CS4330: Combinatorial Methods in Bioinformatics

Genome assembly quality assessment

Wong Limsoon

NUS
National University of Singapore
National University of Singapore

Genome assembly quality

Contiguity

How contiguous the assembly is

Completeness

How much of a reference genome is covered

What fraction of a set of reference genes is covered

Correctness

How many mis-assembled segments there are

What proportion of the assembly is error free

Contiguity

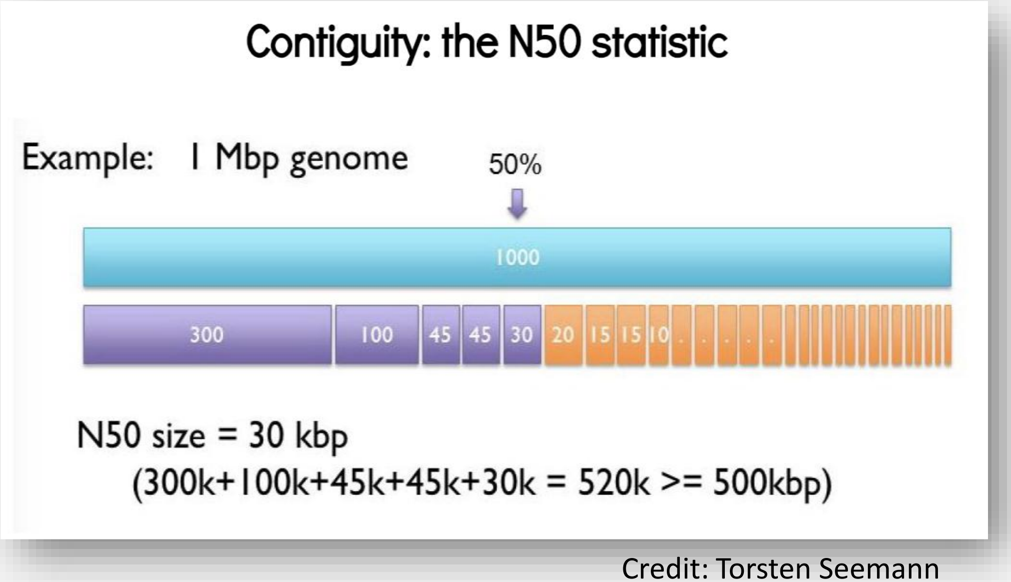
Fewer and longer contigs are desired

Metrics

Ave contig length

Max contig length

N50, NG50, NGA50, …



Contiguity: the N50 statistic

Example: 1 Mbp genome

N50 size = 30 kbp
(300k+100k+45k+45k+30k = 520k >= 500kbp)

Credit: Torsten Seemann

Completeness

Proportion of original genome represented by the assembly

$$\frac{\text{Assembled genome size}}{\text{Estimated genome size}}$$

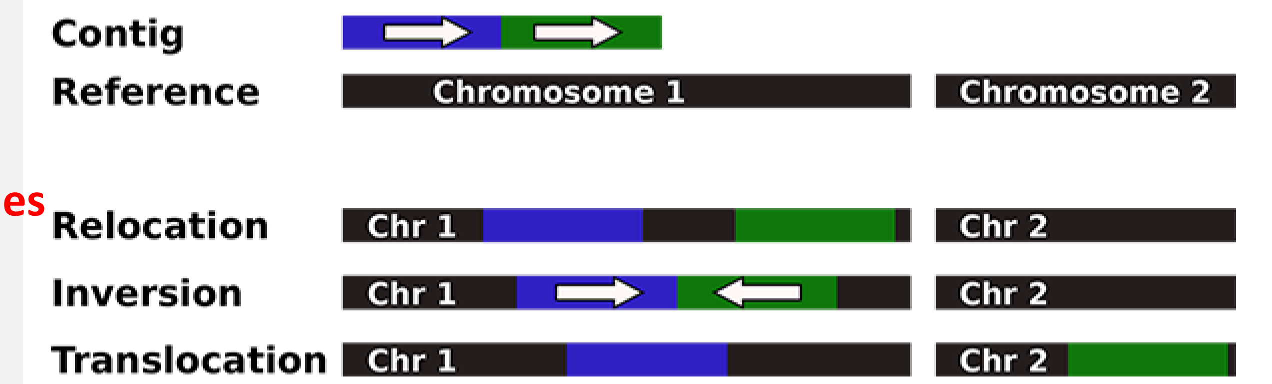
Proportion of core genes covered

$$\frac{\text{\# of core genes in assembly}}{\text{\# of core genes known}}$$

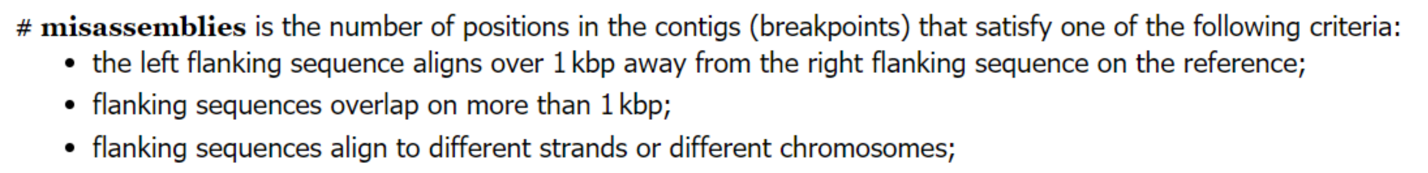
Correctness

Proportion of assembly that is error free

Kinds of Mis-assemblies



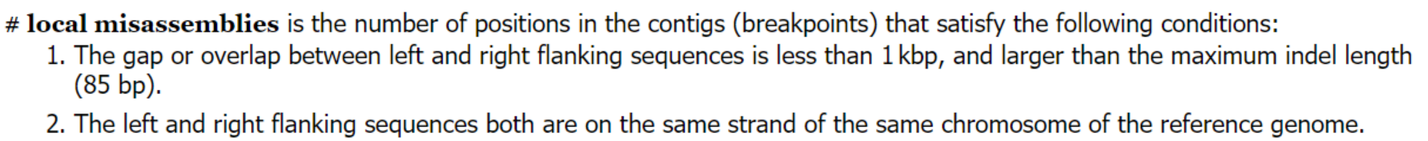
#, not size



misassemblies is the number of positions in the contigs (breakpoints) that satisfy one of the following criteria:

- the left flanking sequence aligns over 1 kbp away from the right flanking sequence on the reference;
- flanking sequences overlap on more than 1 kbp;
- flanking sequences align to different strands or different chromosomes;

#, not size



local misassemblies is the number of positions in the contigs (breakpoints) that satisfy the following conditions:

1. The gap or overlap between left and right flanking sequences is less than 1 kbp, and larger than the maximum indel length (85 bp).
2. The left and right flanking sequences both are on the same strand of the same chromosome of the reference genome.

Credit: QUAST user manual

Exercise

Some “mis-assemblies” may not be mis-assemblies

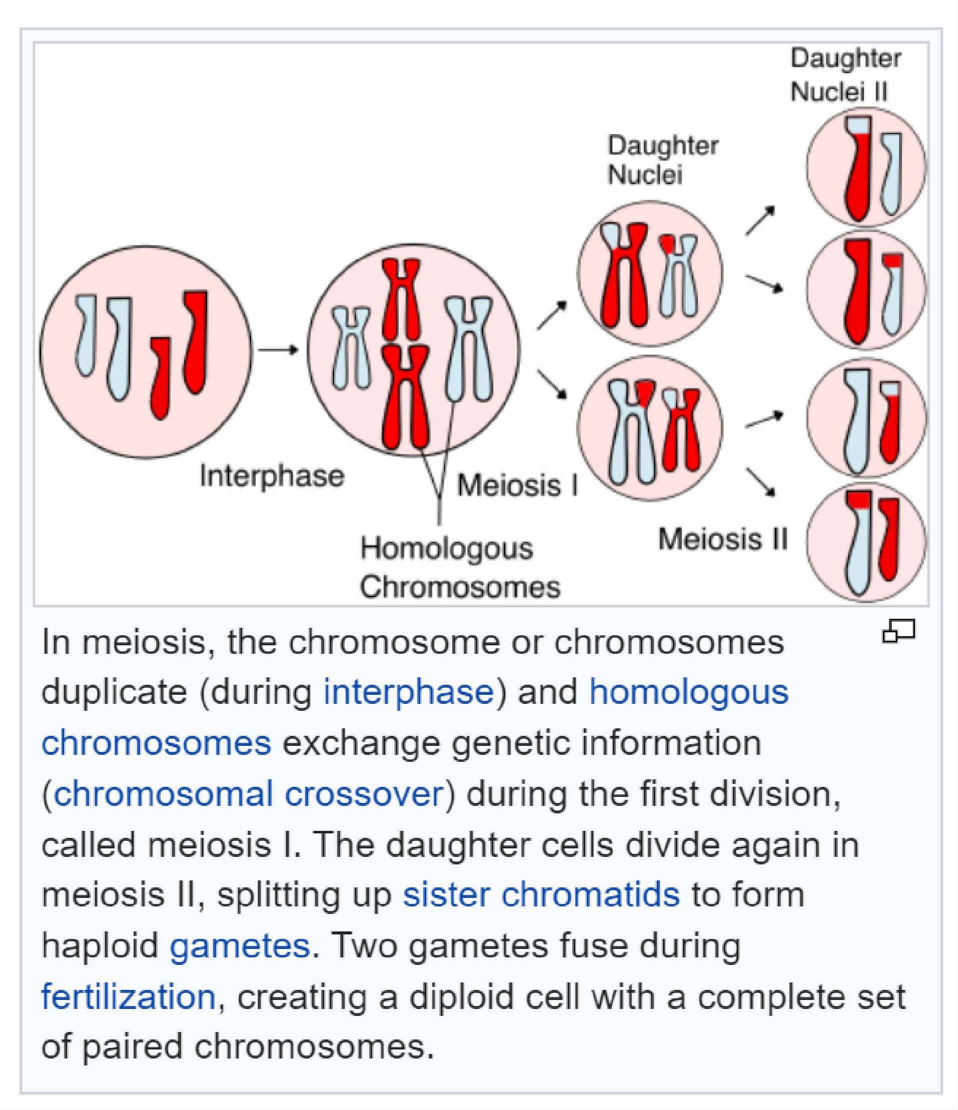
Why?

Exercise

Identify some issues with genome assembly quality measures such as NG50, # mis-assemblies, etc.

Law of genetic linkage

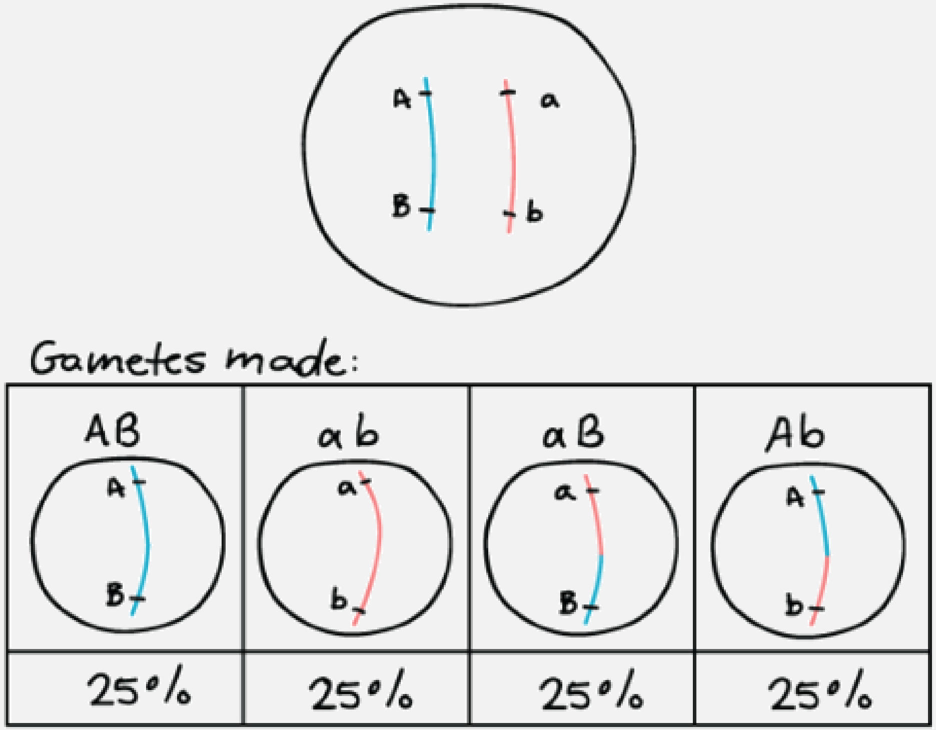
Meiosis



In meiosis, the chromosome or chromosomes duplicate (during interphase) and homologous chromosomes exchange genetic information (chromosomal crossover) during the first division, called meiosis I. The daughter cells divide again in meiosis II, splitting up sister chromatids to form haploid gametes. Two gametes fuse during fertilization, creating a diploid cell with a complete set of paired chromosomes.

Image credit: Wikipedia

When two genes are far apart, this is what happens during meiosis

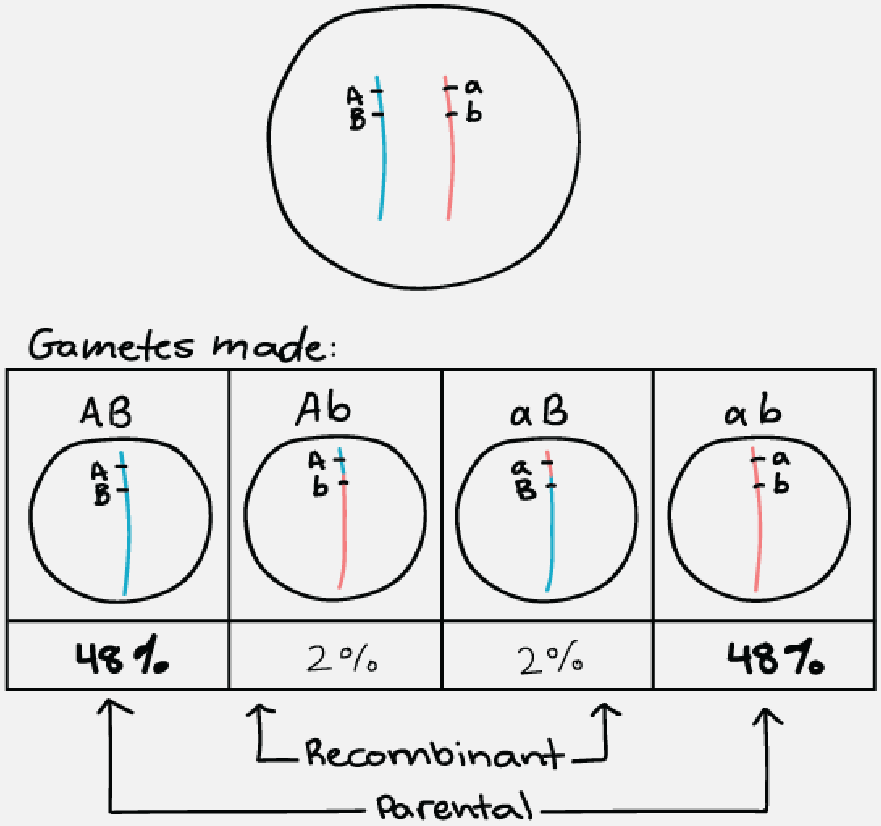


Gametes made:

AB	ab	aB	Ab
25%	25%	25%	25%

Image credit: Khan Academy

When two genes are close together, this is what happens during meiosis



Gametes made:

AB	Ab	aB	ab
48%	2%	2%	48%

Image credit: Khan Academy

Law of genetic linkage

The closer two genes / genomic loci are, the more likely they are passed on to the next generation together

Genome assembly assessment: *Does the assembly allow us to estimate the distance between two loci on ref genome well?*

Genome assembly improvement: *Do two close-by / far-apart loci on the assembly look like they should be close-by / far-apart on ref genome?*

Integrative genome assembly quality assessment

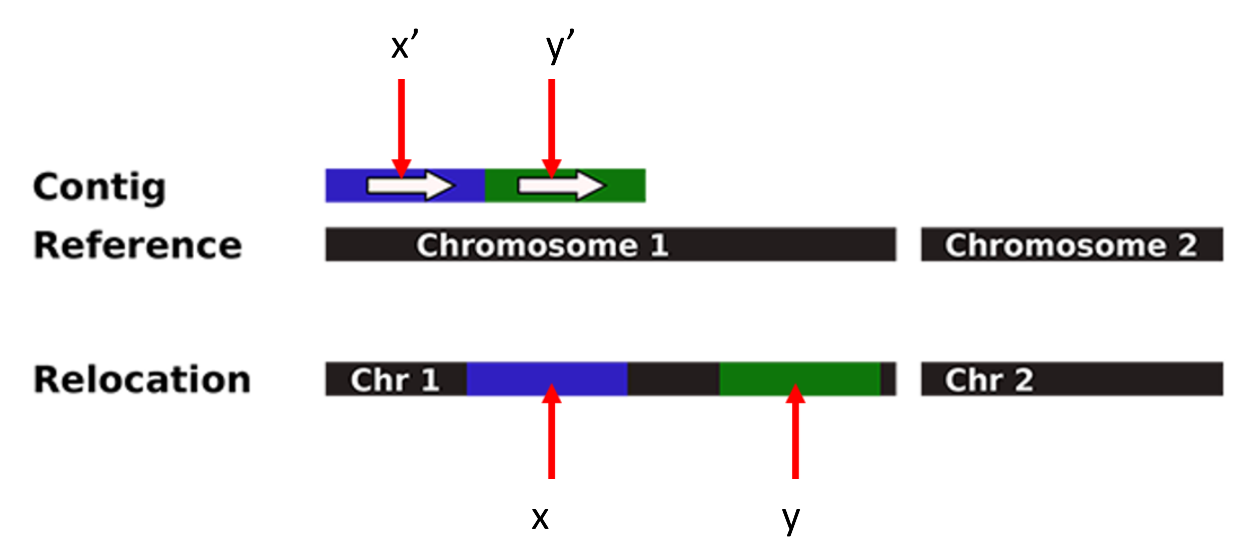
Pairwise distance reconstruction, PDR

Sites on assembly mapped to xy

Sites on ref genome

Size of ref genome

$$PDR = (\Sigma_{xy} \text{Min}(d_{x'y'}, d_{xy}) / \text{Max}(d_{x'y'}, d_{xy})) / G^2$$



Xie & Wong, “PDR: A new genome assembly evaluation metric based on genetics concerns”, *Bioinformatics*, 37(3):289-295, 2021

Intuition of PDR

PDR is designed to answer a basic biology question:

How accurately can the distance of two positions on a genome be determined from the assembly?

PDR integrates contiguity

Smaller contigs make PDR smaller

(x, y) on same chromosome

(x', y') on different contigs

$\Rightarrow d_{xy}$ *is small but* $d_{x'y'} = \infty$

PDR integrates completeness

More missed loci make PDR smaller

(x, y) on same chromosome

x' or y' not on any contig

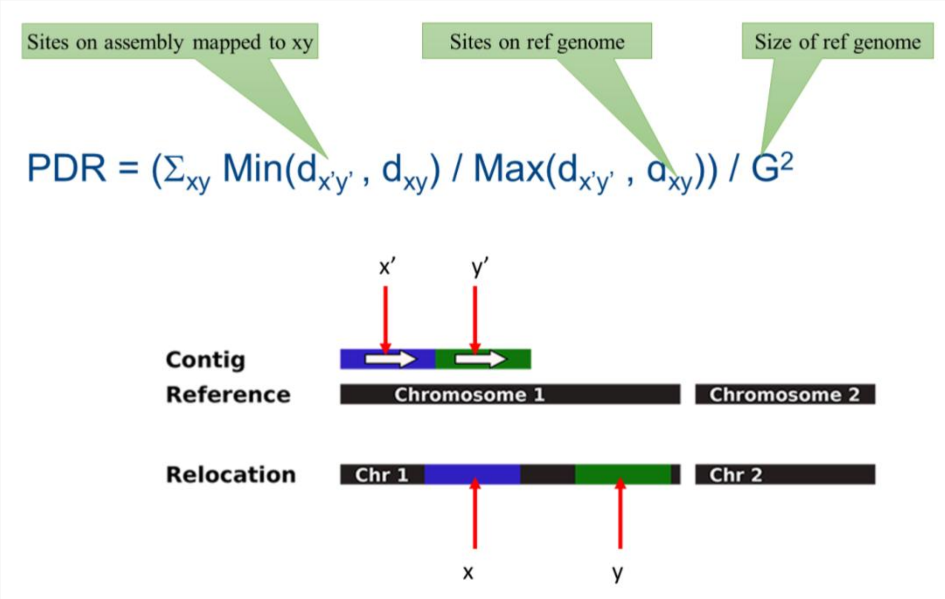
$\Rightarrow d_{xy}$ *is small but* $d_{x'y'} = \infty$

PDR integrates correctness

When a larger genome segment is mis-assembled, $\min(d_{x'y'}, d_{xy})$ is more different from $\max(d_{x'y'}, d_{xy})$

$\Rightarrow$ Make PDR smaller

# and size of mis-assemblies are accounted



Sites on assembly mapped to xy

Sites on ref genome

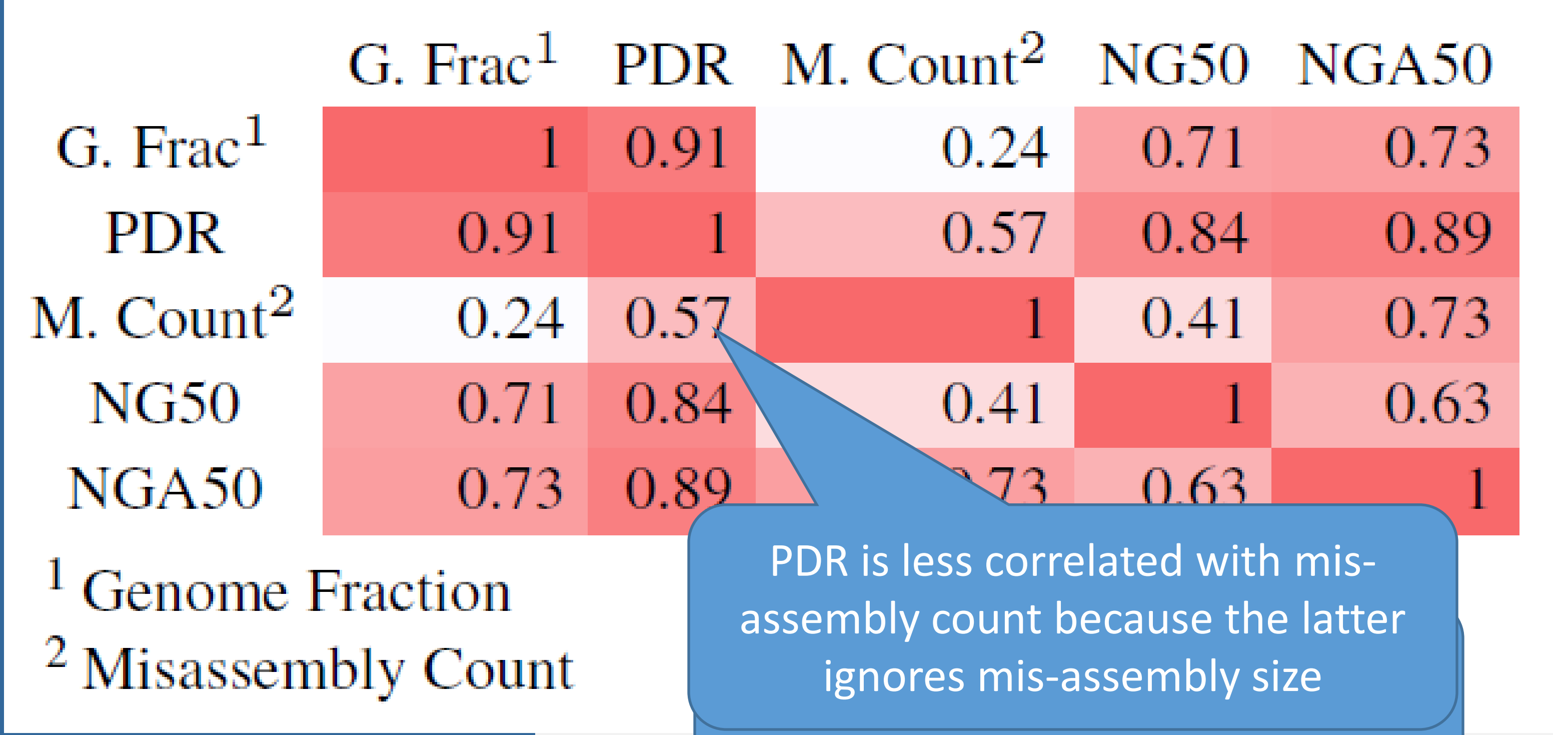
Size of ref genome

$$\text{PDR} = \left(\Sigma_{xy} \text{Min}(d_{x'y'}, d_{xy}) / \text{Max}(d_{x'y'}, d_{xy})\right) / G^2$$

Correlation to contiguity, completeness, & correctness

Dataset	Worm
Genome size (bp)	100.3M
Sequencing platform	Illumina pair-ends and PacBio SMRT
Assemblers	Upperbound, Canu, FALCON, Flye, MaSuRCA, Miniasm

E. Coli dataset from QUAST-LG benchmark



	G. Frac[1]	PDR	M. Count[2]	NG50	NGA50
G. Frac[1]	1	0.91	0.24	0.71	0.73
PDR	0.91	1	0.57	0.84	0.89
M. Count[2]	0.24	0.57	1	0.41	0.73
NG50	0.71	0.84	0.41	1	0.63
NGA50	0.73	0.89	0.73	0.63	1

[1] Genome Fraction

[2] Misassembly Count

PDR is less correlated with mis-assembly count because the latter ignores mis-assembly size

to each other

Computing PDR naively is costly

$$PDR = \left(\Sigma_{xy} \mathrm{Min}(d_{x'y'}, d_{xy}) / \mathrm{Max}(d_{x'y'}, d_{xy})\right) / G^2$$

(x,y) ranges over all possible pairs of loci on a genome.
There are $(3{,}000{,}000{,}000)^2$ pairs on the human genome

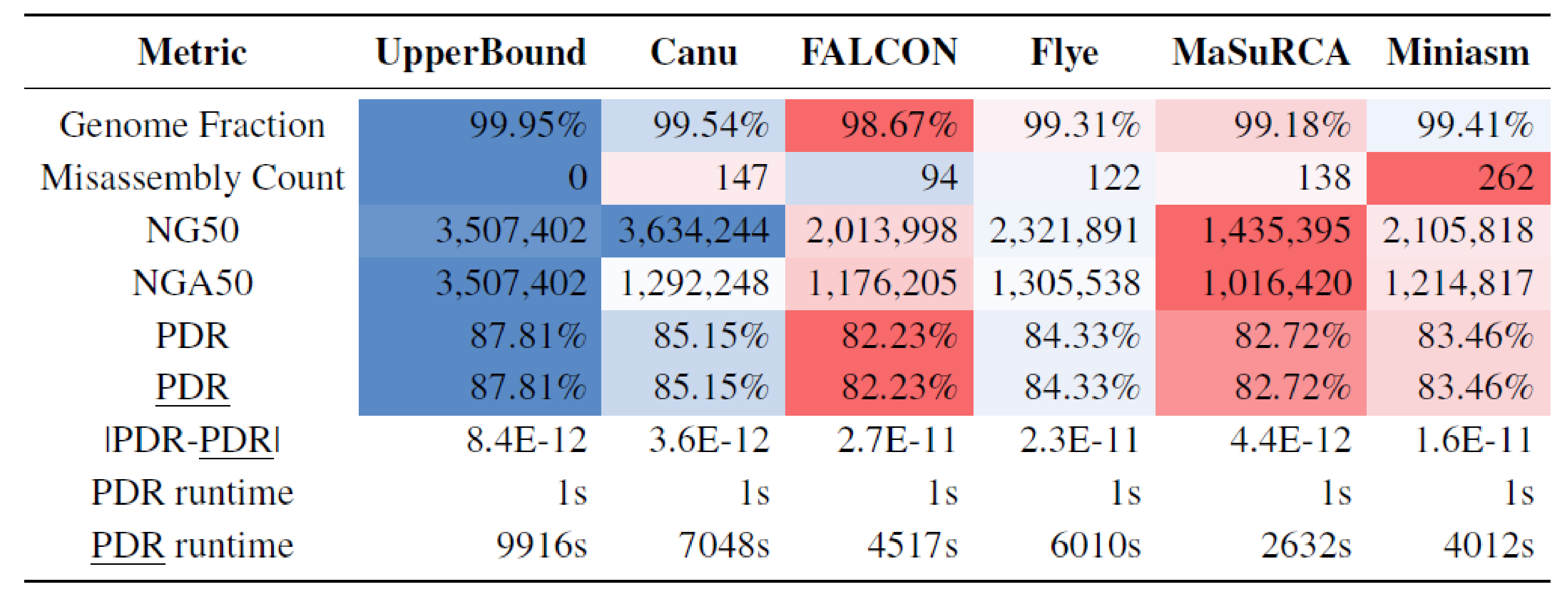
But it can be optimized

Approximate it piece-wise by integrals of “segment” pairs

Segment pair: A segment of contiguous loci on the reference genome that is mapped to a segment of contiguous positions on a contig in the assembly

Xie & Wong, *Bioinformatics*, 37(3):289-295, 2021

Accurate thousand-fold speed-up of PDR computation

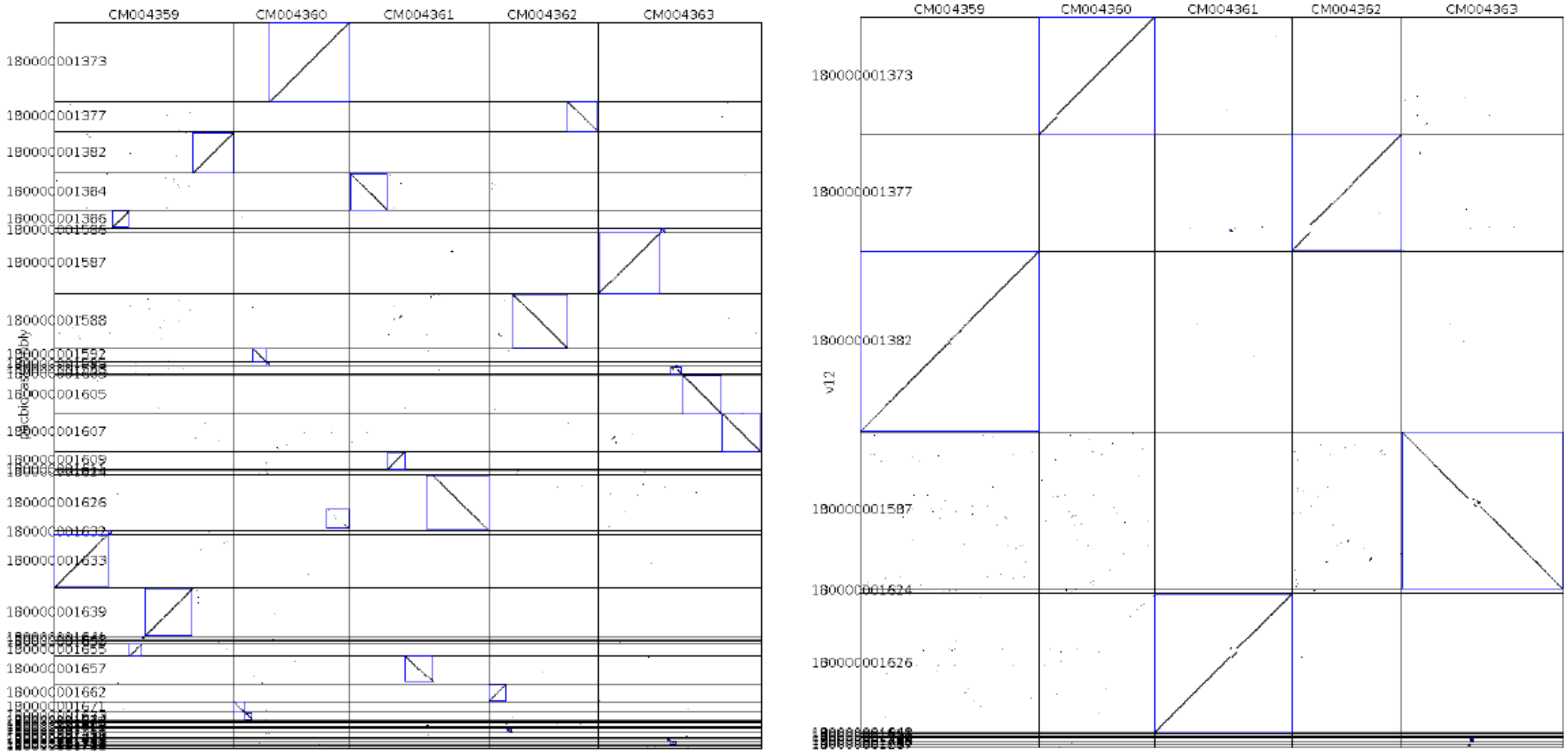


Metric	UpperBound	Canu	FALCON	Flye	MaSuRCA	Miniasm
Genome Fraction	99.95%	99.54%	98.67%	99.31%	99.18%	99.41%
Misassembly Count	0	147	94	122	138	262
NG50	3,507,402	3,634,244	2,013,998	2,321,891	1,435,395	2,105,818
NGA50	3,507,402	1,292,248	1,176,205	1,305,538	1,016,420	1,214,817
PDR	87.81%	85.15%	82.23%	84.33%	82.72%	83.46%
$\underline{\text{PDR}}$	87.81%	85.15%	82.23%	84.33%	82.72%	83.46%
$\lvert\text{PDR}-\underline{\text{PDR}}\rvert$	8.4E-12	3.6E-12	2.7E-11	2.3E-11	4.4E-12	1.6E-11
PDR runtime	1s	1s	1s	1s	1s	1s
$\underline{\text{PDR}}$ runtime	9916s	7048s	4517s	6010s	2632s	4012s

~1hr to compute naively for E. coli

~1s to compute by piece-wise integrals, with approximation error ~10^{-11}

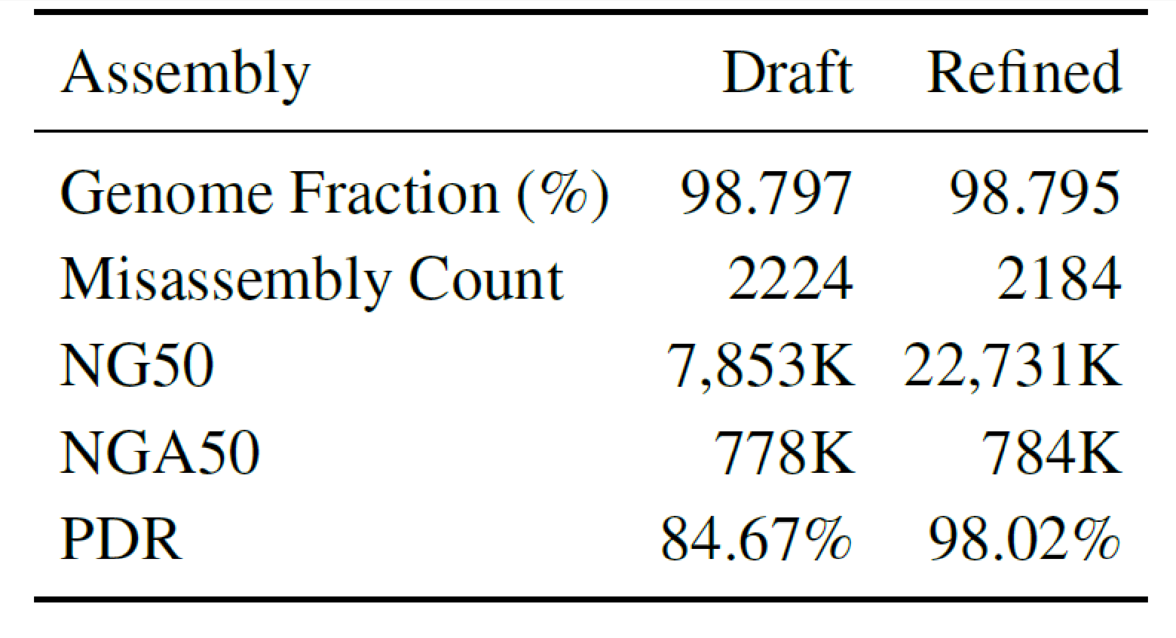
Two assemblies of a *A. thaliana* genome



(a) Draft assembly

(b) Refined assembly

A convincing test of PDR



Assembly	Draft	Refined	
Genome Fraction (%)	98.797	98.795	0% diff
Misassembly Count	2224	2184	2% diff
NG50	7,853K	22,731K	189% diff
NGA50	778K	784K	1% diff
PDR	84.67%	98.02%	15% diff

PDR shows the *A. thaliana* refined assembly is near perfect and more reasonable diff from the draft assembly

Other measures show less informative differences

Good to read

[QUAST] Gurevich et al., “QUAST: quality assessment tool for genome assemblies”, *Bioinformatics*, 29(8):1072-1075, 2013

https://pubmed.ncbi.nlm.nih.gov/23422339/

[PDR] Xie & Wong, “PDR: A new genome assembly evaluation metric based on genetics concerns”, *Bioinformatics*, 37(3):289-295, 2021

https://pubmed.ncbi.nlm.nih.gov/32761066/