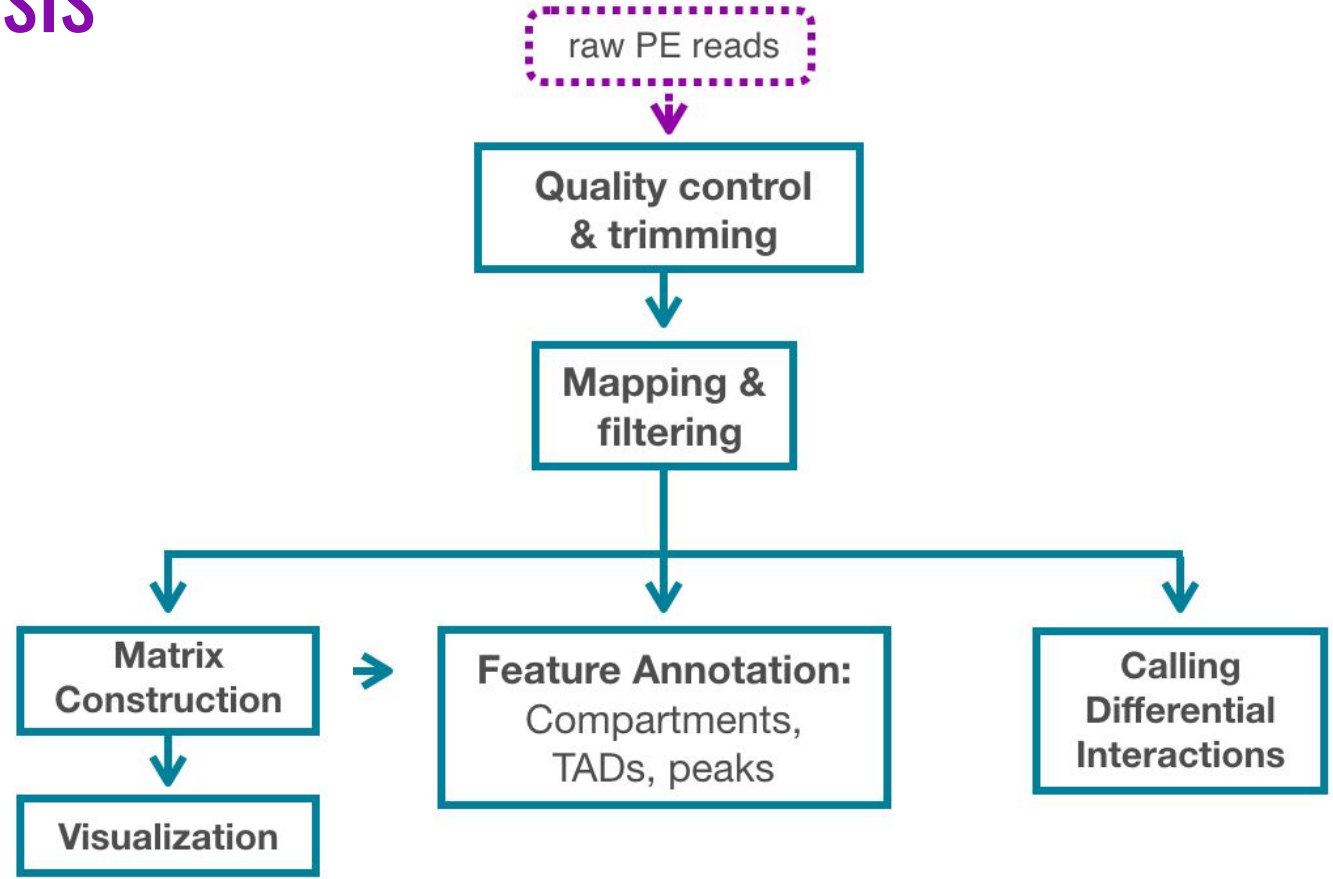
Mapping and processing Hi-C data

Hi-C alignment strategies
Obtaining informative Hi-C pairs

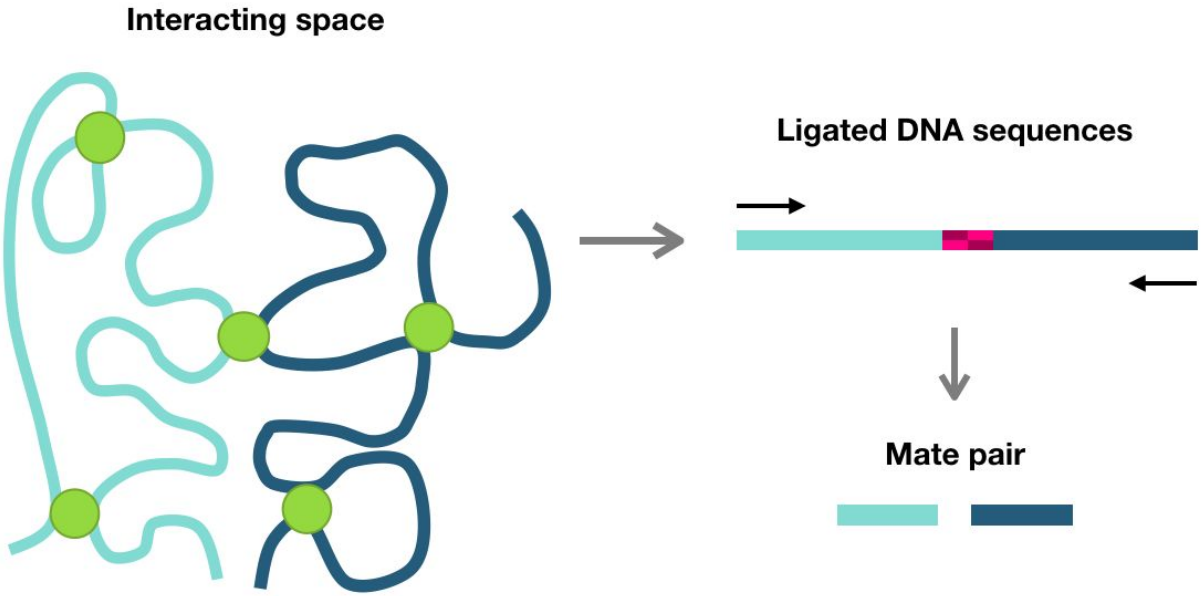
Hi-C Analysis Overview



Learning objectives

- HiC mapping strategies
- HiCUP pipeline
- Troubleshoot the protocol

Hi-C Paired-end reads

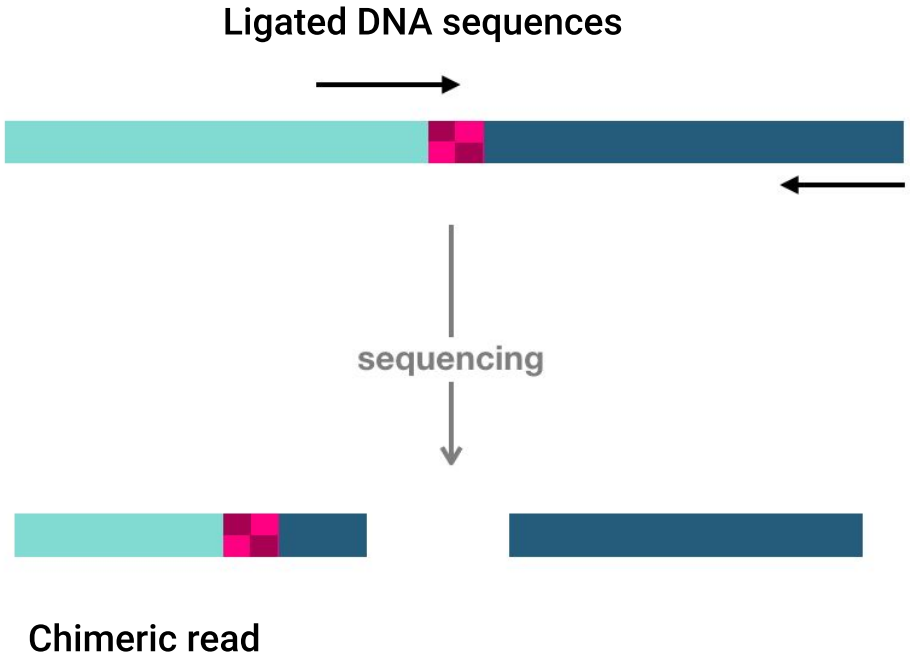


Based on Lieberman-Aiden *et al.* (2009)

Chimeric DNA fragments make Hi-C alignment a challenge

A chimeric read is created when sequencing of the ligation product is performed **across** the ligation junction (modified restriction site).

(Lun & Smyth, 2015)



Mapping strategies

- Iterative
- Truncating
- Local

Iterative mapping

Each read is truncated to a 5´ subsequence (25 bp) and gradually extended from the 3´ end until it aligns uniquely.

Iterative mapping concludes when either the read is uniquely aligned, or the maximal read length is reached

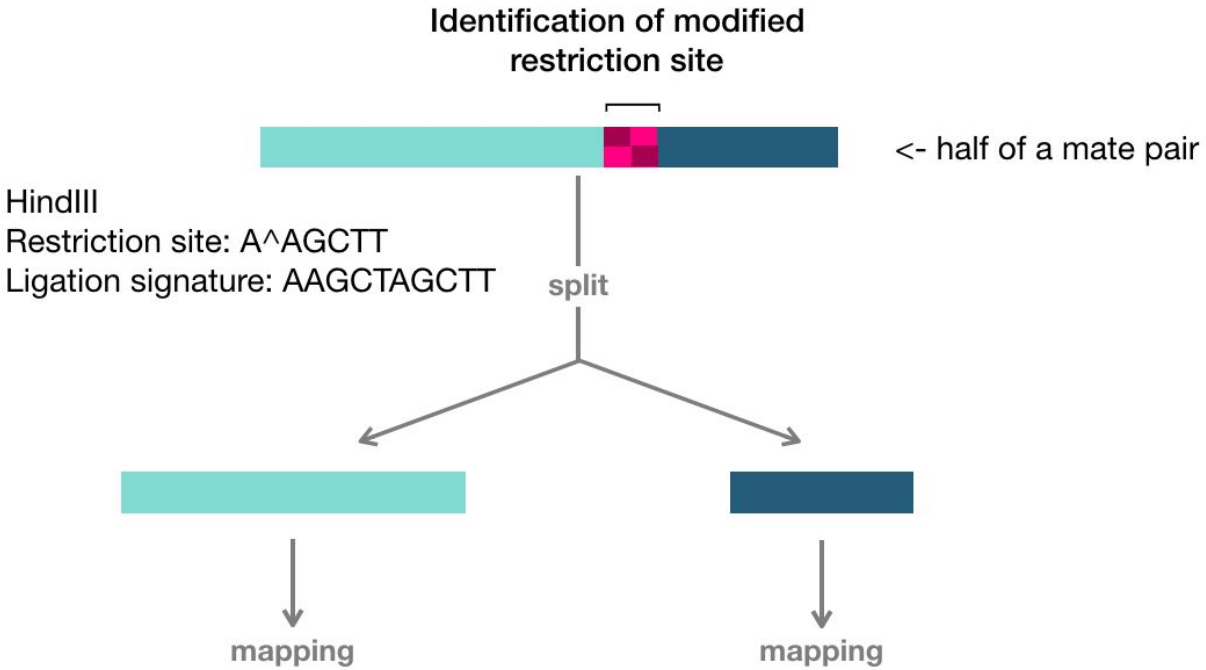
Based on Lajoie, Dekker & Kaplan (2015)

Ligation junction identification

/ Pre-splitting

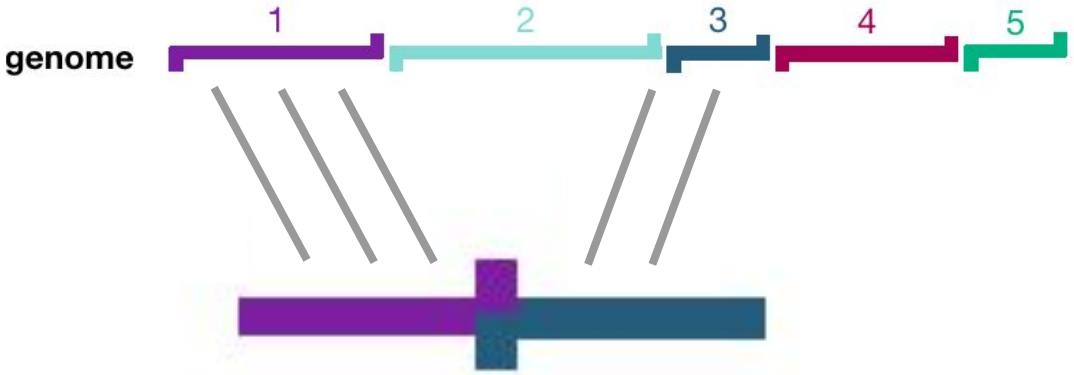
* The junction might not be covered with short reads

* Very useful for chimeric read alignment



Local alignment

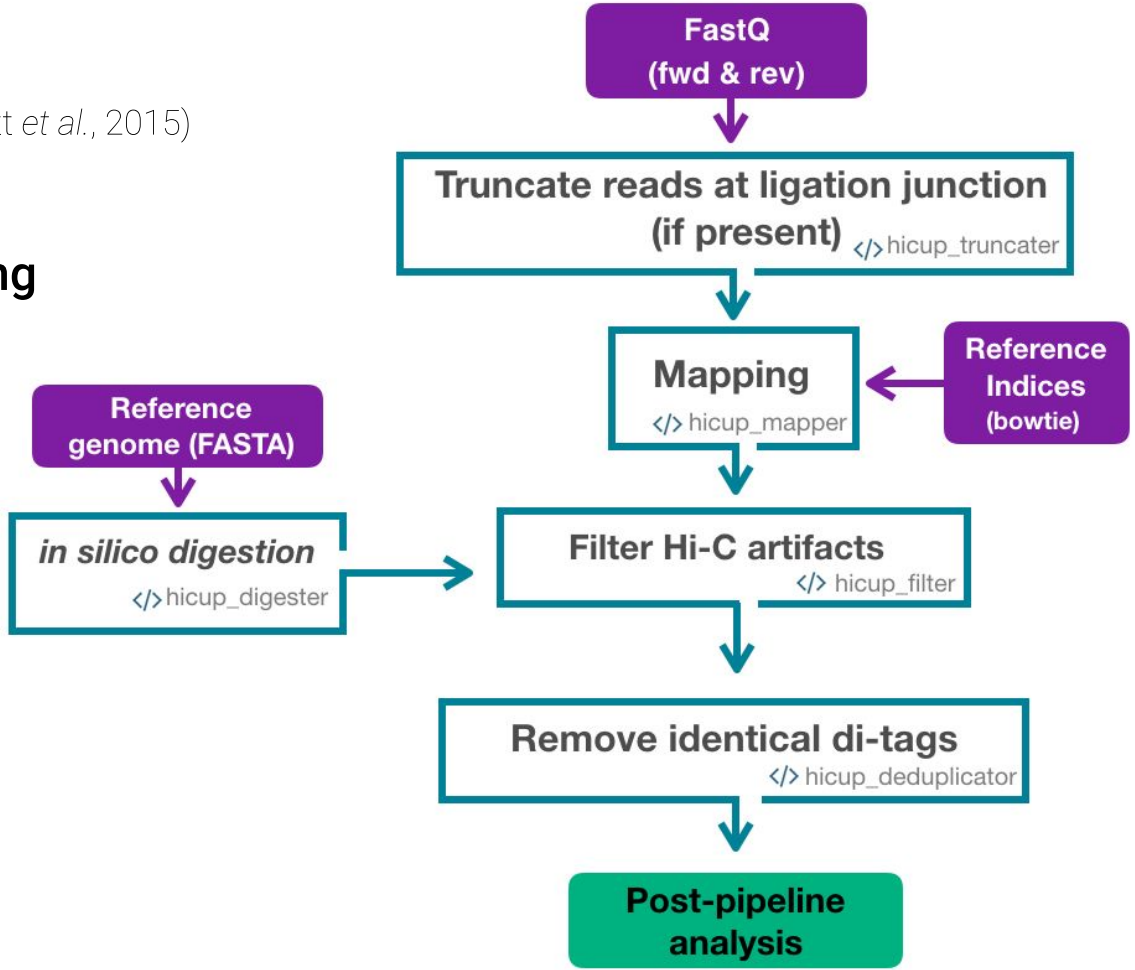
Some aligners
do not try to map the entire read to a single position
Reads can be split if each part map well to different genomic positions (bwa mem, bowtie2, hisat2)



HiCUP (Wingett *et al.*, 2015)

Mapping and pre-processing of Hi-C reads

HiCUP takes PE FASTQ files with a reference genome and associated aligner indices and reports valid di-tags in BAM/SAM format.



Based on Wingett *et al.* (2015)

Mapping with HiCUP

- Truncate reads at the ligation junction (if present) — hicup_truncater
- Choose between bowtie1 and bowtie2, generate index
- In silico digestion of genome for fragment assignment — hicup_digester
- Map Fwd and Rev reads independently — hicup_mapper
- Filter out common Hi-C artefacts — hicup_filter
- Keep unique high quality alignments — hicup_deduplicator
- Re-pair both ends

(Wingett *et al.*, 2015)

Obtaining informative Hi-C pairs (di-tags)

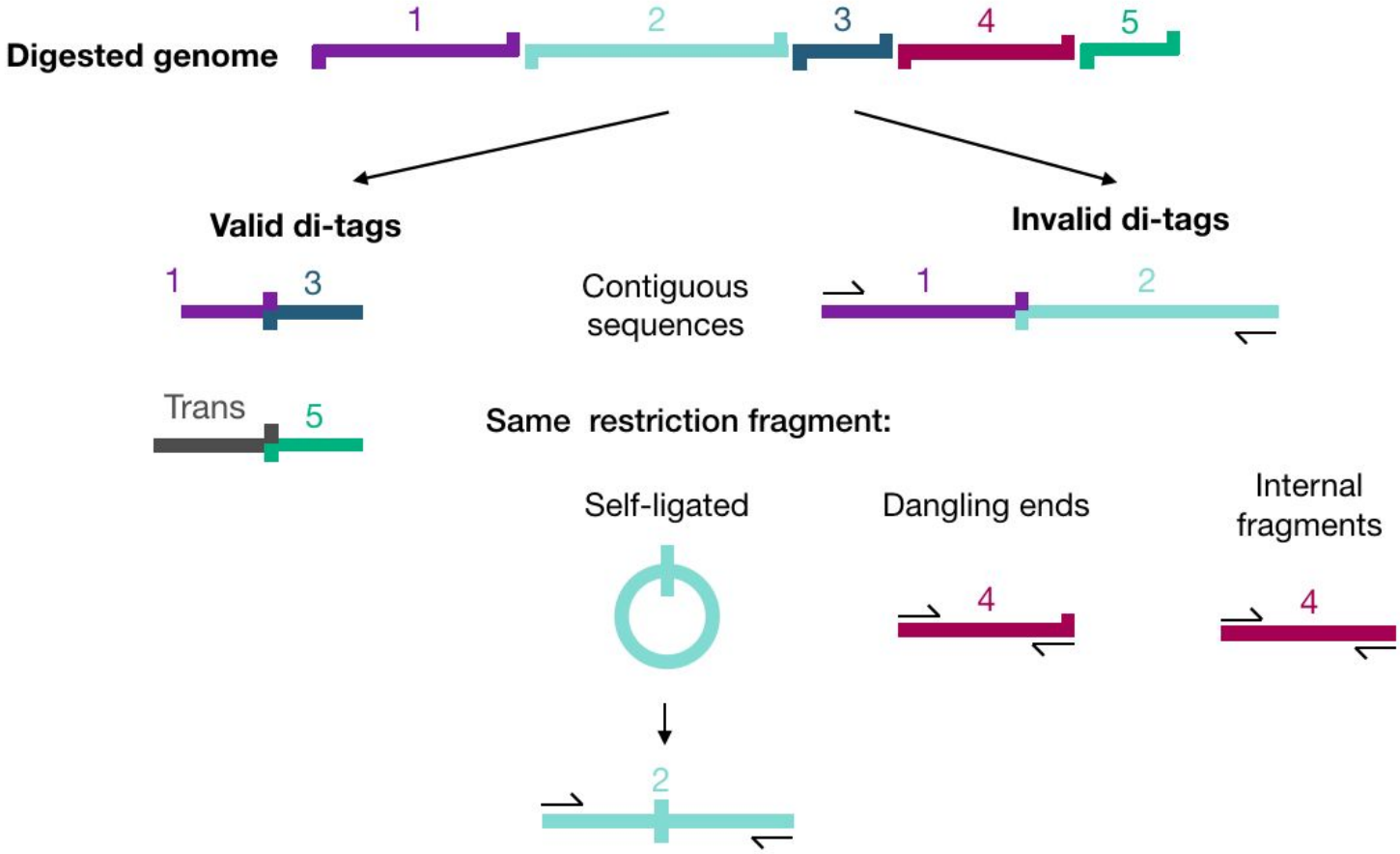
Hi-C products

- Informative HiC pairs
- Re-ligation of adjacent restriction fragments
- Intra-fragment reads
  - Circularization
  - Dangling ends
  - Internal

How to identify Hi-C byproducts?

- Assign each read end to a restriction fragment.
- Classify each read pair with respect to location of restriction fragments and read pair orientation.

Hi-C products



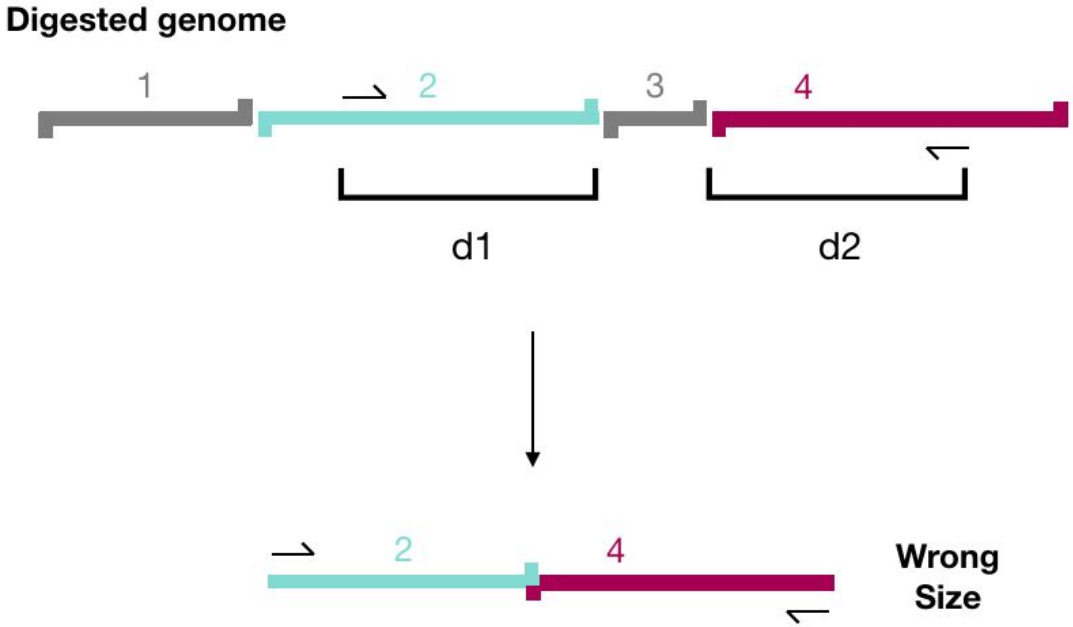
Troubleshoot the protocol with byproduct evidence

- High abundance of contiguous sequences indicates poor digestion
- High abundance of dangling ends indicates failure of “chewback”
- Internal fragments suggest inefficient pull down, digestion at non canonical size

Size distribution

Read pairs should fall on the expected size distribution of the library preparation step

d1 + d2 > expected size range



The cis / trans ratio

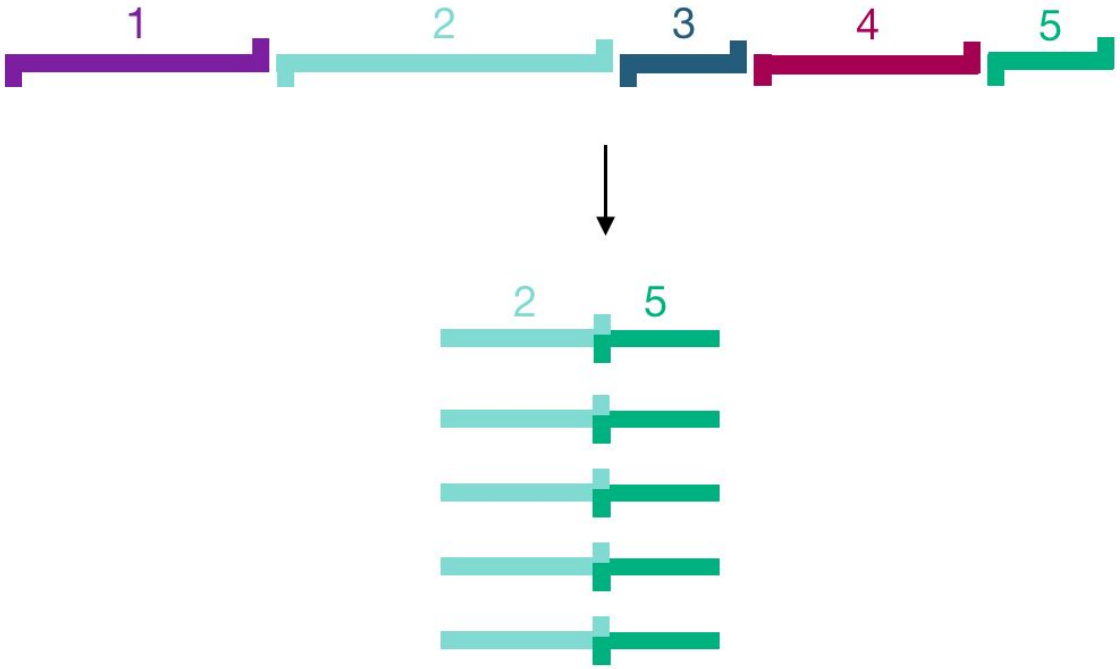
- **Assumes that biologically, cis read pairs are more abundant** (might not be the case for some plant genomes)

- Spurious ligation events tend to enrich for trans ratio

- Might indicate problems with fixation step

De-duplication

Duplicate read pairs
likely arise from PCR



Practical

- Generate genome index
- Generate restriction fragment digested genome
- Run HiCUP
  - HiCUP digester
  - HiCUP wrapper: truncater, mapper, filter, deduplicator
- Interpret the QC report

Resources

- https://www.youtube.com/watch?v=xWpjlXnsOU4
- https://www.bioinformatics.babraham.ac.uk/projects/hicup/read_the_docs/html/index.html