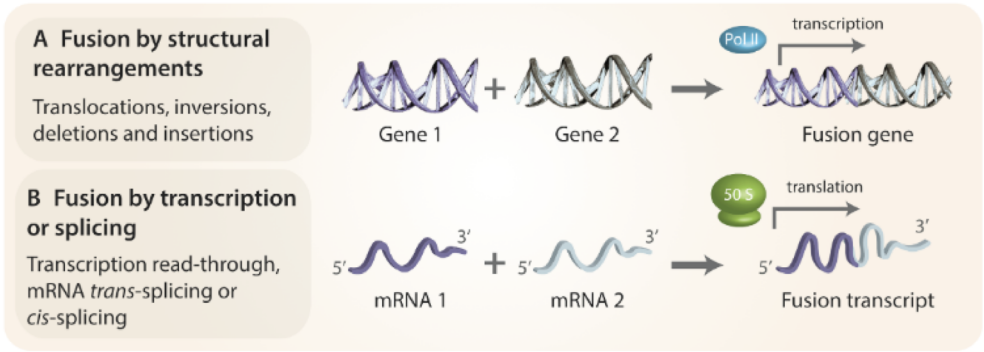
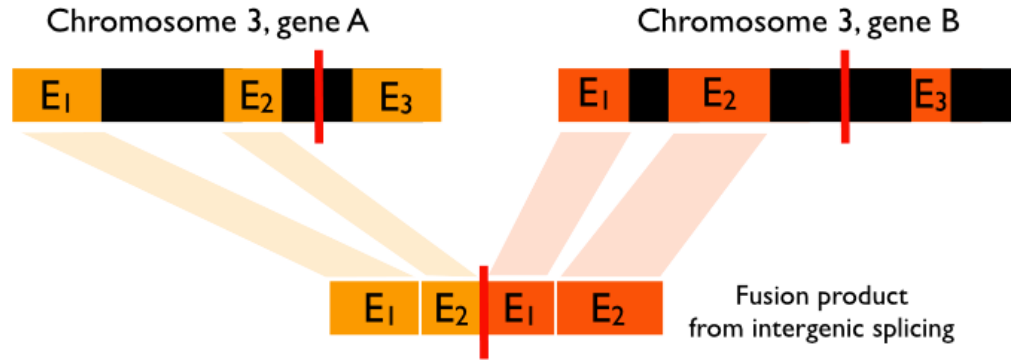
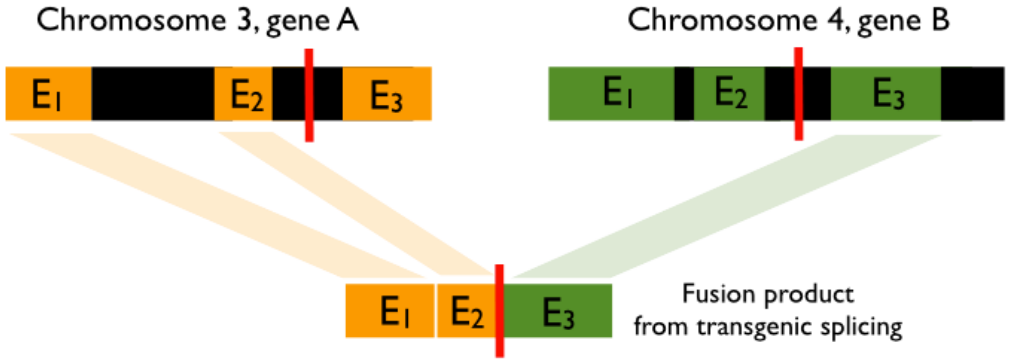
RNAseq applications – Fusion genes

- RNA-seq has the potential to discover genes created by complex chromosomal rearrangements:
  - 'Fusion' genes formed by the breakage and re-joining of two different chromosomes have repeatedly been implicated in the development of cancer.







- Notes:
  - Fusion may not happen at exon boundaries
  - Non-canonical junctions must be considered

RNAseq applications – Fusion genes

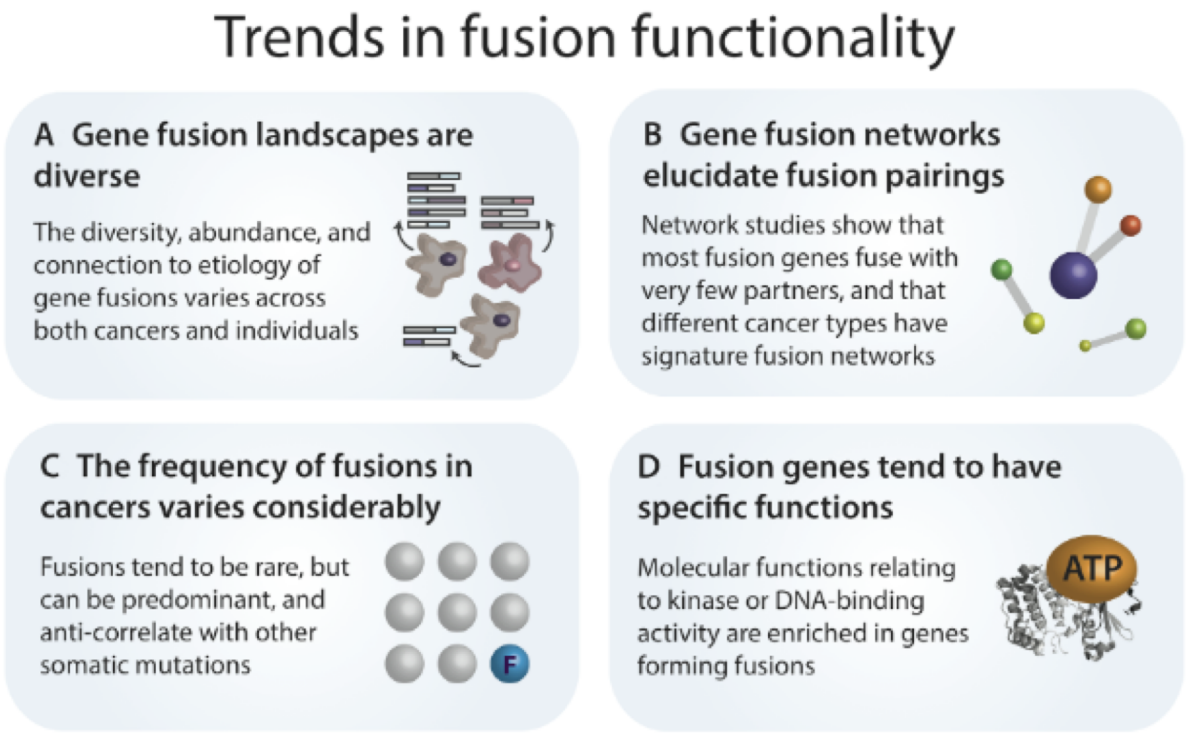
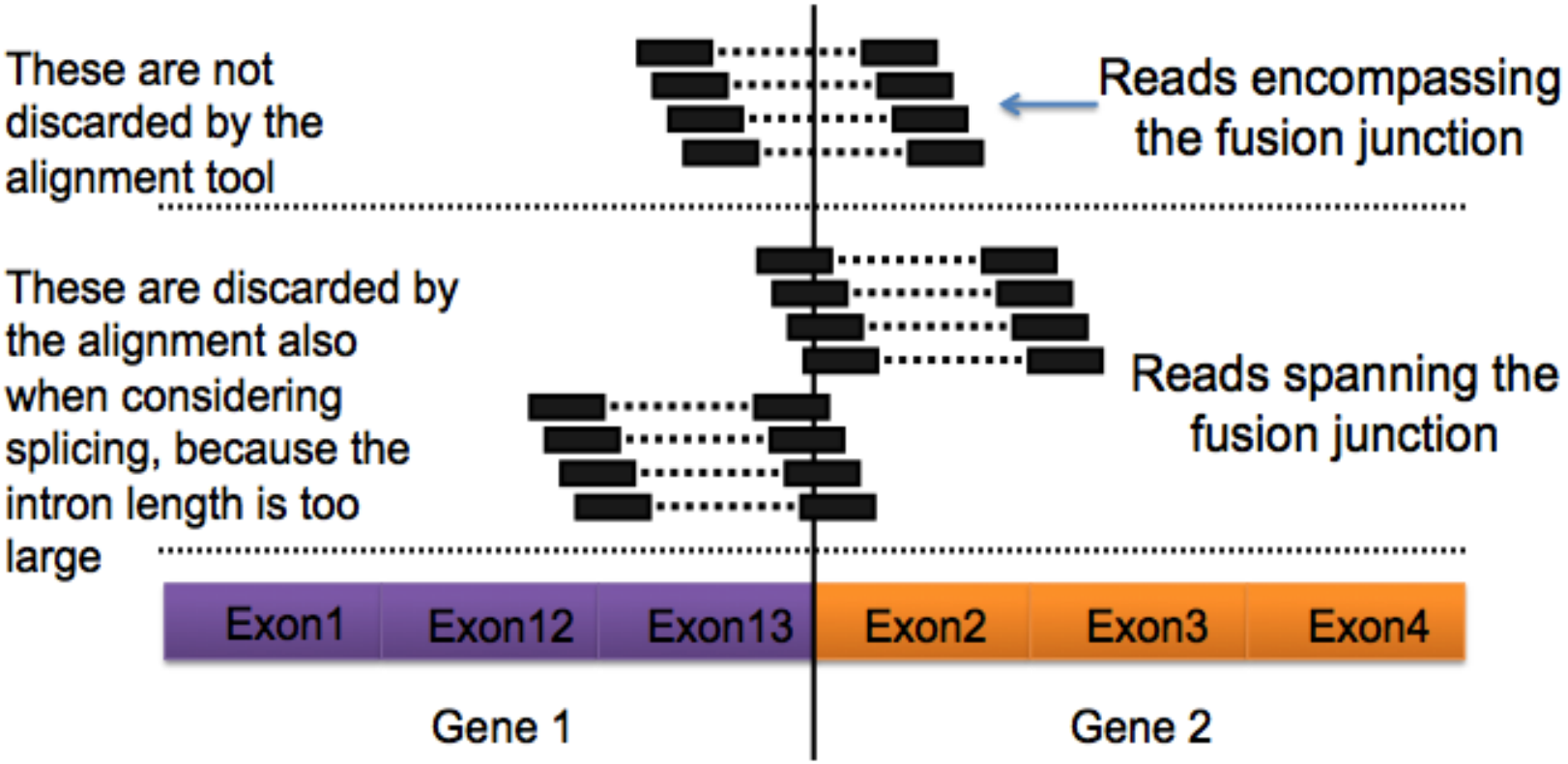
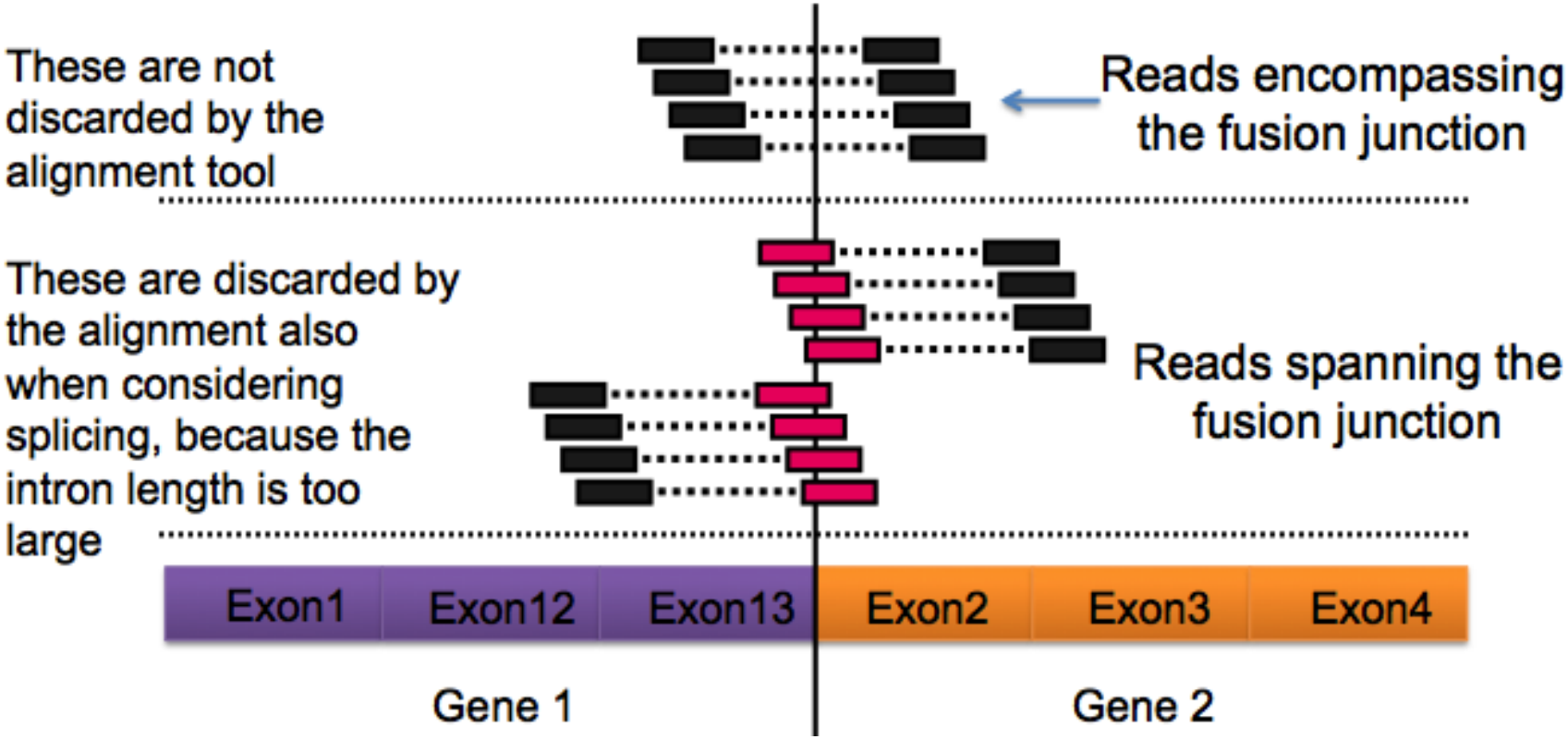


Figure 2. Trends in fusion functionality. (**A**) Recent surveys have uncovered the diverse gene fusion landscapes present in a variety of cancers. (**B**) The frequency of gene fusions varies by cancer type and appears to anti-correlate with frequencies of other somatic mutations at the level of both cancer types and individual tumor samples. (**C**) Gene fusions tend to involve genes with kinase, DNA-binding and chromatin modifying activity. (**D**) Network studies of fusions have identified global and cancer-type-specific patterns in gene partnerships, such as the trend toward most fusion genes only fusing with only one other partner.

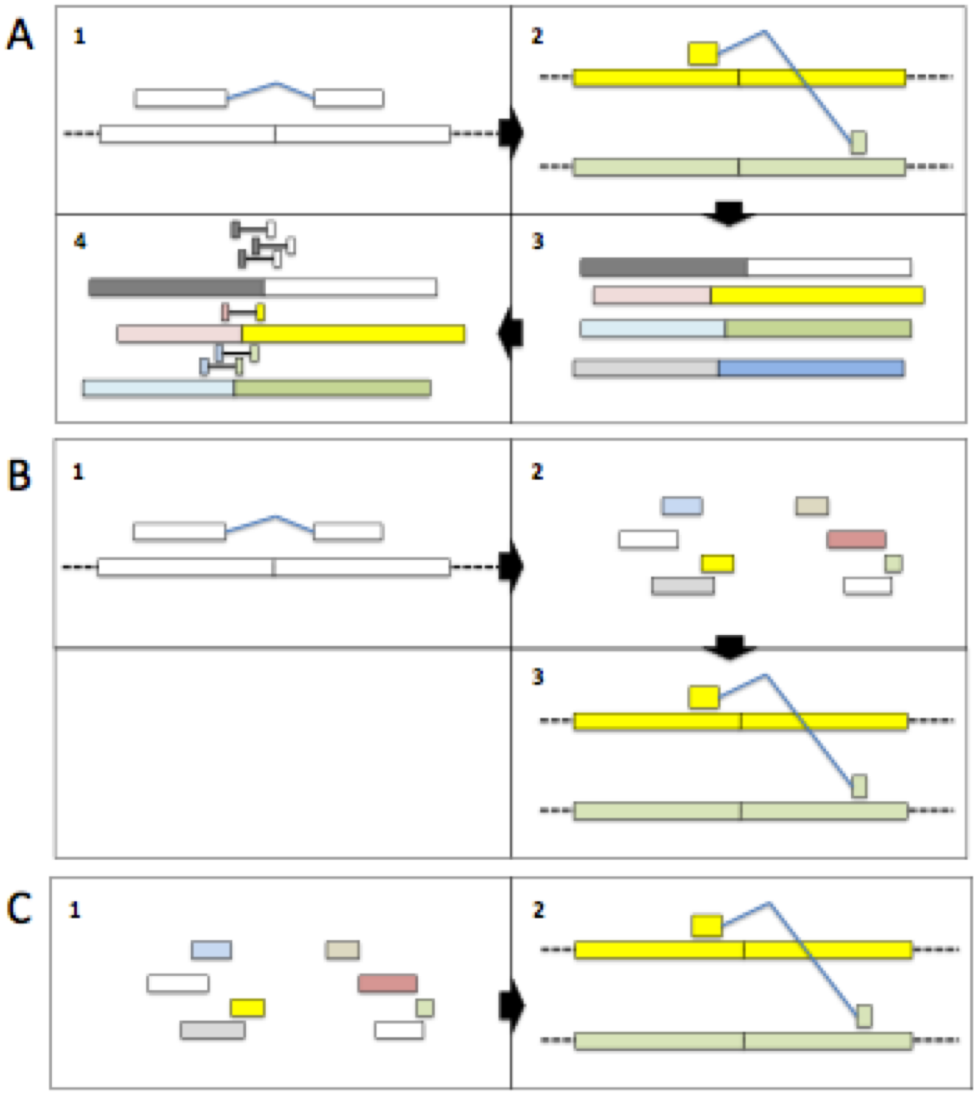
RNAseq applications – Fusion genes



RNAseq applications – Fusion genes



RNAseq applications – Fusion genes



A- Whole pair-end

B- Pair-end + fragmentation

C- Direct fragmentation

Algorithms

- **Whole paired-end**
  - tools align the full-length paired-end reads on a reference and use discordant alignments to generate a set of putative fusion events which are finally selected using several additional pieces of information or filtering steps.

  - DeFuse and FusionHunter

Algorithms

- **Paired-end + fragmentation**
  - the full-length paired-end reads are aligned on a reference and the discordant alignments are used to generate new pseudo-reference.
  - reads unaligned in the first step are fragmented and re-aligned on the pseudo-reference to identify junction-spanning reads. Only the putative fusion events associated with junction-spanning reads are selected as input to the filtering step.
  - TopHat-Fusion, ChimeraScan and Bellerophontes

Algorithms

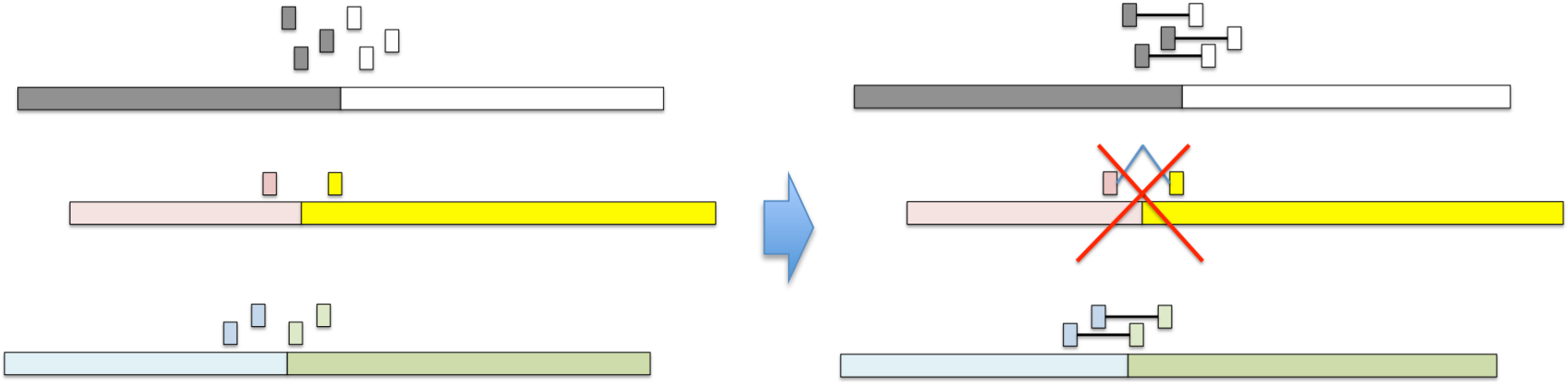
- **Direct fragmentation**
  - Each read is fragmented before the first alignment. The algorithm finds fusion candidates aligning read fragments to a genomic reference.
  - The putative fusion events are then selected implementing a set of filtering steps

  - MapSplice, FusionMap and FusionFinder

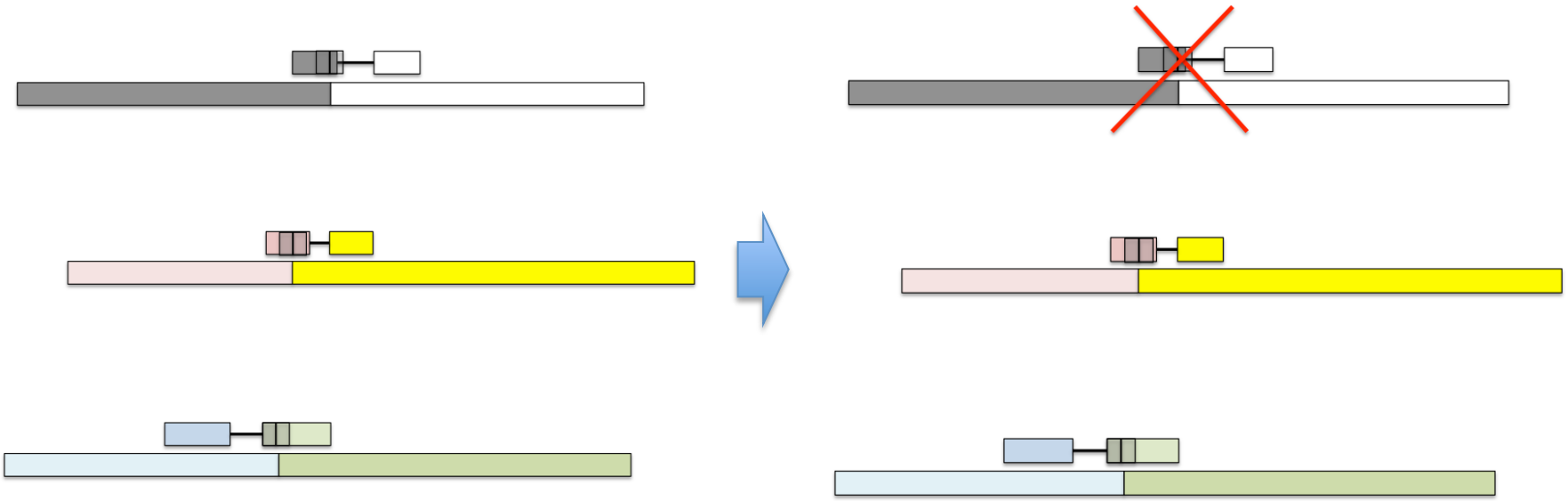
RNAseq applications – Fusion genes

Table 1 Classification of fusion-finder algorithms according to their alignment strategies.	
Tool name	**Macrogroup**
Bellerophontes	Paired-end + Fragmentation
BreakFusion	Whole paired-end
BreakPointer	Statistical information exploiting
ChimeraScan	Paired-end + Fragmentation
deFuse	Whole paired-end
EBARDenovo	Direct fragmentation
EricScript	Whole paired-end
FusionAnalyser	Whole paired-end
FusionFinder	Direct fragmentation
FusionHunter	Whole paired-end
FusionMap	Direct fragmentation
FusionSeq	Whole paired-end
LifeScope	Paired-end + Fragmentation
MapSplice	Direct fragmentation
ShortFuse	Whole paired-end
SnowShoes-FTD	Whole paired-end
SOAPFuse	Whole paired-end
TopHat-Fusion	Paired-end + Fragmentation

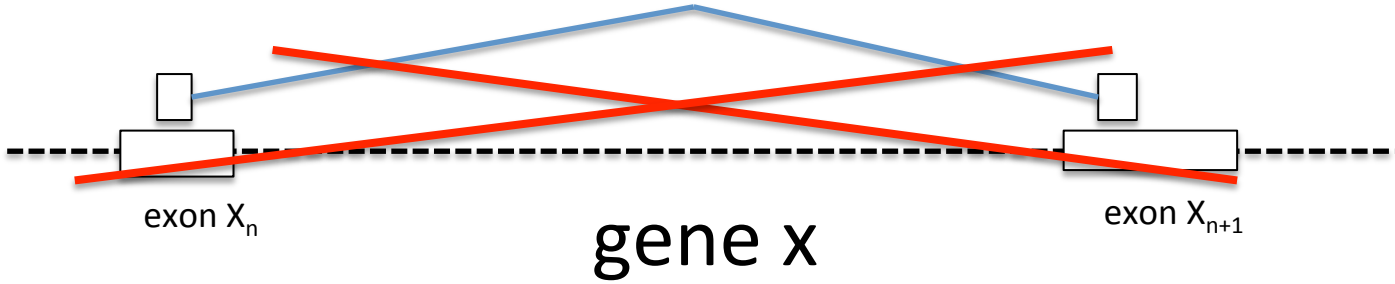
Paired-End Information Filters



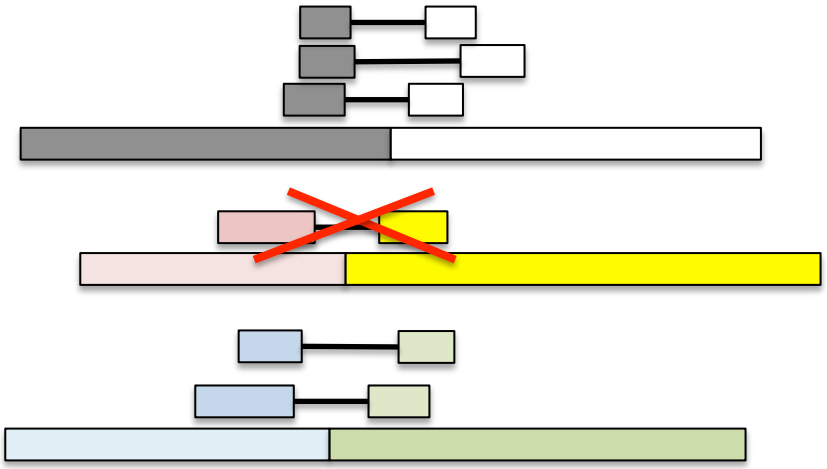
Anchor Length Filters

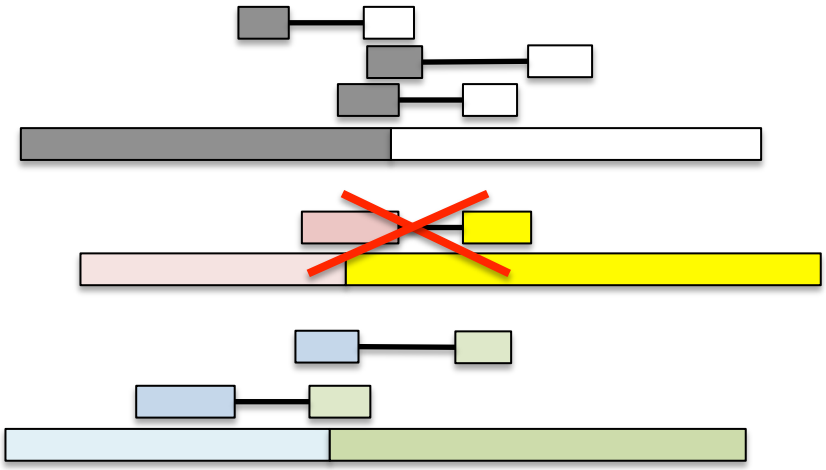


Read Through Transcripts Filter

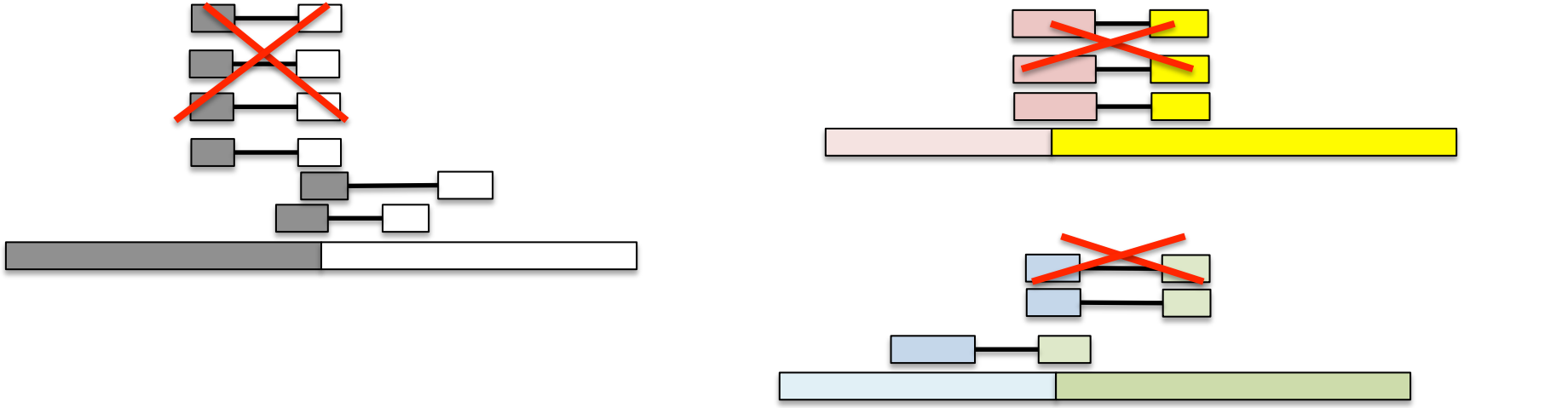


Junction Encompassing Spanning Reads Filter

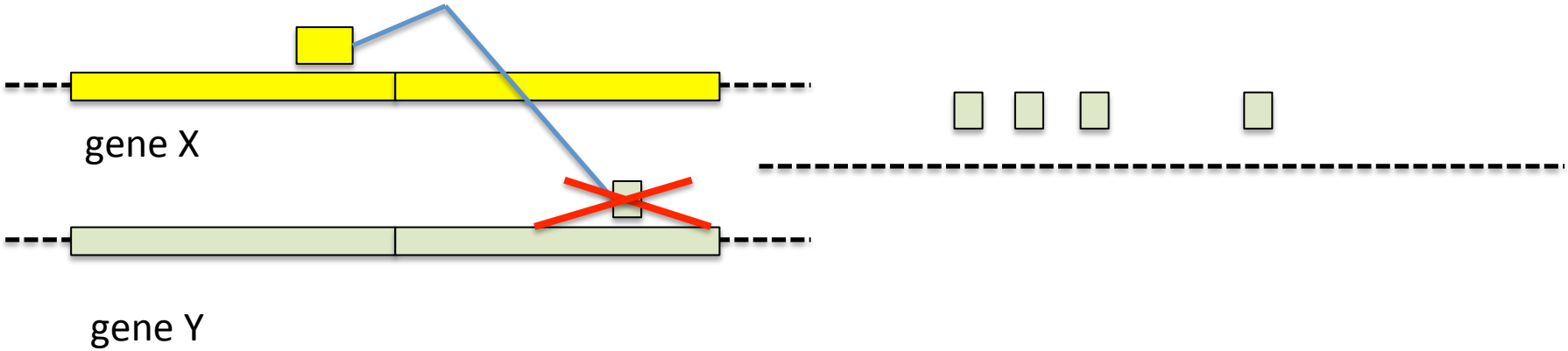




PCR Artifact Filter



Homology Filter

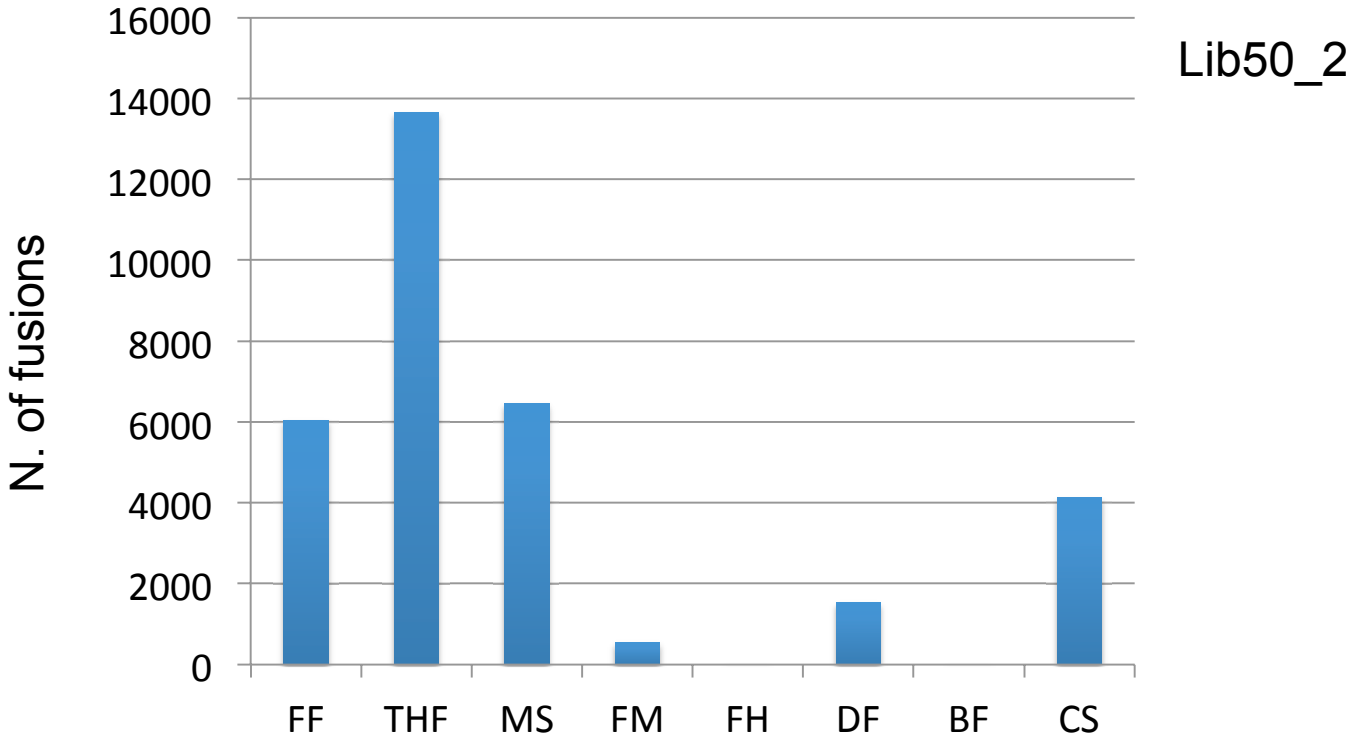


RNAseq applications – Fusion genes

Table 2 Consideration of filters implemented by each fusion-finder algorithm.

Tool name	Paired-end information	Anchor length	Read-through transcripts	Junction spanning reads	PCR artifact	Homology	Scoring	Reads quality	Encompassing reads	Black list	Statistics	Additional filters	
Bellerophontes	X		X	X	X				X			Ambiguous reads	
BreakFusion											X		
BreakPointer					X		X						
ChimeraScan	X	X											
deFuse	X												
EBARDenovo													
EricScript					X	X	X				X	Junction homology	
FusionAnalyser			X			X			X	X			
FusionFinder	X		X			X						Antisense	
FusionHunter		X	X	X	X								
FusionMap			X	X	X		X	X		X			
FusionSeq						X	X	X		X		Comparison chimera expression with general expression	
LifeScope								X				Junction evidence graph	
MapSplice											X	Canonical junctions	Introns length
ShortFuse				X	X		X				X		Reads from Spliceosome components
SnowShoesFTD			X	X		X		X	X			Fusion genes orientation	excessive putative junction point
SOAPFuse	X			X		X						Read trimming	
TopHatFusion		X	X			X							

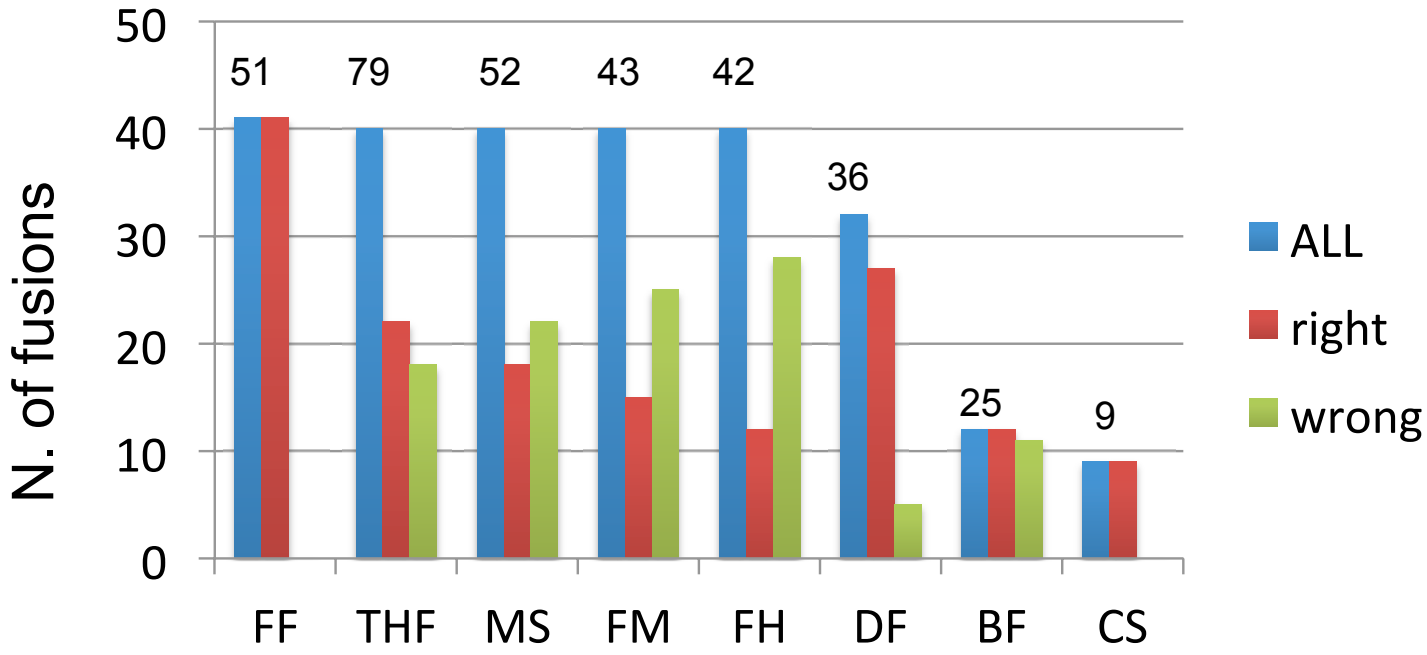
Testing false discovery rate of fusion detection tools *(negative_set)*



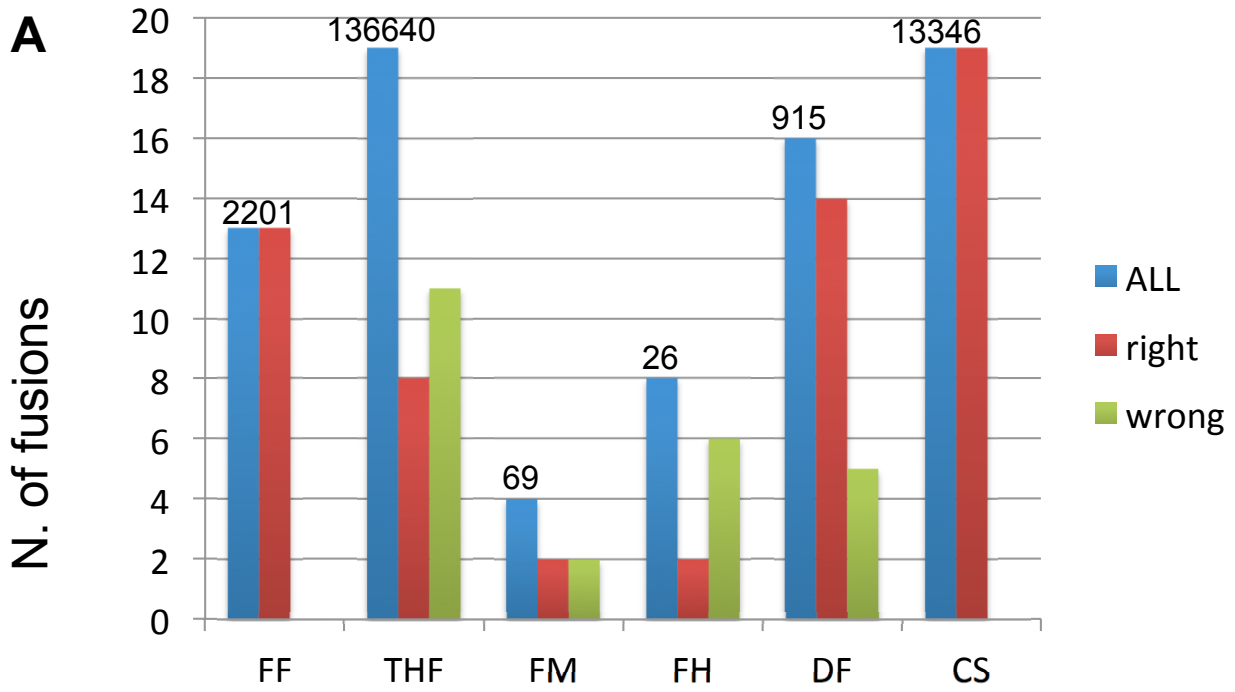
Testing sensitivity of fusion detection tools

- In this analysis of sensitivity we considered three parameters:
  - the total number of true positive fusions detected by the different tools (called **ALL**)
  - the number of true positive fusions detected with the correct orientation of the two genes (called **right**)
  - the number of true positive fusions detected with erroneous orientation of the two genes (called **wrong**).

FM_se: synthetic, 50 fusion events

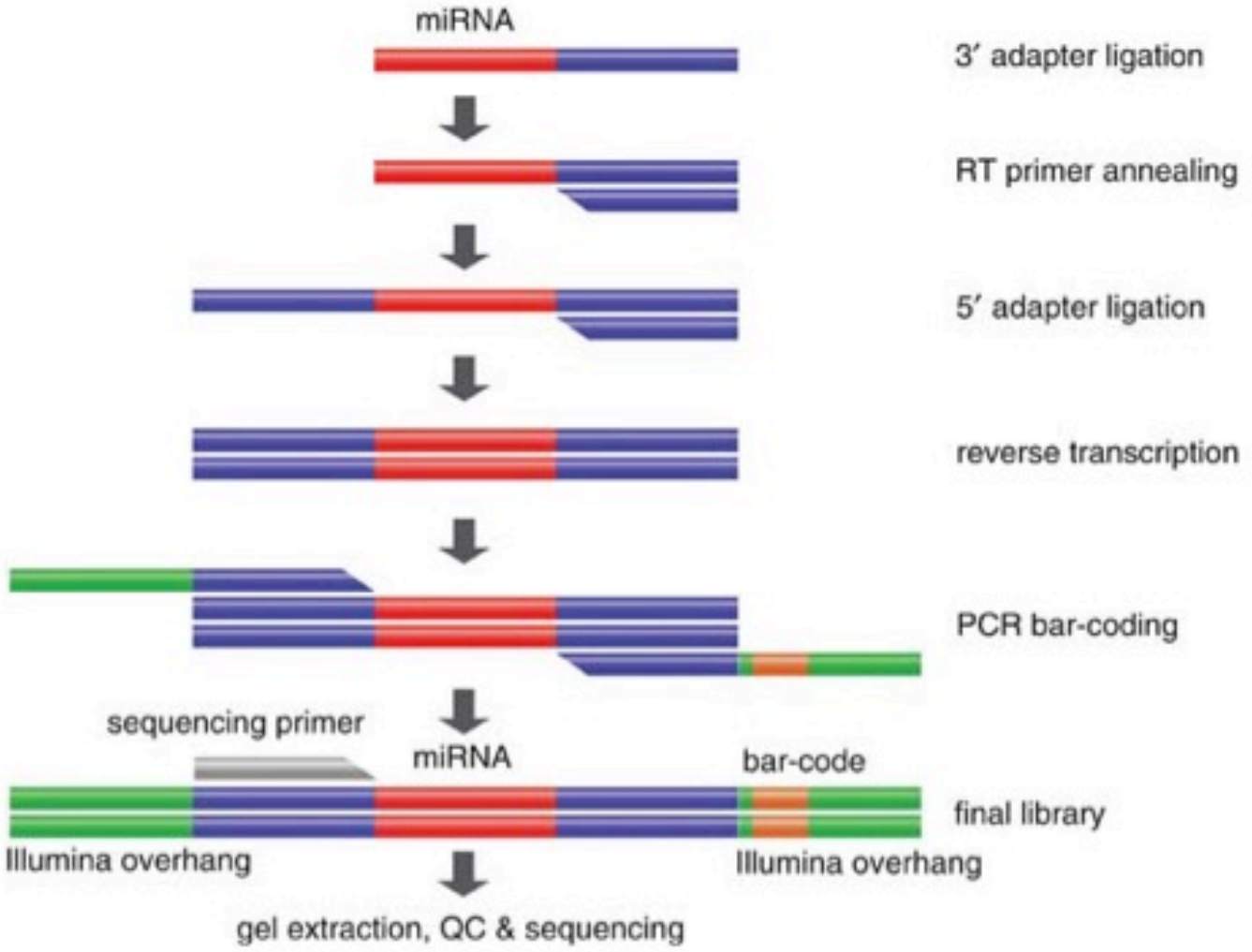


Edgren_se: real, 27 experimentally validated fusion genes

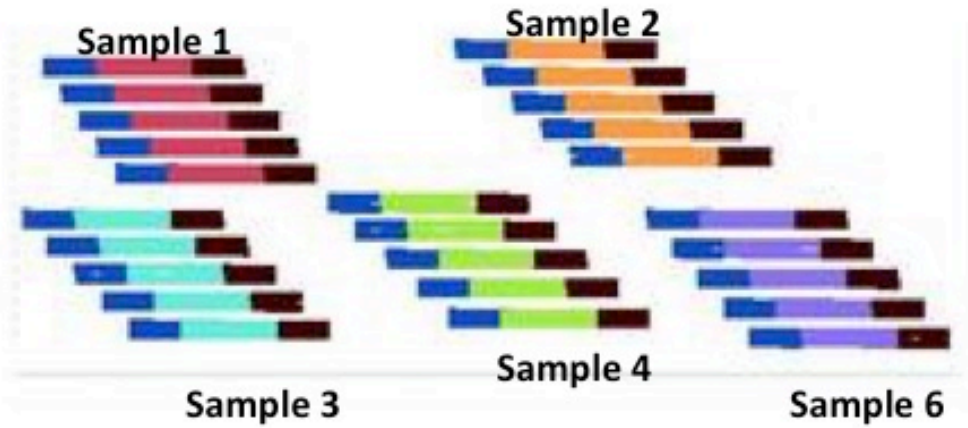


synthetic datasets encompassing fusion events may not fully catch the complexity of a RNA-seq experiment

RNAseq applications – microRNA



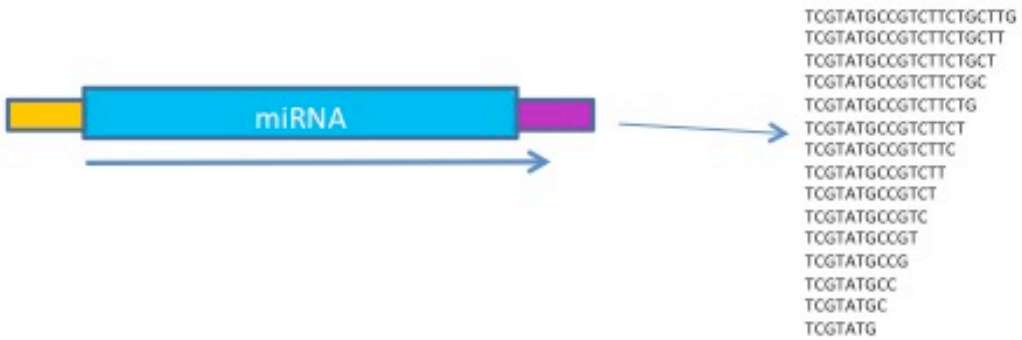
RNAseq applications – microRNA



- Barcodes are unique sequence identifiers added to samples during library construction.
- Once barcodes are added, multiple libraries can be pooled together for emulsion PCR/cluster generation and sequencing.

RNAseq applications – microRNA

- Since miRNAs are shorter then 50 nts some of the sequences collected are contaminated by linkers.

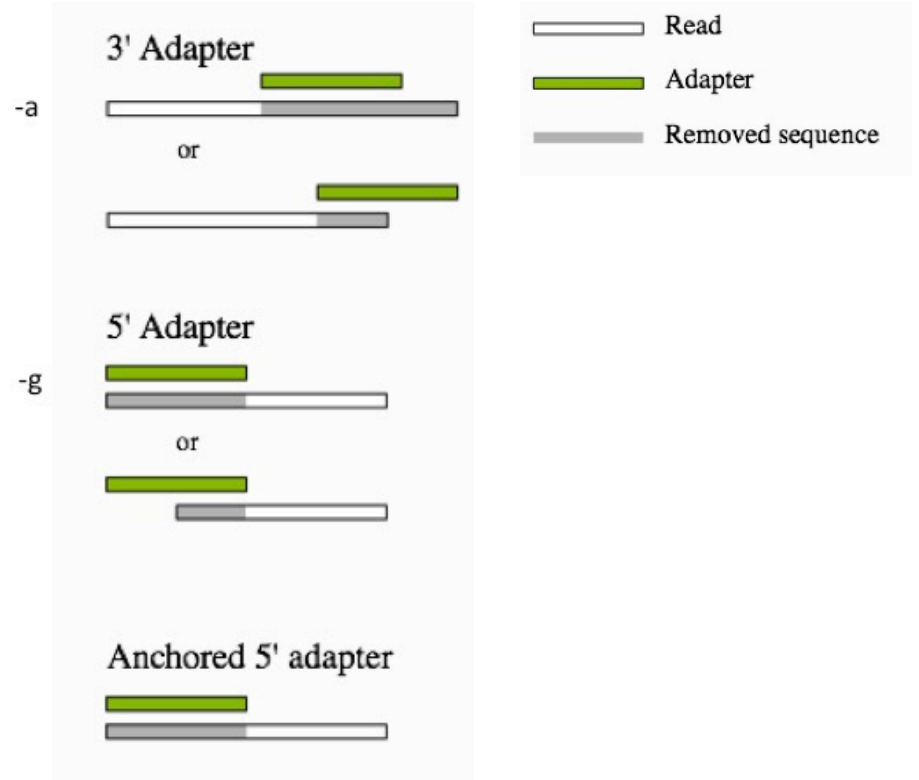


- Adapter_trim/Cutadapt: perl/python scripts for removing adapters

RNAseq applications – microRNA

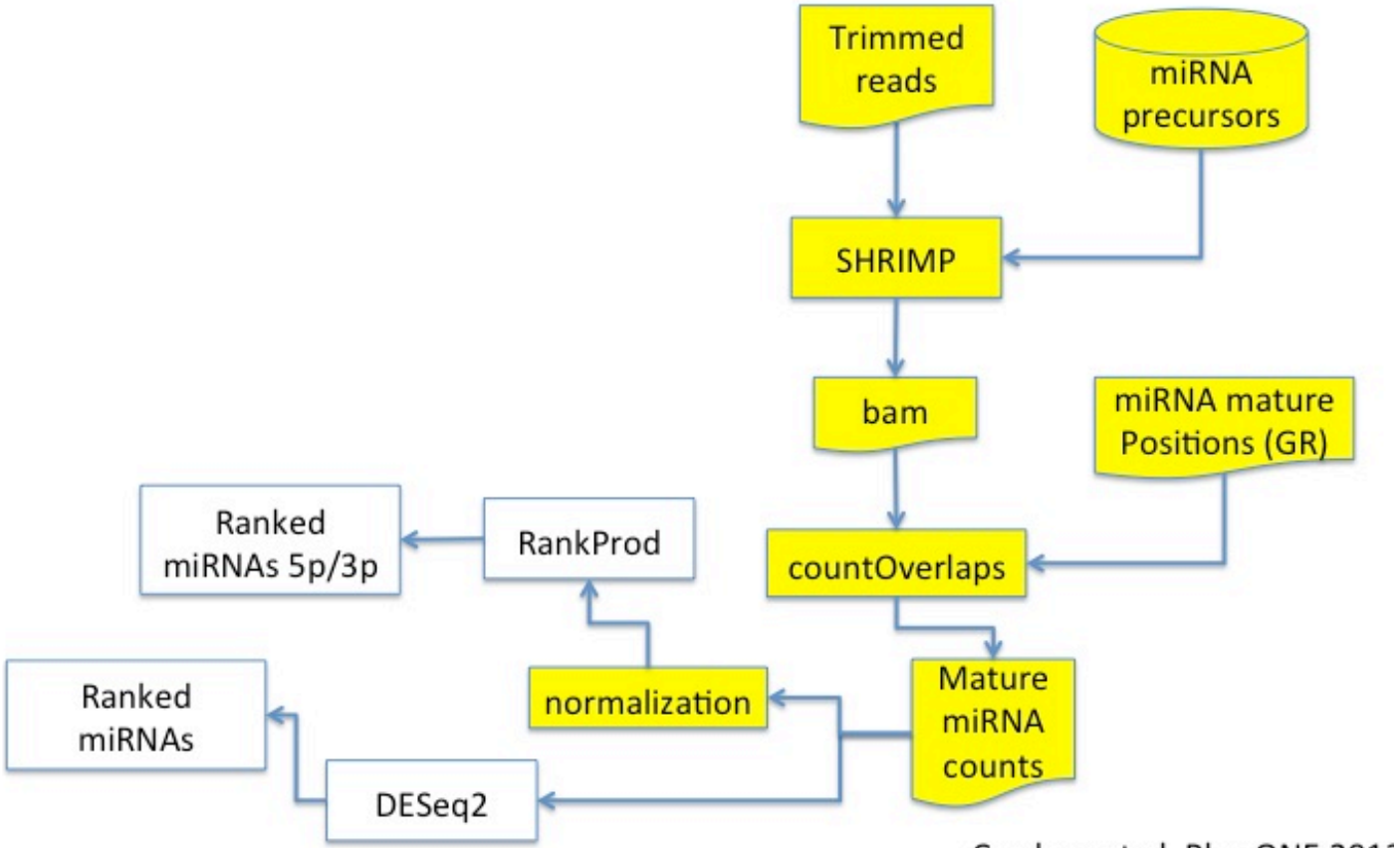
cutadapt

- It is an open-source tool which removes adapter sequences from high-throughput sequencing reads
- It is mainly written in ***Python***, but the alignment algorithm is implemented in ***C*** as a ***Python extension module.***
- It it can be downloaded at https://code.google.com/p/cutadapt/



Beforetrimming	Aftertrimming	Adaptertype
MYSEQUENCEADAPTERSOMETHING	**MYSEQUENCE**	3'adapter
MYSEQUENCE**ADAPTER**	**MYSEQUENCE**	3'adapter
MYSEQUENCE**ADAP**	**MYSEQUENCE**	3'adapter
M**ADAPTER**	**M**	3'adapter
ADAPTER**MYSEQUENCE**	**MYSEQUENCE**	5'adapter
DAPTER**MYSEQUENCE**	**MYSEQUENCE**	5'adapter
TER**MYSEQUENCE**	**MYSEQUENCE**	5'adapter

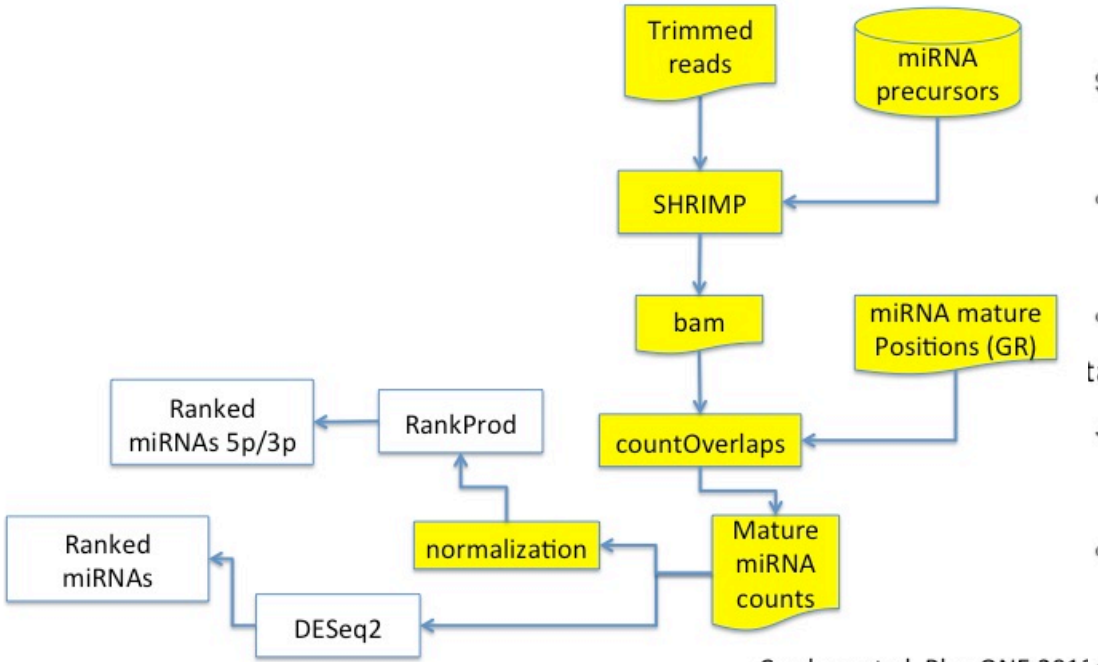
RNAseq applications – microRNA

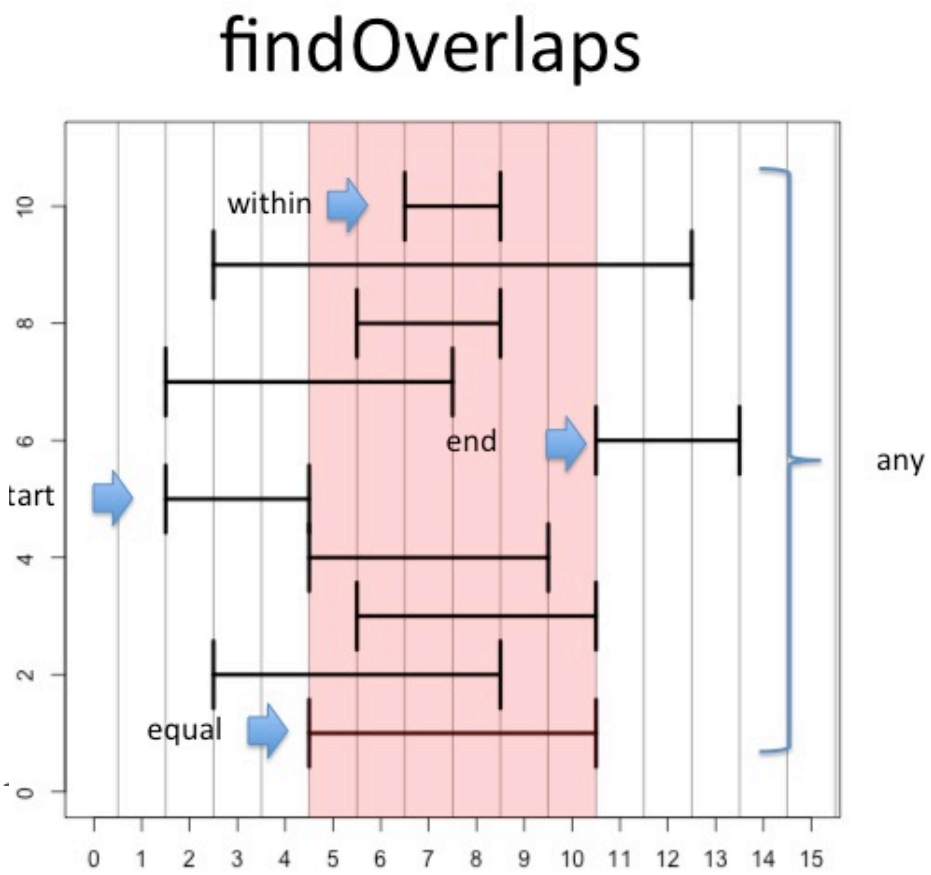


RNAseq applications – microRNA

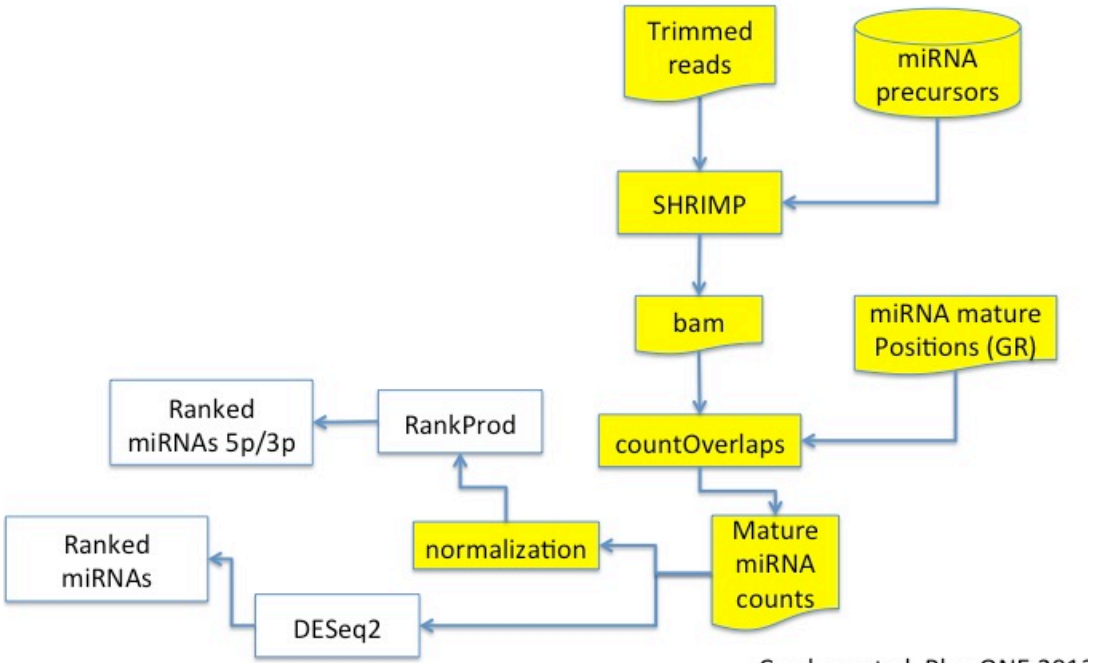
- Whole genome
  - Some microRNAs are duplicated
    - In alignment for counting, reads mapping in multiple genome locations are discarded
    - The use of an efficient segmentation algorithm might discriminate between miR and miR*
- miRbase precursors
  - In alignment for counting, miR*, -5P, -3P miRs are associated to its precursor

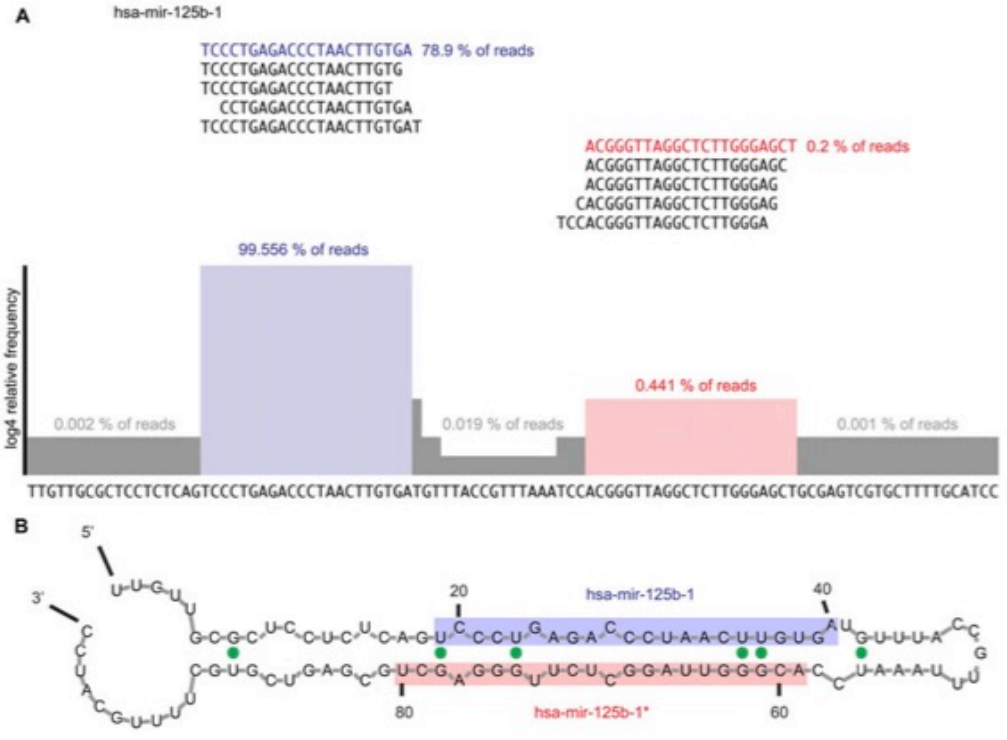
RNAseq applications – microRNA





RNAseq applications – microRNA

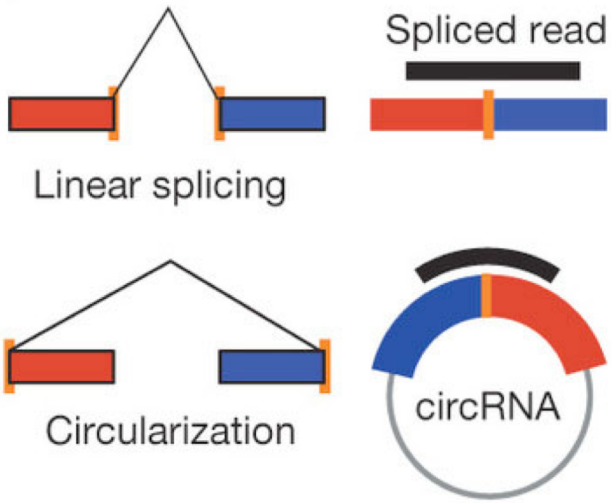


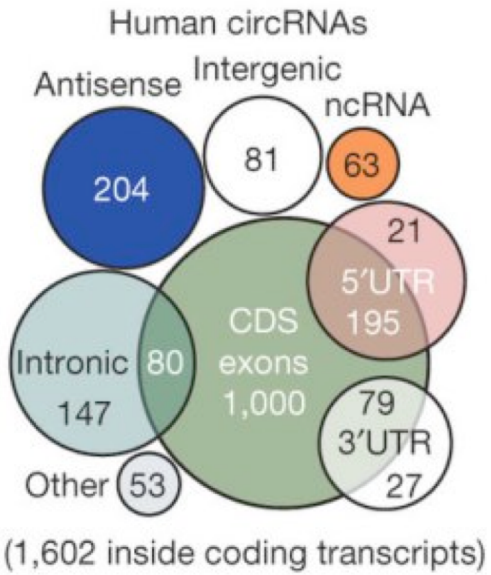


RNAseq applications – circRNA

- **Circular RNA (circRNAs) are abundant, and are found in Human cells**
  - *There are between 25,000 and 100,000 circular RNA species per cell!*
  - *They far outnumber linear RNAs*
- **CircRNAs are transcribed from DNA but are not translated into proteins**
- **CircRNAs explains several phenomena observed in DNA**
  - *including non-colinear splicing, scrambling of introns, and certain non-coding antisense transcripts.*
- **CircRNAs may be implicated in disease processes and aging**.
  - In particular, splice variants of one long circular RNA known as ANRIL located at the exact location of the 9p21.3 SNP reproduce the same phenotype as the 9p21.3 "risk allele" seen with atherosclerotic disease.
- **The research/genetic establishments rejected the idea of circular RNA for a long time**, so a great deal is yet to be learned about them.
- **CircRNAs are evolutionarily conserved**
  - *passed on from generation to generation*
- **CircRNAs live in to cytoplasm and are long lasting CircRNAs offer large number of docking sites for miRNAs**,
  - including ones which are capable of silencing genes – *they are like coat racks for siRNAs*
- **The net impact of circRNAs on gene expression can be significant**
  - *because their siRNA docking sites are competitive with those on genes*

RNAseq applications – microRNA



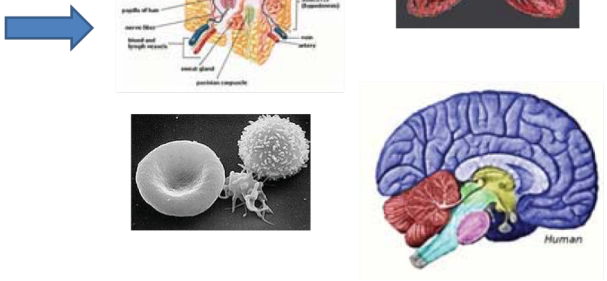


- CircRNAs junction point resembles the fusion break point in chimeras.
- CircRNAs formation mainly involves coding exons.

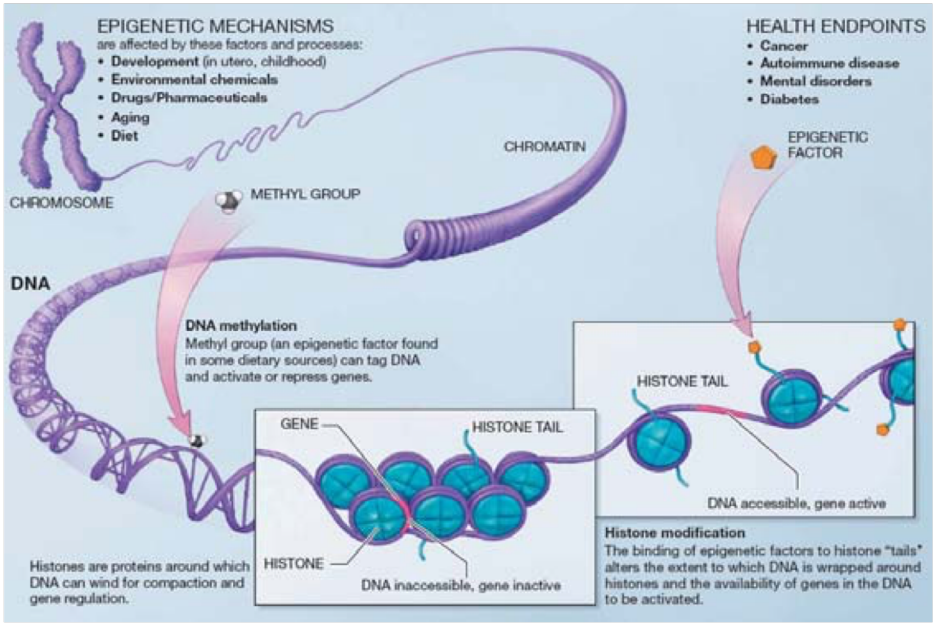
Deep Sequencing technology - other applications

One Genome – Many Cell Types

ACCAGTTACGACGGTCA
GGGTACTGATACCCCAA
ACCGTTGACCGCATTTA
CAGACGGGGTTTGGGTT
TTGCCCCACACAGGTAC
GTTAGCTACTGGTTTAG
CAATTTACCGTTACAAC
GTTTACAGGGTTACGGT
TGGGATTTGAAAAAAAG
TTTGAGTTGGTTTTTTC
ACGGTAGAACGTACCGT
TACCAGTA

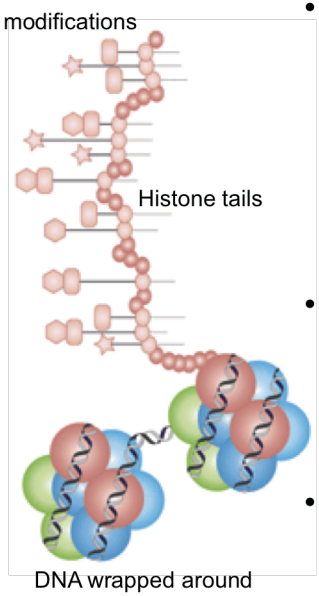


Diverse epigenetic modifications



Deep Sequencing technology - other applications

Diversity of epigenetic modifications



- 100+ different histone modifications
  - Histone protein → H3/H4/H2A/H2B
  - AA residue → Lysine4(K4)/K36…
  - Chemical modification → Met/Pho/Ubi
  - Number → Me-Me-Me(me3)
  - Shorthand: H3K4me3, H2BK5ac
- In addition:
  - DNA methylation primarily at CpG
  - Nucleosome positioning
  - DNA accessibility
- The constant struggle of gene regulation
  - TF/histone/nucleo/GFs/Chrom compete

Ongoing epigenomic mapping projects

- Mapping multiple modifications
- In multiple cell types
- In multiple individuals
- In multiple species
- In multiple conditions
- With multiple antibodies
- Across the whole genome

- First wave published
- Lots more in pipeline
- Time for analysis!

Deep Sequencing technology - other applications

ChIP-chip and ChIP-Seq technology

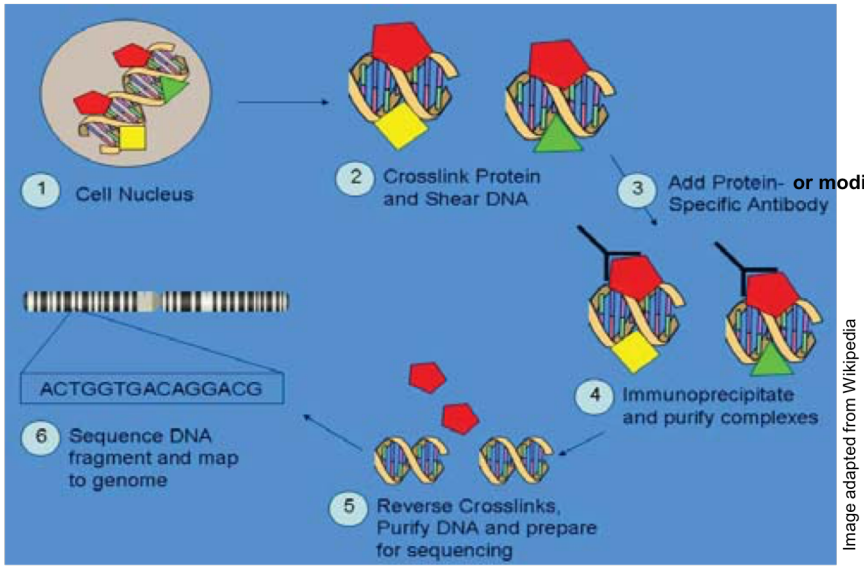


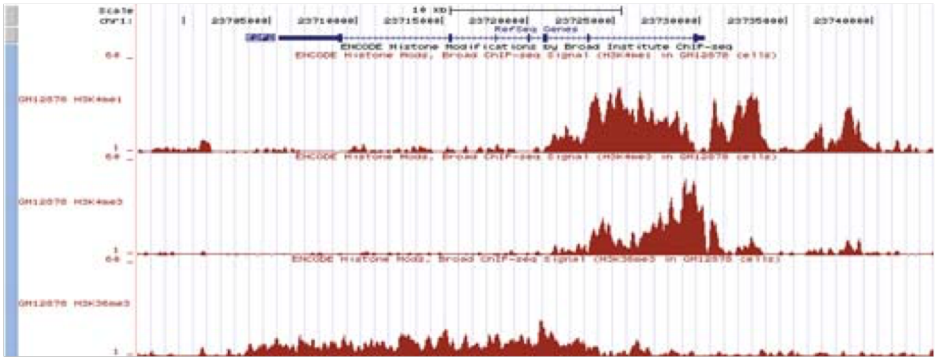
Image adapted from Wikipedia

Modification-specific antibodies → Chromatin Immuno-Precipitation

followed by: ChIP-chip: array hybridization

ChIP-Seq: Massively Parallel Next-gen Sequencing

ChIP-Seq Histone Modifications: What the raw data looks like



- Each sequence tag is 30 base pairs long
- Tags are mapped to unique positions in the ~3 billion base reference genome
- Number of reads depends on sequencing depth. Typically on the order of 10 million mapped reads.

Deep Sequencing technology - other applications