Micro-C Experimental Design

Research Goals

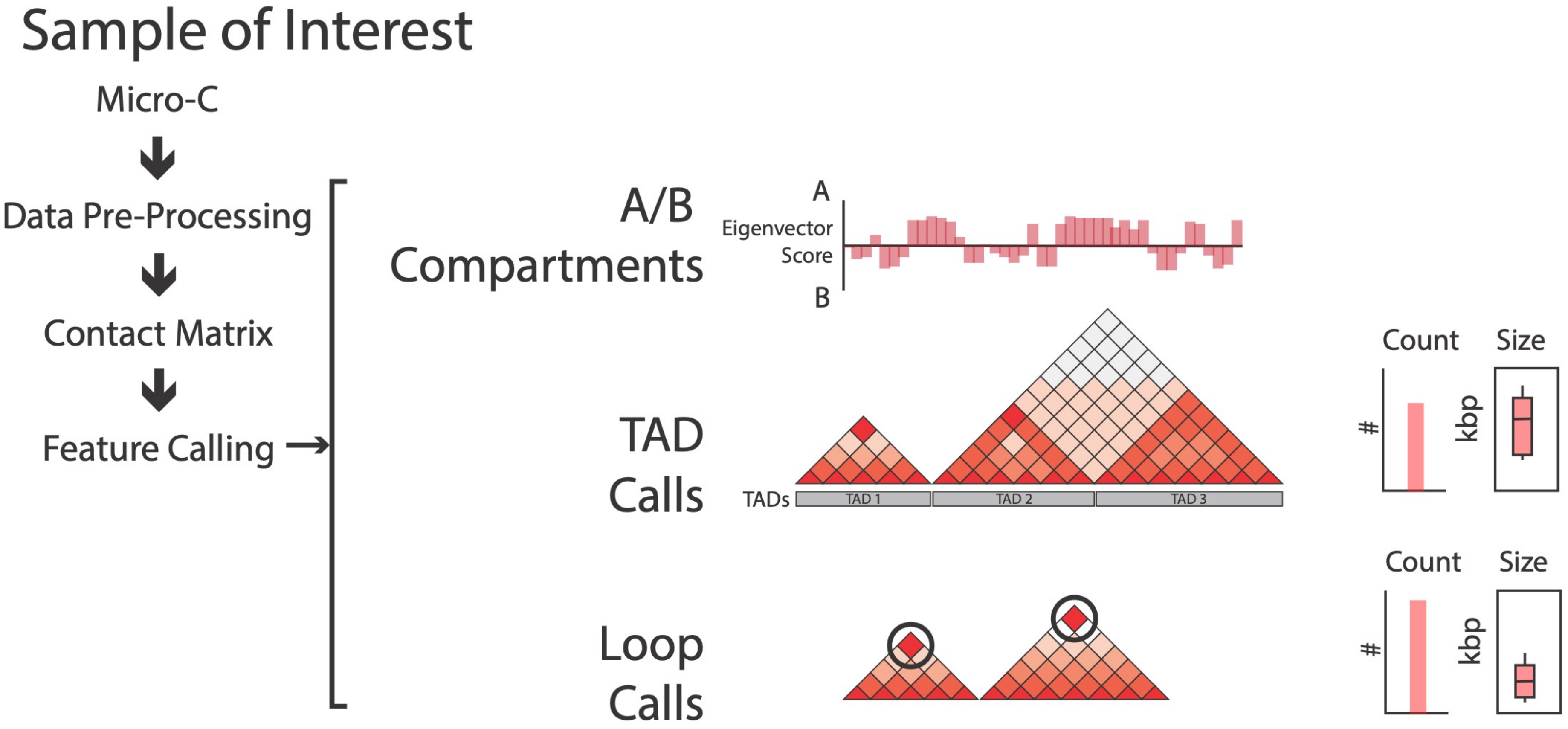
Typically, there are two main types of research goals

1. **Discovery** – Investigate and describe the chromatin conformation of a single sample or condition. The analyses typically consists of data pre-processing, and feature calling (e.g. A/B Compartments, TADs, and Loops). Additional analyses could include contextualizing these features with other data such as RNA-seq, ATAC-seq or ChIP-seq.

2. **Comparison** – Compare chromatin conformation between two samples or conditions such as, tumor-normal, treatment-control, or wildtype-knockout. The analyses typically consists of data pre-processing, and feature calling (e.g. A/B Compartments, TADs, and Loops) followed by a comparison of these feature locations between the two conditions. Such comparisons include A/B Compartment switches, gain or loss of TAD, TAD merging, and shared vs. unique chromatin loops.

Note: Regardless of your research goal we recommend a minimum of 80X-90X coverage (2x150 bp) per condition or sample to enable feature calling across the scale of hierarchical chromatin conformation ranges from 1Mb – 5kb. For example in the human genome that would be 800-900 Million read pairs. Higher resolutions will require more sequencing. 80X -90X coverage is a good place to start.

Example of Discovery-based Analyses



Sample of Interest

Micro-C

Data Pre-Processing

Contact Matrix

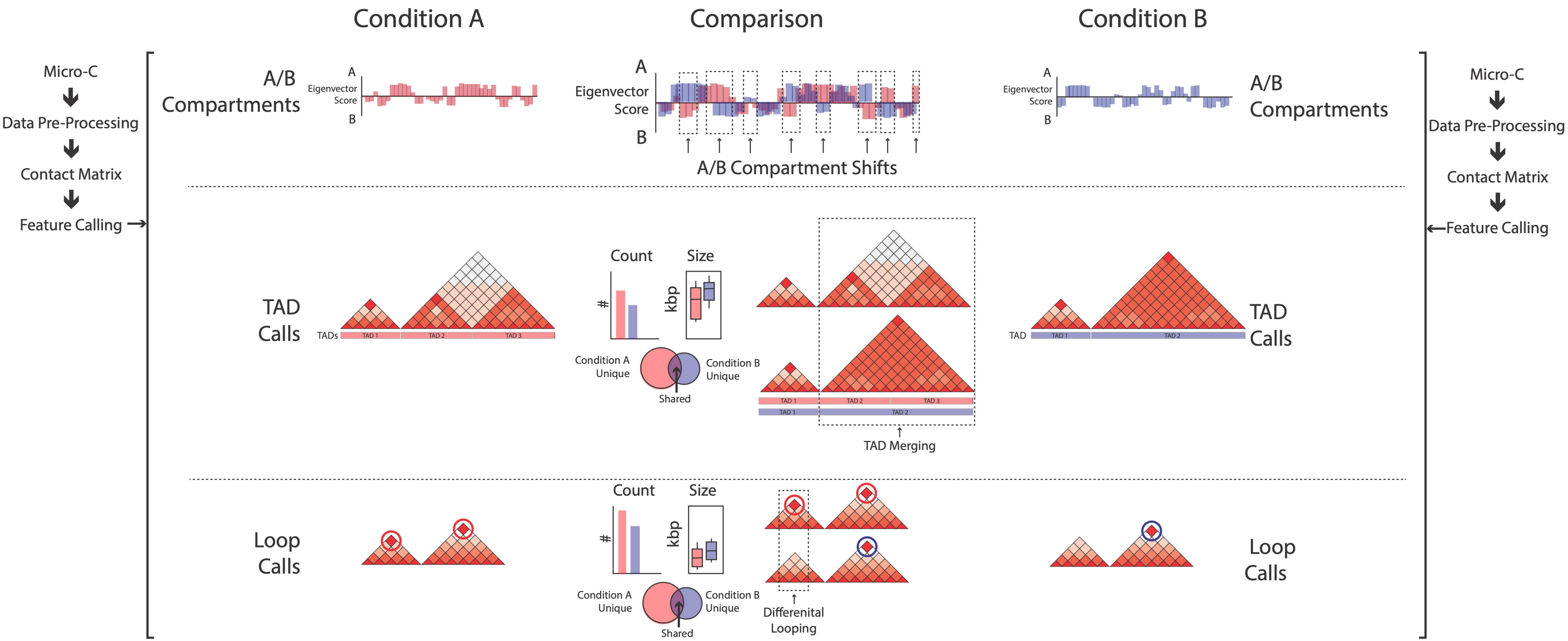
Feature Calling

A/B Compartments

TAD Calls

Loop Calls

Example of Comparison-based Analyses



Condition A

Comparison

Condition B

Micro-C → Data Pre-Processing → Contact Matrix → Feature Calling →

Micro-C → Data Pre-Processing → Contact Matrix → ←Feature Calling

A/B Compartments

A/B Compartment Shifts

TAD Calls

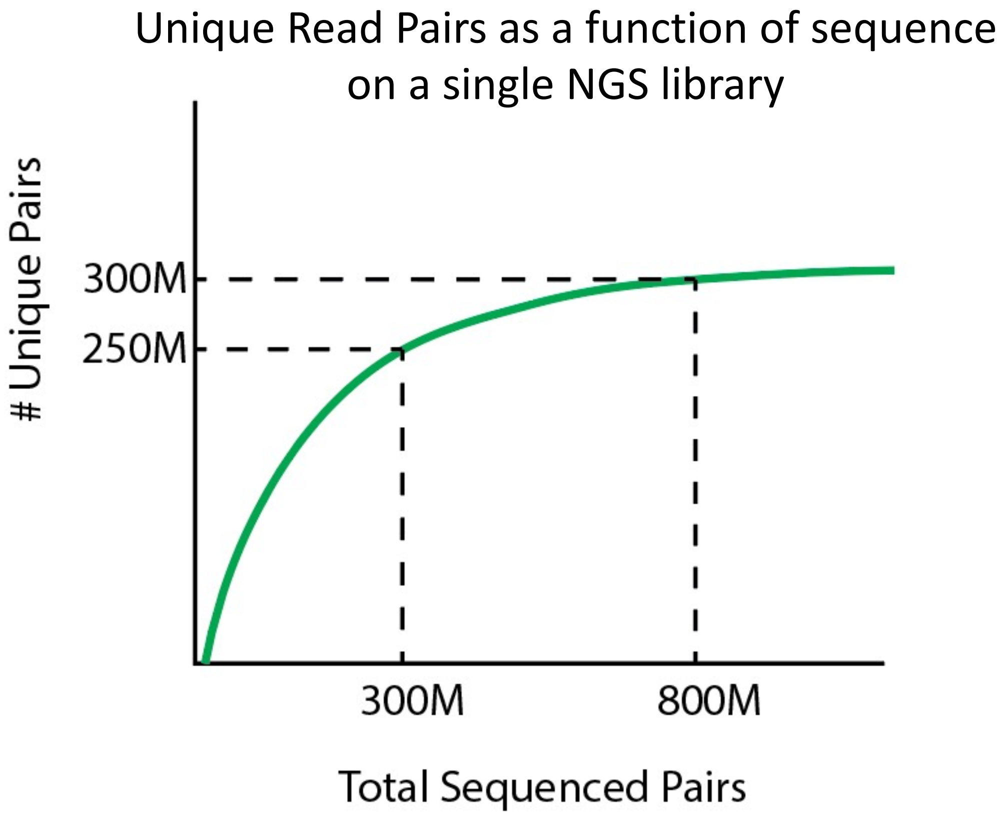
TAD Merging

Loop Calls

Differenital Looping

How to get to 800-900 million read pairs?

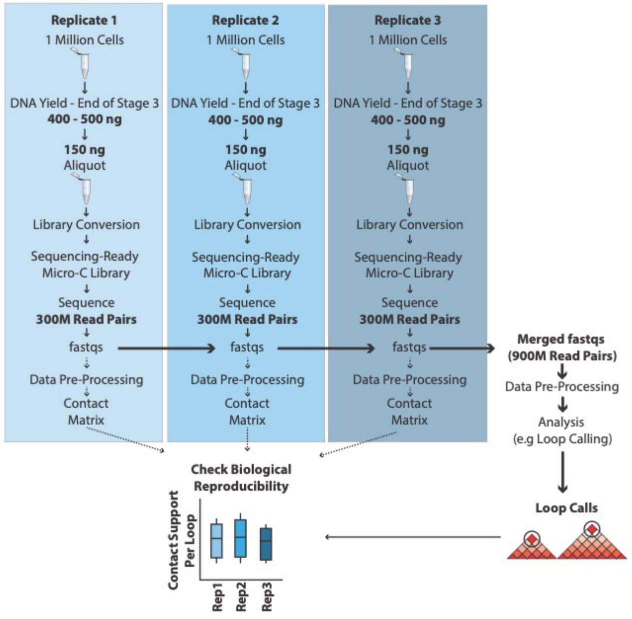
- For feature calling we recommend sequencing to 800-900 million read pairs (2x150bp)
- Sequencing a single library beyond 300 million read pairs typically results in a high PCR duplicate in the fastqs. This is because the number of unique molecules is saturated at this depth, so if you continue to sequence more, the molecules being captured are PCR duplicates.
- In order to avoid high PCR duplication in deep sequencing we recommend replication. This replication can be achieved in <u>three different ways</u>.



Overview Of The Replication Approaches

Dovetail GENOMICS
An EdenRoc Sciences Company

1. Biological Replication



Replicate 1
1 Million Cells
DNA Yield - End of Stage 3
400 - 500 ng
150 ng
Aliquot
Library Conversion
Sequencing-Ready Micro-C Library
Sequence
300M Read Pairs
fastqs
Data Pre-Processing
Contact Matrix

Replicate 2
1 Million Cells
DNA Yield - End of Stage 3
400 - 500 ng
150 ng
Aliquot
Library Conversion
Sequencing-Ready Micro-C Library
Sequence
300M Read Pairs
fastqs
Data Pre-Processing
Contact Matrix

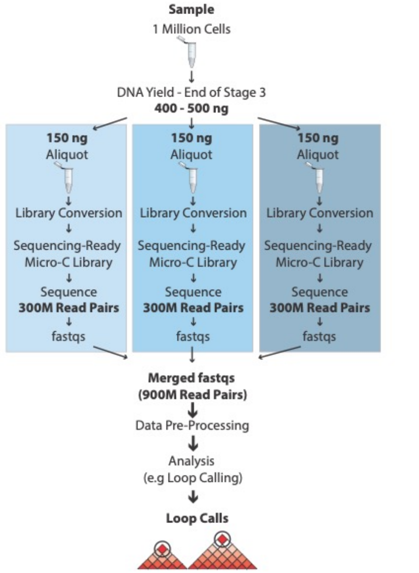
Replicate 3
1 Million Cells
DNA Yield - End of Stage 3
400 - 500 ng
150 ng
Aliquot
Library Conversion
Sequencing-Ready Micro-C Library
Sequence
300M Read Pairs
fastqs
Data Pre-Processing
Contact Matrix

Merged fastqs
(900M Read Pairs)
Data Pre-Processing
Analysis
(e.g Loop Calling)
Loop Calls

Check Biological Reproducibility

Summary
Cells: 3 Million (1 Million per replicate)
Micro-C Proximity Ligation Reactions: 3
Libraries: 3
Read Pairs: 900 Million (300 Million per biological replicate)

2. Technical Replication



Sample
1 Million Cells
DNA Yield - End of Stage 3
400 - 500 ng

150 ng
Aliquot
Library Conversion
Sequencing-Ready Micro-C Library
Sequence
300M Read Pairs
fastqs

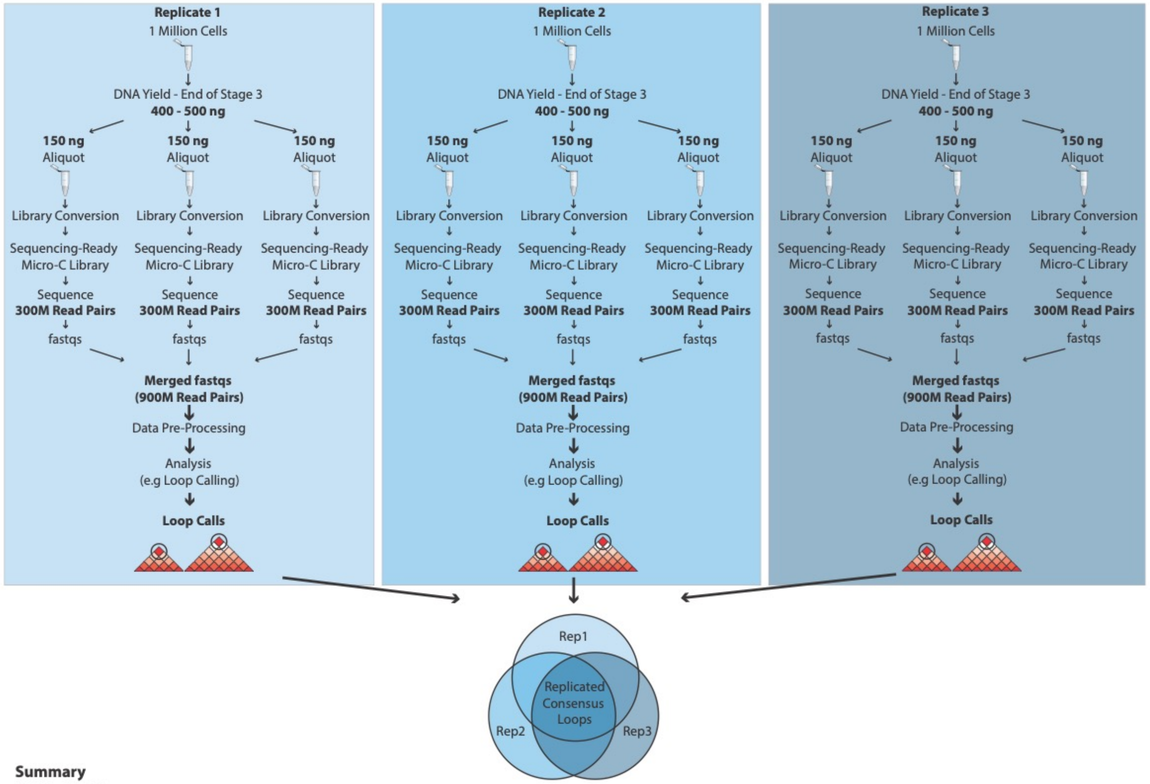
150 ng
Aliquot
Library Conversion
Sequencing-Ready Micro-C Library
Sequence
300M Read Pairs
fastqs

150 ng
Aliquot
Library Conversion
Sequencing-Ready Micro-C Library
Sequence
300M Read Pairs
fastqs

Merged fastqs
(900M Read Pairs)
Data Pre-Processing
Analysis
(e.g Loop Calling)
Loop Calls

Summary
Cells: 1 Million
Micro-C Proximity Ligation Reactions: 1
Libraries: 3
Read Pairs: 900 Million (300 Million per technical replicate)

3. Combination of Biological Replicates and Technical Replicates



Replicate 1
1 Million Cells
DNA Yield - End of Stage 3
400 - 500 ng
150 ng Aliquot → Library Conversion → Sequencing-Ready Micro-C Library → Sequence **300M Read Pairs** → fastqs (×3)
Merged fastqs
(900M Read Pairs)
Data Pre-Processing
Analysis
(e.g Loop Calling)
Loop Calls

Replicate 2
1 Million Cells
DNA Yield - End of Stage 3
400 - 500 ng
150 ng Aliquot → Library Conversion → Sequencing-Ready Micro-C Library → Sequence **300M Read Pairs** → fastqs (×3)
Merged fastqs
(900M Read Pairs)
Data Pre-Processing
Analysis
(e.g Loop Calling)
Loop Calls

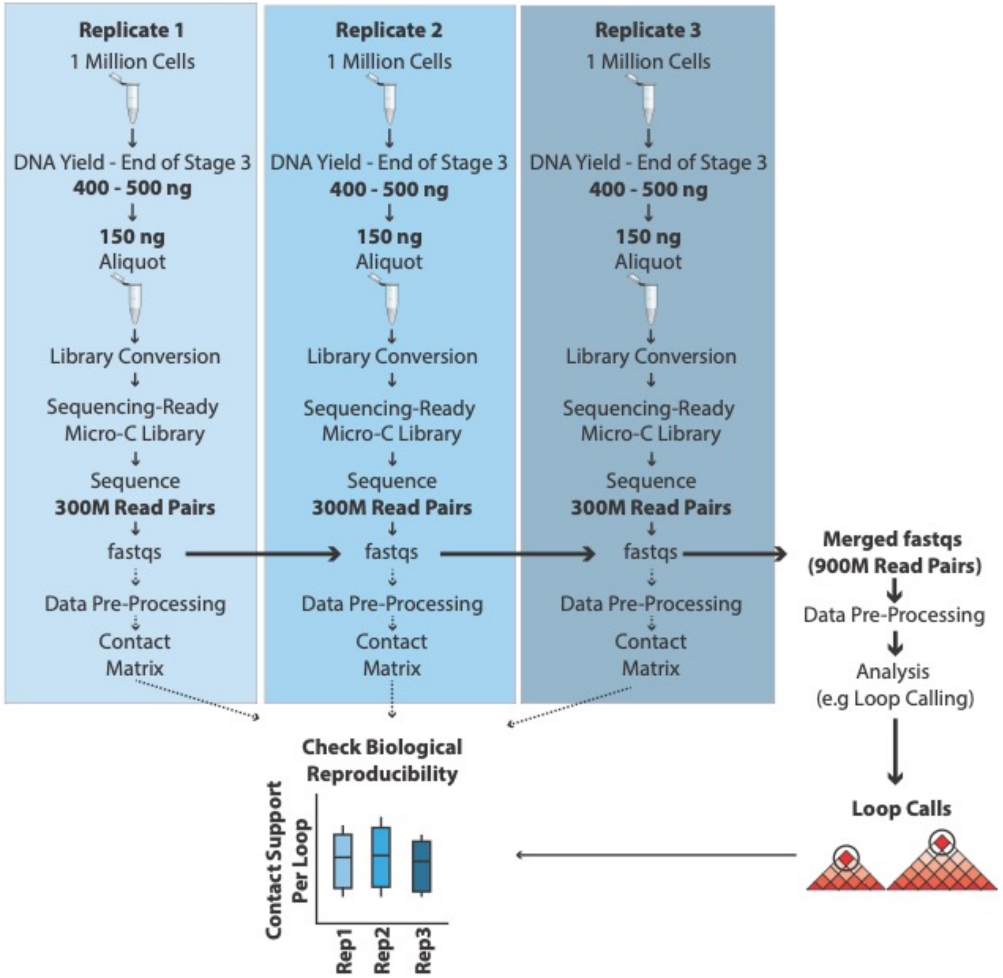
Replicate 3
1 Million Cells
DNA Yield - End of Stage 3
400 - 500 ng
150 ng Aliquot → Library Conversion → Sequencing-Ready Micro-C Library → Sequence **300M Read Pairs** → fastqs (×3)
Merged fastqs
(900M Read Pairs)
Data Pre-Processing
Analysis
(e.g Loop Calling)
Loop Calls

Summary
Cells: 3 Million
Micro-C Proximity Ligation Reactions: 3
Libraries: 9
Read Pairs: 2.7 Billion (900 Million per biological replicate)

The next slides will cover these approaches in more detail

1. Biological Replication

- 3 biological replicates of a sample or experimental condition
- Prepare 1 library per biological replicate
- Sequence each replicate to ~300 M read pairs.
- Merge the sequence data from all three replicates
- Perform feature calling on the combined data (~900 M read pairs)
- Behavior of individual replicates can be assessed for concordance.
- This approach doesn't rely much on complementary data to prove a hypothesis, although having complementary data is helpful in interpretating the findings.
- This approach is commonly used in both discovery-based and comparison-based studies.



Replicate 1
1 Million Cells
DNA Yield - End of Stage 3
400 - 500 ng
150 ng
Aliquot
Library Conversion
Sequencing-Ready
Micro-C Library
Sequence
300M Read Pairs
fastqs
Data Pre-Processing
Contact
Matrix

Replicate 2
1 Million Cells
DNA Yield - End of Stage 3
400 - 500 ng
150 ng
Aliquot
Library Conversion
Sequencing-Ready
Micro-C Library
Sequence
300M Read Pairs
fastqs
Data Pre-Processing
Contact
Matrix

Replicate 3
1 Million Cells
DNA Yield - End of Stage 3
400 - 500 ng
150 ng
Aliquot
Library Conversion
Sequencing-Ready
Micro-C Library
Sequence
300M Read Pairs
fastqs
Data Pre-Processing
Contact
Matrix

Merged fastqs
(900M Read Pairs)
Data Pre-Processing
Analysis
(e.g Loop Calling)

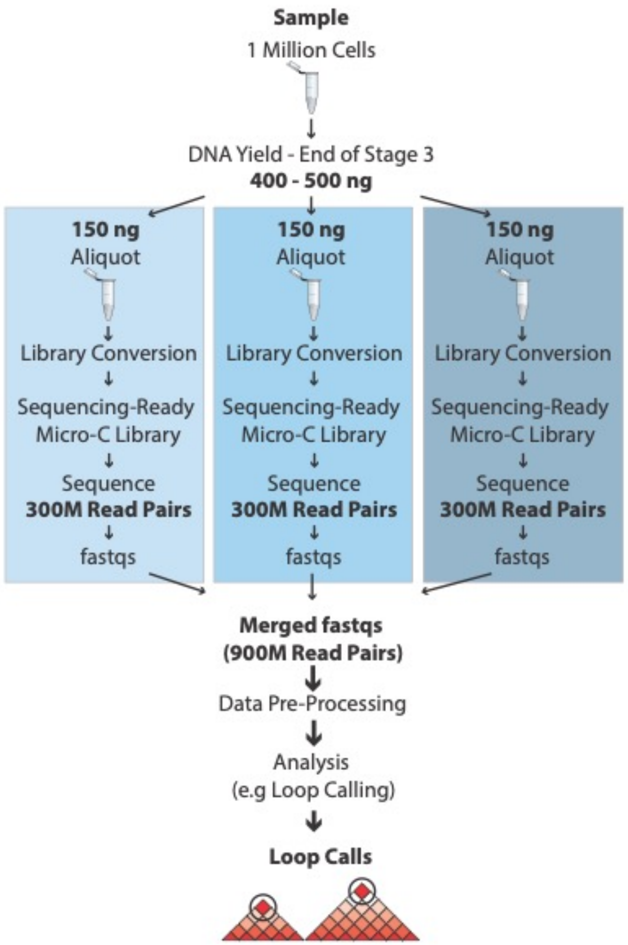
Loop Calls

Check Biological
Reproducibility

Summary
Cells: 3 Million (1 Million per replicate)
Micro-C Proximity Ligation Reactions: 3
Libraries: 3
Read Pairs: 900 Million (300 Million per biological replicate)

2. Technical Replication

- 1 biological replicates of a sample or experimental condition
- Prepare 3 library per biological replicate
- Sequence each replication to ~300 M read pairs.
- Merge the sequence data from all three replicates
- Perform feature calling on the combined data (~900 M read pairs)
- This approach relies heavily on complementary data to prove a hypothesis.
- This approach is commonly used in both discovery-based and occasionally in comparison-based studies where starting material is limiting.



Sample
1 Million Cells
↓
DNA Yield - End of Stage 3
400 - 500 ng

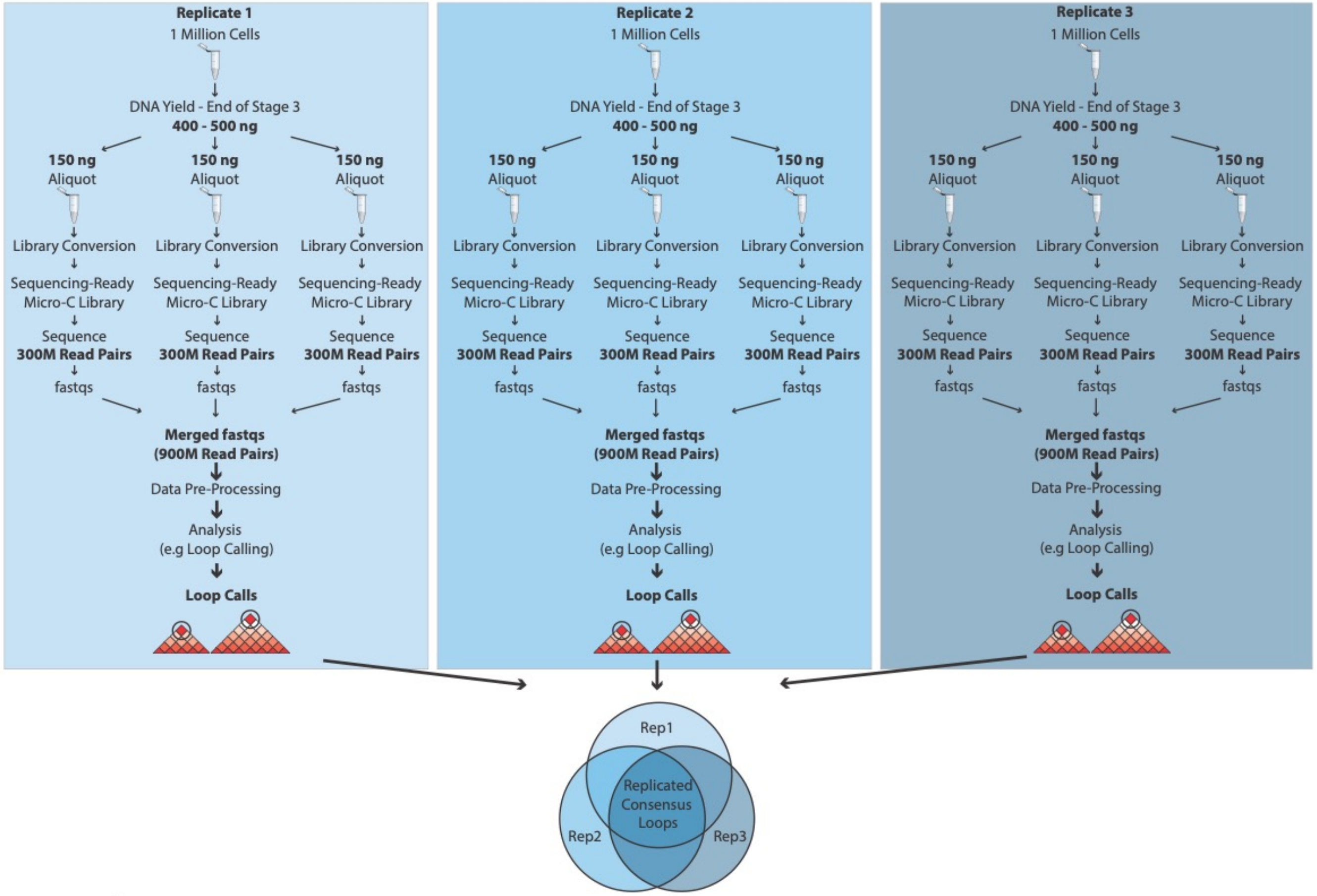
150 ng Aliquot	**150 ng** Aliquot	**150 ng** Aliquot
↓	↓	↓
Library Conversion	Library Conversion	Library Conversion
↓	↓	↓
Sequencing-Ready Micro-C Library	Sequencing-Ready Micro-C Library	Sequencing-Ready Micro-C Library
↓	↓	↓
Sequence **300M Read Pairs**	Sequence **300M Read Pairs**	Sequence **300M Read Pairs**
↓	↓	↓
fastqs	fastqs	fastqs

Merged fastqs
(900M Read Pairs)
↓
Data Pre-Processing
↓
Analysis
(e.g Loop Calling)
↓
Loop Calls

Summary
Cells: 1 Million
Micro-C Proximity Ligation Reactions: 1
Libraries: 3
Read Pairs: 900 Million (300 Million per technical replicate)

3. Combination Of Biological And Technical Replication

- 3 biological replicates of a sample or experimental condition
- Prepare 3 library per biological replicate
- Sequence each replicate to ~300 M read pairs.
- Merge the sequence data from all 3 technical replicates withing a biological replication
- Perform feature calling on each biological replicate data (~900 M read pairs)
- Find shared features across all replicates to get high confidence feature calls
- This approach doesn't rely much on complementary data to prove a hypothesis, although having complementary data is helpful in interpretating the findings.
- This approach is suitable for both discovery-based and comparison-based studies but is rare due to high sequencing burden.



Replicate 1
1 Million Cells
DNA Yield - End of Stage 3
400 - 500 ng
150 ng Aliquot | 150 ng Aliquot | 150 ng Aliquot
Library Conversion | Library Conversion | Library Conversion
Sequencing-Ready Micro-C Library | Sequencing-Ready Micro-C Library | Sequencing-Ready Micro-C Library
Sequence 300M Read Pairs | Sequence 300M Read Pairs | Sequence 300M Read Pairs
fastqs | fastqs | fastqs
Merged fastqs (900M Read Pairs)
Data Pre-Processing
Analysis (e.g Loop Calling)
Loop Calls

Replicate 2
1 Million Cells
DNA Yield - End of Stage 3
400 - 500 ng
150 ng Aliquot | 150 ng Aliquot | 150 ng Aliquot
Library Conversion | Library Conversion | Library Conversion
Sequencing-Ready Micro-C Library | Sequencing-Ready Micro-C Library | Sequencing-Ready Micro-C Library
Sequence 300M Read Pairs | Sequence 300M Read Pairs | Sequence 300M Read Pairs
fastqs | fastqs | fastqs
Merged fastqs (900M Read Pairs)
Data Pre-Processing
Analysis (e.g Loop Calling)
Loop Calls

Replicate 3
1 Million Cells
DNA Yield - End of Stage 3
400 - 500 ng
150 ng Aliquot | 150 ng Aliquot | 150 ng Aliquot
Library Conversion | Library Conversion | Library Conversion
Sequencing-Ready Micro-C Library | Sequencing-Ready Micro-C Library | Sequencing-Ready Micro-C Library
Sequence 300M Read Pairs | Sequence 300M Read Pairs | Sequence 300M Read Pairs
fastqs | fastqs | fastqs
Merged fastqs (900M Read Pairs)
Data Pre-Processing
Analysis (e.g Loop Calling)
Loop Calls

Rep1
Rep2
Rep3
Replicated Consensus Loops

Summary
Cells: 3 Million
Micro-C Proximity Ligation Reactions: 3
Libraries: 9
Read Pairs: 2.7 Billion (900 Million per biological replicate)

Considerations For Making Your Choice

- **The biological question** – some questions require more statistical rigor than others
- **Availability of orthogonal or complementary data** – The ability to have other data points supporting your findings can reduce the need to replicate
- **Targeted Journal** – Higher tier journals are associated with a high degree of rigor and replication may be a requirement to get your manuscript through the review process
- **Available starting material** – If you a minimal starting material, this will impact your ability to do biological replicates, if that is the case technical replication of the libraries may be the only route forward
- **Budget** – Some approaches require a large amount of sequencing, and when doing comparison analyses that sequencing cost is multiplied, sequencing depth is an important component to consider when making your decision

Summary of Approaches

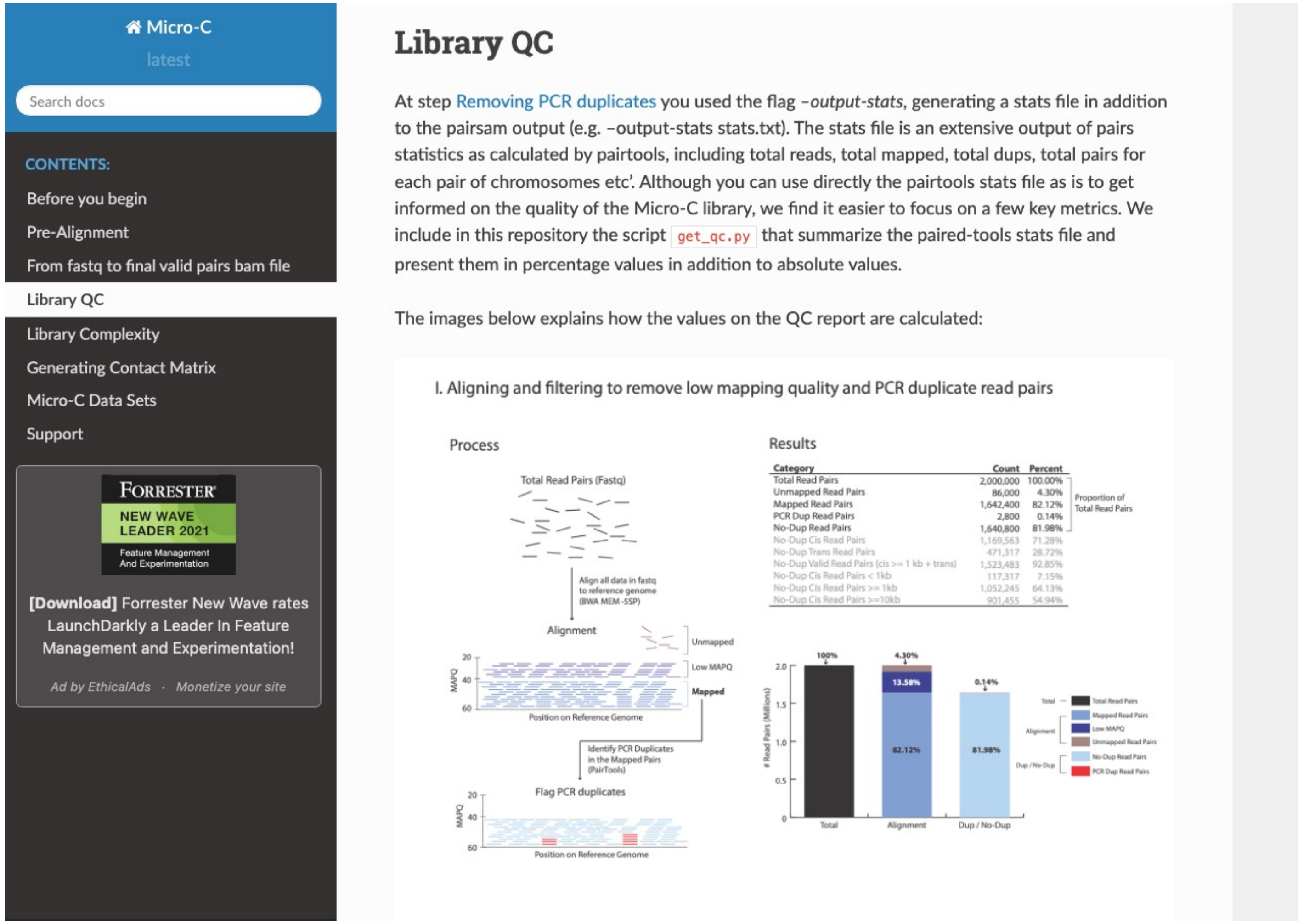
	Biological Replication	Technical Replication	Biological + Technical
Input Requirement	High	Low	High
Reaction Usage	Moderate	Low	High
Sequencing Cost	Minimal	Minimal	Maximal
Statistical Rigor	Moderate	Low	High
Complementary Data Requirement To Publish	Preferred	Required	Not Required, but nice to have
How Often is the Approach Used	Common	Common	Rare

Data Analysis – Data Pre-Processing

- Data pre-processing steps are outlined in our guide to from fastqs to contact matrices here: https://micro-c.readthedocs.io/en/latest/

- These steps follow the best practices guides outlined by the 4D nucleosome consortium

Library QC page from the pre-processing guide



Library QC

At step Removing PCR duplicates you used the flag *–output-stats*, generating a stats file in addition to the pairsam output (e.g. –output-stats stats.txt). The stats file is an extensive output of pairs statistics as calculated by pairtools, including total reads, total mapped, total dups, total pairs for each pair of chromosomes etc'. Although you can use directly the pairtools stats file as is to get informed on the quality of the Micro-C library, we find it easier to focus on a few key metrics. We include in this repository the script `get_qc.py` that summarize the paired-tools stats file and present them in percentage values in addition to absolute values.

The images below explains how the values on the QC report are calculated:

I. Aligning and filtering to remove low mapping quality and PCR duplicate read pairs

Process

Total Read Pairs (Fastq)
Align all data in fastq to reference genome (BWA MEM -5SP)
Alignment
Unmapped
Low MAPQ
Mapped
MAPQ
20
40
60
Position on Reference Genome
Identify PCR Duplicates in the Mapped Pairs (PairTools)
Flag PCR duplicates
MAPQ
20
40
60
Position on Reference Genome

Results

Category	Count	Percent
Total Read Pairs	2,000,000	100.00%
Unmapped Read Pairs	86,000	4.30%
Mapped Read Pairs	1,642,400	82.12%
PCR Dup Read Pairs	2,800	0.14%
No-Dup Read Pairs	1,640,800	81.98%
No-Dup Cis Read Pairs	1,169,563	71.28%
No-Dup Trans Read Pairs	471,317	28.72%
No-Dup Valid Read Pairs (cis >= 1 kb + trans)	1,523,483	92.85%
No-Dup Cis Read Pairs < 1kb	117,317	7.15%
No-Dup Cis Read Pairs >= 1kb	1,052,245	64.13%
No-Dup Cis Read Pairs >=10kb	901,455	54.94%

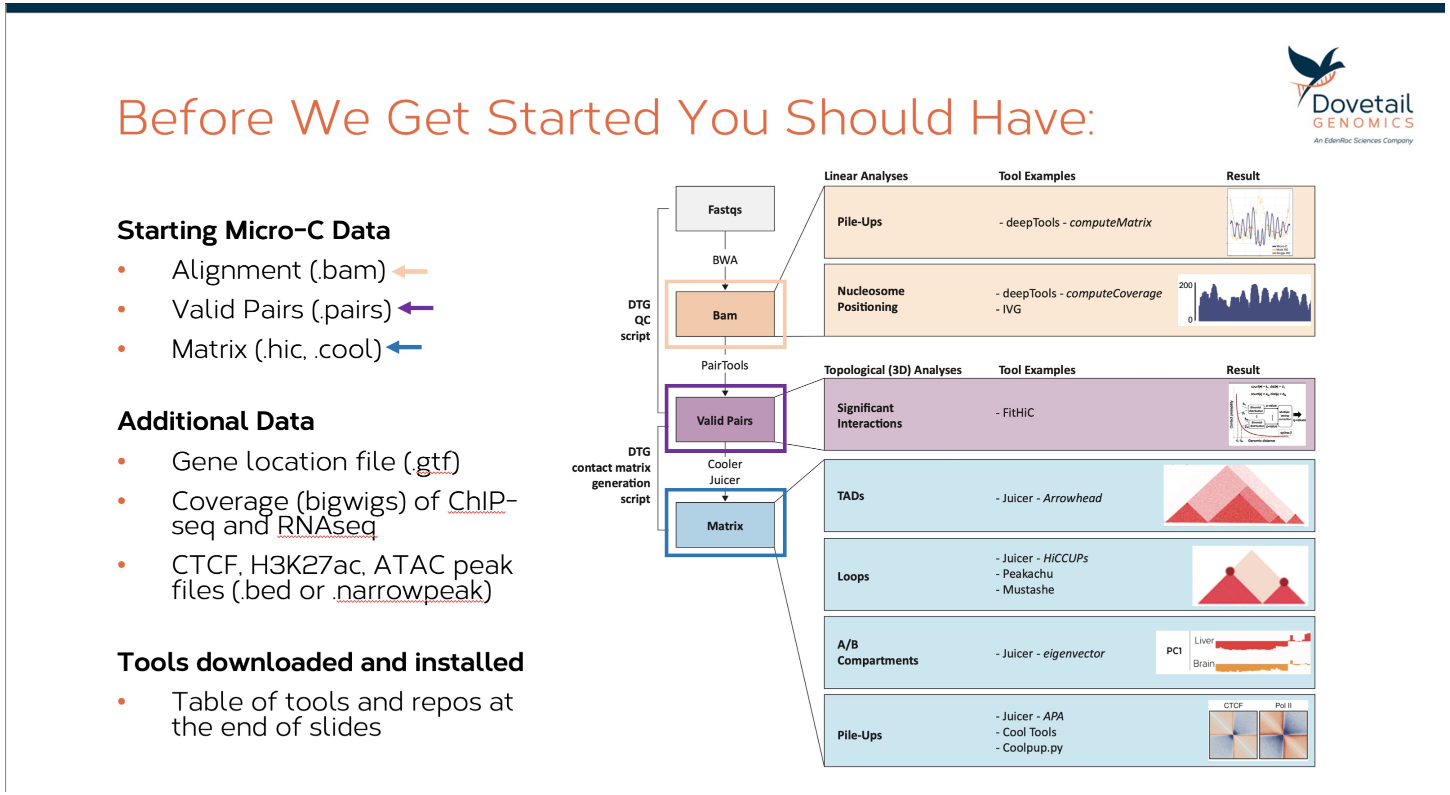
Proportion of Total Read Pairs

Data Analysis – Feature Calling

- We have a virtual workshop available that outlines step-by-step how to feature call your Micro-C data

- Standard tools are used to call features out of a Micro-C contact matrix such as:
  - Juicer –eigenvector for A/B Compartments
  - Juicer –arrowhead for TADs
  - Juicer –HiCCUPs for Loops
  - Mustache a Micro-C specific Loop calling

Screenshot of the workshop



Before We Get Started You Should Have:

Starting Micro-C Data

- Alignment (.bam) ←
- Valid Pairs (.pairs) ←
- Matrix (.hic, .cool) ←

Additional Data

- Gene location file (.gtf)
- Coverage (bigwigs) of ChIP-seq and RNAseq
- CTCF, H3K27ac, ATAC peak files (.bed or .narrowpeak)

Tools downloaded and installed

- Table of tools and repos at the end of slides

Linear Analyses	Tool Examples	Result
Pile-Ups	- deepTools - *computeMatrix*	
Nucleosome Positioning	- deepTools - *computeCoverage* - IVG	

Topological (3D) Analyses	Tool Examples	Result
Significant Interactions	- FitHiC	
TADs	- Juicer - *Arrowhead*	
Loops	- Juicer - *HiCCUPs* - Peakachu - Mustashe	
A/B Compartments	- Juicer - *eigenvector*	
Pile-Ups	- Juicer - *APA* - Cool Tools - Coolpup.py	

Data Analysis - Comparisons

Compare Feature Calls

Most features are either in bed/bigwig or bedpe format so you can use very standard comparison tools

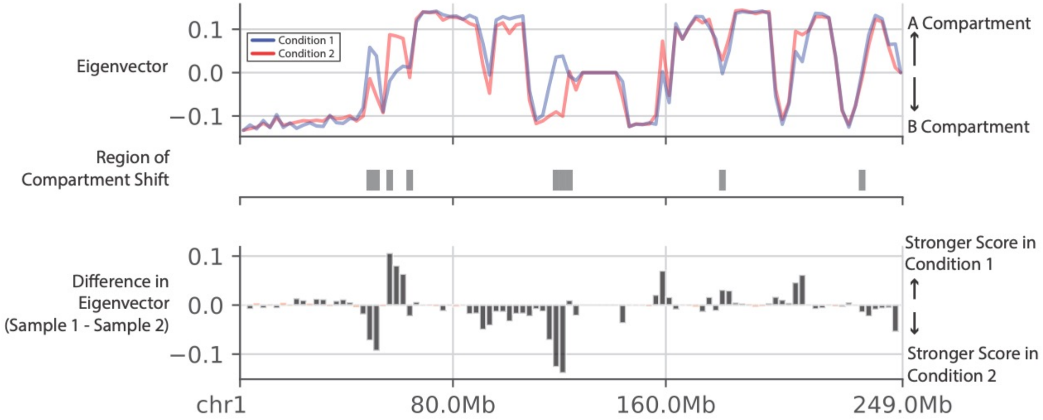
- Compare compartment score – bedtools intersect, deeptools bigwigCompare, with R
- TAD comparison - bedtools intersect
- Loop calls – bedtools pairToPair

Direct Comparison of matrix

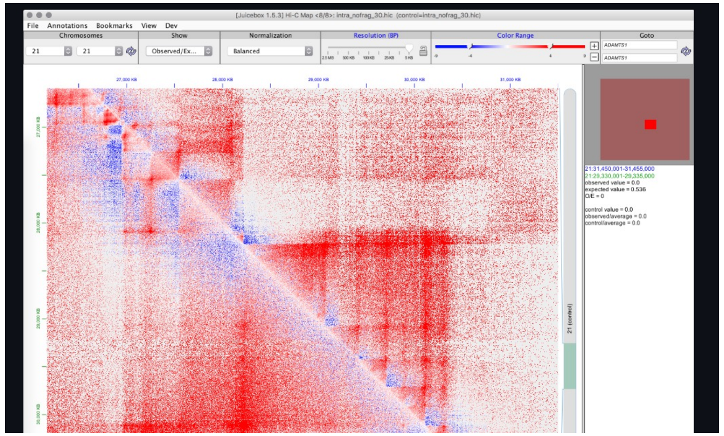
When you have two matrices you can directly compare contacts by dividing one by the other or subtracting one by the other. You need to be careful here to understand what is your question and what you're putting in. A good practice here is normalize the data in the pre-processing steps to have the same number of valid pairs into the matrix

- GUI – Juicebox
- Command line – Fan-C compare
- R - HicCompare

Example – A/B Compartment Shifts through R and bedtools



Example – fold change in Juicebox



Chromosomes: 21 21

Show: Observed/Ex...

Normalization: Balanced

Goto: ADAMTS1 ADAMTS1

21:31,450,001-31,455,000
21:29,330,001-29,335,000
observed value = 0.0
expected value = 0.536
O/E = 0

control value = 0.0
observed/average = 0.0
control/average = 0.0

Data Analysis – Data Visualizing

There are several tools that can be used to visualize data

- Data browsers that allow you to load data and scroll around in real time
  - Juicebox, HiGlass, WashU Epigenome Browser

- Plotters that allow you print specific regions of interests
  - Fan-C, HiCExplorer, R packages like Sushi