UMASS
PROGRAM IN SYSTEMS BIOLOGY

4D Nucleome

Hi-C Data Analysis Bootcamp

Harvard Medical School,
May 8th, 2018

Johan Gibcus

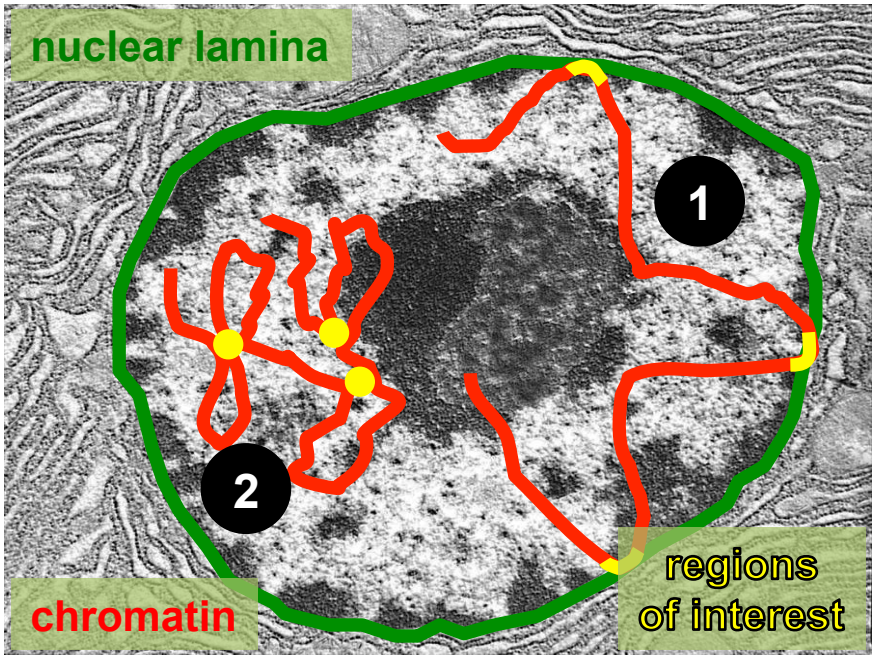
Presentation Overview

- Available (omics) techniques
  - See the forest for the tree
- How Hi-C works
  - The Hi-C protocol
  - Hi-C quality and QC
  - From bench to bites
- Hi-C enrichment
  - single cell and 5C
  - Less to see more

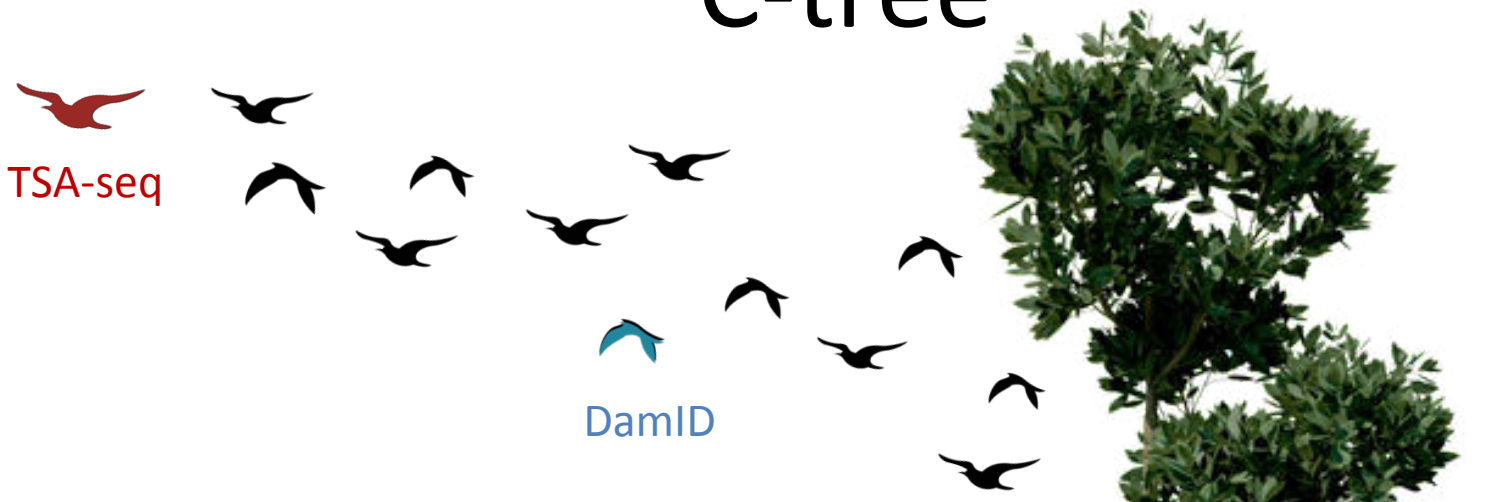
Molecular methods for studying genome organization

1. *Locus-landmark:* Measure interactions of genomic loci with relatively fixed nuclear 'landmarks'
   - Techniques: ChIP, DamID, TSA-seq

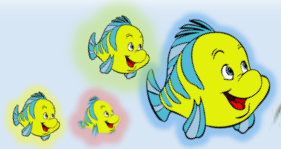
2. *Locus-locus:* Measure interactions between genomic loci
   - Techniques: 3C, Hi-C, ChIA-PET etc.



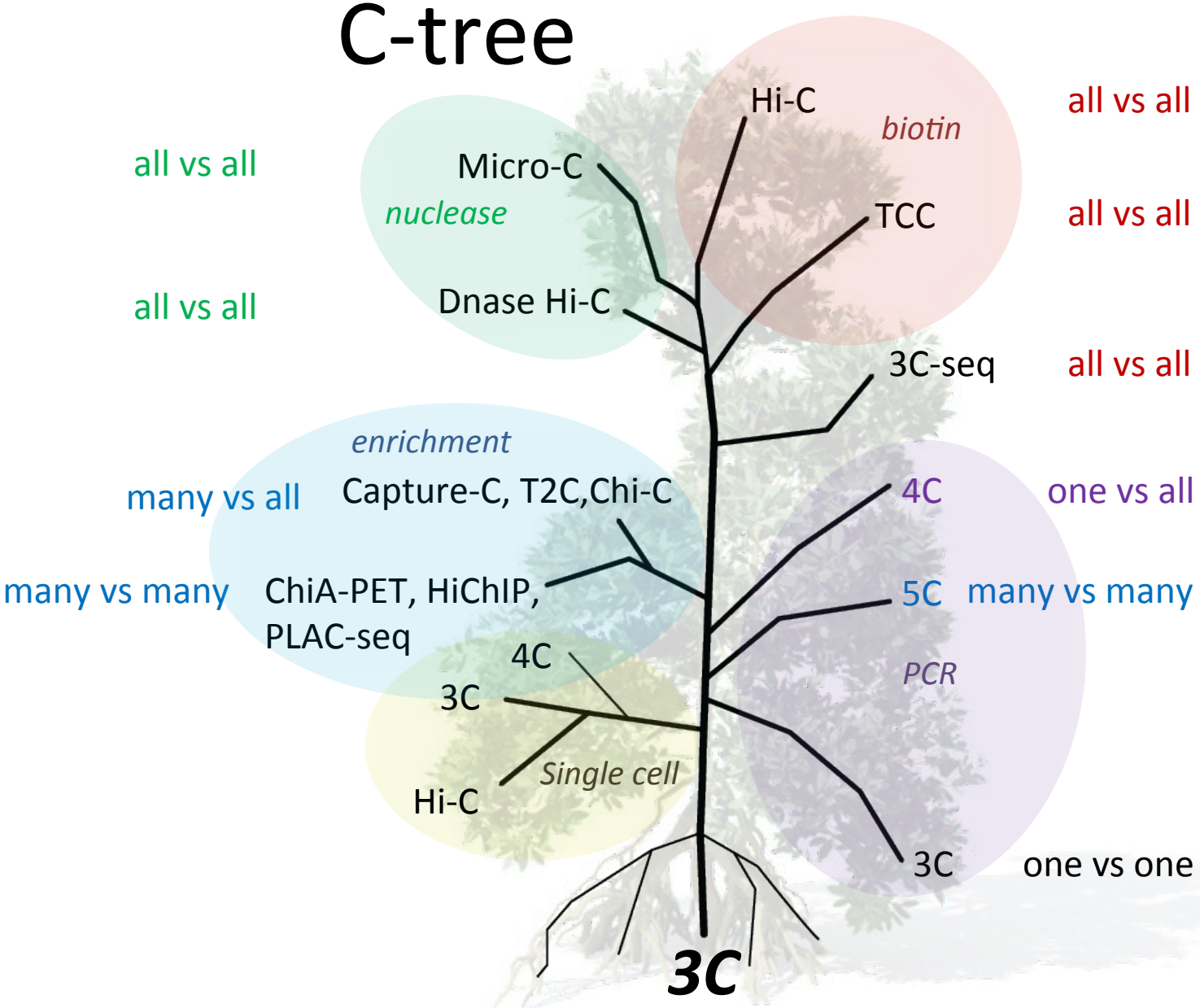
C-tree



FISH



C-tree



For references please refer to: Denker & De Laat, *Genes&Dev.* 2016

3C: Chromosome Conformation Capture

- Detects physical interactions between genomic elements
- Interacting elements are converted into ***ligation products***

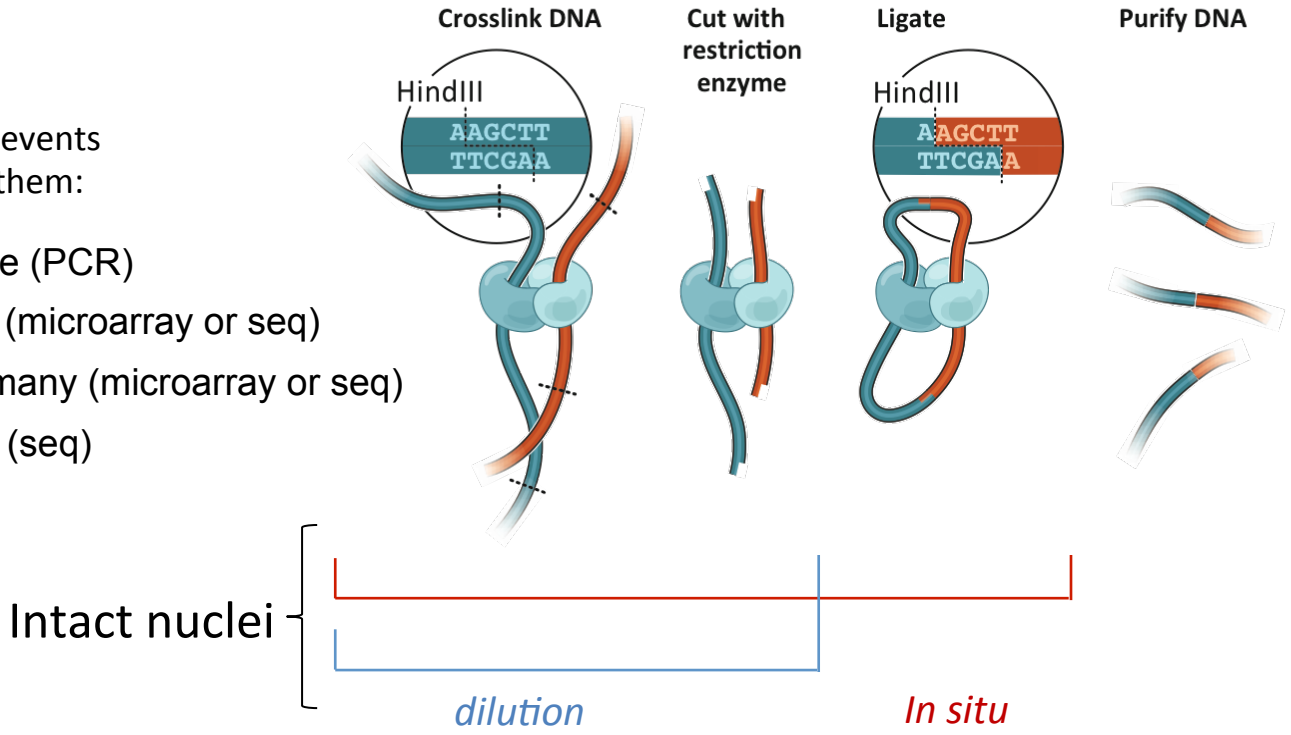
Detect ligation events
by probing for them:

3C: one by one (PCR)

4C: one by all (microarray or seq)

5C: many by many (microarray or seq)

Hi-C: all by all (seq)



Dekker et al. *Science* 2002
Dostie et al., *Genome Res*. 2006
Lieberman-Aiden, Van Berkum et al., *Science* 2009
Rao et al., *Cell* 2014

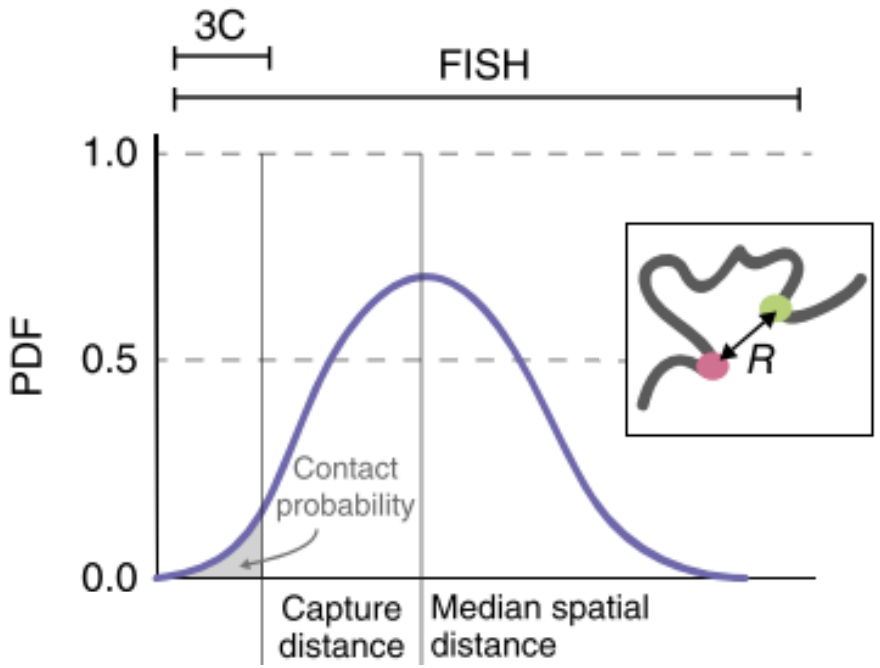
FISH or C?

FISH

- In general: Single cell
- **Spatial distance**
  - Any distance outside probe "glare"

Chromosome Conformation Capture

- In general: population
- **Contact frequency**
  - Capture radius dependent
  - Long distances in close proximity



Finn, *BioRxiv*, 2017
Belmont, *Curr.Opin.Cell Biol*. 2014
Giorgietti, *Gen.Biol* 2016
Fudenberg and Imakaev, *Nat.Methods* 2017

A more detailed look at what makes a good experiment

THE HI-C PROTOCOL

Belaghzal, *Methods* 2017

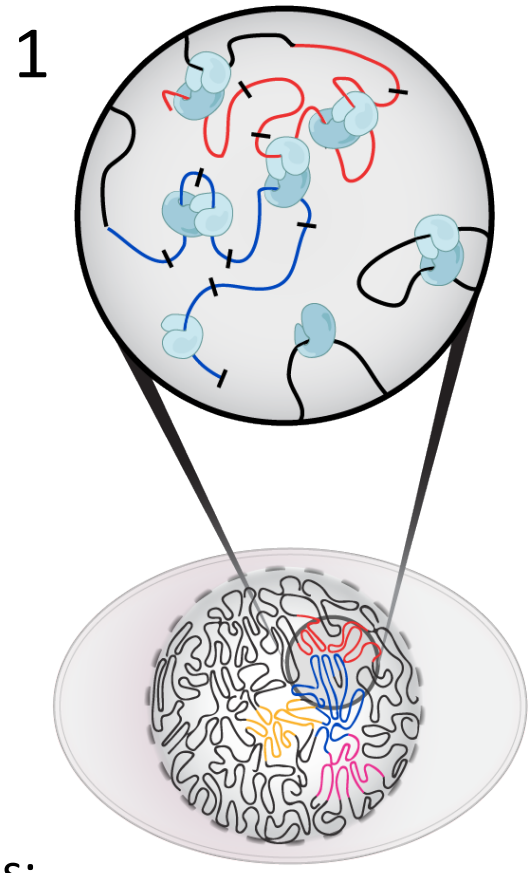
Goal of this presentation

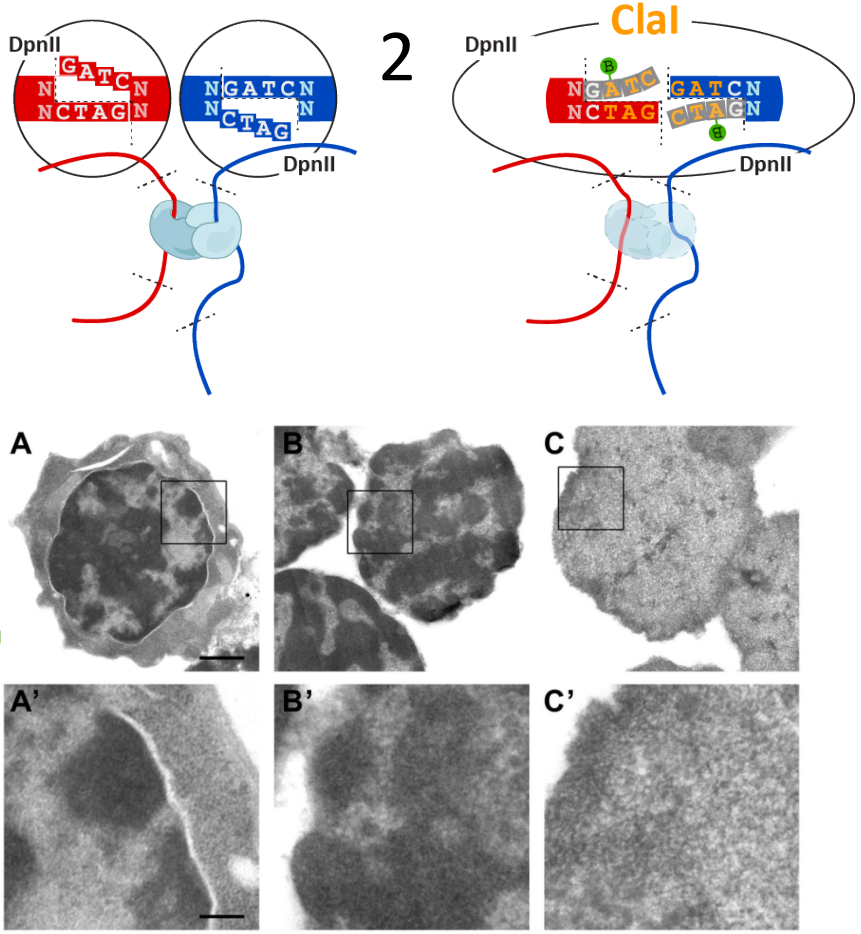
Understand how lab Hi-C procedures affect bioinformatic analyses

Some protocol publications on Hi-C

- Hi-C original: Lieberman-Aiden et al., *Science* 2010
- Hi-C 1.0: Belton-JM et al., *Methods* 2012
- In situ Hi-C: Rao et al., *Cell* 2014
- Single cell: Nagano et al., *Genome Biology* 2015
- Hi-C 2.0: Belaghzal et al., *Methods* 2017
- DLO-Hi-C Lin et al., *Nature Genetics* 2018
- Hi-C improving: Golloshi et al., *Methods* 2018
- Hi-C quality: Oddes et al., *BioRxiv* 2018
- Arima 1-day Hi-C: Ghurye et al., *BioRxiv* 2018

Hi-C: What we C





Critical steps:

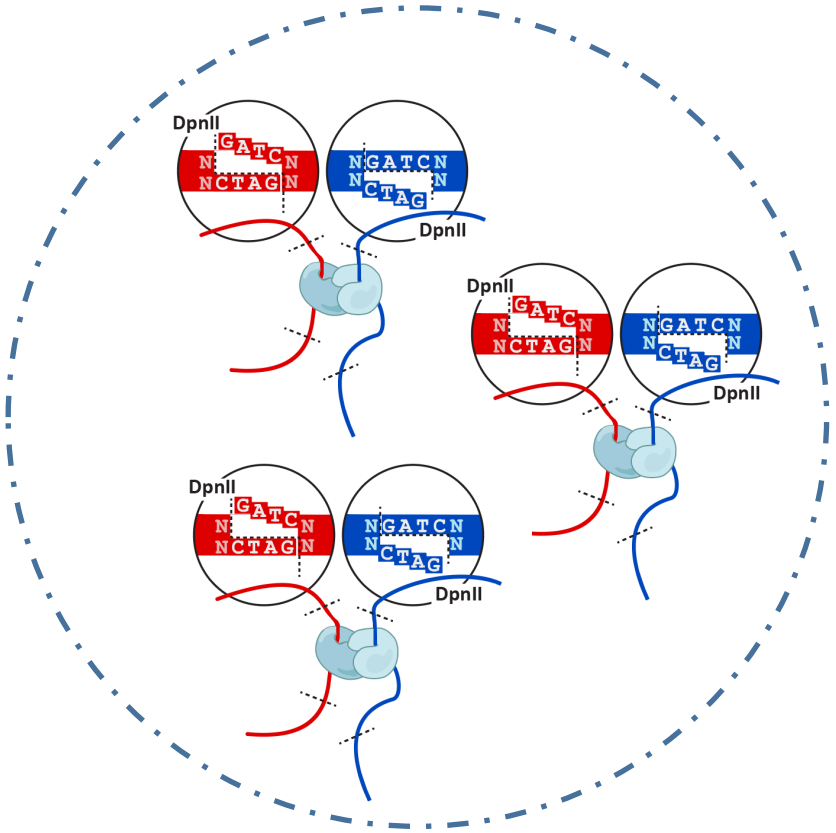
1. Crosslinking to fix conformation
2. Digestion and re-ligation
3. Sequencing (biotinylated) junctions

Lieberman-Aiden et al., *Science* 2009

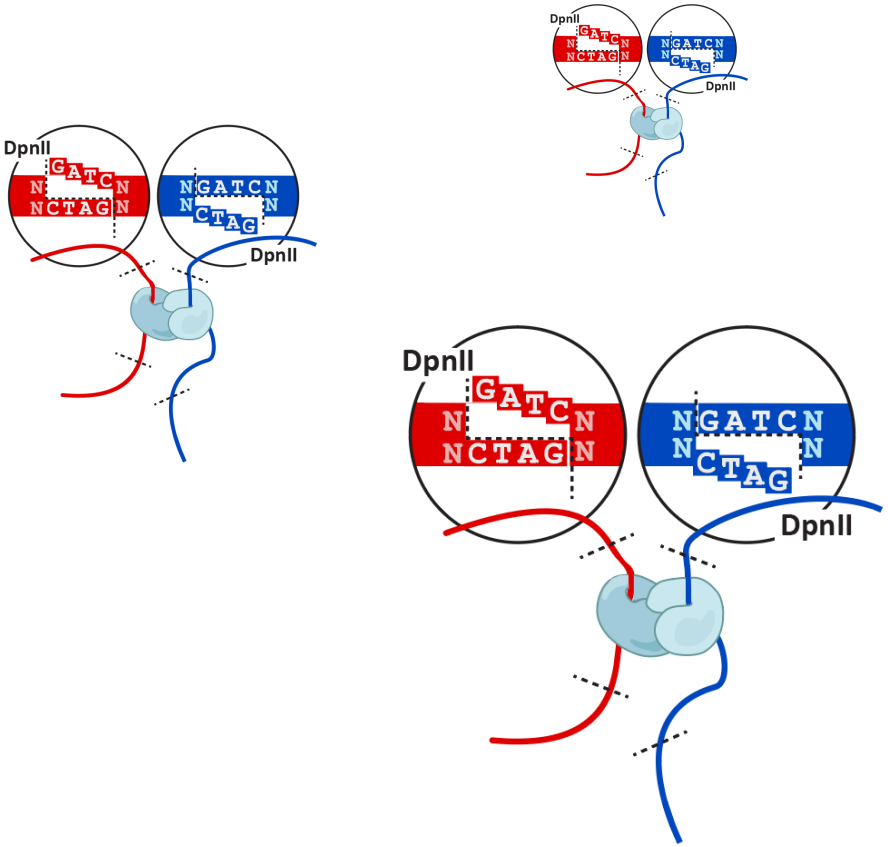
Belaghzal et al., *Methods* 2017 Gavrilov et al., *NAR* 2013

In situ Hi-C versus dilution Hi-C

- In situ Hi-C (2.0)



- Dilution Hi-C (1.0)



Belton-JM et al., *Methods* 2012
Belaghzal et al., *Methods* 2017

Hi-C critical steps

1. Fixation: keep DNA conformed
2. Digestion: enzyme frequency and penetration
3. Fill-in: biotin for junction enrichment
4. Ligation: freeze interactions in sequence
5. Biotin removal: junctions only!
6. Fragment size: small fragments sequence better
7. Adapter ligation: paired-end and indexing
8. PCR: create enough material for flow cell

1. Fixation

- Goal: Fix DNA in 3D position

- Crosslinking of DNA
  - 1% FA most common; up to 3% for yeast and bacteria
  - Freshness, dilution, serum and buffers affect results
  - Formaldehyde crosslinks mostly histones (lysines)
  - Additional crosslinkers can be used to bridge proteins specifically

- Suboptimal fixation
  - Lost interactions (cis)
  - Gained interactions (trans)

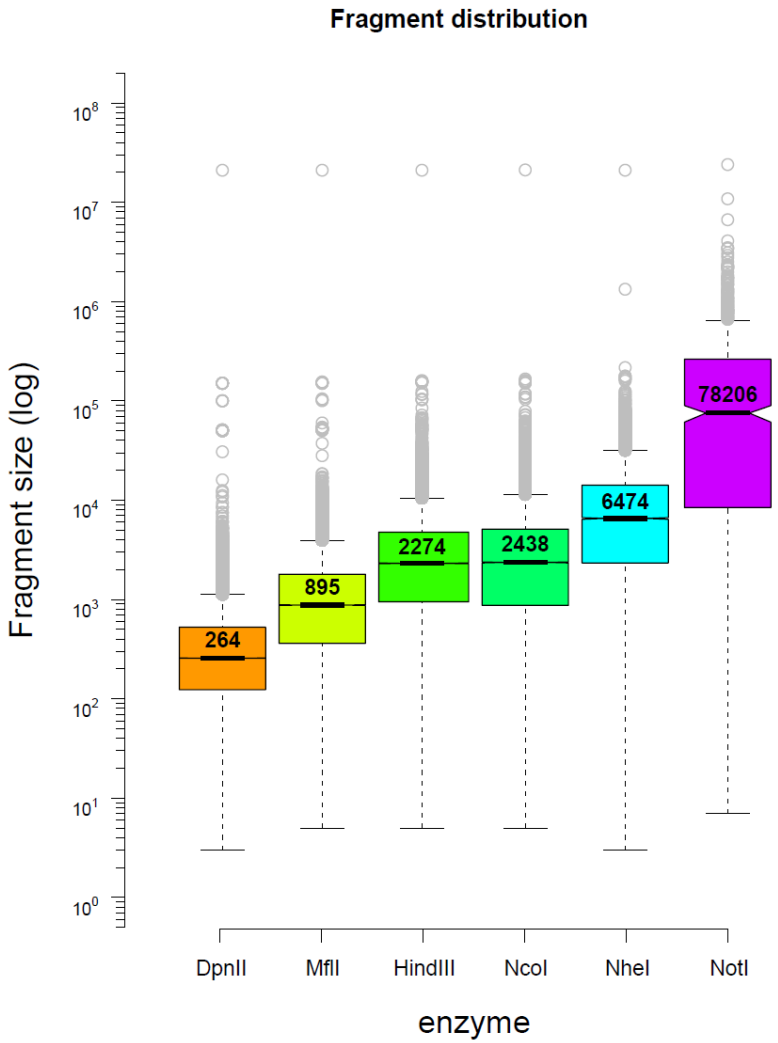
Brutlag et al., *Biochemistry* 1969
Lu et al., *J. Am. Chem. Soc* 2010
Hoffman et al., *JBC* 2015
Golloshi et al., *Mehods* 2018

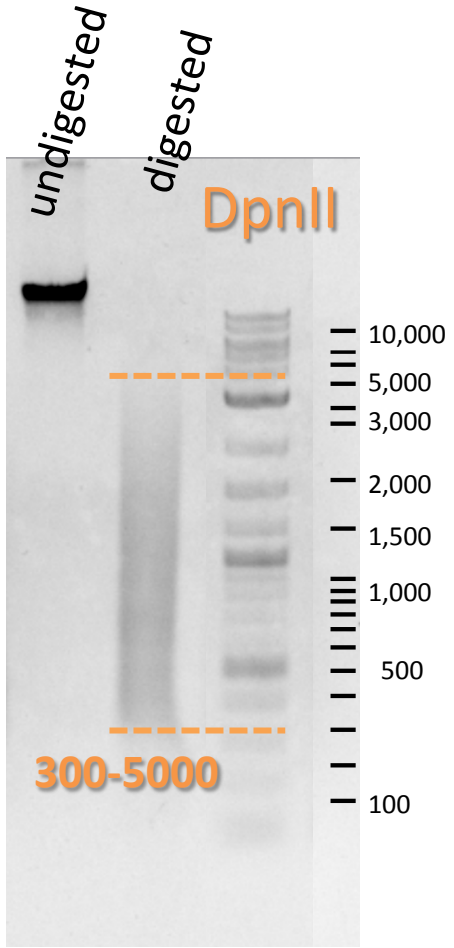
2. Digestion

- Enzyme choice
  - Frequency determines resolution
  - 4-cutters (DpnII/MboI) vs 6-cutters (HindIII, NcoI)
    - Cut more see more?

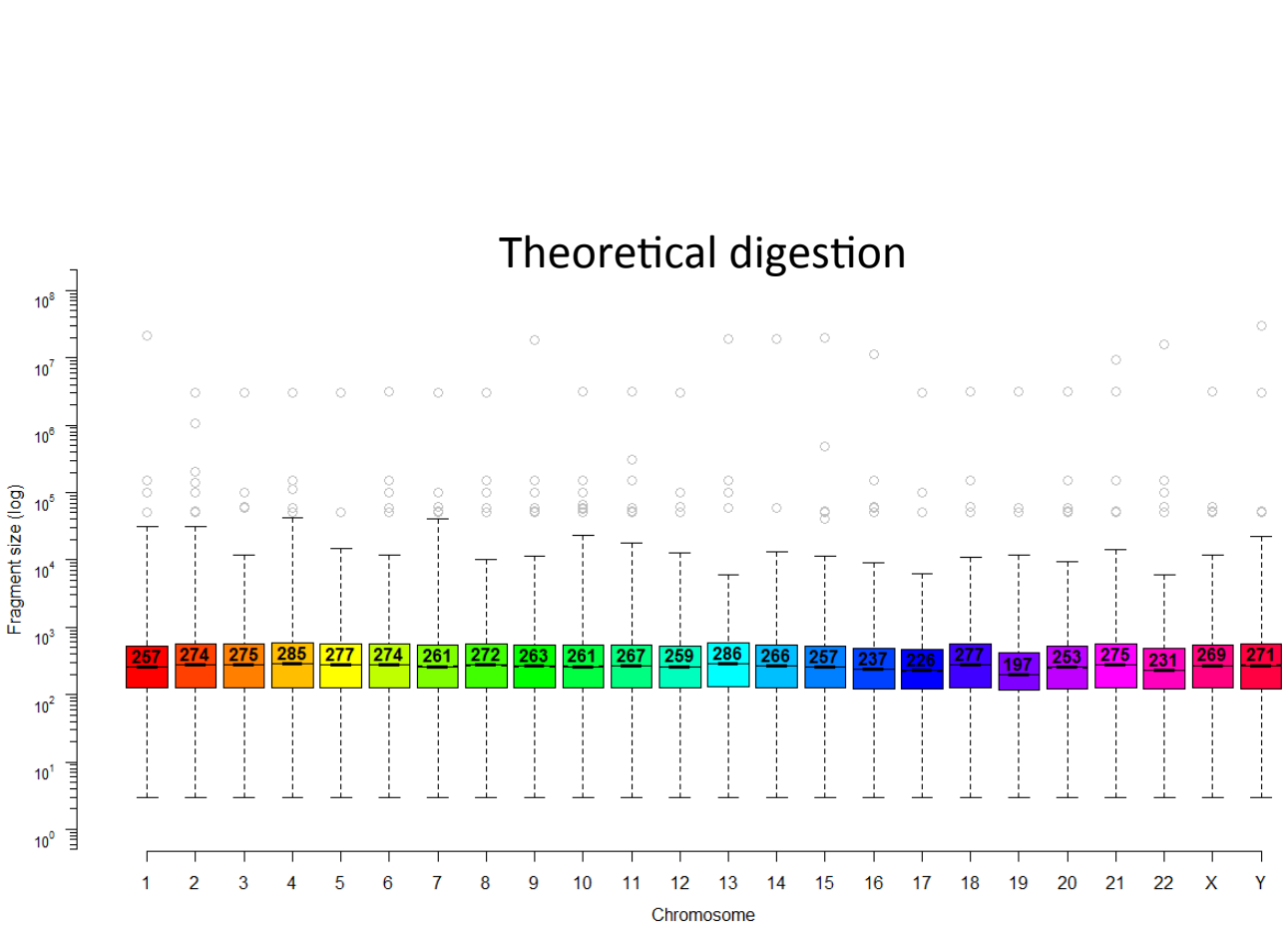
- Enzyme penetration
  - What is the percentage of actual digested sites?
  - What happens in partial digestion?
  - What about random breaks?

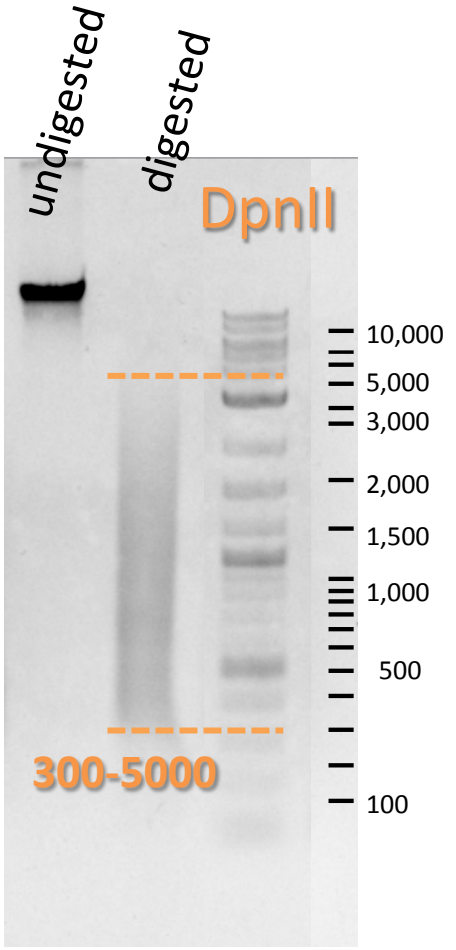
Theory vs actual digestion



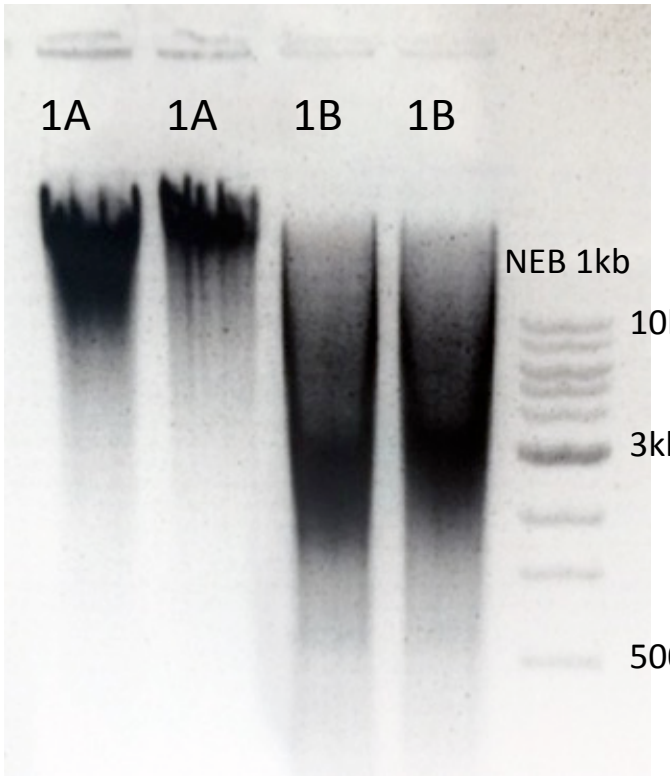


Actual DpnII digestion





Less digestion increases background

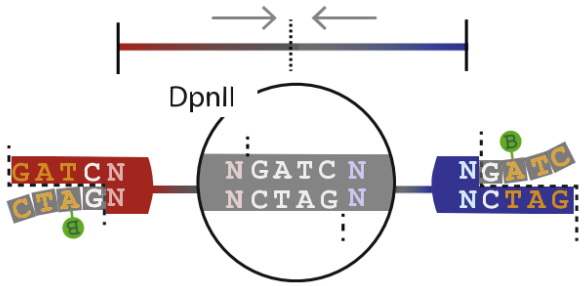


	Digestion gel appearance	Dangling end %	% Cis
Expt 1A	Still quite intact, high MW	69	18
Expt 1B	clearly digested	45	48
Expt 2A	More clearly digested	33	60
Expt 2B	less digested	43	33
Expt 2C	More clearly digested	22	61
Expt 3A*	Still quite intact, high MW	56	64
Expt 3B	More clearly digested	29	64

Golloshi et al., *Mehods* 2018

Partial digestion

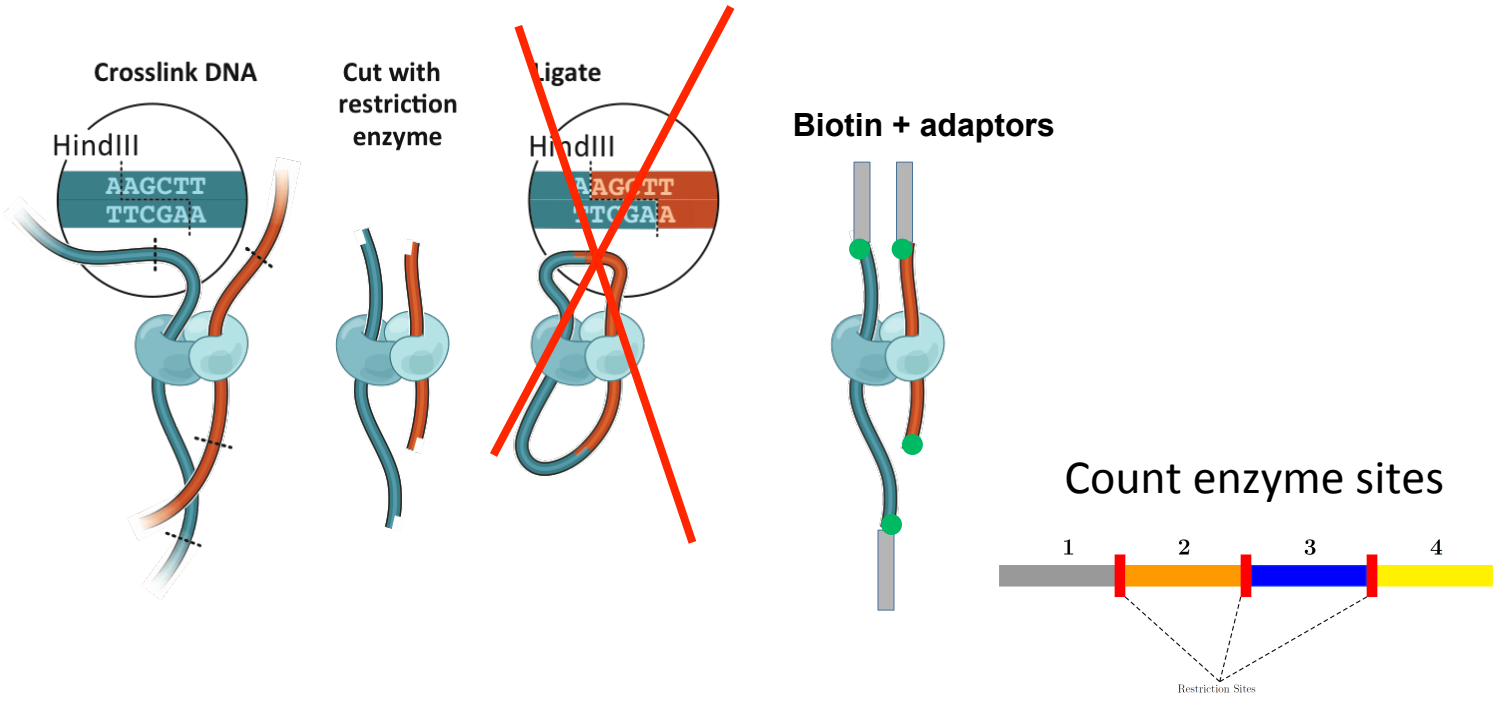
- A special case at close range
  - Undigested sites near digested, unligated sites
    - A significant portion of the anchor fragments is not cross-linked to any other restriction fragment



  - Presence of biotin and DpnII site → pulldown
  - Computationally indistinguishable from digested and re-ligated sites → remove in lab!

Gavrilov et al., *PLoS ONE* 2013
Belaghzal et al., *Mehods* 2017

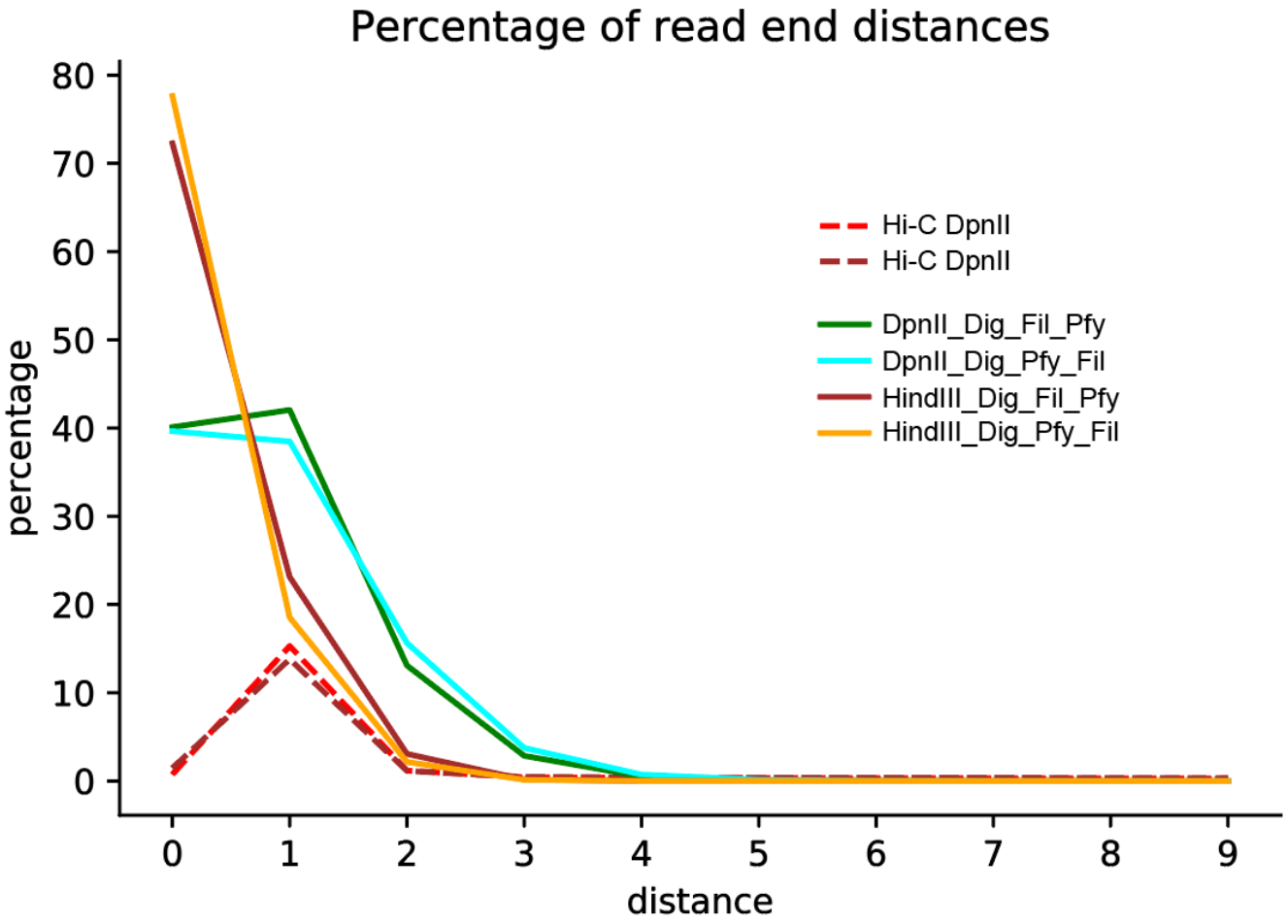
Assessing enzyme penetration



- Need for long reads to read through multiple restriction sites
  - 6-cutter detection less frequent than 4-cutters

Hakan Ozadam, Dekker lab

More sites = less penetration?



Hakan Ozadam, Dekker lab

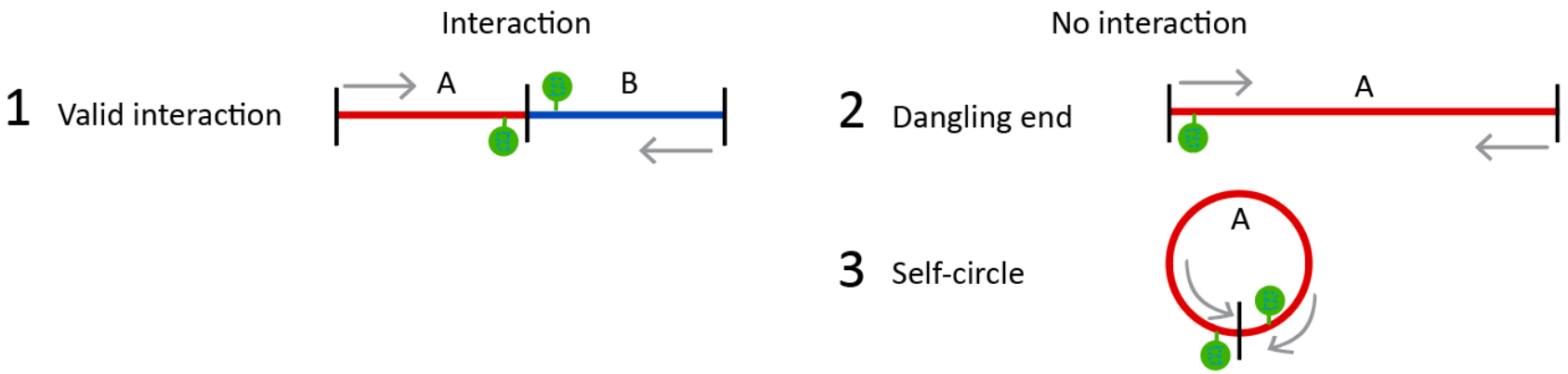
Random breaks

- Hi-C can be performed without addition of enzymes!
  - DNA isolation causes random breaks that will ligate to neighboring, broken DNA
  - DNA conformation still measured!

4. Ligation

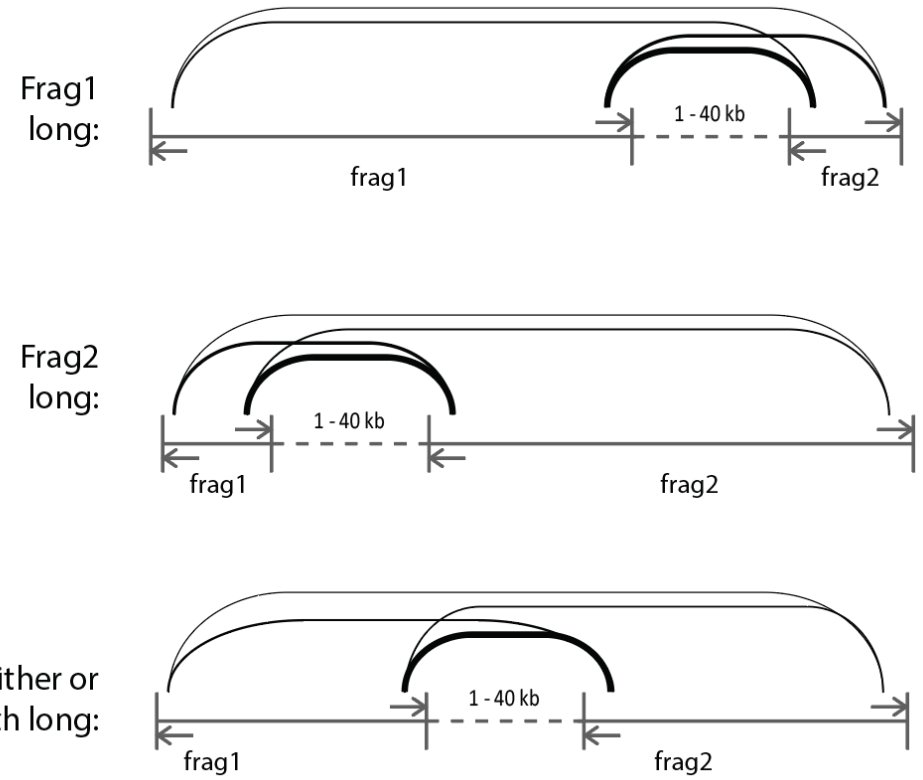
- T4 DNA ligase on blunted overhangs
- Until ligation is completed, fixation needs to stay intact:
    - avoid crosslink reversal from incubating at 65°C for too long

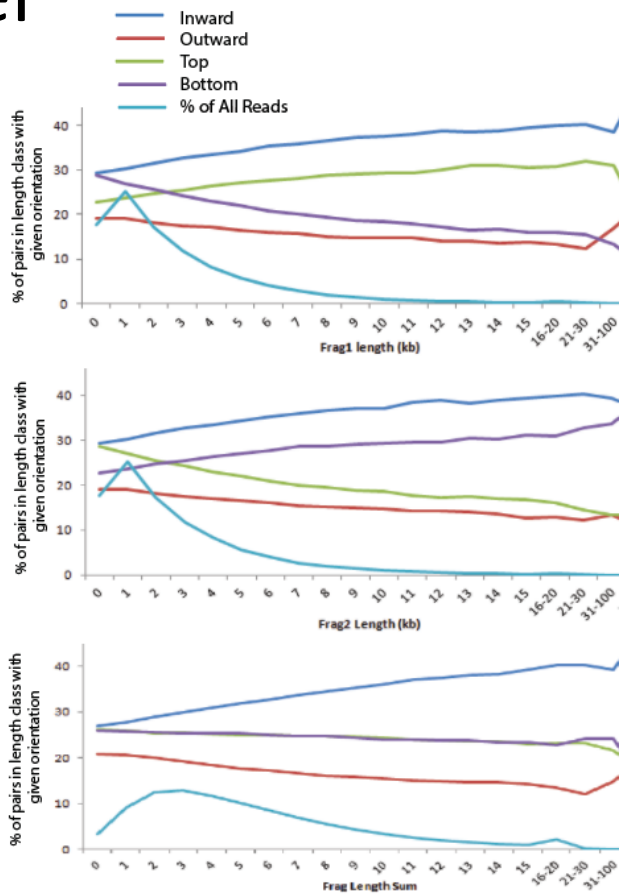
- Ligation relies on successful fill-in
- Ligation may affect cis/trans ratios (background)



Fragment size bias

- Longer fragments ligate better

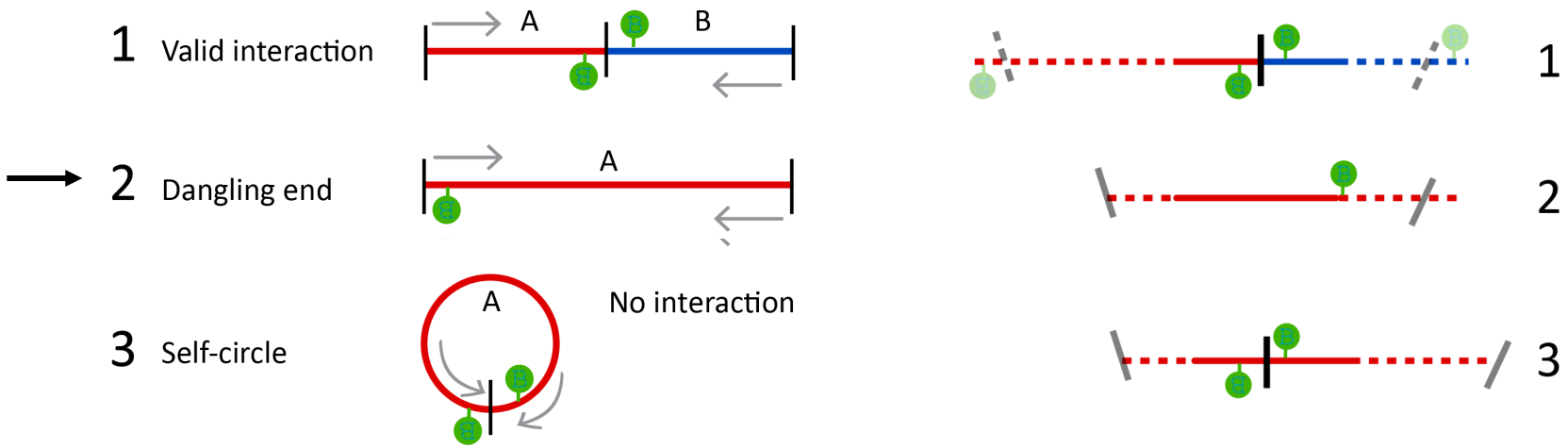




Goloshi et al., *Methods* 2018

5. Biotin removal

- Before sonication, biotins need to be removed from unligated (dangling) ends
  - Reaction uses exonuclease activity of Klenow
  - Sequencing unligated ends is costly!
  - After sonication, unligated ends look like true ligations



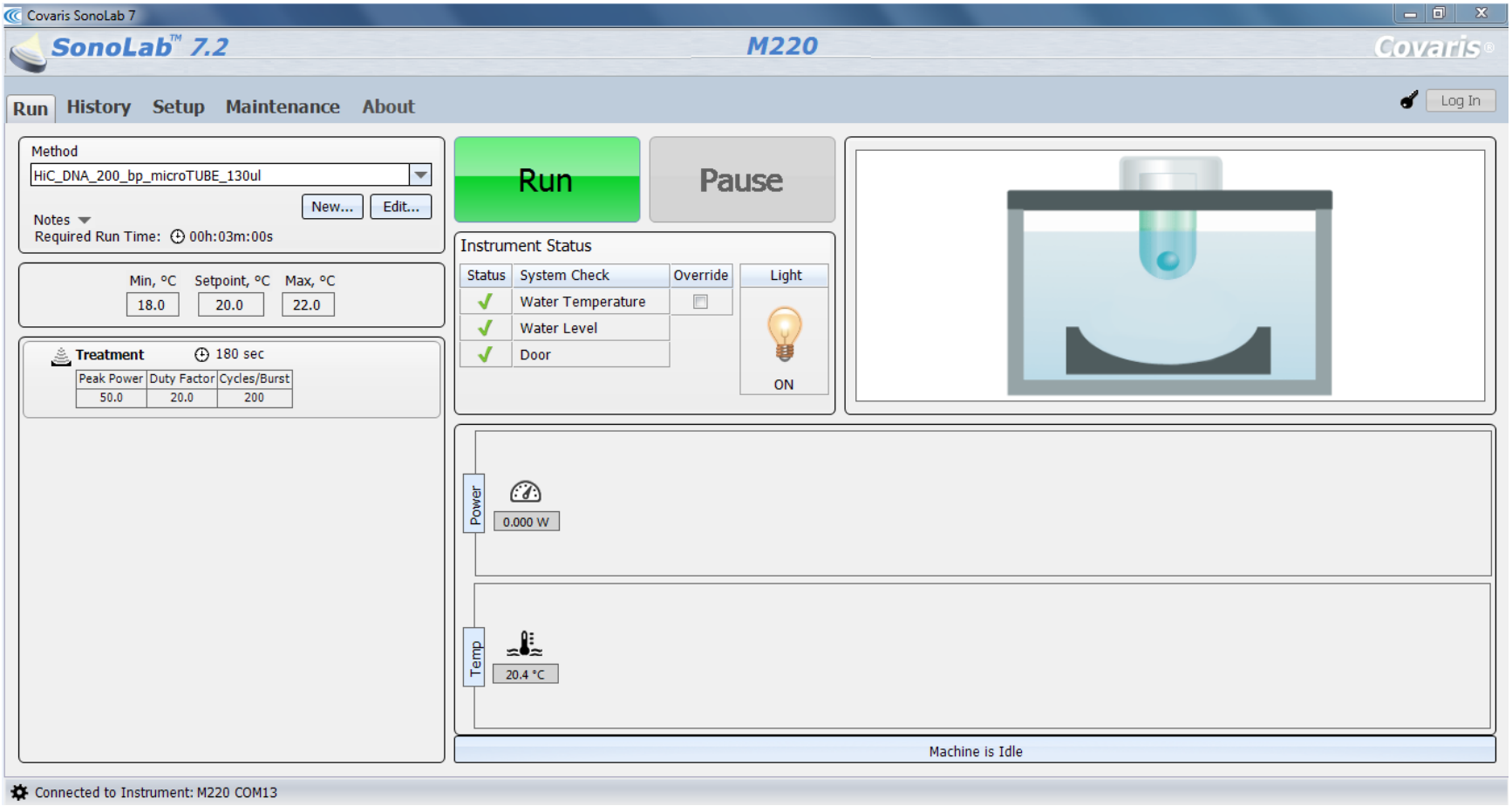
6. Fragment sizing

- Small molecules for illumina sequencing

- Fragmentation methods:
  - Sonication
  - Dnase
  - Restriction digestion

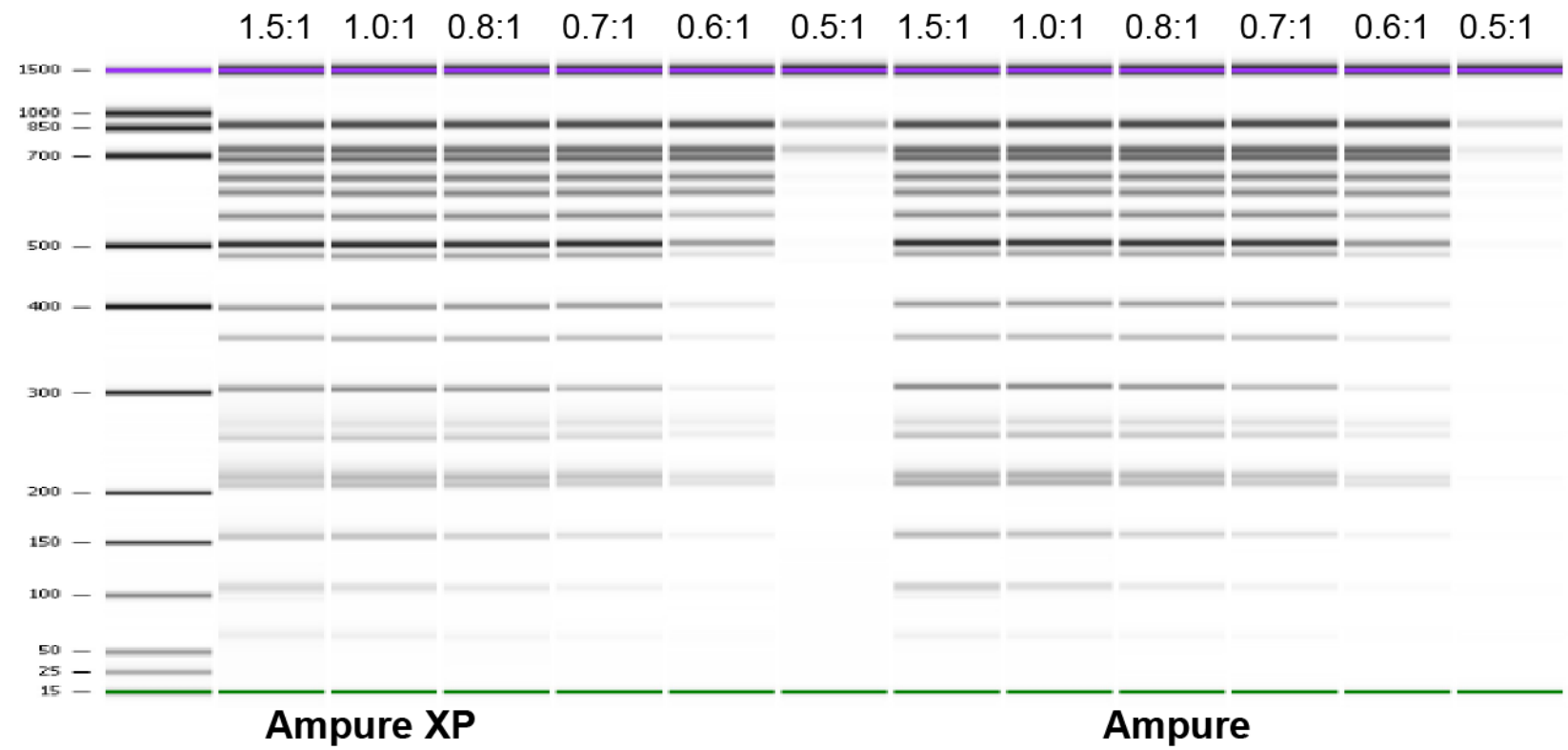
→ we use Covaris sonication + AMpure beads for tight fragment sizing

Covaris sonication



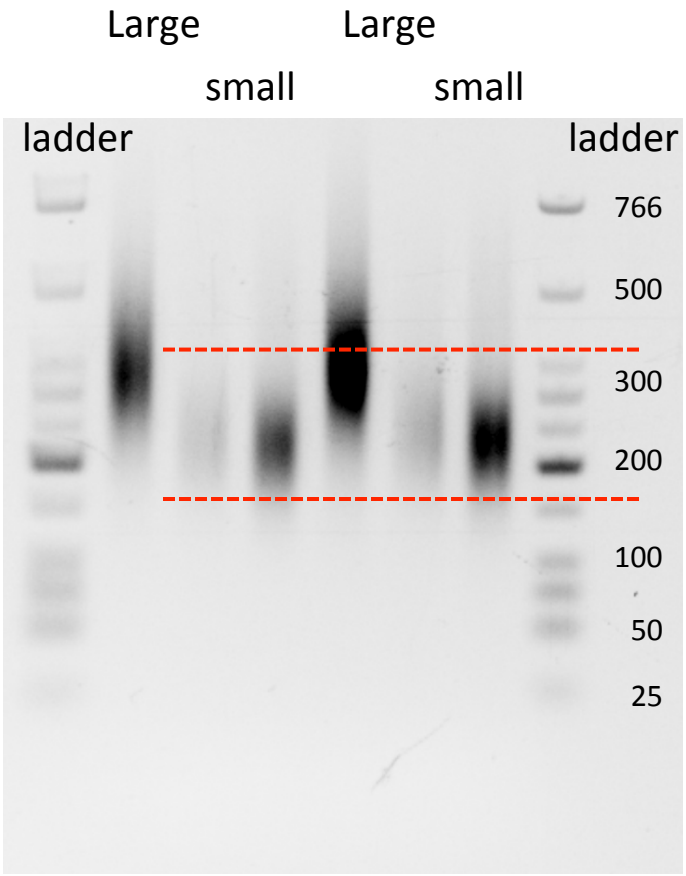
AMpure ratios

- Ratio of PEG8000 : DNA-containing solution
  - PEG expels DNA from solution



Covaris and Ampure

- Covaris to 200 bp fragments
- Use “AMpure ratios” to size select into tight large and small fractions
- Small fractions for adapter ligation

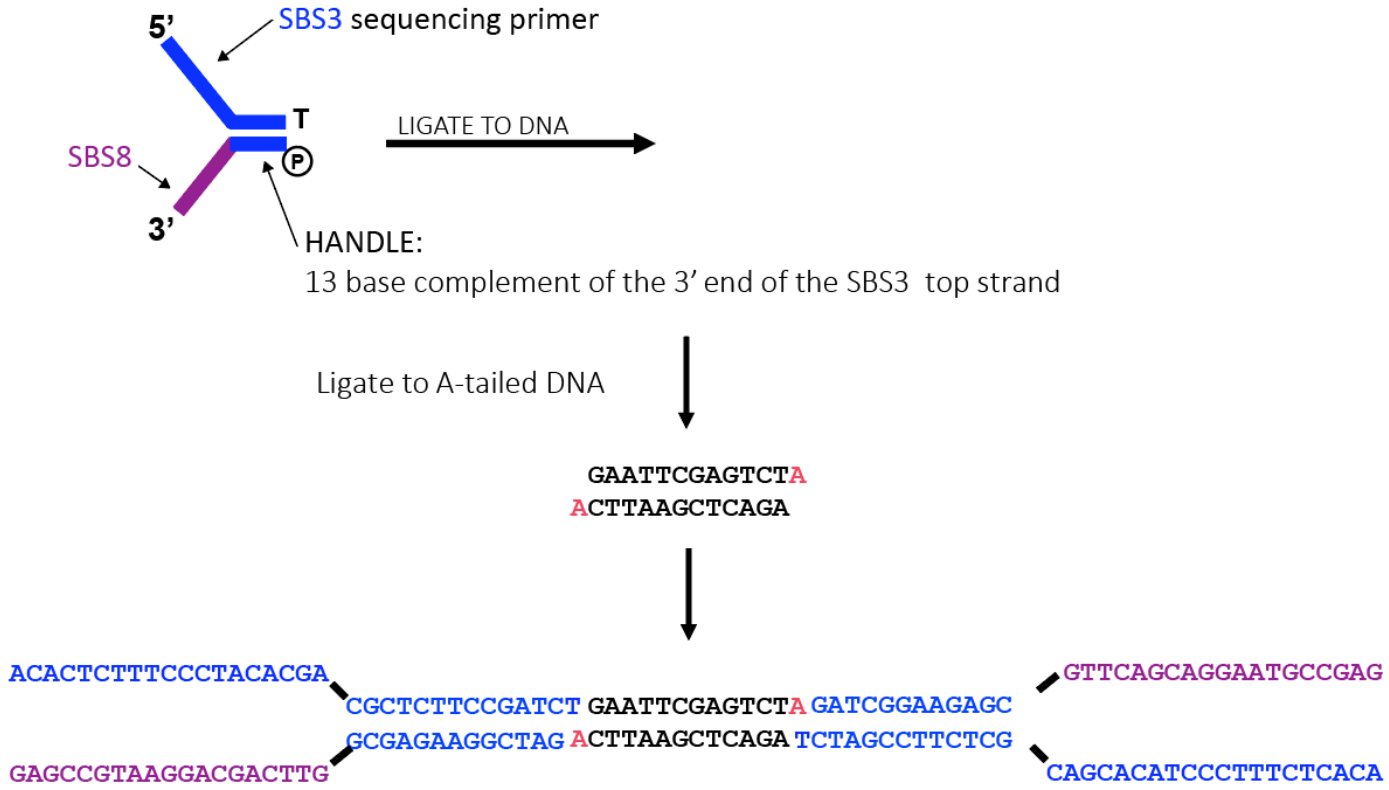


7. Adapter ligation

- **Paired-end adapters**
  - Stich sequence to both ends of interactions to allow for PCR amplification and sequencing
  - Adapters are build to fit on either side and allow PCR with different primers (PE 1.0 and PE 2.0)

- **Indexing**
  - Requires extra run cycle on sequencer (expensive)
  - Allows multiplexing to probe interactions
    - Multiple samples per lane
    - Lane reads divided over number of samples
  - Choose index sequence of minimal overlap in multiplex

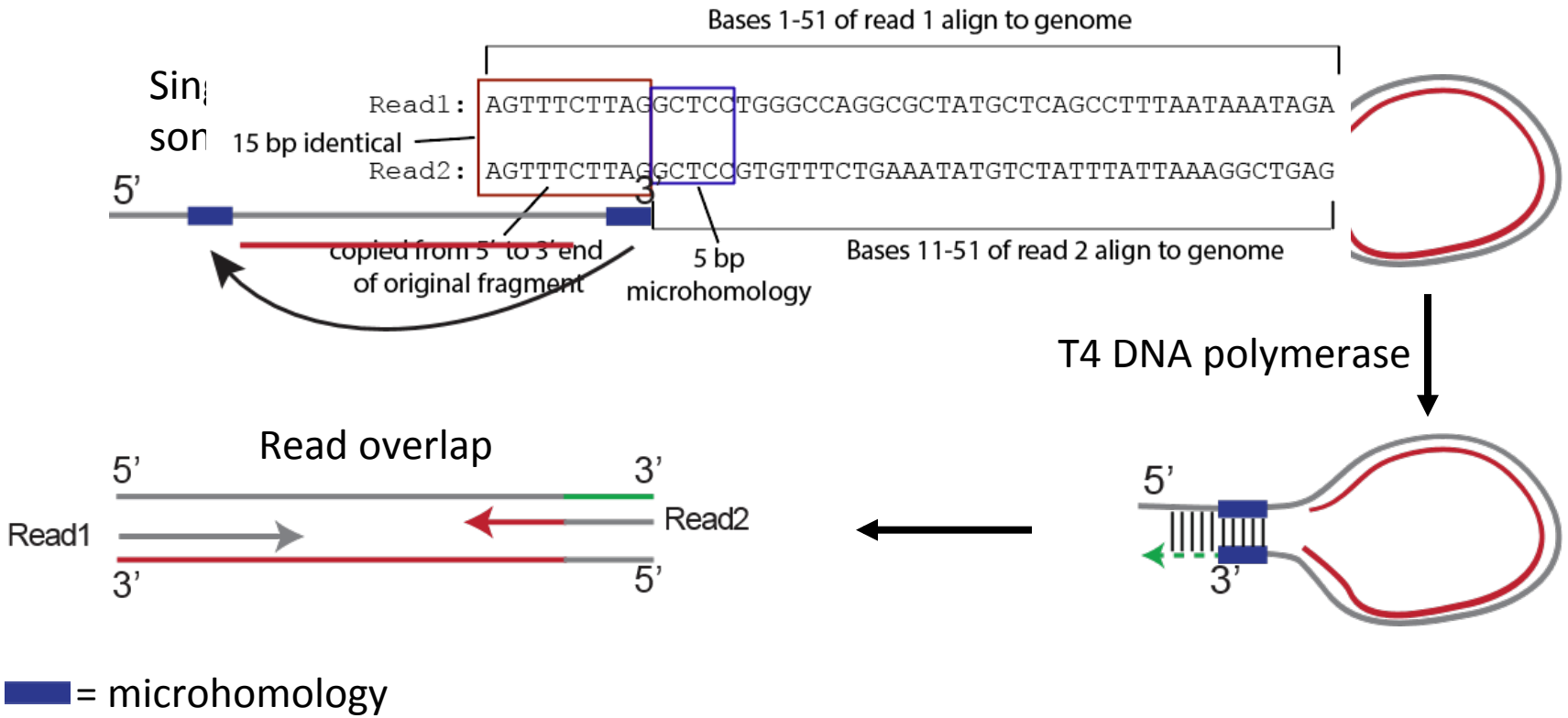
Paired end adapters



PCR w Primer 1.0 and Primer 2.0

Microhomology issues

- Adaptor homology causes sequencing issues



Goloshi et al., *Methods* 2018

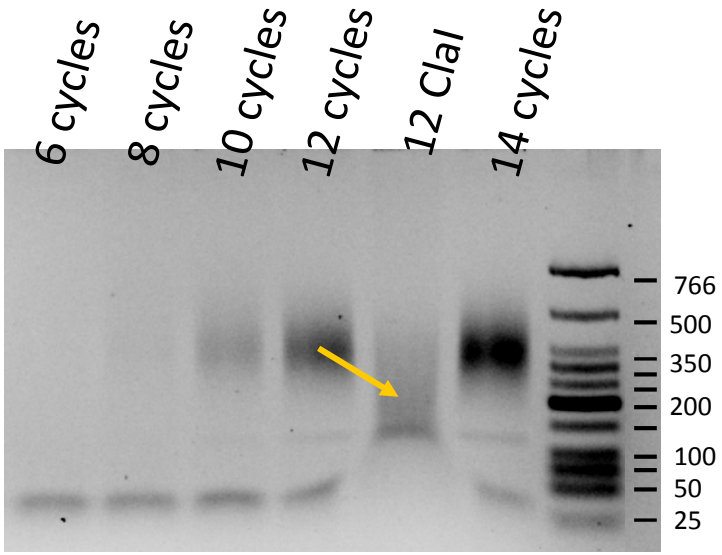
8. PCR amplification

- High fidelity enzymes
  - NEB Q5
  - PfuUltra High-Fidelity

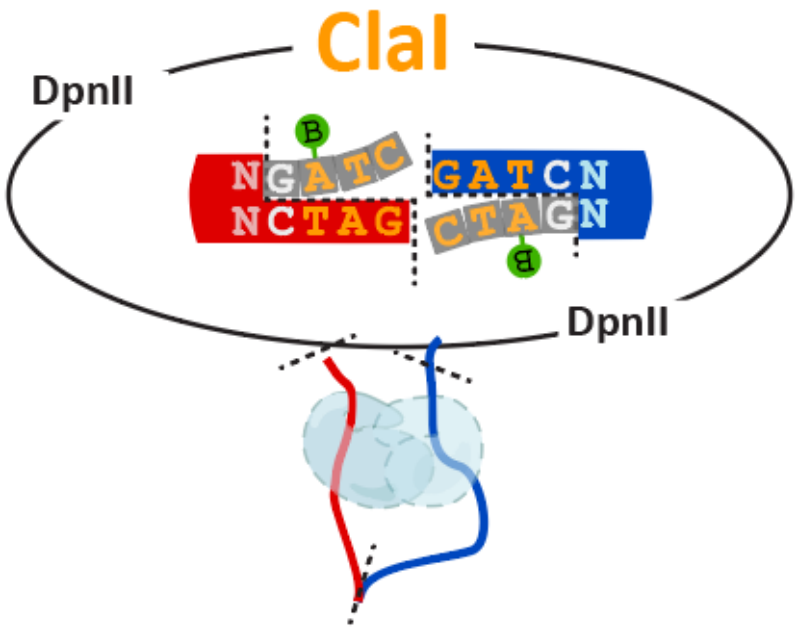
- PCR titration to identify the amount of PCR required

PCR cycle titration

Increasing cycles



Digest with ClaI for validity



What can go wrong?

- No fixation
  - Possible loss of interactions
- No complete (partial) digestion
  - Useless sequence reads between neighbors
- Ligation
  - No ligation:
    - no interactions, dangling ends or self-circles
  - Too many ligations:
    - increase in %trans interactions

How to assess quality from sequence

HI-C QUALITY AND QC

Belaghzal et al., *Methods* 2017
Lajoie et al., *Methods* 2015
Golloshi et al., Methods 2018
Oddes et al., *BioRxiv* 2018

Hi-C quality

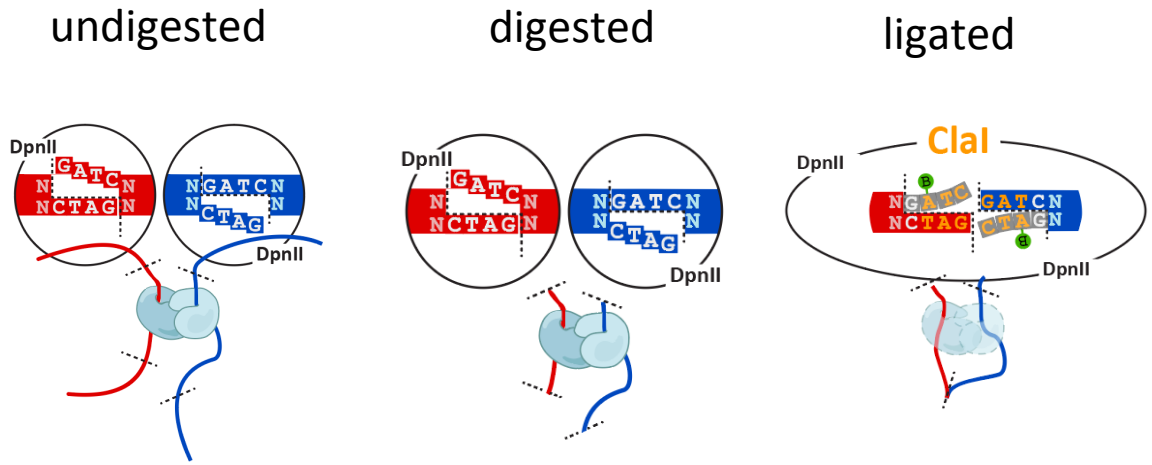
- What determines Hi-C quality?
  1. Cross linking
  2. Digestion (penetration, i.e. DpnII, HindIII)
  3. Ligation

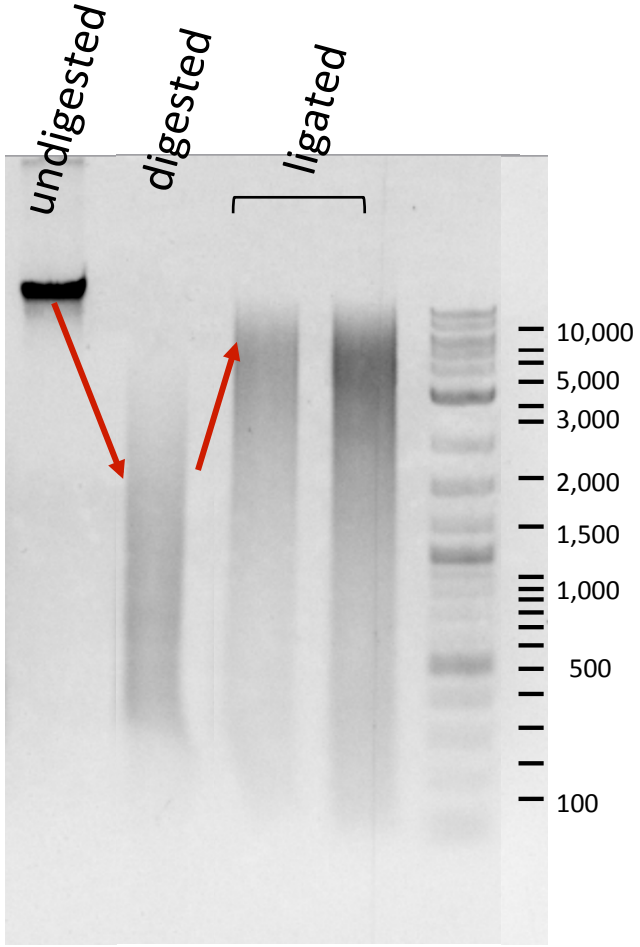
- What is the best quality metric?
  - Number of reads?
  - Cis/trans ratios?

- What is the best quality metric resolution?

Quality control in the lab

- Digestion and ligation

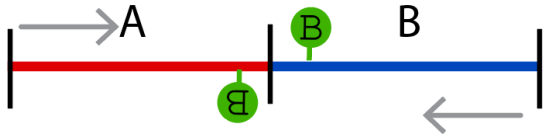


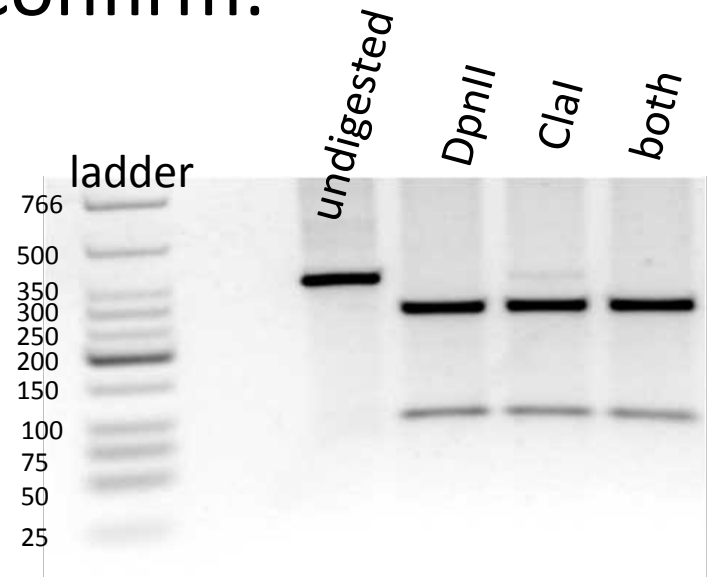


QC-PCR after ligation

- Probe a *specific* interaction from the pool
  - Neighboring fragments

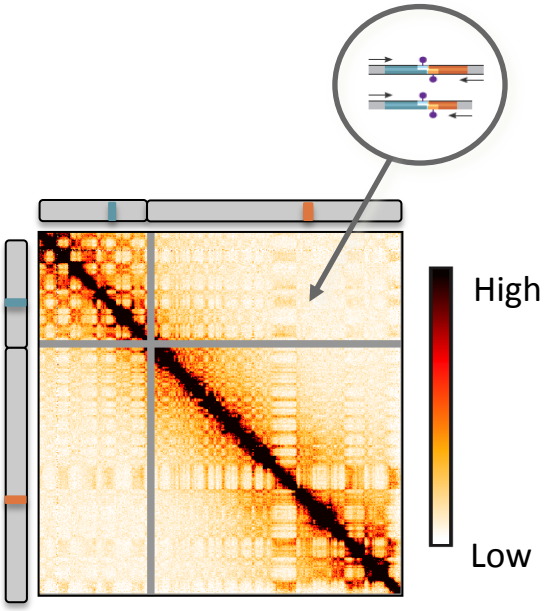
- Digest the PCR product to confirm:
  - Digestion
  - Biotin incorporation
  - Ligation





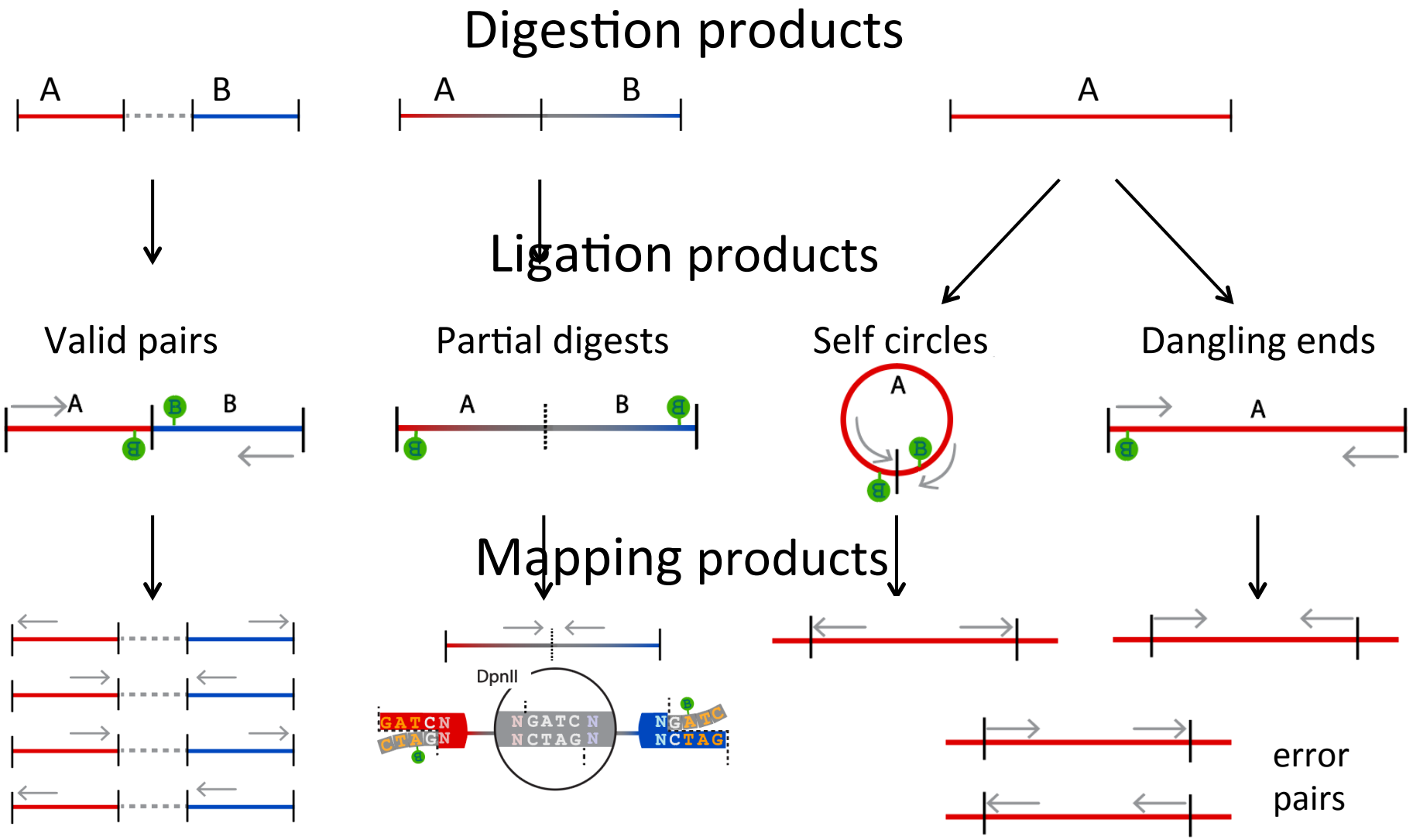
From bench to bites

- Sequence and then what?
- Put the ends back together
  - Map to a reference genome
  - Determine valid pairs
    - Read orientations
- Binning → Matrix
  - Sizes (resolution)

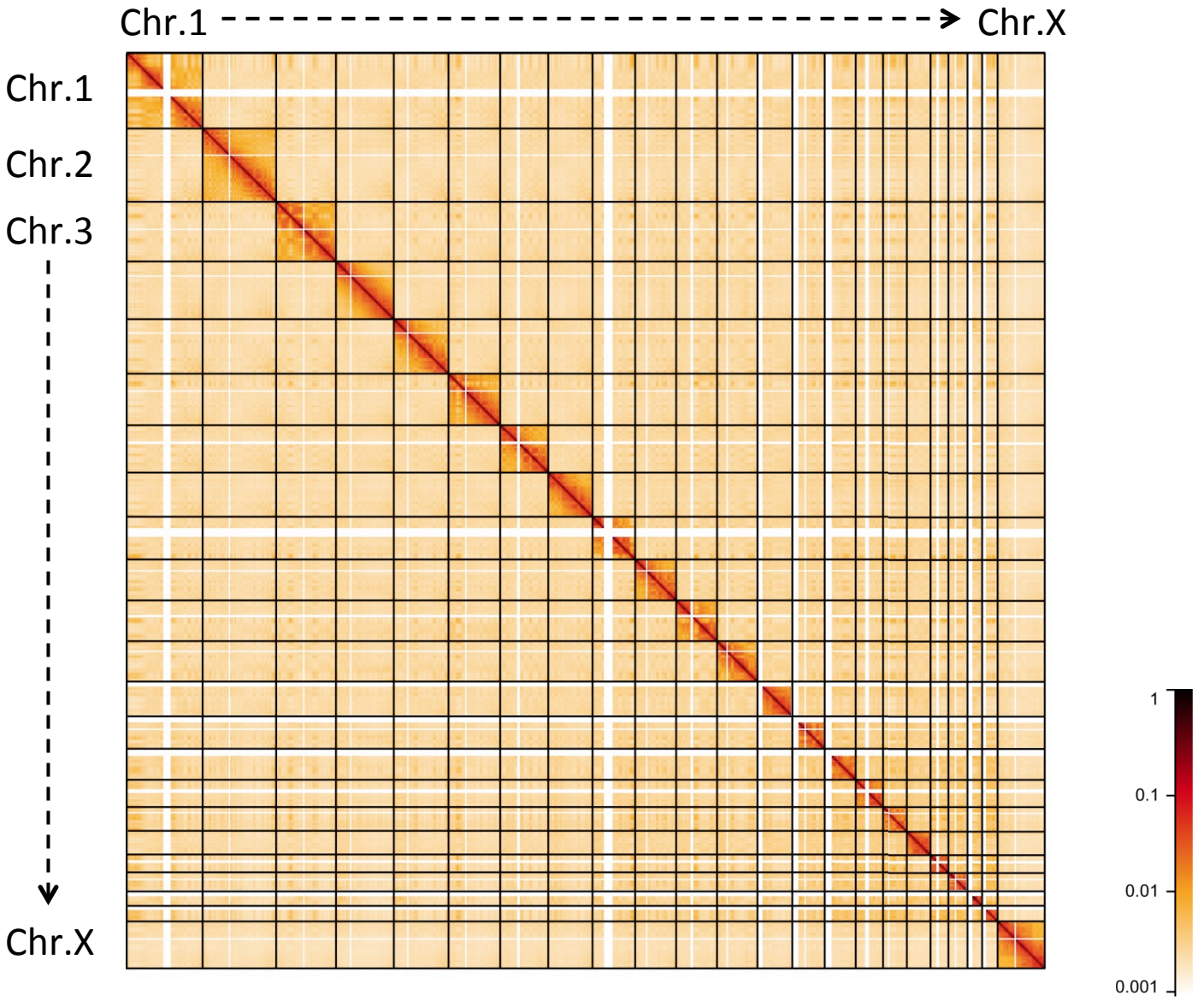


Unless single cell, ensemble average (over 10^7 cells)
Unless phased, averages over homologs.
Unless synchronized, averages over the cell cycle.

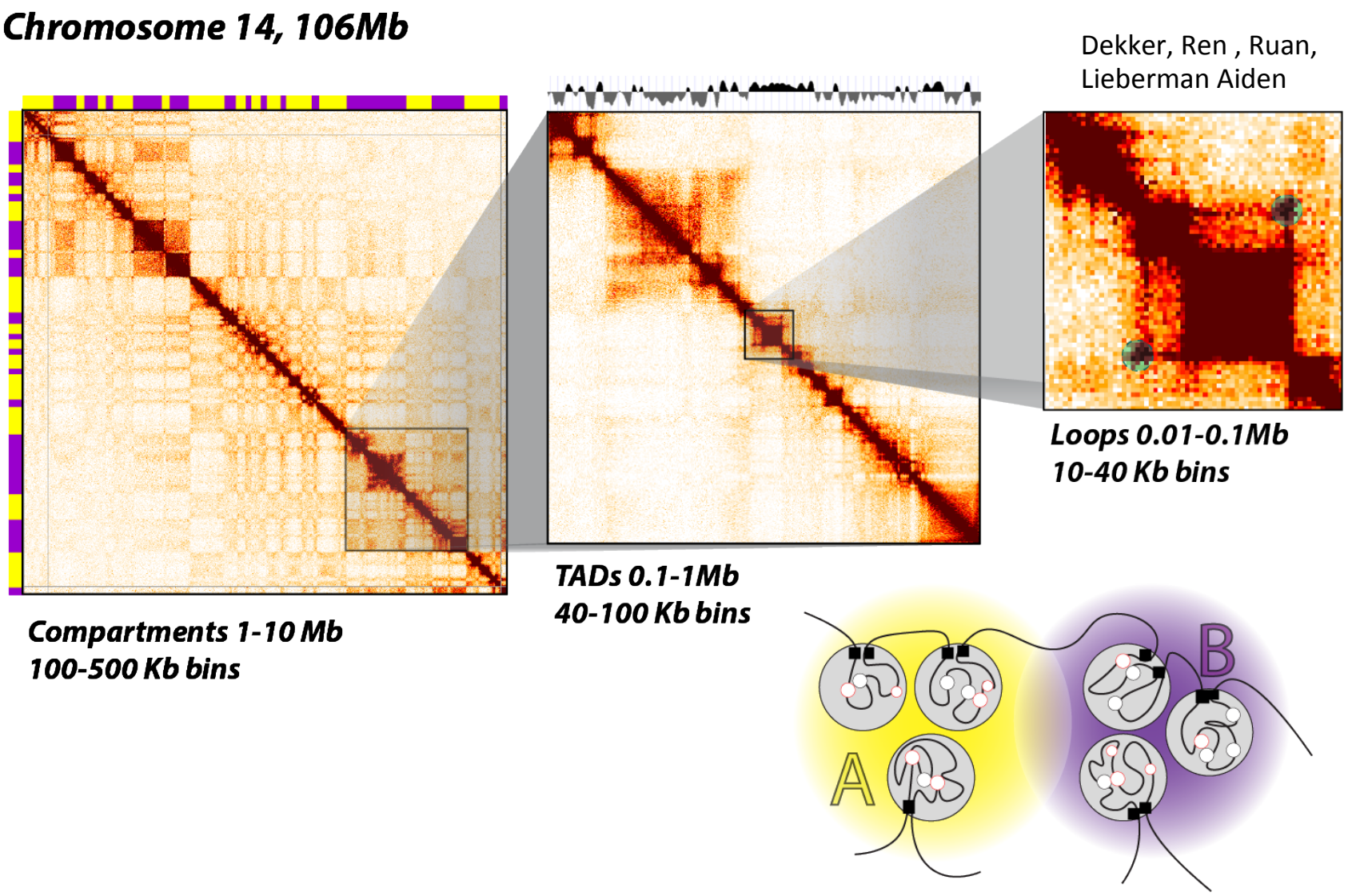
Interactions and product pairs



Hi-C Features



Scales of organization



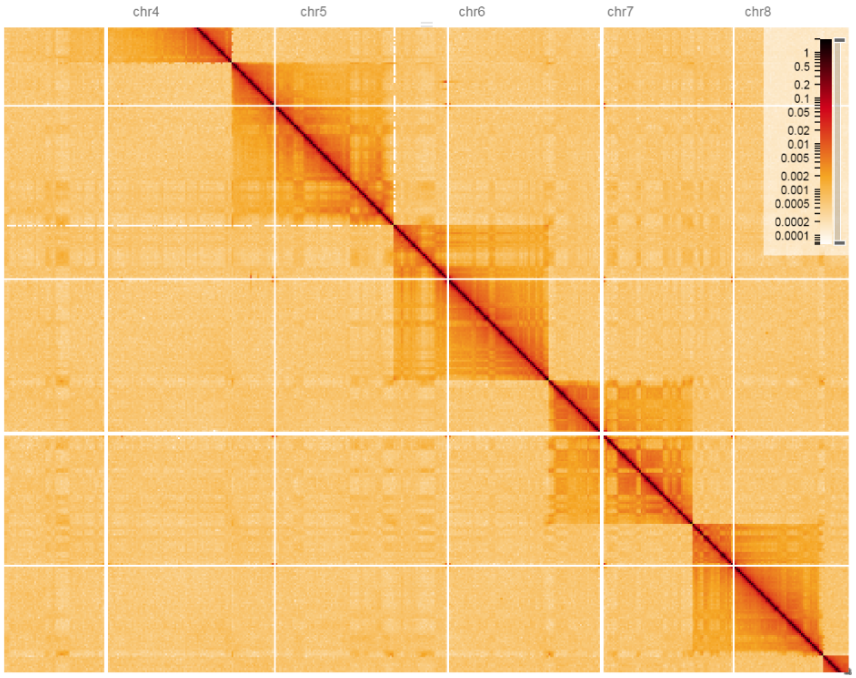
Data quality metrics

- Cis/trans ratios

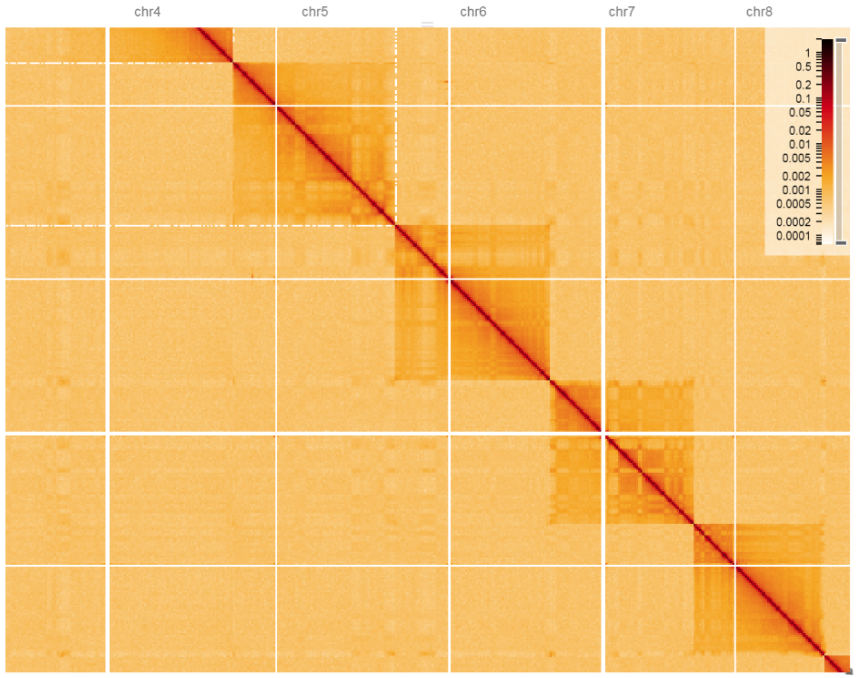
Belaghzal et al., *Methods* 2017
Lajoie et al., *Methods* 2015
Golloshi et al., Methods 2018
Oddes et al., *BioRxiv* 2018

Quality at chromosome level

High %cis (127x10^6 NRVP)



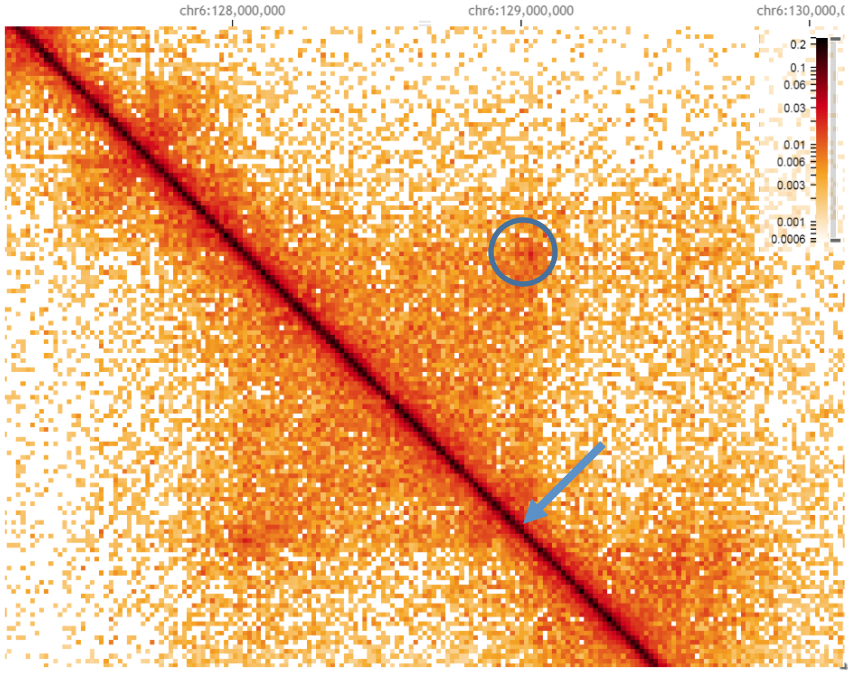
Low %cis (143x10^6 NRVP)



- More signal within chromosomes
- More signal near diagonal

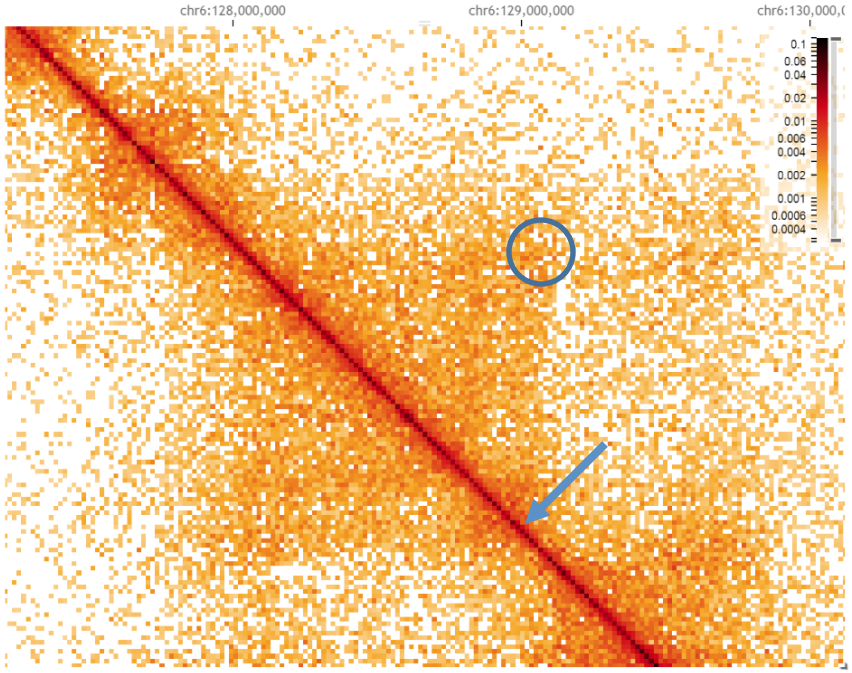
Quality at TAD level

High %cis ($127x10^6$ NRVP)



- More loops
- More at diagonal

Low %cis ($143x10^6$ NRVP)



- Less loops
- Less at diagonal

Data quality metrics

- Cis/trans ratios

- Invariant Hi-C interaction patterns
  1. Intra-chromosomal interaction enrichment
  2. Distance-dependent interaction decay
  3. Local interaction smoothness

Belaghzal et al., *Methods* 2017
Lajoie et al., *Methods* 2015
Golloshi et al., Methods 2018
Oddes et al., *BioRxiv* 2018

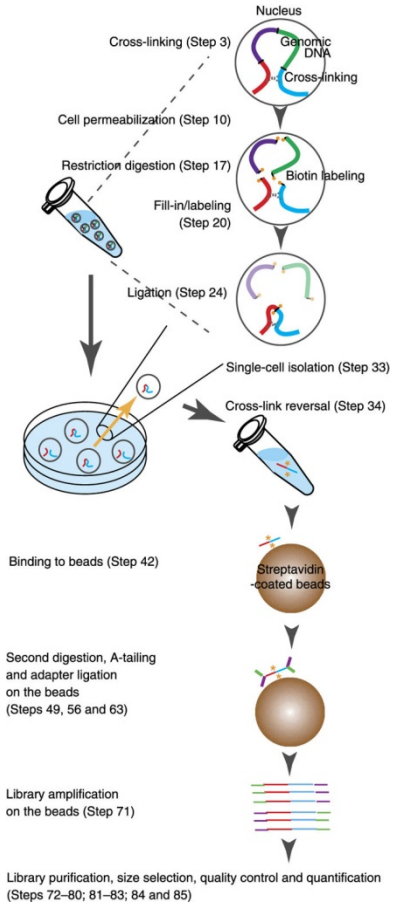
Less to see more!

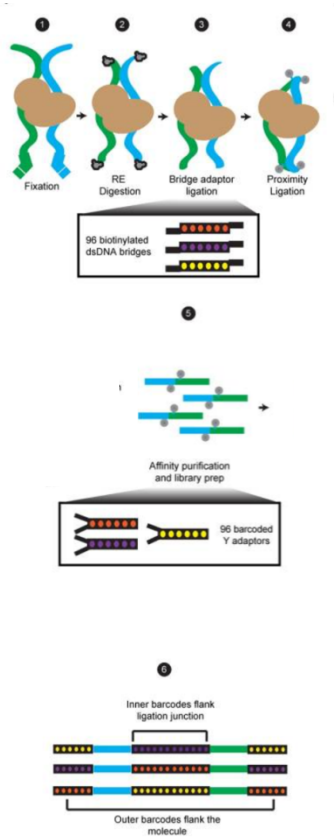
SINGLE CELL AND ENRICHMENT

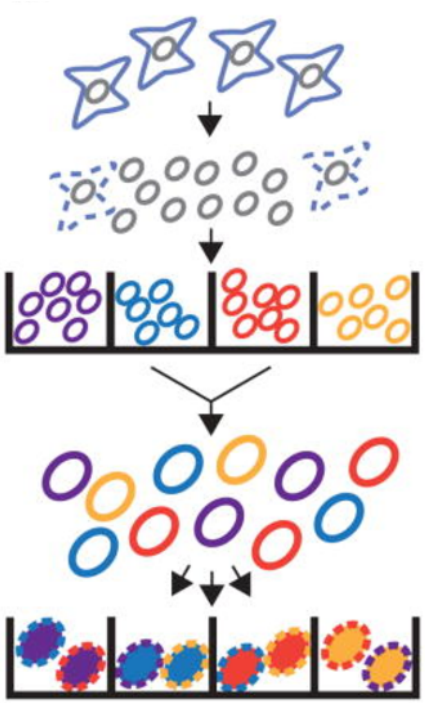
Notes on conformation

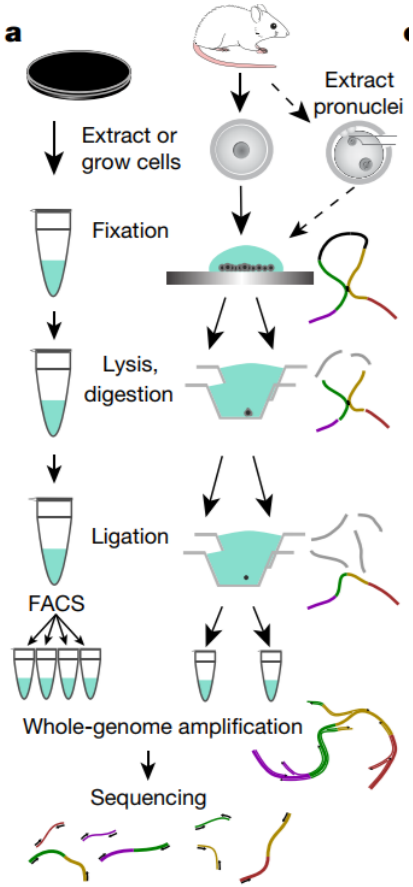
- Epigenetics are more subtle than genetics
  - Capturing epigenetic states (folding) is prone to variation
  - Every cell has the same genome, but cell types vary epigenetically (i.e. fold differently)
  - Single Cell Hi-C capture is cell specific
    - Limited information?
  - Enrichment can better visualize subtle details

Single cell Hi-C









Tjong et al., *PNAS* 2016
Cusanovich et al., *Science* 2015
Nagano et al., *Nat.Protoc* 2015
Ramani et al., *Nat Meth*. 2017
Flyamer et al, *Nature* 2017

Different ways to "enrich"

- The classical 5C
  - Choose a specific region for PCR probe design
- Capture Hi-C
  - Use DNA sequence of a specific region to enrich C-libraries by probe hybridization
- HiChIP/ChIA-PET
  - Enrichment by antibody based selection of proteins.
    - Note that normalization (input) is hard!

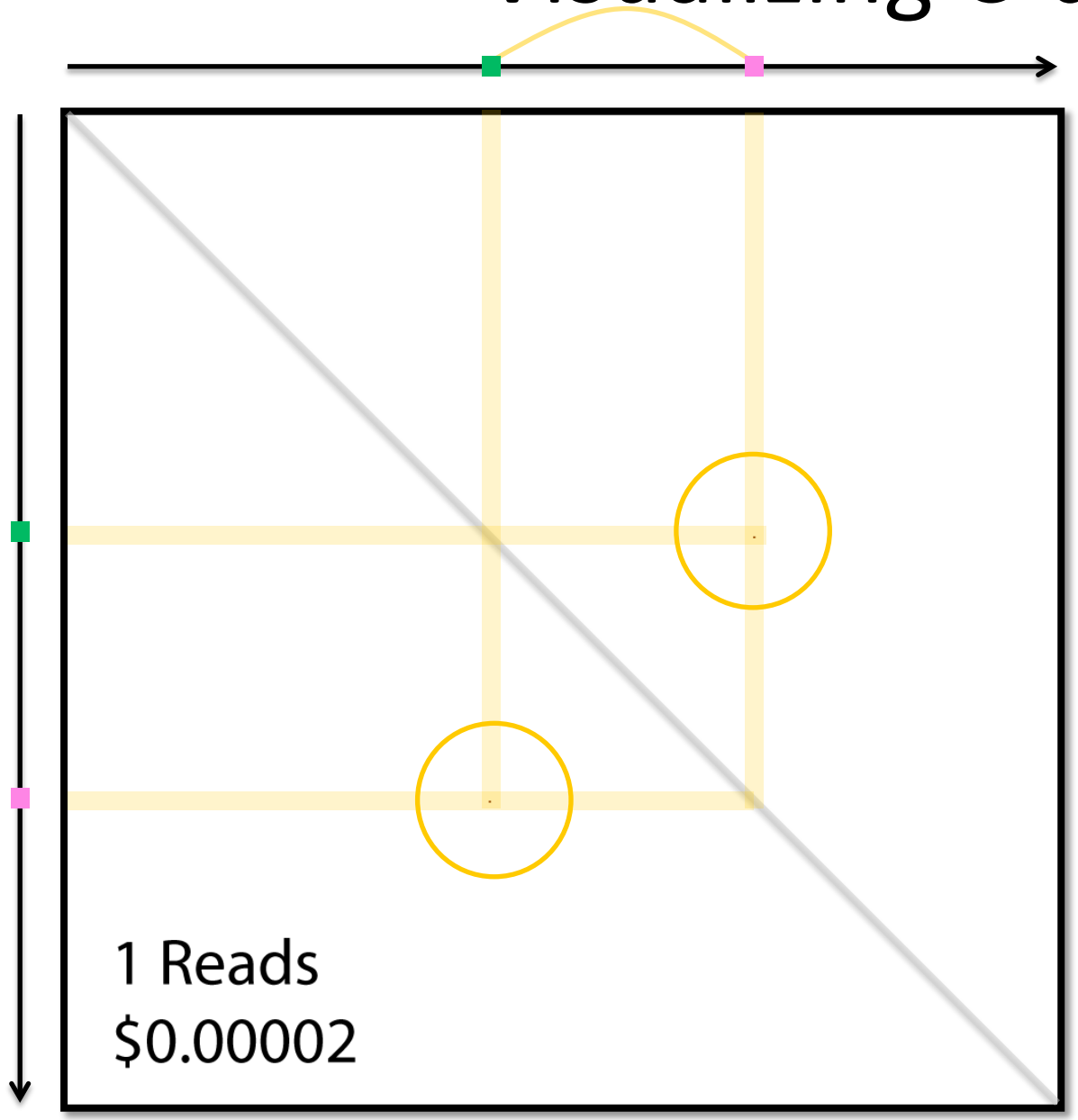
Dosti, *Gen.Res* 2006
Misfud, *Nat Genet*, 2015
Mumbach, *Nat.Methods* 2016
Li, *Cell* 2012

Hi-C protocol choice

- Chose towards your visualization goal
  - High resolution requires frequent cutters and more reads
  - Compartments and TADs can be seen without high-resolution (and this is cheaper!)

- Anticipate in the lab what you will struggle with computationally
  - Bad data is worse than no data
  - Some qualities can not be filtered out computationally

Visualizing C-data



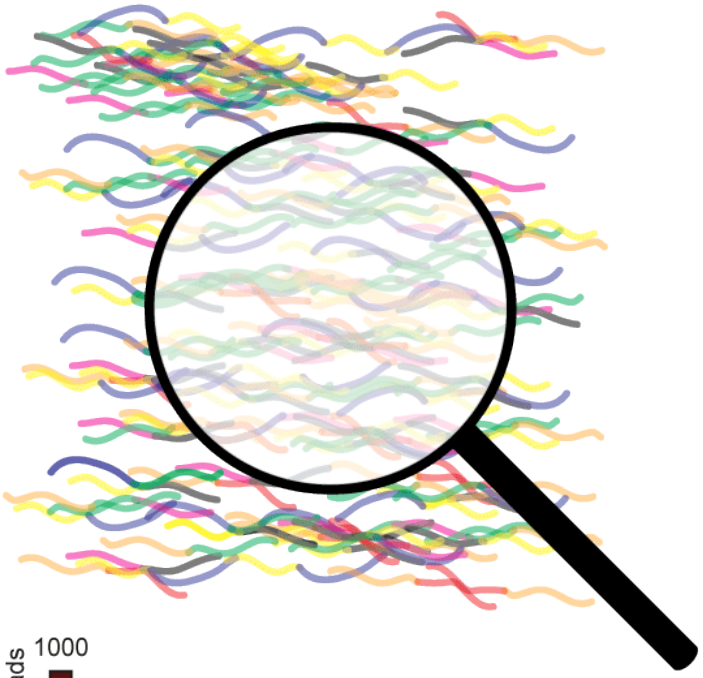
Hi-C Ligation Products
(Chimeric DNA)

Bryan Lajoie, Dekker lab

Visualizing C-data

3 Reads
$0.00005

Hi-C Ligation Products
(Chimeric DNA)

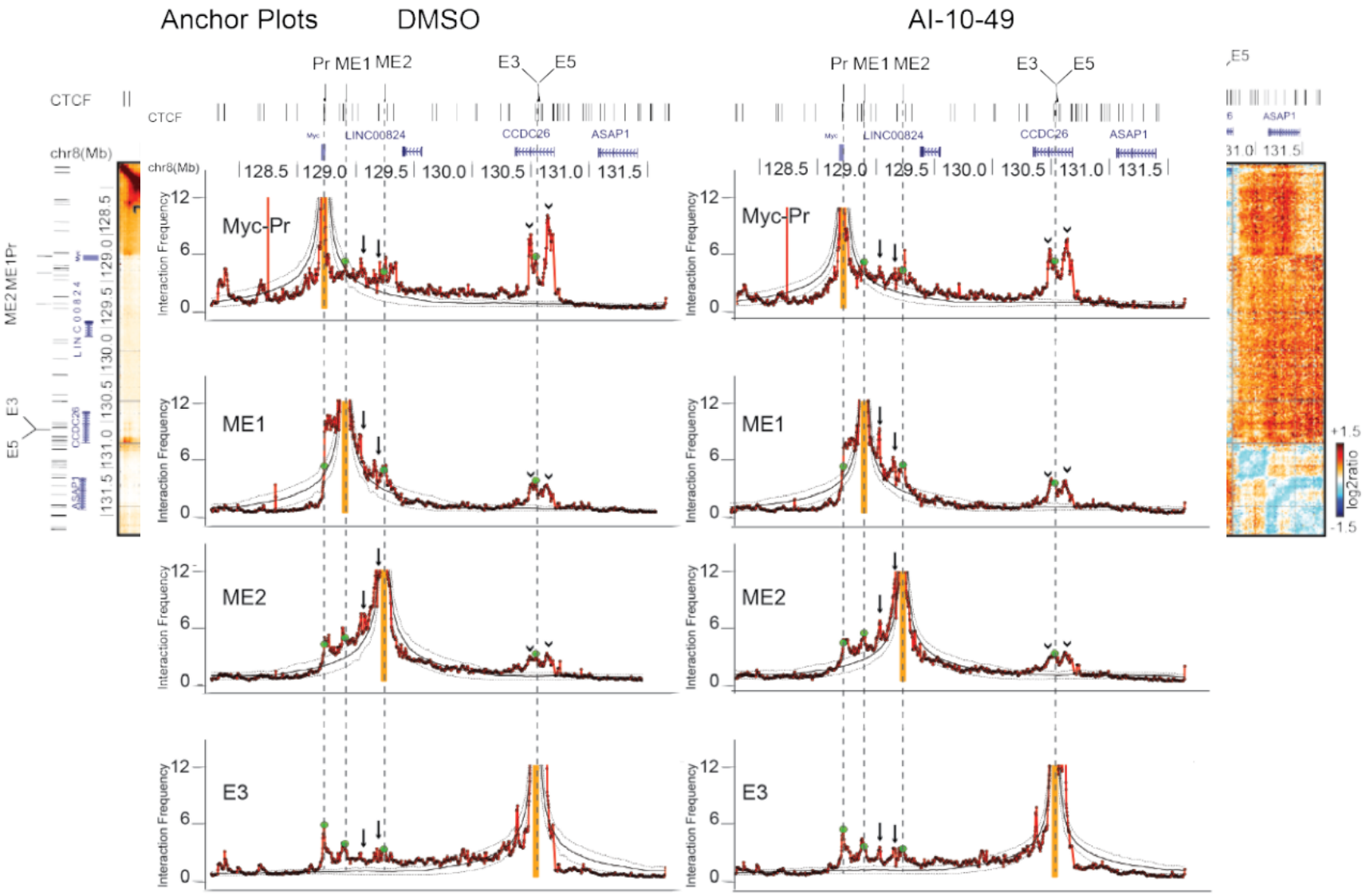


Bryan Lajoie, Dekker lab

The power of enrichment

- Regional data can become extremely high-res.
  - A 3Mb 5C experiment of 30 $x10^6$ mapped reads would require **30x 10^9 valid pairs** genome wide!
  - At this resolution we can quantify (subtle) differences in promoter-enhancer interactions!

5C enrichment



Houda Belaghzal, Dekker lab

Acknowledgements

Dekker lab

- Houda Belaghzal
- Hakan Ozadam
- Job Dekker

McCord Lab

- Rosela Golloshi
- Jacob Sanders
- Rachel McCord

Mirny Lab

- Nezar Abdennur
- Leonid Mirny

Program in Systems Biology

The University of Tennessee Knoxville

MIT Mirny Lab