Supplementary Information

A wall-through of the example 3:

Running HaploMerger2 (HM2) on a real highly-polymorphic diploid assembly

Supplementary Table 1. Statistical summary

The initial diploid assembly			
Total span (Mb)	708		~4% heterozygosity; haploid genome is ~440Mb; assembled from short reads (a mixture of 454 and illumine reads); 30% of the genome has been soft-masked.
Total bases (Mb)	697		
Scaffold count	10354		
Scaffold N50 size (kb)	264		
Contig N50 size (kb)	30		
The reference haploid assembly created by the old HaploMerger pipeline			
Span (Mb)	400		25Mb unpaired sequences not included.
Total bases (Mb)	395		
Scaffold count	1339		
Scaffold N50 size (kb)	919		
Contig N50 size (kb)	34		
Two haploid assemblies created by the new HaploMerger2 pipeline			
	Reference	Alternative	
[A] After removal of potential mis-joins			
Span (Mb)	708		228 mis-joins processed; 126 short scaffolds removed.
Total bases (Mb)	697		
Scaffold count	10456		
Scaffold N50 size (kb)	245		
Contig N50 size (kb)	30		
[B] After rebuilding of two haploid sub-assemblies			
Span (Mb)	410	385	21Mb unpaired sequences not included.
Total bases (Mb)	405	379	
Scaffold count	1442	1442	
Scaffold N50 size (kb)	916	881	
Contig N50 size (kb)	31	32	
[C] After re-scaffolding of two haploid sub-assemblies			
Span (Mb)	412	387	
Total bases (Mb)	405	379	
Scaffold count	573	573	
Scaffold N50 size (kb)	2287	2212	
Contig N50 size (kb)	31	32	
[D] After removal of tandem alleles (for the reference haploid sub-assembly only)			
Span (Mb)	407	387	788 tandems removed; 5.7M bases removed. This is the only step that loses data. The removed data were stored in _D3.tandem_removal_excised_seq.fa.
Total bases (Mb)	400	379	
Scaffold count	573	573	
Scaffold N50 size (kb)	2240	2212	
Contig N50 size (kb)	31	32	
[E] After gap-filling (for the reference haploid sub-assembly only)			
Span (Mb)	407	387	26988 gaps in total; 6286 gaps closed. These final assemblies do not include 21Mb unpaired sequences.
Total bases (Mb)	401	379	
Scaffold count	573	573	
Scaffold N50 size (kb)	2240	2212	
Contig N50 size (kb)	40	32	

Supplementary Text. Description of the full procession

1. The initial diploid genome assembly of the amphioxus (*Branchiostoma belcheri*) is created from a mixture of 454 reads (shotgun and 2/3/8kb mate-pairs, ~30X) and Illumina reads (2*115bp PE, ~30X). The Celera assembler CABOG v6.1 was used to create this diploid assembly, with customized options: utgErrorRate=0.015; overlapper=mer; merSize=22; unitigger=bog; doExtendClearRanges=2; stoneLevel=2; doResolveSurrogates=1; cgwDemoteRBP=1; and doToggle=1.

2. The resulted diploid assembly spans 708Mb, with a scaffold/contig N50 size of 264kb/30kb (Supplementary Table 1). According the k-mer analysis, the estimated haploid genome size is about 440Mb, suggesting that each haplotype is averagely ~80% complete.

3. This diploid assembly has been deposited in GenBank under accession: AYSR00000000.1.

4. To suppress false alignments without losing true alignments, we used WindowMasker to soft-masked this diploid assembly and 30% of the sequences have been masked. A copy of this soft-masked assembly (bbv18wm.fa.gz) can be downloaded from our HM2 website: https://github.com/mapleforest/HaploMerger2/releases. Alternatively, one can use RepeatMasker instead, or use WindowMasker and RepeatMasker together.

5. The knowledge of the allelic polymorphism rate is important. We compared the longest 10% sequences of the diploid assembly against the rest 90%, by doing so we estimated that the average difference rate between alleles is ~4%. Empirically, for stringent 'hard' filtering of alignments, the identity cutoff for alignment can be set to 100% minus 2-3 times of the average polymorphism rate (i.e., to 88-92% in this case). Here we used the default cutoff value (80%), which means imposing relax 'hard' filtering. In this example, there was little difference in the outcomes between stringent and relax 'hard' filtering because we will introduce a more flexible 'soft' filtering – a genome-specific scoring matrix.

6. To achieve optimal alignment sensitivity/specificity and to avoid imposing a 'hard' identity cutoff for alignments, we can use an alignment scoring matrix adjusted to the oplyorphism rates and nucleotide mutational biases in the diploid assembly. Here we used lastz_D_Wrapper.pl from the HM2 package to infer the scoring matrix. In this inference, the longest 10% sequences of the diploid assembly were aligned against the rest 90%, and the identity cutoff for alignments was set to 90%.

7. For comparison, here we first ran the old HaploMerger pipeline on the diploid assembly, with all parameters set to default, except that the unpaired sequences are not included in the final haploid assembly. The statistics of the resulted haploid assembly is presented in Supplementary Table 1, spanning 400Mb with scaffold/contig N50 sizes of 919kb/34kb. It should be noted that the old pipeline can only produce the reference haploid assembly and cannot implement re-scaffolding and gap-filling.

8. To run HM2 on the diploid assembly, we download the soft-masked copy of this assembly as well as the required library files from our HM2 website. The running environment was set up as suggested in HM2's manual. Most parameters were set to default, except that:
   a) 12-20 CPU threads were used when applicable;
   b) a scoring matrix specific to this diploid assembly was used (which happens to be the default matrix);
   c) the option "--step" in Lastz's control file "all_lastz.ctl" was set to 20 in order to speed up the alignment procedure;
   d) unpaired sequences were not included into the haploid sub-assemblies in this test (i.e., "including_unpaired" in hm.batchB5 is set to 0);
   e) only one round was carried out for each MH2 module.

9. We strongly recommend cleaning up the fasta sequences of the input diploid assembly before running HM2 by using faDnaPolishing.pl from HM2:

```
gunzip –c genome.fa.gz | ./bin/faDnaPolishing.pl --legalizing \
--maskShortPortion=1 --noLeadingN --removeShortSeq=1 >genome_cleaned.fa.gz
```

10. We used HM2, module hm.batchA to remove major misjoins from the diploid assembly. The result is shown in Supplementary Table 1. The default setting is to break up potential mis-joins that are flanked by over 50kb sequences. A total 228 mis-joins were found and broken up.

11. We used HM2, module hm.batchB to create two haploid sub-assemblies, the reference and the alternative. The scaffold number has been reduced to 573, and the scaffold N50 size has been increased to 919kb.

    It is worth noting that the reference haploid sub-assembly is more complete than the alternative haploid sub-assembly. The result is shown in Supplementary Table 1.

12. We used HM2, module hm.batchC to re-scaffold the haploid sub-assemblies with 2kb, 3kb, 8kb and 20kb mate-pair libraries (downloadable on our HM2 website). The scaffold N50 size has been increased to 2287kb, an order higher than the initial diploid assembly. The result is show in Supplementary Table 1.

    It should be noted that the 2/3/8kb libraries are the same libraries used in the de novo assembly stage. And the 20kb library (contain ~12 thousand effective mate-pairs) was not used in the de novo assembly because CABOG cannot finish the assembly with this library due to the over-complex graph. Taken together, this result shows that scaffolding on the haploid assembly is much more effective than on the diploid assembly.

13. We used HM2, module hm.batchD to remove potential tandem errors from the reference haploid sub-assembly. In this example, tandems longer than 4kb have been examined, and a total of 788 tandems (~5.7Mb) were removed. The result is shown in Supplementary Table 1.

    It should be noted that this is the only step in HM2 that will lose genomic information, so we strongly recommend being careful with this module. To avoid permanent loss of the genomic information, MH2 will collect all the excised tandem sequences in an independent fasta file (_D3.tandem_removal_excised_seq.fa).

14. Finally, we used HM2, module hm.batchE to fill N-gaps in the reference haploid sub-assembly. The Illumina paired-end sequence libraries for gap-filling can be downloaded from GenBank, under accession SRX1364942 and SRX1364943. More than 1/4 of the gaps have been closed in this example, and the contig N50 size has been increased to 40kb. The result is shown in Supplementary Table 1.

    It is worth noting that we only used a small quantity of Illumina PE reads in this example in order to limit the overall running time. By using more reads and even 2-8kb mate-pairs, HM2 can close many more gaps.