Supplementary Tables for “Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing”

Supplementary Table 1: Raw sequence data read length statistics

Col-0 Dataset, General Statistics
Statistics for all wells of length 500 bases or more

2,309,888 reads out of 2,348,820 (98.3%)
15,218,029,983 base pairs out of 15,226,098,559 (99.9%)
6,588 average read length
4,478 standard deviation

Col-0: Distribution of Read Lengths (Bin size = 1,000)				
Bin:	Count	% Reads	% Bases	Average
41,000:	1	0.0	0.0	41981
40,000:	2	0.0	0.0	40886
39,000:	2	0.0	0.0	40345
38,000:	1	0.0	0.0	40029
37,000:	3	0.0	0.0	39084
36,000:	10	0.0	0.0	37707
35,000:	13	0.0	0.0	36771
34,000:	33	0.0	0.0	35602
33,000:	34	0.0	0.0	34864
32,000:	59	0.0	0.0	33962
31,000:	97	0.0	0.1	32998
30,000:	125	0.0	0.1	32152
29,000:	209	0.0	0.1	31203
28,000:	381	0.0	0.2	30120
27,000:	491	0.1	0.3	29227
26,000:	700	0.1	0.4	28336
25,000:	1,073	0.1	0.6	27385
24,000:	1,632	0.2	0.8	26406
23,000:	2,357	0.3	1.2	25449
22,000:	3,379	0.5	1.7	24498
21,000:	4,755	0.7	2.4	23561
20,000:	6,890	1.0	3.3	22603
19,000:	9,414	1.4	4.5	21673
18,000:	12,897	1.9	6.1	20747

17,000:	17,989	2.7	8.1	19805
16,000:	24,912	3.8	10.8	18856
15,000:	32,799	5.2	14.2	17933
14,000:	43,025	7.1	18.3	17022
13,000:	56,405	9.5	23.3	16112
12,000:	71,645	12.6	29.1	15219
11,000:	88,411	16.4	35.8	14349
10,000:	107,611	21.1	43.2	13495
9,000:	127,238	26.6	51.2	12665
8,000:	140,205	32.7	59.0	11890
7,000:	153,409	39.3	66.5	11147
6,000:	173,893	46.8	73.9	10398
5,000:	196,392	55.3	81.0	9644
4,000:	220,778	64.9	87.5	8885
3,000:	234,022	75.0	92.9	8158
2,000:	245,308	85.7	96.9	7455
1,000:	281,860	97.9	99.7	6714
0:00	49,428	100.0	100.0	6588

Cvi-0 Dataset, General Statistics
Statistics for all wells of length 500 bases or more
2,411,099 reads out of 2,440,205 (98.8%)
14,777,982,500 base pairs out of 14,785,365,866 (100.0%)
6,129 average read length
4,577 standard deviation

Cvi-0: Distribution of Read Lengths (Bin size = 1,000)				
Bin:	Count	% Reads	% Bases	Average
44,000:	1	0.0	0.0	44472
43,000:	0	0.0	0.0	44472
42,000:	0	0.0	0.0	44472
41,000:	0	0.0	0.0	44472
40,000:	0	0.0	0.0	44472
39,000:	2	0.0	0.0	41040
38,000:	0	0.0	0.0	41040
37,000:	1	0.0	0.0	40162
36,000:	3	0.0	0.0	38599
35,000:	7	0.0	0.0	37006
34,000:	20	0.0	0.0	35480
33,000:	25	0.0	0.0	34606

32,000:	55	0.0	0.0	33560
31,000:	73	0.0	0.0	32723
30,000:	111	0.0	0.1	31886
29,000:	207	0.0	0.1	30878
28,000:	323	0.0	0.2	29927
27,000:	442	0.1	0.2	29068
26,000:	670	0.1	0.4	28163
25,000:	1,001	0.1	0.5	27245
24,000:	1,488	0.2	0.8	26309
23,000:	2,204	0.3	1.1	25367
22,000:	3,364	0.4	1.7	24391
21,000:	4,788	0.6	2.3	23443
20,000:	6,740	0.9	3.3	22513
19,000:	9,582	1.3	4.5	21577
18,000:	13,036	1.8	6.2	20661
17,000:	17,795	2.6	8.3	19745
16,000:	24,377	3.6	11.0	18822
15,000:	33,570	5.0	14.5	17885
14,000:	44,442	6.8	18.9	16964
13,000:	58,152	9.2	24.2	16053
12,000:	72,843	12.2	30.3	15172
11,000:	85,481	15.8	37.0	14345
10,000:	95,952	19.8	43.8	13569
9,000:	108,327	24.3	50.7	12813
8,000:	123,427	29.4	57.8	12060
7,000:	140,224	35.2	64.9	11305
6,000:	162,836	42.0	72.1	10529
5,000:	187,662	49.7	79.0	9740
4,000:	205,764	58.3	85.3	8972
3,000:	225,114	67.6	90.6	8215
2,000:	285,609	79.5	95.4	7358
1,000:	409,941	96.5	99.5	6324
0:00	85,440	100.0	100.0	6129

Cvi-0 x Col-0 Dataset, General Statistics
Statistics for all wells of length 500 bases or more
1,607,065 reads out of 1,637,256 (98.2%)
18,543,189,547 base pairs out of 18,551,762,921 (100.0%)

11,538 average read length
8,219 standard deviation

Cvi-0 x Col-0 F1: Distribution of Read Lengths (Bin size = 1,000)				
88,000:	1	0	0	88007
87,000:	0	0	0	88007
86,000:	0	0	0	88007
85,000:	0	0	0	88007
84,000:	0	0	0	88007
83,000:	0	0	0	88007
82,000:	0	0	0	88007
81,000:	0	0	0	88007
80,000:	0	0	0	88007
79,000:	0	0	0	88007
78,000:	0	0	0	88007
77,000:	0	0	0	88007
76,000:	0	0	0	88007
75,000:	1	0	0	81965
74,000:	0	0	0	81965
73,000:	0	0	0	81965
72,000:	0	0	0	81965
71,000:	0	0	0	81965
70,000:	2	0	0	76109
69,000:	1	0	0	74708
68,000:	1	0	0	73669
67,000:	4	0	0	71283
66,000:	4	0	0	69938
65,000:	7	0	0	68402
64,000:	4	0	0	67755
63,000:	10	0	0	66522
62,000:	11	0	0	65588
61,000:	13	0	0	64710
60,000:	20	0	0	63628
59,000:	18	0	0	62848
58,000:	19	0	0	62115
57,000:	38	0	0.1	60976
56,000:	43	0	0.1	59993
55,000:	52	0	0.1	59047
54,000:	56	0	0.1	58209
53,000:	63	0	0.1	57410
52,000:	69	0	0.1	56637
51,000:	99	0	0.2	55677
50,000:	117	0	0.2	54754

49,000:	148	0	0.2	53778
48,000:	159	0.1	0.3	52903
47,000:	209	0.1	0.3	51937
46,000:	256	0.1	0.4	50954
45,000:	275	0.1	0.5	50071
44,000:	321	0.1	0.5	49181
43,000:	443	0.2	0.6	48154
42,000:	506	0.2	0.8	47189
41,000:	623	0.2	0.9	46202
40,000:	776	0.3	1.1	45185
39,000:	893	0.3	1.3	44219
38,000:	1,226	0.4	1.5	43135
37,000:	1,477	0.5	1.8	42086
36,000:	1,717	0.6	2.1	41091
35,000:	2,151	0.7	2.6	40072
34,000:	2,480	0.9	3	39102
33,000:	2,991	1.1	3.6	38131
32,000:	3,679	1.3	4.2	37141
31,000:	4,637	1.6	5	36113
30,000:	5,806	2	5.9	35073
29,000:	7,042	2.4	7.1	34050
28,000:	9,048	3	8.5	32989
27,000:	11,084	3.6	10.1	31947
26,000:	13,961	4.5	12.1	30895
25,000:	17,564	5.6	14.5	29840
24,000:	21,597	7	17.4	28805
23,000:	26,774	8.6	20.7	27775
22,000:	33,490	10.7	24.8	26745
21,000:	41,499	13.3	29.6	25721
20,000:	51,643	16.5	35.3	24701
19,000:	61,925	20.4	41.8	23713
18,000:	64,688	24.4	48.3	22853
17,000:	52,629	27.7	53.3	22221
16,000:	47,252	30.6	57.5	21671
15,000:	48,324	33.6	61.5	21118
14,000:	50,349	36.7	65.4	20553
13,000:	53,757	40.1	69.3	19964
12,000:	58,400	43.7	73.3	19343
11,000:	63,543	47.7	77.2	18692
10,000:	69,103	52	81.1	18013
9,000:	73,373	56.5	84.9	17325

8,000:	74,037	61.1	88.3	16660
7,000:	66,903	65.3	91	16077
6,000:	58,125	68.9	93	15575
5,000:	55,580	72.4	94.7	15093
4,000:	56,669	75.9	96	14601
3,000:	60,483	79.7	97.2	14076
2,000:	81,357	84.7	98.3	13382
1,000:	202,051	97.3	99.8	11837
0:00	43,389	100	100	11538

V. vinifera Dataset, General Statistics
Statistics for all wells of length 500 bases or more
6,877,303 reads out of 7,001,805 (98.2%)
73,666,711,100 base pairs out of 73,701,175,777 (100.0%)

10,711 average read length
7,749 standard deviation

V. vinifera: Distribution of Read Lengths (Bin size = 1,000)				
Bin:	Count	% Reads	% Bases	Average
79,000:	1	0	0	79230
78,000:	0	0	0	79230
77,000:	1	0	0	78345
76,000:	0	0	0	78345
75,000:	2	0	0	76683
74,000:	0	0	0	76683
73,000:	1	0	0	75966
72,000:	2	0	0	74897
71,000:	6	0	0	73320
70,000:	2	0	0	72945
69,000:	4	0	0	72260
68,000:	11	0	0	70913
67,000:	7	0	0	70233
66,000:	12	0	0	69333
65,000:	8	0	0	68812
64,000:	12	0	0	68095
63,000:	30	0	0	66699
62,000:	23	0	0	65903
61,000:	33	0	0	64945

60,000:	44	0	0	63960
59,000:	54	0	0	62990
58,000:	56	0	0	62163
57,000:	65	0	0	61347
56,000:	91	0	0	60397
55,000:	115	0	0	59424
54,000:	144	0	0.1	58449
53,000:	164	0	0.1	57534
52,000:	213	0	0.1	56562
51,000:	268	0	0.1	55572
50,000:	317	0	0.1	54615
49,000:	403	0	0.2	53624
48,000:	535	0	0.2	52571
47,000:	654	0	0.2	51557
46,000:	812	0.1	0.3	50549
45,000:	1,014	0.1	0.3	49544
44,000:	1,325	0.1	0.4	48502
43,000:	1,644	0.1	0.5	47479
42,000:	2,085	0.1	0.6	46451
41,000:	2,690	0.2	0.8	45411
40,000:	3,326	0.2	1	44397
39,000:	4,097	0.3	1.2	43404
38,000:	5,004	0.4	1.5	42429
37,000:	6,310	0.5	1.8	41441
36,000:	7,952	0.6	2.2	40444
35,000:	9,547	0.7	2.6	39478
34,000:	11,761	0.9	3.2	38511
33,000:	14,247	1.1	3.8	37556
32,000:	17,432	1.3	4.6	36601
31,000:	20,967	1.7	5.5	35655
30,000:	25,181	2	6.5	34716
29,000:	30,584	2.5	7.8	33771
28,000:	36,913	3	9.2	32824
27,000:	44,021	3.6	10.8	31884
26,000:	52,622	4.4	12.7	30946
25,000:	62,097	5.3	14.9	30016
24,000:	72,809	6.4	17.3	29096
23,000:	85,331	7.6	20	28181
22,000:	98,502	9	23	27279
21,000:	113,186	10.7	26.3	26387
20,000:	128,635	12.6	29.9	25508

19,000:	144,348	14.7	33.7	24646
18,000:	163,242	17	37.8	23788
17,000:	183,152	19.7	42.2	22936
16,000:	207,348	22.7	46.8	22079
15,000:	235,653	26.1	51.8	21215
14,000:	265,938	30	57	20348
13,000:	288,364	34.2	62.3	19508
12,000:	295,618	38.5	67.3	18725
11,000:	295,964	42.8	71.9	17998
10,000:	300,385	47.2	76.2	17304
9,000:	310,065	51.7	80.2	16622
8,000:	319,726	56.3	83.9	15952
7,000:	330,239	61.1	87.2	15287
6,000:	340,085	66.1	90.2	14629
5,000:	351,942	71.2	92.9	13973
4,000:	370,169	76.6	95.1	13306
3,000:	406,510	82.5	97	12603
2,000:	492,915	89.6	98.7	11792
1,000:	546,266	97.6	99.8	10957
0:00	166,002	100	100	10711

Clavicorona pyxidata Dataset, General Statistics
369,622 reads out of 377,554 (97.9%)
4,084,017,637 base pairs out of 4,086,359,932 (99.9%)
11,049 average read length
7,560 standard deviation

Clavicorona pyxidata data set: Distribution of Read Lengths (Bin size = 1,000)				
Bin:	Count	% Reads	% Bases	Average
66,000:	1	0	0	66177
65,000:	0	0	0	66177
64,000:	0	0	0	66177
63,000:	1	0	0	65056
62,000:	0	0	0	65056
61,000:	2	0	0	63210
60,000:	1	0	0	62741
59,000:	2	0	0	61747
58,000:	2	0	0	61080
57,000:	1	0	0	60758
56,000:	3	0	0	59751

55,000:	5	0	0	58531
54,000:	7	0	0	57405
53,000:	6	0	0	56613
52,000:	4	0	0	56130
51,000:	10	0	0.1	55108
50,000:	12	0	0.1	54126
49,000:	12	0	0.1	53311
48,000:	14	0	0.1	52489
47,000:	21	0	0.1	51463
46,000:	36	0	0.2	50182
45,000:	33	0	0.2	49280
44,000:	50	0.1	0.3	48204
43,000:	66	0.1	0.3	47113
42,000:	75	0.1	0.4	46156
41,000:	96	0.1	0.5	45180
40,000:	125	0.2	0.6	44174
39,000:	172	0.2	0.8	43113
38,000:	222	0.3	1	42064
37,000:	260	0.3	1.2	41098
36,000:	318	0.4	1.5	40159
35,000:	408	0.5	1.9	39189
34,000:	484	0.7	2.3	38258
33,000:	633	0.8	2.8	37277
32,000:	746	1	3.4	36342
31,000:	966	1.3	4.2	35360
30,000:	1,177	1.6	5	34396
29,000:	1,422	2	6.1	33449
28,000:	1,830	2.5	7.3	32461
27,000:	2,121	3.1	8.8	31529
26,000:	2,622	3.8	10.5	30580
25,000:	3,293	4.7	12.5	29607
24,000:	3,974	5.7	14.9	28648
23,000:	4,786	7	17.6	27699
22,000:	5,857	8.6	20.9	26741
21,000:	7,033	10.5	24.6	25790
20,000:	8,750	12.9	29	24815
19,000:	10,384	15.7	33.9	23862
18,000:	12,035	19	39.4	22940
17,000:	12,592	22.4	44.8	22111
16,000:	12,586	25.8	49.8	21369
15,000:	12,822	29.2	54.7	20672

14,000:	12,869	32.7	59.3	20015
13,000:	13,527	36.4	63.7	19359
12,000:	14,116	40.2	68.1	18707
11,000:	15,320	44.3	72.4	18032
10,000:	16,476	48.8	76.6	17344
9,000:	17,543	53.5	80.7	16648
8,000:	18,445	58.5	84.5	15953
7,000:	19,182	63.7	88	15264
6,000:	19,208	68.9	91.1	14603
5,000:	19,115	74.1	93.7	13968
4,000:	19,933	79.5	95.9	13325
3,000:	20,392	85	97.6	12687
2,000:	21,739	90.9	98.9	12027
1,000:	23,792	97.3	99.8	11332
0:00	9,887	100	100	11049

Supplementary Table 2. Concordance of *Arabidopsis* TAIR10 with Falcon assembly of *Arabidopsis* Col-0

Ref start	Ref end	Query start	Query end	R-alignment len	Q-alignmenr leng	Alignment Identity	Reference length	Query length	Reference name	Query name
7078998	7349609	270614	1	270612	270614	99.99	7352871	30427671	000006F|quiver	Chr1
4891941	7078731	2457266	270496	2186791	2186771	99.99	7352871	30427671	000006F|quiver	Chr1
1144744	1157712	6191529	6204497	12969	12969	99.95	7352871	30427671	000006F|quiver	Chr1
1151846	4886129	6197390	2463078	3734284	3734313	99.99	7352871	30427671	000006F|quiver	Chr1
1	1150604	7349246	6198632	1150604	1150615	99.99	7352871	30427671	000006F|quiver	Chr1
1325616	1403650	7437344	7359313	78035	78032	99.95	1403650	30427671	000017F|quiver	Chr1
1086349	1326124	7675971	7436196	239776	239776	99.99	1403650	30427671	000017F|quiver	Chr1
215492	1086247	8546721	7675959	870756	870763	99.99	1403650	30427671	000017F|quiver	Chr1
1	215363	8762075	8546715	215363	215361	99.99	1403650	30427671	000017F|quiver	Chr1
1	160129	8840477	9000590	160129	160114	99.99	3935192	30427671	000011F|quiver	Chr1
160081	531303	9000591	9371815	371223	371225	99.99	3935192	30427671	000011F|quiver	Chr1
531380	3052185	9371718	11892533	2520806	2520816	99.99	3935192	30427671	000011F|quiver	Chr1
3051726	3935192	11891849	12775311	883467	883463	99.99	3935192	30427671	000011F|quiver	Chr1
1	388699	12812556	13201252	388699	388697	99.99	1278374	30427671	000020F|quiver	Chr1
388696	1180431	13201353	13993095	791736	791743	99.99	1278374	30427671	000020F|quiver	Chr1
1180423	1274678	13993196	14087450	94256	94255	99.99	1278374	30427671	000020F|quiver	Chr1
208199	254729	14139823	14093292	46531	46532	99.99	254729	30427671	000033F|quiver	Chr1
92090	208159	14255823	14139755	116070	116069	99.99	254729	30427671	000033F|quiver	Chr1
55425	91727	14292120	14255818	36303	36303	100	254729	30427671	000033F|quiver	Chr1
41349	58910	14309681	14292121	17562	17561	99.99	254729	30427671	000033F|quiver	Chr1
379	34445	14343812	14309746	34067	34067	100	254729	30427671	000033F|quiver	Chr1
1	66899	14394916	14461807	66899	66892	99.87	66899	30427671	000040F|quiver	Chr1
1	47412	14461808	14509220	47412	47413	99.95	109436	30427671	000036F|quiver	Chr1
57618	84465	14581720	14608570	26848	26851	99.74	109436	30427671	000036F|quiver	Chr1
27627	48531	14685672	14664773	20905	20900	99.92	288146	30427671	000031F|quiver	Chr1
1	27504	14713209	14685704	27504	27506	99.98	288146	30427671	000031F|quiver	Chr1
87024	99326	14745975	14733674	12303	12302	99.89	109436	30427671	000036F|quiver	Chr1
47660	126278	14864080	14942697	78619	78618	99.67	288146	30427671	000031F|quiver	Chr1
115496	148597	14929379	14962555	33102	33177	96.39	288146	30427671	000031F|quiver	Chr1
158076	174997	14968509	14985435	16922	16927	97.63	288146	30427671	000031F|quiver	Chr1
167323	221752	14976654	15031086	54430	54433	98.15	288146	30427671	000031F|quiver	Chr1
225384	246942	15025629	15047216	21559	21588	98.89	288146	30427671	000031F|quiver	Chr1
13269	23903	15042083	15052791	10635	10709	98.37	28341	30427671	000055F|quiver	Chr1
1	30590	15055154	15085737	30590	30584	99.88	30605	30427671	000054F|quiver	Chr1
1	17372	15104420	15087046	17372	17375	99.8	31698	30427671	000051F|quiver	Chr1
10894	25271	15148808	15134443	14378	14366	97.77	31698	30427671	000051F|quiver	Chr1
1	34047	15178509	15144444	34047	34066	99.9	34047	30427671	000050F|quiver	Chr1
48460	74989	15215965	15189428	26530	26538	99.95	481716	30427671	000026F|quiver	Chr1
15011	49657	15215966	15250682	34647	34717	98.62	60279	30427671	000041F|quiver	Chr1
439897	469971	15290002	15259919	30075	30084	99.86	469973	30427671	000028F|quiver	Chr1
416769	459922	15313131	15269984	43154	43148	99.98	469973	30427671	000028F|quiver	Chr1
47896	60279	15346528	15359009	12384	12482	97.55	60279	30427671	000041F|quiver	Chr1
406034	419491	15424579	15411105	13458	13475	99.35	469973	30427671	000028F|quiver	Chr1
340305	377827	15474161	15436643	37523	37519	99.64	469973	30427671	000028F|quiver	Chr1
175477	358718	15645930	15462688	183242	183243	99.96	469973	30427671	000028F|quiver	Chr1
1	175382	15821252	15645865	175382	175388	99.99	469973	30427671	000028F|quiver	Chr1
988769	1207586	16053016	15834196	218818	218821	99.98	1207586	30427671	000022F|quiver	Chr1
813951	988766	16227776	16052939	174816	174838	99.92	1207586	30427671	000022F|quiver	Chr1
746514	824594	16291877	16213797	78081	78081	99.72	1207586	30427671	000022F|quiver	Chr1
731138	743043	16296261	16284350	11906	11912	99.78	1207586	30427671	000022F|quiver	Chr1
743071	754042	16307255	16296289	10972	10967	99.75	1207586	30427671	000022F|quiver	Chr1
299963	734577	16738427	16303816	434615	434612	99.98	1207586	30427671	000022F|quiver	Chr1
21004	299937	17016715	16737793	278934	278923	99.99	1207586	30427671	000022F|quiver	Chr1
1	41028	17037718	16996697	41028	41022	99.99	1207586	30427671	000022F|quiver	Chr1
1	91641	17046923	17138573	91641	91651	99.91	4248075	30427671	000010F|quiver	Chr1
88850	295844	17142724	17349749	206995	207026	99.92	4248075	30427671	000010F|quiver	Chr1
275785	585461	17329682	17639368	309677	309687	99.99	4248075	30427671	000010F|quiver	Chr1
585453	964409	17639469	18018425	378957	378957	99.99	4248075	30427671	000010F|quiver	Chr1
968178	1709890	18013162	18754885	741713	741724	99.99	4248075	30427671	000010F|quiver	Chr1
1689856	2407900	18734845	19452896	718045	718052	99.99	4248075	30427671	000010F|quiver	Chr1
2408216	2501755	19452852	19546433	93540	93582	99.93	4248075	30427671	000010F|quiver	Chr1
2507169	2723277	19531199	19747344	216109	216146	99.93	4248075	30427671	000010F|quiver	Chr1
2703292	3066150	19727350	20090217	362859	362868	99.99	4248075	30427671	000010F|quiver	Chr1
3066187	3451888	20090212	20475918	385702	385707	99.99	4248075	30427671	000010F|quiver	Chr1
3452050	3916879	20475910	20940729	464830	464820	99.99	4248075	30427671	000010F|quiver	Chr1
3920184	3955847	20944034	20979699	35664	35666	99.98	4248075	30427671	000010F|quiver	Chr1
3956213	4248075	20979650	21271521	291863	291872	99.99	4248075	30427671	000010F|quiver	Chr1
1	298479	21312897	21611374	298479	298478	99.97	471830	30427671	000027F|quiver	Chr1
298035	467221	21607610	21776799	169187	169190	99.99	471830	30427671	000027F|quiver	Chr1
1	231218	21830915	22062137	231218	231223	99.99	8608368	30427671	000004F|quiver	Chr1

229872	4070856	22059127	25900116	3840985	3840990	99.99	8608368	30427671	000004F|quiver	Chr1
4050864	6708196	25880118	28537489	2657333	2657372	99.99	8608368	30427671	000004F|quiver	Chr1
6708192	6960071	28537590	28789478	251880	251889	99.99	8608368	30427671	000004F|quiver	Chr1
6959654	7476465	28788838	29305647	516812	516810	99.99	8608368	30427671	000004F|quiver	Chr1
7477803	7522330	29300709	29345245	44528	44537	99.84	8608368	30427671	000004F|quiver	Chr1
7522413	8604838	29345233	30427671	1082426	1082439	99.99	8608368	30427671	000004F|quiver	Chr1
4042	31492	28151	55627	27451	27477	99.87	3248763	19698289	000012F|quiver	Chr2
34624	876102	49788	891285	841479	841498	99.99	3248763	19698289	000012F|quiver	Chr2
876756	909102	891939	924284	32347	32346	99.99	3248763	19698289	000012F|quiver	Chr2
909232	2182823	924279	2197880	1273592	1273602	99.99	3248763	19698289	000012F|quiver	Chr2
2182532	2527744	2197060	2542272	345213	345213	99.99	3248763	19698289	000012F|quiver	Chr2
2528046	2614941	2542216	2629114	86896	86899	99.99	3248763	19698289	000012F|quiver	Chr2
2615223	2666716	2629109	2680595	51494	51487	99.96	3248763	19698289	000012F|quiver	Chr2
2667027	3097575	2680354	3110909	430549	430556	99.99	3248763	19698289	000012F|quiver	Chr2
3097744	3248763	3110724	3261751	151020	151028	99.98	3248763	19698289	000012F|quiver	Chr2
1	21020	3364793	3385826	21020	21034	99.85	21020	19698289	000060F|quiver	Chr2
16156	46452	3375632	3405894	30297	30263	99.77	46452	19698289	000045F|quiver	Chr2
32707	97636	3571884	3506953	64930	64932	99.99	97636	19698289	000037F|quiver	Chr2
1	32708	3604206	3571500	32708	32707	99.99	97636	19698289	000037F|quiver	Chr2
1006206	1377912	4005117	3633424	371707	371694	99.92	1399302	19698289	000018F|quiver	Chr2
783706	884384	4092541	3991886	100679	100656	99.96	884384	19698289	000023F|quiver	Chr2
226007	803682	4650238	4072593	577676	577646	99.99	884384	19698289	000023F|quiver	Chr2
90866	225983	4785358	4650212	135118	135147	99.97	884384	19698289	000023F|quiver	Chr2
1	110848	4876224	4765348	110848	110877	99.97	884384	19698289	000023F|quiver	Chr2
1	111563	4885389	4996951	111563	111563	99.99	312162	19698289	000030F|quiver	Chr2
111734	312161	4996915	5197342	200428	200428	99.99	312162	19698289	000030F|quiver	Chr2
410560	481716	5276667	5205497	71157	71171	99.94	481716	19698289	000026F|quiver	Chr2
94940	409480	5572058	5257490	314541	314569	99.97	481716	19698289	000026F|quiver	Chr2
6986697	7015408	5589101	5560386	28712	28716	99.94	7015408	19698289	000008F|quiver	Chr2
6625687	6986669	5950039	5589048	360983	360992	99.99	7015408	19698289	000008F|quiver	Chr2
6562929	6625831	6012456	5949561	62903	62896	99.95	7015408	19698289	000008F|quiver	Chr2
5955923	6583110	6619471	5992283	627188	627189	99.99	7015408	19698289	000008F|quiver	Chr2
5836896	5955968	6738546	6619472	119073	119075	99.99	7015408	19698289	000008F|quiver	Chr2
5565086	5836796	7010187	6738474	271711	271714	99.99	7015408	19698289	000008F|quiver	Chr2
4548577	5565067	8021663	7005164	1016491	1016500	99.99	7015408	19698289	000008F|quiver	Chr2
4188052	4548524	8382023	8021556	360473	360468	99.99	7015408	19698289	000008F|quiver	Chr2
1936903	4188191	10633155	8381942	2251289	2251214	99.99	7015408	19698289	000008F|quiver	Chr2
1918169	1947456	10651889	10622602	29288	29288	99.84	7015408	19698289	000008F|quiver	Chr2
1	1917990	12569883	10651873	1917990	1918011	99.99	7015408	19698289	000008F|quiver	Chr2
1	516520	12578849	13095378	516520	516530	99.99	2001157	19698289	000014F|quiver	Chr2
516801	1380669	13095175	13959056	863869	863882	99.99	2001157	19698289	000014F|quiver	Chr2
1382202	2001157	13959050	14578029	618956	618980	99.99	2001157	19698289	000014F|quiver	Chr2
1	1057239	14566543	15623786	1057239	1057244	99.99	5134905	19698289	000009F|quiver	Chr2
1057279	1282939	15623761	15849390	225661	225630	99.98	5134905	19698289	000009F|quiver	Chr2
1283038	1773267	15849360	16339598	490230	490239	99.99	5134905	19698289	000009F|quiver	Chr2
1773242	2536545	16339617	17102936	763304	763320	99.99	5134905	19698289	000009F|quiver	Chr2
2536645	3380148	17102929	17946444	843504	843516	99.99	5134905	19698289	000009F|quiver	Chr2
3382857	3506493	17946028	18069649	123637	123622	99.95	5134905	19698289	000009F|quiver	Chr2
3486510	4537408	18049682	19100540	1050899	1050859	99.99	5134905	19698289	000009F|quiver	Chr2
4537528	5134900	19100450	19697819	597373	597370	99.99	5134905	19698289	000009F|quiver	Chr2
10232144	12193761	1961732	104	1961618	1961629	99.99	12197075	23459830	000000F|quiver	Chr3
10111995	10232145	2079348	1959228	120151	120121	99.97	12197075	23459830	000000F|quiver	Chr3
8055240	10131973	4136119	2059400	2076734	2076720	99.99	12197075	23459830	000000F|quiver	Chr3
6853390	8055179	5337862	4136084	1201790	1201779	99.99	12197075	23459830	000000F|quiver	Chr3
4335779	6873406	7855486	5317872	2537628	2537615	99.99	12197075	23459830	000000F|quiver	Chr3
4127044	4339843	8073283	7860518	212800	212766	99.98	12197075	23459830	000000F|quiver	Chr3
3299009	4147061	8901305	8053302	848053	848004	99.99	12197075	23459830	000000F|quiver	Chr3
3092578	3298164	9106883	8901293	205587	205591	99.99	12197075	23459830	000000F|quiver	Chr3
3027612	3092586	9171882	9106911	64975	64972	99.99	12197075	23459830	000000F|quiver	Chr3
2746751	3027619	9452856	9171983	280869	280874	99.99	12197075	23459830	000000F|quiver	Chr3
2204536	2755019	9986808	9436328	550484	550481	99.99	12197075	23459830	000000F|quiver	Chr3
3451	18233	10545940	10560722	14783	14783	99.95	18233	23459830	361|quiver	Chr3
1634081	2224565	10557296	9966786	590485	590511	99.97	12197075	23459830	000000F|quiver	Chr3
1	14807	10560723	10575507	14807	14785	98.96	18233	23459830	361|quiver	Chr3
855441	1645415	11354159	10564173	789975	789987	99.99	12197075	23459830	000000F|quiver	Chr3
687864	850871	11522400	11359418	163008	162983	99.98	12197075	23459830	000000F|quiver	Chr3
608498	707833	11601766	11502462	99336	99305	99.95	12197075	23459830	000000F|quiver	Chr3
447775	609764	11761232	11599235	161990	161998	99.99	12197075	23459830	000000F|quiver	Chr3
40767	467804	12168252	11741195	427038	427058	99.99	12197075	23459830	000000F|quiver	Chr3
1	40567	12208744	12168178	40567	40567	99.99	12197075	23459830	000000F|quiver	Chr3
906201	1364645	12680434	12221989	458445	458446	99.99	1364645	23459830	000019F|quiver	Chr3
826695	926198	12759944	12660443	99504	99502	99.99	1364645	23459830	000019F|quiver	Chr3
723684	825948	12862195	12759927	102265	102269	99.99	1364645	23459830	000019F|quiver	Chr3
563254	743669	13022636	12842202	180416	180435	99.99	1364645	23459830	000019F|quiver	Chr3
182881	563255	13403107	13022737	380375	380371	99.99	1364645	23459830	000019F|quiver	Chr3
169991	182231	13415292	13403052	12241	12241	100	1364645	23459830	000019F|quiver	Chr3

1	169960	13585245	13415281	169960	169965	99.99	1364645	23459830	000019F|quiver	Chr3
1	89109	13596544	13685689	89109	89146	99.74	90176	23459830	000038F|quiver	Chr3
1	30630	13809497	13840151	30630	30655	99.2	39023	23459830	000048F|quiver	Chr3
14775	276624	14132041	13870192	261850	261850	99.99	288346	23459830	000032F|quiver	Chr3
1	20777	14162404	14141629	20777	20776	99.99	288346	23459830	000032F|quiver	Chr3
18990	30077	14192869	14203967	11088	11099	99.7	31939	23459830	000052F|quiver	Chr3
10119	20905	14194393	14205200	10787	10808	99.33	31939	23459830	000052F|quiver	Chr3
1	282748	14235095	14517844	282748	282750	99.99	9254201	23459830	000002F|quiver	Chr3
272710	897158	14507806	15132294	624449	624489	99.98	9254201	23459830	000002F|quiver	Chr3
917169	1523928	15132546	15739312	606760	606767	99.99	9254201	23459830	000002F|quiver	Chr3
1524080	7034002	15739177	21249142	5509923	5509966	99.99	9254201	23459830	000002F|quiver	Chr3
7014004	7496469	21229150	21711594	482466	482445	99.99	9254201	23459830	000002F|quiver	Chr3
7476466	7789968	21691602	22005064	313503	313463	99.98	9254201	23459830	000002F|quiver	Chr3
7769997	9244669	21985125	23459789	1474673	1474665	99.99	9254201	23459830	000002F|quiver	Chr3
1035849	1694392	659544	1001	658544	658544	99.99	1712756	18585056	000015F|quiver	Chr4
936348	1035623	758786	659517	99276	99270	99.99	1712756	18585056	000015F|quiver	Chr4
828213	956404	866927	738735	128192	128193	99.98	1712756	18585056	000015F|quiver	Chr4
1	828426	1694948	866516	828426	828433	99.99	1712756	18585056	000015F|quiver	Chr4
1748	171212	1739739	1909213	169465	169475	99.93	177588	18585056	000034F|quiver	Chr4
981247	1015859	1934152	1899590	34613	34563	99.16	1399302	18585056	000018F|quiver	Chr4
1	1001222	2915404	1914205	1001222	1001200	99.99	1399302	18585056	000018F|quiver	Chr4
143	59479	2958326	3017650	59337	59325	98.79	59479	18585056	000042F|quiver	Chr4
28550	58790	3053301	3023061	30241	30241	99.99	58790	18585056	000043F|quiver	Chr4
110021	152297	3099998	3057723	42277	42276	99.97	152297	18585056	000035F|quiver	Chr4
4	110326	3210300	3099999	110323	110302	99.87	152297	18585056	000035F|quiver	Chr4
1	25913	3219134	3245079	25913	25946	99.71	40192	18585056	000047F|quiver	Chr4
1	16043	3231147	3215031	16043	16117	99.29	16043	18585056	000074F|quiver	Chr4
21882	40192	3231961	3250305	18311	18345	99.1	40192	18585056	000047F|quiver	Chr4
373908	626050	3507966	3255803	252143	252164	99.98	626050	18585056	000025F|quiver	Chr4
343306	369181	3533832	3507956	25876	25877	99.96	626050	18585056	000025F|quiver	Chr4
220786	343905	3656043	3532934	123120	123110	99.98	626050	18585056	000025F|quiver	Chr4
203464	220012	3671924	3655376	16549	16549	99.86	626050	18585056	000025F|quiver	Chr4
1	203471	3874892	3671445	203471	203448	99.87	626050	18585056	000025F|quiver	Chr4
1	71253	3947540	3876287	71253	71254	99.99	71253	18585056	000039F|quiver	Chr4
6990773	7047406	4062866	4006232	56634	56635	99.3	7047406	18585056	000007F|quiver	Chr4
6470220	6994859	4583291	4058618	524640	524674	99.83	7047406	18585056	000007F|quiver	Chr4
6270402	6472635	4782139	4579906	202234	202234	99.96	7047406	18585056	000007F|quiver	Chr4
5446841	6256186	5586563	4777208	809346	809356	99.99	7047406	18585056	000007F|quiver	Chr4
5154323	5448862	5876111	5581567	294540	294545	99.98	7047406	18585056	000007F|quiver	Chr4
4371611	5173406	6676187	5874321	801796	801867	99.99	7047406	18585056	000007F|quiver	Chr4
3823127	4381573	7224669	6666225	558447	558445	99.99	7047406	18585056	000007F|quiver	Chr4
1864772	3843117	9182963	7204688	1978346	1978276	99.97	7047406	18585056	000007F|quiver	Chr4
1822441	1853899	9223909	9192434	31459	31476	99.88	7047406	18585056	000007F|quiver	Chr4
1725808	1842409	9320553	9203921	116602	116633	99.93	7047406	18585056	000007F|quiver	Chr4
1700541	1720810	9343615	9323346	20270	20270	99.97	7047406	18585056	000007F|quiver	Chr4
1	1697731	11043496	9345767	1697731	1697730	99.98	7047406	18585056	000007F|quiver	Chr4
1	115712	11093724	11209436	115712	115713	99.97	7496166	18585056	000005F|quiver	Chr4
114890	661889	11208536	11755551	547000	547016	99.99	7496166	18585056	000005F|quiver	Chr4
666402	1458787	11754866	12547283	792386	792418	99.99	7496166	18585056	000005F|quiver	Chr4
1438787	3138046	12527275	14226523	1699260	1699249	99.99	7496166	18585056	000005F|quiver	Chr4
3118040	7496166	14206543	18584640	4378127	4378098	99.99	7496166	18585056	000005F|quiver	Chr4
1	900838	5724	906564	900838	900841	99.99	11244802	26975502	000001F|quiver	Chr5
880806	965858	886537	971581	85053	85045	99.98	11244802	26975502	000001F|quiver	Chr5
966561	2605361	971392	2610123	1638801	1638732	99.99	11244802	26975502	000001F|quiver	Chr5
2606398	3368473	2610118	3372196	762076	762079	99.99	11244802	26975502	000001F|quiver	Chr5
3369147	4000850	3372146	4003846	631704	631701	99.99	11244802	26975502	000001F|quiver	Chr5
4006891	5780360	4002894	5776363	1773470	1773470	99.99	11244802	26975502	000001F|quiver	Chr5
5778356	7878819	5776359	7876830	2100464	2100472	99.99	11244802	26975502	000001F|quiver	Chr5
7879158	9664524	7876822	9662202	1785367	1785381	99.99	11244802	26975502	000001F|quiver	Chr5
9644524	11050293	9642208	11047996	1405770	1405789	99.99	11244802	26975502	000001F|quiver	Chr5
11030278	11186821	11027971	11184521	156544	156551	99.99	11244802	26975502	000001F|quiver	Chr5
13085	407947	11662555	11267687	394863	394869	99.99	407947	26975502	000029F|quiver	Chr5
1	13040	11675590	11662551	13040	13040	99.98	407947	26975502	000029F|quiver	Chr5
1	46415	11733135	11779522	46415	46388	99.81	52259	26975502	000044F|quiver	Chr5
1	14443	11801133	11786699	14443	14435	99.59	14443	26975502	000069F|quiver	Chr5
1	10887	11810723	11821602	10887	10880	99.25	15739	26975502	000075F|quiver	Chr5
1	23423	11927266	11903841	23423	23426	99.97	23423	26975502	000057F|quiver	Chr5
241804	286274	11990458	11945965	44471	44494	98.82	288146	26975502	000031F|quiver	Chr5
1	661249	12049413	12710659	661249	661247	99.99	780821	26975502	000024F|quiver	Chr5
662148	779150	12710436	12827448	117003	117013	99.99	780821	26975502	000024F|quiver	Chr5
758917	780821	12849930	12871838	21905	21909	98.01	780821	26975502	000024F|quiver	Chr5
1	274108	12891085	13165189	274108	274105	99.99	1440677	26975502	000016F|quiver	Chr5
254056	313092	13145143	13204170	59037	59028	99.91	1440677	26975502	000016F|quiver	Chr5
306241	545869	13195371	13435011	239629	239641	99.89	1440677	26975502	000016F|quiver	Chr5
545953	561070	13434417	13449555	15118	15139	95.38	1440677	26975502	000016F|quiver	Chr5
566394	1273916	13450628	14158161	707523	707534	99.87	1440677	26975502	000016F|quiver	Chr5

1274087	1440677	14158135	14324731	166591	166597	99.99	1440677	26975502	000016F|quiver	Chr5
5646	605659	14449557	15049574	600014	600018	99.97	1226169	26975502	000021F|quiver	Chr5
585639	1226169	15029547	15670079	640531	640533	99.99	1226169	26975502	000021F|quiver	Chr5
1	103675	15697254	15800919	103675	103666	99.99	2448551	26975502	000013F|quiver	Chr5
83678	706634	15780930	16403876	622957	622947	99.93	2448551	26975502	000013F|quiver	Chr5
703088	2166278	16395164	17858350	1463191	1463187	99.99	2448551	26975502	000013F|quiver	Chr5
2146303	2433457	17838408	18125530	287155	287123	99.99	2448551	26975502	000013F|quiver	Chr5
8626061	8822328	18321767	18125491	196268	196277	99.99	8822328	26975502	000003F|quiver	Chr5
7364932	8646164	19582912	18301658	1281233	1281255	99.99	8822328	26975502	000003F|quiver	Chr5
6411554	7384953	20536290	19562883	973400	973408	99.99	8822328	26975502	000003F|quiver	Chr5
6350358	6431550	20597484	20516307	81193	81178	99.98	8822328	26975502	000003F|quiver	Chr5
5660420	6383700	21320768	20597484	723281	723285	99.98	8822328	26975502	000003F|quiver	Chr5
4714343	5655718	22258333	21316985	941376	941349	99.99	8822328	26975502	000003F|quiver	Chr5
2949713	4734327	24022974	22238383	1784615	1784592	99.99	8822328	26975502	000003F|quiver	Chr5
1	2949385	26972228	24022834	2949385	2949395	99.99	8822328	26975502	000003F|quiver	Chr5

dnadiff output

	[REF]	[QRY]
[Sequences]		
TotalSeqs	377	7
AlignedSeqs	60(15.92%)	7(100.00%)
UnalignedSeqs	317(84.08%)	0(0.00%)
[Bases]		
TotalBases	119959929	119667750
AlignedBases	117215989(97.71%)	117131679(97.88%)
UnalignedBases	2743940(2.29%)	2536071(2.12%)
[Alignments]		
1-to-1	239	239
TotalLength	118258577	118259429
AvgLength	494805.76	494809.33
AvgIdentity	99.98	99.98
M-to-M	239	239
TotalLength	118258577	118259429
AvgLength	494805.76	494809.33
AvgIdentity	99.98	99.98
[Feature	Estimates]	
Breakpoints	396	475
Relocations	6	6
Translocations	4	62
Inversions	4	2
Insertions	154	122
InsertionSum	668258	2633180
InsertionAvg	4339.34	21583.44

TandemIns	37	34
TandemInsSum	47812	63943
TandemInsAvg	1292.22	1880.68
[SNPs]		
TotalSNPs	5860	5860
AT	577(9.85%)	545(9.30%)
AC	350(5.97%)	319(5.44%)
AG	652(11.13%)	779(13.29%)
TA	545(9.30%)	577(9.85%)
TC	615(10.49%)	747(12.75%)
TG	410(7.00%)	412(7.03%)
CA	319(5.44%)	350(5.97%)
CT	747(12.75%)	615(10.49%)
CG	237(4.04%)	217(3.70%)
GA	779(13.29%)	652(11.13%)
GT	412(7.03%)	410(7.00%)
GC	217(3.70%)	237(4.04%)
TotalGSNPs	1424	1424
AT	122(8.57%)	121(8.50%)
AC	94(6.60%)	80(5.62%)
AG	135(9.48%)	222(15.59%)
TA	121(8.50%)	122(8.57%)
TC	150(10.53%)	188(13.20%)
TG	114(8.01%)	100(7.02%)
CA	80(5.62%)	94(6.60%)
CT	188(13.20%)	150(10.53%)
CG	57(4.00%)	41(2.88%)
GA	222(15.59%)	135(9.48%)
GT	100(7.02%)	114(8.01%)
GC	41(2.88%)	57(4.00%)
TotalIndels	6896	6896
H.	0(0.00%)	1(0.01%)
A.	1042(15.11%)	1568(22.74%)
T.	1041(15.10%)	1532(22.22%)
C.	419(6.08%)	389(5.64%)
G.	427(6.19%)	477(6.92%)
.H	1(0.01%)	0(0.00%)
.A	1568(22.74%)	1042(15.11%)
.T	1532(22.22%)	1041(15.10%)

.C	389(5.64%)	419(6.08%)
.G	477(6.92%)	427(6.19%)
TotalGIndels	3386	3386
H.	0(0.00%)	1(0.03%)
A.	471(13.91%)	940(27.76%)
T.	475(14.03%)	888(26.23%)
C.	154(4.55%)	131(3.87%)
G.	166(4.90%)	160(4.73%)
.H	1(0.03%)	0(0.00%)
.A	940(27.76%)	471(13.91%)
.T	888(26.23%)	475(14.03%)
.C	131(3.87%)	154(4.55%)
.G	160(4.73%)	166(4.90%)

Supplementary Table 3: BUSCO results for all assemblies

	Complete Single-Copy BUSCOs	Complete Phased BUSCOs	Fragmented BUSCOs	Missing BUSCOs	Total BUSCO groups searched
TAIR10	915	149	13	28	956
(%)	96%	16%	1%	3%	
Col-0 Assembly (FALCON)	914	153	10	32	956
(%)	96%	16%	1%	3%	
Cvi-0 Assembly (FALCON)	909	151	12	35	956
(%)	95%	16%	1%	4%	
Col-0 x Cvi-0 F1 Assembly (FALCON + FALCON-Unzip)	906	877	11	39	956
(%)	95%	92%	1%	4%	
Col-0 x Cvi-0 F1 Assembly (Platanus)	895	150	21	40	956
(%)	94%	16%	2%	4%	
Col-0 x Cvi-0 F1 Assembly (SOAPdenovo, k=93)	524	91	147	285	956
(%)	55%	10%	15%	30%	
Cabernet Sauvignon assembly (FALCON + FALCON-Unzip)	894	766	23	39	956
	94%	80%	2%	4%	
Cabernet Sauvignon assembly (SOAPdenovo, k=33)	123	8	138	659	956
	13%	1%	14%	69%	
Cabernet Sauvignon assembly (SOAPdenovo, k=43)	52	5	97	807	956
	5%	1%	10%	84%	
Clavicorona pyxidata assembly (FALCON + FALCON-Unzip)	366	277	33	30	429
(%)	85%	65%	8%	7%	
Clavicorona pyxidata assembly (Platanus)	365	29	35	29	429
(%)	85%	7%	8%	7%	
Clavicorona pyxidata assembly (SOAPdenovo, k=19)	15	2	92	322	429
(%)	3%	0%	21%	75%	

Supplementary Table 4. Comparison of structural variation calls of F1 long and short read assemblies to TAIR10

		Falcon Unzip (long reads)		Platanus (short reads)		SOAPdenovo (short reads)	
	Size range	Count	Total bp	Count	Total bp	Count	Total bp
Insertion	1-9 bp	85,223	177,966	38,280	80,577	1,165	1,405
	10-49 bp	7,319	132,311	3,293	59,094	6	131
	50-499 bp	856	143,041	352	50,044	1	107
	500-10,000 bp	239	174,575	44	30,627	0	0
	Total	93,637	627,893	41,969	220,342	1,172	1,643
Deletion	1-9 bp	111,741	209,765	39,673	84,890	1,201	1,669
	10-49 bp	7,352	132,172	3,423	58,610	32	612
	50-499 bp	860	146,067	255	41,402	6	434
	500-10,000 bp	243	173,339	62	44,010	0	0
	Total	120,196	661,343	43,413	228,912	1,239	2,715
Tandem expansion	1-9 bp	10	63	0	0	0	0
	10-49 bp	4	115	0	0	0	0
	50-499 bp	95	20,285	12	2,170	2	461
	500-10,000 bp	33	23,974	1	849	0	0
	Total	142	44,437	13	3,019	2	461
Tandem contraction	1-9 bp	3	19	1	6	0	0
	10-49 bp	13	362	4	173	0	0
	50-499 bp	72	12,972	31	5,026	0	0
	500-10,000 bp	14	9,392	5	3,703	0	0
	Total	102	22,745	41	8,908	0	0
Repeat expansion	1-9 bp	42	215	13	72	0	0
	10-49 bp	107	3,122	37	1,097	0	0
	50-499 bp	480	98,978	133	27,657	1	356
	500-10,000 bp	223	160,197	48	34,068	0	0
	Total	852	262,512	231	62,894	1	356
Repeat contraction	1-9 bp	36	180	12	74	0	0
	10-49 bp	104	2,752	35	982	0	0
	50-499 bp	500	109,907	128	27,847	0	0
	500-10,000 bp	232	167,932	57	38,981	0	0
	Total	872	280,771	232	67,884	0	0
Total variants		215,801	1,899,701	85,899	591,959	2,414	5,175
Total structural variants (>50 bp)		3,847	1,240,659	1,128	306,384	10	1,358

Counts and numbers of base-pairs affected for all variants called using Assemblytics on three Arabidopsis F1 assemblies created using Falcon Unzip, Platanus, and SOAPdenovo.

Supplementary Table 5. Haplotig SNP rate against two parental inbred lines

Haplotig	Haplotig Length	Cvi-0 SNP count	Col-0 SNP count	Total SNP count	Minority SNP Percent
000002F_048|quiver	11648134	51332	67	51399	0.13%
000001F_020|quiver	10223221	71	46926	46997	0.15%
000003F_043|quiver	8549752	40239	9	40248	0.02%
000004F_017|quiver	8151251	39237	42	39279	0.11%
000007F_009|quiver	7363020	20	31712	31732	0.06%
000006F_002|quiver	6920133	36	28565	28601	0.13%
000000F_023|quiver	4980446	17666	23	17689	0.13%
000005F_002|quiver	4634947	20290	21	20311	0.10%
000000F_067|quiver	4264247	58	20338	20396	0.28%
000003F_003|quiver	2951262	22	14508	14530	0.15%
000008F_024|quiver	2474970	4	11107	11111	0.04%
000000F_024|quiver	2269985	0	11999	11999	0.00%
000012F_008|quiver	1915216	5918	2235	8153	27.41%
000010F_015|quiver	1907498	7617	55	7672	0.72%
000009F_003|quiver	1756424	7098	18	7116	0.25%
000005F_014|quiver	1549736	0	5796	5796	0.00%
000014F_001|quiver	1441662	1	6361	6362	0.02%
000011F_019|quiver	1195815	5849	79	5928	1.33%
000011F_001|quiver	1185774	5301	60	5361	1.12%
000017F_001|quiver	1037616	37	4841	4878	0.76%
000008F_013|quiver	1023039	13	4165	4178	0.31%
000006F_022|quiver	999021	3702	12	3714	0.32%
000001F_014|quiver	967088	37	4054	4091	0.90%
000021F_001|quiver	888598	1	3462	3463	0.03%
000009F_010|quiver	857101	3690	14	3704	0.38%
000028F_001|quiver	819332	4484	0	4484	0.00%
000016F_001|quiver	815207	198	3361	3559	5.56%
000000F_048|quiver	791525	3	4396	4399	0.07%
000008F_030|quiver	597307	10	1201	1211	0.83%
000029F_001|quiver	571770	2	2152	2154	0.09%
000036F_001|quiver	556366	2082	179	2261	7.92%
000008F_028|quiver	471244	1880	26	1906	1.36%
000018F_006|quiver	468656	0	2387	2387	0.00%
000004F_013|quiver	424432	1	1388	1389	0.07%
000047F_001|quiver	389013	2101	588	2689	21.87%
000009F_005|quiver	350100	1624	1	1625	0.06%
000000F_008|quiver	328329	1676	5	1681	0.30%
000005F_021|quiver	317751	2	2099	2101	0.10%
000010F_013|quiver	267325	1314	49	1363	3.60%
000013F_003|quiver	254323	2	649	651	0.31%
000019F_001|quiver	244975	32	866	898	3.56%
000000F_029|quiver	235982	477	3	480	0.63%
000015F_004|quiver	222960	1357	1	1358	0.07%
000014F_008|quiver	215295	0	899	899	0.00%
000018F_005|quiver	149807	0	817	817	0.00%
000010F_002|quiver	142762	0	308	308	0.00%
000018F_003|quiver	132892	596	4	600	0.67%
000010F_009|quiver	124749	648	0	648	0.00%
000032F_001|quiver	109741	4	357	361	1.11%
000036F_006|quiver	98869	46	0	46	0.00%
000026F_005|quiver	96326	159	0	159	0.00%
000002F_053|quiver	81554	4	0	4	0.00%
000082F_001|quiver	74889	217	13	230	5.65%
000051F_001|quiver	69253	2102	0	2102	0.00%

000009F_009|quiver	62928	309	1	310	0.32%
000008F_022|quiver	62415	0	82	82	0.00%
000107F_001|quiver	60667	17	2	19	10.53%
000026F_006|quiver	56116	1	0	1	0.00%
000000F_053|quiver	55481	17	434	451	3.77%
000143F_001|quiver	47951	0	25	25	0.00%
000040F_001|quiver	43707	222	1	223	0.45%
000053F_001|quiver	43417	55	0	55	0.00%
000050F_001|quiver	42565	0	186	186	0.00%
000097F_001|quiver	42449	0	386	386	0.00%
000009F_018|quiver	41320	4	153	157	2.55%
000026F_002|quiver	41157	1	0	1	0.00%
000009F_017|quiver	40786	90	0	90	0.00%
000011F_018|quiver	40469	1	0	1	0.00%
000026F_001|quiver	38716	23	0	23	0.00%
000015F_005|quiver	38585	0	1	1	0.00%
000013F_004|quiver	37531	4	111	115	3.48%
000000F_001|quiver	37300	285	19	304	6.25%
000044F_001|quiver	33412	155	7	162	4.32%
000010F_007|quiver	33020	146	19	165	11.52%
000012F_004|quiver	32961	81	0	81	0.00%
000015F_001|quiver	32675	1	170	171	0.58%
000003F_051|quiver	31417	55	0	55	0.00%
000056F_001|quiver	30771	807	0	807	0.00%
000011F_008|quiver	30548	194	0	194	0.00%
000010F_012|quiver	29282	197	1	198	0.51%
000018F_004|quiver	28808	280	2	282	0.71%
000005F_013|quiver	26271	37	19	56	33.93%
000066F_001|quiver	25194	23	202	225	10.22%
000189F_001|quiver	22194	17	2	19	10.53%
000143F_002|quiver	22091	0	34	34	0.00%
000109F_001|quiver	21623	155	105	260	40.38%
000031F_002|quiver	21378	0	1	1	0.00%
000046F_001|quiver	20553	0	2	2	0.00%
000033F_002|quiver	20195	1	157	158	0.63%
000017F_004|quiver	18591	0	117	117	0.00%
000001F_018|quiver	18157	12	0	12	0.00%
000058F_002|quiver	16353	64	18	82	21.95%
000002F_059|quiver	16172	53	0	53	0.00%

Supplementary Table 6. Haplotig concordance against two parental inbred lines

Haplotig	Haplotig Length	SNP vs Cvi-0	SNP vs Col-0	DEL vs Cvi-0	DEL vs Col-0	INS vs Cvi-0	INS vs Col-0	estimated SNP concordance(phred scale)	estimated indel concordance(phred scale)
000002F_048|quiver	11648134	51332	67	25268	1284	25707	523	52	38
000001F_020|quiver	10223221	71	46926	1004	23012	602	22127	52	38
000003F_043|quiver	8549752	40239	9	18984	782	19860	250	59	39
000004F_017|quiver	8151251	39237	42	18298	687	18002	364	53	39
000007F_009|quiver	7363020	20	31712	641	16191	438	15110	55	38
000006F_002|quiver	6920133	36	28565	432	14184	310	13666	53	40
000000F_023|quiver	4980446	17666	23	9265	370	9679	145	53	40
000005F_002|quiver	4634947	20290	21	9359	452	10130	196	53	39
000000F_067|quiver	4264247	58	20338	672	8340	379	7943	49	36
000003F_003|quiver	2951262	22	14508	312	6542	184	5697	51	38
000008F_024|quiver	2474970	4	11107	197	4503	93	4741	57	39
000000F_024|quiver	2269985	0	11999	127	5673	107	5483	64	40
000012F_008|quiver	1915216	5918	2235	2438	1283	2518	803	29	30
000010F_015|quiver	1907498	7617	55	3102	283	2675	91	45	37
000009F_003|quiver	1756424	7098	18	3049	269	3084	41	50	38
000005F_014|quiver	1549736	0	5796	56	3054	65	2716	62	41
000014F_001|quiver	1441662	1	6361	80	2768	75	3135	59	40
000011F_019|quiver	1195815	5849	79	1857	308	1592	62	42	35
000011F_001|quiver	1185774	5301	60	2219	332	2489	121	43	34
000017F_001|quiver	1037616	37	4841	140	1233	51	1211	44	37
000008F_013|quiver	1023039	13	4165	304	1207	72	1251	49	34
000006F_022|quiver	999021	3702	12	1947	115	2204	50	49	38
000001F_014|quiver	967088	37	4054	383	1711	67	1531	44	33
000021F_001|quiver	888598	1	3462	85	1126	23	929	56	39
000009F_010|quiver	857101	3690	14	1564	154	1755	26	48	37
000028F_001|quiver	819332	4484	0	2035	42	2060	17	59	41
000016F_001|quiver	815207	198	3361	239	1418	111	1387	36	34
000000F_048|quiver	791525	3	4396	44	1934	50	1945	53	39
000008F_030|quiver	597307	10	1201	127	577	39	515	47	36
000029F_001|quiver	571770	2	2152	63	966	41	771	53	37
000036F_001|quiver	556366	2082	179	496	282	511	237	35	30
000008F_028|quiver	471244	1880	26	904	227	835	59	42	32
000018F_006|quiver	468656	0	2387	21	1004	15	1234	57	41
000004F_013|quiver	424432	1	1388	68	770	23	695	53	37
000047F_001|quiver	389013	2101	588	478	72	431	174	28	32
000009F_005|quiver	350100	1624	1	599	70	418	8	52	36
000000F_008|quiver	328329	1676	5	739	77	522	14	47	36
000005F_021|quiver	317751	2	2099	71	676	34	735	50	35
000010F_013|quiver	267325	1314	49	423	55	640	11	37	36
000013F_003|quiver	254323	2	649	53	208	16	215	49	36
000019F_001|quiver	244975	32	866	85	213	17	374	39	34
000000F_029|quiver	235982	477	3	185	88	206	10	48	34
000015F_004|quiver	222960	1357	1	347	46	246	13	50	36
000014F_008|quiver	215295	0	899	12	399	7	172	53	40
000018F_005|quiver	149807	0	817	13	327	5	323	52	39
000010F_002|quiver	142762	0	308	30	83	11	46	52	35
000018F_003|quiver	132892	596	4	223	47	243	17	44	33
000010F_009|quiver	124749	648	0	143	15	170	0	51	39
000032F_001|quiver	109741	4	357	58	68	4	164	43	32
000036F_006|quiver	98869	46	0	75	0	18	0	50	50
000026F_005|quiver	96326	159	0	54	0	86	0	50	50
000002F_053|quiver	81554	4	0	37	0	6	0	49	49
000082F_001|quiver	74889	217	13	62	14	52	19	37	33
000051F_001|quiver	69253	2102	0	290	2	69	0	48	44
000009F_009|quiver	62928	309	1	212	66	191	5	45	29
000008F_022|quiver	62415	0	82	0	65	0	77	48	48
000107F_001|quiver	60667	17	2	19	19	6	3	43	34
000026F_006|quiver	56116	1	0	69	0	34	0	47	47
000000F_053|quiver	55481	17	434	26	119	14	144	35	31

000143F_001|quiver	47951	0	25	0	98	0	164	47	47
000007F_004|quiver	47307	0	0	0	1	0	0	47	47
000040F_001|quiver	43707	222	1	97	58	83	11	43	28
000053F_001|quiver	43417	55	0	198	0	53	0	46	46
000050F_001|quiver	42565	0	186	0	97	0	63	46	46
000097F_001|quiver	42449	0	386	0	137	0	78	46	46
000014F_006|quiver	41387	0	0	0	57	0	14	46	46
000009F_018|quiver	41320	4	153	34	54	18	35	39	29
000026F_002|quiver	41157	1	0	33	0	26	0	46	46
000009F_017|quiver	40786	90	0	14	11	26	1	46	35
000011F_018|quiver	40469	1	0	1	16	0	2	46	43
000026F_001|quiver	38716	23	0	33	0	43	0	46	46
000015F_005|quiver	38585	0	1	0	48	0	9	46	46
000013F_004|quiver	37531	4	111	16	33	8	8	39	32
000006F_030|quiver	37337	0	0	0	2	0	1	46	46
000000F_001|quiver	37300	285	19	138	58	141	16	33	27
000014F_004|quiver	35424	0	0	0	8	0	3	45	45
000044F_001|quiver	33412	155	7	49	37	35	11	36	28
000010F_007|quiver	33020	146	19	106	73	12	4	32	26
000012F_004|quiver	32961	81	0	103	0	33	0	45	45
000015F_001|quiver	32675	1	170	31	28	7	74	42	30
000003F_051|quiver	31417	55	0	34	0	6	0	45	45
000056F_001|quiver	30771	807	0	236	0	64	0	45	45
000011F_008|quiver	30548	194	0	53	0	75	0	45	45
000010F_012|quiver	29282	197	1	98	32	63	5	42	29
000018F_004|quiver	28808	280	2	83	1	28	4	40	37
000002F_015|quiver	28220	0	0	8	0	0	0	45	45
000005F_013|quiver	26271	37	19	69	36	14	28	31	27
000066F_001|quiver	25194	23	202	100	86	31	26	30	23
000189F_001|quiver	22194	17	2	15	5	1	3	39	35
000002F_030|quiver	22178	0	0	21	0	9	0	43	43
000143F_002|quiver	22091	0	34	0	149	0	0	43	43
000109F_001|quiver	21623	155	105	109	75	24	21	23	23
000031F_002|quiver	21378	0	1	0	16	0	1	43	43
000017F_006|quiver	20664	0	0	0	21	0	7	43	43
000046F_001|quiver	20553	0	2	0	25	0	5	43	43
000033F_002|quiver	20195	1	157	76	123	6	109	40	24
000039F_001|quiver	19189	0	0	23	0	4	0	43	43
000017F_004|quiver	18591	0	117	0	198	0	16	43	43
000001F_018|quiver	18157	12	0	26	0	6	0	43	43
000058F_002|quiver	16353	64	18	97	32	6	2	29	27
000002F_059|quiver	16172	53	0	9	0	9	0	42	42
000002F_045|quiver	14026	0	0	1	0	1	0	41	41
000028F_004|quiver	10031	0	0	10	0	1	0	40	40

Supplementary Table 7. Arabidopsis AUGUSTUS CDS prediction results

Assemblies		TAIR 10	Col-0	Cvi-0
	Number of predicted CDS	27,954	28,176	27,797
		100% indel-free full length overlaps		
Col-0	28,176	26,897		
		96.2%		
Cvi-0	27,797	18,422	18,473	
		65.9%	65.6%	
Col-0 x Cvi-0	56,806	26,812	26,931	26,582
		95.9%	95.6%	95.6%

Supplementary Table 8. V. vinifera CDS alignment summary

	CDS aligned to p_ctgs	CDS aligned to h_ctgs	CDS aligned to both p_ctg and h_ctg	CDS aligned to both p_ctg or h_ctg	Total CDS predicted in PN40024
Full and partial alignemnts	27288	22447	20841	28894	29971
	91%	75%	70%	96%	100%
Full alignemnts with clipping	24251	19436	16981	26706	29971
	81%	65%	57%	89%	100%

Supplementary Table 9. FALCON-Unzip phase concordance with short read dataset

Variant Filtering Criterion	**total pairs**	**phase discordant pairs**	**phase discordant %**	**sum of lengths between pairs**
QUAL > 10, All	136348	782	0.574%	9726672
QUAL > 500, SNP only	109149	425	0.389%	9110210
QUAL >1000, SNP only	92780	270	0.291%	8062835

QUAL = Freebayes variant call "Quality value"

Variant Filtering Criterion	QUAL > 10, All	QUAL > 500, SNP only	QUAL >1000, SNP only
Total number of block	8896	8662	8590
Total phased length	9647518	9056219	8018724
phasing BLOCK N50 (bp)	2339	2155	1831
Total variant phased	145112	117732	101242
phased variant count N50	53	37	28
Number of 100% concordant phased blocks	8582	8458	8429
Total concordant phased length	8945335	8593313	7677149
N50 of 100% concordant phased blocks	2174	2078	1764
Number of variant in 100% concordant phased blocks	132279	111177	97359
N50 of 100% concordant phased variants	48	36	27

Supplemental Table 10: Selected Heterozygous Genomes

Latin Name	**Genome Size (Mbase)**	**Heterozygosity (%)**	**Citation and URL**
Ananas comosus (Pineapple)	526	1.89%	Ming, R, et al. "The pineapple genome and the evolution of CAM photosynthesis" Nature Genetics 47, 1435–1442 (2015) doi:10.1038/ng.3435 http://www.nature.com/ng/journal/v47/n12/full/ng.3435.html
Boa constrictor constrictor (Boa Constricter)	1,600	0.36%-0.42%	Bradnam, Keith R., et al. "Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species." GigaScience 2.1 (2013): 1. https://gigascience.biomedcentral.com/articles/10.1186/2047-217X-2-10
Ciona intestinalis (Sea Squirt)	160	1.20%	Dehal, Paramvir, et al. "The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins." Science 298.5601 (2002): 2157-2167. http://science.sciencemag.org/content/298/5601/2157
Crassostrea gigas (Oyster)	637	1.30%	Zhang, Guofan, et al. "The oyster genome reveals stress adaptation and complexity of shell formation." Nature 490.7418 (2012): 49-54. http://www.nature.com/nature/journal/v490/n7418/full/nature11413.htm
Lates calcarifer (Asia Seabass)	700	0.50%	Vij, Shubha, et al. "Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-layered Scaffolding." PLoS Genet 12.4 (2016): e1005954. http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005954
Melopsittacus undulatus (Budgerigar)	1,200	0.1%-0.8%	Bradnam, Keith R., et al. "Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species." GigaScience 2.1 (2013): 1.

			https://gigascience.biomedcentral.com/articles/10.1186/2047-217X-2-10
Olea europea sbsp. *Europea* (Olive Tree)	1,320	0.50%	Cruz, Fernando, et al. "Genome sequence of the olive tree, Olea europaea."GigaScience 5.1 (2016): 1. https://gigascience.biomedcentral.com/articles/10.1186/s13742-016-0134-5
Phoenix dactylifera (Date Palm)	380	0.46%	Al-Dous, EK, et al. "De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera)" Nature Biotechnology 29, 521–527 (2011) doi:10.1038/nbt.1860 http://www.nature.com/nbt/journal/v29/n6/full/nbt.1860.html
Picea abies (Norway Spruce)	20000	0.82%	Nystdt, B, et al. "The Norway spruce genome sequence and conifer genome evolution" Nature 497, 579–584 (30 May 2013) doi:10.1038/nature12211 http://www.nature.com/nature/journal/v497/n7451/full/nature12211.html
Populus tremula (Poplar Tree)	500	1.30%	Wang, Jing, et al. "Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related Populus species." Genetics 202.3 (2016): 1185-1200. http://www.genetics.org/content/202/3/1185.abstract
Pyrus bretschneideri Rehd. (Pear)	527	1.02%	Wu, J et al. "The genome of the pear (Pyrus bretschneideri Rehd.)" Genome Research. 2013. 23: 396-408 http://genome.cshlp.org/content/23/2/396.full
Solanum tuberosum L (Potato)	844	2.50%	The Potato Genome Sequencing Consortium. "Genome sequence and analysis of the tuber crop potato" Nature 475, 189–195 (14 July 2011) doi:10.1038/nature10158 http://www.nature.com/nature/journal/v475/n7355/full/nature10158.html

Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing

Supplementary Note

Additional details for *Arabidopsis* sequencing and assembly

We individually sequenced the inbred Col-0 and Cvi-0 genomes using 49 and 60 SMRT Cells with P4-C2 sequencing chemistry, generating 15.2 Gbp (~130x coverage) and 14.7 Gbp (~120x coverage) of raw sequence data, respectively. The raw data produced was commensurate with P4-C2 performance (Kim et al. 2014), with average insert read length of 6.5 kbp and 6.1 kbp, and maximum read-lengths of 44,472 bp (**Supplementary Table 1**). The two inbreds were assembled independently using FALCON and both generated ~120 Mb of assembled sequence in 377 (Col-0) and 260 (Cvi-0) contigs. (**Table 1**). The contig N50 sizes of the assemblies were 7.4 Mb (Col-0) and 6.0 Mb (Cvi-0) (**Table 1**). We used the whole genome alignment algorithm nucmer and the associated dnadif tool (MUMmer3 package(Kurtz et al. 2004)) to evaluate the accuracy by comparing our Col-0 assembly to the TAIR10 assembly. The nucleotide sequence accuracy was greater than 99.98%.

For the Col-0 x Cvi-1 F1 cross sample, we collected 18.5 Gbp sequence data (~ 140X haploid size from 29 SMRTcells, P6-C4 chemistry) and 60x short-read (paired-end 250bp) dataset. For all three samples, we collect high coverage long read datasets in order to build the best assemblies and provide the datasets for future down-sampling experiments to understand the tradeoff between assembly contiguity or accuracy and sequencing coverage. We performed a 60x down-sampling assembly for the F1 strain, FALCON results an assembly of contig N50=2.28Mb which is still much more contiguous than the assembly form similar coverage of the short read data.

Arabidopsis assembly gene space completeness evaluation

Since Arabidopsis genome is A/T (~64%) rich, when examining the sequence alignments around the discordant sites in detail, we found that most indel discordant sites were within long homopolymers (**Supplementary Data 1**: col-0_TAIR10_1_l10000.snp_with_context.txt) or in simple tandem duplications. There were 68,036 A or T homopolymer blocks at least 10bp long from in our diploid F1 assembly (haplotigs and primary contigs), and 34,483 such blocks in the haploid TAIR10 assembly up to 48bp long (**Supplementary Fig. 6**). While the SMRT Sequencing can processively read through very long, e.g. greater than 20bp, 100% A/T homopolymers, the exact length of the longest homopolymer regions may not be correct. We caution, however, that even Sanger sequencing of very long homopolymers is unreliable, and some of the discordance could arise from errors in the reference or reflect true polymorphisms existing between the samples.

To assess the quality of the assembly of the gene space, and the impact of the residual homopolymer indels on gene prediction, we first applied the gene prediction tool AUGUSTUS (Stanke and Waack 2003) on TAIR10, our FALCON Cvi-0 and Col-0 assemblies, and our Cvi-0 x Col-0 F1 FALCON/FALCON-Unzip assembly and compared the predicted coding sequences (CDS) (**Supplementary Table 7**). We focus on the fully aligned CDS predictions without indels as mismatches are less likely to cause frameshift errors. Complete predicted CDS from the inbred and F1 assemblies aligned completely to 95 - 96% of the 27,954 CDS of TAIR10 without indels. To avoid potential bias due to *ab initio* prediction, we performed an additional analysis by directly aligning the manually curated TAIR10 CDS to the entire assembly using the STAR aligner (Dobin et al. 2013). Overall, 96% of the 35,386 TAIR10 CDS were fully aligned without indels or truncations to the F1 assembly contigs. Another 1.04% CDS were successfully aligned allowing for only one base insertion or deletion: 0.18% CDS had a 1 base insertion and 0.86% CDS had a 1 base deletion.

Analysis of the variations between haplotypes in *Arabidopsis* Col-0 x Cvi-0 F1 genome

With both haplotype sequences robustly assembled, we analyzed the differences between the homologous chromosomes using nucmer and Assemblytics (**Table 2**). When we compared the haplotigs to the primary contigs in the diploid F1 assembly, we identified 450,680 SNPs, compared to 501,243 found by aligning the Col-0 and Cvi-0 inbred assemblies. As expected, many of the het-SNP pairs between the comparisons were indeed identical SNPs. We identify the context (20bp + 20bp each side around the SNPs) between the het-SNP pairs between (1) the haplotigs to the primary contigs and (2) the Col-0 and Cvi-0 contigs. Out of 449,789 unique contexts of the het-SNPs between the haplotigs to the primary contigs, 422,588 (94.0 %) of them are identical to those pairs between Col-0 and Cvi-0 contigs.

Using Assemblytics, we identified 966 SV events (>50bp indels or tandem repeat contractions and expansions) between the haplotigs and primary contigs, compared to 1,051 between the Cvi-0 and Col-0 assemblies. Thus, FALCON-Unzip phased 85.7% of all SNPs and 91.9% of all SVs directly from the shotgun sequence assembly. Interestingly, 5039 of the 31,679 Augustus predicted coding regions (full transcript predictions) intersected structural variants at least 50bp in size identified on the primary contigs, which may have important effect on gene expression regulation and/or functionality.

We estimated the amount of variation affecting coding sequences by comparing the predicted CDS (**Table 2**) between the Col-0 and Cvi-0 inbred lines. We found about 184,000 (0.45%) SNPs within the 40.7Mbp predicted CDS of the inbred assemblies, compared to 148,000 (0.41%) SNPs within the 36 Mbp CDS between the haplotigs and the primary contigs in the FALCON-Unzip assembly of the F1 (**Table 2**). The number of heterozygous SNPs and SVs present in the F1 assembly is marginally lower than those from the comparison between Col-0 and Cvi-0, mostly because the collection of the haplotigs does not fully represent the full haploid chromosome set. In particular, the number of variants between the haplotigs and the primary contigs is consistent with the total haplotig size (105 Mbp) that is about 87% of the estimated genome size.

Comparison of the long-read and short-read assemblies of the *Arabidopsis* Col-0 x Cvi-0 F1 genome

By comparing the short and long read assemblies to the TAIR10 genome, we can assess the quality differences between the assemblies and the ability to identify variants. We used Assemblytics to call insertion and deletion variants from the three *Arabidopsis* F1 assemblies with FALCON-Unzip, Platanus, and SOAPdenovo to the TAIR10 reference. We aligned the contig sequences for the short-read assemblies since aligning the scaffolds may introduce artificial variants due sequence gaps marked with Ns. Assemblytics identified a total of 215,801 variants from the FALCON-Unzip assembly, of which 3,847 were structural variants larger than 50bp (**Supplementary Fig. 4**, **Supplementary Table 4**). In contrast, Assemblytics detected 85,899 variants (1,128 sites >50bp) in the Platanus assembly and only 2,414 variants (10 sites >50bp) from the SOAPdenovo assembly. The variants from the FALCON-Unzip assembly captured 89% of the Platanus variants and 90% of the SOAP variants at a stringent requirement of the exact same variant type, size, and genomic location. However, the Platanus and SOAP assemblies captured only 37% and 1% of the FALCON-Unzip variants, respectively. The contiguity of the assemblies greatly affects the numbers and sizes of variants that can be called, and since differing haplotypes can result in mis-assemblies, constructing proper haplotypes can be an important factor in accurate variant-calling from an assembly.

BUSCO plant ortholog set selection

The BUSCO plant set is comprised of 956 putative single copy orthologs totaling 587,335 amino acid residues from a set of 30 embryophyte lineages as accessioned in OrthoDB V9.

Clavicorona pyxidata gene space evaluation with BUSCO

In lieu of a reference, we evaluated the assemblies using BUSCO and genomic sequencing data (SRA accession: SRR1800147, 86X, 150 bp reads) (**Supplementary Table 3**). The Platanus assembly contains approximately the same number of single copy eukaryote BUSCO sets as those in the FALCON-Unzip assembly (**Supplementary Table 3**). However, the Platanus assembly missed nearly all of the homologous copies in the diploid genome (29/429 eukaryotic BUSCO proteins are duplicated compared to 277/429 in the FALCON-Unzip assembly) and the regulatory context around the genes are much more limited. The best SOAPdenovo assembly (with *k*-mer size 19) yields less then 3% single complete genes copies and had a shorter N50 than the Platanus assembly.

The effect of different level of heterozygosity on primary contig and haplotig layout

By design, primary contigs are only locally phased within the homologous region of each haplotig but some regions in the primary contigs have no corresponding haplotig due to low

heterozygosity. FALCON-Unzip generates the primary contig to maintain the continuity through those regions, but it does not maintain the haplotype phases across them. The type and the amount heterozygosity between homologous chromosomes will affect the segregation of haplotigs from the primary contigs.

Fundamentally, to separate the haplotypes reliably, two conditions are necessary: (1) enough read coverage to call het-SNPs with random sequence error (Myers Jr 2016) and (2) enough number of het-SNPs within the read lengths. For (1), we only call het-SNPs on positions with coverage larger than 10. This minimum coverage requirement allows reducing false positives of the het-SNP calls even with the sequence or alignment errors. For (2), if the neighboring het-SNPs are too distant, we might not have reads connecting them we will not be able build haplotype relationship between those SNPs. And if a read covers less than two het-SNPs, we will not be able to assign the haplotype reliably. In such case, we will have regions that are un-phased and the regions will have no corresponding alternative haplotigs from the primary contigs. From this study and earlier studies, many genomes have heterozygosity level of het-SNP rate from 0.5% to more than 1.0% **(Supplementary Table 10).** If the overall het-SNP rate is roughly 0.5% to 1%, we should get at least 25 to 50 het-SNP sites for reads longer than 5kb. However, the het-SNP may not be evenly distributed through the genomes and it affects the performance of the haplotype segregation.

Supplementary Fig. 8 shows the distribution of observed number of het-SNP per reads for the longest contig of each genome. If we only focus on reads of length between 9kb to 11k to avoid the confounding factor from variable read lengths, we find the number of het-SNPs per read with the specific read length range can be fitted with exponential distributions. This implies that certain regions will have many fewer het-SNPs than the genome-wide average.

With 20 to 50 het-SNPs in 10kb region, we will have enough information to reliably assign the haplotype of the reads. However, the variations between the haplotypes are subjected to evolution constraints and such variations cannot be characterized by a single average het-SNP rate. Even with high average het-SNP rate, there could be some regions with lower het-SNP rates, perhaps even zero het-SNPs, that we cannot make unambiguous haplotype assignment. For example, if we need three or more het-SNPs for reliably haplotyping the reads of 10kb, we will need the local het-SNP rate at least higher than 1 in 3kb. With a random mutation assumption, the requirement of the het-SNP rate for haplotyping can be modeled mathematically. How the performance for haplotype segregation depends on local het-SNP rates across the tree of life is certainly an important subject for future research.

In extreme cases, if there are many structural variations between the haplotypes or the sequence identity is lower than the pre-specified overlapping identity threshold, we might not even find proper overlaps between the reads from different haplotypes. When this happens, the paths corresponding to different haplotypes will not even cross each other and FALCON will layout them as separated primary contigs.

Our assemblies of the three organisms we assembled in this work reflect how such different levels of heterozygosity affect the results at qualitative level. In term of heterozygosity observed from the data, *C. pyxidata* has the lowest heterozygosity genome-wide, as about 50% of the (haploid) genome has low het-SNP density that cannot be phased (see **Supplementary Fig. 8** for an example of a low heterozygosity region identified with orthogonal data).

Next, *Arabidopsis* Col-0 x Cvi-0 F1 genome has a medium rate of heterozygosity, with an overall SNP rate of ~1/200bp and ~1000 structural variations at least 50bp in size. The total haplotig size (105Mbp) is less than the haploid genome size, indicating small portion of the genome has low heterozygosity. Furthermore, the p-contig size is 140Mb that is modestly larger then haploid genome size. Using Satsuma (Grabherr et al. 2010), we find about 18.9 Mbp of separate primary contigs were actually syntenic to each other and mapped to the same region along the TAIR10 reference. This can be explained by the variations between the parental haplotypes being so extensive such that the sub-graphs of the two haplotypes are no longer associated. Thus, the total primary assembly size may be modestly larger than the haploid genome size and the total haplotig size might be smaller than the haploid genome size.

Finally, in V. vinifera the p-contig assembly size (590 Mb) is about 18% bigger and the haplotig size (368 Mb) is 26% smaller than the estimated genome size (~500 Mb). We use Satsuma to find syntenic p-contigs. We identified 25 Mbp of syntenic blocks between the p-contigs. Given the two parent strains are actually different species, it maybe possible there are some other syntenic regions that are not identified by our method as the divergence between them are high beyond the sensitivity of the tool we use.

With the default FALCON parameter setting, it requires at least 96% identity between error-corrected reads for overlapping. If the sequence identity between homologous chromosomes drops below that or there are many larger structure variations between the reads, the reads will not have any reported overlaps, and the assembler will construct separate p-contigs. While it is possible to turn the parameters to be more permissive to catch weaker homologies between haplotypes, the parameters would depend on the underlying biological sequence structure of the chromosome as well as the sequencing characteristics. We summarize how the different level heterozygosity may affect the contig layout in **Supplementary Fig. 10**.With the observation of different degrees of heterozygosity from 3 different genomes in this study, we would need to survey a broader range of organisms to understand the nature of the diploid and polyploidy heterozygosity across the different kingdoms of life.

Updated FALCON consensus algorithm

FALCON's consensus module (FALCON-sense) takes a set of seeding sequences and their supporting reads to generate consensus for each of the seeding sequences. Here is the updated consensus process:

1. For each seed sequence, the supporting reads are aligned to the seed sequences with a variation of Gene Myers' O(ND) aligner (Myers 1986). We don't allow for mis-matches in

the alignments. All differences are encoded as insertions and deletions.

2. For each position in an alignment between a supporting read and its seed read, we generate an "alignment tag" specific to the position and the supporting read. The tag consists of 4 fields: (1) the anchoring position in the seed sequence, (2) the "delta" from the anchoring position on the seed sequence, (3) the base of the seed sequence and (4) the base of the supporting read in the alignment. When there are insertions in the supporting reads at a given position, the inserted base in the alignment has a non-zero "delta" value indicating the number of base from the anchoring bases. Note that field (3) and (1) contain the same information encoded differently. We keep field (3) just for convenience.

3. We create the alignment graph from all alignment tags of all supporting reads. The alignment graph is created as following:
    1. For each distinct tag, we create a corresponding node in the graph.
    2. For consecutive tags from a supporting read, we create an edge in graph and assign a variable "`edge_count`" associated with the edge to be 1. If the edge already exists, we simply increase the "`edge_count`" by 1. The result is a directed acyclic graph where each node has a tag, and the "`edge_count`" indicates how many reads supporting the connection between the two nodes. A higher number for "`edge_count`" of an edge indicates higher confidence about the "connection" between the nodes from the supporting reads. A high quality consensus sequence is simply a chain of nodes connected by those high confidence edges.

4. A standard dynamic programming algorithm is applied on the alignment graph to find the highest confidence path. We generate the consensus sequence by concatenating the bases of the tags through the best path.

Supplementary Fig. 12 summarizes the process. In the highlighted box of the path, we see one insertion error corrected and two missing bases recovered.

Two paths are highlighted in the graph with dashed curves. The cyan dashed curve indicates the path of the seed sequence. The green dashed curve indicates the final consensus path. Node "`0092:0:T:T`" is identified as an erroneous extra base in the seed read. The new nodes, "`0091:0:-:C`" and "`0091:0:-:A`" are added to the consensus path. The green arrows highlight the path of the original template sequence. The "CA" insertion to the seed sequence concurs with the original template sequence. Namely, the consensus indeed corrects 3 simulated errors in the seed read for this case.

Assembly string graph reduction process summary

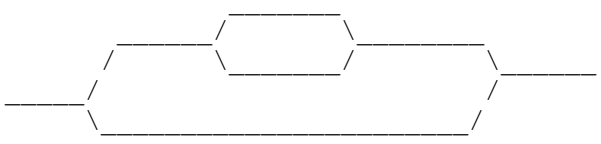
Constructing the initial string graph

After the reads are error-corrected, “daligner” was used to obtain the overlaps between the reads. An overlap-filtering step is applied to remove contained reads or reads that appear to be from high copy repetitive regions. We adapt the method described in Myers’ paper (Myers 2005) to build the string graph from the filtered overlap pairs of reads. There are some differences in our implementation of the string graph from the original paper. Instead of using the bi-directed edge graph described in the paper, we construct an ordinary directed graph from the overlaps. Each pair of overlap actually creates two directed edges. However, we only trace the graph along one direction to represent both contigs and their reverse complement counterpart. Namely, for each pair of edges, only one of the two edges will be in the final contigs. We explicitly track such duality of the edge-pairs to remove redundant reverse compliment edges during the layout stage.

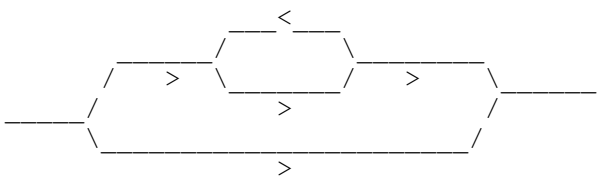
A heuristic algorithm for constructing compound paths

Here we describe some heuristic algorithms used to identify “bubbles” in the initial assembly string graph. Conceptually, a "bubble" refers to a sub-graph that (1) has only one source node and one sink node, (2) has multiple paths from the source to the sink and (3) remains weakly connected after any single edge removal. Complicated heterozygous structural variation or repetitive regions between two haplotypes in a diploid dataset may result in more complicated structures:

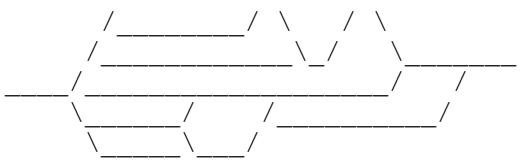
(1) For example, it is possible to have nested bubbles:



(2) Or, there can be loops in such graph



(3) We also encounter some very tangled nested bubbles, which might indicate some local complicated repeats that are hard to resolve:



We adopt a pragmatic approach to solve the problem of finding such local structures. A sketch of the heuristic used by Falcon to find the “compound path” of “bubbles” is presented here:

1) Simplify the initial assembly string graph to a graph with simple paths. Namely, all edges in a path with out any branching node become single edge. We call this graph UG_0.
2) For each node with multiple out-edges in UG_0, we start a search to find a local "bundle" of edges. We set up "active tracers" to trace down each branch from the source node.
   a) In each iteration step, we check each single node that has an active tracer. If every in-node of a node already has a tracer, we add active tracers to its offspring nodes and the tracer of the parent node becomes inactive. If there is only one active tracer left, the subgraph of all traced nodes and the edges between them is considered a "compound path".
   b) We detect a loop by checking if any of the offspring of a node with an active tracer already have an occupied tracer. When a loop is detected, the search stops and we don’t generate a compound path.
   c) In some complicated repetitive parts of the genome, the number of active tracers can increase quickly. We limit the number of active tracers in order to avoid excessive computation unnecessarily. If the number of active tracers is beyond a pre-specified limit (3 tracers in the current implementation), we stop the search.
   d) For each step, we can calculate the number of nodes and the length of the path as longest number of sequence bases from the source to all nodes with an active tracer. If those numbers of nodes or the longest path length is bigger than pre-specified numbers, the search stops.
3) Some of the "compound paths" are overlapped with others or one contains the other, e.g., the smaller bubble is part of the big bubble in the case of nested bubbles. In such case we choose the big one and ignore the small one.

Each "compound path" is a set of the simple paths in UG_0. The simple paths in UG_0 are then replaced by the newly identified “compound path” resulting in a unitig graph where each edge is either a simple path or a compound path. The next step is to identify the un-branched paths in the unitig graph and generate the contigs by concatenating the sequences along the edges as shown in **Supplementary Fig. 14**.

From sequence overlaps to primary and associated contigs

Supplementary Fig. 14 summarizes the whole process from the corrected reads to primary and associated contigs through multiple stages of graph simplification. Basically, we decompose the string graph into compound and simple paths. A primary contig can be constructed by finding

the path of a non-branching collection of compound and simple paths. For each compound path, we can generate the associated contigs that typically have a large number of variations relative to the primary contigs. The primary contigs are still haplotype-fused contigs. Namely, different parts of the primary contigs might come from different haplotypes. The layout rules provide no guarantee to keep the haplotype consistent through the primary contig at this stage. Further, when there is little variation between two haplotypes, there will be no associated contigs generated. We can further use the "FALCON-Unzip" code to separate the haplotype sequences and generate haplotype coherent contigs (haplotigs).

A greedy SNPs phasing algorithm

To reconstruct the haplotype sequences, we need to separate the reads that are homologous to each other but belong to different haplotypes into different groups. In FALCON-Unzip, this is done with the following steps:

1) Track PacBio reads to each primary contig and generate alignments between the reads and the contigs. There are two strategies: (1) examining the overlapping data generated during the assembly process to identify reads that belong to each specific contig and re-aligning reads only to each specific contig or (2) simply aligning the reads back to all assembled contigs. For (1), initially, we examine the tiling path of each of the contigs. The reads corresponding to the tiling path are assigned to the contig. After that we examine the overlapping data generated from the assembly process. Each read may overlap with other contig-assigned reads or other un-assigned reads in the original overlap data. The overlapped contig-assigned reads may not be from the same contig. We need to score the overlaps and decide the best contig assignment of the query. We score each pair overlap by the overlap length. If the best pair overlap read already has been assigned to a contig, the unassigned query read will be assigned to the same contig. The reads are grouped into sets that are associated with each primary contig. Each contig and its associated reads are aligned independently. In contrast to strategy (2), where the alignments are done globally, method (1) is much more computationally efficient as it re-uses overlap data that already exists.
2) Call het-SNPs for each contig. As most haplotype reconstruction algorithms (Halldórsson et al. 2004; Bansal and Bafna 2008), we need a set of het-SNP calls as input. We use a simple alignment and counting method for calling het-SNPs. We align the reads to the draft contigs with BLASR (Chaisson and Tesler 2012) and generate SAM/BAM alignment files. We only focus on SNP calls and we also ignore SNP calls that have insertions or deletions nearby. Namely, we only count variations within the "M" block of the CIGAR string in the BAM files (https://samtools.github.io/hts-specs/SAMv1.pdf). For each base in the contig, we count the number each "A", "C", "G", and "T" bases from the aligned reads. If the highest count is less than 75% and the second count is higher than 25%, we call a het-SNP site and the top two bases are used as the two allele bases for further SNP phasing downstream.
3) Phase het-SNPs for each contig. We use a simple greedy algorithm to phase the SNPs into different haplotype groups. The general method is shown in the **Supplementary Fig. 15**. For

each het-SNP site, the initial phase of the variants are randomly chosen. For any two sites that have shared reads mapped on, we can calculate a "coupling score". The "coupling score" is defined simply as the number of reads supporting the particular phase. Along the reference, we scan the het-SNP sites from the 5'-end to 3'-end. We test the two possible choices of the phase at the given site for each het-SNP site. The choice with more "coupling score" support is kept. When there are fewer than 6 reads connecting two het-SNP sites, we break the haplotype blocks. We apply this scanning and updating process iteratively until no more optimization of the score is possible or reach a pre-specified limit of iterations. At the end of this process, the het-SNP sites along the reference will be grouped into different haplotype blocks and we have a set of phased SNPs associated with each block.

4) Assign phases to the raw reads. For each read, we examine the alignment to the reference. If the read contain het-SNP calls against the reference, we count which phases it agrees the most and assign a tuple of "(block identifier, phase identifier)" to the read. For reads where no haplotype block and phase assignment is possible, e.g. reads without enough het-SNP calls, we simply assign a sentinel value for the block identifier to trace the un-phased reads.

The phasing method presented here is rather generic. It performs reasonably as shown in the evaluation work presented in the main manuscript. We anticipate if we integrate more sophisticate algorithms for phasing can further increase the accuracy and phasing contiguity.

Incorporate phased reads to haplotype-fused string graph for unzipping collapsed paths

Supplementary Fig. 13(**a**) shows the schematic of how to merge the haplotype specific assembly graph H_c to the initial haplotype-fused graph $G_c^{(f)}$ and reconstruct the haplotype tiling paths for generating the haplotigs. The haplotype specific assembly graph has 4 disconnected subgraphs inside the cyan boxes, two for each haplotype. If we add the edges and nodes from the H_c to the haplotype-fused graph in the yellow box, we get the intermediate graph shown in the upper right panle. In the graph, the red nodes and blue nodes indicate the reads from the two different haplotypes. There are some edges connecting the nodes from two different haplotypes. Once the haplotype of the reads are known, we can remove those edges. This results in the graph in lower panel. We can construct the tiling paths for the two haplotigs from the two disconnected and haplotype specific components.

Supplementary Fig. 13(**b**) shows an example constructing the primary contig and haplotigs from the *Clavicorona pyxidata* assembly. Along the initial primary contig, we can find 4 haplotype blocks and an un-phased region. By applying the FALCON-Unzip process, we bring back some reads to fully reconstruct the haplotigs. In this example, the four initial haplotype block can be further merged by the overlaps across the bubbles in the assembly graph.

Augmented alignments for haplotype specific quiver consensus

Once we construct the final primary contigs and the associated haplotigs, each of them may contain one of more known haplotype blocks and specific phase of each block. Each read also has a tuple "(block identifier, phase identifier)". The tuple label is used to separate the read into different groups and aligned to the specific primary contigs or haplotigs specifically for generating the final consensus with Quiver. Since each set of the alignments on each haplotigs or primary contigs is already specific to the correspond haplotypes, we can apply the Quiver consensus algorithm that is designed for single haplotype to get accurate consensus sequences. In the future, we can design an aligner taking the extra information of the reads and the references to do haplotype match alignment without manually separating reads into different files and aligning separately.

Analysis parameters and commands for the analysis

Daligner and FALCON parameter choices for error correction

Here are the algorithm-related assembly parameters used for the *in silico* FALCON F1 *Arabidopsis* assembly:

```
length_cutoff = 4000
length_cutoff_pr = 4000
pa_HPCdaligner_option =  -v -dal128  -e0.75 -M24 -l1800 -k18 -h240 -w8 -s100
ovlp_HPCdaligner_option =  -v -dal128  -M24 -k24 -h750 -e.96 -l1500 -s100
pa_DBsplit_option = -a -x500 -s400
ovlp_DBsplit_option = -s400
falcon_sense_option = --output_multi --output_dformat --min_idt 0.70 --min_cov 4 --
max_n_read 400 --n_core 8
falcon_sense_skip_contained = False
overlap_filtering_setting = --max_diff 80 --max_cov 120 --min_cov 4 --n_core 1
```

In the parameters passed to "daligner" for error correction, we use "-l1800 –k18 –h240" to reduce the number of local alignments reported by daligner to increase computational efficiency.

Here are the algorithm related assembly parameters for the *V. vinifera* Assembly:

```
length_cutoff = 5000
length_cutoff_pr = 5000

pa_HPCdaligner_option =  -v -dal128  -e0.75 -M24 -l2500 -k18 -h1250 -w8 -s100
```

```
ovlp_HPCdaligner_option =  -v -dal128  -M24 -k24 -h1250 -e.96 -l1500 -s100
pa_DBsplit_option = -a -x500 -s200
ovlp_DBsplit_option = -s200
falcon_sense_option = --output_multi --output_dformat --min_idt 0.70 --min_cov 4 --
max_n_read 400 --n_core 8
falcon_sense_skip_contained = False
overlap_filtering_setting = --max_diff 120 --max_cov 120 --min_cov 4 --n_core 12
```

Here are the algorithm related assembly parameters for the *Clavicorona pyxidata* Assembly:

```
length_cutoff = 7500
length_cutoff_pr = 7500
pa_HPCdaligner_option =  -v -dal128 -t16 -e0.75 -M24 -l3200 -k18 -h480 -w8 -s100
ovlp_HPCdaligner_option =  -v -dal128  -M24 -k24 -h1024 -e.96 -l2500 -s100
pa_DBsplit_option = -a -x500 -s200
ovlp_DBsplit_option = -s200
falcon_sense_option = --output_multi --output_dformat --min_idt 0.70 --min_cov 4 –
max_n_read 200 --n_core 8
falcon_sense_skip_contained = False
overlap_filtering_setting = --max_diff 120 --max_cov 120 --min_cov 2 --n_core 12
```

In the parameters passed to “daligner” for error correction, we use “-l3200 -k18 -h480” to reduce the number of local alignment reported by daligner to increase computational efficiency.

Canu assemblies

Canu (https://github.com/marbl/canu), a fork of Celera Assembler designed specifically for noisy single molecule data was used to generate assemblies for the *Arabidopsis* Col-0 x Cvi-0 F1 cross, *Vitis vinifera* cv. Cabernet Sauvignon as well as *Clavicorona pyxidata.* Refer to the Canu documentation for details about the assembly method. In brief there are 3 steps that occur in the Canu pipeline:

1) Read correction. The MHAP algorithm (Berlin et al. 2015) is used to detect overlaps in the raw noisy subreads which are subsequently corrected using the overlapping data.
2) Read trimming. Overlap data is used to identify and remove low quality regions from the previously corrected consensus sequences.

3) Unitig construction. Celera Assembler is used to generate consensus Unitig sequences from the corrected high quality sequencing data.

An error rate of 0.025 coupled with the estimated haploid genome sizes of 120 Mb, 475 Mb and 44 Mb were the parameters used to assemble *Arabidopsis, Vitis vinifera* cv. Cabernet Sauvignon and *Clavicorona pyxidata.* This equates to roughly ~165X, 158X and 88X starting input coverage for the three species respectively.

Short-read assemblies

The illumina reads were error corrected using Lighter(Song et al. 2014) (version 1.0.7). To obtain the best achievable results we used once the raw and once the error corrected reads for the assembly with SOAPdenovo(Li et al. 2010) (version 2.04) and Platanus(Kajitani et al. 2014) (version 1.2.4). SOAPdenovo was run with the "all" option and a range of kmer sizes "-K from 19 to 127 in order to obtain the best results. Platanus was executed with the default parameters since it internally evaluates several values of k.

The resulting contigs and scaffolds were compared to the existing reference genomes or in the case of the *Clavicorona pyxidata* to the existing PacBio based de novo assembly using an in house script to compute the N50 and NG50.

Overall both short-read assemblers run successfully on our data. However, in only one case did we not obtain a successful run from Platanus, and that was with the *V. vinifera* cv. Cabernet Sauvignon genome. We identified that the coverage (here 45x) must be the problem since recently a successful de novo assembly of this *V. vinifera* cv. Cabernet Sauvignon was reported(Patel et al. 2015) using over 1,577 million reads in three different insert size libraries.

Supplementary Fig. 5 shows the different results for SOAPDeNovo using different kmer sizes as parameters and their impact on N50 and NG50. In our case we selected those assemblies having the largest N50 or NG50 for comparison to FALCON-Unzip.

BUSCO command line options

A pre-release set of plant orthologs obtained from BUSCO authors is used for evaluations. The BUSCO outputs are generated by the following command:

```
python3 BUSCO_plants.py -in <asm>.fasta -o <output dir name> -l
./plant_early_release/
```

The `./plant_early_release/` directory contains the plant ortholog set. AUGUSTUS v.3.0.3 is used along with the BUSCO script for *de novo* gene prediction.

CDS predictions and comparisons

We use AUGUTUS 3.2.1 for predicting transcripts and CDS in *de novo* mode (with `--species=arabidopsis` option, but no RNA seq or transcript data provided.) The GFF files contain the predictions are generated by:

```
augustus --species=arabidopsis --codingseq=on <assembly>.fasta  > <assembly>.gff
```

Coding sequences are extracted with the `getAnnoFasta.pl` script, e.g.,

```
perl ../../augustus-3.2.1/scripts/getAnnoFasta.pl  cns_ph_ctg_gene.gff
```

The predicted coding sequences between different assemblies are aligned against each other with BLASR to generate SAM files. The CIGAR string in the SAM is scanned to filter out any entry that has insertion, deletion, hard clipping and soft clipping tags for identifying full alignments (mismatch allowed). Here is an example comparing the predicted coding sequences from TAIR10 to those of the primary contigs:

```
blasr TAIR10_all.codingseq.fa cns_p_ctg_gene.codingseq.fa -noSplitSubreads -bestn 1 -m 4 -nproc 48 -maxLCPLength 15 -nCandidates 24 -sam -out TAIR10_cns_p.sam

cat TAIR10_cns_p.sam TAIR10_cns_h.sam | awk '$0 !~/@/ && $6 !~ /[ISHD]/ && $6 != "*" {print $1}' |  sort -u | wc
```

Synteny analysis of *Arabidopsis* and Cabernet Sauvignon

In order to identify syntenic regions between primary contigs in both the Arabidopsis F1 and the cabernet assemblies, Satsuma(Grabherr et al. 2010), an alignment algorithm that identifies sequence matches through cross-correlation and fast Fourier transfrom was applied. In separate analyses, primary contigs from each genome were used as targets while all other contigs were subsequently checked for syntenic relationships.

The command to generate the alignments was:

SatsumaSynteny -self 1 -q <query>.fasta -t <target>.fasta -o <output_dir> -ni 12 -n 24

The matches for each contig are collected in a satsuma_summary.chained.out file and subsequently concatenated with other target contig matches. Reciprocal (redundant) hits were removed from the table which is available as **Supplementary Data 2-3** (`synteny_arabF1_160425.csv, synteny_cabernet_160421.csv`). To find the total

quantity of syntenic bases, tend – tstart was summed across all matches to yield roughly ~18.9Mb for Arabidopsis and ~25.3 Mb for Cabernet.

BWA alignment for Illumina short reads to PacBio *Clavicorona pyxidata* assembly

We used the following command line options for BWA (v.0.7.12, http://sourceforge.net/projects/bio-bwa/files/bwa-0.7.12.tar.bz2/download) alignment of Illumina short reads to PacBio *Clavicorona pyxidata* assembly contigs:

(1) Aligning the short reads to all contigs:

```
bwa mem  -M -t 48 all_quivered_ctg_cns.fasta SRR1800147_1.fastq SRR1800147_2.fastq |
samtools view -@ 24 -1 -b -S - | samtools sort -@ 8 -m 8G - ilmn_to_all_aln_sorted
```

(2) Aligning the short reads to primary contigs:

```
bwa mem  -M -t 48 all_p_cns.fasta SRR1800147_1.fastq SRR1800147_2.fastq |  samtools
view -@ 24 -1 -b -S - | samtools sort -@ 8 -m 8G - ilmn_to_pcns_aln_sorted
```

Note that there are many regions larger than 500bp regions between the haplotigs and their primary contigs that are 100% identical if there are no het-SNPs in those regions. `bwa mem` will not be able to uniquely map some reads as those homologous regions are in both primary contigs and the associated contigs . 'bwa mem' will assign low map QV of the aligned reads in the homologous regions.

(3) Align RNA-seq reads to the assembly

```
mkdir -p ./cns_p_ctg; STAR --genomeFastaFiles ../../4-
quiver_keep/cns_output/cns_p_ctg.fasta  --genomeDir cns_p_ctg --runMode
genomeGenerate

STAR --genomeDir cns_p_ctg/ --runThreadN 48 --readFilesIn SRR1589642_1.fastq
SRR1589642_2.fastq --outSAMtype BAM SortedByCoordinate --alignIntronMax 5000 --
outReadsUnmapped Fastx
```

(4) Align CDS annotation (Clapy2_GeneCatalog_CDS_20150910.fasta) to the contig:

```
STARlong --genomeDir ../RNA_aln/cns_p_ctg/  --runThreadN 48  --outFilterMismatchNmax
100 --seedSearchLmax 30   --seedSearchStartLmax 30 --seedPerReadNmax 100000 --
seedPerWindowNmax 100 --alignTranscriptsPerReadNmax 100000 --
alignTranscriptsPerWindowNmax 10000 --readFilesIn
Clapy2_GeneCatalog_CDS_20150910.fasta --outSAMtype BAM SortedByCoordinate
```

Genome-wide Alignment for Col-0 and Cvi-0 to TAIR10 comparsion and Assemblytics analysis

The “delta” files are generated using the following MUMmer3 commands for Assemblytics analysis and a list of variations at SNP level is generated by `show-snps` command:

```
nucmer -maxmatch -l 100 -c 1000 asm.fa TAIR10_all.fa -p asm
delta-filter -l 10000 -1 asm.delta > asm_l10000.delta
show-snps -C -x 20 -H -T -l –q asm_l10000.delta > asm_l10000.delta.snps
```

For **Fig. 2**, the map between the assembly contigs to TAIR10 is generated by first identifying maximal exact matches between the genomes. For computational efficiency, we used a re-implementation of the Mummer algorithm (https://github.com/godotgildor/sparseMEM-big). The generated max-exact-matches are clustered using “mgaps” from the Mummer3 package. The chains clustered are parsed and the range of the contig mapped to TAIR10 is output as bed files and used for plotting **Fig. 2**. A similar procedure for aligning contigs to a reference and performing draft calls for large genome structural variations is documented (https://github.com/PacificBiosciences/FALCON/wiki/Some-Tricks-Using-Falcon-To-Detect-Structure-Variations). We only use this process to anchor the contigs to the reference. The SV calls in paper are generated by the Assemblytics using the standard MUMmer/nucmer alignments.

Assembly example in the cloud

We generated an Amazon AWS EBS volume that has the data and the configuration for generating the *Clavicorona pyxidata* assembly results end-to-end with a local single node configuration. The step-by-step instruction to execute FALCON and FALCON-Unzip with AWS EC2 is in the **Supplementary Data 4** (`AWS_FALCON-Unzip_Example.pdf`). For large genomes, a HPC cluster environment with Sun Grid Engine is currently supported by FALCON and FALCON-Unzip and special optimization for local cluster environment to optimize CPU and I/O throughput will be necessary in general. For example, see recent FALCON assembly for gorilla genome (Gordon et al. 2016).

Software revision used for this work

```
FALCON-Integrate:
URL: https://github.com/PacificBiosciences/FALCON.git
Git Revision: ffbc609057741896dbe9d09f65bc5c8ef8d57f5a
Git Tag: https://github.com/PacificBiosciences/FALCON-
integrate/tree/funzip_052016

FALCON:
URL: git://github.com/PacificBiosciences/FALCON.git
Git Revision: a1dd4d45bbe0144842cabdd378c46278744231f7
Git Tag:
https://github.com/PacificBiosciences/FALCON/tree/funzip_052016
```

Daligner:
URL: git://github.com/PacificBiosciences/DALIGNER.git
Git Revision: 029bfa8a40b456bad5499f6e456e6a215bfd307c

DAZZ_DB:
URL: git://github.com/PacificBiosciences/DAZZ_DB.git
Git Revision: ed0b85e6f14dc394fbabd0731187f98e08a79a0a

pypeFLOW:
URL: git://github.com/PacificBiosciences/pypeFLOW.git
Git Revision: e880e2b3cebe7ae19f3101a497a0a8cc6685588e

FALCON-Unzip:
URL: https://github.com/PacificBiosciences/FALCON_unzip
Git Revision: 21b1df3491e3bb7b9d8ecd13fc0c9c1a45b6393f
Git Tag:
https://github.com/PacificBiosciences/FALCON_unzip/tree/funzip_052016

Canu: https://github.com/marbl/canu
Revision: Canu v1.0 r7237
Git Tag:
https://github.com/marbl/canu/commits/f00ca16e9d83a225e8f94c42cc52900456a4a6b0

Lighter: version 1.0.7

SOAPdenovo: version 2.0.4

Platanus: version 1.2.4

References

Bansal V, Bafna V. 2008. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**: i153-159.

Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology*.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC bioinformatics* **13**: 238.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Gordon D, Huddleston J, Chaisson MJ, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW. 2016. Long-read sequence assembly of the gorilla genome. *Science* **352**: aae0344.

Grabherr MG, Russell P, Meyer M, Mauceli E, Alföldi J, Di Palma F, Lindblad-Toh K. 2010. Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* **26**: 1145-1151.

Halldórsson BV, Bafna V, Edwards N, Lippert R, Yooseph S, Istrail S. 2004. A Survey of Computational Methods for Determining Haplotypes. In *COMPUTATIONAL METHODS FOR SNPS AND HAPLOTYPE INFERENCE*, (ed. S Istrail, et al.), pp. 26--47. Springer.

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H et al. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res* **24**: 1384-1395.

Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, Chin C-S, Rapicavoli NA, Rank DR, Li J. 2014. Long-read, whole-genome shotgun sequence data for five model organisms. *Scientific data* **1**.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome biology* **5**: R12.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* **20**: 265-272.

Myers EW. 1986. An O (ND) difference algorithm and its variations. *Algorithmica* **1**: 251-266.

Myers EW. 2005. The fragment assembly string graph. *Bioinformatics* **21**: ii79-ii85.

Myers Jr EW. 2016. A history of DNA sequence assembly. *it-Information Technology* **58**: 126-132.

Patel S, Swaminathan P, Fennell A, Zeng E. 2015. De novo genome assembly tool comparison for highly heterozygous species Vitis vinifera cv. Sultanina. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pp. 1771-1773. IEEE.

Song L, Florea L, Langmead B. 2014. Lighter: fast and memory-efficient sequencing error correction without counting. *Genome Biol* **15**: 509.

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**: ii215-ii225.