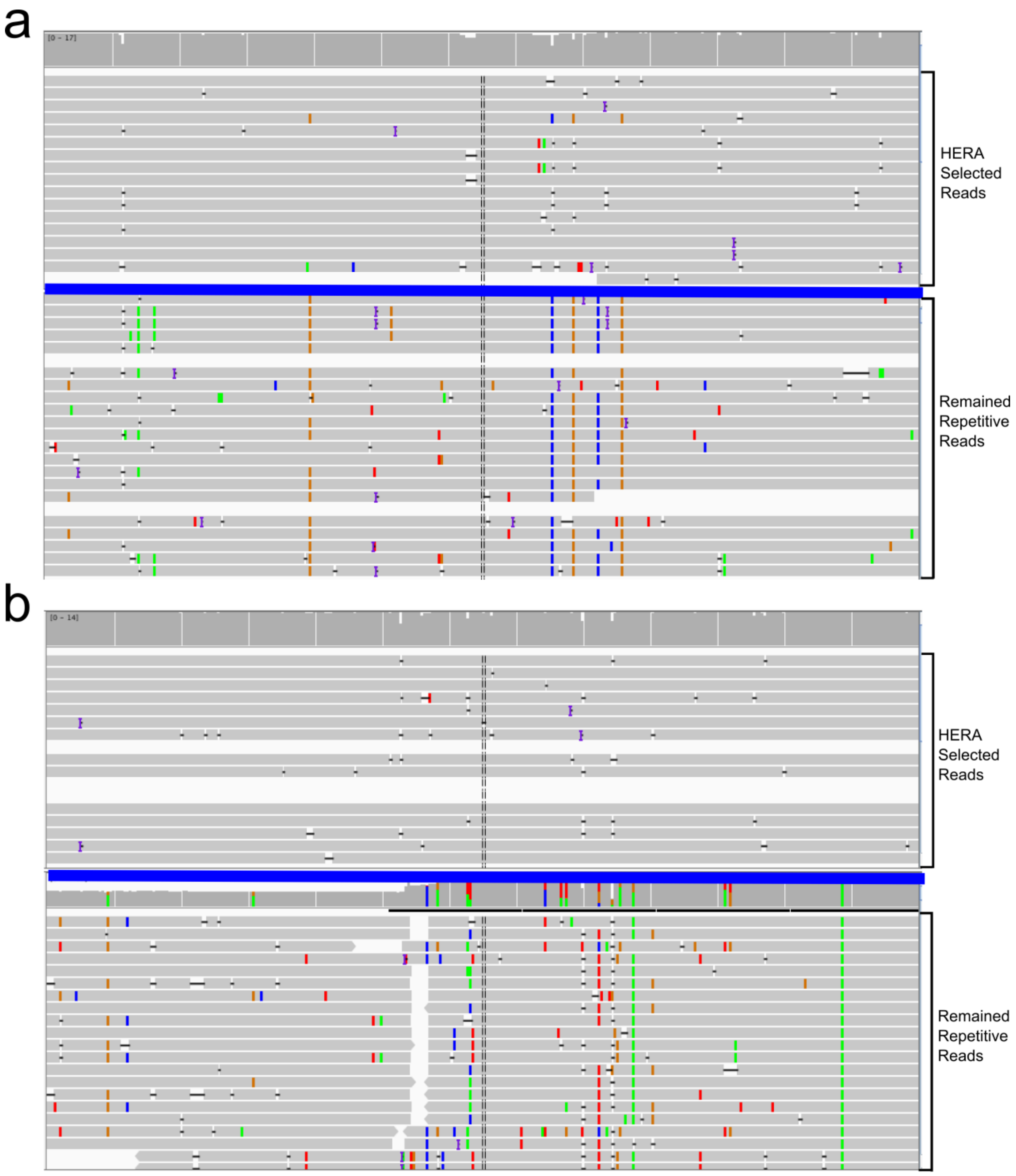
Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads

Huilong Du[1,2] and Chengzhi Liang[1,2]*

[1]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, 1 Beichen West Road No. 2, Beijing 100101, China

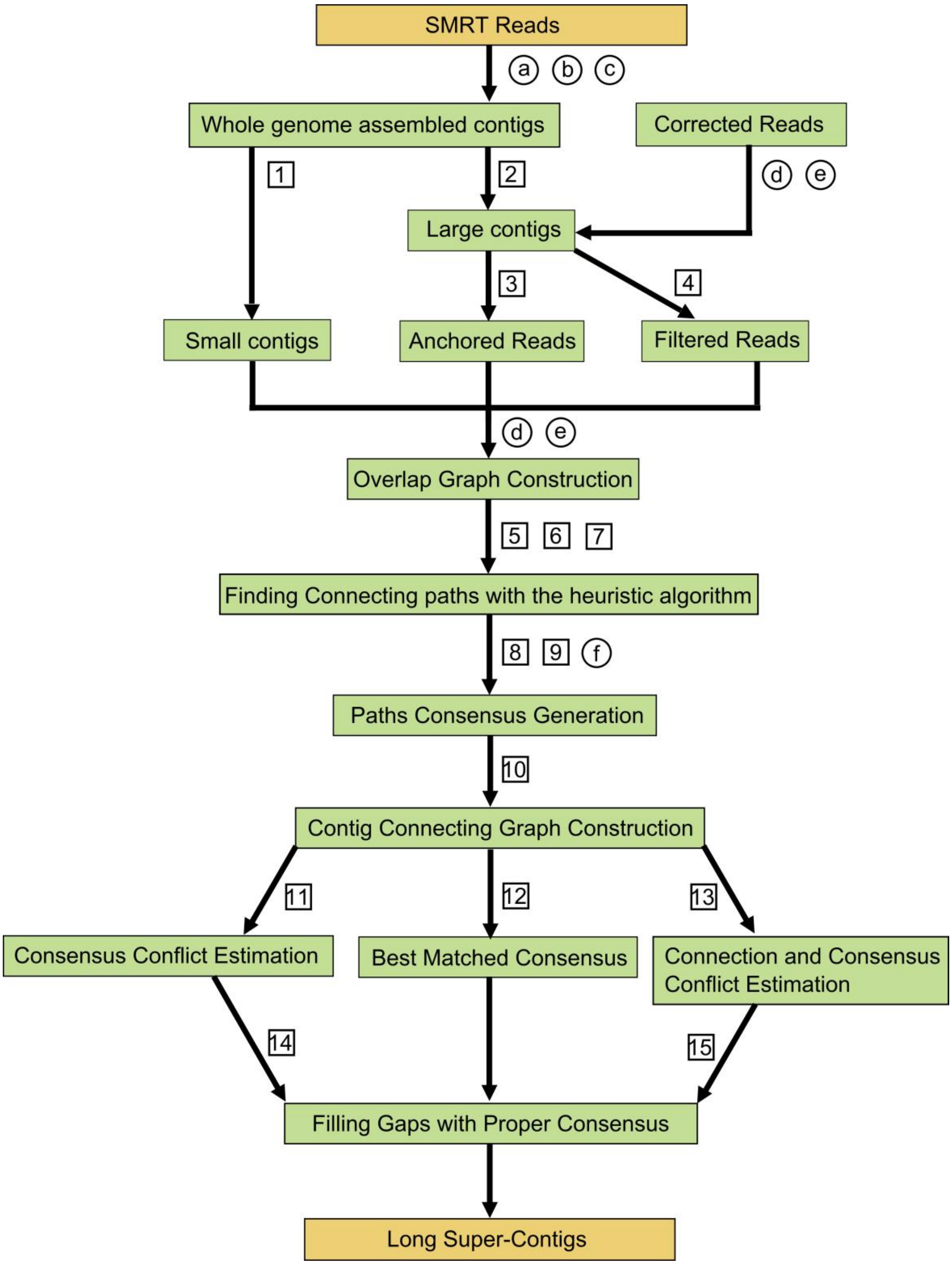
[2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

*Corresponding to: Chengzhi Liang (cliang@genetics.ac.cn)



Supplementary Figure 1: An example of repetitive sequences in R498 genome that can be distinguished at read level due to sequence variations.

(**a**)(**b**) Two repetitive sequences of ~23 kb on R498 chromosome 2 assembled by HERA shared sequence identity of 97.83%. The colored bars from IGV (http://software.broadinstitute.org/software/igv/) screenshot indicated SNPs between the HERA assembled sequences and the reads. Notably the reads can be classified into two groups that can be distinguished from each other based on the SNP distribution: the reads originated from the same copy (upper panel) and the reads originated from the other repeat copy (lower panel). CANU failed to assemble the two repeats and generated several contigs for each region.

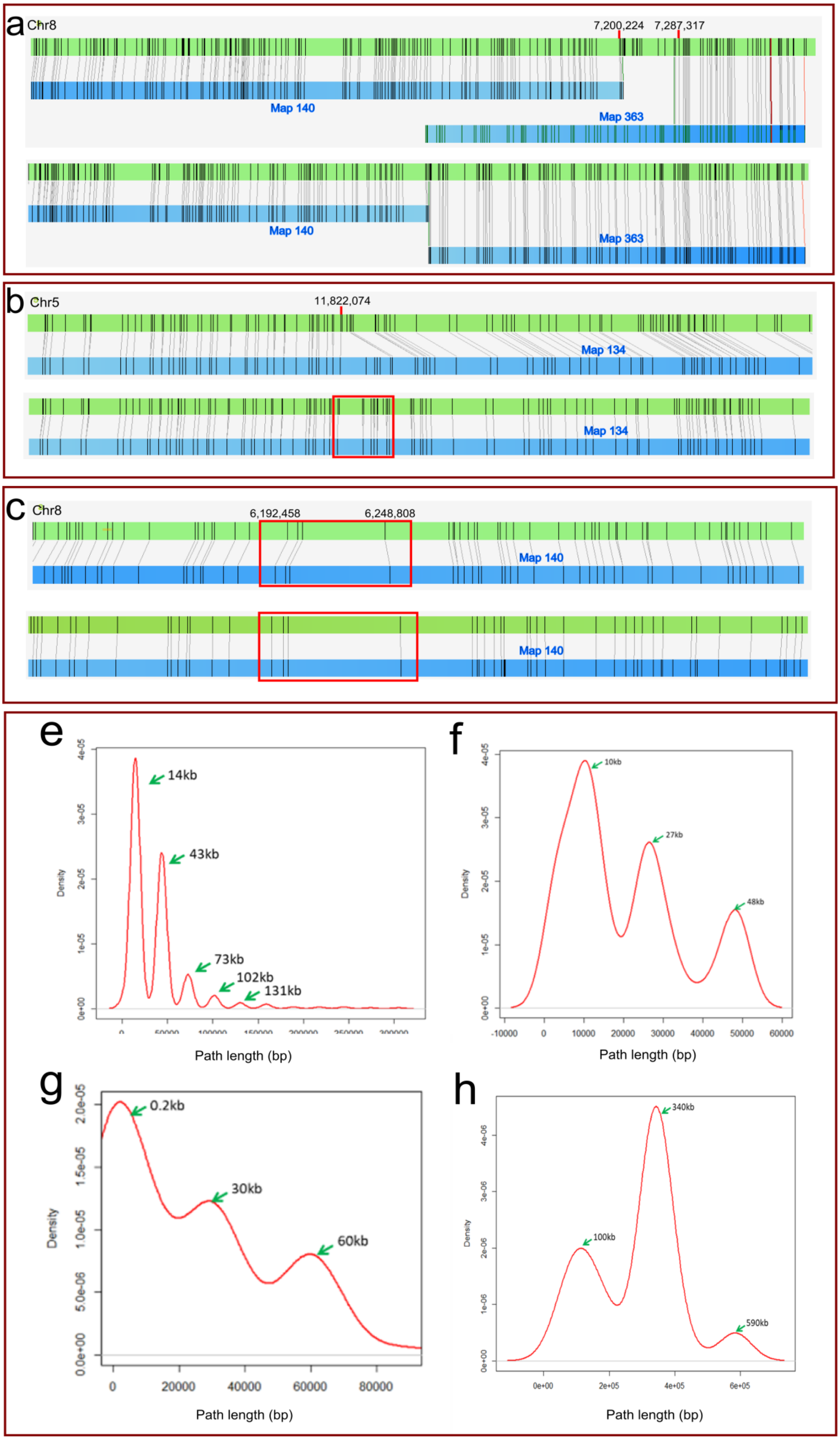


SMRT Reads
ⓐ ⓑ ⓒ
Whole genome assembled contigs
Corrected Reads
1
2
ⓓ ⓔ
Large contigs
3
4
Small contigs
Anchored Reads
Filtered Reads
ⓓ ⓔ
Overlap Graph Construction
5 6 7
Finding Connecting paths with the heuristic algorithm
8 9 ⓕ
Paths Consensus Generation
10
Contig Connecting Graph Construction
11
12
13
Consensus Conflict Estimation
Best Matched Consensus
Connection and Consensus Conflict Estimation
14
15
Filling Gaps with Proper Consensus
Long Super-Contigs

Supplementary Figure 2: Work flow of data process and HERA assembly.

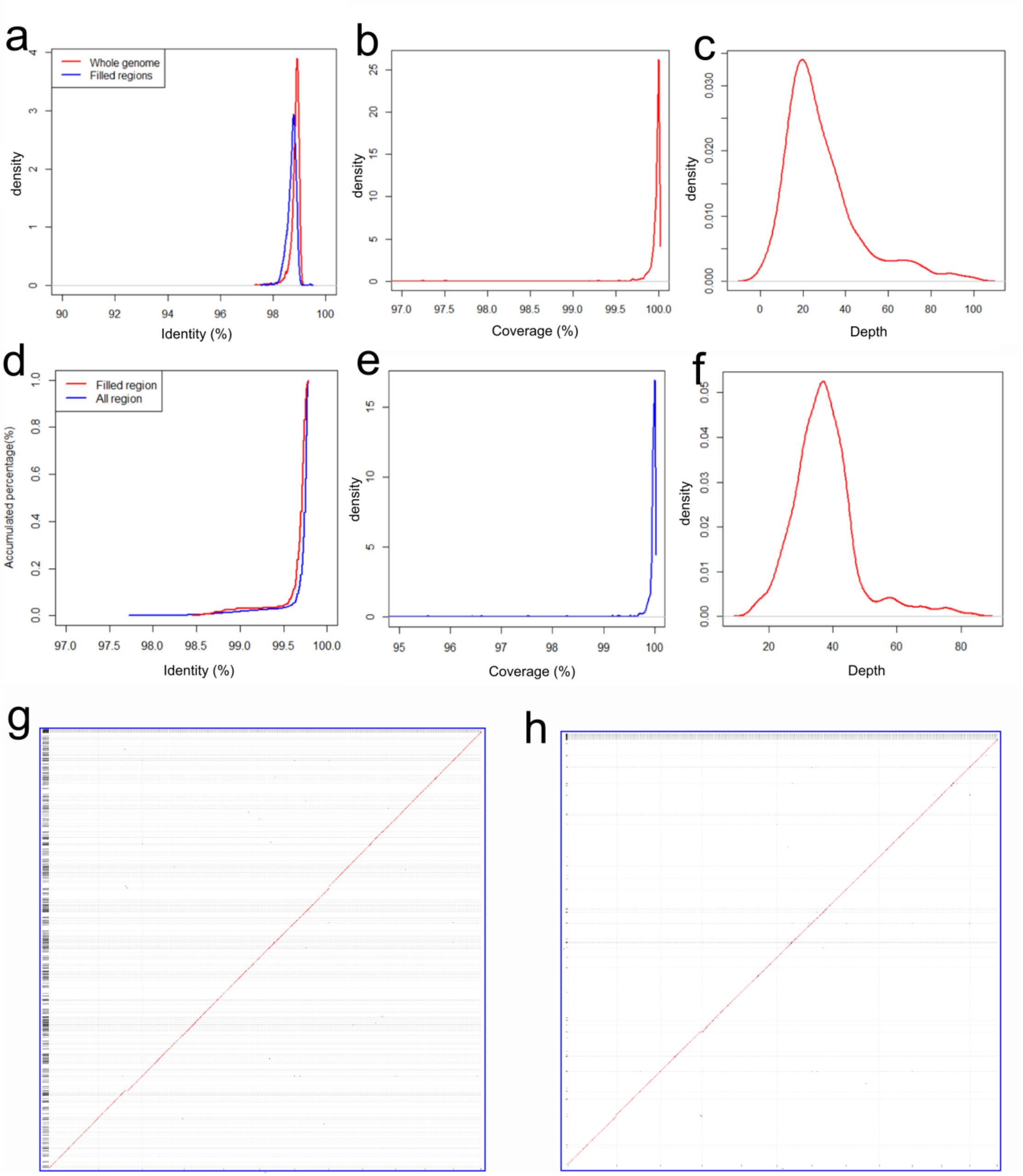
External software: (a) PBcR. (b) CANU. (c) FALCON. (d) MiniMap2. (e) BWA. (f) Daligner.

(1)(2) Selecting sequences by length. (3) Selecting the best matched reads which are aligned to the head or tail of contigs as starting/ending reads. (4) Removing the reads that are completely aligned to the internal region of contigs. (5) Using the highest overlap score for read extension. (6) Using the longest extension length for read extension. (7) A Monte Carlo approach for read extension. (8)(9)(10) Construction of connection graph. (11) Using the grouping information of contigs such as genetic maps. (12) Using the gap length information between ordered and oriented contigs. (13) Not using any additional information. (14) (15) Determining the proper contig pairs based on the conflicting indices and the proper consensus sequences for connection.


a
Chr8
7,200,224
7,287,317
Map 140
Map 363
Map 140
Map 363
b
Chr5
11,822,074
Map 134
Map 134
c
Chr8
6,192,458
6,248,808
Map 140
Map 140
e
14kb
43kb
73kb
102kb
131kb
Density
0e+00
1e-05
2e-05
3e-05
4e-05
0
50000
100000
150000
200000
250000
300000
Path length (bp)
f
10kb
27kb
48kb
Density
0e+00
1e-05
2e-05
3e-05
4e-05
-10000
0
10000
20000
30000
40000
50000
60000
Path length (bp)
g
0.2kb
30kb
60kb
Density
0.0e+00
5.0e-06
1.0e-05
1.5e-05
2.0e-05
0
20000
40000
60000
80000
Path length (bp)
h
340kb
100kb
590kb
Density
0e+00
1e-06
2e-06
3e-06
4e-06
0e+00
2e+05
4e+05
6e+05
Path length (bp)

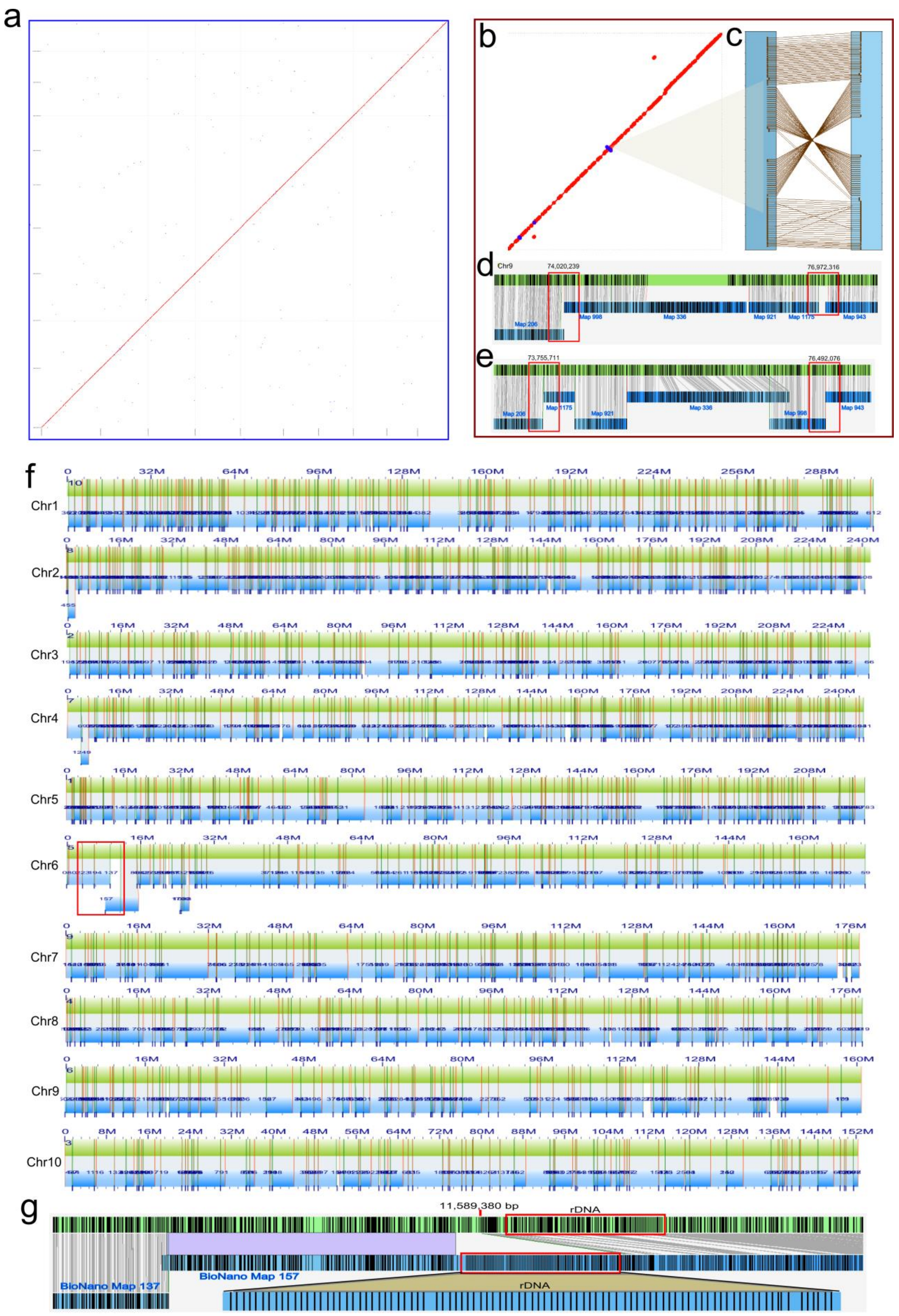
Supplementary Figure 3: The examples of improvements on the published R498 reference genome.

(**a**) The assembly of the region containing a missing sequence of 387 kb on chromosome 8 in our previously published R498 genome. (**b**) The improvement of a repetitive region on chromosome 5 around 11,822,074 bp. (**c**) The improvement of the repetitive region on chromosome 8 around 6,195,389 bp. (**e-h**) The length distribution of connecting paths revealed the presence of tandem repeats. (**e**) Around 7,986,535 bp on chromosome 1. (**f**) Around 16,466,293 bp on chromosome 7. (**g**) Around 6,022,125 bp on chromosome 8. (**h**) Around 14,965,388 bp on chromosome 9. Multiple peaks in the path length distribution plots indicated the complexity of these regions.



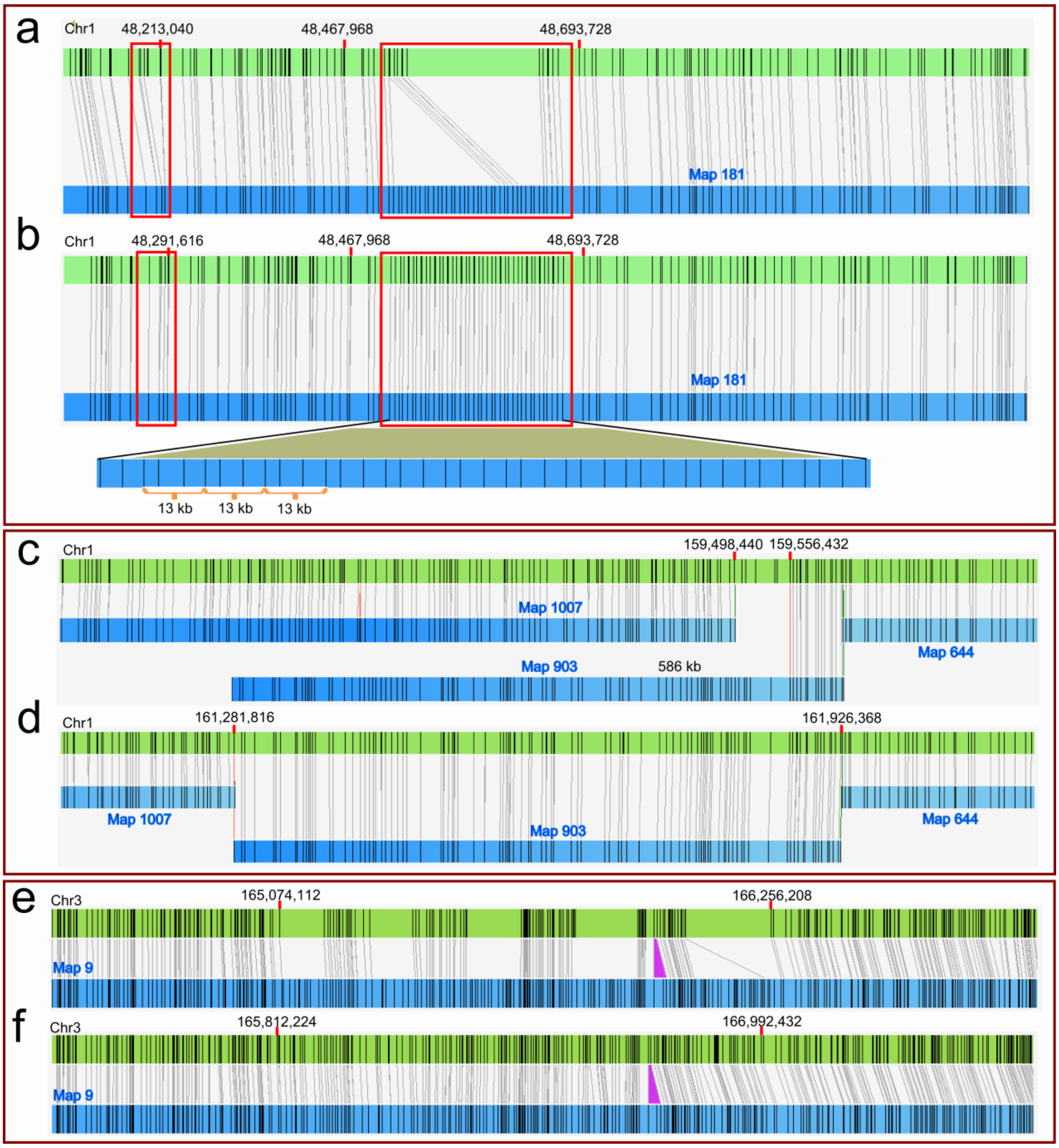
Supplementary Figure 4: The quality validation of the HERA assembled sequences in R498 genome.

(**a**-**c**) Validated by aligning to the corrected PacBio long reads. (**a**) The sequence identity. (**b**) The sequence coverage of the filled regions. (**c**) The sequence depth is peaked around 22x (genome average: 25x). (**d**-**f**) Validated by aligning to the Illumina short reads. Sequencing depth of whole genome: 41x. (**g**) CANU-assembled contigs aligned to R498 reference. (**h**) HERA-assembled contigs aligned to R498 reference.



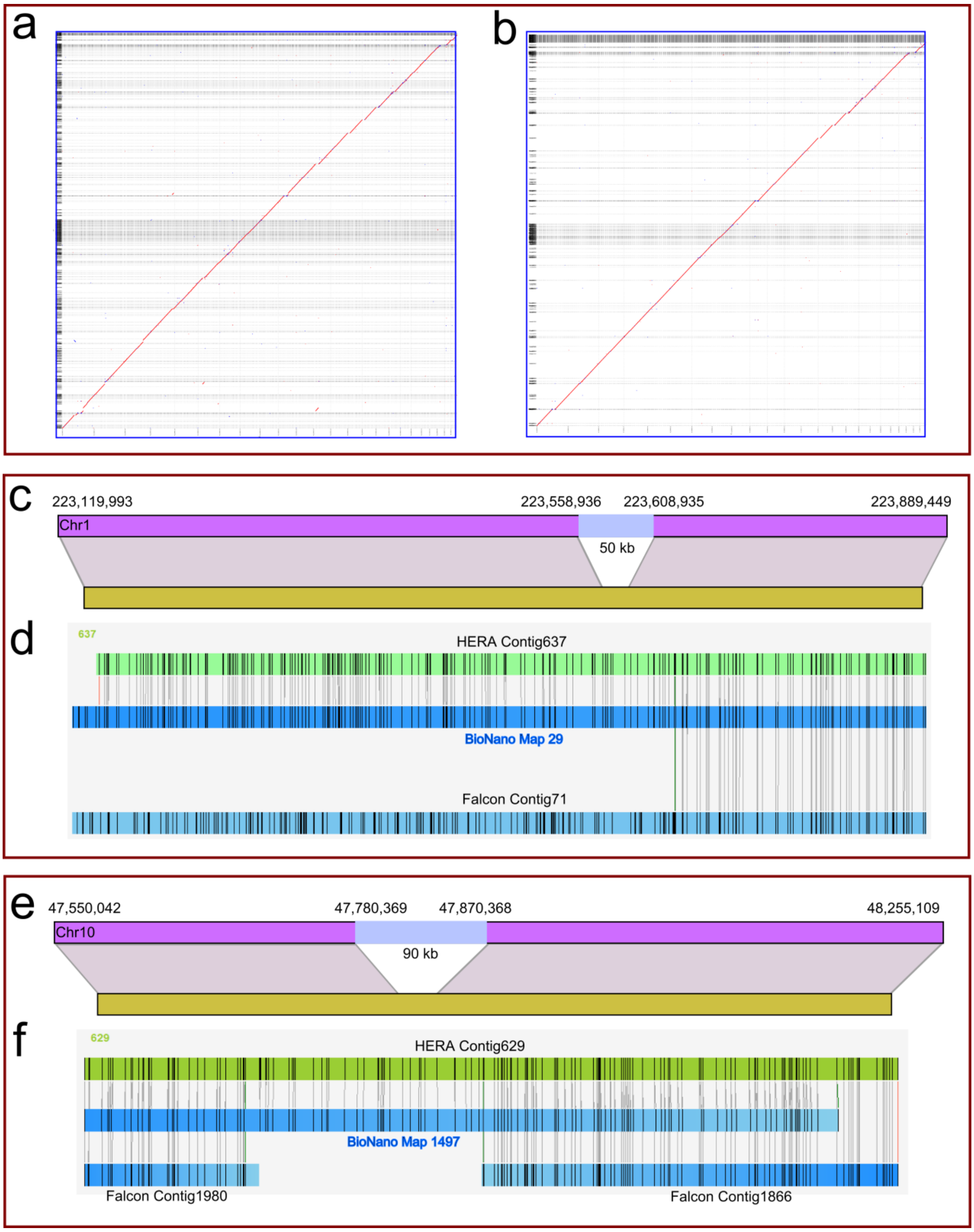
Supplementary Figure 5: Overview of the HERA-constructed maize genome.

(**a**) Comparison of HERA assembly to B73 RefGen_v4. (**b-e**) An inversion in RefGen_v4 was fixed. (**f**) comparison to BioNano maps. (**g**) A deletion in HERA contig caused by a long tandem repeat of rDNA.



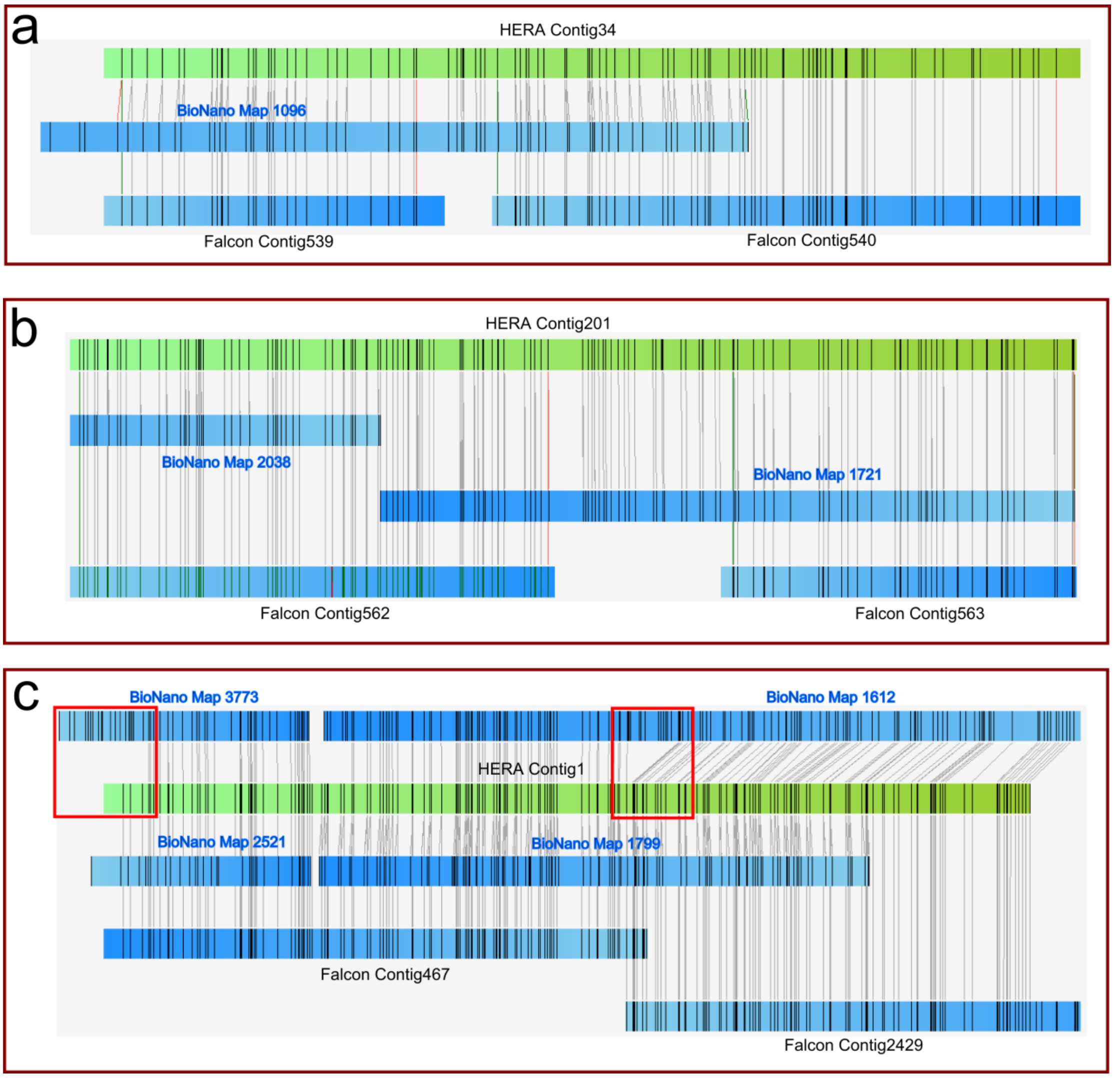
Supplementary Figure 6: The examples of missing sequences in B73 RefGen_v4 that were corrected by HERA.

(**a**)(**b**) A missing sequence of ~150 kb on chromosome 1 of B73 RefGen_v4 is a tandem repeat with the unit length of 13 kb. As shown in the right box in (**b**), all units of the tandem repeat are correctly assembled by HERA. (**c**)(**d**) A missing sequence of ~586 kb on chromosome 1 of B73 RefGen_v4 was assembled correctly by HERA. (**e**)(**f**) A region with five gaps on chromosome 3 of RefGene_v4 from 165.07 Mb to 166.25 Mb was assembled by HERA. Note that HERA did not try to fix the deletion introduced by PBcR in the purple triangle in (e) and (f).



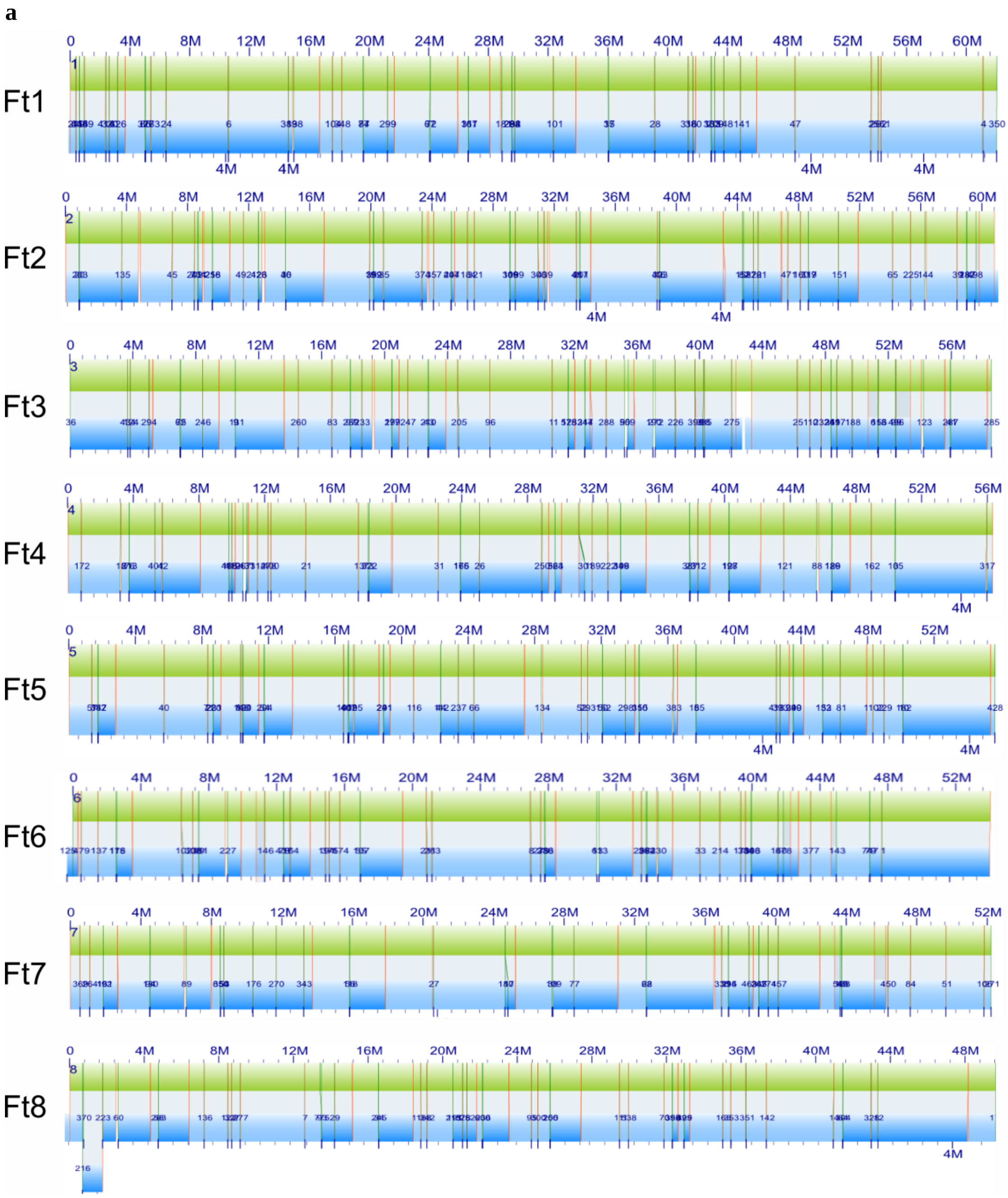
Supplementary Figure 7: The comparison of HERA assembly with the published HX1 genome (HX1_FALCON) and GRCh38 reference genome.

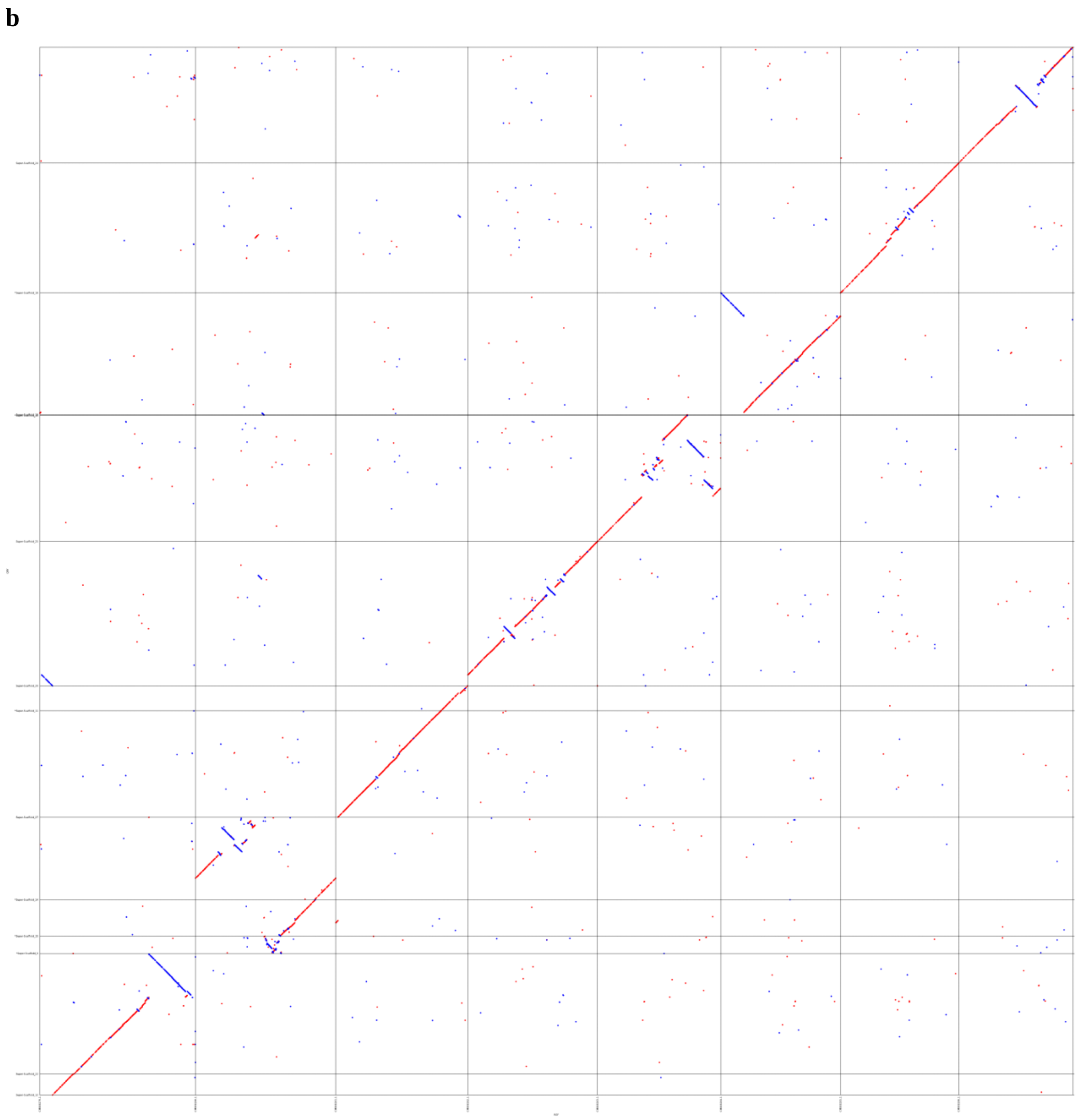
(**a**) The alignment of HX1_FALCON to GRCH38. (**b**) The alignment of HX1_HERA (HERA assembly of HX1) to GRCH38. (**c**)(**d**) A gap on GRCh38 chromosome 1 could be closed based on HX1_HERA. HERA correctly assembled a whole region that has a missing sequence in the HX1_FALCON. The gap in GRCH38 is in the missing sequence of HX1_FALCON. (**e**)(**f**) Another gap on GRCh38 chromosome 10 could be closed using HX1_HERA. HERA correctly assembled the region that contained two contigs and a missing sequence in HX1_Falcon. The gap on GRCh38 chromosome 10 is within the missing sequence of HX1_FALCON.



Supplementary Figure 8: The examples of potential falsely anchored sequences in either chromosome X or chromosome Y in GRCh38.

The HERA contigs Contig34, Contig201 and Contig1 were each partially aligned to GRCh38 chromosomes X and Y with some segments matching to only one chromosome and some other segments matching to both chromosomes (Supplementary Table 5). (**a**)(**b**) A single BioNano map completely covered the HERA contigs in each case which proved their correctness. (**c**) A single HERA contig Contig1 matched to two BioNano maps. One is consistent with Contig1 while the other one contained several indels and overhangs compared with Contig1, suggesting that the corresponding sequences in GRCh38 chromosomes X and Y were misassembled or mispositioned.


a
Ft1
Ft2
Ft3
Ft4
Ft5
Ft6
Ft7
Ft8
0
4M
8M
12M
16M
20M
24M
28M
32M
36M
40M
44M
48M
52M
56M
60M



Supplementary Figure 9: Comparison of the hybrid scaffolds based on HERA assembly of Tartary buckwheat to BioNano genome maps (a) and old pseudomolecules (b).

(a) No structural variations other than indels were found between the sequence and genome maps. (b) A number of misoriented and several mispositioned sequences (including one on different chromosome) were found in the old pseudomolecules, incorrectly incorporated due to the small contig size.