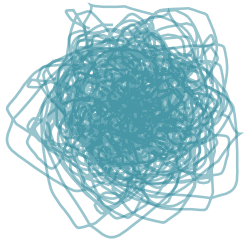
@ColauttiLab

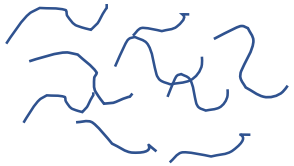
BIOL 432

De novo assembly

Whole Genome Shotgun (WGS) assembly



Extract

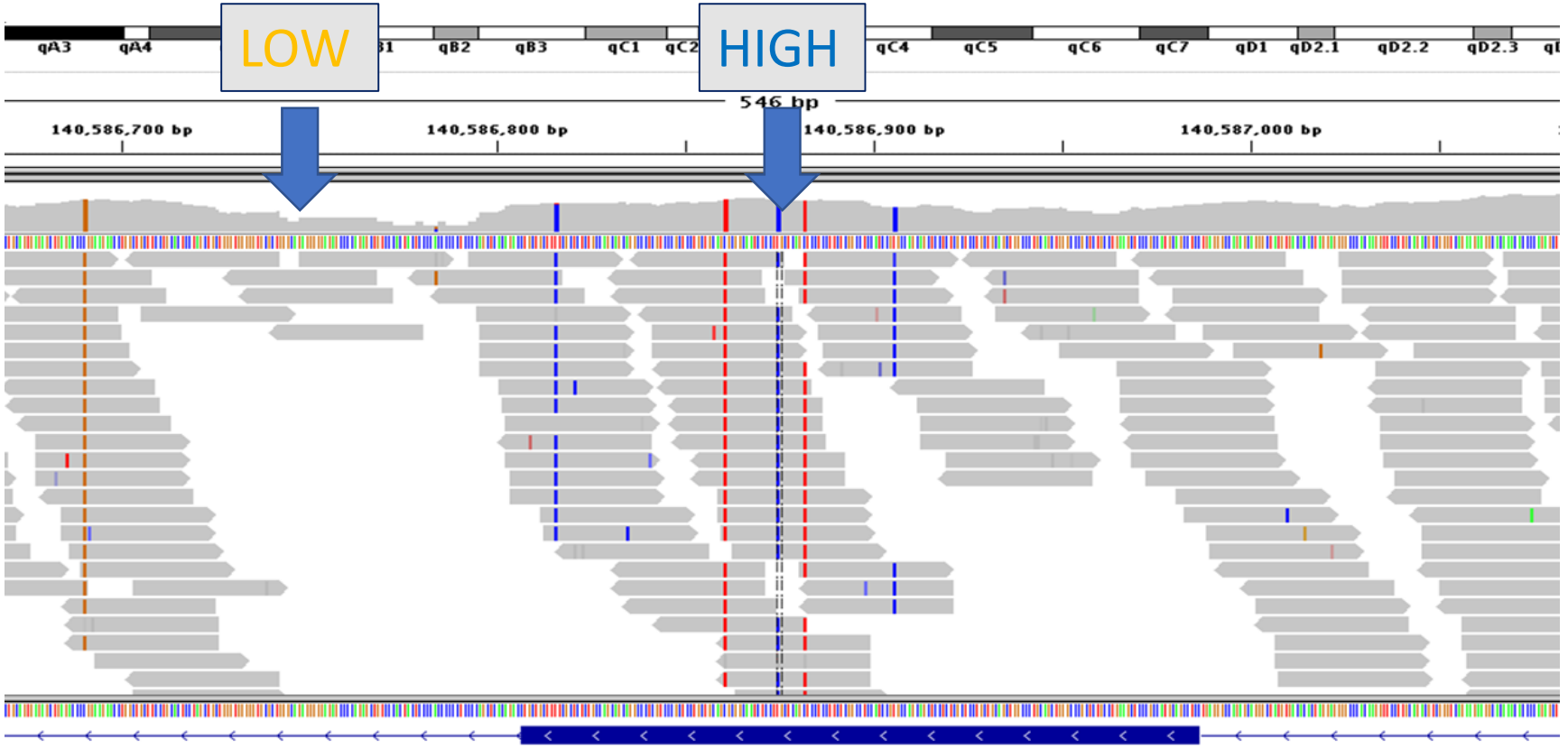


Fragment

Sequence

Assemble

Alignment coverage



- = average # reads at any given location (base pair)
- A 'random' sampling process
- (Higher is better)

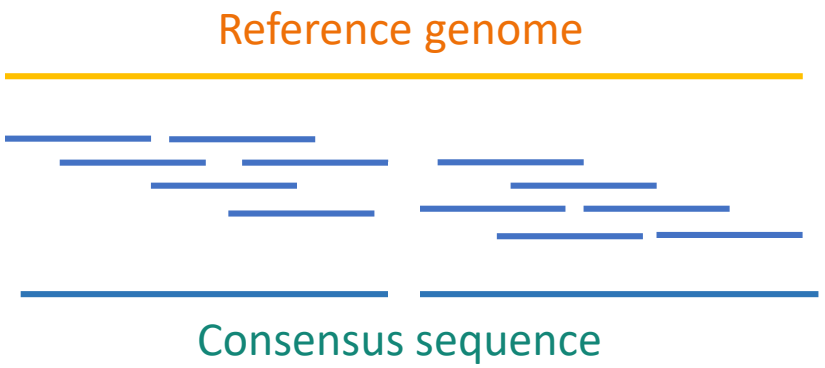
WGS without a reference genome

@ColauttiLab

Queen's University

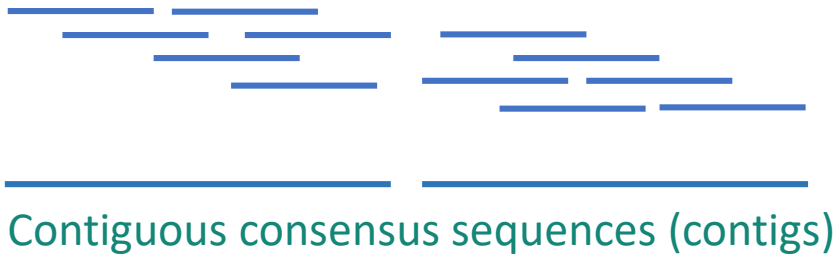
Alignment to reference

- sometimes called 'resequencing'
- common in 'model systems'
- e.g. human, mouse*, Drosophila, C. elegans*, *Arabidopsis*



de novo assembly

- 'non-model' organisms

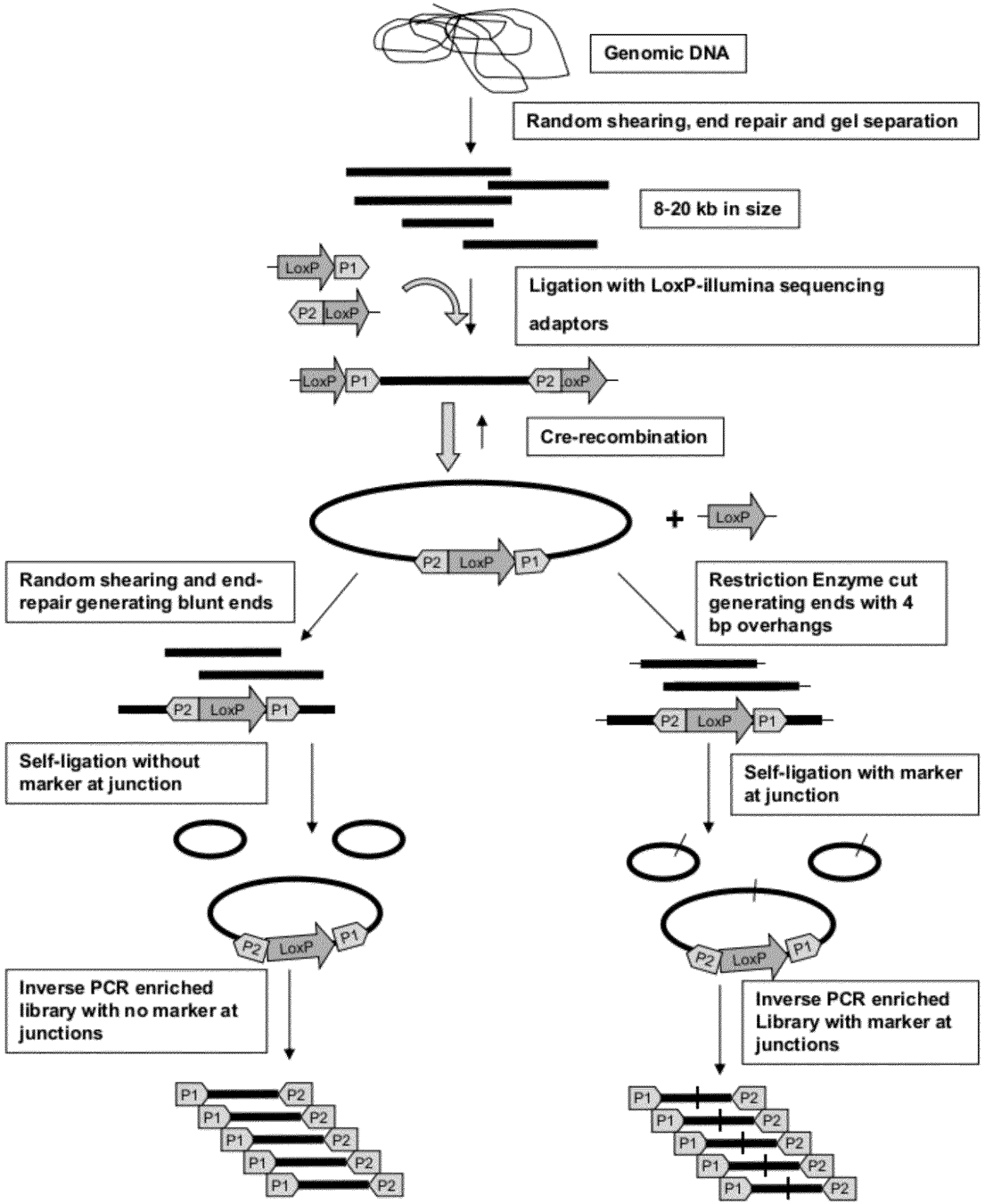


‘Paired End’ vs ‘Mate Pair’

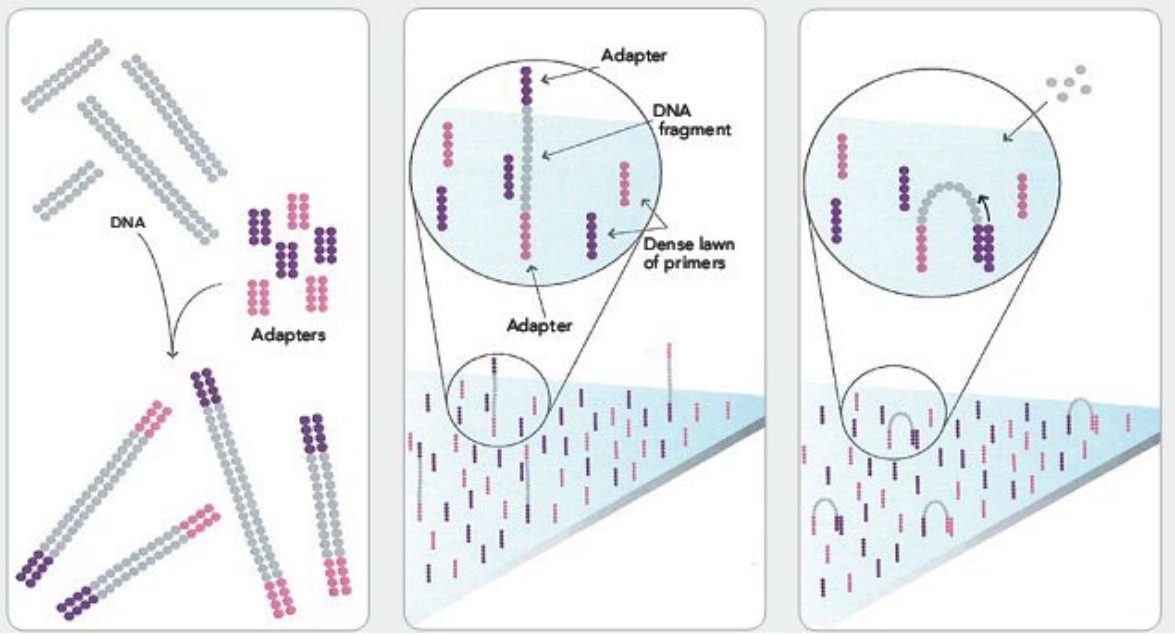
Figure 5. *De Novo* Assembly with Mate Pairs

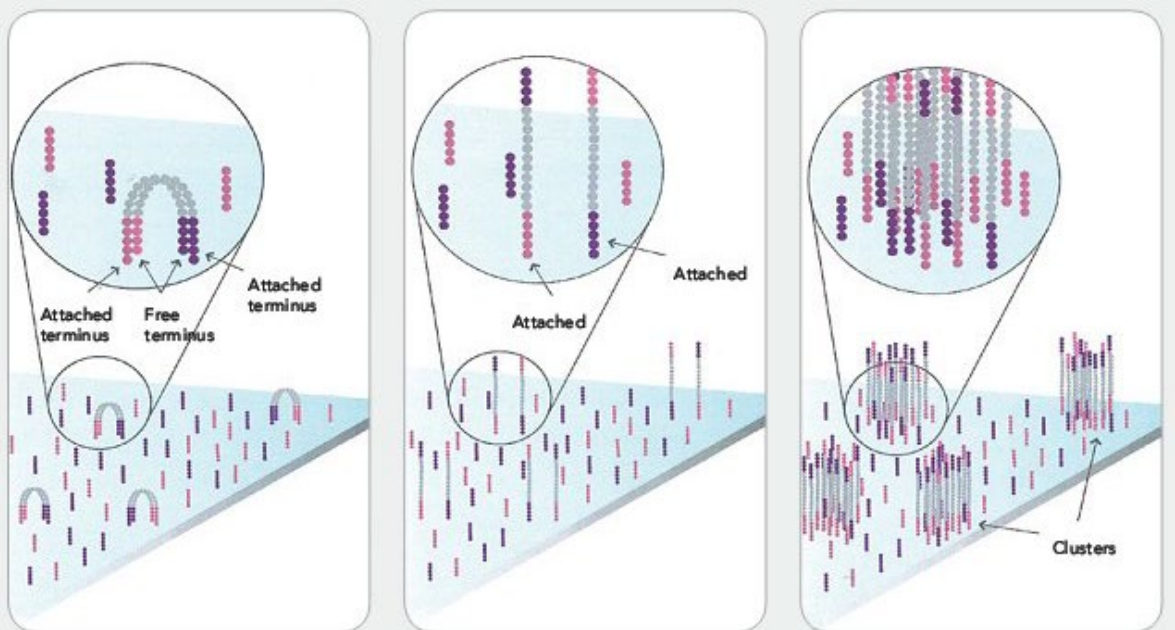
Using a combination of short and long insert sizes with paired-end sequencing results in maximal coverage of the genome for de novo assembly. Because larger inserts can pair reads across greater distances, they provide a better ability to read through highly repetitive sequences and regions where large structural rearrangements have occurred. 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.

PE protocol


Genomic DNA
Random shearing, end repair and gel separation
8-20 kb in size
LoxP
P1
P2
LoxP
Ligation with LoxP-illumina sequencing adaptors
Cre-recombination
+
Random shearing and end-repair generating blunt ends
Restriction Enzyme cut generating ends with 4 bp overhangs
Self-ligation without marker at junction
Self-ligation with marker at junction
Inverse PCR enriched library with no marker at junctions
Inverse PCR enriched Library with marker at junctions

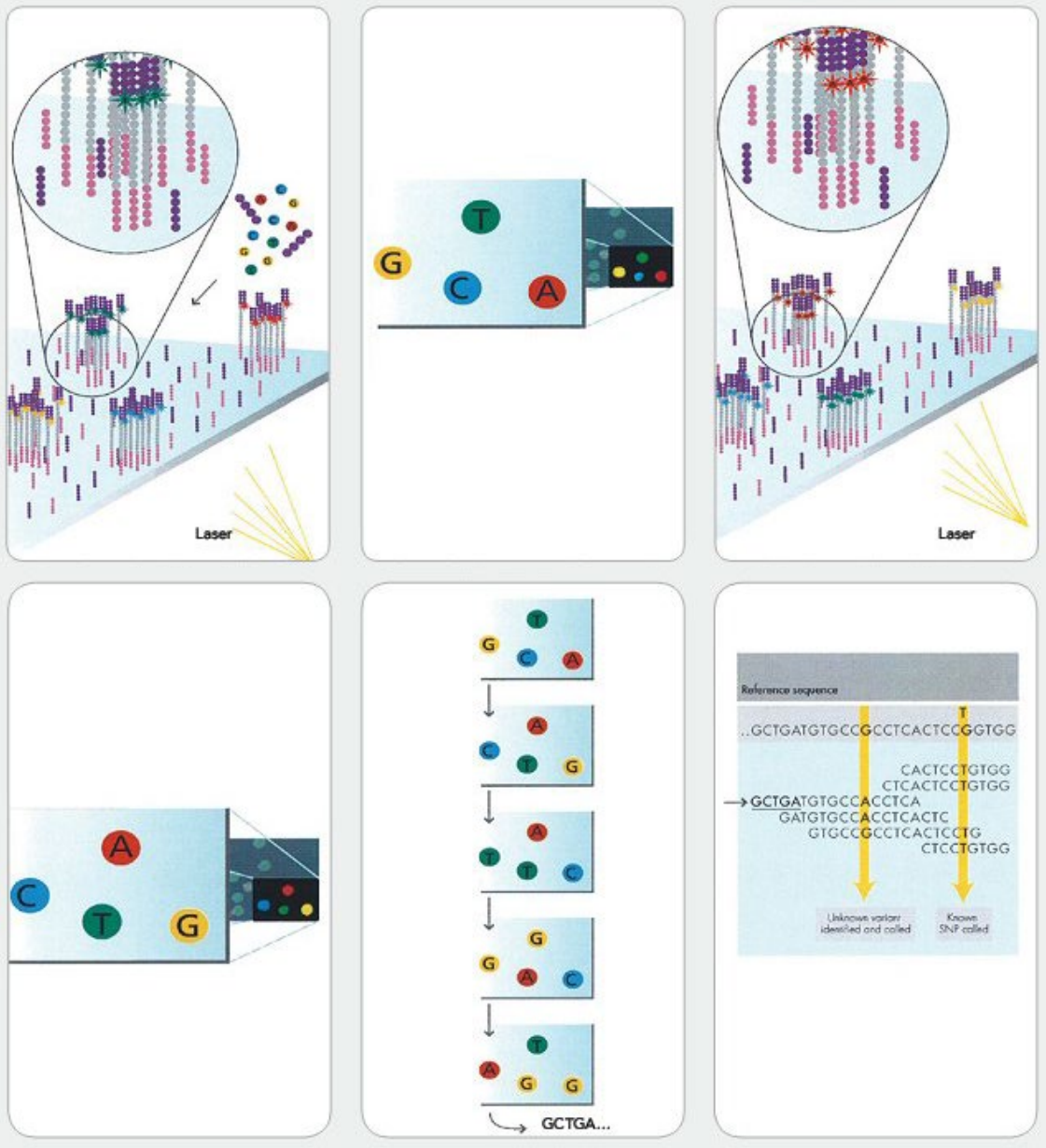
Illumina sequencing (formerly Solexa)





1. Prepare genomic DNA
2. Attach DNA to surface
3. Bridge amplification
4. Fragment become double stranded
5. Denature the double stranded molecules
6. Complete amplification

Illumina sequencing



7. Determine first base
8. Image first base
9. Determine second base
10. Image second base
11. Sequence reads over multiple cycles
12. Align data

A few de novo assemblers

- ABySS
- ALLPATHS-LG
- CORTEX
- CLC Genomics Workbench
- DISCOVAR de novo
- Geneious
- IDBA
- MaSuRCA
- MIRA
- PLATANUS
- RAY
- SOAP de novo

Group Activity:

Assemble a genome by hand

What are the ‘read lengths’, ‘coverage’ and ‘genome size’ (aka assembly size) of your genome assembly?

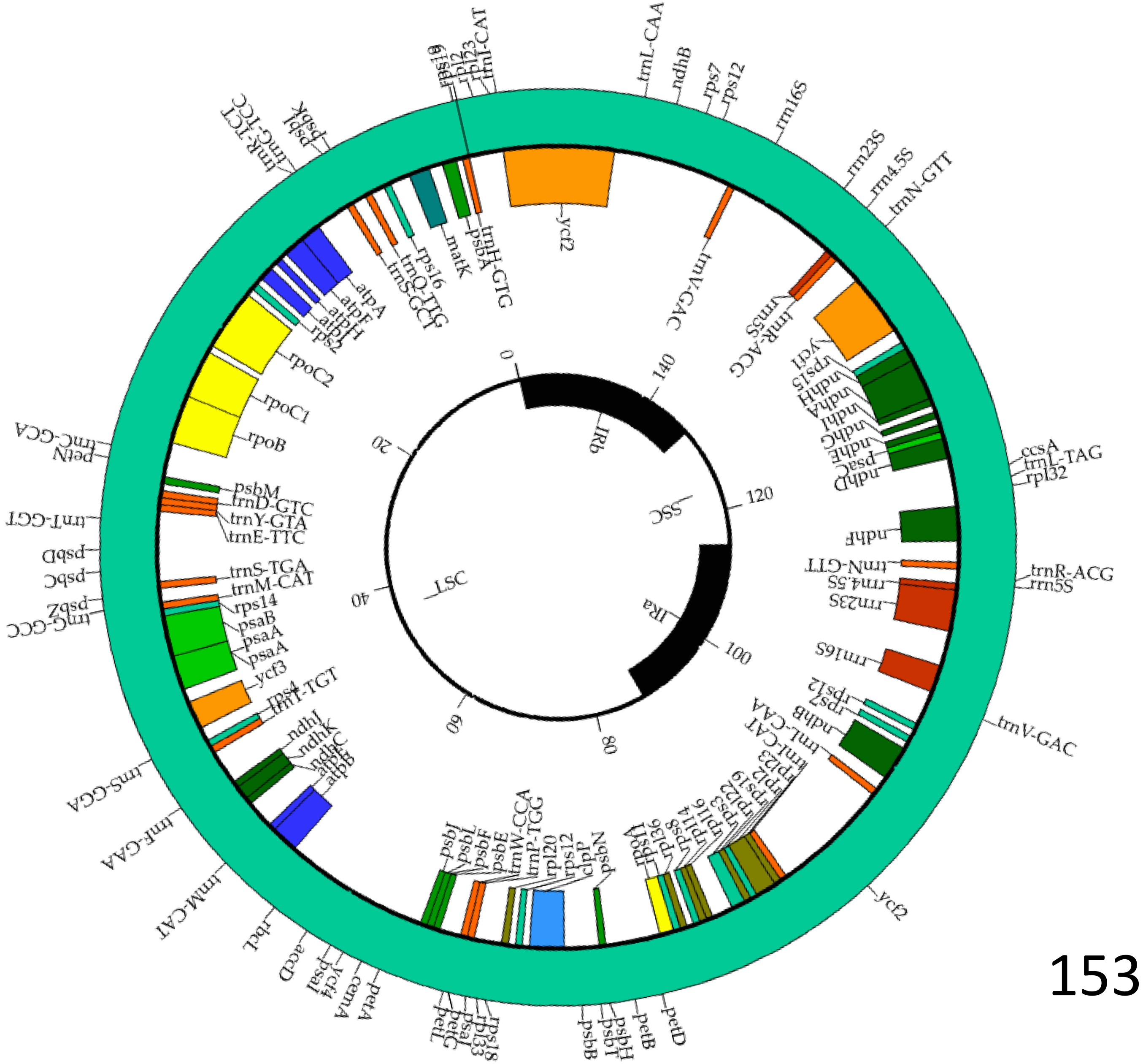
Why/how would these three factors affect the quality of the genome assembly?

What else might affect assembly quality?

Assemble the garlic mustard chloroplast genome

@ColauttiLab

Queen's University



153,190bp