Article:
Bancroft, Ian orcid.org/0000-0001-7707-1171 and He, Zhesi orcid.org/0000-0001-8335-9876 (2018) Organisation of the genome sequence of the polyploid crop species Brassica juncea. Nature genetics. pp. 1496-1497. ISSN 1546-1718

https://doi.org/10.1038/s41588-018-0239-0

Organisation of the genome sequence of the polyploid crop species *Brassica juncea*

To the Editor:

A draft genome sequence of *Brassica juncea*, a member of the Brassicaceae and therefore a species benefiting from the functional genomics advances in the "model" species *Arabidopsis thaliana*, was reported recently by Yang et al[1]. *B. juncea* is a recently-formed allotetraploid, the diploid progenitors of which were mesohexaploids: *B. rapa* (which contributed the A genome) and *B. nigra* (which contributed the B genome). In addition to underpinning future trait-oriented work in this important crop species, which includes both vegetable and oil types, the sequences were analysed for characteristics of genome evolution under crop selection. For both purposes, the genome sequences must represent with high fidelity (though not perfectly in "draft" form), both the gene complement and gene order of the species. As a model for addressing the challenges of achieving an adequate representation of the latter for allopolyploid crops, the construction methodology employed short shotgun sequence reads, single-molecule long reads, BioNano sequencing and high-resolution genetic mapping.

A particular problem in genetic mapping in polyploids is the confounding effects of single nucleotide polymorphisms (SNPs) resulting from inter-homoeologue polymorphisms (IHPs), which are much more abundant than the allelic SNPs that are needed for genetic (linkage) mapping. In species such as *B. juncea* and *B. napus* (which also contains the A genome contributed by a *B. rapa* progenitor, but in this species in combination with the C genome contributed by a *B. oleracea* progenitor), further complications arise from the mesohexaploid nature of the genomes of the diploid progenitors, resulting in inter-paralogue polymorphisms (IPPs). However, so long as sufficient sequencing redundancy has been obtained to overcome stochastic sampling effects and differentiate allelic SNPs (which will segregate across a linkage mapping population) from IHPs and IPPs (which should be invariant), the confounding effects can be overcome. Even using transcriptome sequence data, robust methodologies have been developed in *B. napus* to score allelic SNPs for high resolution linkage map construction and to underpin association genetics[2-4].

We aimed to test the fidelity with which the genome sequence reported by Yang et al[1] represents the gene order of *B. juncea* by comparing that with our own estimates using an AB *Brassica* genomics platform constructed as for our AC *Brassica* genomics platform[5], based on the sequences of the progenitor species *B. rapa* (A genome) and *B. nigra* (B genome) (Supplementary Note). For the test, we used the CDS gene models from (1) the AB *Brassica* genomics platform and (2) the *B. juncea* genome sequence of Yang et al[1]

(denoted J genome) as the reference sequences for mapping Illumina mRNAseq reads from 106 lines of the *B. juncea* VHDH mapping population[6, 7] with variant-calling essentially as described previously for *B. napus*[2, 3, 4] (Supplementary Note, Life Sciences Reporting Summary). The SNP scoring strings were filtered to retain only simple SNPs (i.e. polymorphisms between resolved bases) and displayed in genome sequence order as genome-ordered graphical genotypes (GOGGs). If the order in the genome sequence of the genes in which the polymorphisms are scored is correct, the result should resemble a genetic linkage map, i.e. with few instances of nearby alternating parental alleles in individual recombinant lines. The GOGGs generated comprised 33,059 scored SNP markers for the AB *Brassica* genomics platform and 29,834 scored SNP markers for the *B. juncea* genome sequence reported by Yang et al[1] (Supplementary Figure 1). An example, for chromosome J1 of Yang et al[1] compared with A1 from the AB *Brassica* genomics platform, is shown in Figure 1. The results of this simple quality control assessment show that the authentic arrangement of genes in *B. juncea* matches very well that of their orthologues in the AB reference, and hence in the progenitor species, but they also show that the *B. juncea* genome sequence reported by Yang et al[1] is extensively mis-assembled. We note also that the internationally-agreed nomenclature for B genome chromosomes[8], which we followed for the AB resource, was not followed for the *B. juncea* genome sequence.

The assembly and validation methodology described by Yang et al[1] sounds plausible and may well be taken as a model to follow for other polyploid crops, so why was it ineffective? Detailed inspection of the GOGGs suggests two problems: chimeric assemblies (where collinearity with the genome of *A. thaliana* breaks down) and mistaking IHPs or IPPs for allelic SNPs when undertaking the linkage mapping with the 5,333 “bin markers” or in the pre-existing linkage map (where collinearity with the genome of *A. thaliana* is maintained). The bin markers appear to have been scored on the basis of only ~0.7-fold redundant genome re-sequencing, which wouldn’t be sufficient (in SNP scoring) to differentiate the differing types of polymorphisms (IHPs, IPPs and allelic SNPs) in polyploid genomes. It is less clear why use of the single-molecule long reads and BioNano sequencing failed to detect the chimerism.

Although the draft of the *B. juncea* genome sequence reported by Yang et al[1] does not appear to faithfully represent the organization of that genome, undermining analyses requiring positional information (such as illustrated in Figures 1, 2a, 3 and 4a in the report of Yang et al[1]), it could easily be improved by exploiting the linkage mapping information depicted by the GOGGs. Indeed, the B genome component of our AB *Brassica* genomics

platform was based on the *B. nigra* genome sequence reported by Yang et al[1] alongside that of *B. juncea* and was developed by splitting it (into 175 segments) and re-organising based on the transcriptome SNPs scored across the *B. juncea* VHDH mapping population. The assessment of genome assemblies based on GOGGs therefore not only represents an important quality control measure, it also provides a solution where problems are found. Linkage mapping populations have been a fundamental resource for the genetic analyses of traits in crop so will usually be available already in crop species for which genome sequencing is being undertaken. To help assure the quality of genome sequences, we would like to propose an expectation that validation by means of GOGGs should be incorporated into the assembly workflow for polyploid crop genomes.

ACKNOWLEDGEMENTS

This work was supported by UK Biotechnology and Biological Sciences Research Council (BB/L002124/1, BB/L011751/1), including work carried out within the ERA-CAPS Research Program (BB/L027844/1). We would like to thank Isobel Parkin and Andrea Harper for their valuable comments on a draft of this manuscript.

AUTHOR CONTRIBUTIONS

Z.H. and I.B. designed the study and analysed the data. I.B. wrote the manuscript and Z.H. read and approved the manuscript for publication.

COMPETING FINANIAL INTERESTS

The authors declare no competing financial interests.

DATA AVAILABILITY

The *B. juncea* mRNAseq data used for production of the graphical genotypes have been deposited in the SRA data library under project ID PRJNA471033.

Zhesi He & Ian Bancroft

Department of Biology, University of York, Heslington, York, YO10 5DD, UK

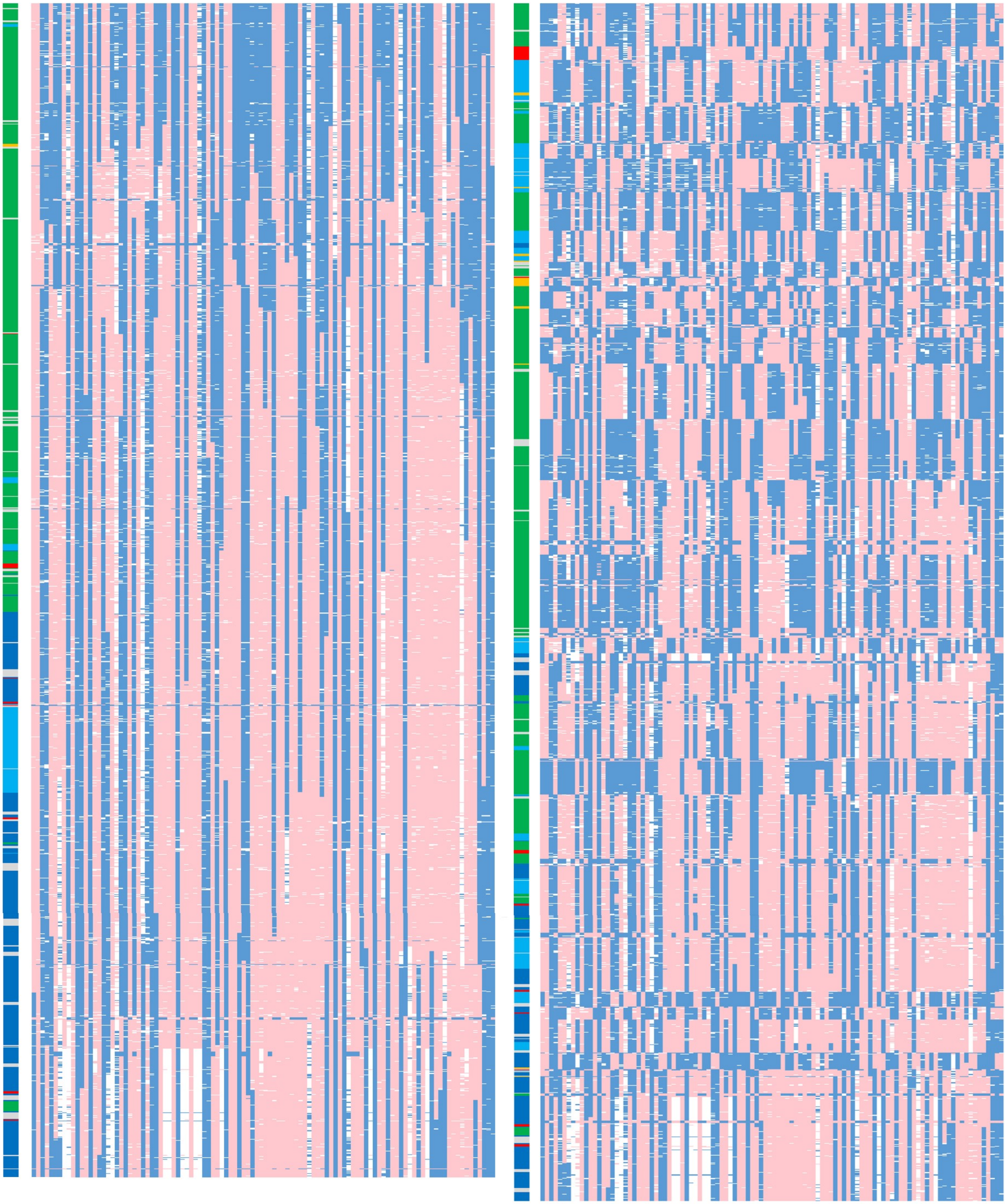
Email: ian.bancroft@york.ac.uk

1. Yang, J. *et al. Nat. Genet.* **48**, 1225-1232 (2016).
2. Trick, M. *et al, Plant Biotechnol. J.* **7**, 334-346 (2009).
3. Bancroft, I *et al, Nat. Biotechnol.* **29**, 762-766 (2011).
4. Harper, A.L. *et al, Nat. Biotechnol.* **30**, 798-802 (2012).

5. He, Z. *et al. Data in Brief* **4**, 357-362 (2015).
6. Paritosh *et al. BMC Genomics* **15**, 396 (2014).
7. He, Z. *et al. Plant Biotechnol. J.* **15**, 594-604 (2017).
8. King, G. https://dx.doi.org/10.4226/47/5afb8519d194c (2010).

FIGURE LEGENDS

Figure 1. Quality control assessment of genome sequence organisation of *B. juncea* using genome-ordered graphical genotypes, chromosomes A1 and J1 as an example. Graphical genotypes are shown for transcriptome SNP markers scored across 106 lines of the VHDH mapping population with Heera alleles in coral, Veruna alleles in blue and missing scores in white. The genotypes for 2004 and 2040 markers are shown for chromosomes A1 and J1, respectively. The multi-coloured bars are colour-coded to the chromosome of the top BLAST sequence similarity match in *Arabidopsis thaliana* of the *Brassica* gene model in which the SNP is scored (light blue = chromosome 1, orange = chromosome 2, dark blue = chromosome 3, green = chromosome 4, red = chromosome 5, light grey = no BLAST hit with E-value < 1e30).


A1
J1

Organisation of the genome sequence of the polyploid crop species *Brassica juncea*

Zhesi He & Ian Bancroft
Department of Biology, University of York, Heslington, York, YO10 5DD, UK
Email: ian.bancroft@york.ac.uk

Supplementary note

Production of genome-ordered graphical genotypes for *B. juncea*

For our genomics platform, we used the A genome component of the *B. napus* AC Pan-transcriptome resource[1], which was based on the version 2.0 *B. rapa* genome sequence[2], with minor updates, and a newly-developed B genome component. A total of 88,713 CDS models were extracted from the genome resources to form the AB transcriptome reference sequence. Illumina mRNAseq reads from 106 lines of the *B. juncea* VHDH mapping population[3] were mapped with this AB transcriptome reference and SNPs scored using methodology developed and described previously for *B. napus*[4-7]. The SNP scoring strings were filtered to remove hemi-SNPs (i.e. instances where the most frequent or second most frequent allele scored is an ambiguity code representing more than one base). The remaining SNPs were output to MS Excel files with each row representing, in order: (1) the SNP identifier; (2) genome coordinate (chromosome_start nucleotide_end nucleotide) of the CDS gene model in which the SNP was scored; (3) best BLAST nucleotide sequence similarity match of the gene model with *Arabidopsis thaliana* gene models (with conditional formatting coded to the chromosome of the *A. thaliana* gene model); (4) the name of the gene model in which the SNP was scored; (5) simple SNP flag; (6) nucleotide allele in Heera parent; (7) nucleotide allele in Veruna parent; (8-113) the graphical genotypes as the parental allele calls for each of the 106 lines of the VHDH mapping population (with A corresponding to the Heera allele, B corresponding to the Veruna allele and conditional formatting coded to A or B allele). The spreadsheet was sorted by genome coordinate of the gene models in which the SNPs were scored, row height was set to 1 pixel and screen shot images compiled in MS PowerPoint to display the genome-ordered graphical genotypes (GOGGs).

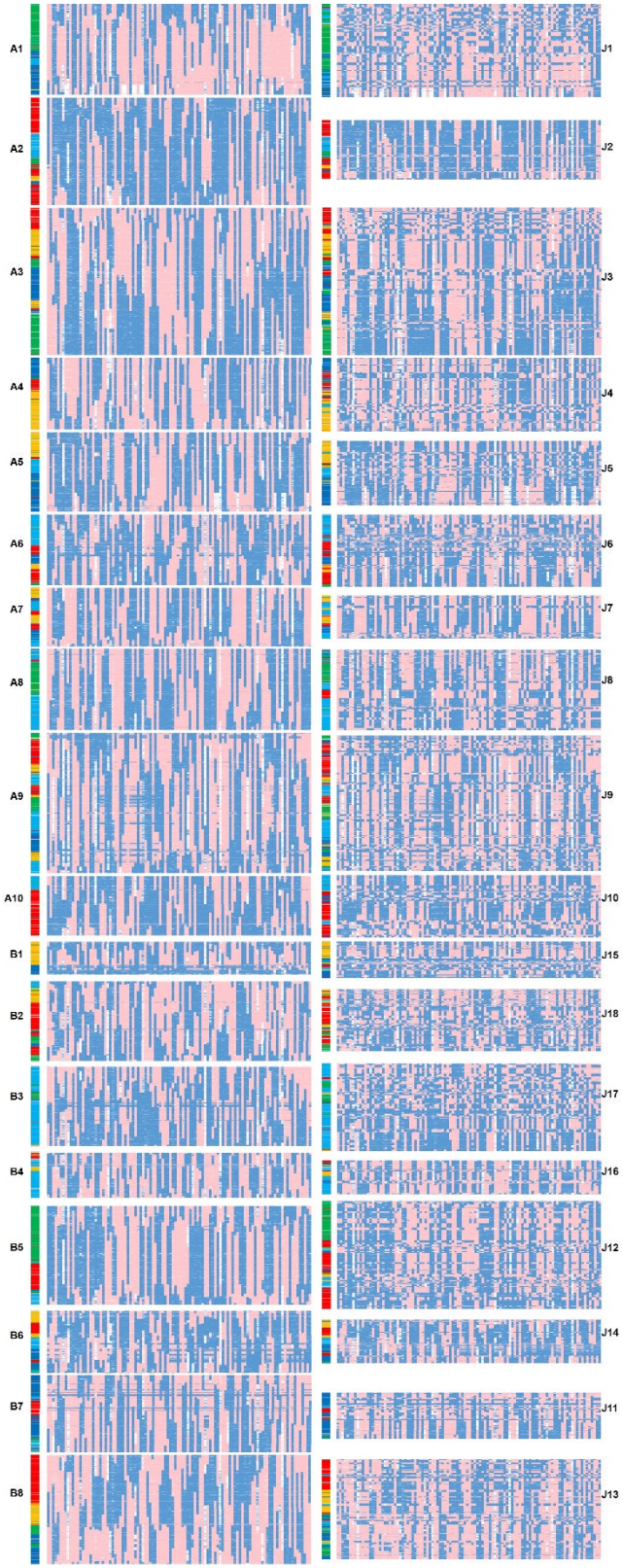
Re-assembly of the *B. nigra* genome

The *B. nigra* genome sequence as reported by Yang et al[8] was first imaged as a GOGG, in combination with the *B. rapa*-derived A genome and using Illumina mRNAseq reads from 106 lines of the *B. juncea* VHDH mapping population[3], as described above. This revealed extensive mis-assembly as disjoint blocks of markers with consistent graphical genotypes.

The mis-assembled blocks of scored markers were rearranged manually in the MS Excel spreadsheet underlying the GOGG. Based on the end-most genes in the blocks with consistent genotypes, positions in the chromosome assemblies were identified visually for splitting, using an MS Excel spreadsheet list of gene models arranged by genome coordinate. The split sites were chosen either as the mid-point between genes representing the positions of discontinuities in collinearity with the *A. thaliana* genome (indicative of chimeric scaffolds) or, where collinearity with *A. thaliana* was maintained, as the mid-point between genes representing the positions of discontinuities in gene model nomenclature (indicative of mis-mapping of scaffolds to homoeologous/paralogous positions). We developed the new B genome resource by splitting the published genome into 175 segments and re-concatenating them to be consistent with the linkage mapping shown by the SNP scoring strings in the graphical genotypes (Supplementary Table 1). The chromosomes were then re-numbered to match the international convention[9]. New coordinates for the gene models were generated based on best BLAST similarity match in the B genome re-assembly (E-value < 1e30). Finally, the assembly was validated by producing a GOGG based on the new genome coordinates of the gene models, as shown in the B genome section of Supplementary Figure 1.

1. He, Z. *et al. Data in Brief* **4**, 357-362 (2015).
2. Cai, C. *et al. Mol. Plant.* **10**, 649-651 (2017).
3. Paritosh *et al. BMC Genomics* **15**, 396 (2014).
4. Trick, M. *et al, Plant Biotechnol. J.* **7**, 334-346 (2009).
5. Harper, A.L. *et al, Nat. Biotechnol.* **30**, 798-802 (2012).
6. Bancroft, I *et al, Nat. Biotechnol.* **29**, 762-766 (2011).
7. He, Z. *et al. Plant Biotechnol. J.* **15**, 594-604 (2017).
8. Yang, J. *et al. Nat. Genet.* **48**, 1225-1232 (2016).
9. King, G. https://dx.doi.org/10.4226/47/5afb8519d194c (2010).

Supplementary Figures



Supplementary Figure 1. Quality control assessment of genome sequence organisation of *B. juncea* using genome-ordered graphical genotypes. Graphical genotypes are shown for transcriptome SNP markers scored across 106 lines of the VHDH mapping population with

Heera allele in coral, Veruna alleles in blue and missing scores in white. The graphical genotypes are organised by linkage group and labelled using the international convention for *Brassica* chromosome nomenclature (A1 to A10 and B1 to B8; genotypes for 33,059 markers shown) or the nomenclature used by Yang et al[8] (J1 to J18; genotypes for 29,834 markers shown). The multi-coloured bars are colour-coded to the chromosome of the top BLAST sequence similarity match in *Arabidopsis thaliana* of the *Brassica* gene model in which the SNP is scored (light blue = chromosome 1, orange = chromosome 2, dark blue = chromosome 3, green = chromosome 4, red = chromosome 5, light grey = no BLAST hit with E-value < 1e30).

Supplementary tables

Supplementary Table 1. Re-build specification for the *B. nigra* genome. Nucleotide coordinates for blocks of genome sequence refer to the original chromosome assemblies of Yang et al[8]. The chromosome (Chr) nomenclature in the re-assembly corresponds to that of Yang et al[8]. The international nomenclature[9] B1, B2, B3, B4, B5, B6, B7 and B8 correspond to Yang et al[8] chromosomes B5, B8, B7, B6, B2, B4, B1 and B3, respectively.

Chr	Block	Start nucleotide	Stop nucleotide	Orientation
B01	1	B03_007582801	B03_007885645	fwd
B01	2	B08_032495400	B08_032871719	fwd
B01	3	B01_009064427	B01_009778984	rev
B01	4	B01_007582779	B01_008477874	rev
B01	5	B01_007217776	B01_007582778	fwd
B01	6	B01_006942609	B01_007217775	rev
B01	7	B01_008477875	B01_009064426	rev
B01	8	B01_002065798	B01_006485396	rev
B01	9	B01_000000001	B01_000498900	fwd
B01	9.5	B02_013890010	B02_014921067	fwd
B01	10	B08_018602377	B08_018802048	fwd
B01	11	B06_010596212	B06_010889573	fwd
B01	12	B06_027053519	B06_027293964	rev
B01	13	B07_004574023	B07_004606846	fwd
B01	14	B06_002493704	B06_002720701	fwd
B01	15	B04_015647478	B04_015901751	fwd
B01	16	B01_009778985	B01_010343995	rev
B01	17	B01_000498901	B01_002065797	fwd
B01	18	B01_012483451	B01_013591901	fwd
B01	19	B01_010343996	B01_012483450	rev
B01	20	B01_006485397	B01_006942608	fwd
B01	21	B01_013591902	B01_015582396	rev
B01	22	B01_015582397	B01_029179896	fwd
B01	23	B04_031974885	B04_032085266	rev
B01	24	B01_029564619	B01_029798227	fwd
B01	25	B01_029798228	B01_030315192	fwd
B01	26	B01_030411022	B01_030653157	rev
B01	27	B01_030653158	B01_999999999	rev
B02	28	B07_021173313	B07_022247982	fwd
B02	29	B02_003989485	B02_005436514	rev
B02	30	B02_005436515	B02_006692969	rev
B02	31	B02_002437988	B02_003989484	rev
B02	32	B02_007736853	B02_009122239	fwd
B02	33	B06_017792993	B06_017875548	fwd
B02	34	B02_000450561	B02_002437987	fwd
B02	35	B07_004606847	B07_004869051	fwd
B02	36	B02_016512896	B02_017518240	rev
B02	37	B02_019008138	B02_020832038	rev
B02	38	B02_000000001	B02_000450560	fwd

B02	39	B02_017518241	B02_019008137	rev
B02	40	B07_018021702	B07_018298751	fwd
B02	41	B02_015825373	B02_016512895	rev
B02	42	B02_010033202	B02_010585834	fwd
B02	43	B02_013467424	B02_013890009	fwd
B02	44	B06_005692483	B06_006252255	rev
B02	45	B02_020832039	B02_021732776	fwd
B02	46	B02_010585835	B02_011874152	rev
B02	47	B02_009492751	B02_010033201	rev
B02	48	B02_032414389	B02_033026168	fwd
B02	49	B02_021732777	B02_025864381	fwd
B02	50	B05_019757534	B05_019790035	fwd
B02	51	B02_025864382	B02_032414388	fwd
B02	52	B02_033026169	B02_035498868	fwd
B02	53	B02_035710914	B02_037794611	fwd
B02	54	B02_039028808	B02_039273359	fwd
B02	55	B02_037794612	B02_039028807	fwd
B02	56	B02_039273360	B02_042454026	fwd
B02	57	B02_043886028	B02_044029580	fwd
B02	58	B02_014921068	B02_015825372	rev
B02	59	B02_042454027	B02_043130084	fwd
B02	60	B02_043643036	B02_043886027	rev
B02	61	B02_043130085	B02_043277069	fwd
B02	62	B02_044029581	B02_999999999	fwd
B02	63	B07_012423924	B07_012534505	fwd
B03	64	B03_000000001	B03_000578874	fwd
B03	64.5	B02_043277070	B02_043643035	rev
B03	65	B05_037355021	B05_038578986	fwd
B03	66	B03_000578875	B03_004697856	fwd
B03	67	B03_005434935	B03_007331587	fwd
B03	68	B03_007331588	B03_007572334	rev
B03	69	B03_007572335	B03_007582800	fwd
B03	69.1	B03_013520776	B03_014015714	rev
B03	69.2	B03_016414470	B03_017979820	fwd
B03	69.3	B03_014015715	B03_016414469	fwd
B03	69.4	B03_013400580	B03_013520775	rev
B03	69.5	B08_024905159	B08_025092390	rev
B03	69.6	B03_012624233	B03_012929964	rev
B03	69.7	B06_026686062	B06_026697744	rev
B03	69.8	B03_009392611	B03_011730707	rev
B03	69.9	B03_007885646	B03_008708174	rev
B03	70	B03_004697857	B03_005434934	rev
B03	77.2	B03_017979821	B03_020071193	fwd
B03	78	B02_035498869	B02_035525737	fwd
B03	79	B08_037718029	B08_037955293	rev
B03	80	B02_035525738	B02_035710913	rev

B03	81	B03_029973613	B03_030697003	fwd
B03	82.1	B03_020071194	B03_025868248	fwd
B03	82.2	B03_029098122	B03_029973612	fwd
B03	82.3	B03_025868249	B03_026922726	fwd
B03	83	B03_027152666	B03_027645303	fwd
B03	84	B04_027715113	B04_027734630	rev
B03	85	B03_027645304	B03_029098121	fwd
B03	86	B03_030697004	B03_038201514	fwd
B03	87	B03_038327909	B03_042763060	fwd
B03	88	B03_042883456	B03_043298766	fwd
B03	89	B03_043837672	B03_043951567	fwd
B03	90	B03_043298767	B03_043837671	fwd
B03	91	B03_042763061	B03_042883455	fwd
B03	92	B03_043951568	B03_999999999	fwd
B04	93	B04_017635956	B04_018939349	fwd
B04	94	B04_011931681	B04_012235066	fwd
B04	95	B04_002903707	B04_003179714	fwd
B04	96	B03_026922727	B03_027152665	fwd
B04	97	B04_003358314	B04_009557364	fwd
B04	98	B04_009557365	B04_011931680	rev
B04	98.5	B03_012929965	B03_013400579	fwd
B04	99	B04_000000001	B04_002903706	rev
B04	100	B04_015901752	B04_017375991	fwd
B04	101	B04_012235067	B04_015647477	fwd
B04	102	B04_020400645	B04_021339711	fwd
B04	103	B04_021339712	B04_021890388	rev
B04	104	B04_021890389	B04_022645012	fwd
B04	105	B03_008708175	B03_009392610	fwd
B04	106	B04_022645013	B04_027715112	fwd
B04	107	B04_027734631	B04_028054252	fwd
B04	108	B04_028775153	B04_031697322	fwd
B04	109	B01_030315193	B01_030411021	fwd
B04	110	B04_031697323	B04_031974884	fwd
B04	111	B04_032085267	B04_032392229	fwd
B04	112	B04_032392230	B04_032449791	rev
B04	113	B04_032449792	B04_999999999	fwd
B05	114	B01_029179897	B01_029564618	rev
B05	115	B05_000000001	B05_011253844	fwd
B05	116	B05_011253845	B05_019757533	fwd
B05	117	B05_019790036	B05_031539983	fwd
B05	118	B05_033059644	B05_037355020	fwd
B05	119	B05_038578987	B05_999999999	fwd
B05	120	B04_003179715	B04_003358313	fwd
B05	121	B04_017375992	B04_017635955	fwd
B06	122	B06_006252256	B06_010596211	fwd
B06	123	B06_000000001	B06_002493703	fwd

B06	124	B06_002720702	B06_005692482	fwd
B06	125	B06_010889574	B06_015082735	rev
B06	126	B07_009623277	B07_010823080	fwd
B06	127	B02_006907966	B02_007736852	rev
B06	128	B02_006692970	B02_006907965	fwd
B06	129	B08_003100631	B08_003440549	fwd
B06	130	B07_028680112	B07_029749605	fwd
B06	131	B06_015082736	B06_017792992	fwd
B06	132	B06_017875549	B06_026686061	fwd
B06	133	B06_027293965	B06_999999999	fwd
B07	134	B07_000000001	B07_004574022	fwd
B07	135	B06_026697745	B06_027053518	rev
B07	136	B07_004869052	B07_009623276	fwd
B07	137	B07_010823081	B07_012423923	fwd
B07	138	B07_012534506	B07_015544136	fwd
B07	139	B02_009122240	B02_009492750	fwd
B07	140	B07_015544137	B07_018021701	fwd
B07	141	B07_018298752	B07_018944746	fwd
B07	142	B07_018944747	B07_021173312	rev
B07	143	B07_022247983	B07_026120552	rev
B07	144	B07_026120553	B07_028680111	fwd
B07	145	B07_029749606	B07_041264391	fwd
B07	146	B08_030818428	B08_032495399	fwd
B07	147	B03_038201515	B03_038327908	fwd
B07	148	B07_041264392	B07_999999999	fwd
B08	149	B04_018939350	B04_020400644	rev
B08	150	B03_011730708	B03_012624232	fwd
B08	151	B05_032225572	B05_033059643	fwd
B08	152	B08_002743317	B08_003100630	rev
B08	153	B04_028054253	B04_028775152	fwd
B08	154	B08_003440550	B08_005633715	rev
B08	155	B08_014156055	B08_015272518	fwd
B08	156	B08_005633716	B08_014156054	fwd
B08	157	B08_000000001	B08_002743316	fwd
B08	157.5	B02_011874153	B02_013467423	fwd
B08	158	B08_015272519	B08_018602376	fwd
B08	159	B08_018802049	B08_021295848	fwd
B08	160	B05_031539984	B05_032225571	fwd
B08	161	B08_021295849	B08_024905158	fwd
B08	162	B08_025092391	B08_030117930	fwd
B08	163	B08_032871720	B08_032887280	fwd
B08	164	B08_030117931	B08_030818427	fwd
B08	165	B08_032887281	B08_037718028	fwd
B08	166	B08_037955294	B08_999999999	fwd