Supplementary_Materials

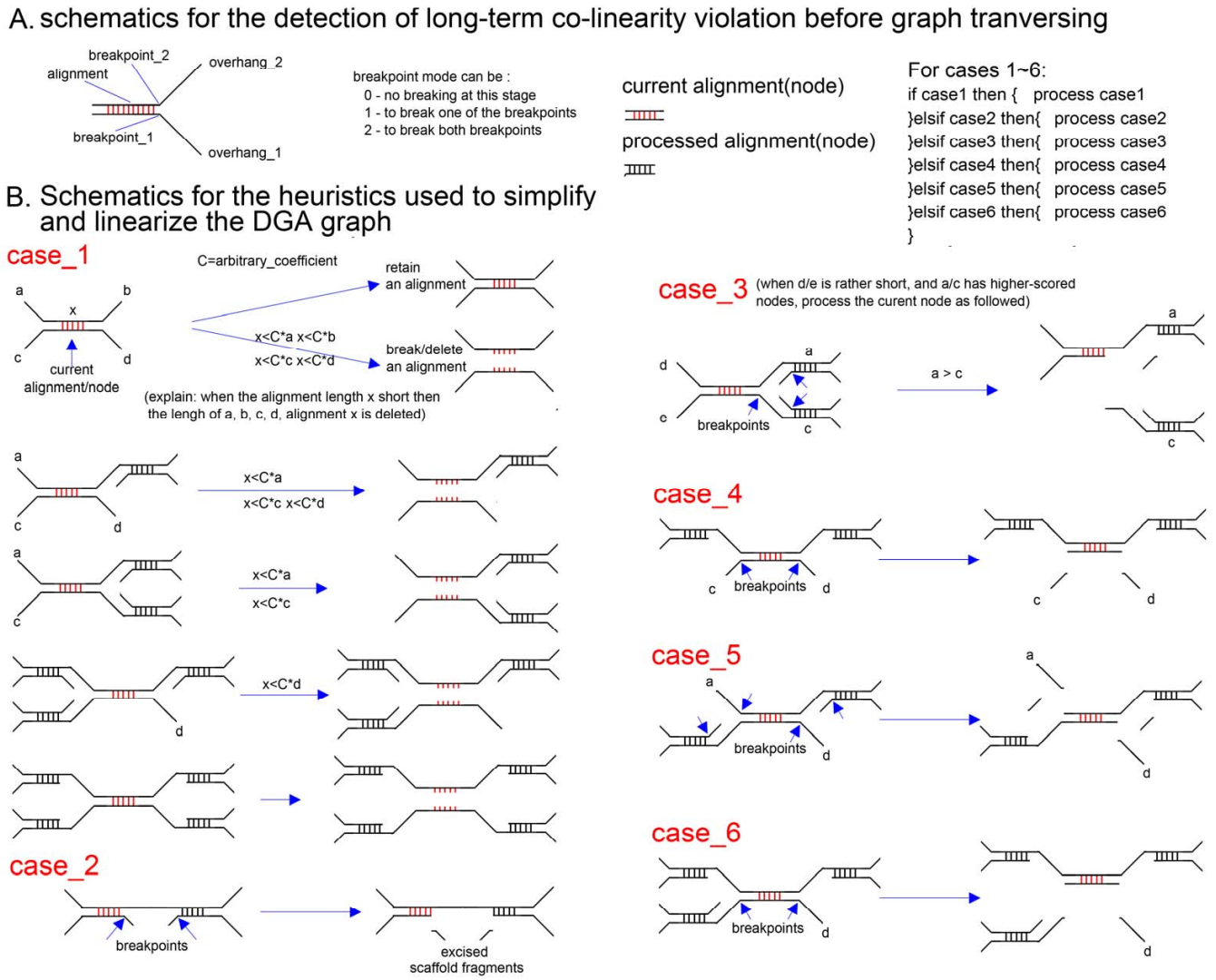


Figure S1.

(A) The diagram shows how to identify long-term co-linearity violation events before traversing the DGA graph. The detection is based on the pairwise alignment of two scaffolds. In simple words, if two scaffolds share a reliable alignments (for example, score >5000000), but there are two long overhangs or un-aligned sequences (for example, >50000bp) following the alignment, then we deem it a violation events. Then HaploMerger offers four choices:

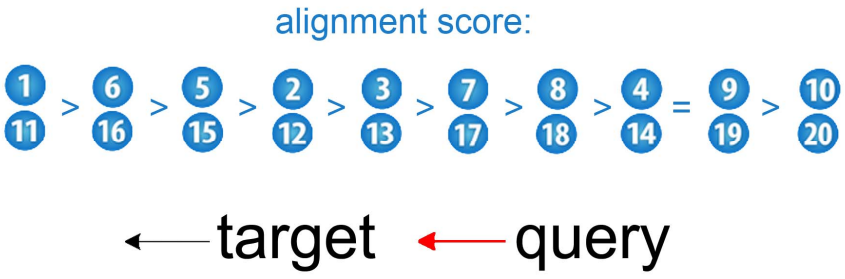
1) does nothing but leave to the subsequent graph traversing;

2) automatically breaks one of the scaffolds based on heuristics (with ⩾50% chance to break the problematic scaffold);

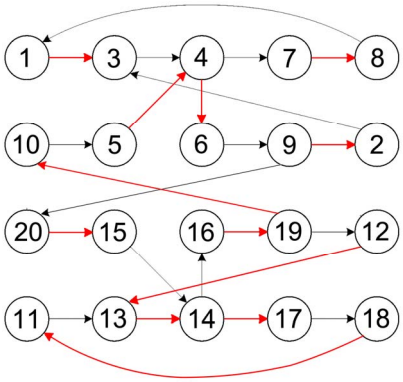
3) automatically breaks both scaffolds (at the expense of losing some scaffold continuity);

4) leaves it to user decision (users may visually inspect the event, consult a mate-pair graph, or do extra analyses/experiments before passing the decision to HaploMerger).

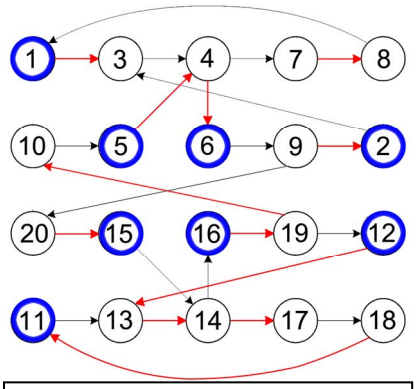
(B) The schematic diagram shows the heuristics used to simplify and linearize the DGA graph in the course of the traversing. There are totally six cases. The red alignment indicates the current alignment (or the node in the DGA graph) for study. For each node/alignment, the node and the context of four ends (a, b, c, d) will be evaluated together. Four ends of the node, a, b, c and d, can be only sequences (if there is no adjacent node on that end), or links to adjacent nodes. A decision for the current node will be made based on the evaluation of the node and the context, such as deleting the node (case 1), keeping the alignment while cutting off the overhang sequences (case 2, 4, and 5), breaking one of the scaffold to remove dichotomous paths (case 3 and 6).



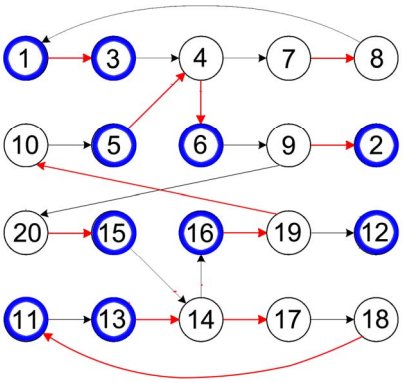
The ranking of the scores of each alignment (node). Black arrow shows the 5->3 direction of the target sequences, whereas the red arrow shows the 5->3 direction of the query sequences.



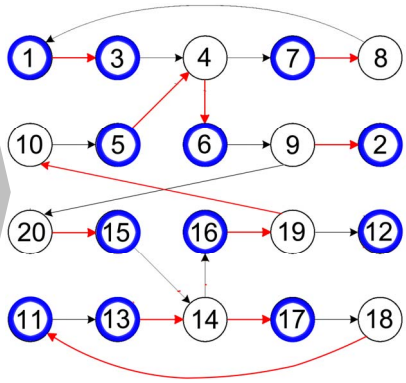
Step 0. The initial DGA graph. Graph traversing is from the highest-score to the lowest.



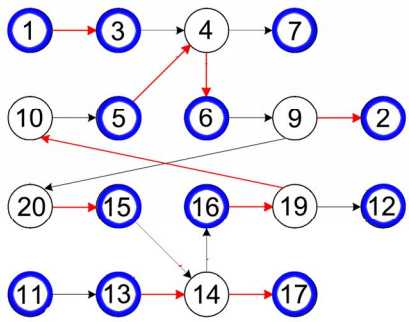
Step 1. Nodes 1/11, 6/16, 5/15 and 2/12 have no conflicts with each other, so retained.



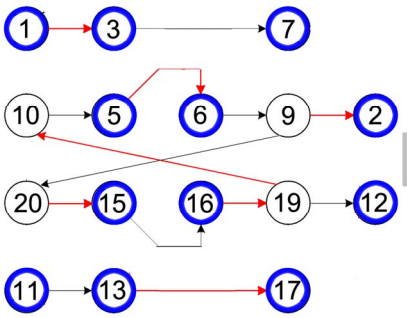
Step 2. Node 3 has two inward edges (1->3,2->3), so Figure_S1B_case3 is used to solve the problem: node 3 is retained; the link 2->3 is broken (i.e., scaffold d is broken, see Figure 2F). So does for node 13.



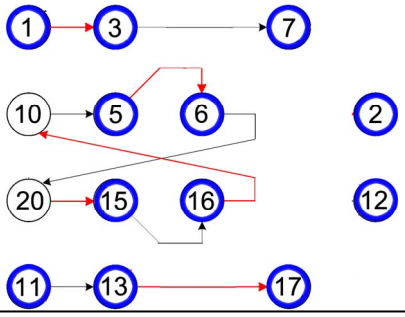
Step 3. Nodes 7 and 17 have no conflicts with previous nodes, so retained.



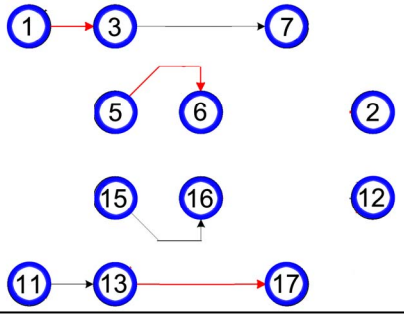
Step 4. Node 8 creates a loop: 8->1->3->7->8, so node 8 and its two edge are deleted. So does for Node 18.



Step 5. Node 4 is surround by four nodes with higher score (3,5,6,7), so Figure_S1B_case1 is applied: node 4 is deleted, and the edges are bridged along. So does for node 14.



Step 6. Node 9 creates a violation of long-term co-linearity: 9->the_end_of_scaffold_g versus 9->2. So Figure_S1B-case1 is applied: node 9 is deleted, and the 6->20 edge is bridged along. So does for node 19.



Step 7. Node 10 and 20 creates a loop: 10->5->6->20->15->16->10. So Node 10 and 20, and their related edges are deleted. Traversing ends.

Figure S2. A schematic diagram shows the step-by-step simplification and linearization of the initial DGA graph. The initial DGA graph is the same as shown in the Figure 2D. A reduced DGA graph is obtained in the final step, which is corresponding to the graph shown in Figure 2E.

Figure S3. Alignment dot plots show all 24 major (>50Kb) assembly errors remained in the reference haploid assembly for the 274Mb simulated genome (on the left panels). The right panels show the corresponding errors in the original diploid assembly. The red circle shows the mis-joining point. Note that some scaffolds contain 2 errors.

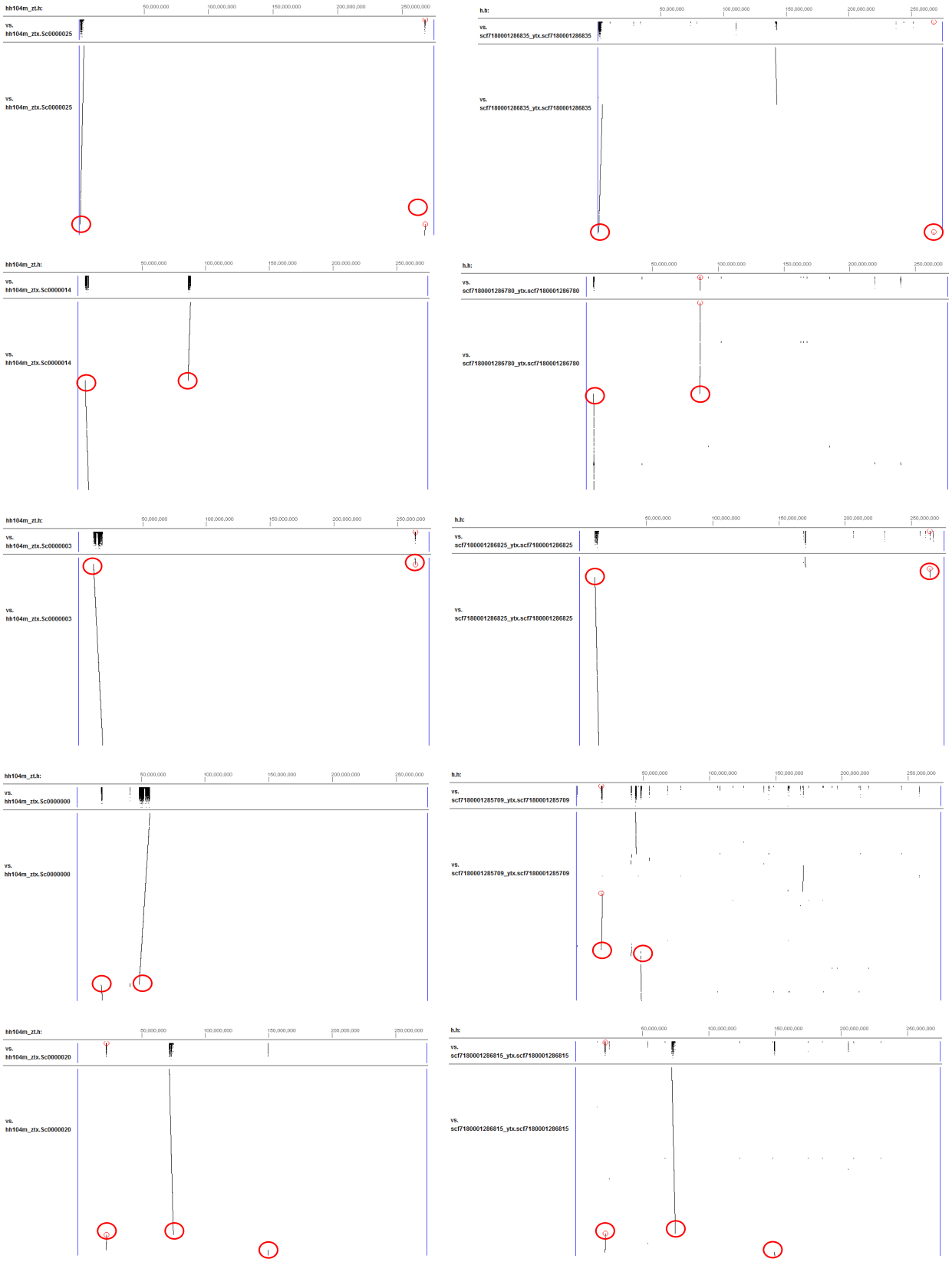


Figure S3 (continued). Alignment dot plots show all 24 major (>50Kb) assembly errors remained in the reference haploid assembly for the 274Mb simulated genome (on the left panels). The right panels show the corresponding errors in the original diploid assembly. The red circle shows the mis-joining point. Note that some scaffolds contains 2 errors.

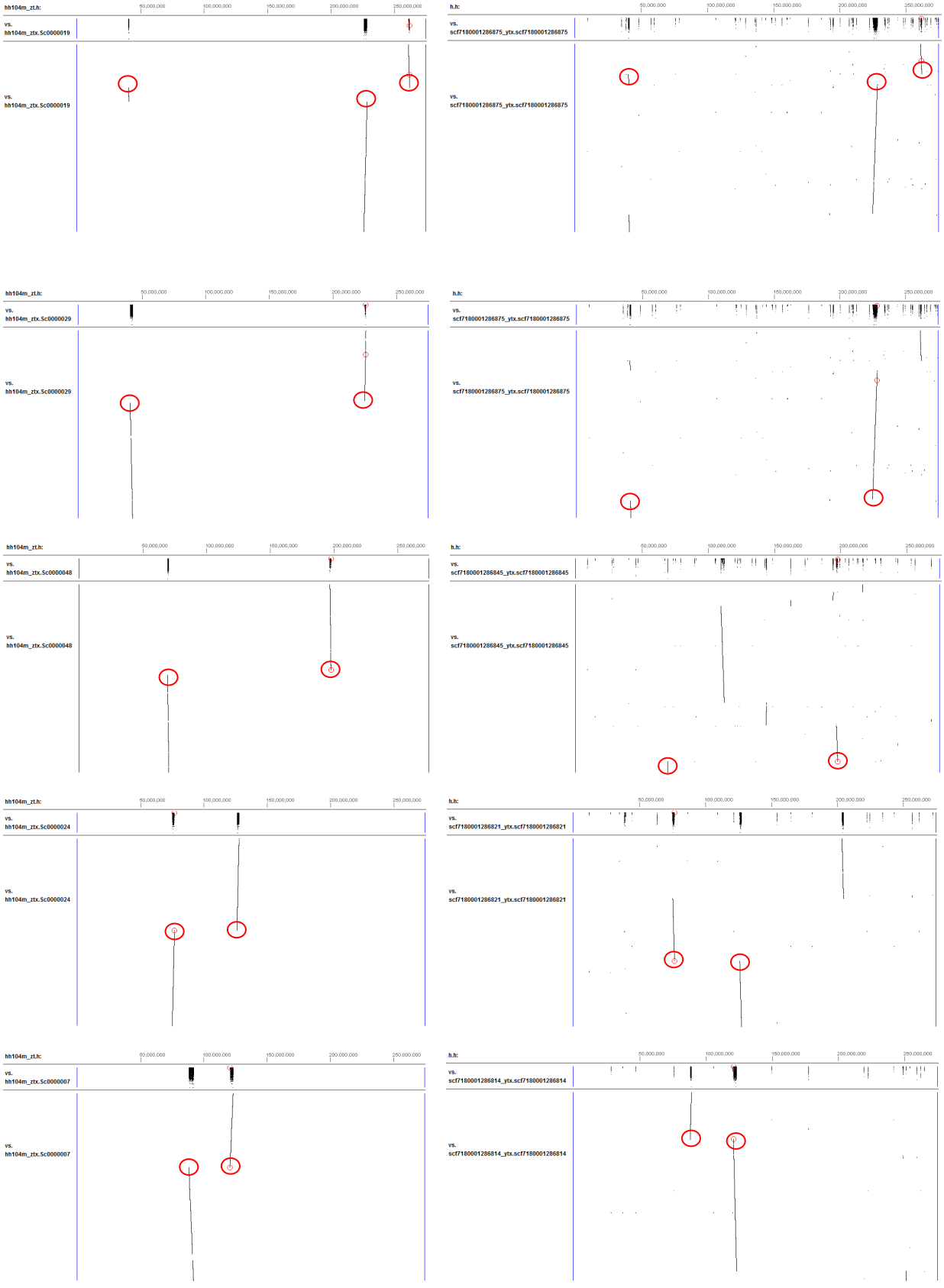


Figure S3 (continued). Alignment dot plots show all 24 major (>50Kb) assembly errors remained in the reference haploid assembly for the 274Mb simulated genome (on the left panels). The right panels show the corresponding errors in the original diploid assembly. The red circle shows the mis-joining point. Note that some scaffolds contains 2 errors.

Figure S3 (continued). Alignment dot plots show all 24 major (>50Kb) assembly errors remained in the reference haploid assembly for the 274Mb simulated genome (on the left panels). The right panels show the corresponding errors in the original diploid assembly. The red circle shows the mis-joining point. Note that some scaffolds contains 2 errors.

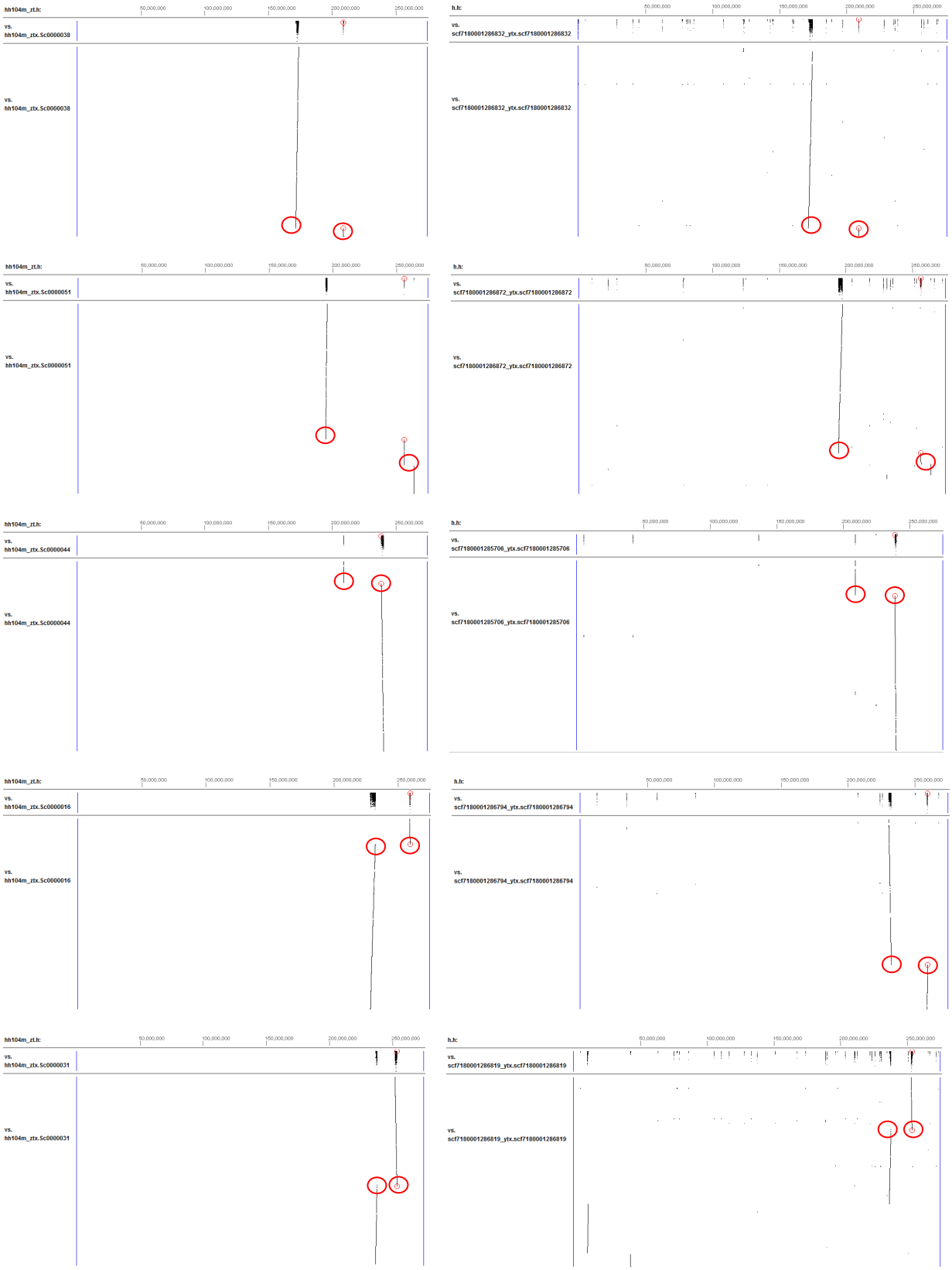


Table S1. The assemblies of twenty-five simulated small genomes.
Reference assemblies were created with HaploMerger. Assembly errors were identified by visual inspection of the whole-genome alignments.

	1	2	3	4	5	6	7	8	9	10	11	12	13
The diploild assembly													
TotalScaffolds	90	53	63	75	85	55	83	49	62	57	72	72	73
TotalContigsInScaffolds	362	260	286	284	336	231	356	287	253	282	325	321	394
TotalBasesInScaffolds	19251067	19197293	19240252	19276428	19240159	19234351	19281245	19247655	19238081	19253852	19243778	19221330	19231572
MaxBasesInScaffolds	1363298	3131518	2471614	2005797	1711220	2263581	977570	1745191	1593876	1745670	1408260	1546711	1701364
N25ScaffoldBases	953228	1146270	1232015	1237217	971957	1148119	659383	1399006	1206533	1493609	840090	808700	1390525
N50ScaffoldBases	580565	707931	816514	845456	603320	846030	499676	1241544	716685	865845	564923	607080	881444
N75ScaffoldBases	361771	552696	433556	393946	280161	527206	331411	623392	322467	578379	383503	441076	423902
MaxContigLength	775523	397980	463726	669077	389868	898979	421618	843010	875956	653936	455035	910619	553152
N25ContigBases	206656	231866	221499	246144	185381	375620	206677	290371	366760	270858	214327	269734	191897
N50ContigBases	126049	163005	142945	155065	127405	223444	133297	172581	184027	162157	129026	132474	111718
N75ContigBases	74822	103197	91224	84485	76563	119412	81447	97669	94452	106563	76162	78314	63006
TotalBigContigs	211	167	188	187	219	144	213	178	154	170	211	191	254
MeanBigContigLength	89564	113460	101016	101793	86425	131897	89350	106714	122889	111789	89639	99009	74036
TotalDegenContigs	3499	4191	4379	4581	7751	4025	3630	3596	4031	4180	6810	4639	4632
TotalSurrogates	13773	15039	13732	14103	18563	13780	13226	12646	14137	13975	17556	14018	15926
ReadsWithGoodMate	3478426(74.56%)	3463386(74.23%)	3068650(69.89%)	3065616(69.66%)	3087702(69.46%)	3120364(70.30%)	3167362(71.20%)	3143198(70.81%)	3136224(70.51%)	3151032(70.83%)	3187566(70.73%)	3067618(69.69%)	3435334(73.65%)
TotalUsableReads	4665159	4666003	4390800	4401138	4445257	4438424	4448333	4438749	4447837	4448993	4506703	4402053	4664147
The reference haploid assembly													
Ref_TotalScaffolds	4	5	4	3	3	4	3	3	4	3	4	4	7
Ref_TotalBasesInScaffolds	9689719	11227746	9803374	9784466	9821939	9789143	9803560	9931100	9879505	9888407	9746856	9631994	10079472
Ref_MaxBasesInScaffolds	5339283	7977394	4184853	6023700	5262238	3499574	7406262	4474079	6748227	5506290	5602758	4459990	3559311
Ref_N50ScaffoldBases	5339283	7977394	2294982	6023700	5262238	3143867	7412226	3819090	6797526	5505359	5602758	2538344	3420842
AssemblyErrors	2	1	4	1	2	3	1	2	1	3	1	1	3
AssemblyErrors_Inherited	2	1	4	1	2	3	1	2	1	3	1	1	3
AssemblyErrors_By_HaploMerger	0	0	0	0	0	0	0	0	0	0	0	0	0

	14	15	16	17	18	19	20	21	22	23	24	25
The diploild assembly												
TotalScaffolds	55	93	55	81	79	54	60	58	64	77	78	97
TotalContigsInScaffolds	288	407	290	307	361	328	282	284	314	287	309	394
TotalBasesInScaffolds	19251179	19207712	19220007	19310429	19264469	19198388	19288079	19216445	19236890	19256635	19226314	19181721
MaxBasesInScaffolds	1830840	1977620	1899802	1561133	1432292	2146984	1481720	1964301	2140323	1470030	2049474	1248517
N25ScaffoldBases	1637224	831624	1508819	721400	882495	1253690	1060919	1618230	1312928	1002614	1365977	645001
N50ScaffoldBases	936874	484506	910956	418657	579474	910986	626007	924190	1009754	561359	1182573	422641
N75ScaffoldBases	549529	291865	549346	269433	348866	566262	438296	533211	635294	284405	552347	267941
MaxContigLength	736802	409851	609139	539503	616266	539772	541445	644092	920556	746719	1039557	491025
N25ContigBases	265128	191026	300833	305304	195172	255310	235884	345807	299306	275563	269999	182134
N50ContigBases	164085	107873	150453	156690	116570	149064	161601	179944	188095	162313	179892	106375
N75ContigBases	95223	59754	85658	94780	65600	74877	78193	97250	87910	85388	106906	63237
TotalBigContigs	176	252	191	187	235	206	199	165	171	186	177	246
MeanBigContigLength	107920	74557	99380	101558	80602	91615	95711	114640	110593	102132	107144	76211
TotalDegenContigs	4192	7775	4480	4257	7409	4556	4130	3899	3643	4162	3882	8269
TotalSurrogates	15215	19793	16117	14057	17852	14165	14459	14642	14781	13648	14450	20756
ReadsWithGoodMate	3458746(74.14%)	3378258(72.40%)	3425634(73.46%)	3234446(71.76%)	3088824(69.64%)	3174038(70.98%)	3229272(71.66%)	3135286(70.52%)	3453474(74.05%)	3119682(70.28%)	3231368(71.74%)	3374192(72.34%)
TotalUsableReads	4664971	4665859	4663337	4507472	4435666	4472032	4506458	4445892	4663889	4439080	4504435	4664357
The reference haploid assembly												
Ref_TotalScaffolds	4	5	4	5	6	2	4	5	5	4	4	5
Ref_TotalBasesInScaffolds	9524882	9686420	9650706	9860222	9821037	9847995	9884899	10039934	10333767	9654473	9830787	9789745
Ref_MaxBasesInScaffolds	5856979	3419825	8015129	6299963	7027965	8164965	8581495	7514680	3601100	4824288	4125823	5518318
Ref_N50ScaffoldBases	5861198	3274740	8015129	6314090	7026623	8168334	8585898	7535570	3539499	3611371	2488781	5518318
AssemblyErrors	1	2	1	2	2	3	1	1	1	1	3	2
AssemblyErrors_Inherited	1	2	1	2	2	3	1	1	1	1	3	2
AssemblyErrors_By_HaploMerger	0	0	0	0	0	0	0	0	0	0	0	0

Table S2. A comparison of the completeness of difference assemblies (a total of 52961 EST contigs encoding single proteins were used for analysis).

		Full-length transcripts*			Coding regions**			5'-UTRs***			3'-UTRs***		
	Cut-off values->	cov85%; id85%	cov80%; id80%	cov80%; id70%	cov85%; id85%	cov80%; id80%	cov80%; id70%	cov85%; id85%; len100	cov80%; id80%; len50	cov80%; id70%; len50	cov85%; id85%; len100	cov80%; id80%; len50	cov80%; id70%; len50
The original diploid assembly	number	45014	46191	46395	45071	46494	46700	8771	12873	12905	23794	26465	26539
	%	85.0	87.2	87.6	85.1	87.8	88.2	16.6	24.3	24.4	44.9	50.0	50.1
Haploid	number	43723	45305	45680	43533	45359	45755	8693	12754	12809	22936	25734	25897
	%	-2.4	-1.7	-1.4	-2.9	-2.1	-1.8	-0.1	-0.2	-0.2	-1.6	-1.4	-1.2
Haploid+unpaired	number	44128	45685	46051	44001	45806	46189	8754	12841	12895	23144	25959	26120
	%	-1.7	-1.0	-0.6	-2.0	-1.3	-1.0	0.0	-0.1	0.0	-1.2	-1.0	-0.8
Haploid+gap filling	number	43778	45355	45732	43656	45467	45869	8717	12780	12833	22987	25776	25947
	%	-2.3	-1.6	-1.3	-2.7	-1.9	-1.6	-0.1	-0.2	-0.1	-1.5	-1.3	-1.1
Haploid+gap filling+unpaired	number	44205	45756	46111	44144	45933	46308	8781	12873	12925	23200	26003	26170
	%	-1.5	-0.8	-0.5	-1.8	-1.1	-0.7	0.0	0.0	0.0	-1.1	-0.9	-0.7
Haploid+long haplotype	number	44091	45627	45968	43924	45696	46049	8758	12836	12900	23090	25934	26083
	%	-1.7	-1.1	-0.8	-2.2	-1.5	-1.2	0.0	-0.1	0.0	-1.3	-1.0	-0.9
Haploid+long haplotype +unpaired	number	44460	45963	46297	44332	46059	46403	8811	12909	12971	23282	26144	26293
	%	-1.0	-0.4	-0.2	-1.4	-0.8	-0.6	0.1	0.1	0.1	-1.0	-0.6	-0.5
Haploid+long haplotype+gap filling	number	44042	45582	45930	43933	45699	46062	8754	12829	12891	23119	25941	26091
	%	-1.8	-1.1	-0.9	-2.1	-1.5	-1.2	0.0	-0.1	0.0	-1.3	-1.0	-0.8
Haploid+long haplotype+gap filling+unpaired	number	44431	45943	46276	44365	46092	46437	8807	12905	12965	23320	26162	26310
	%	-1.1	-0.5	-0.2	-1.3	-0.8	-0.5	0.1	0.1	0.1	-0.9	-0.6	-0.4

* Full-length transcripts refer to EST contigs; cov=full-length sequence sequence coverage including gaps, indels and Ns; id=identity.

** cov=sequence coverage for coding regions, including gaps, indels and Ns; id=identity.

*** cov=sequence coverage for UTRs, including gaps, indels and Ns; id=identity; len=the sequence length of UTRs.

Haploid = the initial reference haploid assembly; +unpaired = add unpaired sequences; +gap filling = implement the refining operation of gap filling; +long haplotype = implement the refining operation of selecting long haplotypes.

Haploid+long haplotype+gap filling+unpaired = the reference haploid assembly reported in the text and Table 1, with all three refining operation done.

Table S3. The difference for transcript alignments between the original and the reference assemblies.

	Exon number difference		Alignment length difference*			
	full-length	coding region	full-length 5%	full-length 10%	full-length 15%	full-length 20%
transcript coverage=80% and identity=80% (46191 transcripts)**						
more exons or alignment length on bbv7xt, number***	2538	1905	3133	2065	1659	1402
more exons or alignment length on bbv7xt, %	5.5	4.1	6.8	4.5	3.6	3.0
more exons or alignment length on bbv7hlg, number****	1961	1013	615	355	199	114
more exons or alignment length on bbv7hlg, %	4.2	2.2	1.3	0.8	0.4	0.2
total difference, %	9.7	6.3	8.1	5.2	4.0	3.3
transcript coverage=85% and identity=85% (45014 transcripts)						
more exons or alignment length on bbv7xt, number	2408	1789	2959	1910	1523	1275
more exons or alignment length on bbv7xt, %	5.3	4.0	6.6	4.2	3.4	2.8
more exons or alignment length on bbv7hlg, number	1830	932	484	249	108	69
more exons or alignment length on bbv7hlg, %	4.1	2.1	1.1	0.6	0.2	0.2
total difference, %	9.4	6.0	7.6	4.8	3.6	3.0

* Alignment length excludes gaps, Ns and indels.

** Transcripts refer to EST contigs here, only those mapped to the genome sequences with 80% coverage and 80% identity were considered.

*** bbv7xt refers to the original diploid assembly.

**** bbv7hlg refers to the reference haploid assembly with N-gaps filled and long haplotypes selected.

Table S4. The allelic difference for transcript alignment against genome sequences.

	full-length transcript * (alignment length coverage>=80%**)		full-length transcripts (alignment length coverage>=60%)		5'-UTR (alignment length coverage>=60% ***, length>100bp****)		3'-UTR (alignment length coverage>=60%, length>100bp)		3'-UTR (alignment length coverage>=60%, length>300bp)	
number of transcripts	number	%	number	%	number	%	number	%	number	%
best hit to the same pair of alleles	31530	100	34268	100	6662	100	18600	100	12909	100
alignment length difference between two alleles <10%										
identity >=0%	28180	89.4	30311	88.5	5285	79.3	14661	78.8	10409	80.6
identity >=90%	28038	88.9	30141	88.0	5231	78.5	14249	76.6	10141	78.6
identity >=95%	26324	83.5	28323	82.7	4636	69.6	12294	66.1	8784	68.0
alignment length difference between two alleles <5%										
identity >=0%	26014	82.5	27960	81.6	4743	71.2	12681	68.2	9052	70.1
identity >=90%	25971	82.4	27901	81.4	4723	70.9	12564	67.5	8987	69.6
identity >=95%	25243	80.1	27123	79.1	4481	67.3	11602	62.4	8317	64.4
exon number difference between two alleles <1										
identity >=0%	25174	79.8	27112	79.1						
identity >=90%	24801	78.7	26686	77.9						
identity >=95%	23423	74.3	25215	73.6						

* Full-length transcript is referred to a EST contig.

** Alignment length coverage is referred to the proportion of alignment length (gaps, Ns, indels excluded) to the full transcript length.

*** Alignment length coverage for 5'-UTR and 3'-UTR is referred to the alignment length coverage of the full-length transcript.

**** Length here is referred to the length of the 5'-UTR or 3'-UTR.