CS4330: Combinatorial Methods in Bioinformatics

Correction & scaffolding for progeny sequencing projects

Wong Limsoon

NUS
National University of Singapore
National University of Singapore

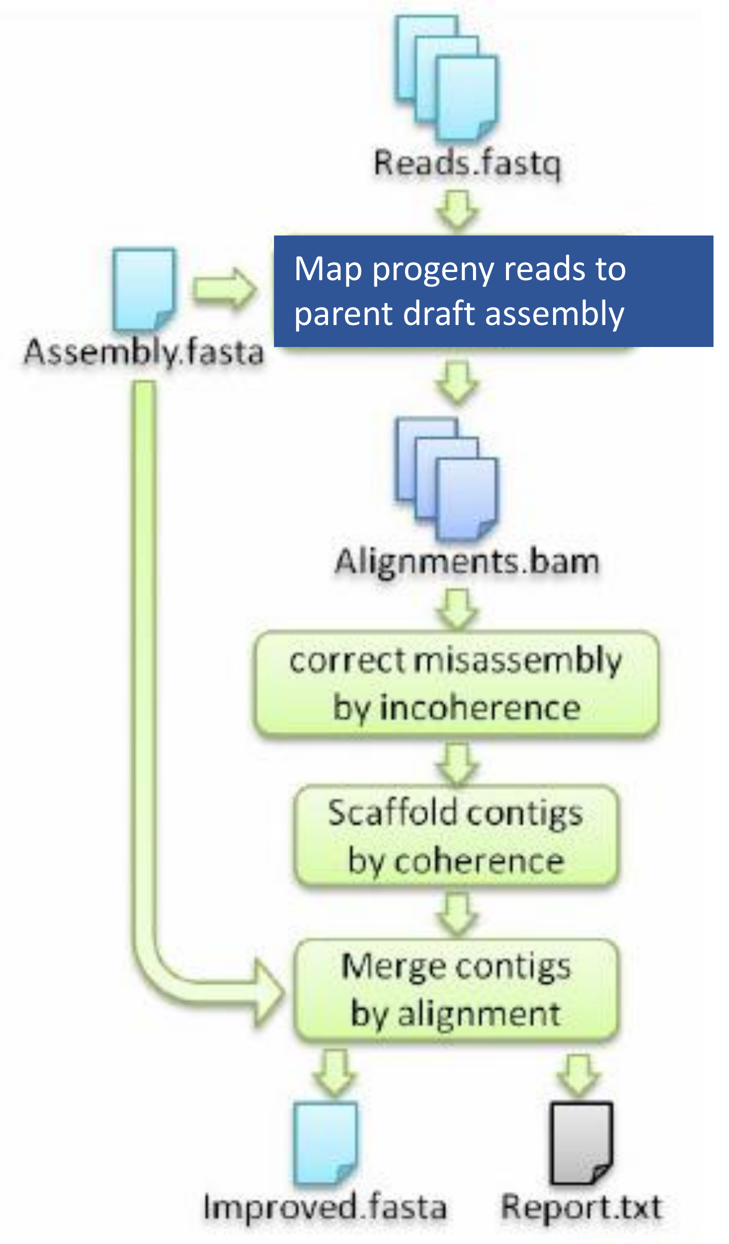
A law of genetic linkage

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

Observation: Two genomic loci have alleles that are (not) highly correlated in closely related strains (e.g. progeny data)

Abduction: The two genomic loci are (not) close together on the same chromosome

CAST, a correction and scaffolding tool for progeny sequencing projects


Reads.fastq
Map progeny reads to parent draft assembly
Assembly.fasta
Alignments.bam
correct misassembly by incoherence
Scaffold contigs by coherence
Merge contigs by alignment
Improved.fasta
Report.txt

Polymorphic positions

Check adjacent polymorphic positions for correlation in progenies

Position is polymorphic if progenies differ on this position

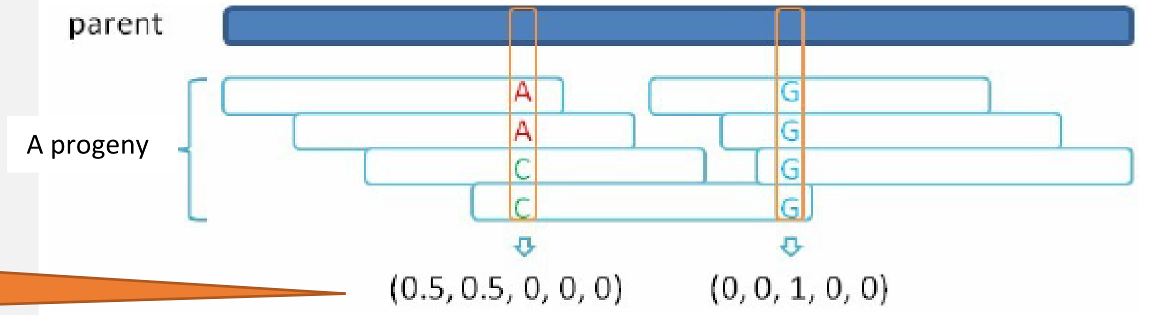


Fig. 2. Examples of genotyping for one progeny. First, reads of this progeny are mapped to the parent's contigs. Then, each column in the alignment is considered. The genotypes of two columns in orange boxes are ratios, in order, of base A, C, G, T, and gap.

Nice way to handle polyploids

Recall law of genetic linkage

Distance at a genomic loci betw the genotypes of two progenies

The distance D_{xjk} betw the genotypes of progeny j & progeny k at a locus x is defined as half the Manhattan distance of their genotypes at this locus

Manhattan distance

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

Example

Progeny j has genotype (0, 1, 0, 0, 0) @ x

Progeny k has genotype (0.5, 0.5, 0, 0, 0) @ x

D_{xjk} = (0.5 + 0.5) / 2 = 0.5

Split by incoherence

Split contig if correlation (in progenies) between adjacent polymorphic positions is low; i.e. mis-assembly is assumed

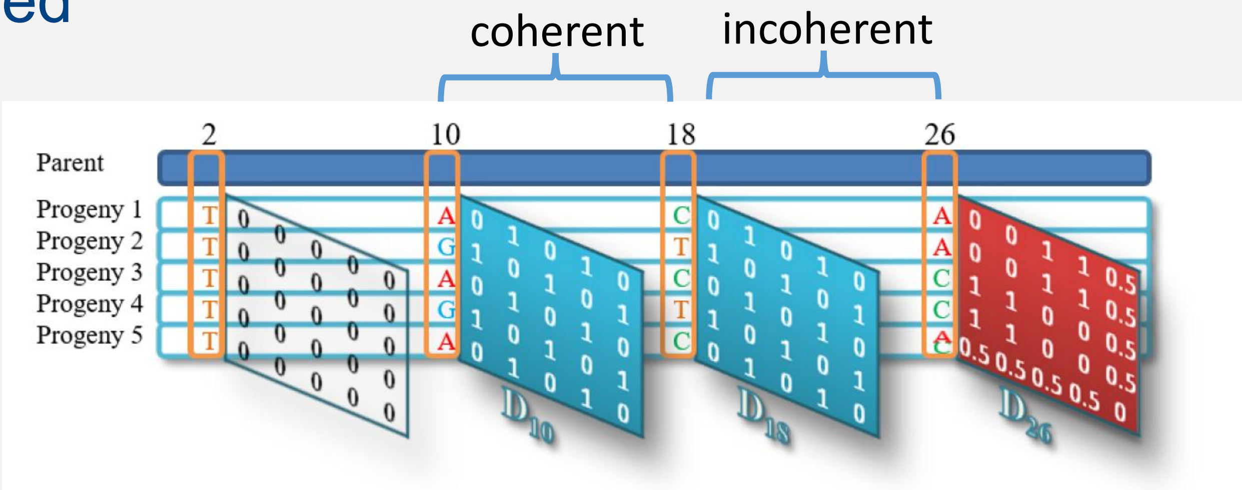


Figure 3.3: Examples of distance matrices. The dark blue bar represents a parent contig, while each cyan bar here is not a read but genotype information of one progeny from its reads. At position 2, all progenies have same genotype (0,0,0,1,0), thus D_2 is a zero matrix (transparent). At position 10, three progenies have (1,0,0,0,0) while another two progenies have (0,0,1,0,0). At position 18, three progenies have (0,1,0,0,0) while another two progenies have (0,0,0,1,0). D_{10} and D_{18} have the same color because they are similar. At position 26, progeny 5 has genotype (0.5,0.5,0,0,0). D_{26} is differentially colored as it has different pattern from D_{18} and D_{10}.

Merge by coherence

Compare all possible pairs of (split) contigs

Select two (split) contigs whose ends are coherent to each other for merging

i.e., the last polymorphic position on the right end of one contig and the first polymorphic position on the left end of the other contig have high correlation in progenies

Merge by overlapping aligned regions

Recall law of genetic linkage

Accounting for sequencing noise

L_i = avg of polymorphic positions within H bases (= 0.1% of genome length) on the left of the i-th polymorphic position

$$L_i = Avg\{D_x | i - H \leq x \leq i\}$$

R_i = avg of polymorphic positions within H bases (= 0.1% of genome length) on the right of the i-th polymorphic position

$$R_i = Avg\{D_x | i \leq x \leq i + H\}$$

$\delta(L_p, R_q) = \Sigma_{i,j} \mid L_p[i,j] - R_q[i,j] \mid$

Null distribution generated by permuting progeny labels

Small δ implies genetic linkage between adjacent positions p & q

Accounting for sequencing noise

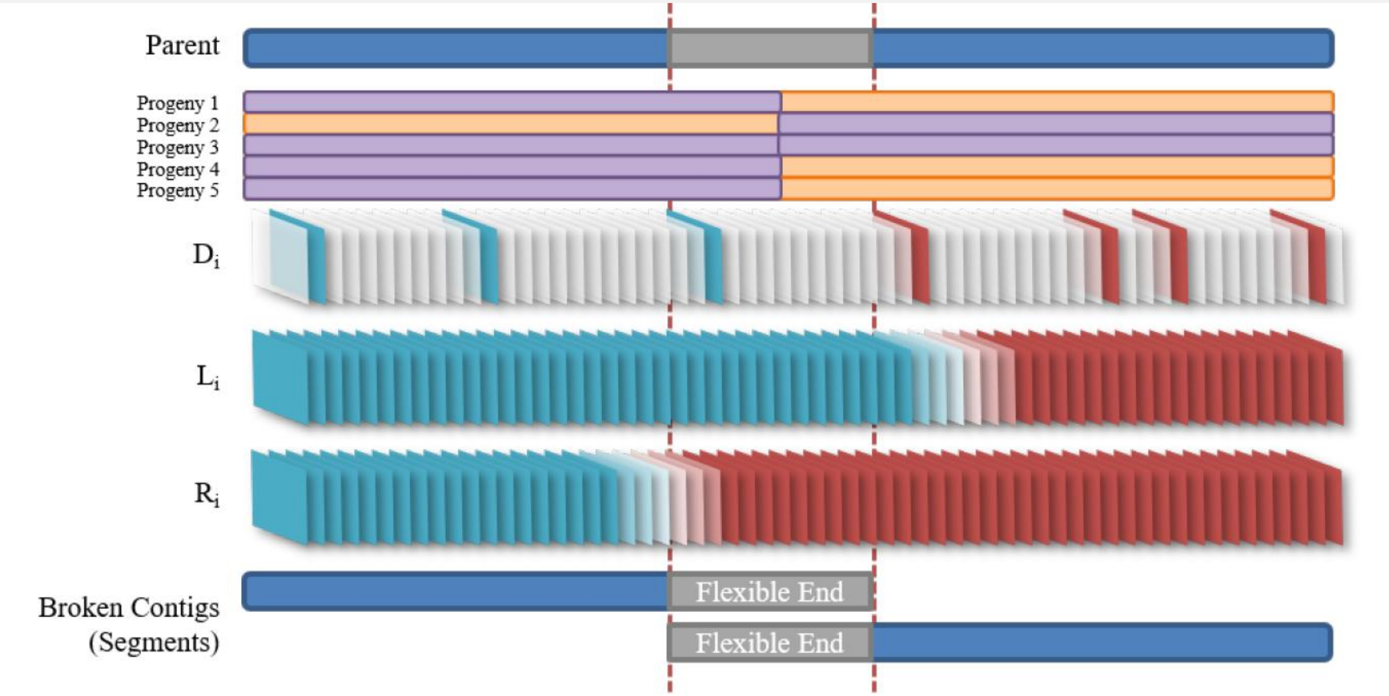


Figure 3.4: Examples of haplotype matrices and splitting. The top layer is a parent contig, on which the grey region harbors a misassembly. The tracks in the second layer shows inheritance of five progenies. Orange segments come from one parent; purple segments come from the other parent. The third layer is distance matrices. Grey matrices are zero matrices due to non-polymorphism, while cyan and red matrices are from different chromosomes (misassembled into this contig). Fourth and fifth layers show L_i and R_i respectively. Around the misassembled region, haplotype matrices are converted gradually as indicated by gradient colors. Finally, the parent contig is split into 2 segments as shown in bottom layer.

Flexible ends

When a contig is split at a pair of adjacent polymorphic positions v and w, each “broken contig” retains a copy of the sequence from v to w

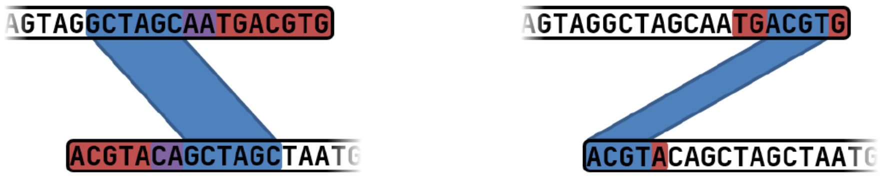
Flexible ends are free to be partially or totally discarded during merging because the assumed mis-assembly may happen at an arbitrary position between v and w

Note that a contig can be split multiple times

Merging

During merging, BLAST is used to find significant aligned fragments between the two involved contigs

The loss of an alignment is the length of the sequence between flexible end and aligned region



(a) Alignment with loss=4

(b) Alignment without loss

Figure 3.5: Two potential alignments between a pair of segments. Red regions indicate flexible ends while blue regions are alignments. (a) shows an alignment where aligned regions are out of flexible ends on both segments. The regions between aligned regions and flexible ends count as loss (marked purple). So, loss=4 for this case. (b) is another alignment where the aligned regions harbour within flexible ends and thus loss=0.

Flexible ends incur no loss because we duplicated them

If there is no alignment, the two contigs are concatenated with 50 N's in the middle

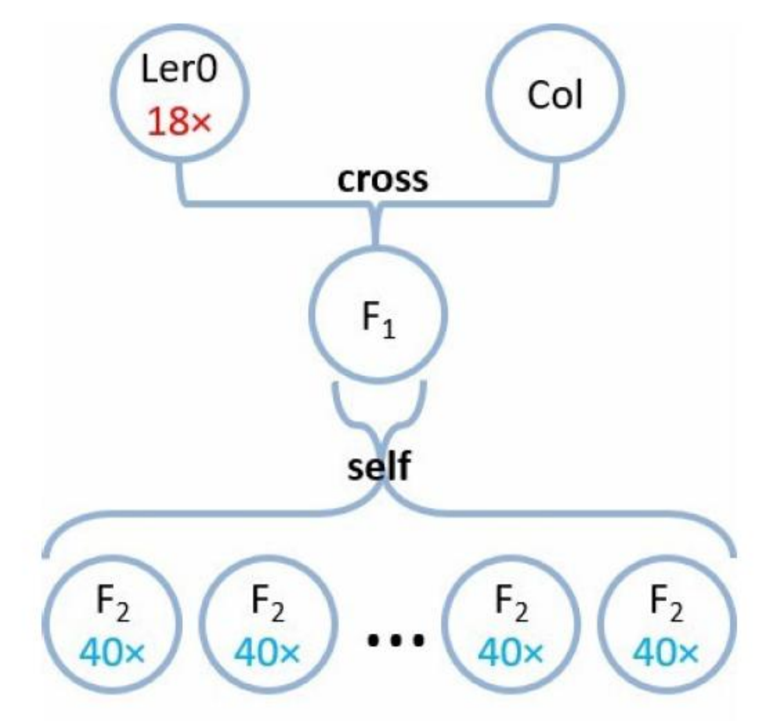
A. thaliana datasets

Parent Ler0 was Pacbio sequenced at 18x

F2 progenies were Illumina sequenced at 40x

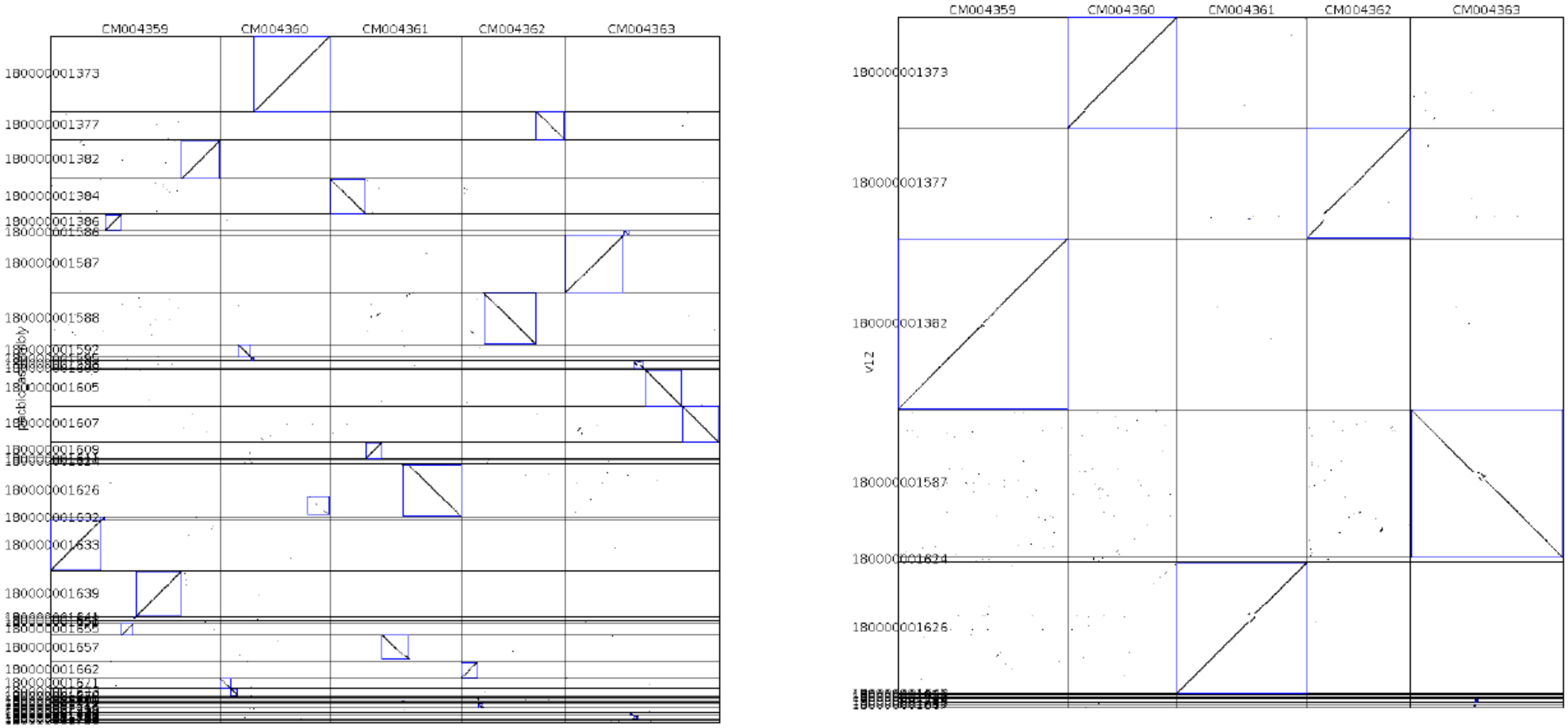
Draft assembly* of Ler0 was CAST-refined using 31 F2 progenies

Reference genome is Ler, GCA_001651475.1



* Pacbio public sample, https://github.com/.PacificBiosciences/DevNet/wiki/Arabidopsis-P5C3

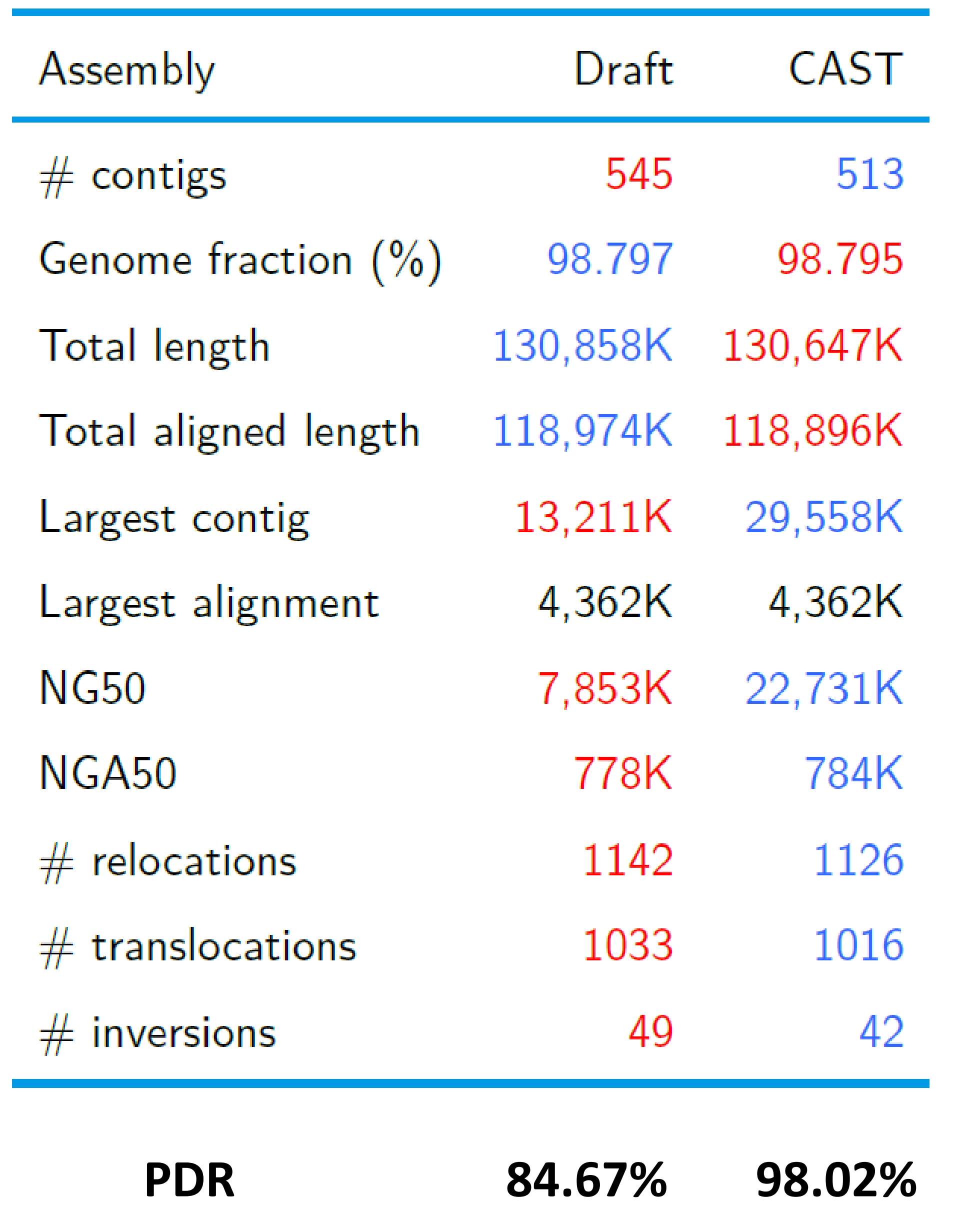
CAST-refined assembly looks obviously better



(a) MUMmer plot: reference genome vs. Draft assembly

(b) MUMmer plot: reference genome vs. CAST-improved assembly

Genome assembly quality metrics



Assembly	Draft	CAST
# contigs	545	513
Genome fraction (%)	98.797	98.795
Total length	130,858K	130,647K
Total aligned length	118,974K	118,896K
Largest contig	13,211K	29,558K
Largest alignment	4,362K	4,362K
NG50	7,853K	22,731K
NGA50	778K	784K
# relocations	1142	1126
# translocations	1033	1016
# inversions	49	42
PDR	**84.67%**	**98.02%**

CAST resolves long repeats

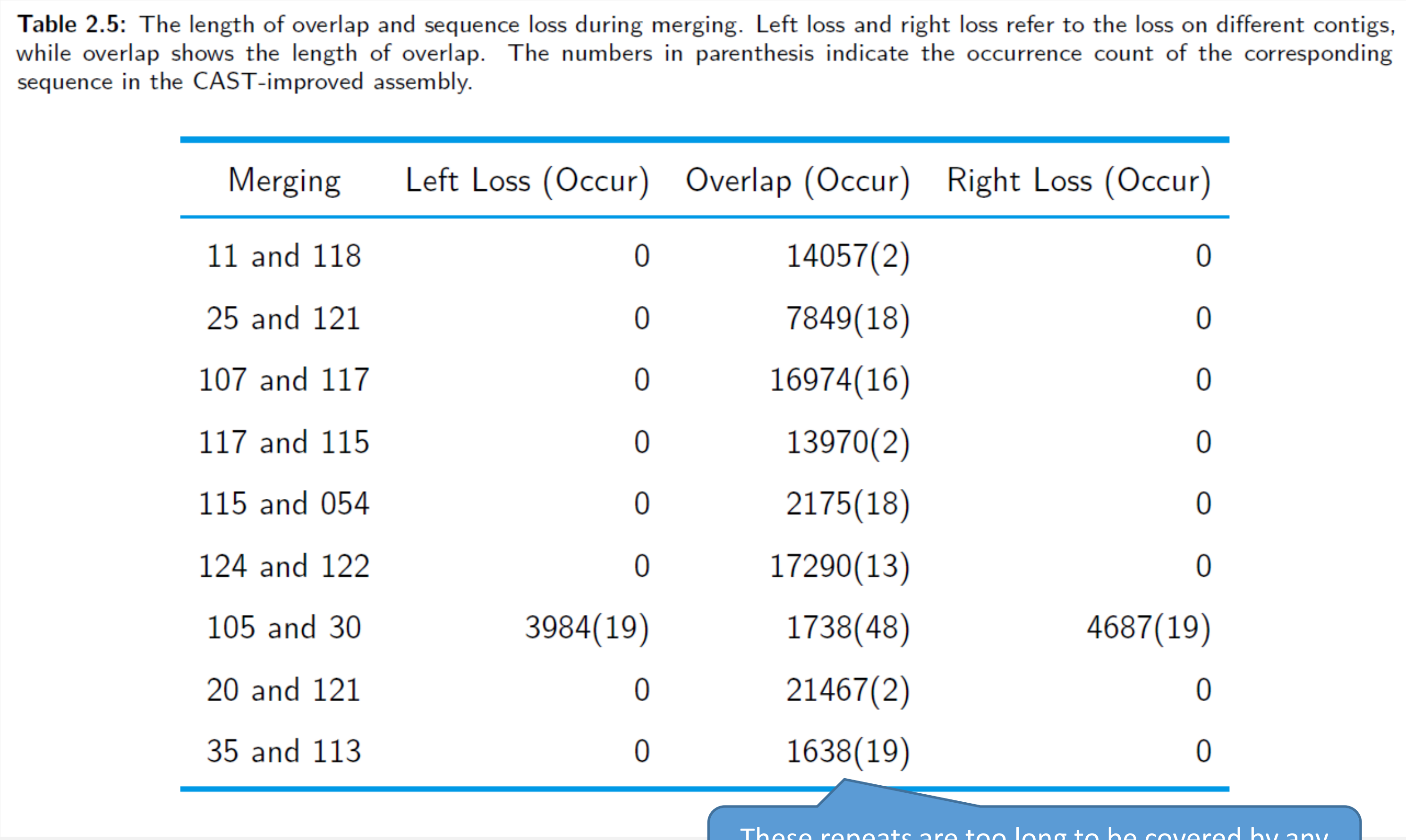


Table 2.5: The length of overlap and sequence loss during merging. Left loss and right loss refer to the loss on different contigs, while overlap shows the length of overlap. The numbers in parenthesis indicate the occurrence count of the corresponding sequence in the CAST-improved assembly.

Merging	Left Loss (Occur)	Overlap (Occur)	Right Loss (Occur)
11 and 118	0	14057(2)	0
25 and 121	0	7849(18)	0
107 and 117	0	16974(16)	0
117 and 115	0	13970(2)	0
115 and 054	0	2175(18)	0
124 and 122	0	17290(13)	0
105 and 30	3984(19)	1738(48)	4687(19)
20 and 121	0	21467(2)	0
35 and 113	0	1638(19)	0

These repeats are too long to be covered by any single read and/or repeat more than 10 times

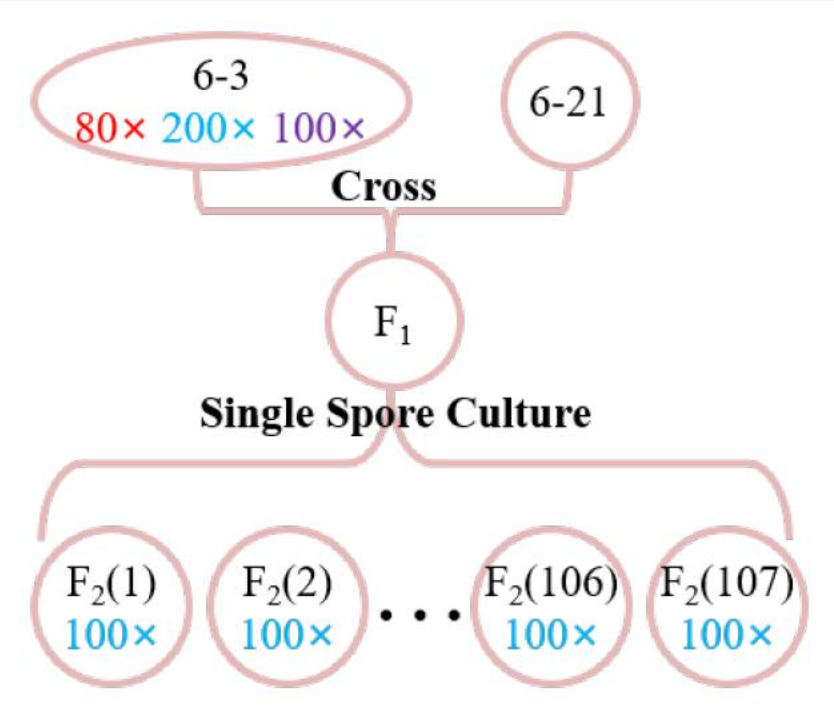
F. velutipes (enoki mushroom) datasets

Parent 6-3 was sequenced by Pacbio at 80x, illumina at 200x, and Hi-C at 100x

F2 progenies were Illumina sequenced at 100x

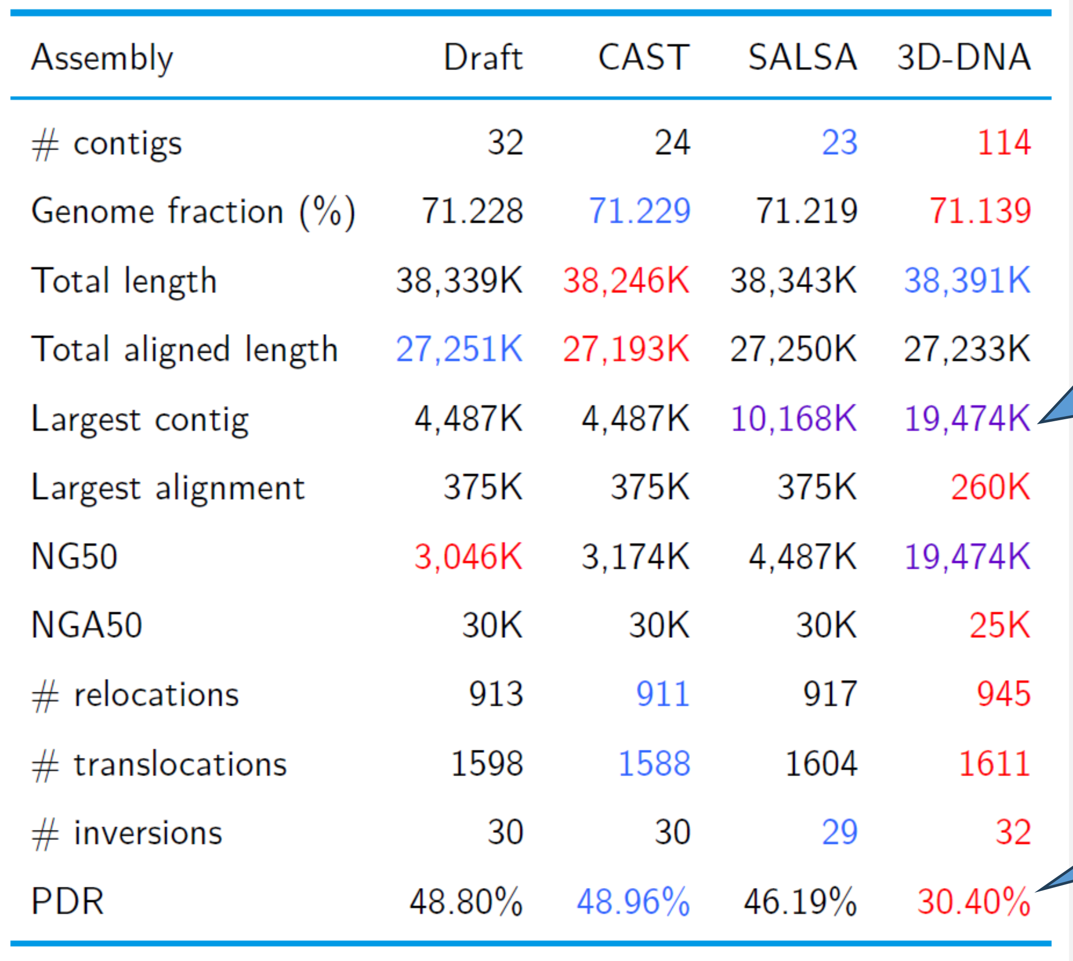
Draft assembly of 6-3 was CAST-refined using 31 of F2 single-spore cultures

Ref genomeis KACC42780, GCA_000633125.1, not 6-3



* Pacbio public sample, https://github.com/.PacificBiosciences/DevNet/wiki/Arabidopsis-P5C3

Genome assembly quality metrics



Assembly	Draft	CAST	SALSA	3D-DNA
# contigs	32	24	23	114
Genome fraction (%)	71.228	71.229	71.219	71.139
Total length	38,339K	38,246K	38,343K	38,391K
Total aligned length	27,251K	27,193K	27,250K	27,233K
Largest contig	4,487K	4,487K	10,168K	19,474K
Largest alignment	375K	375K	375K	260K
NG50	3,046K	3,174K	4,487K	19,474K
NGA50	30K	30K	30K	25K
# relocations	913	911	917	945
# translocations	1598	1588	1604	1611
# inversions	30	30	29	32
PDR	48.80%	48.96%	46.19%	30.40%

The largest F. velutipes chromosome is ~5Mb

SALSA & 3D-DNA mistakenly merged >3 chromosomes!

Ref genome is not 6-3

PDR ~50 already max

Draft vs CAST assemblies

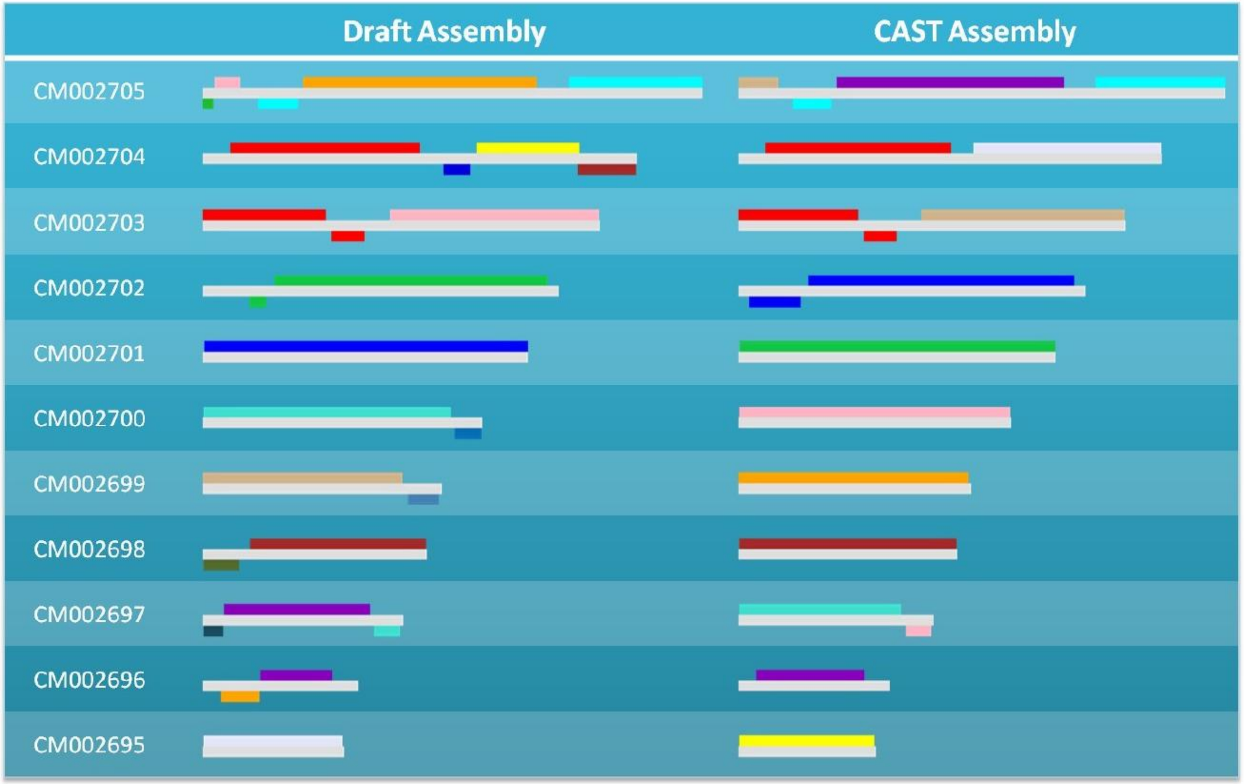


Figure 3.10: Assemblies aligned on the reference genome. Each row indicates a chromosome in the reference genome. Grey bars in the same row represent the same chromosome. Each column is an assembly. The colored bars attached to a grey bar are aligned blocks in this assembly to the reference chromosome. Within an assembly (i.e. a column in the table), bars with identical color are from the same contig.

Reference vs CAST assembly

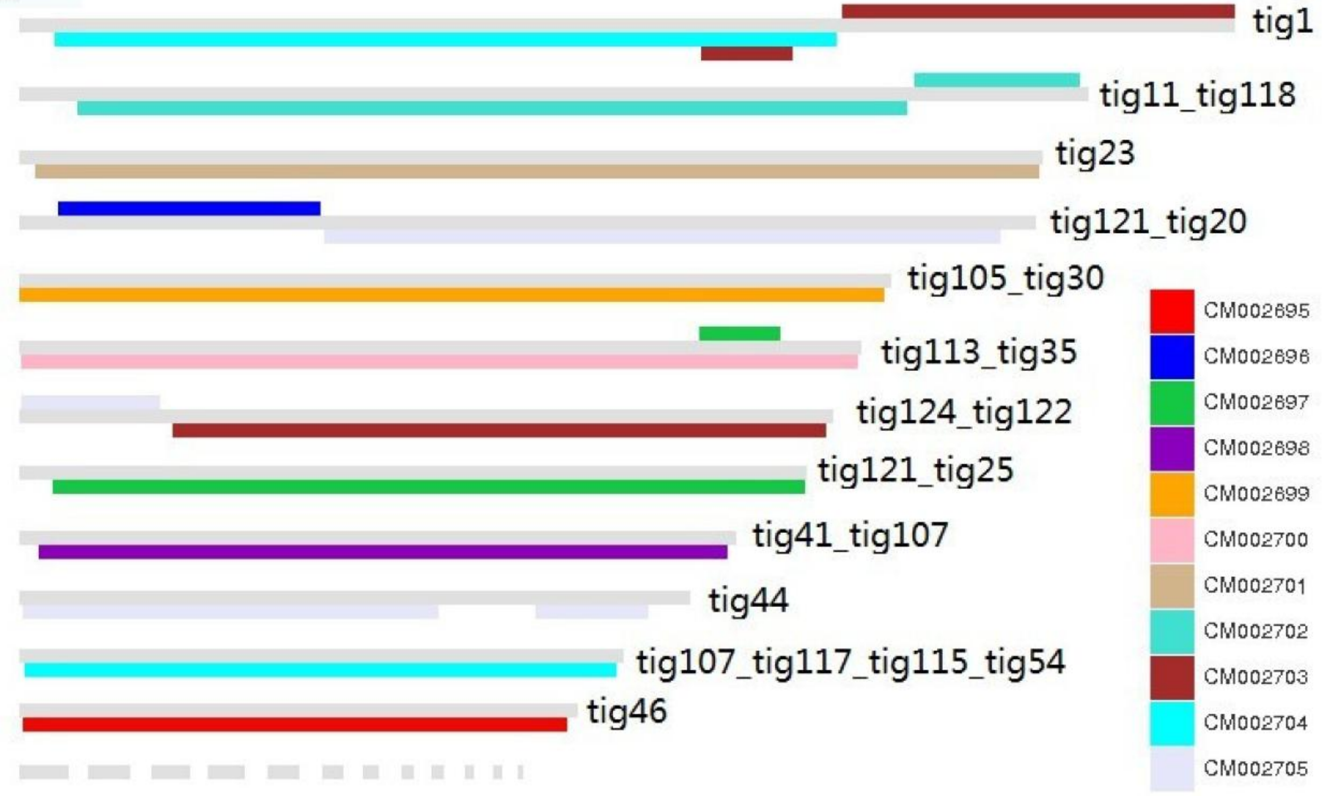
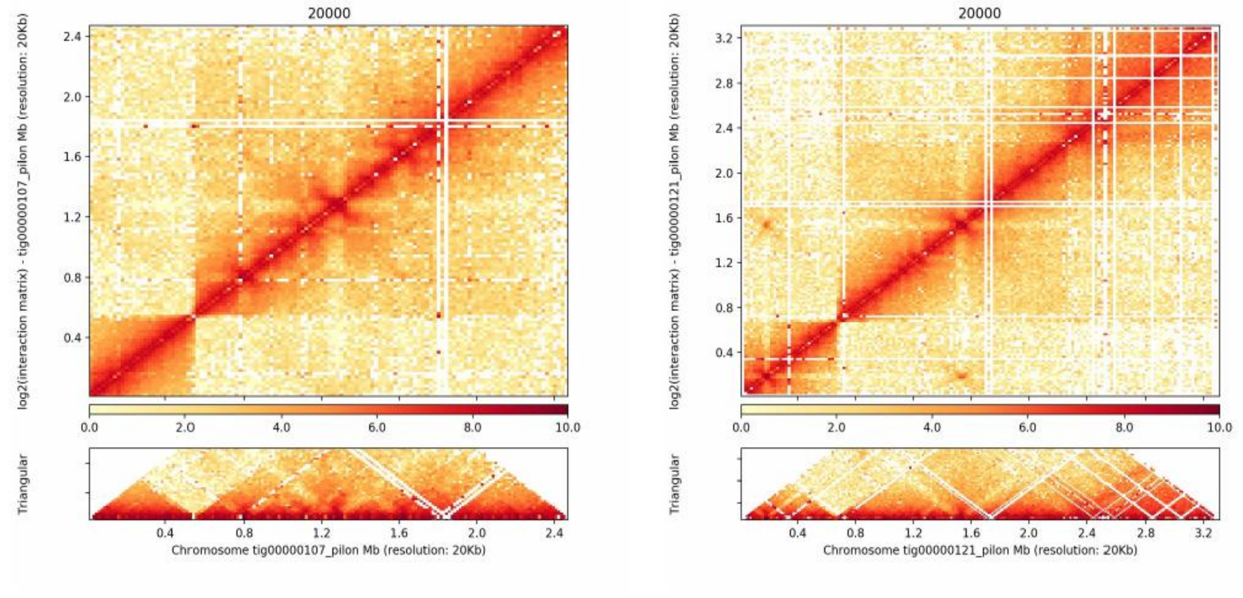


Figure 3.11: The reference genome aligned on CAST assembly. Each grey bar represents a contig in the CAST assembly, and the colored bars on it are aligned blocks from the reference genome. Each contig in the CAST assembly is named by draft contigs which form it. Small contigs without alignment are compacted.

CAST splits tig107 & tig121 of Draft
Are these splits correct?



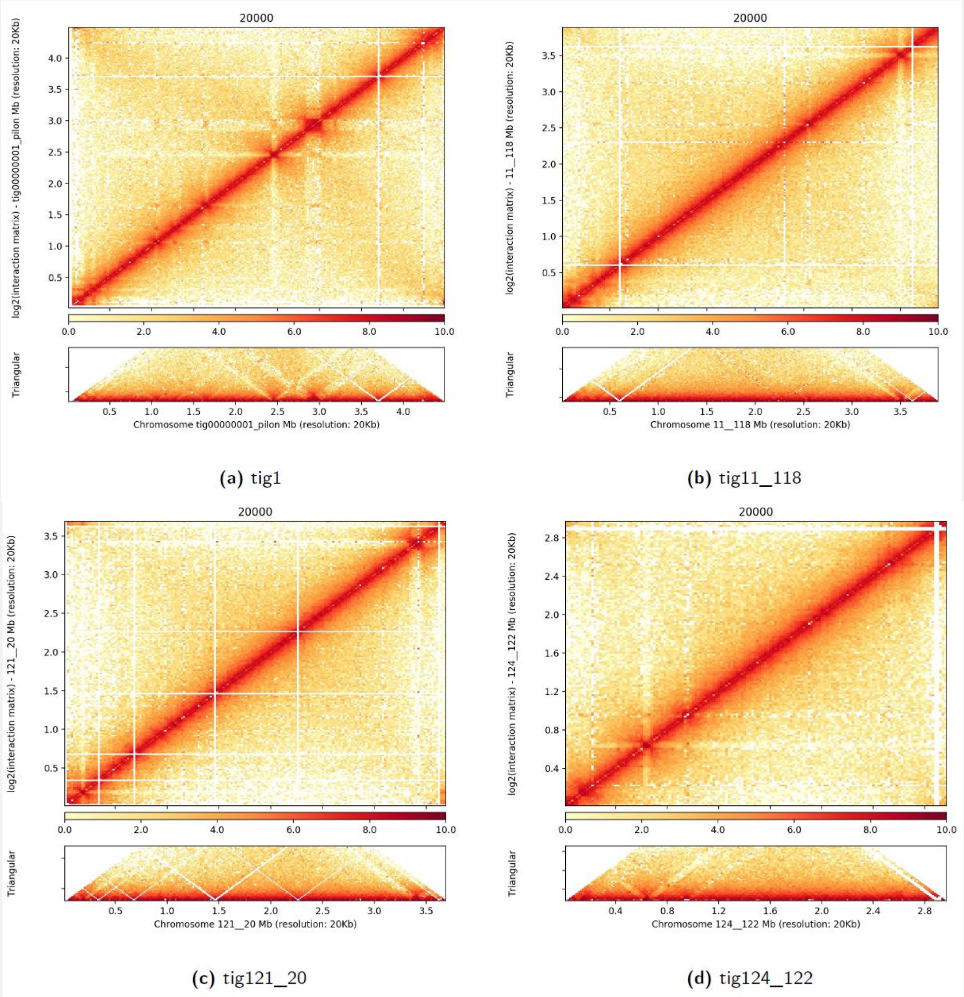
(a) tig107

(b) tig121

Hi-C shows clear discontinuity at 0.5Mb in tig107 and 0.7Mb in tig121, precisely where CAST split them

CAST merges tig124_tig122, tig121_tig20, tig1, tig11_tig118, of Draft.

Are these correct?



(a) tig1

(b) tig11__118

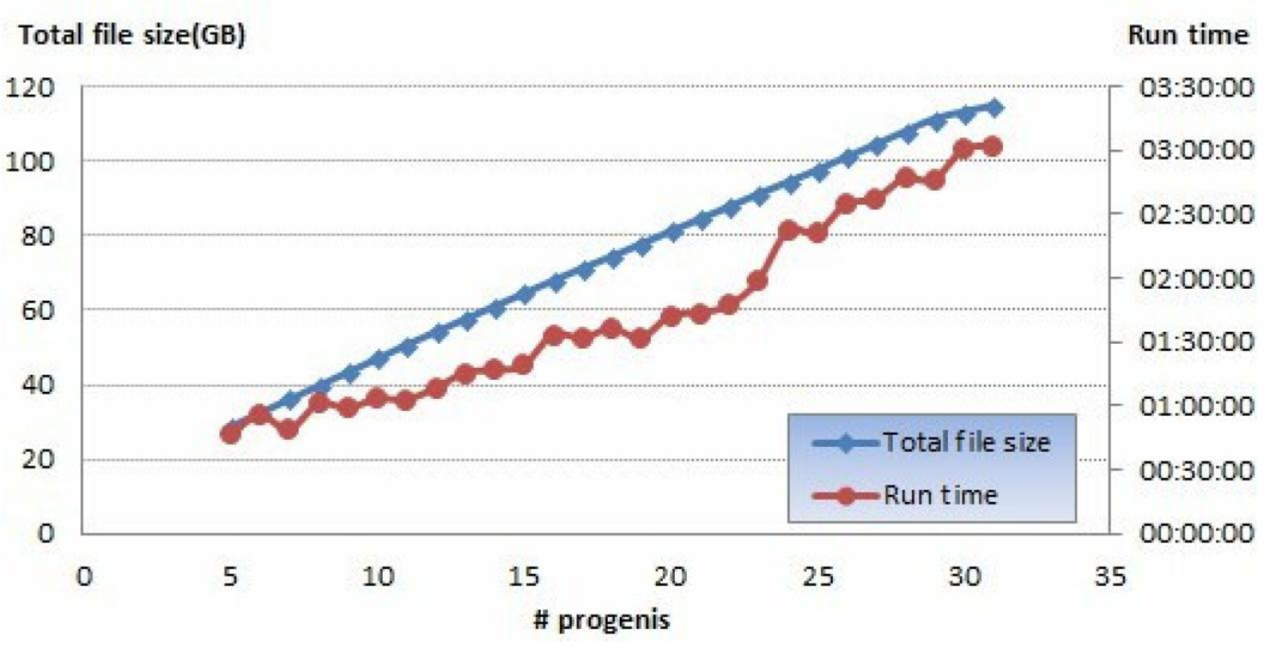
(c) tig121__20

(d) tig124__122

Hi-C data show clear continuity of CAST-merged contigs

These merges are likely correct

CAST is linear wrt file size



~3 hours to process 100GB on two six-core Xeon E5-2620 v3 2.4GHz CPUs, 64GB RAM, running CentOS78

Good to read

[CAST] Luyu Xie, Evaluation and improvement of genome assembly, PhD thesis, NUS, 2020

https://www.comp.nus.edu.sg/~wongls/psZ/luyu-thesis-v5.pdf