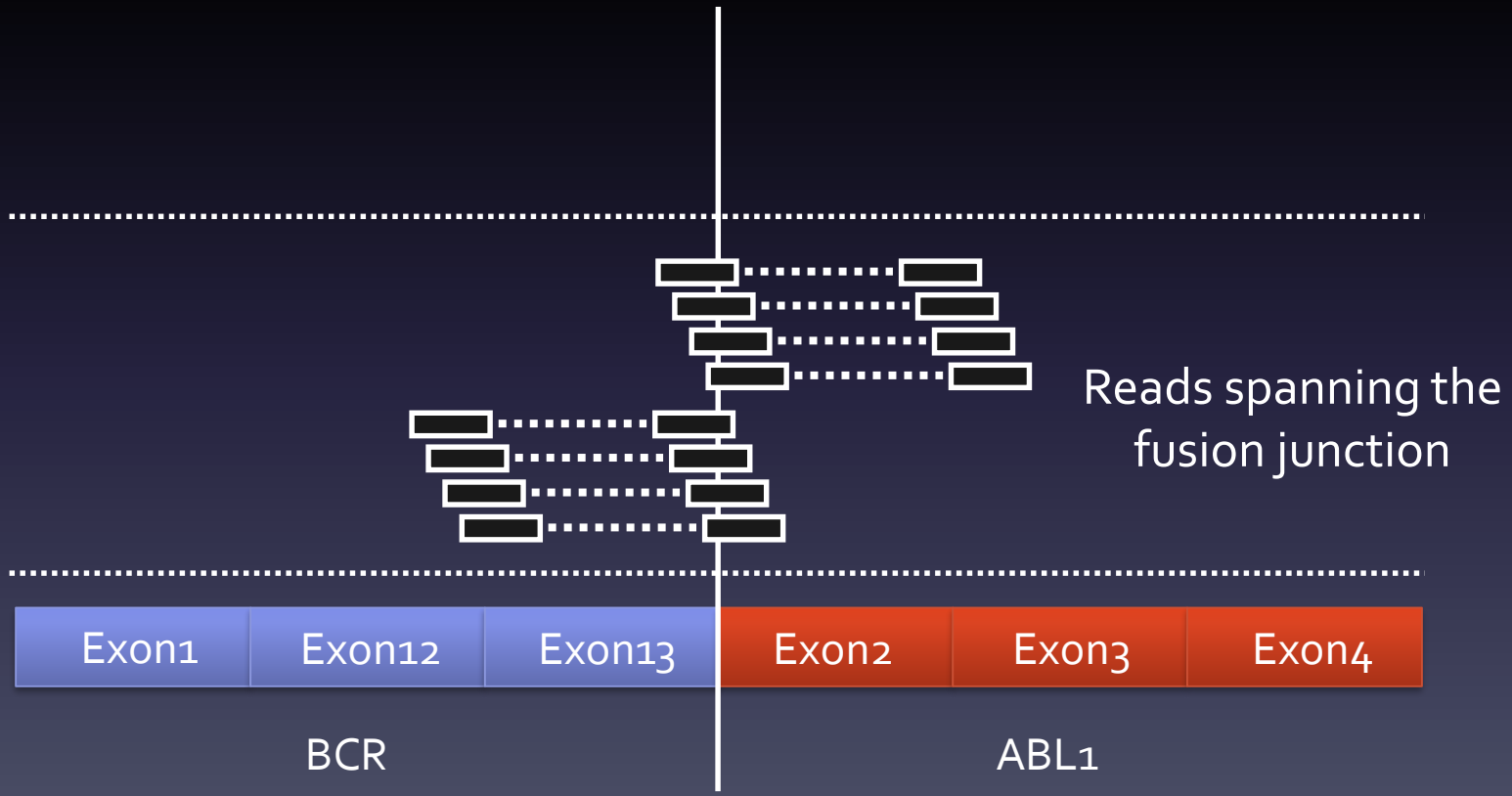
- Asymmetric consensus regions filter
  - Recent experiments demonstrated that asymmetric consensus regions are due to PCR artifacts (*)
  - These candidates are discarded



(*) Edgren et al, Identification of fusion genes in breast cancer by paired-end RNA-sequencing, Genome Biology, 2011

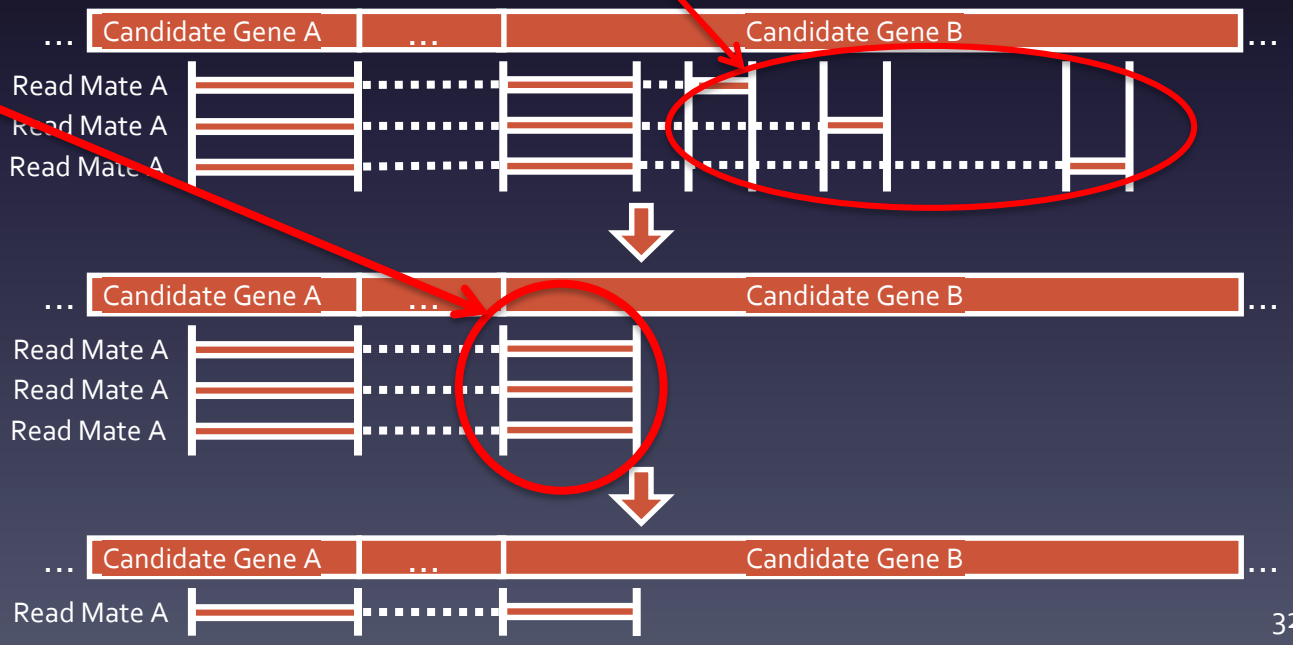
Filtering

- Spanning



Filter on Spanning Reads

- Floating Fragments
  - Mate fragments maps on multiple locations due to homologous regions
  - Alignment tools report multiple read mates
  - Multiple mates are redundant, keeping the single mate mapping on the same region



Filters on Spanning Reads

- PCR artifacts removal
  - Stack-like pattern due to PCR amplification errors(*)
  - Ladder-like pattern -> true positives

False

True

(*) Edgren et al, Identification of fusion genes in breast cancer by paired-end RNA-sequencing, Genome Biology, 2011

Overview on the Features

Features of computational tools for fusion gene detection

Method	Input data				Reference[f]		Fusion junction detection[g]		Assembly[h]
	Type[d]		Format[e]						
	WGS	RNA-Seq	Single-end	Paired-end	Transcriptome	Genome	Split-read	Spanning-read	
Fusion detection specific									
BreakFusion[a]		•		•	•	•			•
ChimeraScan		•		•	•	•	•	•	
Comrad[b]	•	•		•	•	•	•	•	
FusionAnalyser[a]		•		•	•	•	•	•	
deFuse		•		•	•	•	•	•	
FusionMap	•	•	•	•	•	•	•		
FusionHunter		•		•	•	•	•	•	
FusionSeq		•		•	•	•	•	•	
ShortFuse		•		•	•	•	•	•	
SnowShoes-FTD		•		•	•	•	•	•	
SOAPfusion		•		•	•	•	•	•	
Tophat-Fusion		•	•	•	•	•	•		
Structural variant detection									
BreakDancer[c]	•			•		•	•	•	
CREST	•			•		•	•		
GASV	•			•		•		•	
HYDRA	•			•		•		•	
PEMer	•			•		•		•	
R453PlusIToolbox	•		•	•		•	•		
SVDetect	•			•		•		•	
VariationHunter	•			•		•		•	
Others									
R-SAP		•	•	•	•	•	•	•	
Trans-ABySS		•		•			•	•	•
Trinity		•		•			•	•	•