Introduction to bioinformatics: working with strings

Brian Kissmer

USU Department of Biology

Nov. 19th, 2024

Learning objectives

1. Gain general familiarity with the field of bioinformatics
2. Develop a level of comfort working with strings in R

Outline

1. Introduction to bioinformatics
2. Some examples: genome assembly, alignment, variant detection, gene expression
3. Strings in R handout (Thursday)

Next unit

So far, we have learned about the building blocks for computational analysis of biological concepts

1. Simulations
2. Computational statistics
3. Linear/penalized regression

Over the next (and final) unit, we will explore how these and other methods are used to handle biological data, making inferences and predictions on complex (and often very large) data sets

Opening discussion

What do you think of when you hear the term ‘bioinformatics’? How is it different from what we’ve covered so far? Discuss with your group (~3 minutes).

What is bioinformatics?

Application of computational tools to biological data, usually involves:

1. Large, complex data sets
2. -omics data, e.g., genomics, transcriptomics, metabolomics, etc.
3. DNA or RNA sequences (i.e., text data)

Let's look at some examples of bioinformatics

1. Genome assembly and DNA sequence alignment
2. Variant calling
3. Gene expression analysis

Don't get bogged down by the small details. Instead, think about how we are using computational tools to make use of large biological data sets

Example 1

Genome assembly and DNA sequence alignment

Example 1

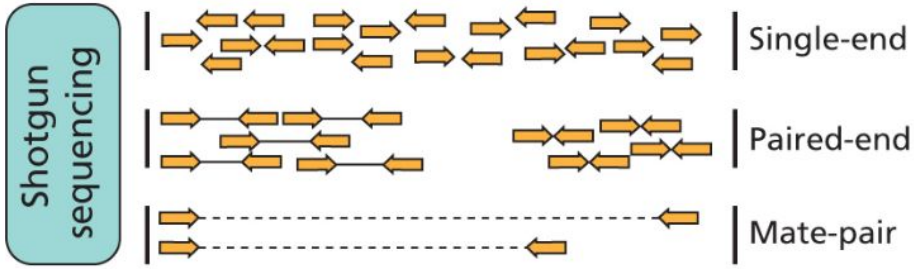
Genome assembly and DNA sequence alignment

How do we turn DNA sequence data into useful information?

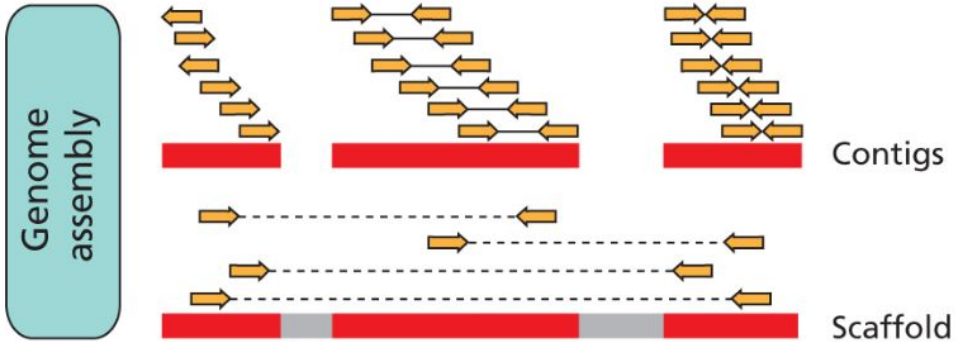
Typical DNA sequence data - Illumina

https://www.youtube.com/watch?v=fCd6B5HRaZ8

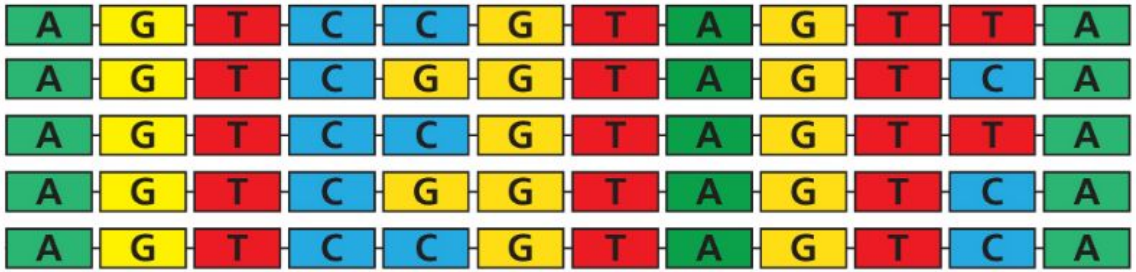
De novo genome sequencing and assembly

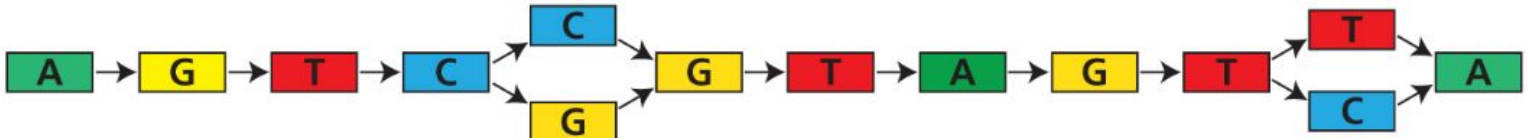


[Særte and Ravinet, 2019]

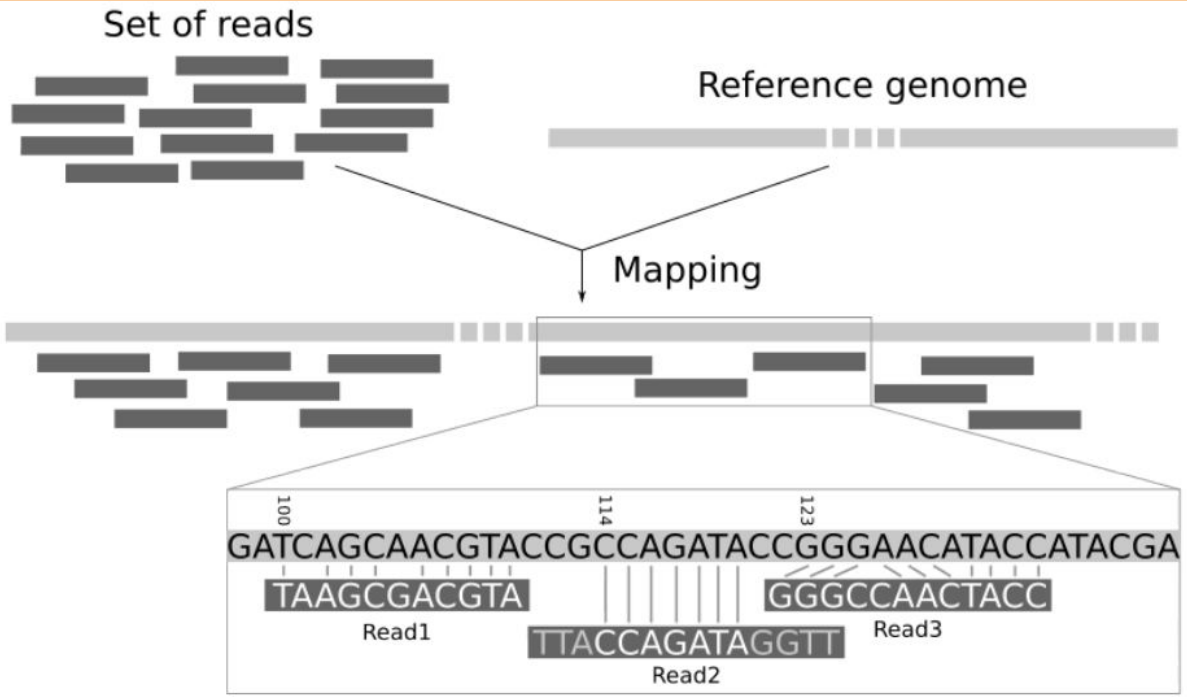


Linear versus non-linear genome models





Aligning short DNA sequences to a reference



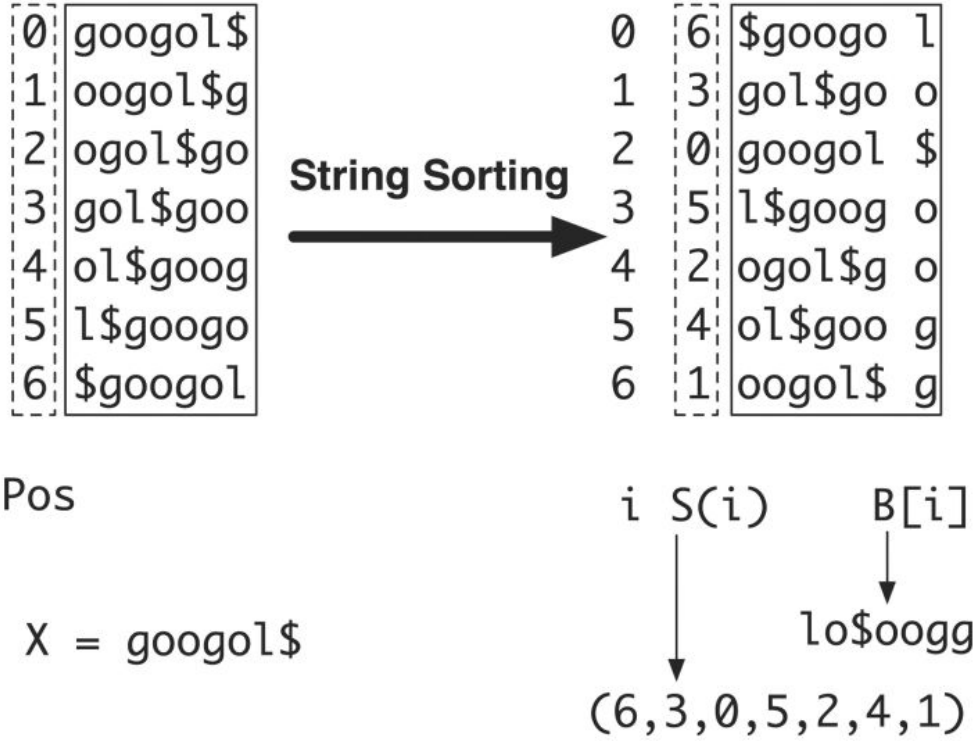
alignment or mapping is simpler than genome assembly

Aligning short DNA sequences to a reference

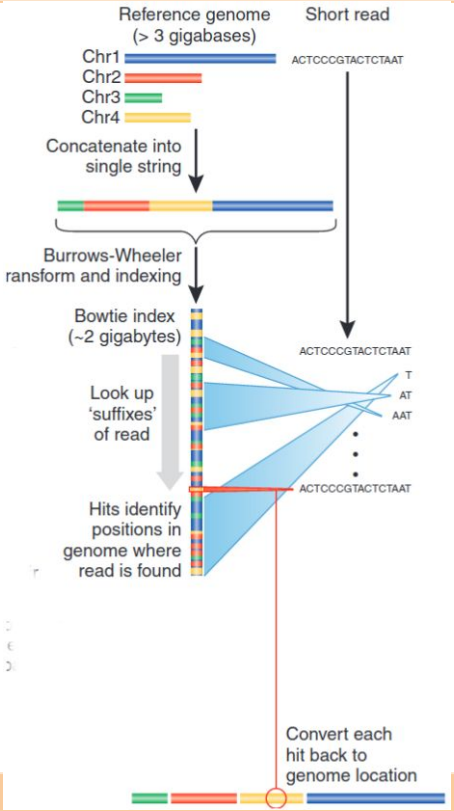
Problem: Genome files can be HUGE (billions of base pairs in a single text file), how do we work with them quickly?

Burrows-Wheeler transform speeds reference-based alignment

Constructing the suffix array and BWT for X=googol$ [Li and Durbin (2009)]



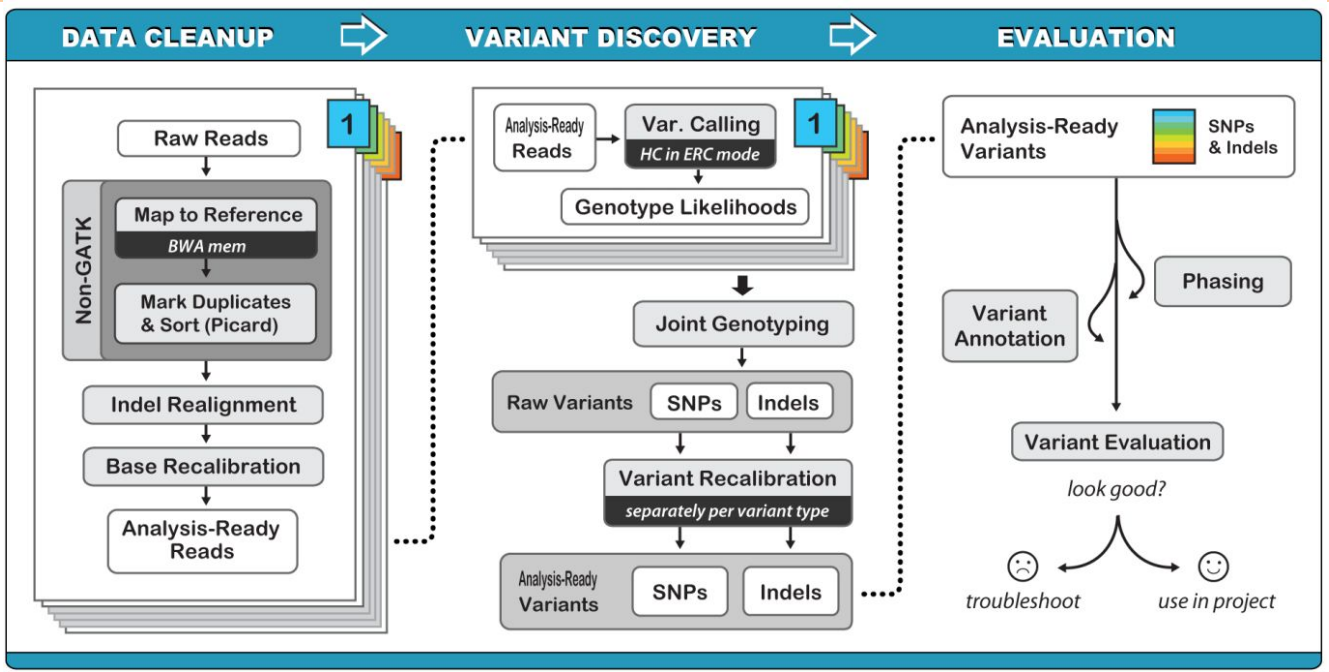
Burrows-Wheeler transform speeds reference-based alignment



Example 2

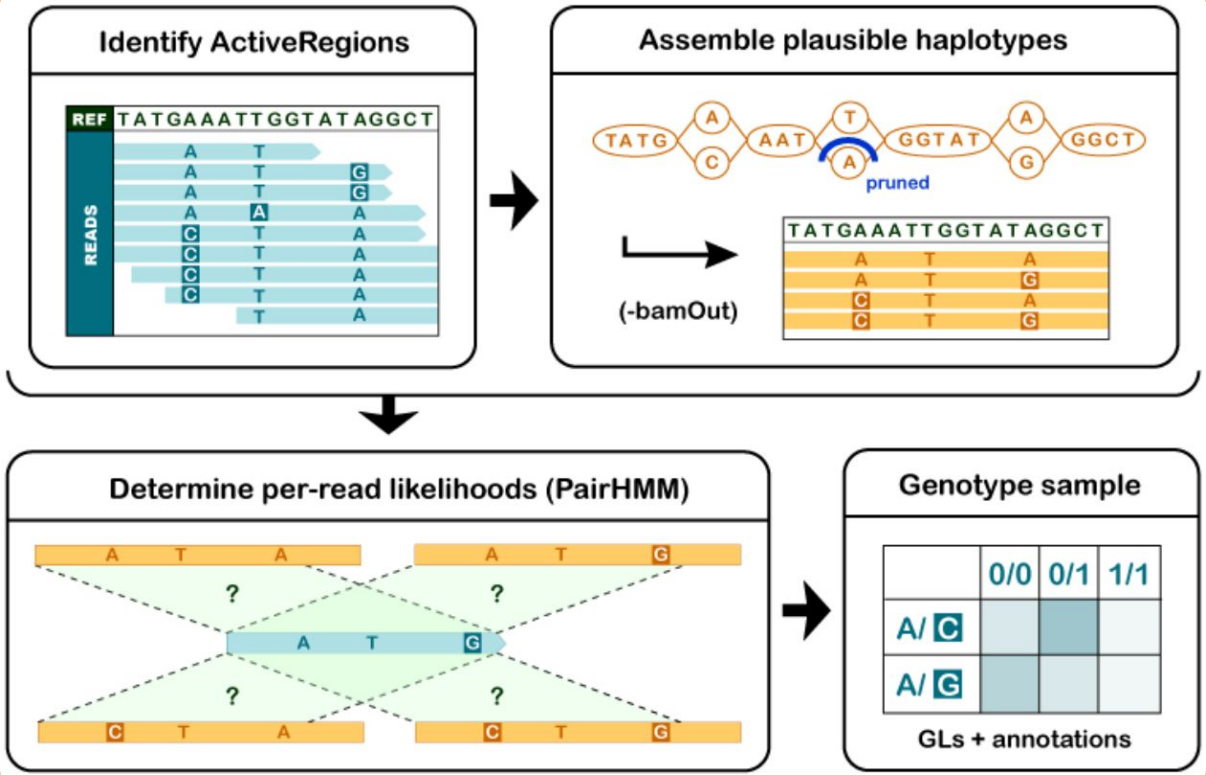
Genetic variant calling

How to get from DNA sequences to SNPs



GATK [Genome Analysis Tool Kit] best practices

Variant detection and genotyping with GATK's `HaplotypeCaller`



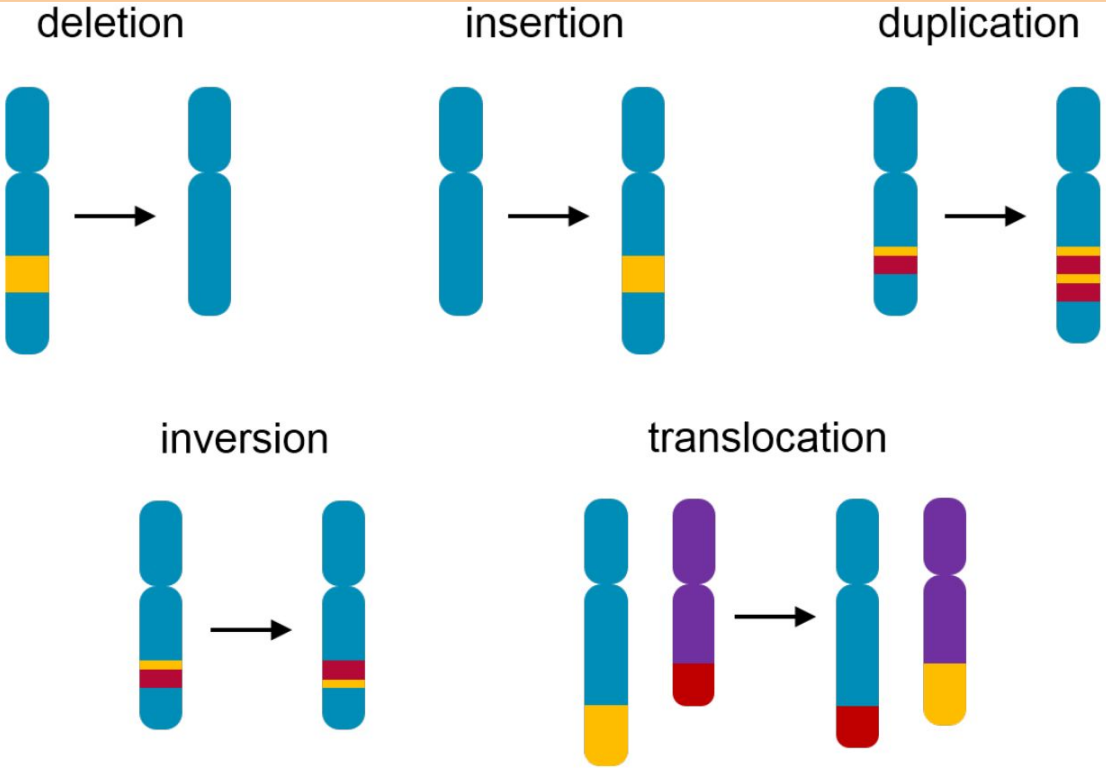
Variant detection and genotyping with GATK's `HaplotypeCaller`

This is great for single nucleotide polymorphisms (SNPs), but what about other kinds of genetic variation?

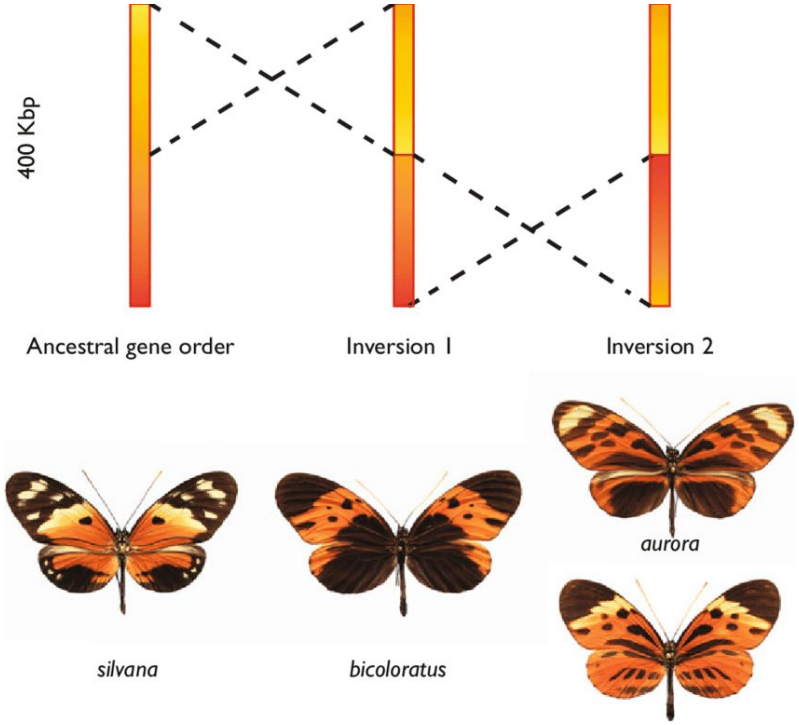
Structural variants, what they are, and why they matter

1. Structural variants comprise various forms of genome rearrangements
2. Structural variants are COMMON and pervasive (including in humans!)
3. Structural variants can affect phenotypes via several mechanisms
4. Human disease studies suggest they are at least as important as SNPs in explaining trait/disease variation

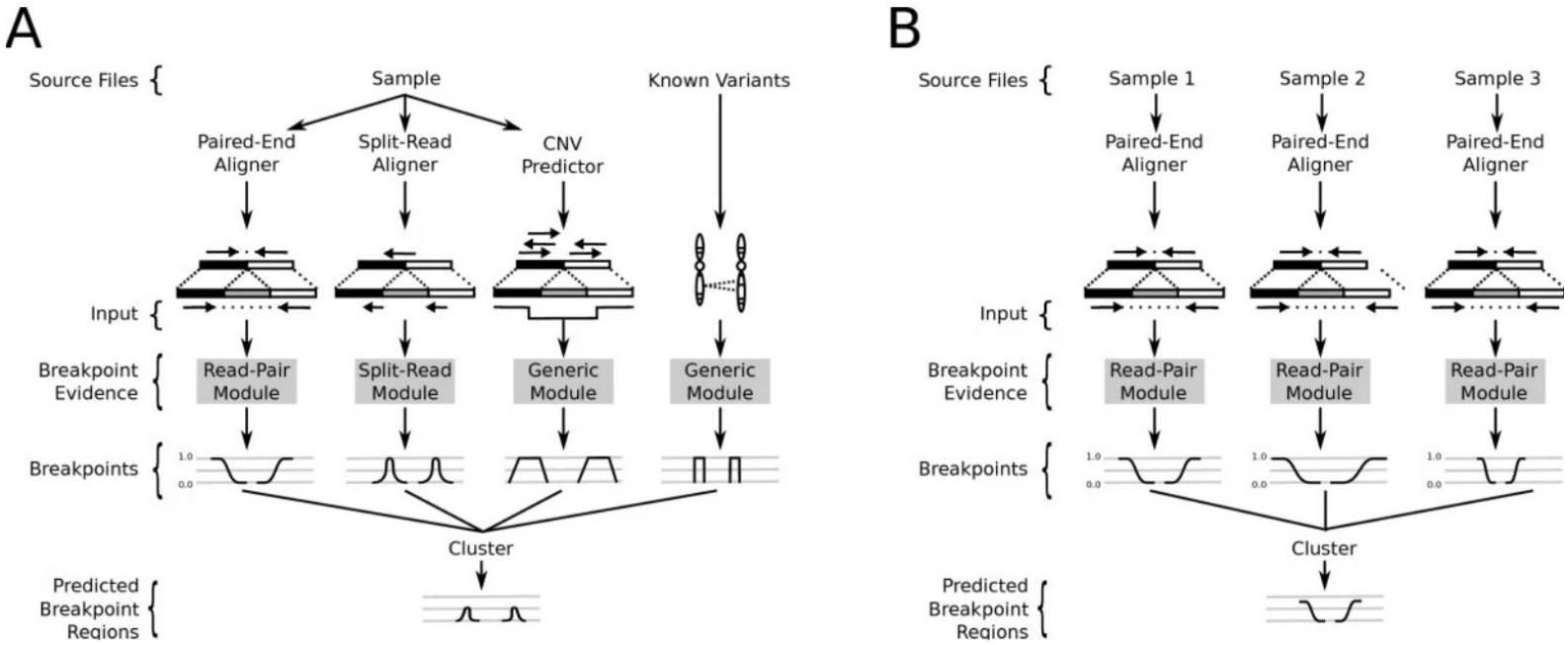
Types of structural variants



Impact of structural variants

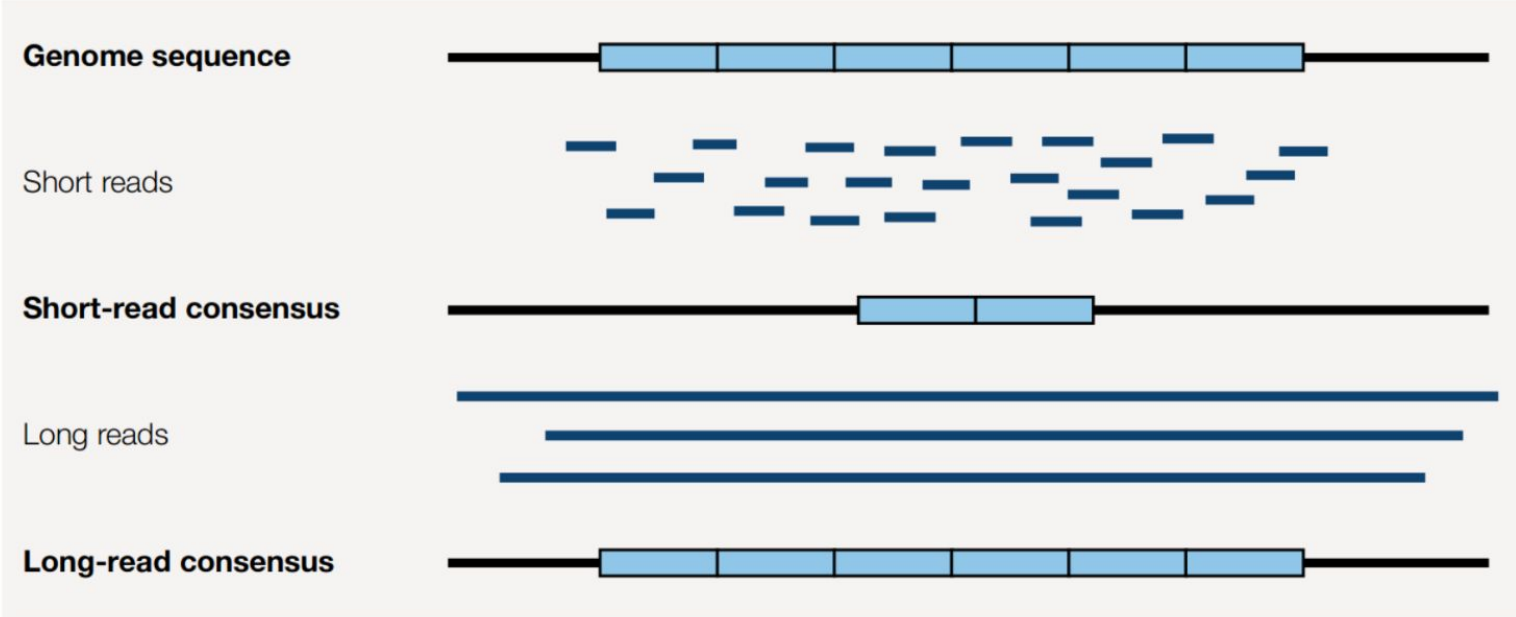


Structural variant signals in standard alignments



SV calls usually based on multiple lines of evidence

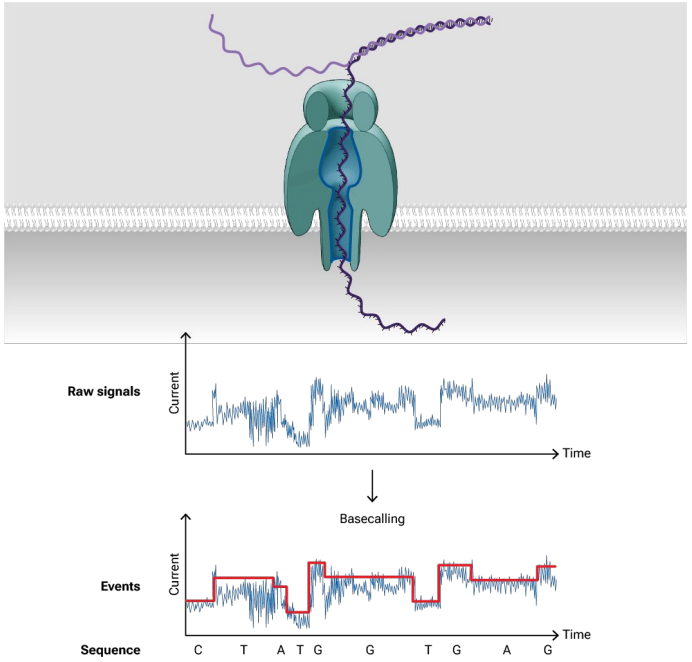
Long reads facilitate identifying structural variants



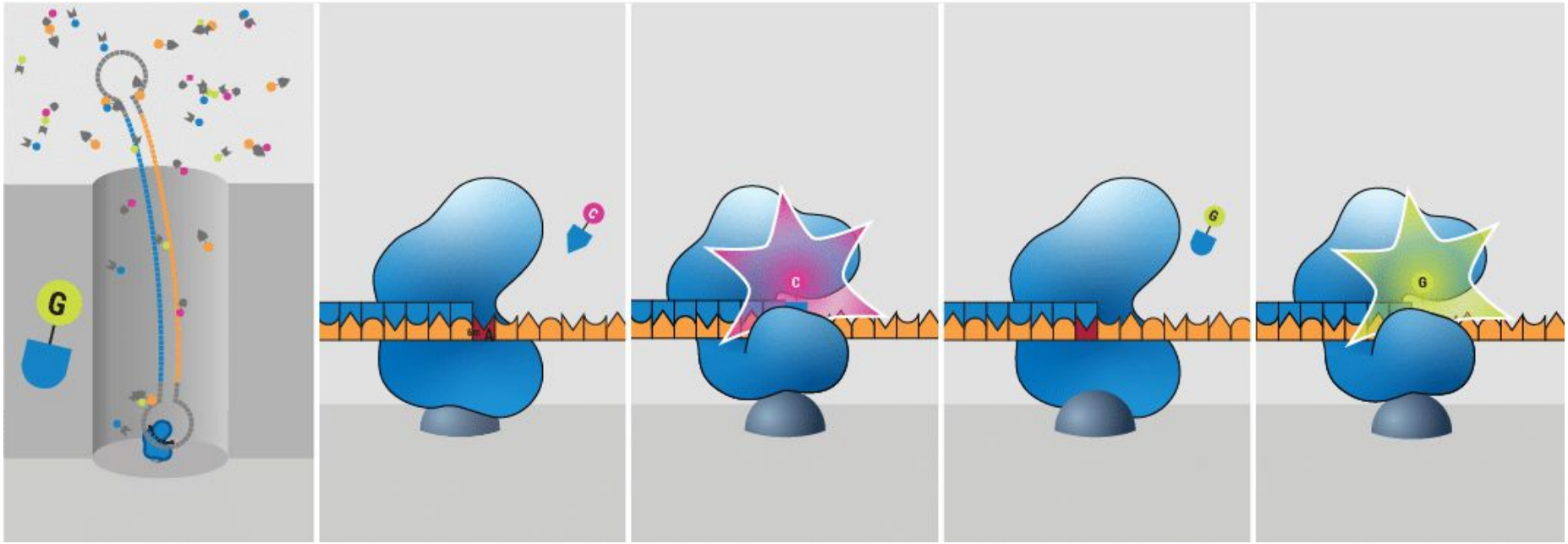
long reads recover tandem duplication not resolved with short reads

Long reads facilitate identifying structural variants- Oxford Nanopore

Reads can exceed 4Mb, while typical sequencing produces reads ~500bp in length



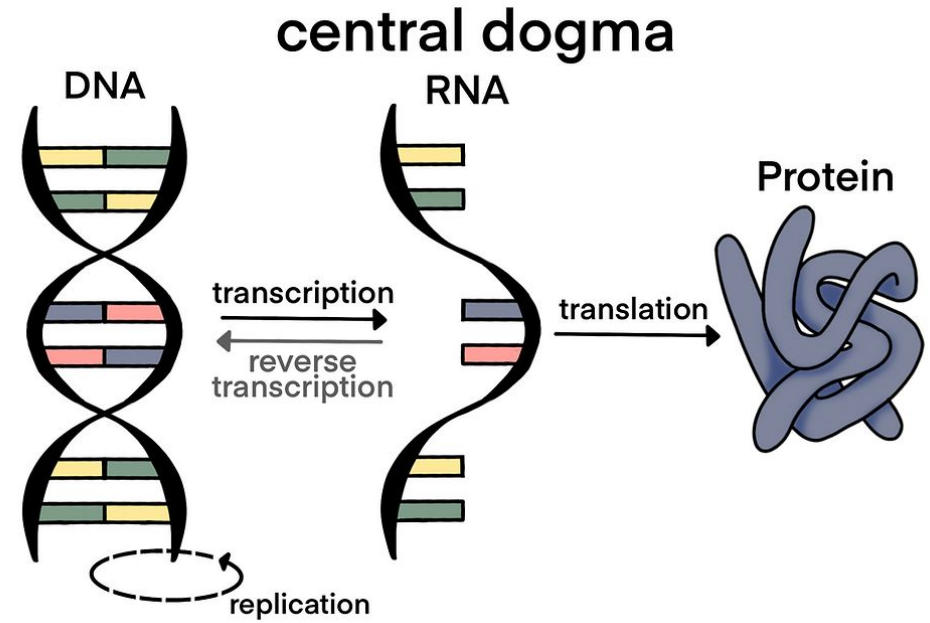
Long reads facilitate identifying structural variants- PacBio Hifi

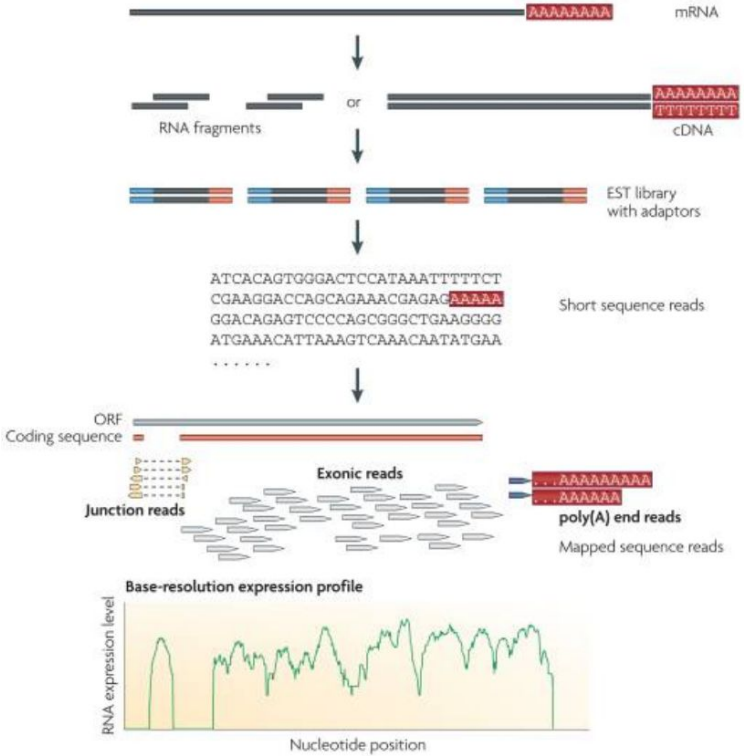


Example 3

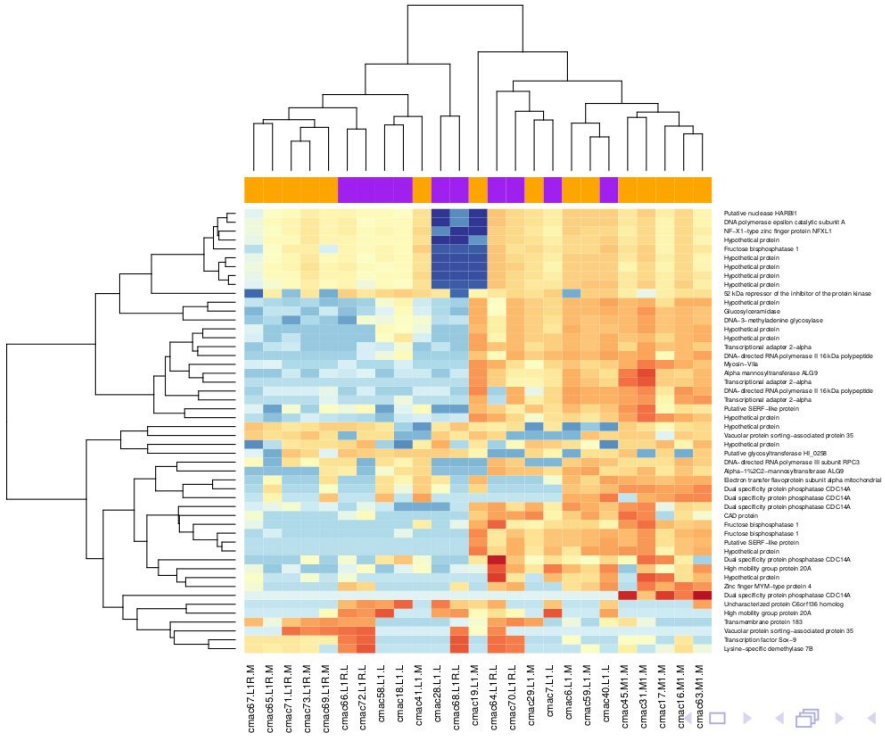
Gene expression analysis

A typical gene expression (RNAseq) experiment

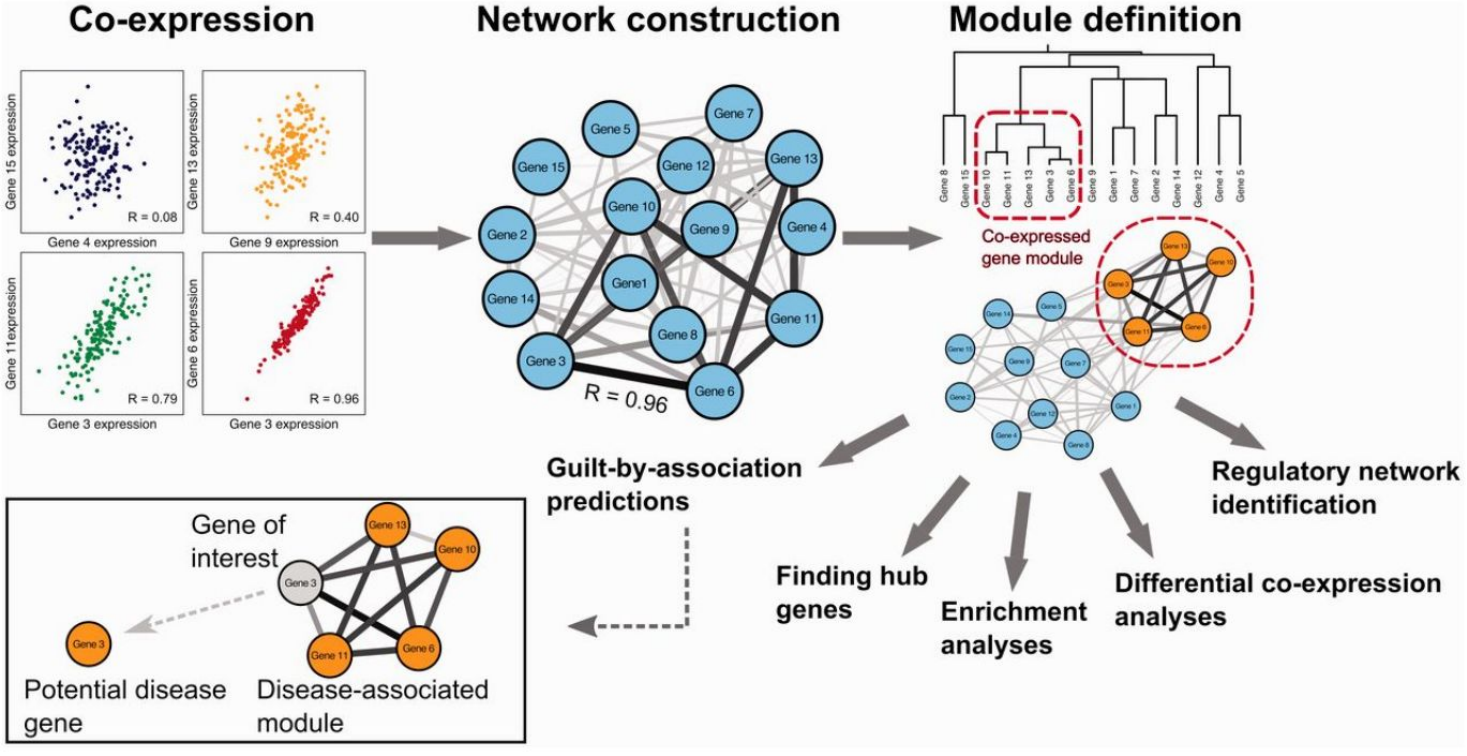




Differential expression of genes in seed beetles adapted to different hosts



Differential expression of genes in seed beetles adapted to different hosts



Bioinformatics and strings

Many of these examples involve analyzing textual data (or strings), rather than numerical data

Strings in R

On Thursday, we will explore strings in R