Repetitive DNA and next-generation sequencing: computational challenges and solutions

Todd J. Treangen, Steven L. Salzberg

Nature Reviews Genetics 13, 36-46 (January 2012)
doi:10.1038/nrg3117

nature REVIEWS GENETICS

Speaker: 黃建龍, 黃元鴻

Date: 2012.06.04

Outline

- Abstract
- Genome resequencing projects
- De novo genome assembly
- RNA-seq analysis
- Conclusions

Abstract

- Repetitive DNA are abundant in a broad range of species, from bacteria to mammals, and they cover nearly half of the human genome.
- Repeats have always presented technical challenges for sequence alignment and assembly programs.
- Next-generation sequencing projects, with their short read lengths and high data volumes, have made these challenges more difficult.
- We discuss the computational problems surrounding repeats and describe strategies used by current bioinformatics systems to solve them.

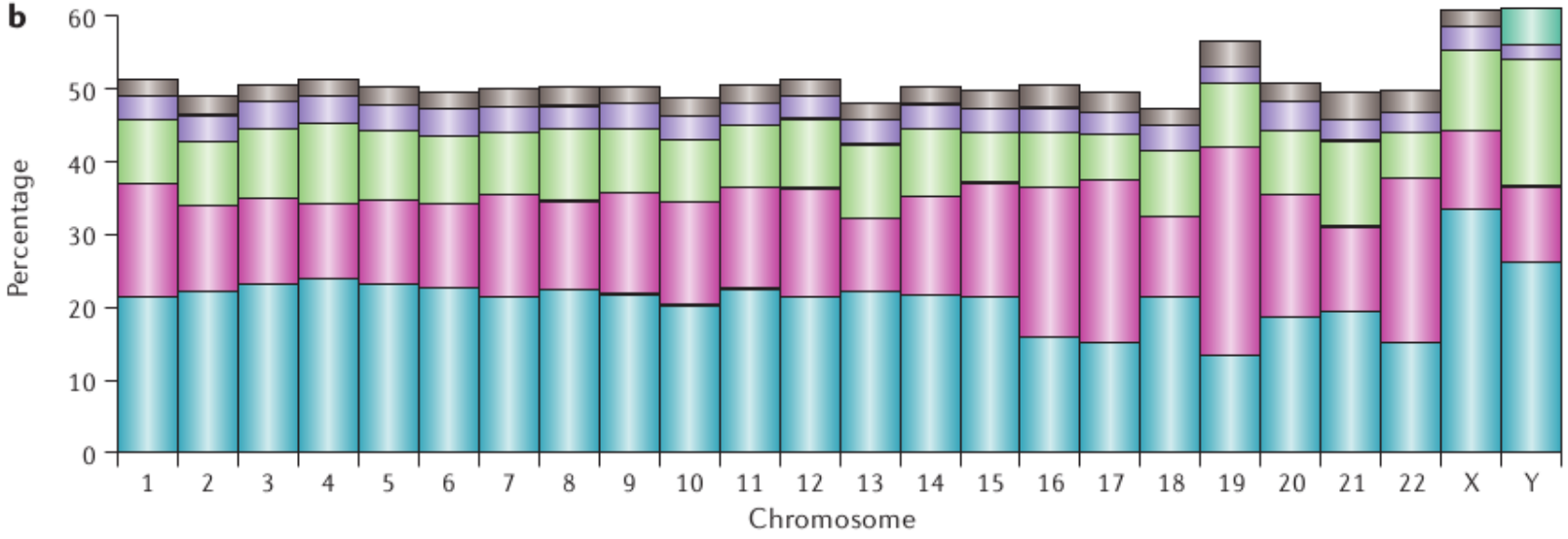
Repeats

- A repetitive sequence in the genome. (> 50% in human genome)
- Although some repeats appear to be nonfunctional, others have played a part in human evolution, at times creating novel functions, but also acting as independent, ‘selfish’ sequence elements.
- Arised from a variety of biological mechanisms that result in extra copies of a sequence being produced and inserted into the genome.

a

Repeat class	Repeat type	Number (hg19)	Cvg	Length (bp)
Minisatellite, microsatellite or satellite	Tandem	426,918	3%	2–100
SINE	Interspersed	1,797,575	15%	100–300
DNA transposon	Interspersed	463,776	3%	200–2,000
LTR retrotransposon	Interspersed	718,125	9%	200–5,000
LINE	Interspersed	1,506,845	21%	500–8,000
rDNA (16S, 18S, 5.8S and 28S)	Tandem	698	0.01%	2,000–43,000
Segmental duplications and other classes	Tandem or interspersed	2,270	0.20%	1,000–100,000

b



Box 1 | Repetitive DNA in the human genome

Genome resequencing projects

- Study genetic variation by analysing many genomes from the same or from closely related species.
- After sequencing a sample to deep coverage, it is possible to detect SNPs, copy number variants (CNVs) and other types of sequence variation without the need for de novo assembly.
- A major challenge remains when trying to decide what to do with reads that map to multiple locations (that is, multi-reads).

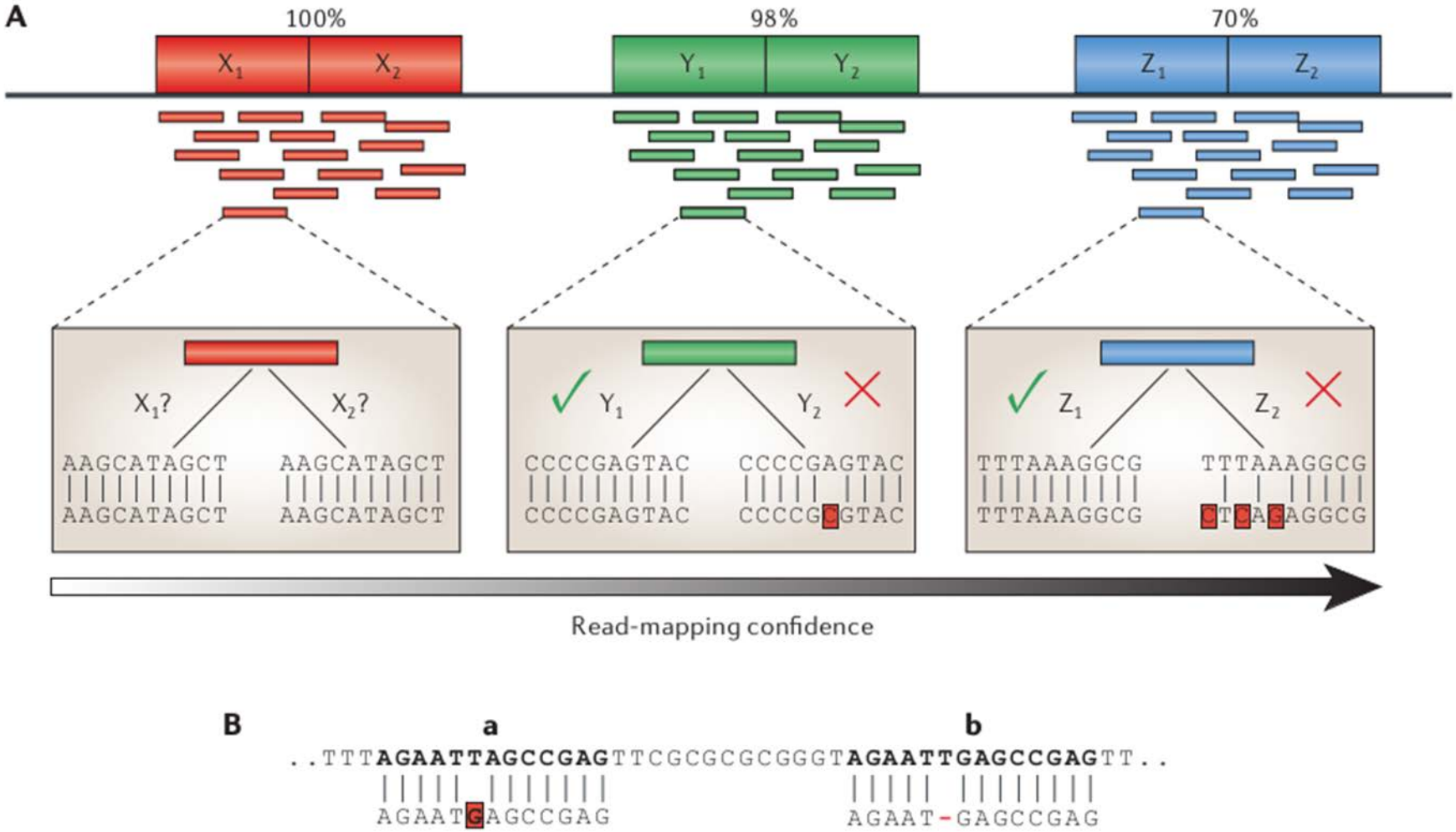


Figure 1 | Ambiguities in read mapping.

Multi-read mapping strategies

- Essentially, an algorithm has three choices for dealing with multi-reads:
  1. Ignore them
  2. The best match approach (If equally good, then choose one at random or report all of them)
  3. Report all alignments up to a maximum number, d (multi-reads that align to > d locations will be discarded)

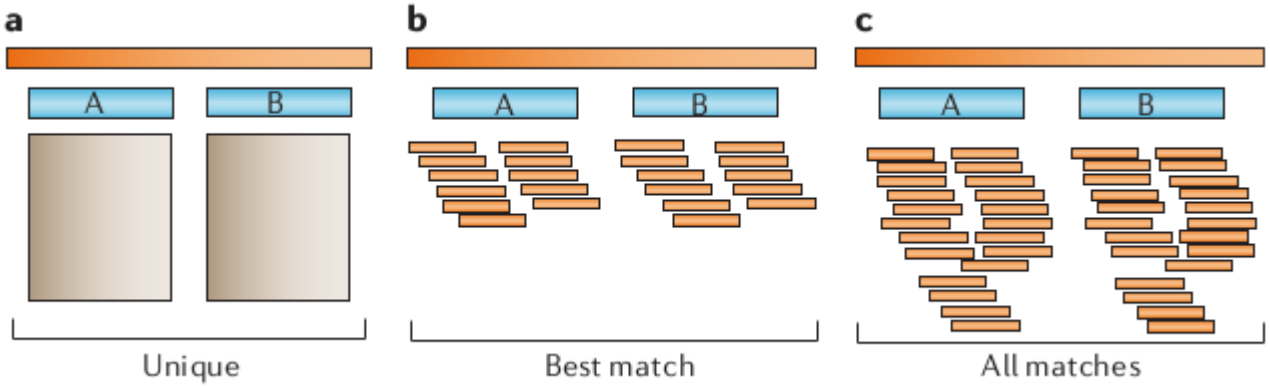


Figure 2 | Three strategies for mapping multi-reads.

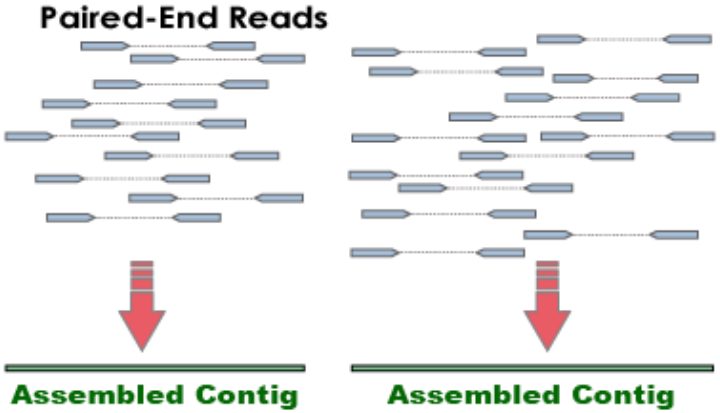
De novo genome assembly

- Set of reads and attempt to reconstruct a genome as completely as possible without introducing errors.

- NGS vs. Sanger sequencing

	NGS	Sanger
Length	50~150 bp	800~900 bp
Depth of coverage	**High**	Lower

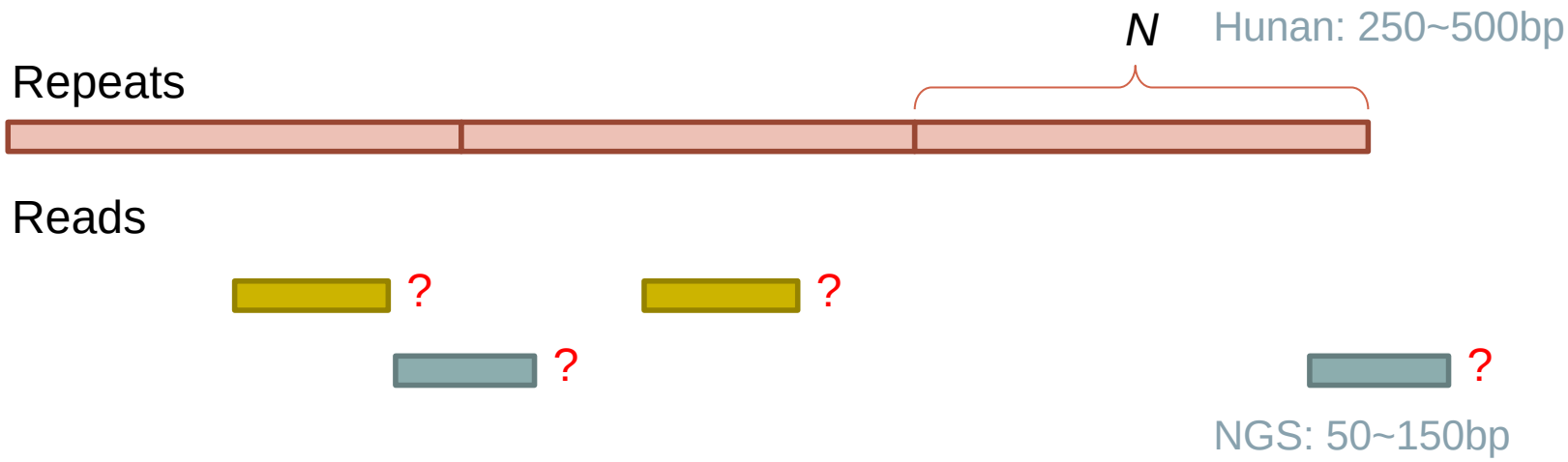
Hard!



http://www.data2bio.com/images/assembly_bg.png

Problems caused by repeats

- Caused by short length of NGS sequences
  - Repeat length > Read Length



- If a species has a common repeat of length ***N***, then assembly of the genome of that species will be **far better** if read lengths are longer than ***N***.

Problems caused by repeats

- Current Assemblers
  - Overlap-based assembler
  - De Bruijn Graph assembler

- Reads → Graph → Traverse & Reconstruct

- Repeats cause branches → **Guess!**
  1. False Joins
  2. Accurate but fragmented assembly. (Short contigs)

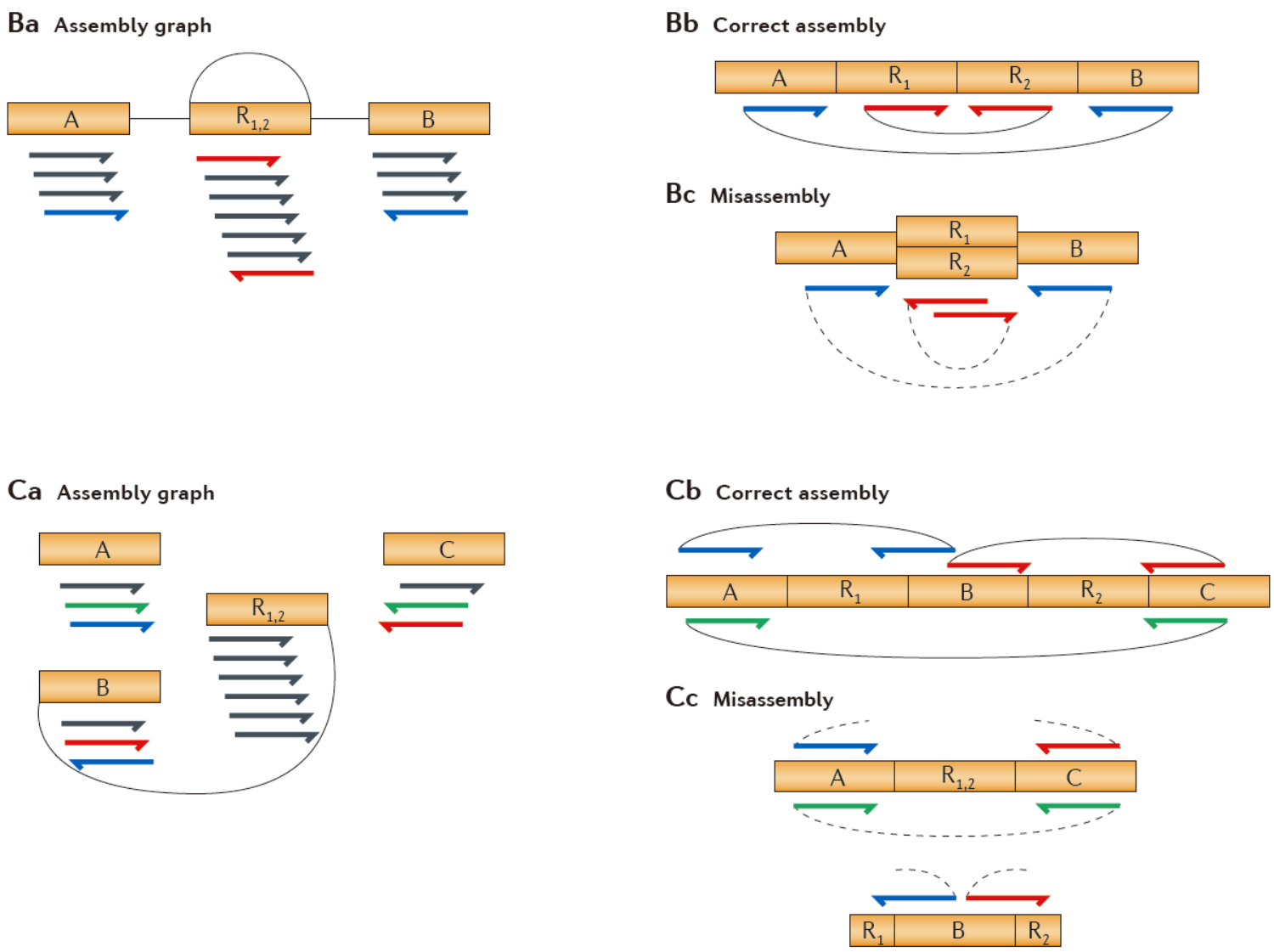
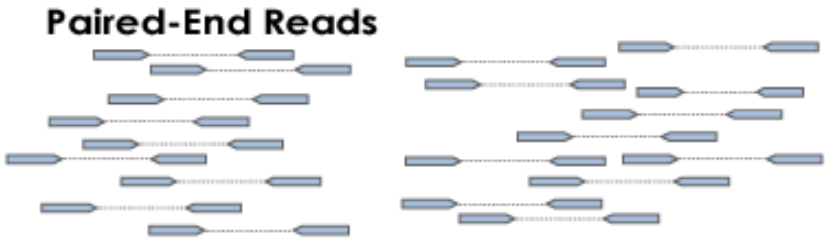


Figure 3 | Assembly errors caused by repeats (B, C)

Problems caused by repeats

- The essential problem with repeats is that an assembler cannot distinguish them.

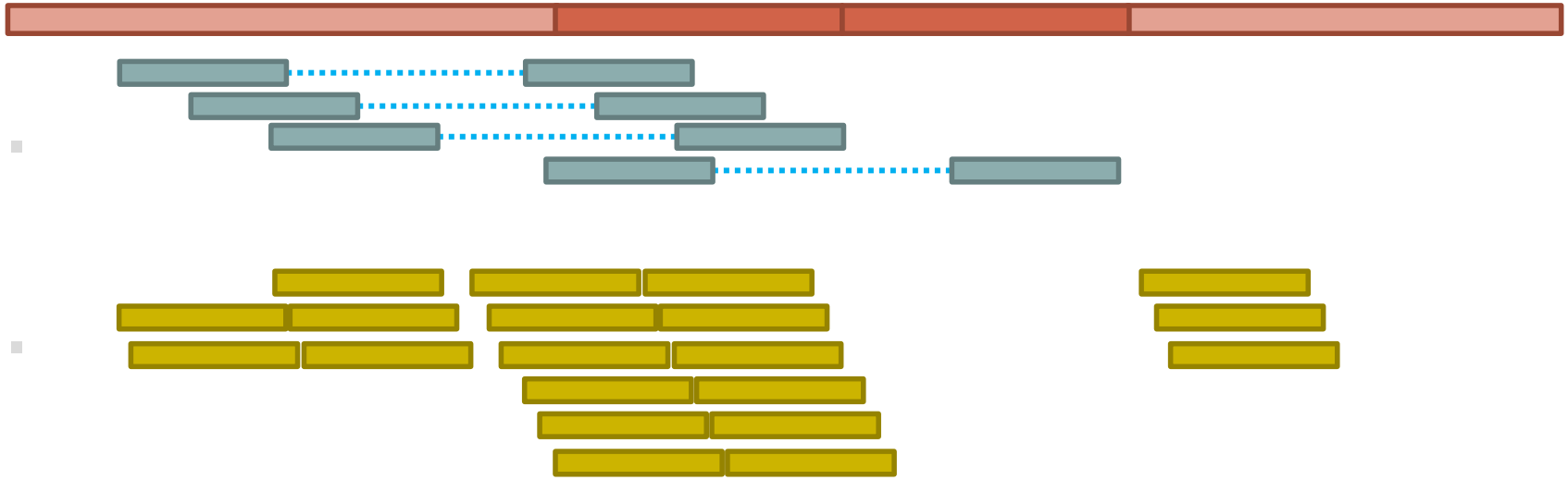
- The only hint of a problem is found in the paired-end links.



- Recent human genome assemblies were found 16% shorter than the reference genome. The NGS assemblies were lacking 420 Mbp of common repeats.

Strategies for handing repeats

1. Use mate-pair information from reads that were sequenced in pairs.
2. The second main strategy: compute statistics on the depth of coverage for each contig
   - Assume that the genome is uniformly covered.



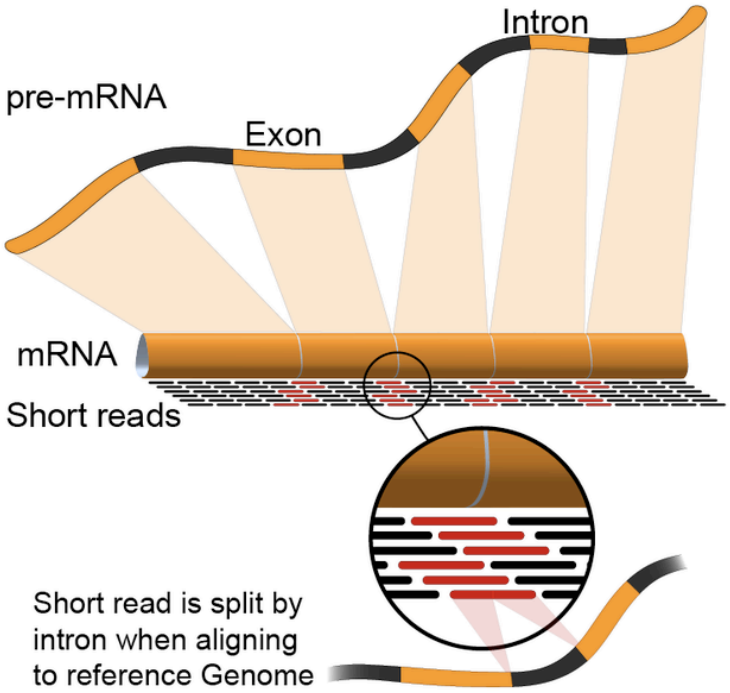
RNA-Seq Analysis

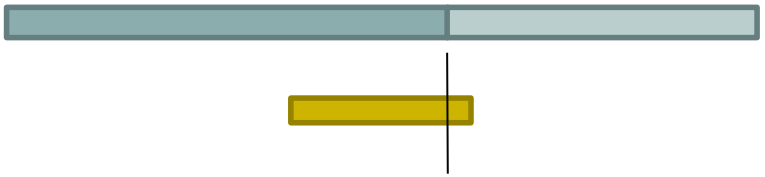
- High-throughput sequencing of the transcriptome provides a detailed picture of the genes that are expressed in a cell.

- Three main computational tasks:
  - **Mapping the reads to a reference genome**
  - **Assembling the reads into full-length or partial transcripts**
  - Quantifying the amount of each transcript.

Splicing

- Spliced alignment is needed for NGS reads.
  - → Aligning a read to two physically separate locations on the genome.
  - For example, if an intron interrupts a read so that only 5 bp of that read span the splice site, then there may be many equally good locations to align the short 5 bp fragment.
  - Another mapping problem.



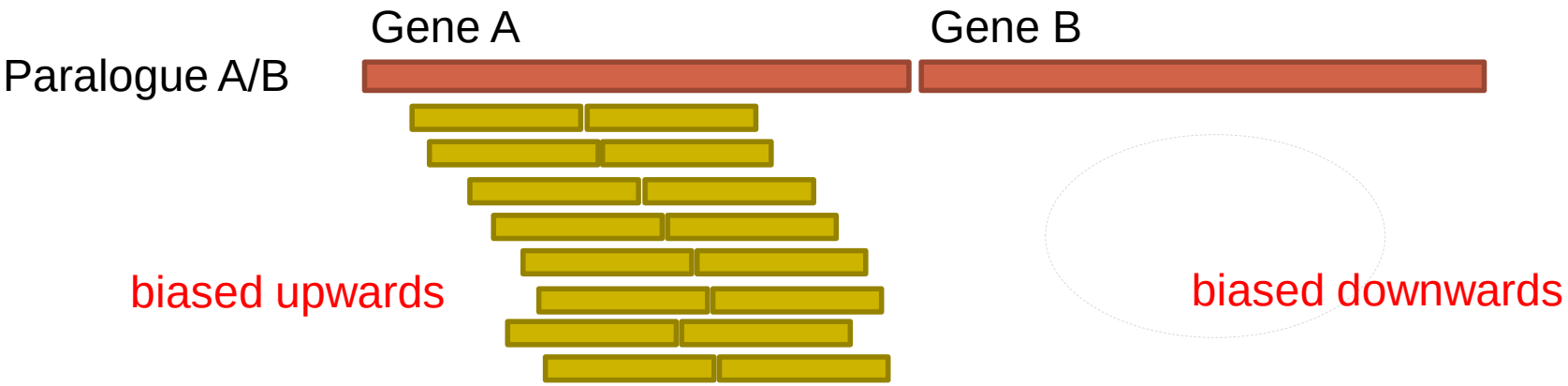


http://en.wikipedia.org/wiki/File:RNA-Seq-alignment.png

Gene expression

- Gene expression levels can be estimated from the number of reads mappig to each gene.

- For gene families and genes containing repeat elements, multi-reads can introduce errors in estimates of gene expression.



Conclusions

- Repetitive DNA sequences present major obstacles to accurate analysis in most of sequencing-based experimental data research.

- Prompted by this challenge, algorithm developers have designed a variety of strategies for handling the problems that are caused by repeats.

Conclusions

- Current algorithms rely heavily on paired-end information to resolve the placement of repeats in the correct genome context.

- All of these strategies will probably rapidly evolve in response to changing sequencing technologies, which are producing ever-greater volumes of data while slowly increasing read lengths.

Thank you very much.

The end.