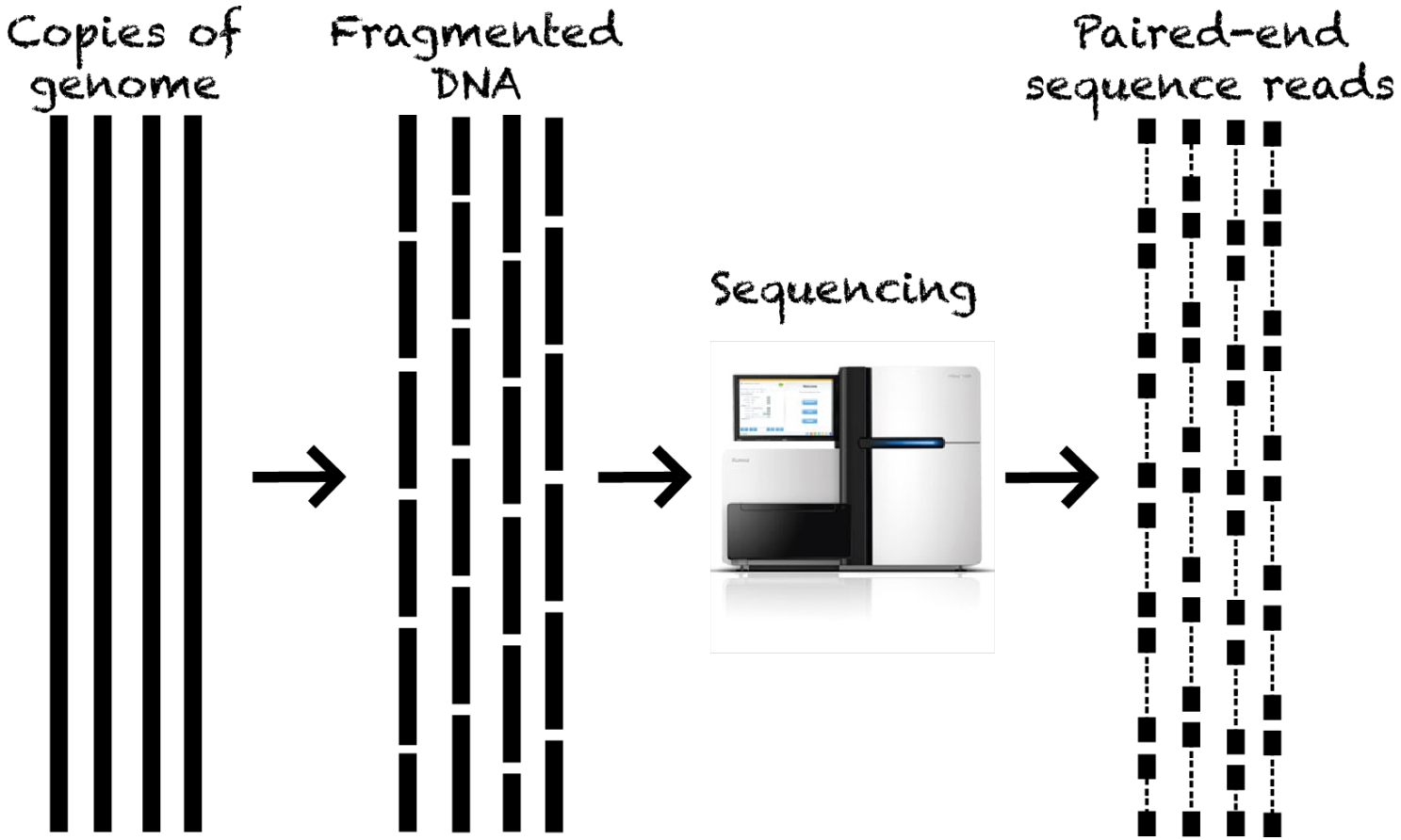
Background info for GENE315 Lab:

Module 2 - Analysis of genetic variation in humans

Identifying genetic variation

- Next Generation Sequencing (NGS) has revolutionized how we study genetic variation.
- It is now routine to generate a whole-genome characterisation of SNVs, InDels and SVs for an individual.
- Until recently, most sequencing was performed using short read (100-150bp) technologies (e.g., Illumina).
- More recently, long read technologies have become popular (e.g., PacBio and Oxford Nanopore).

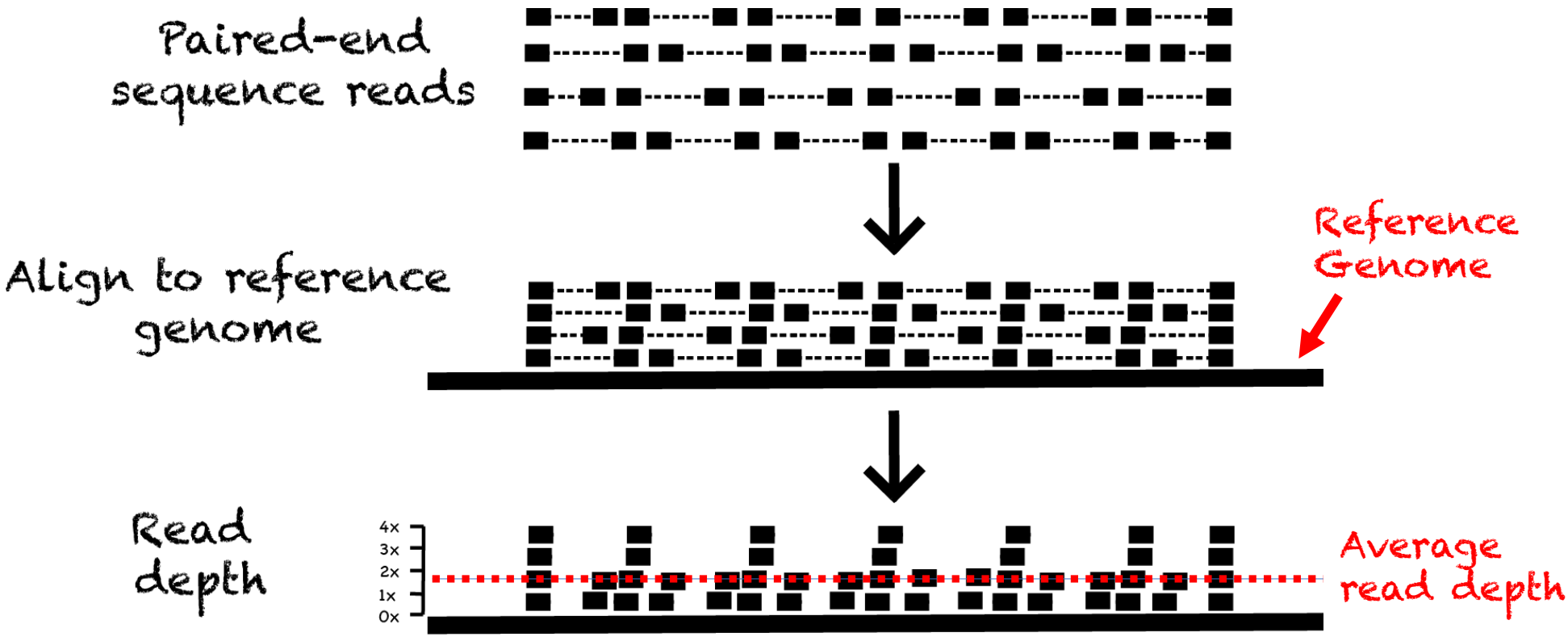
Recap - generating NGS data



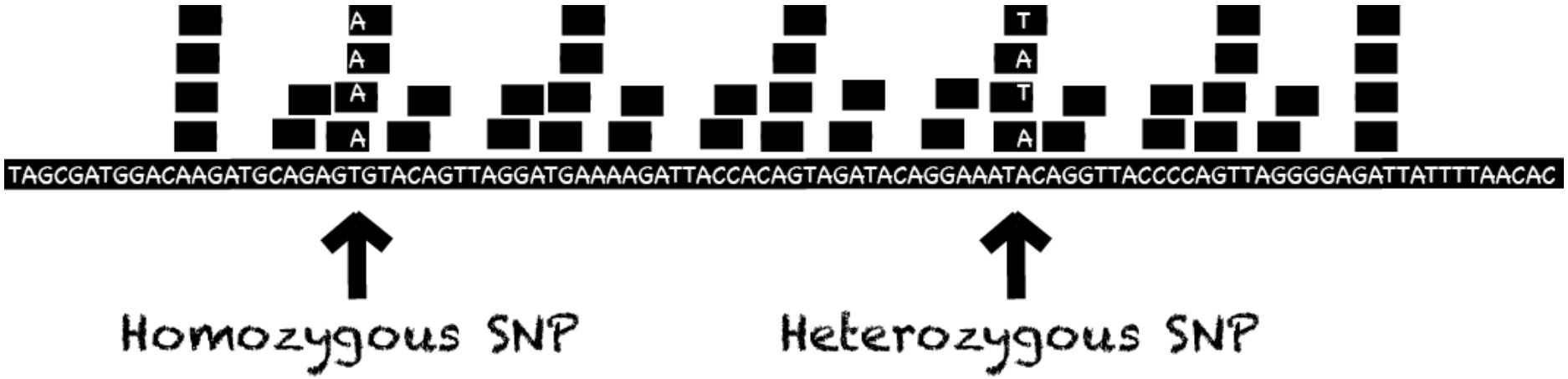
Identifying variation

- Sequencing in humans currently relies on the use of a Reference Genome to identify genetic variants (i.e., we don't (yet) *de novo* assemble human sequence to create a sequence assembly for each individual).
- In order to identify variation in the sequenced samples, the sequence data is compared to the existing reference genome to allow differences to be identified.
- This involves finding the place in the reference genome that each read matches to.
- Due to high sequence similarity within members of the same species, most reads should map to the reference.

Mapping reads to the reference genome



Aligned data - finding SNPs

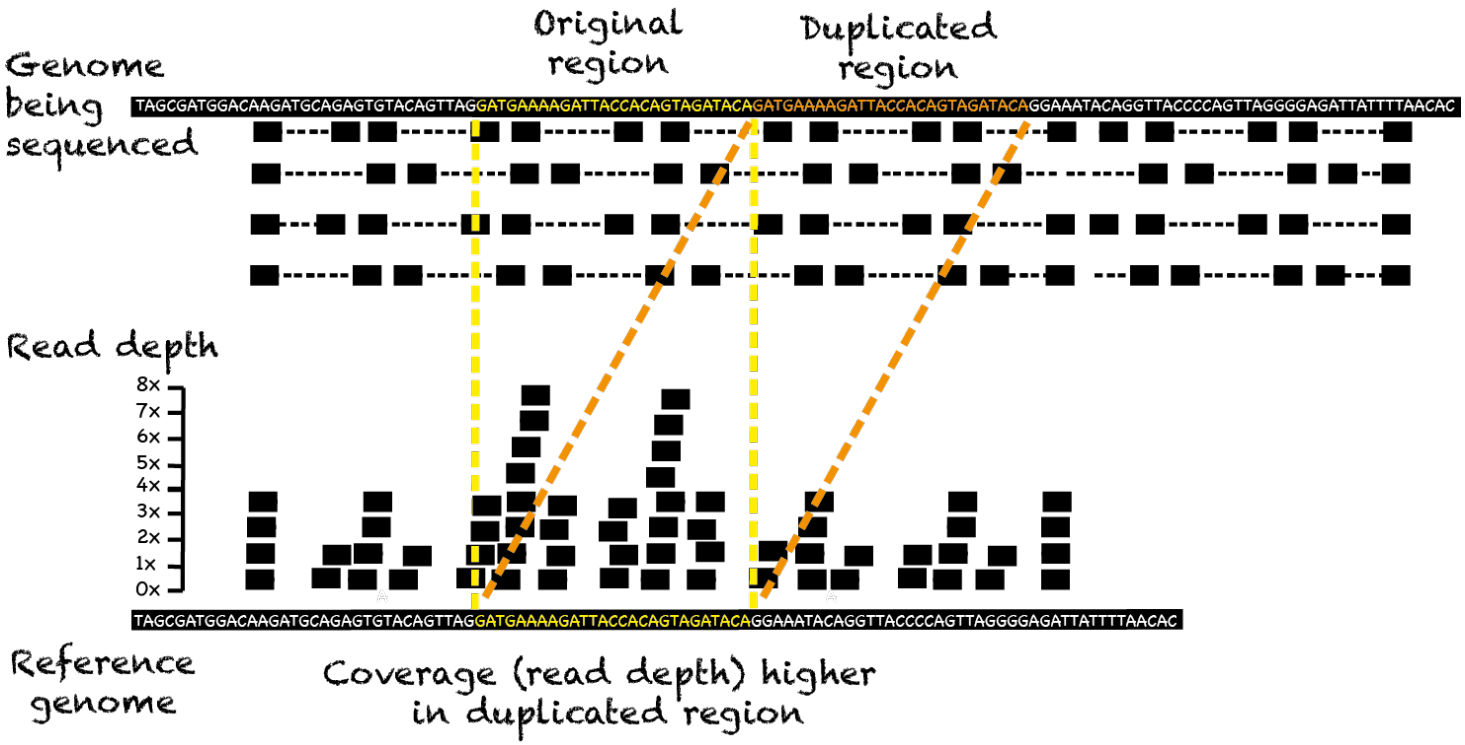


- Alignment to the reference genome allows identification of single base differences (i.e., SNVs) in a new sample.
- There are **MANY** software tools that can perform SNV detection.
- A reasonable read depth is required for accurate variant detection, particularly for heterozygotes.

Using read depth to find copy number changes

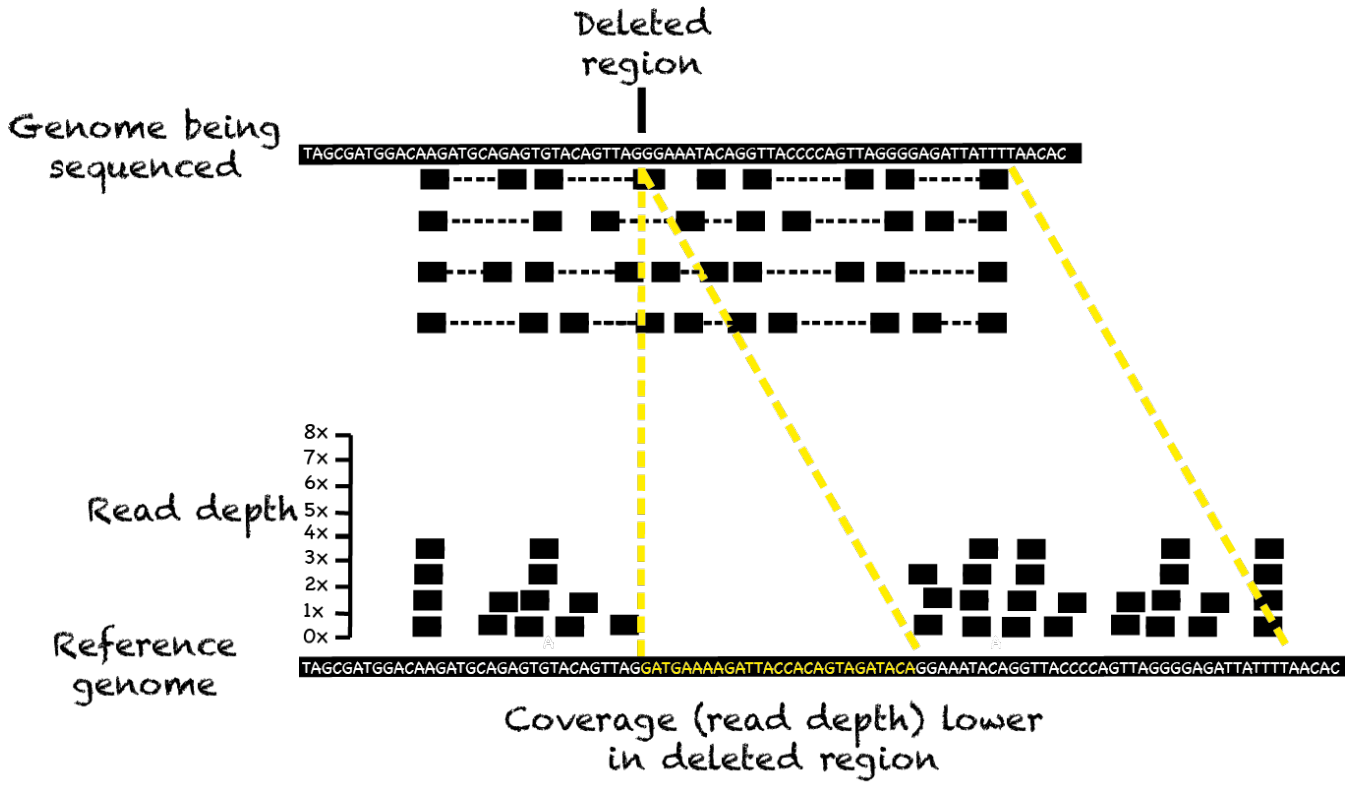
- Sequence data can also be used to identify changes in copy number for specific regions of DNA.
- Various methods exist, but a relatively simple approach involves identifying changes in read depth that are consistent with an increase or decrease in DNA copy number.

Detecting DNA duplication



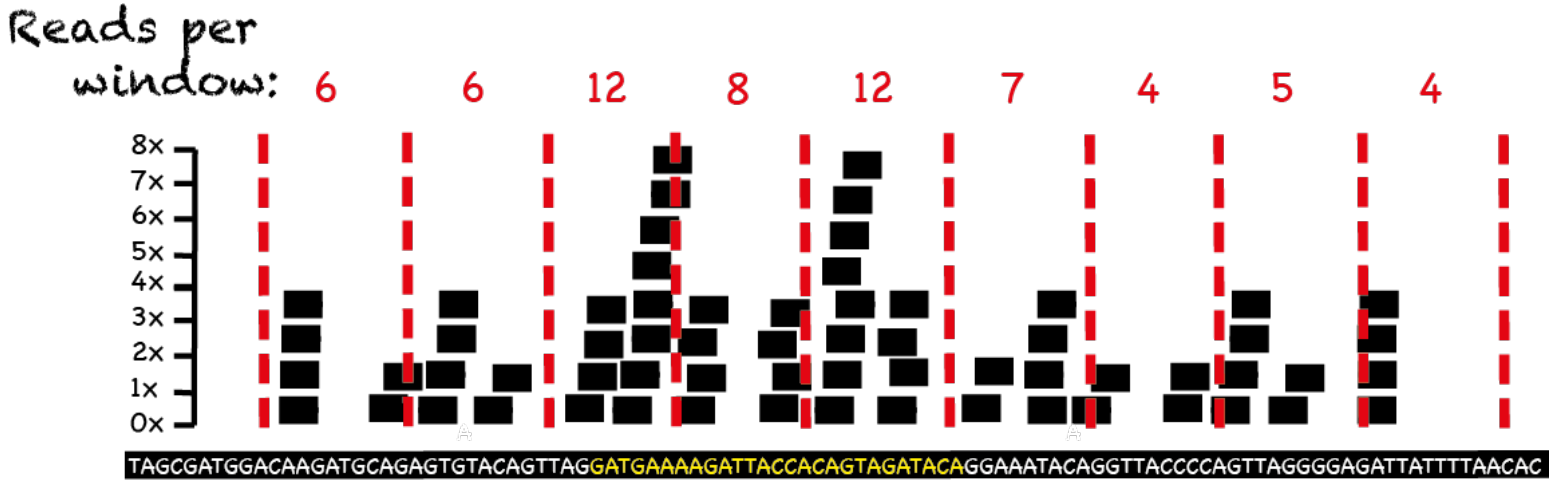
HIGHER read depth in duplicated regions

Detecting DNA deletion



LOWER read depth in deleted regions

Quantifying read depth across region



- Break the region of interest into evenly sized windows, and count the number of reads that start in each window.
- Use a statistical algorithm to identify "change points" - places where the average number of reads per window changes (up or down), over a substantial distance.
- Consistent changes in read depth over a region can be indicative of duplications or deletions.