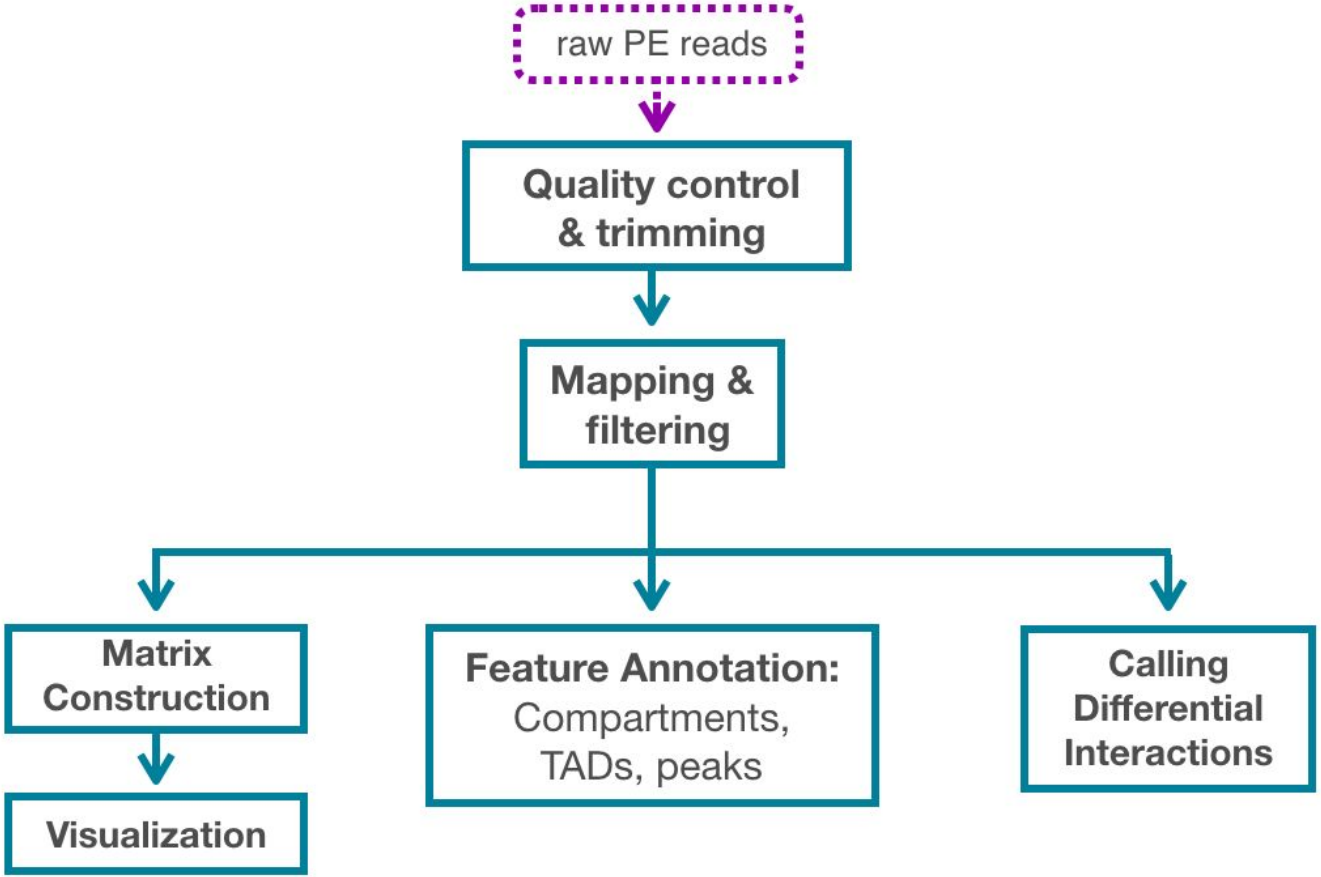
Generating a HiC matrix

Recap


raw PE reads
Quality control & trimming
Mapping & filtering
Matrix Construction
Visualization
Feature Annotation: Compartments, TADs, peaks
Calling Differential Interactions

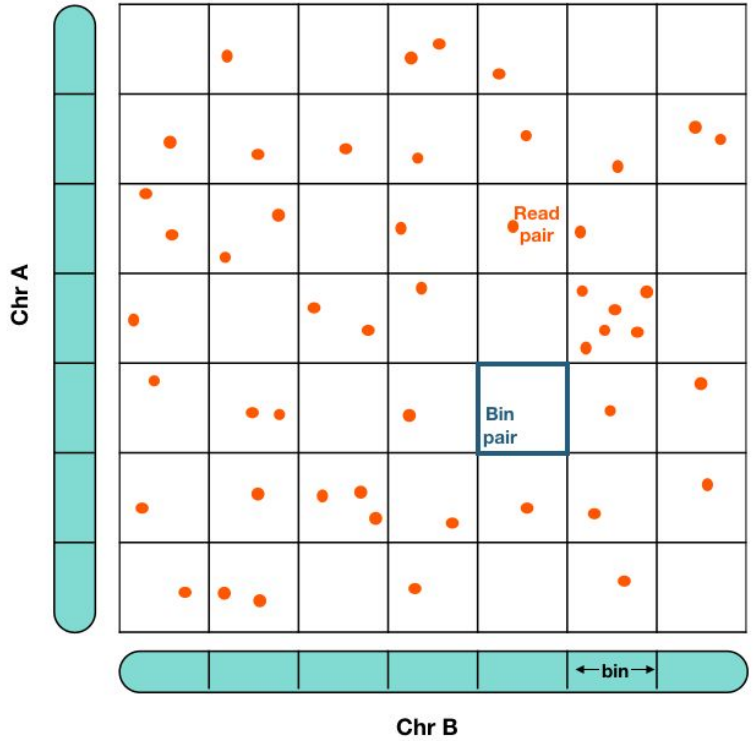
Learning objectives

- From mapped filtered reads to corrected contact maps
- Understanding biases affecting Hi-C matrices
- How to correct for those biases

Matrix binning

- 2D space results in low coverage for a single specific interaction
- Summarize read pairs into genomic windows of fixed size

- Each bin size yields different levels for analysis:
  - Peaks 500bp - 25kb
  - TADs / Sub compartments 500bp - 50 kb
  - Compartments 100kb - 500kb



Based on Lun & Smyth (2015)

Sources of bias

- Spurious ligation
- Length of restriction fragments
- Nucleotide composition
- Mappability

Matrix correction

- Equal visibility assumption:

- if all the genome had equal visibility (i.e. no bias), then each row and each column of the matrix should have the same number of contacts

Bin level filtering prior to balancing

- Remove low count bins
- Remove bins that overlap a blacklist
- Distance filter
- Use a variance cutoff

Practical

- Obtaining multi-resolution balanced HiC matrix
  - .hic
  - .cool

Resources

Common matrix formats

- .hic
  - Binary format compatible with juicer / juicebox
  - Stores many bin sizes and normalizations
  - Extract data: API straw, dump
- .cool
  - Binary format based on hdf5
    - Bin matrix
    - Contacts matrix
    - Chromosome matrix
    - .mcool is analogous to .hic
    - Extract data: cooler dump, python API