CS4330: Combinatorial Methods in Bioinformatics

Scaffolding

Wong Limsoon

Why scaffolding is needed

Sequencing reads are assembled into contigs

Contigs correspond to parts of a genome

They may be on either strand of the genome

They are unordered

Gaps between them have unknown size

What scaffolding is

Arrange contigs into correct order and orientation along chromosomes

Bridge gaps by estimating distance between contigs

Resolve regions of repetitive sequences which are hard to assemble

Proper scaffolding enhances the quality and completeness of genome assemblies, leading to more accurate genomic analyses

Scaffolding techniques

Paired-end sequencing

Mate-pair sequencing

Optical sequencing

Read up this one yourself

Hi-C sequencing

Commonly used techniques

Next lecture

Progeny and/or sibling sequencing

Interesting special context

Paired-end sequencing

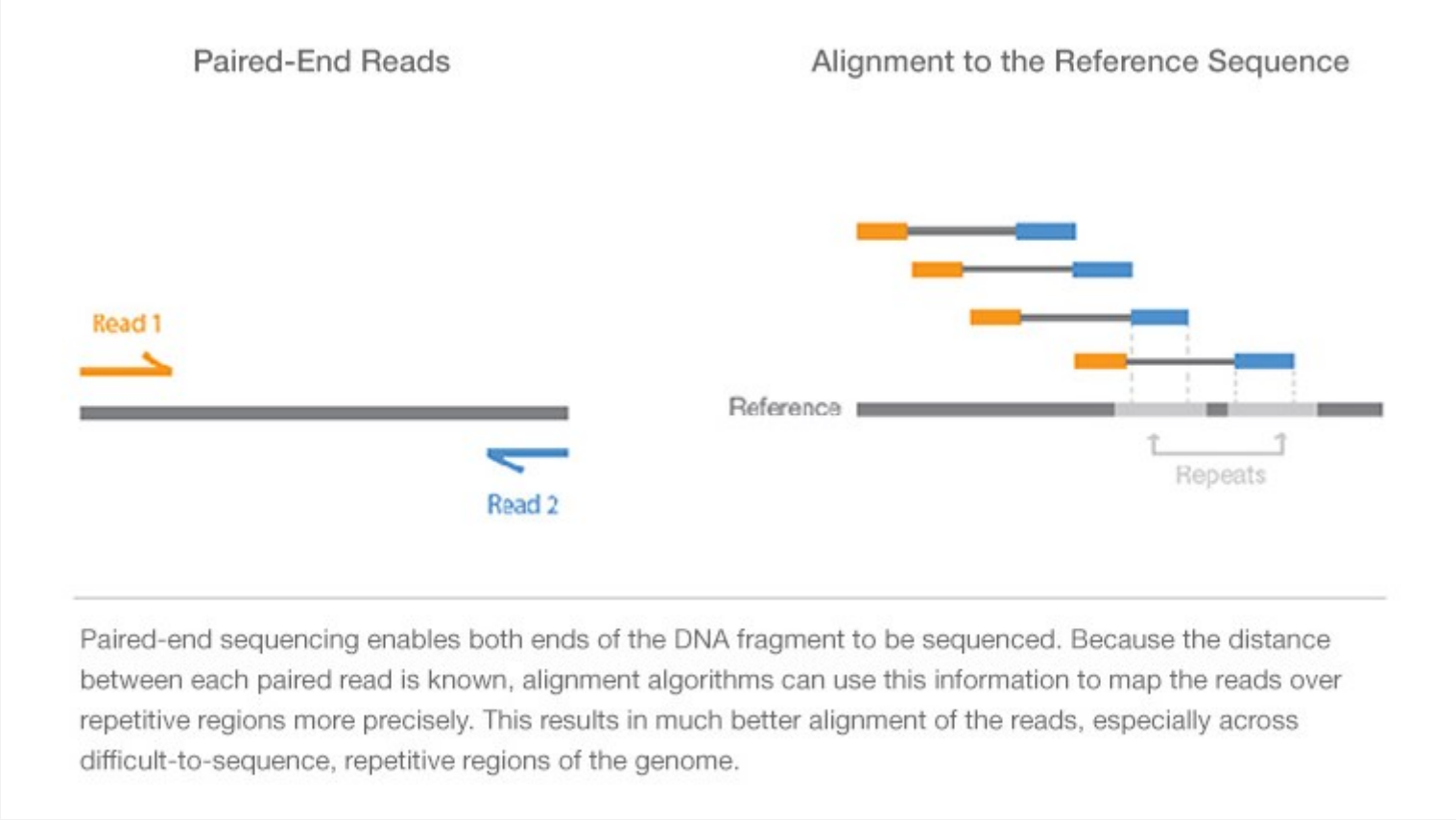
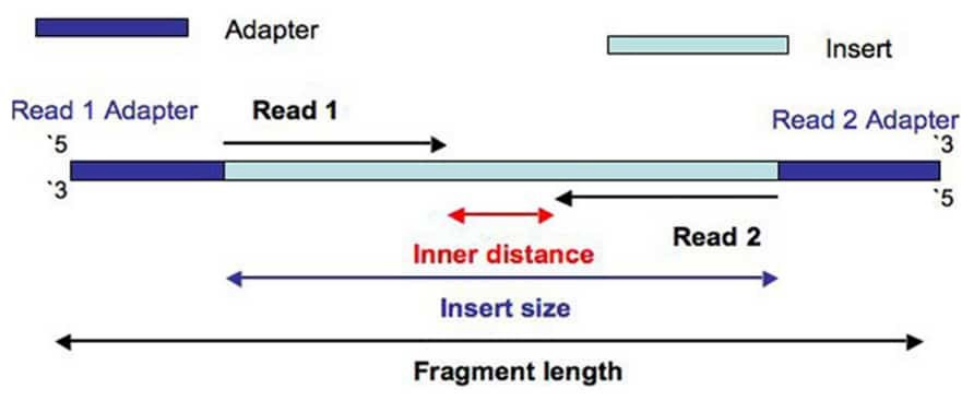
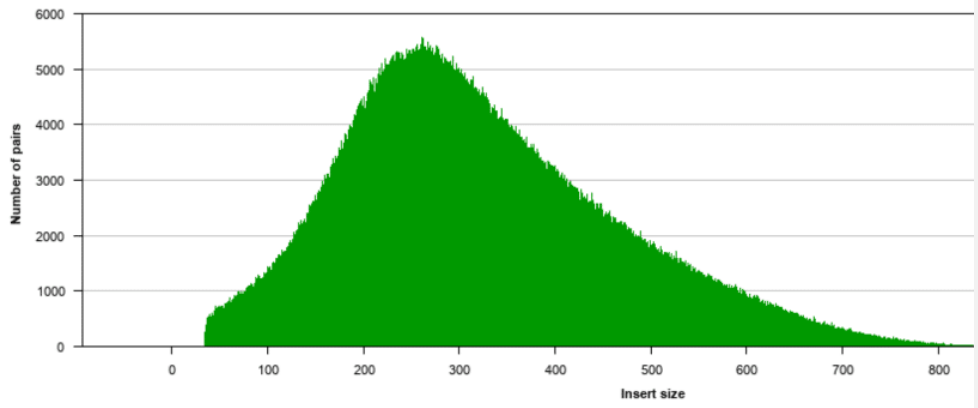


Image credit: Illumina

Paired-end sequencing, cont'd



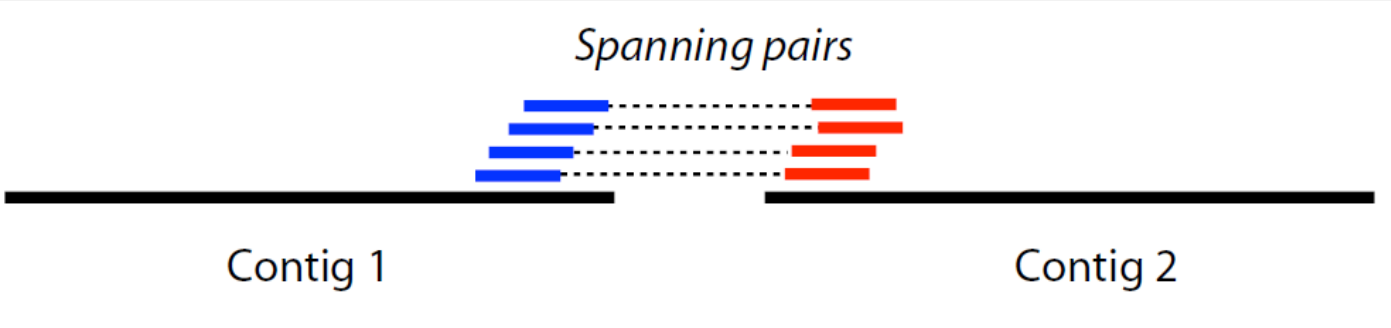


Read1 & read2 are sometimes called "mates", to indicate they form a pair

The Sequencing Center

Scaffolding, adjacent contigs

Say, we have a set of pairs which are assembled into two contigs like this:



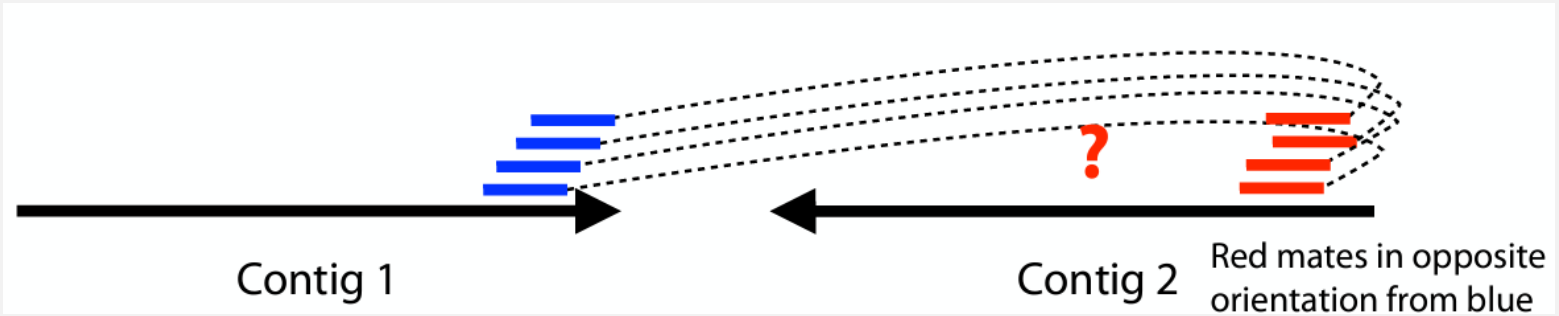
Some of the mates at one end of contig 1 are paired with mates in contig 2; these are called spanning pairs

Spanning pairs suggest the two contigs are close to each other, separated by the insert size of the mate pairs

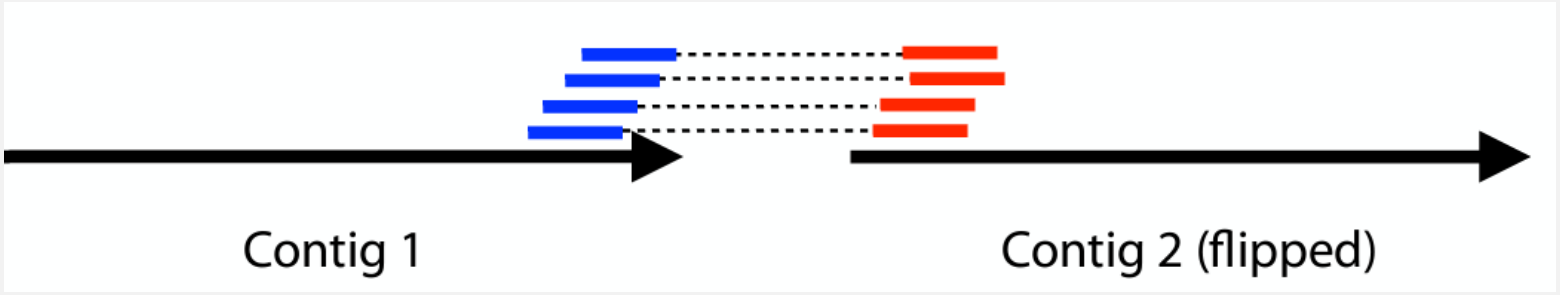
Taken from the lecture slides of Ben Langmead

Scaffolding, flipped contigs

Contig 2 assembled backwards

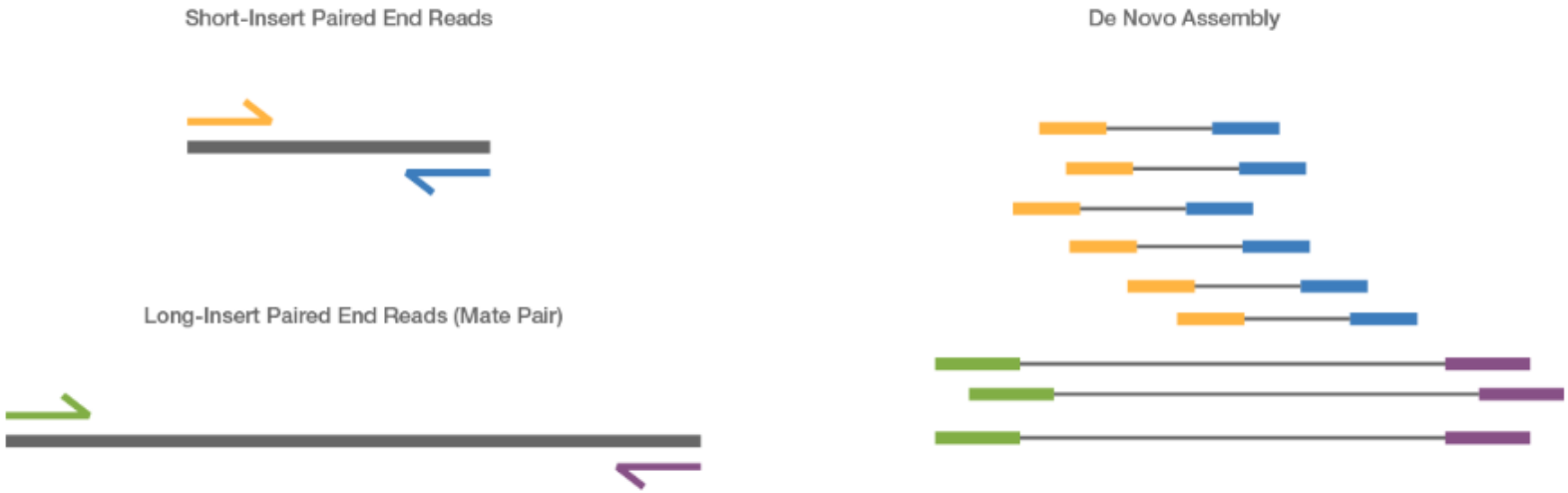


Flip (reverse complement) it



Taken from the lecture slides of Ben Langmead

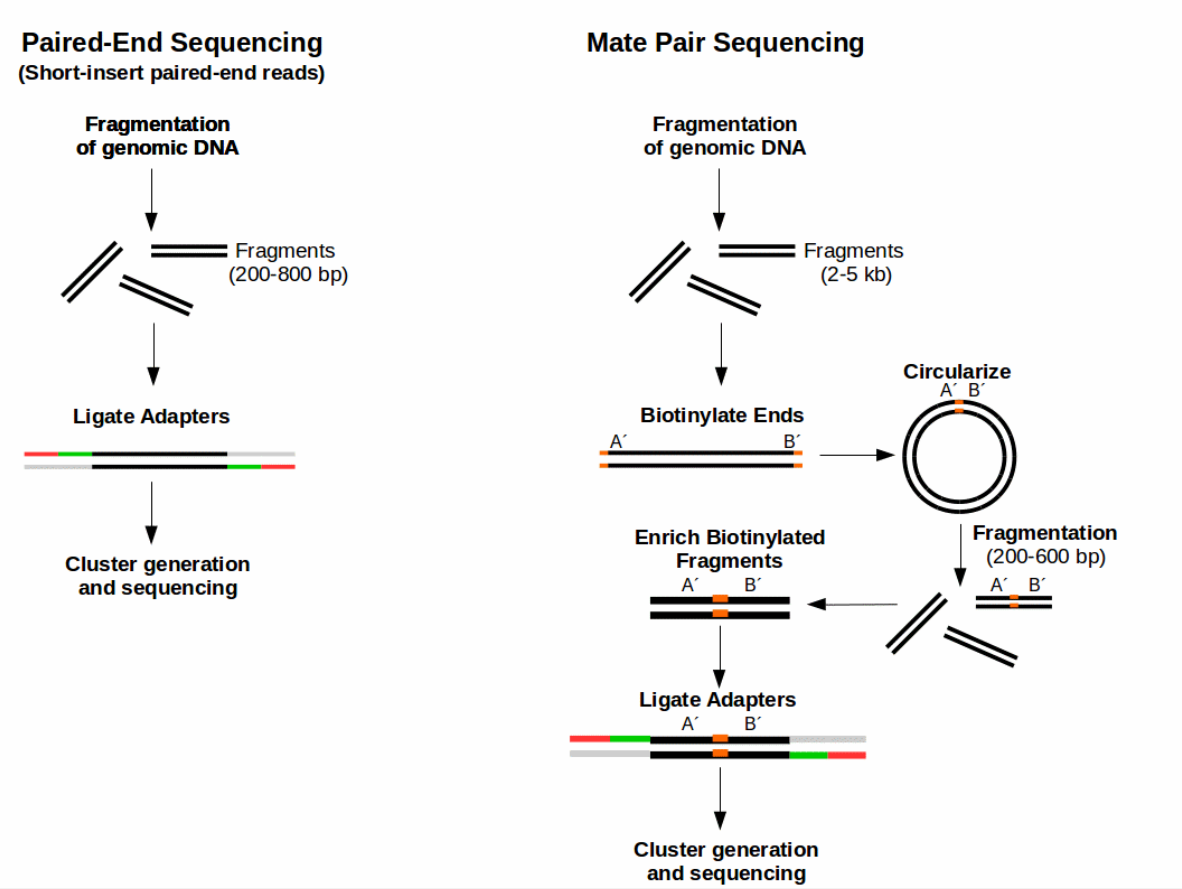
Mate-pair sequencing



Using a combination of short and long insert sizes with paired-end sequencing results in maximal coverage of the genome for de novo assembly. Because larger inserts can pair reads across greater distances, they provide a better ability to read through highly repetitive sequences and regions where large structural rearrangements have occured. Shorter inserts sequenced at higher depths can fill in gaps missed by larger inserts sequenced at lower depths. Thus a diverse library of short and long inserts results in better de novo assembly, leading to fewer gaps, larger contigs, and greater accuracy of the final consensus sequence.

Image credit: Illumina

Mate-pair sequencing, cont'd



https://www.ecseq.com/support/ngs/what-is-mate-pair-sequencing-useful-for

Scaffolding using mate pairs

Similar to scaffolding using paired-end reads, but much bigger insert size

However, bigger insert size is not always better; it depends on the distribution of repeat elements of different sizes

Usually, a mix of insert sizes is needed to achieve optimal outcome

Paired-end scaffolding tools

Paired-End Sequencing Tools:

Tools such as SOAPdenovo, SPAdes, and ABySS utilize paired-end sequencing data to infer the relative order and orientation of contigs. They employ algorithms that analyze the paired-end reads to estimate the distance and orientation between contigs, facilitating scaffold construction.

Mate-Pair Sequencing Tools:

Software packages like ALLPATHS-LG, MIRA, and BESST are specifically designed for mate-pair sequencing data. These tools use mate-pair information, which consists of longer DNA fragments with known distances between paired-end reads, to scaffold contigs. They employ algorithms that incorporate mate-pair information to extend contigs and bridge gaps, improving scaffold continuity.

Hi-C sequencing

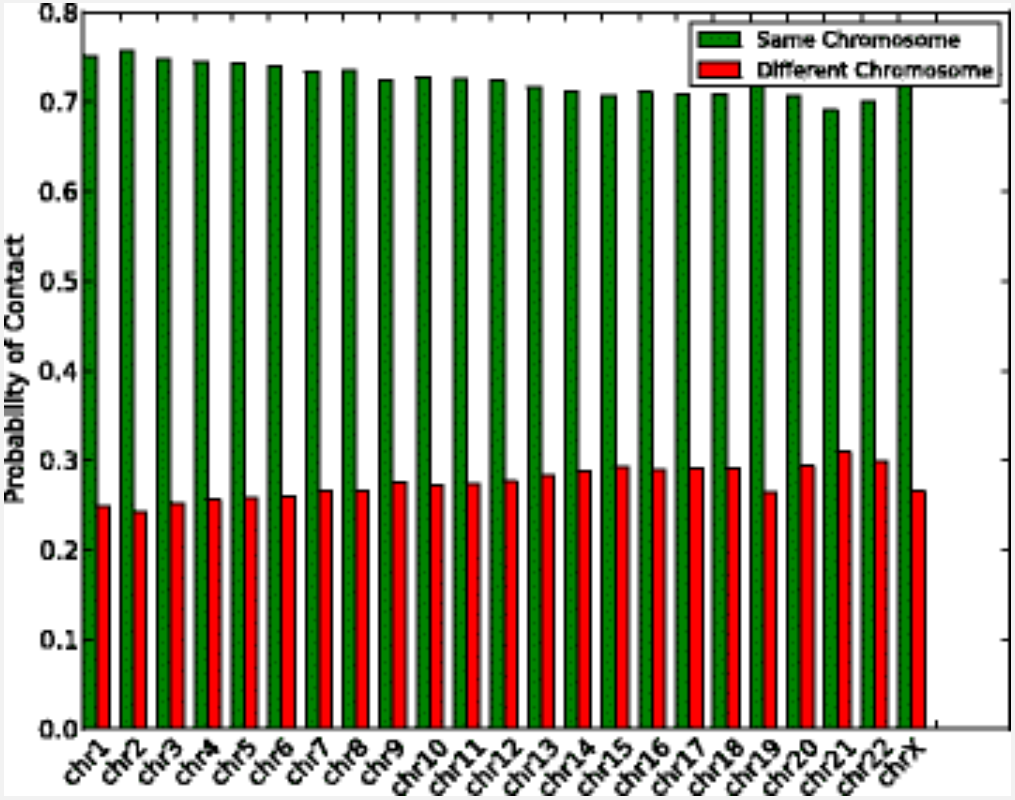
Hi-C measures the frequency (as an average over a cell population) at which two DNA fragments physically associate in 3D space, linking chromosomal structure directly to the genomic sequence

DNA near each other has more contacts

Discontinuity in Hi-C contact map along a contig suggests mis-assembly

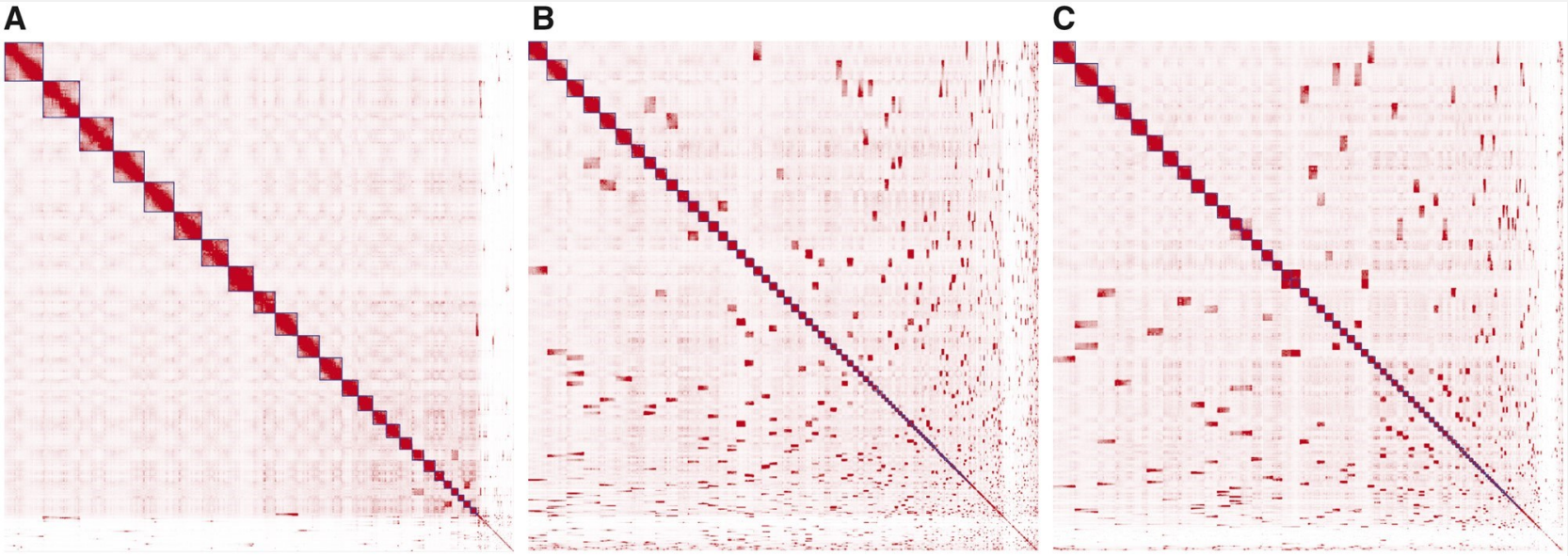
Off-diagonal contacts suggests mis-assembly

Probability of intra vs inter-chromosomal contact in Hi-C mate pairs



Churye et al., “Scaffolding for long read assemblies using long range contact information”, BMC Genomics, 18:527, 2017

Hi-C contact map



Hi-C contact maps of genome assemblies constructed with YaHS (A), SALSA2 (B) and pin_hic (C) for the simulated T2T data without contig errors. The intensity of colour indicates the density of Hi-C read pairs shared between the positions on the x- and y-axis, with darker pixels indicating higher densities. The blocks highlighted with squares along the main diagonal are scaffolds constructed by the tools. The dark off-diagonal blocks indicate scaffold pairs that could be further joined for construction of larger scaffolds. The contact maps were plotted with Juicebox

Taken from Zhou et al., "YaHS: yet another Hi-C scaffolding tool", *Bioinformatics*, 39(1):btac808, 2023

Hi-C scaffolding tools

Juicer: Juicer is a popular tool for analyzing Hi-C data. It processes raw sequencing data to generate Hi-C contact maps and offers various utilities for normalization, visualization, and scaffolding of genomes.

3D-DNA: 3D-DNA is a software package designed specifically for de novo assembly of genomes using Hi-C data. It uses a combination of proximity ligation data and sequence information to produce chromosome-scale scaffolds.

HiRise: HiRise is a component of the software package "SALSA" (Statistical Analysis of LArge-Scale chromosomal interactions) developed by the Dudchenko Lab. It is used for scaffolding genome assemblies by leveraging Hi-C data to order and orient contigs into chromosome-scale scaffolds.

GRAAL: GRAAL (Genome Rearrangement and Annotation Lite) is a tool that employs a combination of Hi-C data and other genomic features to scaffold genomes, while also providing functionalities for structural variant detection and annotation.

SALSA

Align Hi-C reads to contigs

Misjoin correction

Detect discontinuity in Hi-C contact map

Excise it from contig, splitting contig into 2

Ordering and orientation

Overlap merging

Merge contigs which overlap at their ends

Churye et al., BMC Genomics, 18:527, 2017

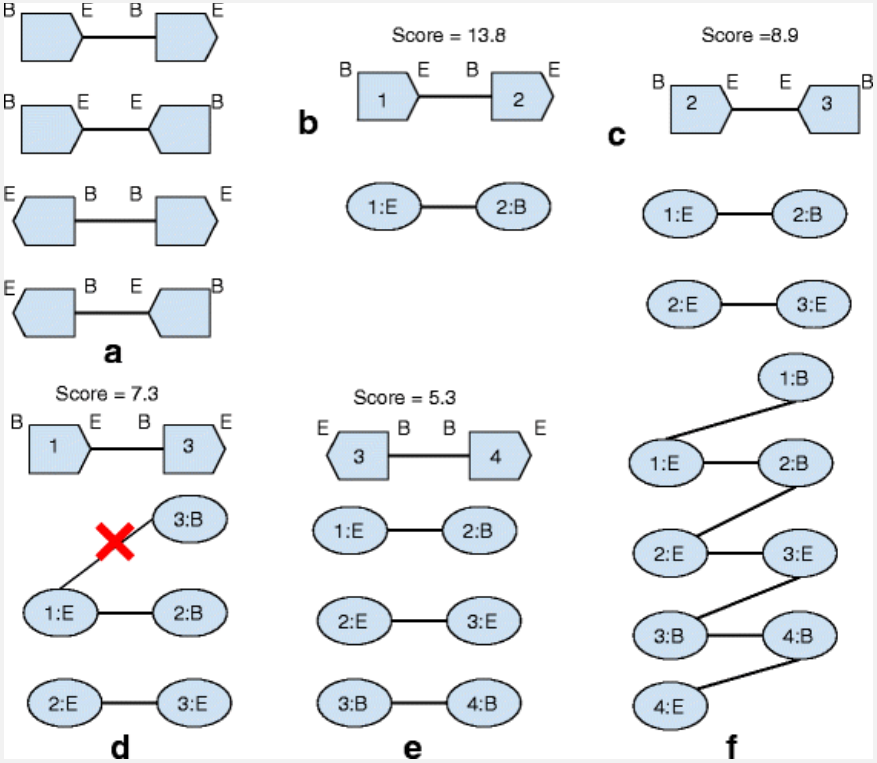
SALSA, ordering and orientating contigs

Contig has two ends (B, E)

Ends of two contigs can be connected in 4 ways (fig a)

Edge weighted by # of Hi-C read pairs mapped to region of length ℓ at ends of two contigs

Add edges greedily, remove low-weight edges to break cycles and avoid out-deg > 1



Churye et al., BMC Genomics, 18:527, 2017

Example

Metric	NA12878
Number of contigs	18903
NG50	26.83 Mb

Original NA12878 assembly

Metric	SALSA
Number of scaffolds	1555
Total bases	2.92 Gb
NG50	60.02 Mb
% Aligned bases	94.52%
Breakpoints	33079
Relocations	136
Translocations	96
Inversions	408

SALSA corrected scaffold

Churye et al., BMC Genomics, 18:527, 2017

Integrating scaffolding techniques

Take advantage of complementary info

Pair-end sequencing provides short-range contig info
Mate-pair sequencing provides longer-range info

Resolve ambiguities

Use multiple lines of evidence to disambiguate complex genomic regions (e.g., repetitive regions and genomic rearrangements)

$\Rightarrow$ Higher contiguity, higher fidelity

Good to read

Overview

Ghurye & Pop, “Modern technologies and algorithms for scaffolding assembled genomes”, *PLoS Comput Biol*, 15(6):e1006994, 2019

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6550390/

SALSA

Ghurye et al., “Scaffolding of long read assemblies using long range contact information”, BMC Genomics, 18:527, 2017

https://pubmed.ncbi.nlm.nih.gov/28701198/