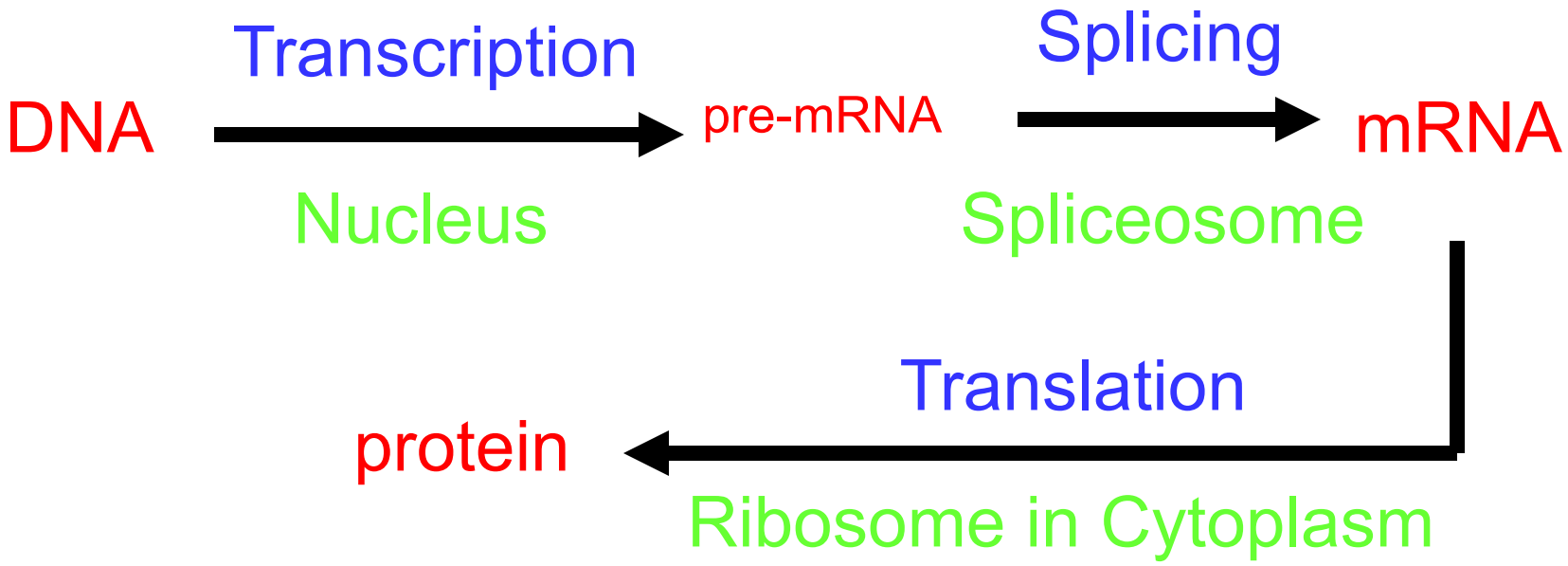
CS681: Advanced Topics in Computational Biology

Week 8 Lectures 2-3

Can Alkan
EA224
calkan@cs.bilkent.edu.tr

http://www.cs.bilkent.edu.tr/~calkan/teaching/cs681/

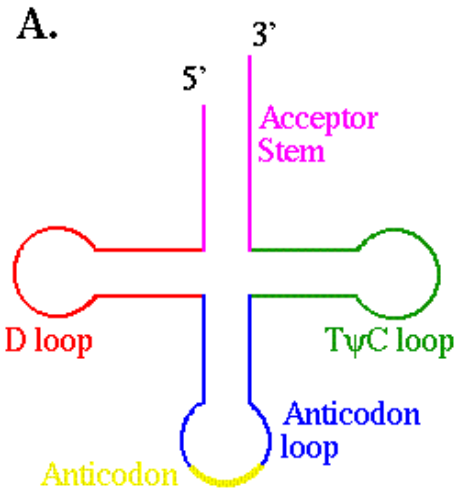
Central dogma of biology

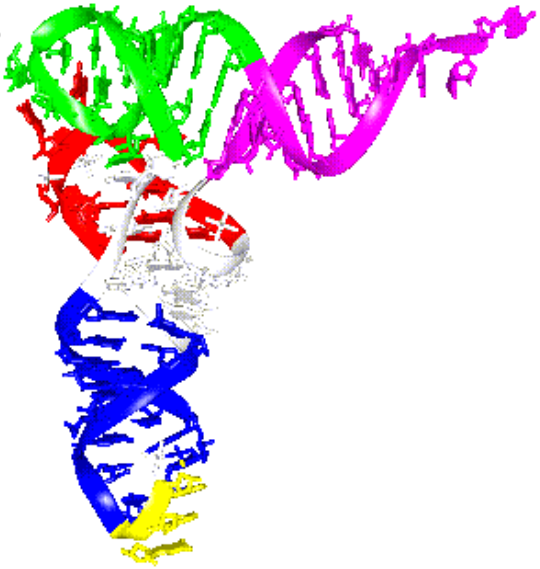


- Base Pairing Rule: A and T or U is held together by 2 hydrogen bonds and G and C is held together by 3 hydrogen bonds.
- Note: Some RNA stays as RNA (ie tRNA,rRNA, miRNA, snoRNA, etc.).

RNA

- RNA is similar to DNA chemically. It is usually only a single strand. T(hymine) is replaced by U(racil)
- Some forms of RNA can form secondary structures by "pairing up" with itself. This can have change its properties





DNA and RNA can pair with each other.

tRNA linear and 3D view: http://www.cgl.ucsf.edu/home/glasfeld/tutorial/trna/trna.gif

RNA, continued

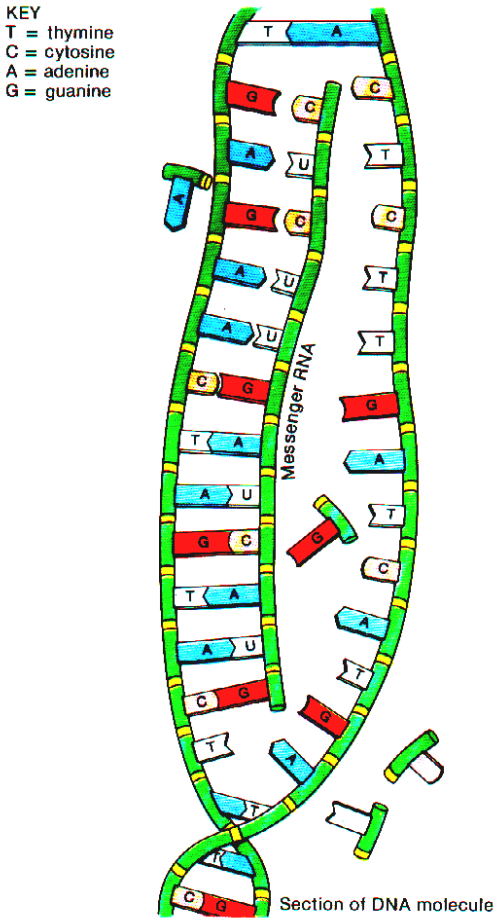
- Several types exist, classified by function
  - mRNA – this is what is usually being referred to when a Bioinformatician says “RNA”. This is used to carry a gene’s *m*essage out of the nucleus.
  - tRNA – *t*ransfers genetic information from mRNA to an amino acid sequence
  - rRNA – *r*ibosomal RNA. Part of the ribosome which is involved in translation.
  - Non-coding RNAs (ncRNA): not translated into proteins, but they can regulate translation
    - miRNA, siRNA, snoRNA, piRNA, lncRNA

RNA vs DNA

- DNA:
  - Double helix
  - Alphabet = {A, C, G, T}
- RNA:
  - Single strand
  - Alphabet = {A, C, G, U}
  - Folding
    - Since RNA is single stranded, it folds onto itself
      - secondary and tertiary structures are important for function

Transcription

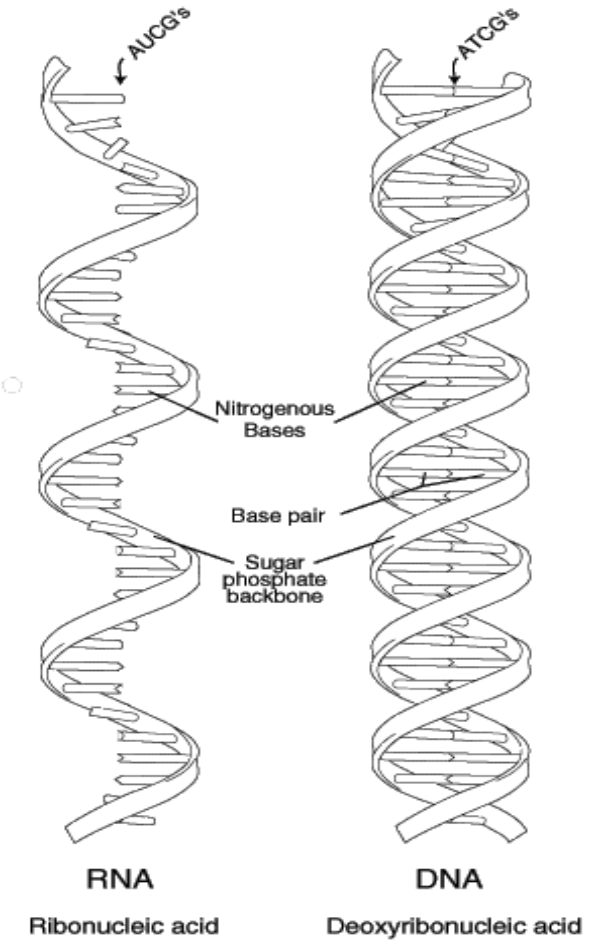
- The process of making RNA from DNA
- Catalyzed by “transcriptase” enzyme
- Needs a promoter region to begin transcription.
- ~50 base pairs/second in bacteria, but multiple transcriptions can occur simultaneously



http://ghs.gresham.k12.or.us/science/ps/sci/ibbio/chem/nucleic/chpt15/transcription.gif

DNA → RNA: Transcription

- DNA gets transcribed by a protein known as *RNA-polymerase*
- This process builds a chain of bases that will become mRNA
- RNA and DNA are similar, except that RNA is single stranded and thus less stable than DNA
  - Also, in RNA, the base uracil (U) is used instead of thymine (T), the DNA counterpart



Transcription, continued

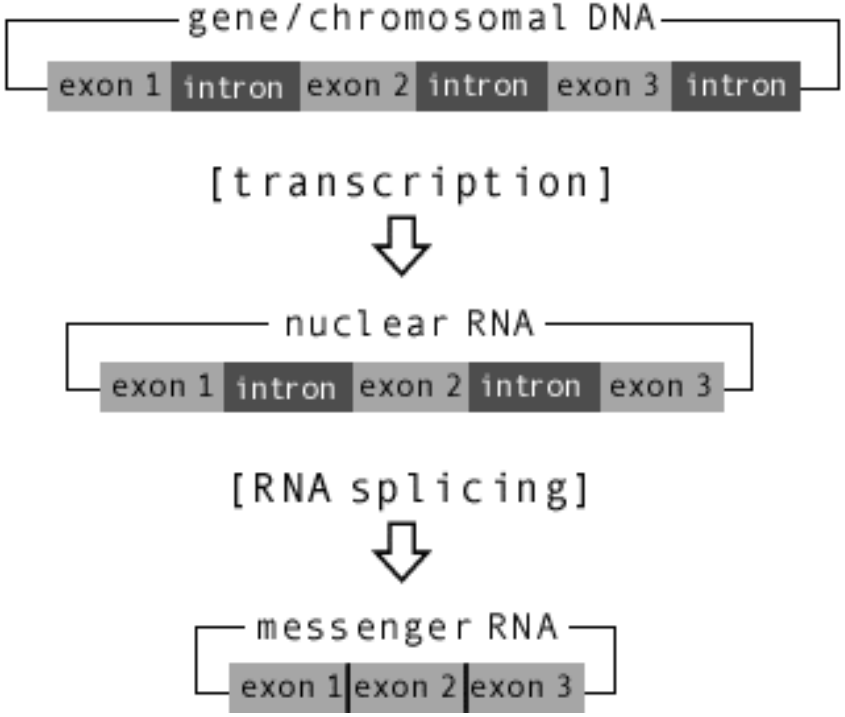
- Transcription is highly regulated. Most DNA is in a dense form where it cannot be transcribed.
- To begin transcription requires a promoter, a small specific sequence of DNA to which polymerase can bind (~40 base pairs “upstream” of gene)
- Finding these promoter regions is a partially solved problem that is related to motif finding.
- There can also be repressors and inhibitors acting in various ways to stop transcription. This makes regulation of gene transcription complex to understand.

Splicing and other RNA processing

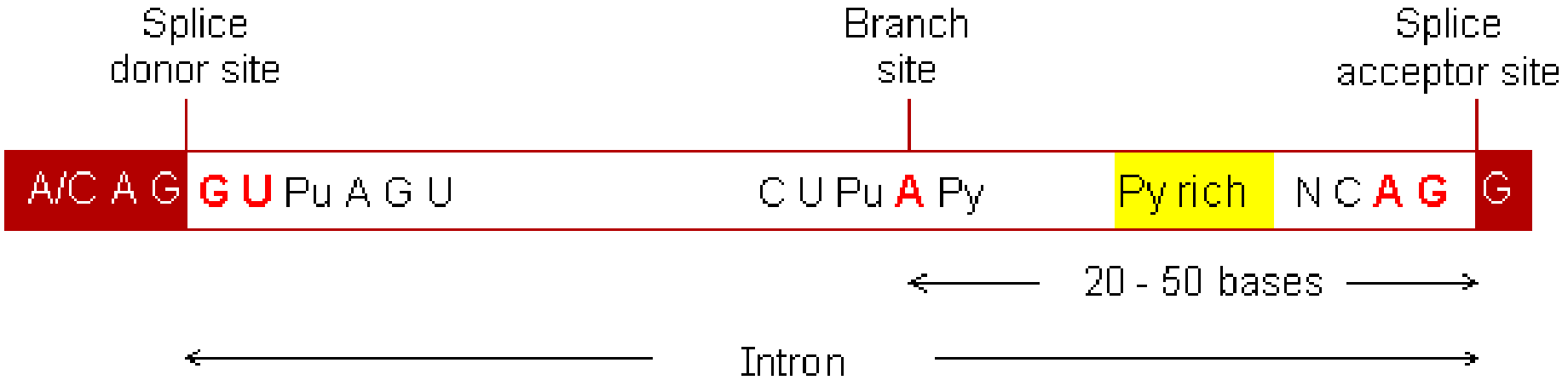
- In Eukaryotic cells, RNA is processed between transcription and translation.
- This complicates the relationship between a DNA gene and the protein it codes for.
- Sometimes alternate RNA processing can lead to an alternate protein as a result. This is true in the immune system.

Splicing (Eukaryotes)

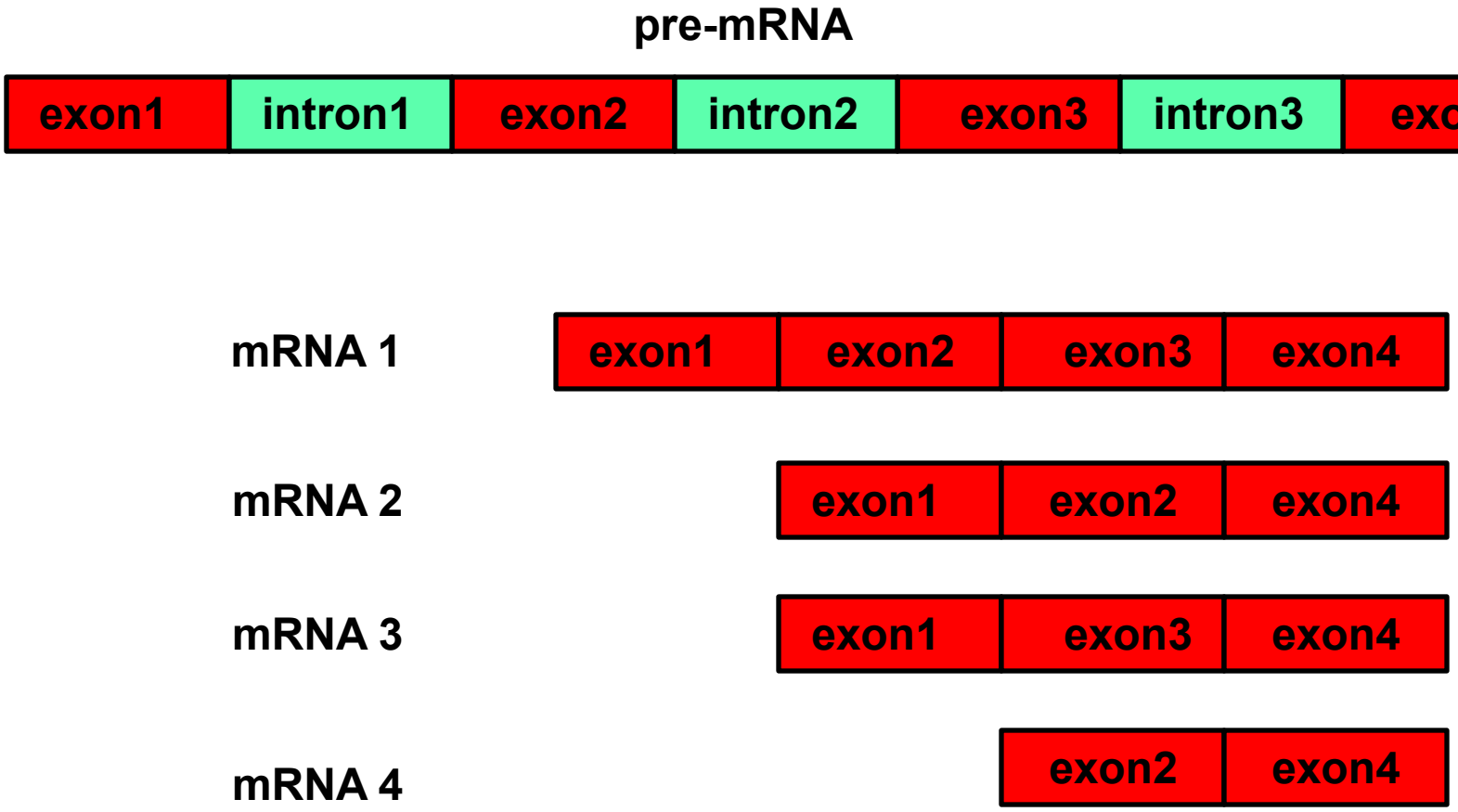
- Unprocessed RNA is composed of Introns and Extrons. Introns are removed before the rest is expressed and converted to protein.
- Sometimes alternate splicings can create different valid proteins.
- A typical Eukaryotic gene has 4-20 introns. Locating them by analytical means is not easy.



Splicing


Splice
donor site
Branch
site
Splice
acceptor site
A/C A G G U Pu A G U
C U Pu A Py
Py rich
N C A G
G
20 - 50 bases
Intron

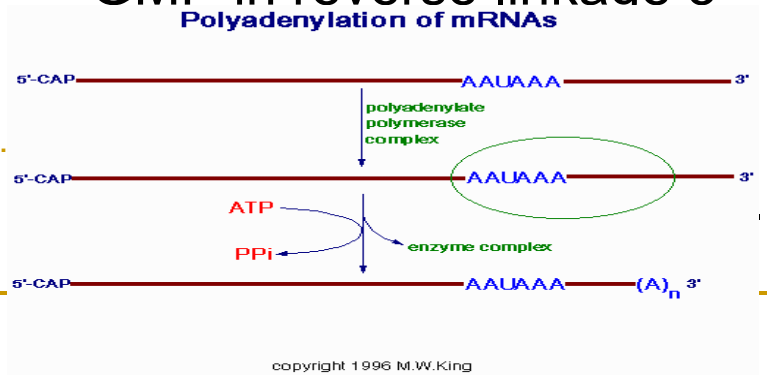
Alternative splicing



Posttranscriptional Processing: Capping and Poly(A) Tail

Capping

- Prevents 5’ exonucleolytic degradation.
- 3 reactions to cap:

1. Phosphatase removes 1 phosphate from 5’ end of pre-mRNA
2. Guanyl transferase adds a GMP in reverse linkage 5’
3. 

Poly(A) Tail

- Due to transcription termination process being imprecise.
- 2 reactions to append:

1. Transcript cleaved 15-25 past highly conserved AAUAAA sequence and less than 50 nucleotides before less conserved U rich or GU rich sequences.
2. Poly(A) tail generated from ATP by poly(A) polymerase which is activated by cleavage and polyadenylation specificity factor (CPSF) when CPSF recognizes AAUAAA. Once poly(A) tail has grown approximately 10 residues, CPSF disengages from the recognition site.

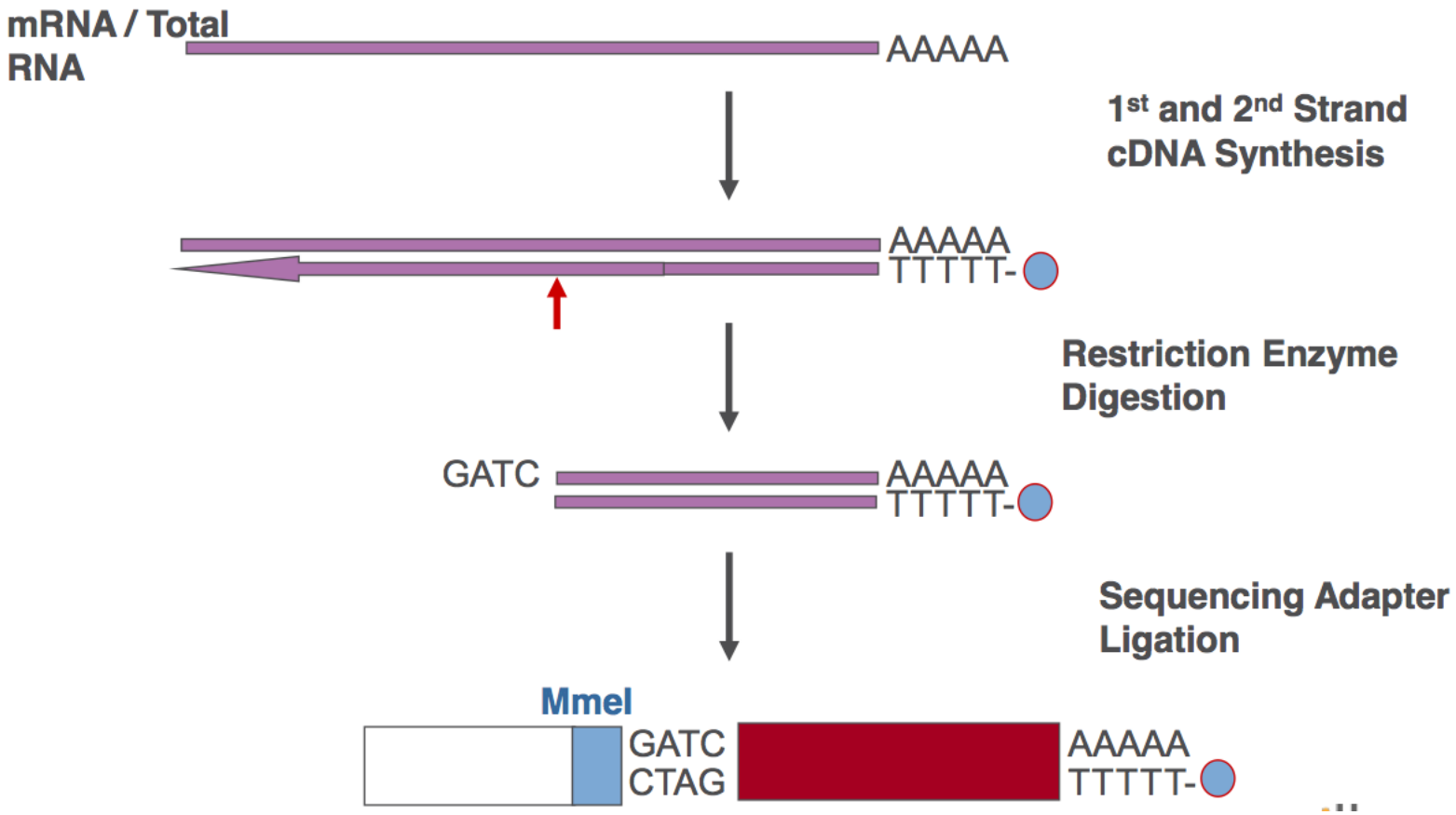
Transcriptome

- Collection of all RNA sequences in the cell
  - mRNA: messenger RNA, encodes for proteins
  - Non-coding RNAs:
    - tRNA: transfer RNA
    - rRNA: ribosomal RNA
    - miRNA, snoRNA, siRNA, etc: micro RNAs
    - lncRNA: long non-coding RNA

RNASeq

- High throughput sequencing of transcriptome
- RNA is not sequenced directly, converted to cDNA first
  - cDNA: coding DNA
- Essential for:
  - Understanding functional and regulatory elements
  - Revealing molecular structures of cells
  - Understanding development and disease

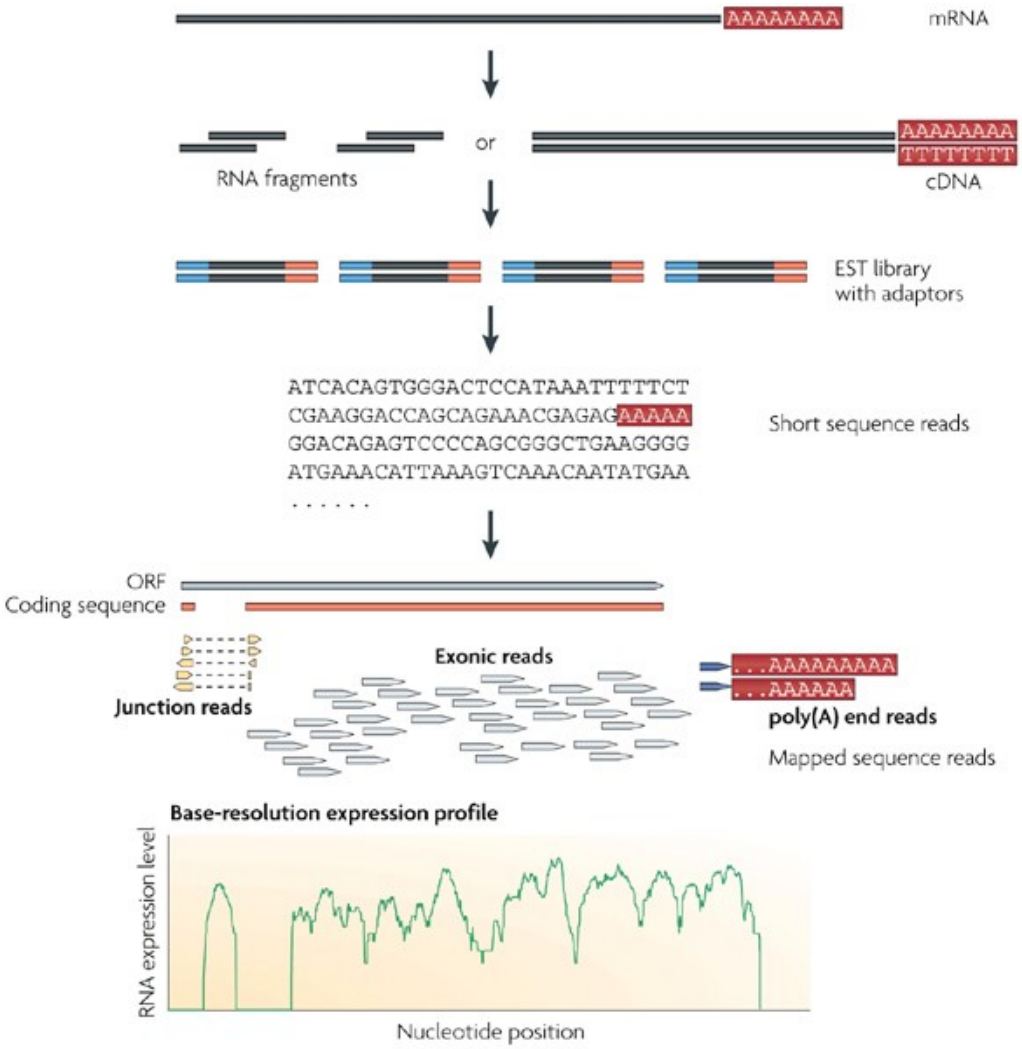
cDNA Synthesis



Aims

- Quantify RNA abundance
  - mRNA or non-coding RNA
- Determine transcriptional structures of genes
  - Start/stop sites
  - Splicing patterns
  - Different isoforms
- Quantify changing expression levels of each transcript in a time frame
  - Developmental stages or under different conditions
- Discover structural variants and/or transcriptional errors: fusion genes

RNASeq

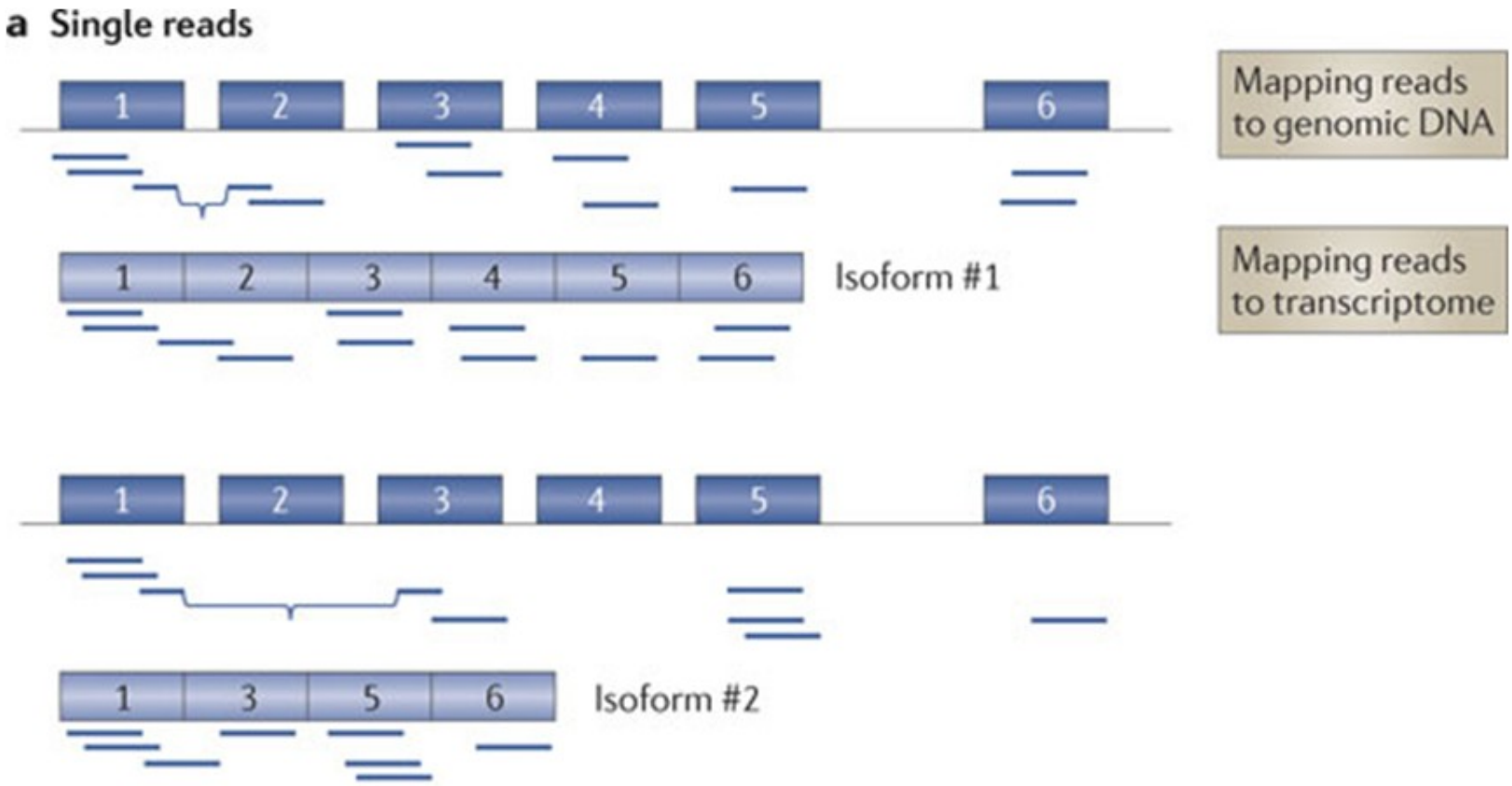


Nature Reviews | Genetics

RNASeq Alignment

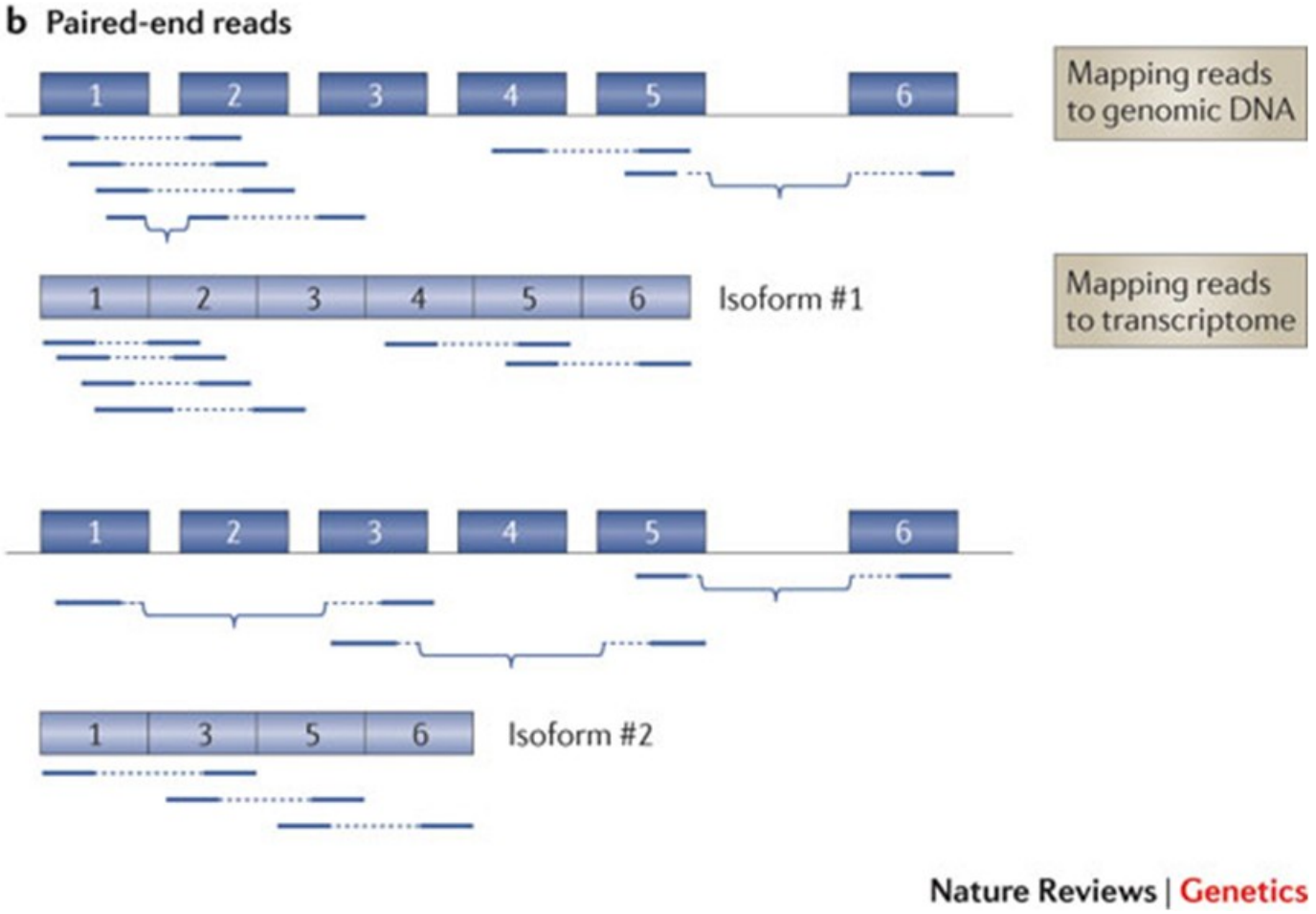
- RNASeq aligners must be able to map across intron/exon junction
  - Essentially split read mapping
  - Also consider the splicing donor/acceptor motifs
- Issues
  - If exon length is shorter than the read length
- Examples:
  - TopHat, GEM, RUM

Isoform detection



Ozsolak et al, Nat Rev Genet, 2011

Isoform detection

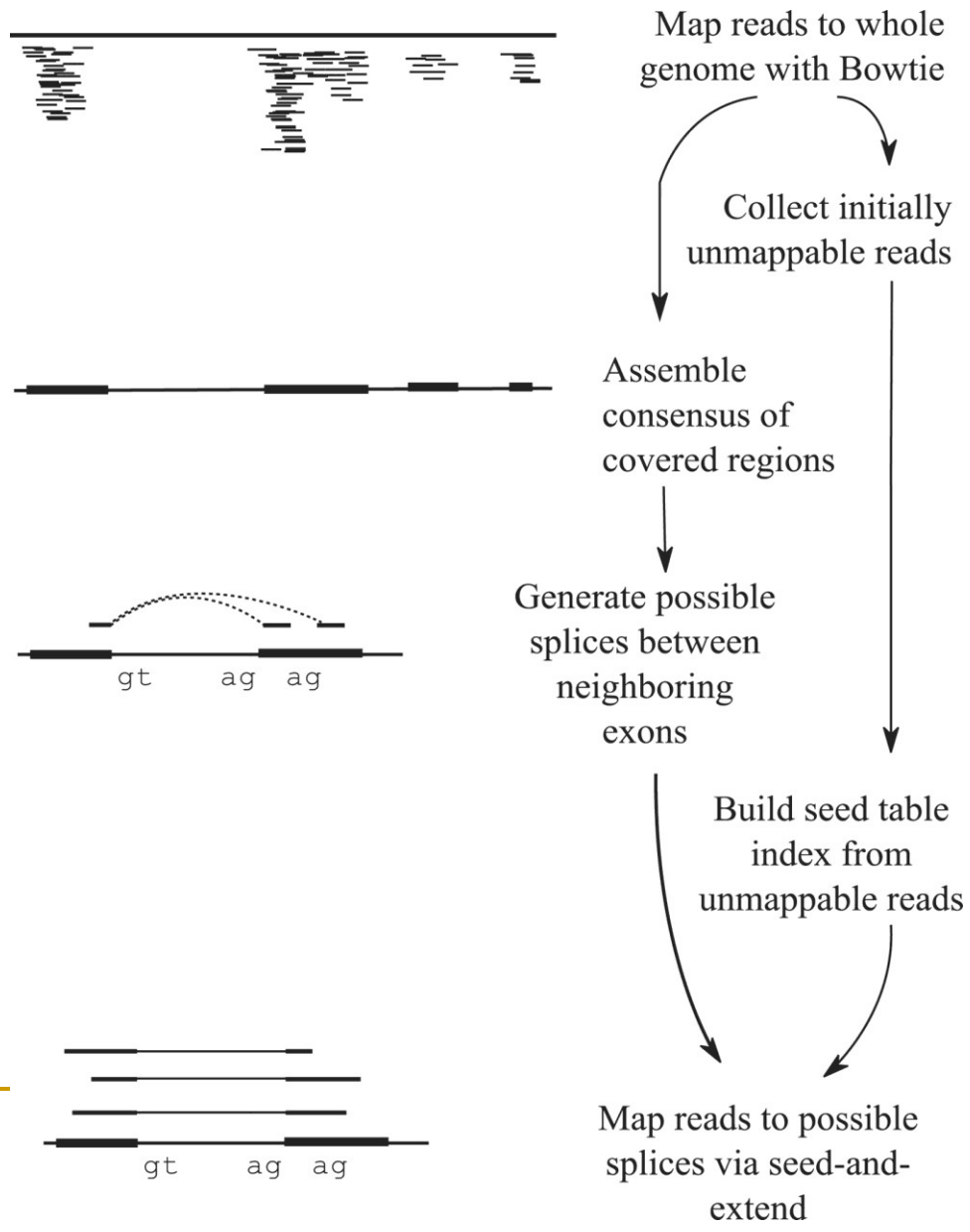


Ozsolak et al, Nat Rev Genet, 2011

TopHat

1. Including flanking seq on both sides of each island to capture donor and acceptor sites from flanking introns.

2. To prevent psedo-gaps of low-expressed genes, merge islands within 70bp of each other (Introns > 70bp)

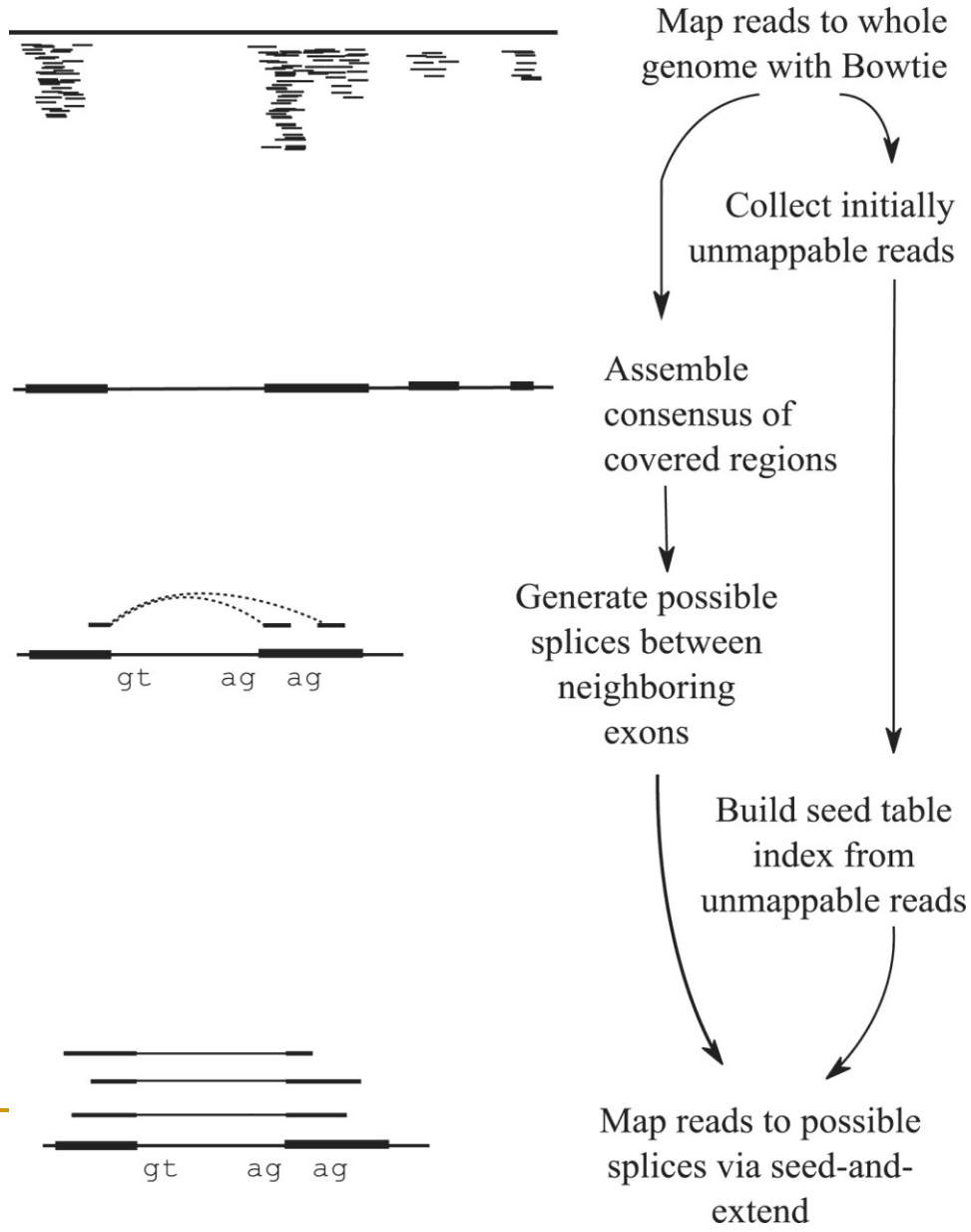


Trapnell et al., Bioinformatics 2009

TopHat: splice junctions

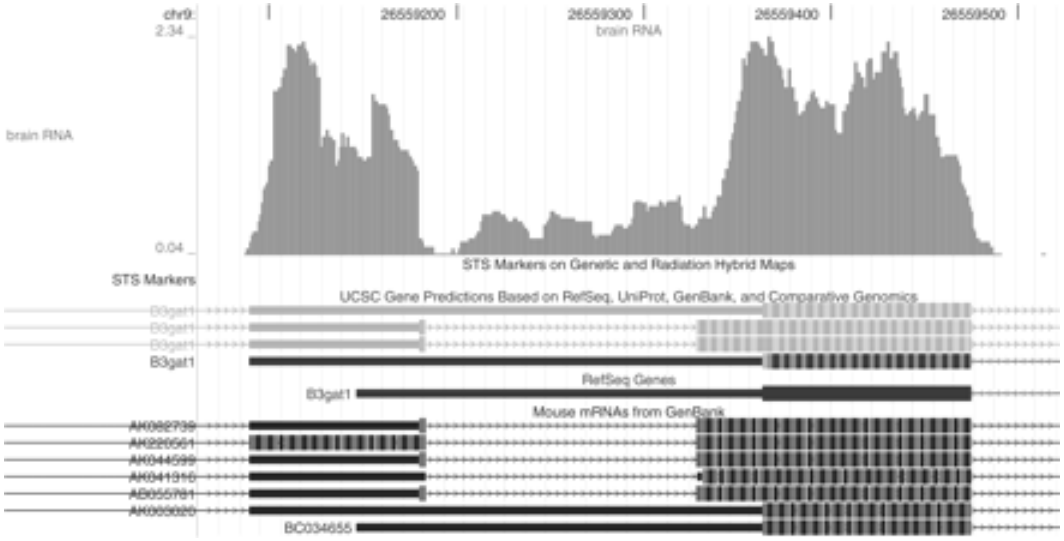
Find GT-AG pairing sites between neighboring (not adjacent) islands

The distance between two sites should > 70bp and <20k bp, as intron length lies within this range



Trapnell et al., Bioinformatics 2009

TopHat: single island junction



Isoforms transcribed at low level -> low coverage

For each island spanning coordinates i to j

$$D_{ij} = \frac{\sum_{m=i}^{j} d_m}{j-i} \cdot \frac{1}{\sum_{m=0}^{n} d_m}$$

D value represents the normalized depth of coverage for an island.
Single-island junctions tend to fall within islands with high *D*

Trapnell et al., Bioinformatics 2009

TopHat: Initially Unmapped Reads

Align s length initially unmapped reads to potential splice junctions

Seed-and-extend strategy:

1. Find IUM span junctions at least k bases on each side
2. 2k-mer 'seed' is constructed by concatenating the k bases on left and right islands
3. Mismatches are allowed except seed regions

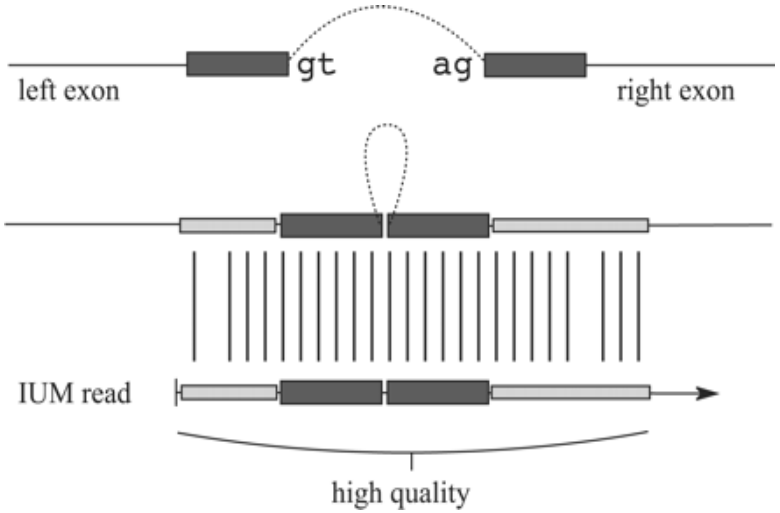


Fig: Dark gray is seeds

Trapnell et al., Bioinformatics 2009c

TopHat: build splice junctions

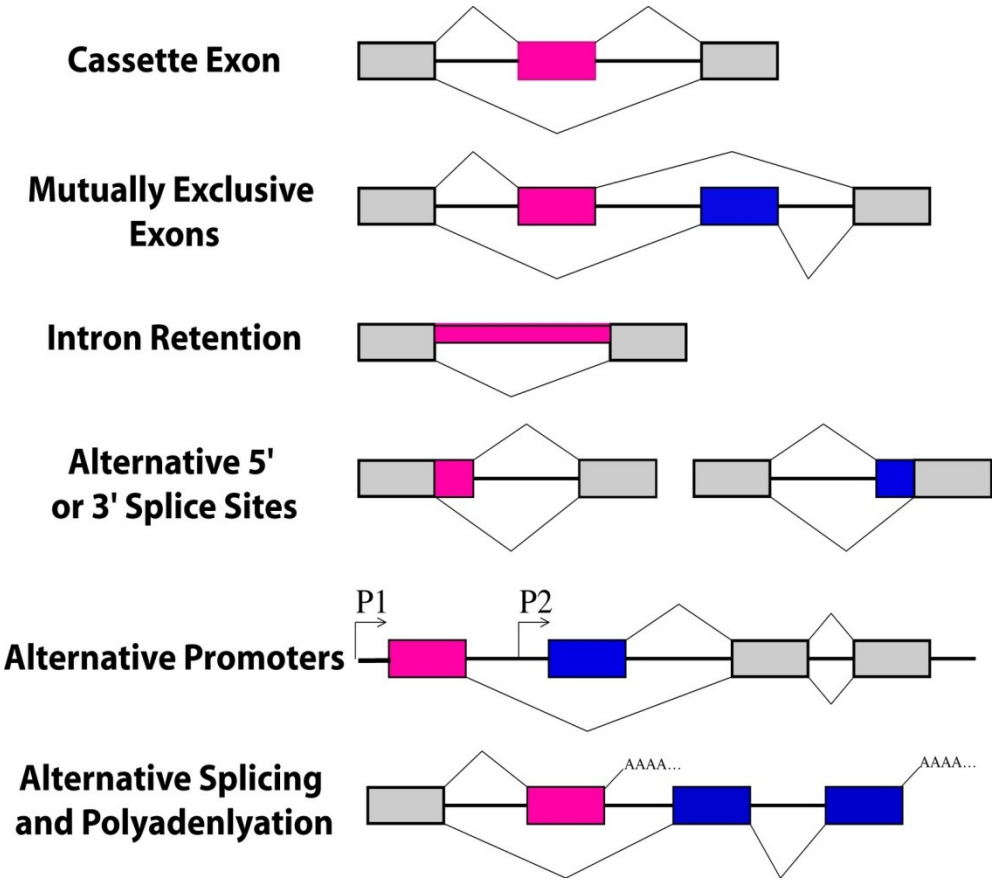
1. **Summarize all the spliced alignment from prior step**

2. **Filter the junctions occurs at <15% of the depth of the exons flanking it**

Trapnell et al., Bioinformatics 2009

GENE AND ISOFORM ABUNDANCE

Alternative splicing & isoforms



Expression Values

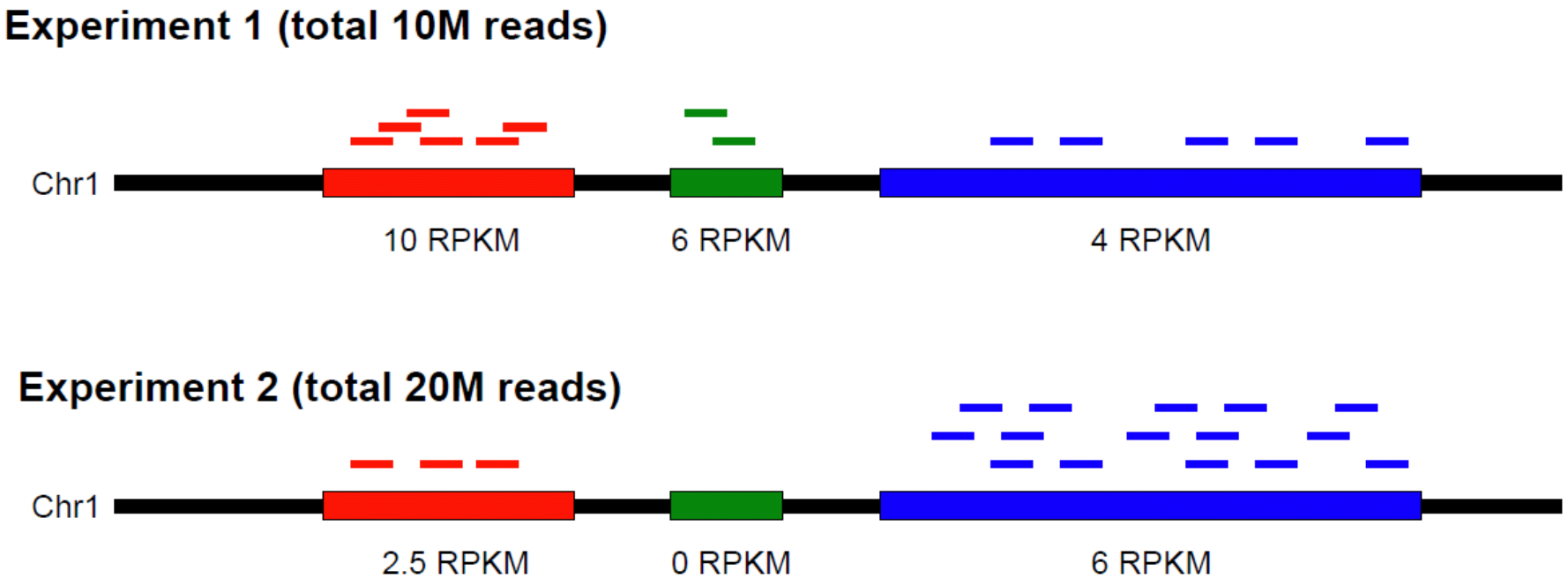
- **R**eads **P**er **K**ilobase of exon model per **M**illion mapped reads
- Nat Methods. 2008, Mapping and quantifying mammalian transcriptomes by RNA-Seq. Mortazavi A et al.

$$RPKM = 10^9 \times \frac{C}{NL}$$

C= the number of reads mapped onto the gene's exons
N= total number of reads in the experiment
L= the sum of the exons in base pairs.

Mortazavi et al, Nat Methods, 2008

RPKM



1 RPKM ~= 0.3 to 1 transcript per cell

Mortazavi et al, Nat Methods, 2008

Cufflinks

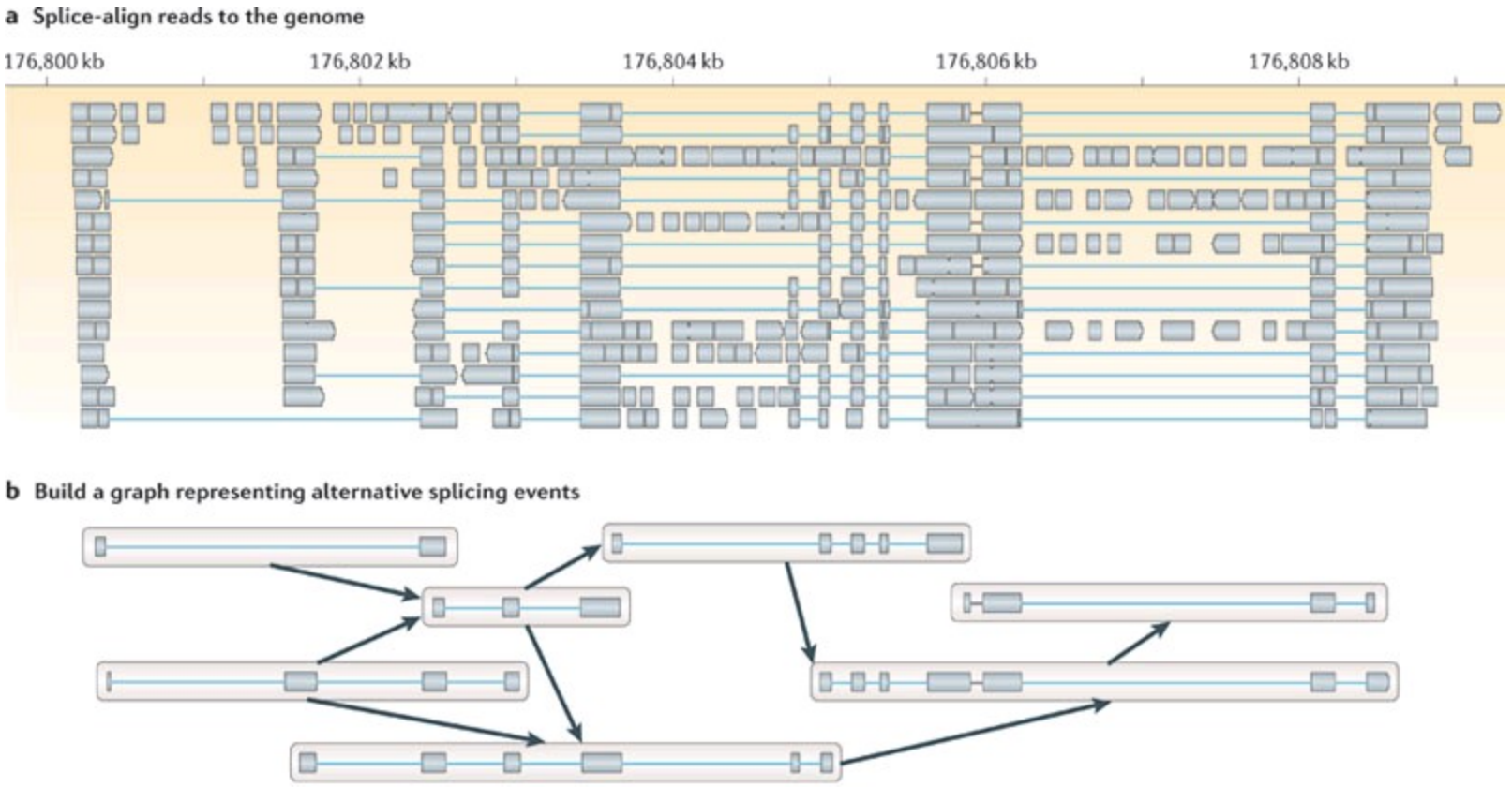
- Similar to RPKM
- Instead define FPKM: fragments per kilobase of exon model per million mapped fragments
- Also can estimate isoform abundance using either:
  - Known annotation
  - Transcriptome assembly

TRANSCRIPTOME ASSEMBLY

Transcriptome assembly

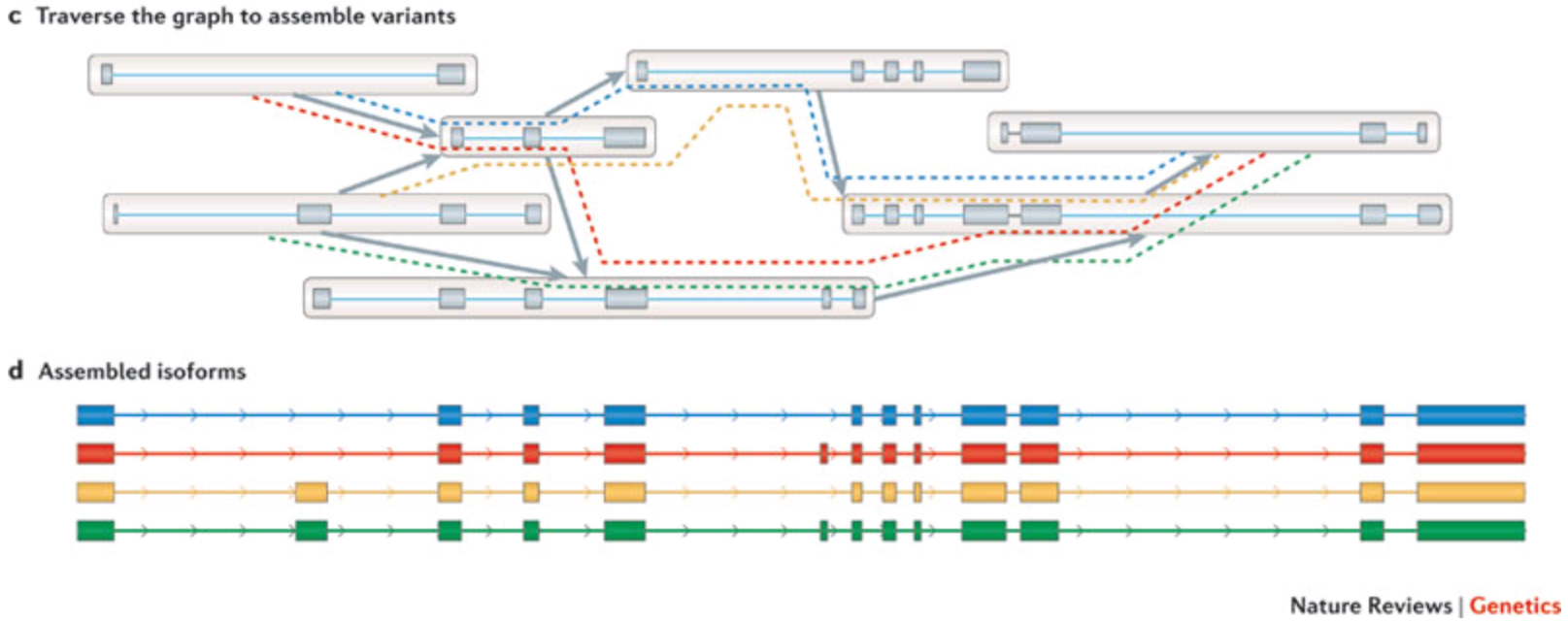
- Similar to genome assembly, but the end-product will be the transcripts
- Lower effect by repeats
- Isoforms:
  - Identical reads coming from different isoforms of the same gene!
  - Reconstruct alternate transcripts
- Assemblers:
  - Reference based: Cufflinks, ERANGE
  - *de novo*: Trans-ABySS, Oases

Reference based



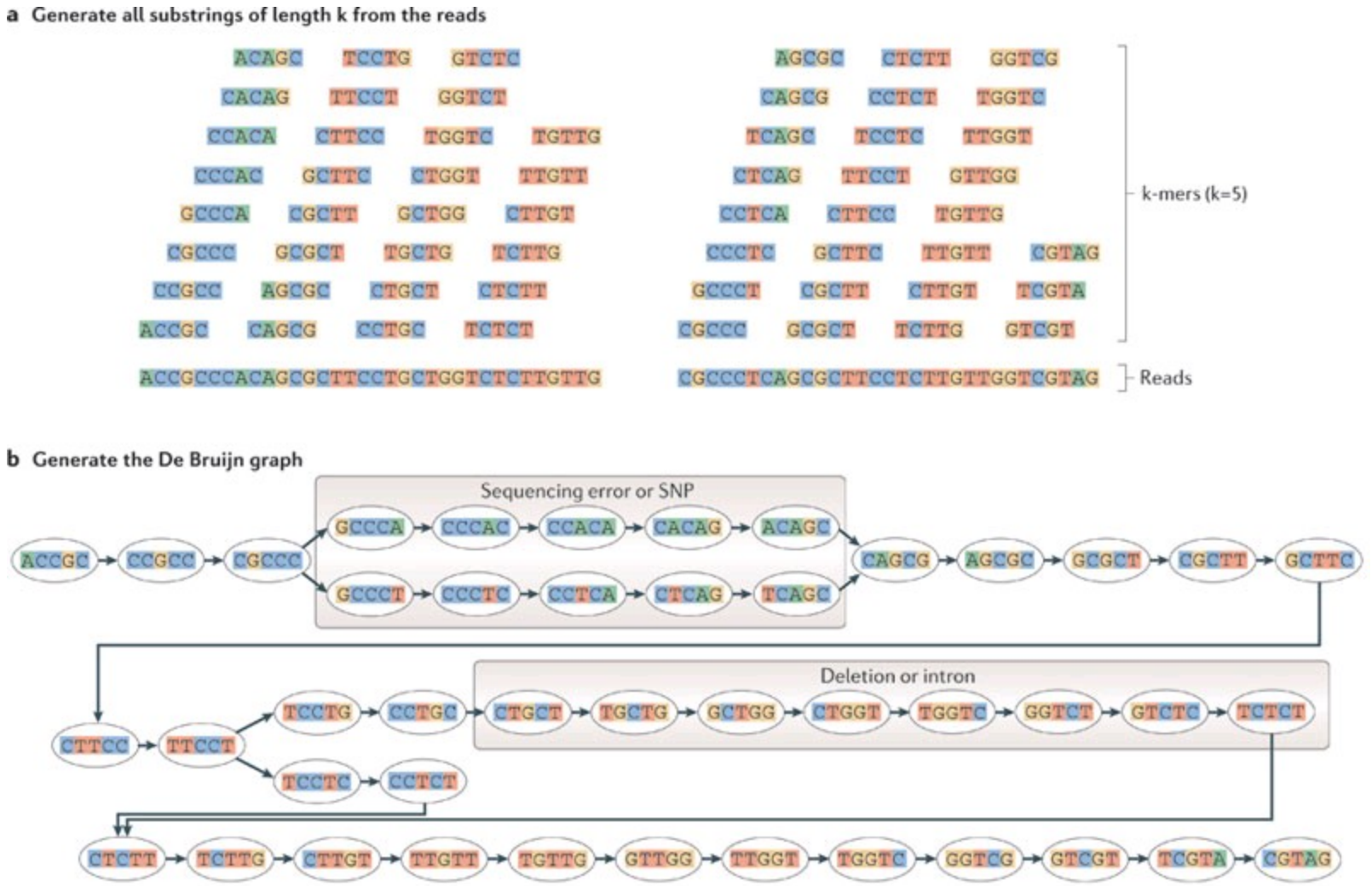
Martin et al., Nat Rev Genet, 2011

Reference based



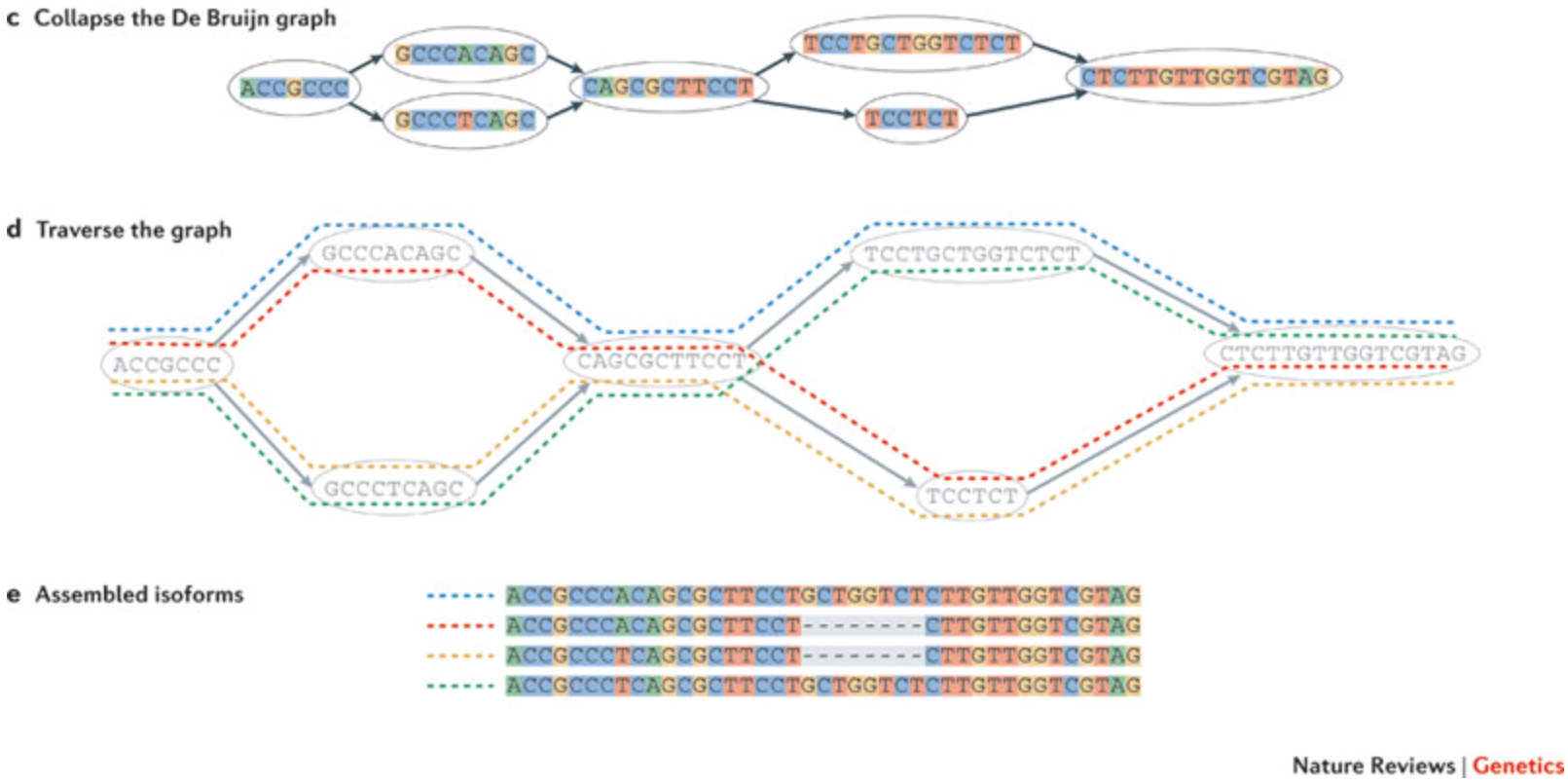
Martin et al., Nat Rev Genet, 2011

De novo



Martin et al., Nat Rev Genet, 2011

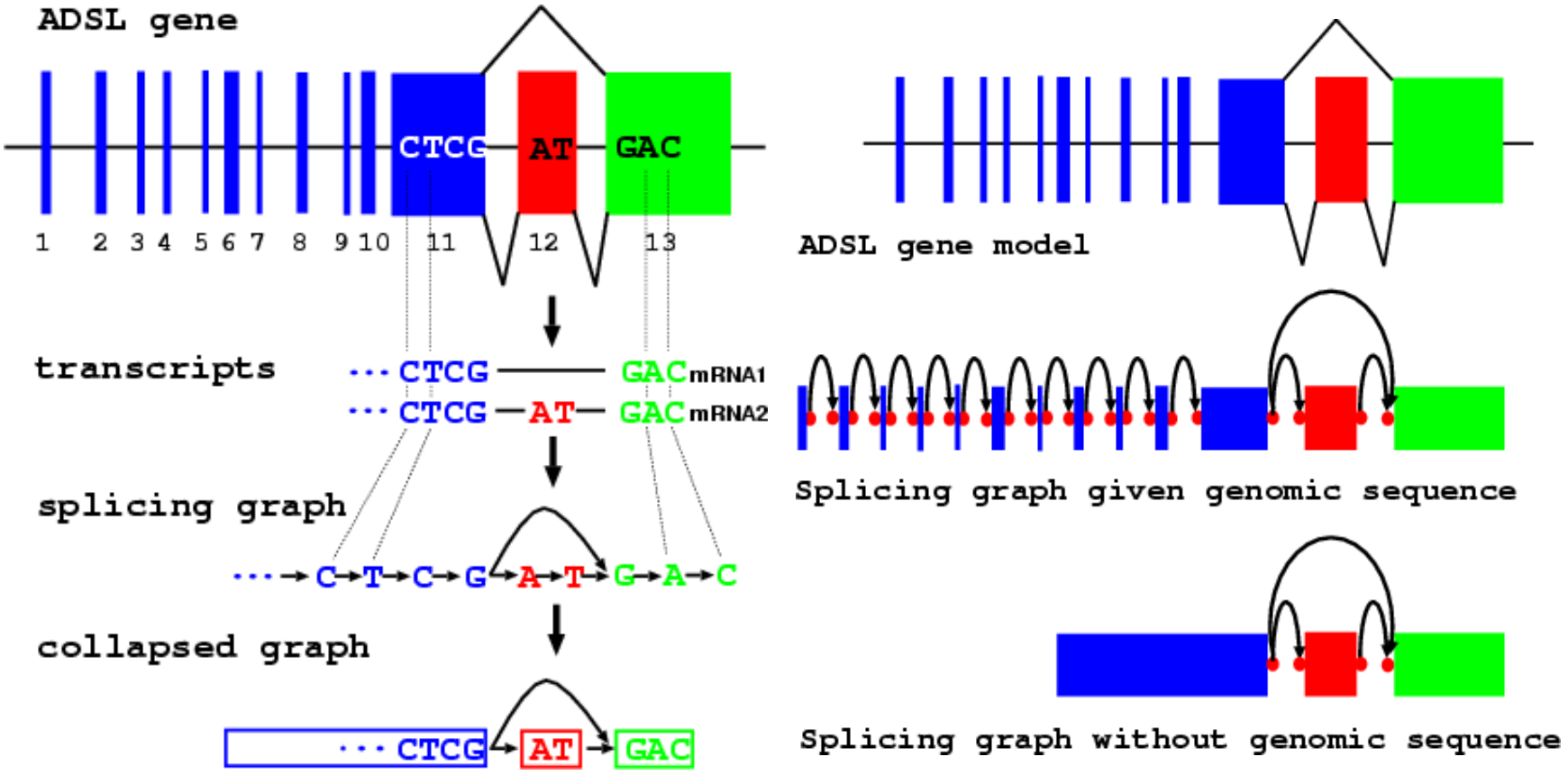
De novo



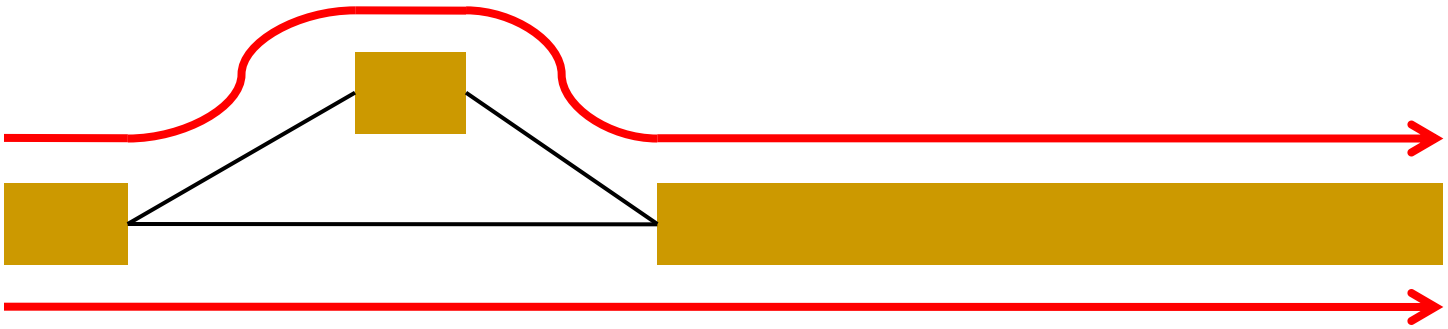
Martin et al., Nat Rev Genet, 2011

De Bruijn graphs ~ splice graphs

Heber et al, 2002

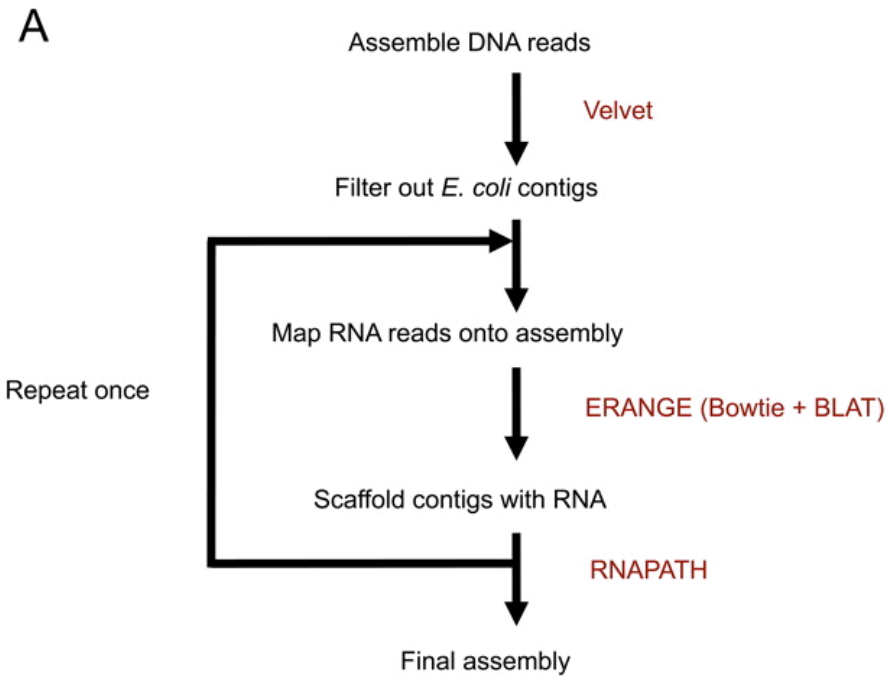


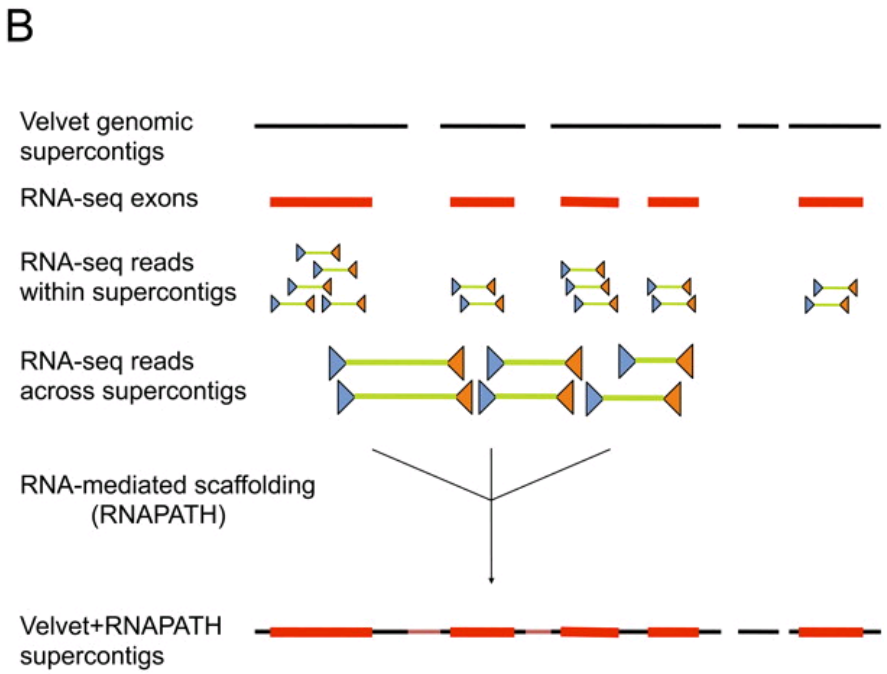
Oases – *de novo* RNAseq assembly



Slide courtesy if Dan Zerbino

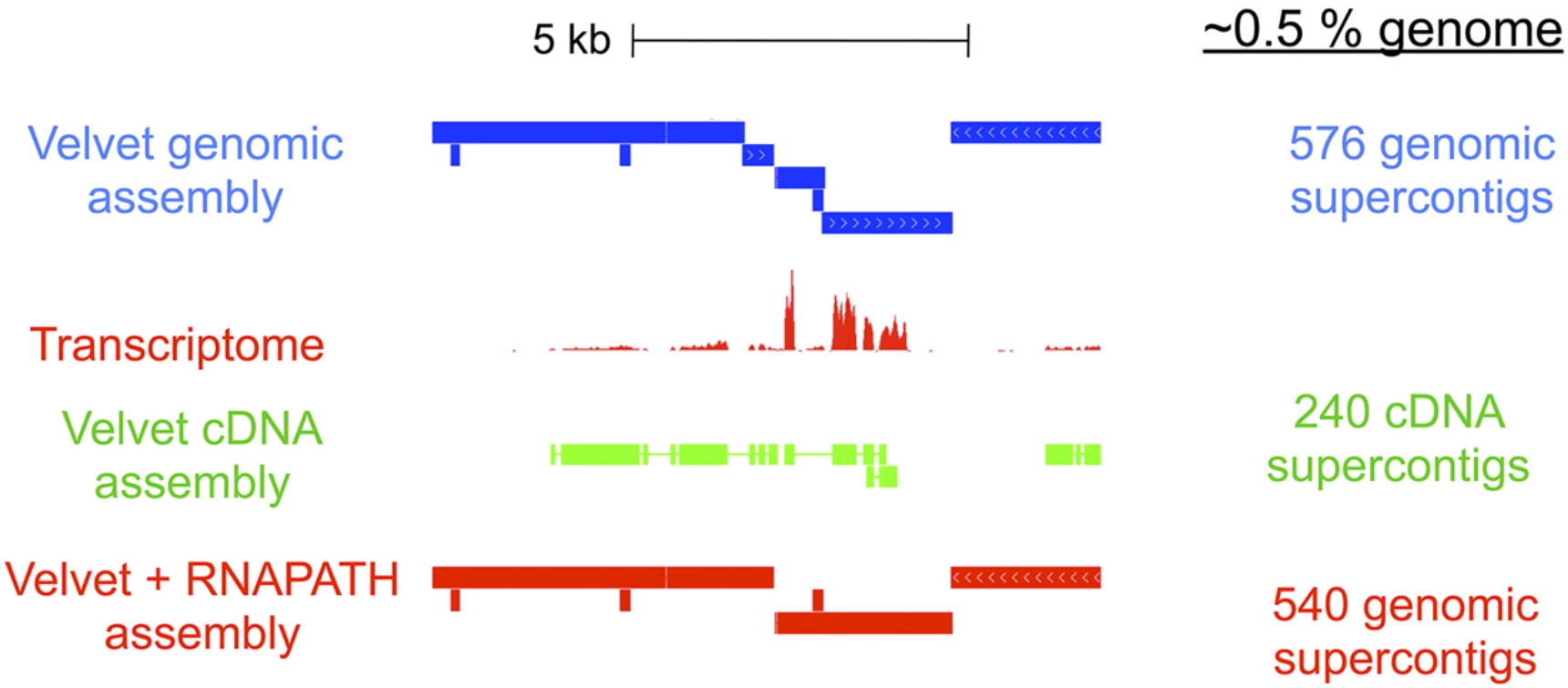
Genome scaffolding using RNAseq





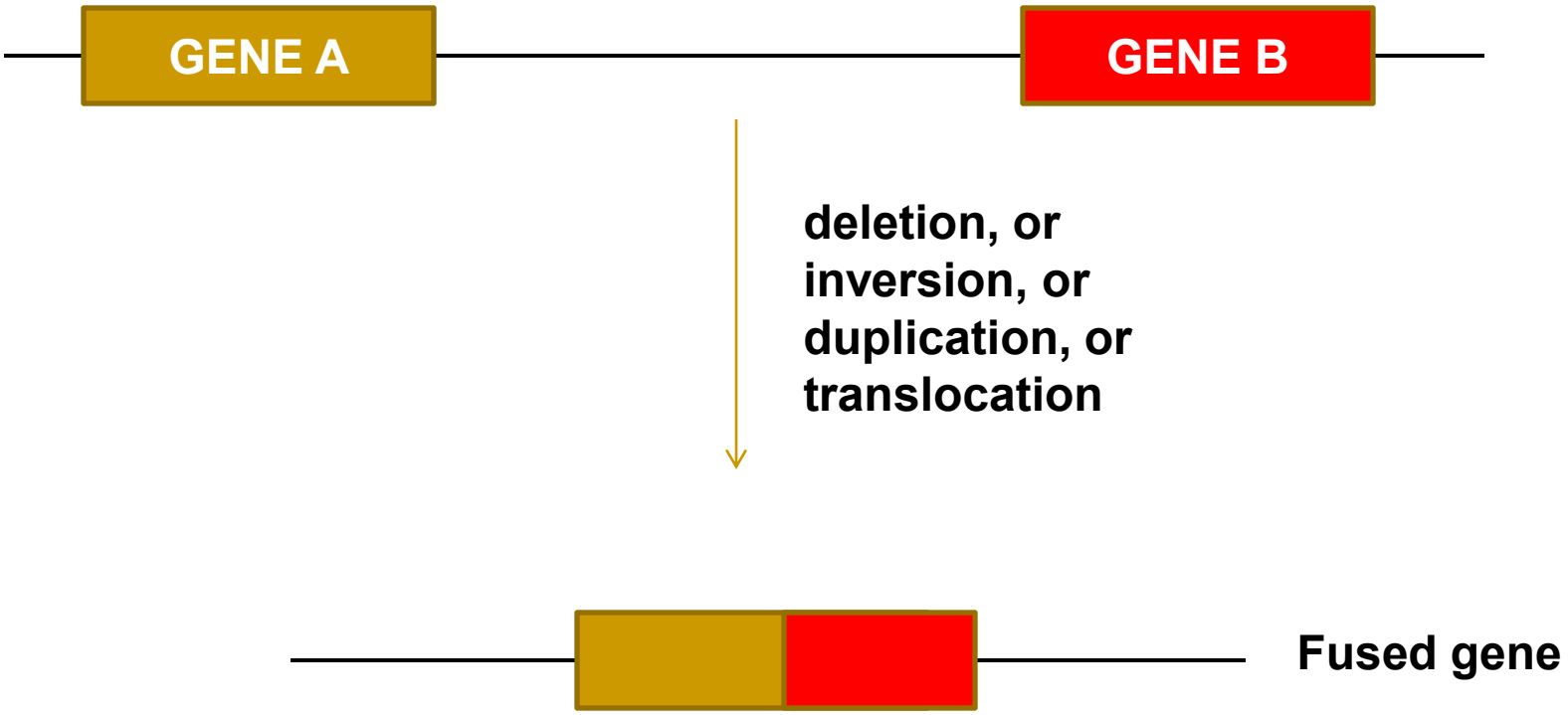
Mortazavi et al, Genome Res., 2010

Genome scaffolding using RNAseq



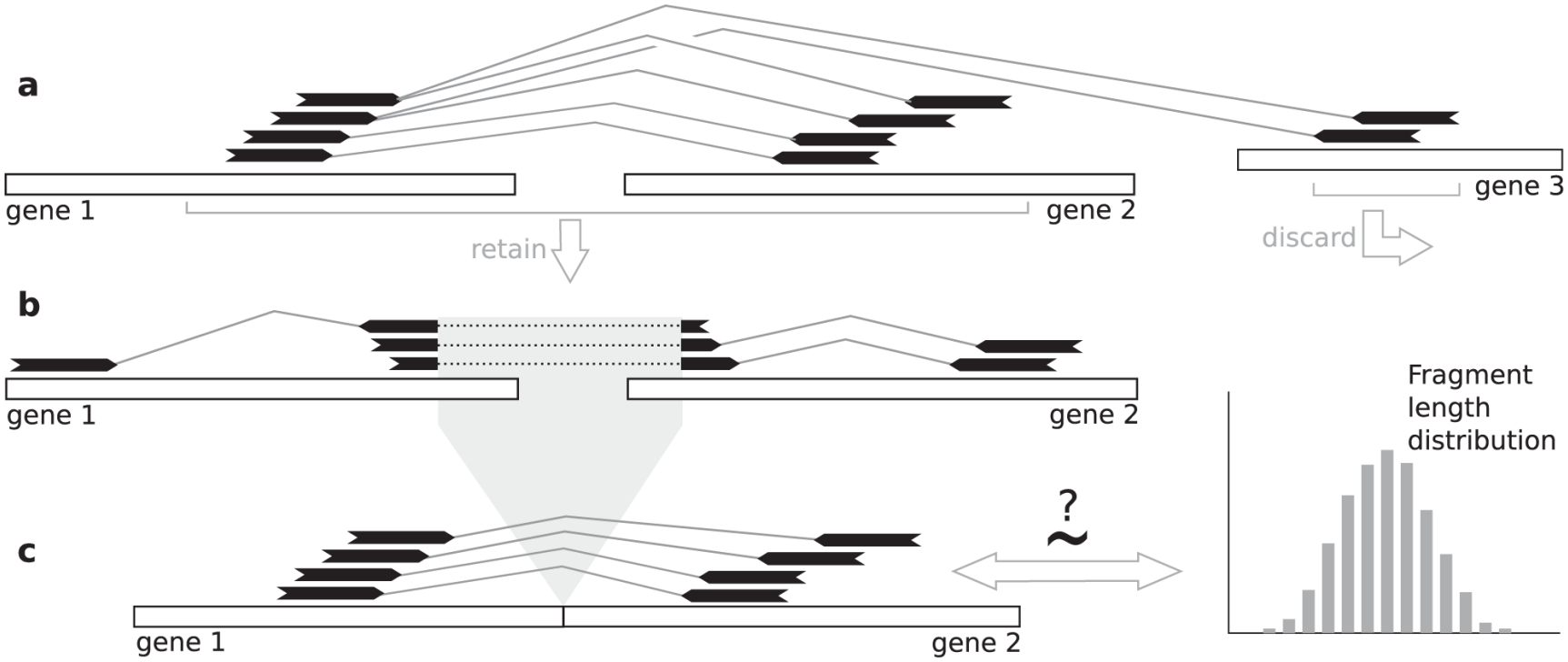
Mortazavi et al, Genome Res., 2010

Fusion genes



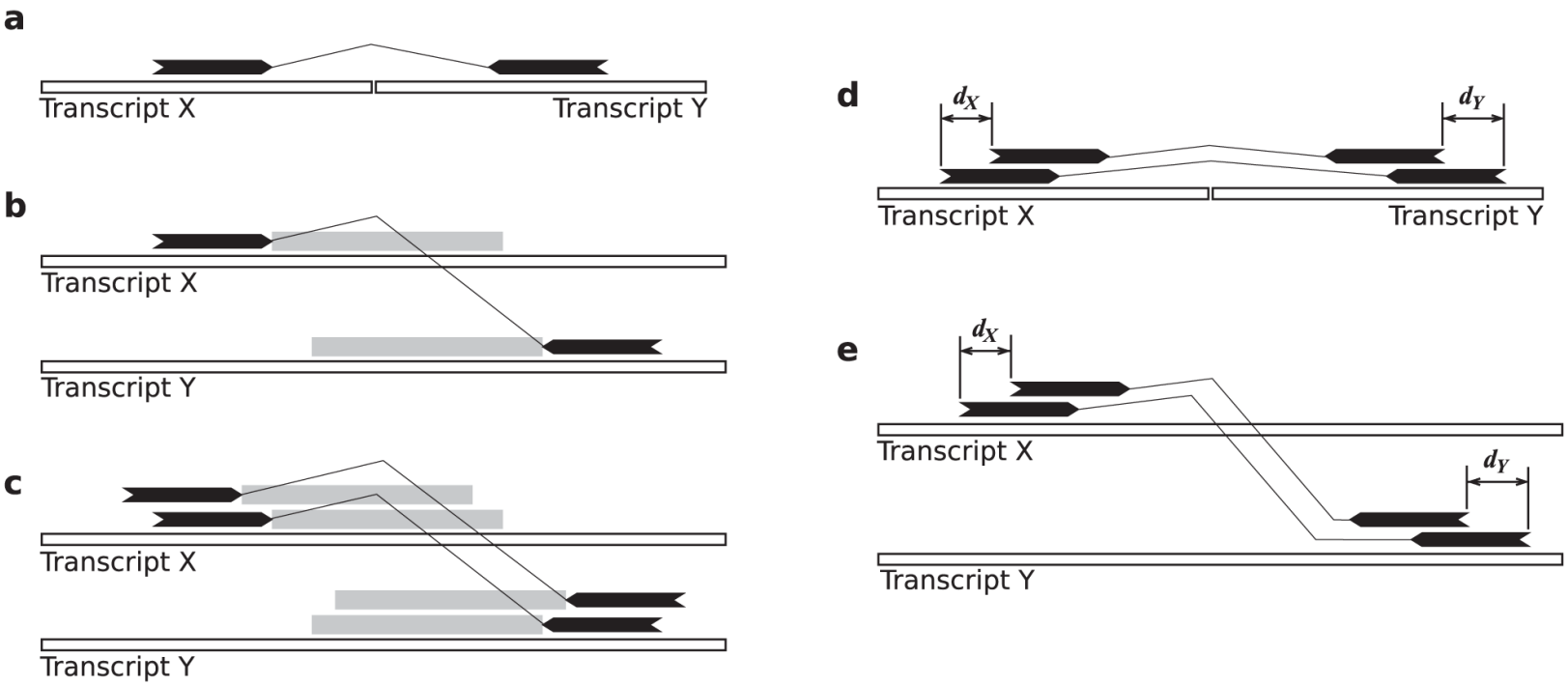
Example: Chronic myelogeneous leukemia (chr9-chr22)
***BCR-ABL* fusion**

Fusion genes: deFuse



McPherson et al., PLoS Comp Biol, 2011

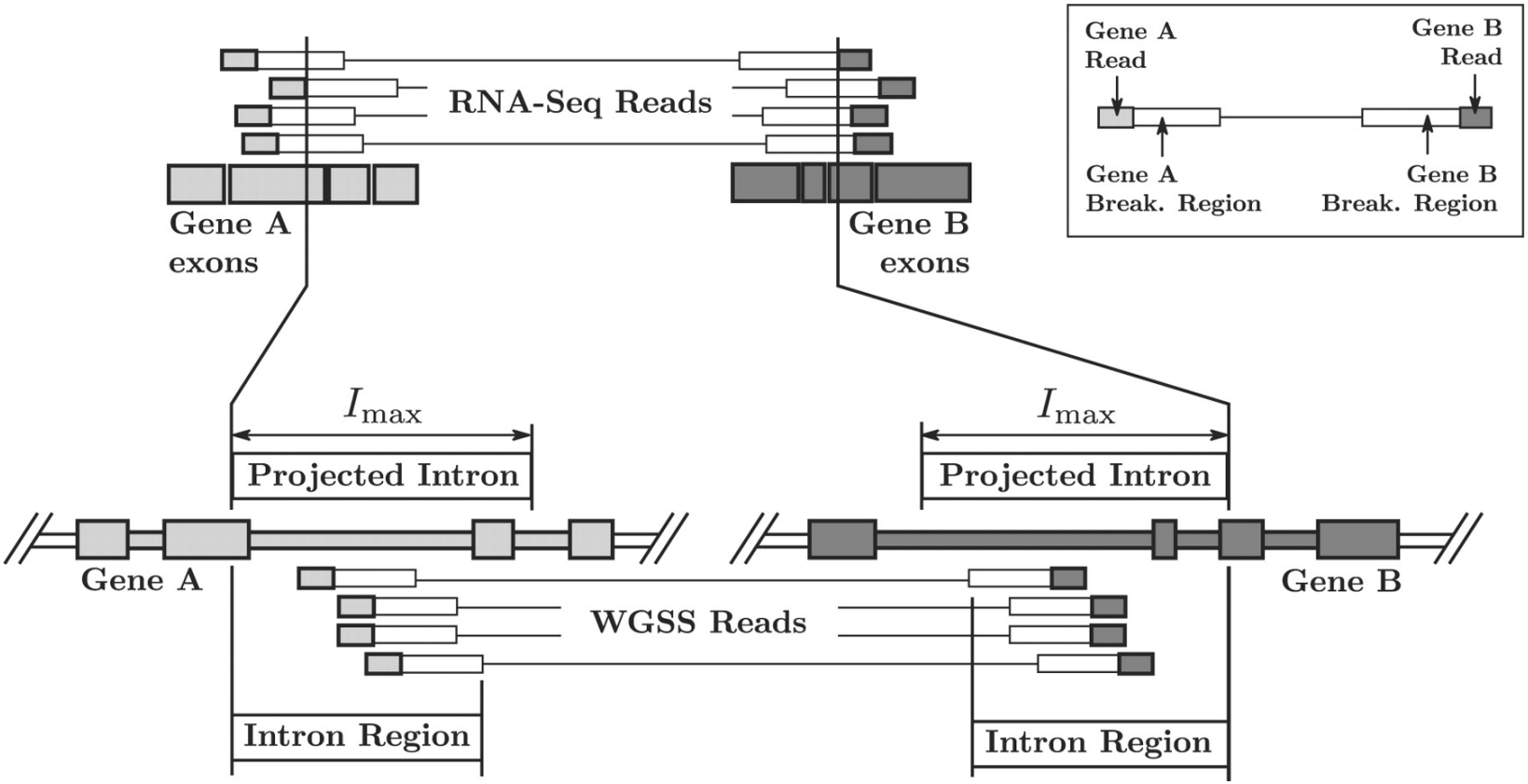
Fusion genes: deFuse



McPherson et al., PLoS Comp Biol, 2011

Comrad: integrate RNASeq+WGS

Good to discover & differentiate genome-level & transcript-level fusions



McPherson et al., Bioinformatics, 2011