Plant Biotechnology Journal

Plant Biotechnology Journal (2014) **12**, pp. 778–786 doi: 10.1111/pbi.12182

A chromosomal genomics approach to assess and validate the *desi* and *kabuli* draft chickpea genome assemblies

Pradeep Ruperao[1,2,3], Chon-Kit Kenneth Chan[1], Sarwar Azam[3], Miroslava Karafiátová[4], Satomi Hayashi[1], Jana Čížková[4], Rachit K. Saxena[3], Hana Šimková[4], Chi Song[5], Jan Vrána[4], Annapurna Chitikineni[3], Paul Visendi[1,2], Pooran M. Gaur[3], Teresa Millán[6], Karam B. Singh[7,8], Bunyamin Taran[9], Jun Wang[5], Jacqueline Batley[1], Jaroslav Doležel[4,*], Rajeev K. Varshney[3,*] and David Edwards[1,2,*]

[1]*University of Queensland, St. Lucia, Queensland, Australia*
[2]*Australian Centre for Plant Functional Genomics, University of Queensland, St. Lucia, Queensland, Australia*
[3]*International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, Andhra Pradesh, India*
[4]*Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Olomouc-Holice, Czech Republic*
[5]*Beijing Genomics Institute (BGI), Shenzhen, China*
[6]*Department of Genetics, University of Cordoba, Cordoba, Spain*
[7]*The University of Western Australia Institute of Agriculture, The University of Western Australia, Crawley, Australia*
[8]*CSIRO Plant Industry, Private Bag 5, Wembley, WA, Australia*
[9]*Crop Development Centre, Plant Sciences, University of Saskatchewan,Saskatoon, Saskatchewan, Canada*

Received 2 October 2013;
revised 21 January 2014;
accepted 9 February 2014.
**Correspondence* (Tel +420 585 238 703, +91 4030713305 and +61 7 3346 7084; fax +420 585 238 704, +91 4030713071 and +61 7 3365 1176; emails dolezel@ueb.cas.cz; r.k.varshney@cgiar.org; Dave.Edwards@uq.edu.au)

Keywords: chickpea, genome assembly, cytogenetics, cicer.

Summary

With the expansion of next-generation sequencing technology and advanced bioinformatics, there has been a rapid growth of genome sequencing projects. However, while this technology enables the rapid and cost-effective assembly of draft genomes, the quality of these assemblies usually falls short of gold standard genome assemblies produced using the more traditional BAC by BAC and Sanger sequencing approaches. Assembly validation is often performed by the physical anchoring of genetically mapped markers, but this is prone to errors and the resolution is usually low, especially towards centromeric regions where recombination is limited. New approaches are required to validate reference genome assemblies. The ability to isolate individual chromosomes combined with next-generation sequencing permits the validation of genome assemblies at the chromosome level. We demonstrate this approach by the assessment of the recently published chickpea *kabuli* and *desi* genomes. While previous genetic analysis suggests that these genomes should be very similar, a comparison of their chromosome sizes and published assemblies highlights significant differences. Our chromosomal genomics analysis highlights short defined regions that appear to have been misassembled in the *kabuli* genome and identifies large-scale misassembly in the draft *desi* genome. The integration of chromosomal genomics tools within genome sequencing projects has the potential to significantly improve the construction and validation of genome assemblies. The approach could be applied both for new genome assemblies as well as published assemblies, and complements currently applied genome assembly strategies.

Introduction

Efforts to sequence and characterize crop genomes have been boosted in recent years by unprecedented developments in next-generation DNA sequencing (NGS). These technologies have dramatically reduced the cost of generating genome sequence data and present exciting new opportunities for crop genetics and breeding (Edwards and Batley, 2010; Varshney *et al.*, 2009). NGS technologies, currently dominated by the Illumina sequencing platforms, have seen a steady increase in read length, data quality and data quantity since their introduction less than a decade ago. The bioinformatics analysis of this data has been a challenge (Batley and Edwards, 2009); however, an increasing number of tools are now available to interrogate and analyse these data (Lai *et al.*, 2012b; Lee *et al.*, 2012; Marshall *et al.*, 2010).

One consequence of the growth of genome sequencing projects is a general decrease in accepted genome quality. The aim of any genome sequencing project should be to produce a genome that is fit for purpose, and often rough drafts are all that are required to answer important biological questions. Basic assemblies that produce the sequence of all genes, promoters and low copy or unique regions are relatively inexpensive and provide valuable biological insights, while more robust pseudomolecule assemblies have greater utility in the identification of gene variation underlying traits, and for use in genomics-assisted breeding (Duran *et al.*, 2010; Varshney *et al.*, 2005). However, the production of valid pseudomolecules representing individual chromosomes is the ultimate aim of many genome projects and remains a significant challenge, even in the age of NGS (Imelfort and Edwards, 2009).

Since the sequencing of the first plant genome, *Arabidopsis thaliana* (Arabidopsis_Genome_Initiative, 2000), and the first crop genome, rice (Yu *et al.*, 2002), genome sequencing methods have advanced significantly (Berkman *et al.*, 2012a; Edwards and Batley, 2010; Edwards and Wang, 2012; Edwards *et al.*, 2013).

Maize was the first large crop genome to be published (Schnable *et al.*, 2011), and maize genome resequencing has demonstrated a huge diversity in the genome structure between different varieties. Other less complex crop genomes have been sequenced, including the 1.1 Gbp soybean genome (Schmutz *et al.*, 2010) and the 844 Mbp autotetraploid genome of potato (Xu *et al.*, 2011). The soybean genome was sequenced using a whole-genome shotgun approach, while the relatively small potato genome was resolved by sequencing a homozygous doubled-monoploid potato clone using data from the Illumina and Roche 454 platforms.

Generating draft genome sequence assemblies of the simpler crop genomes, such as pigeonpea, are feasible and almost routine using whole-genome shotgun sequencing and Illumina sequencing technology (Varshney *et al.*, 2012). For complex genomes such as bread wheat, the complexity and size of the 17 Gbp genome, comprising three homoeologous subgenomes, necessitates alternative approaches to whole-genome *de novo* sequencing. These include the isolation of individual chromosome arms using flow cytometry and a two-stage sequencing approach which aims to initially generate draft shotgun assemblies of individual isolated chromosome arms (Berkman *et al.*, 2011, 2012b, 2013; Hernandez *et al.*, 2012), followed by the sequencing of BAC tiling paths representing each of these arms (Lai *et al.*, 2012a). The highly complex canola genome, which combines polyploidy with recent triplication in the diploid progenitors, presents a significant challenge for assembly. A public assembly of one diploid progenitor genome was published in 2011 (Wang *et al.*, 2011), while the second is near completion (http://www.brassica.info/). An initial draft genome for canola was produced in 2009, although this remains proprietary and efforts are currently underway to produce a public canola genome sequence (http://www.brassica.info/).

Chickpea (*Cicer arietinum*) is the second most important grain legume crop in the world, grown on about 12 million hectares in Asia, Latin America and Australia. This crop is represented by two main market types: large seeded *kabuli* and small seeded *desi*. These two types share a common ancestry, with *kabuli* evolving from *desi* in the Mediterranean basin, with subsequent selection for traits such as flower colour and seed tannins (Jana and Singh, 1993; Maesen, 1972; Moreno and Cubero, 1978). Genome assemblies have recently become available for both *kabuli* (Varshney *et al.*, 2013) and *desi* (Jain *et al.*, 2013) types. Surprisingly, these genome assemblies appear to be significantly different. To resolve these differences, we have developed and applied a chromosomal genomics approach for genome assembly validation. Using flow cytometry, we isolated individual chromosomes of chickpea for the generation of Illumina NGS sequence data. Mapping the resulting sequence reads from isolated *kabuli* and *desi* chickpea chromosomes to the reference genome assemblies allowed us to assess the quality of assembly of the two published genome sequences.

Results and discussion

Estimation of nuclear genome size

Knowledge of genome size is critical to estimate the quality of a genome sequence assembly. To estimate the genome size of both *desi* and *kabuli* chickpea types, we used DNA flow cytometry, which is currently considered the most reliable method (Doležel and Bartoš, 2005). This analysis revealed that chickpea has a medium-sized genome of less than 900 Mbp and that both types of chickpea do not differ significantly in genome size (Table 1). Our estimates are similar to the 1.9 pg DNA/2C (929 Mbp/1C) reported by Bennett and Smith (Bennett and Smith, 1976), greater than the kmer-based estimate of CDC Frontier (Varshney *et al.*, 2013), but significantly lower than the average 2C value of 3.41 pg DNA as predicted by Ohri and Pal (Ohri and Pal, 1991). This difference may be attributed to different methods (Feulgen microdensitometry was used in the older study) and to different reference standards (Doležel and Bartoš, 2005). Nevertheless, it is worth noting that Ohri and Pal (Ohri and Pal, 1991) did not observe significant differences in genome size between *kabuli* and *desi*.

Comparison of the published *kabuli* and *desi* chickpea draft genomes

An initial comparison of assembly statistics for the two draft chickpea genomes suggests differences in assembly quality. Both draft genomes were assembled from NGS data. The *kabuli* assembly was constructed mostly from Illumina data (Varshney *et al.*, 2013) supported by BAC-end sequences generated using Sanger-based methods, while the *desi* assembly applied a hybrid approach, combining Roche/454 and Illumina data. The *kabuli* assembly captured 532 Mbp (60.3% of the estimated genome size) in scaffolds greater than 1000 bp compared to 519 Mbp for *desi* (59.8% of the estimated genome size) in scaffolds greater than 200 bp. Thus, both assemblies represented similar genome fractions. However, the *desi* genome assembly was far more fragmented, with a total of 32 935 scaffolds greater than 1000 bp and an N50 of 106 Kbp, compared to 7163 scaffolds and an N50 of 39 989 Kbp for *kabuli* (Table 2). The method applied to place the scaffolds into pseudomolecules was similar for both genomes, although genotyping by sequencing (GBS) markers were included to validate the *kabuli* assembly.

Pairwise comparison of each of the pseudomolecules from the two assemblies revealed numerous structural variations (Figure 1). These differences include both long and short regions where the orientations of the sequence differed, for example the region from 9.33 Mb to 24.96 Mb on *kabuli* pseudomolecule Ca1 is inverted compared to the equivalent region on the *desi* assembly. There were differences in the position of regions within a pseudomolecule, for example the first half of *desi* pseudomolecule Ca5 is inverted and matches the centre of *kabuli* pseudomolecule Ca5. Of particular interest, we observed several large regions of similarity between unrelated pseudomolecules. These include *desi* pseudomolecule Ca8 matching a region at the start of *kabuli* pseudomolecule Ca7, while *kabuli* pseudomolecule Ca8

Table 1 Estimation of 2C DNA amounts and genome size in chickpea

Cultivar / Genotype	Type	2C DNA amount (pg)		Mean genome size (Mbp/1C)
		Mean	±SD	
ICC 1882	*desi*	1.773	0.012	867
ICC 283	*desi*	1.741	0.009	851
ICC 8261	*desi*	1.793	0.009	877
ICC 4958*	*desi*	1.775	0.008	868
CDC Frontier*	*kabuli*	1.803	0.007	882

*Genotypes used for chromosome sorting and sequencing.

Table 2 A comparison of *desi* and *kabuli* reference genome assembly statistics

Features	*desi* draft genome	*kabuli* draft genome
Total assembly size (Mb)	456 (52.5%*)	532 (60.3%**)
Number of scaffolds	32 935	7163
Minimum reported scaffold length (bp)	1000	1000
Maximum scaffold length (Kbp)	23 376	59 460
Average scaffold length (bp)	13 857	74 311
N50 length (Kb)	106	39 989
GC content (%)	25.6	30.8%
Genome captured in pseudomolecule (Kbp)	124 386 (14.33%*)	347 247 (39.37%**)
Protein coding genes	27 571	28 269
Average gene length (bp)	3122	3055
Average coding sequence length (bp)	962	1166

*Considering 1C = 868 Mbp (Table 1).
**Considering 1C = 882 Mbp (Table 1).

matches the last third of *desi* pseudomolecule Ca3. A large portion of *kabuli* pseudomolecule Ca6 matched the second half of *desi* pseudomolecule Ca2. These differences suggest misassembly of one or both draft genome assemblies.

Isolation and sequencing of chickpea chromosomes

To assess and validate the assembled pseudomolecules from the two genome assemblies, we isolated and sequenced individual chromosomes from both *kabuli* and *desi* varieties of chickpea and mapped the resulting sequence reads to the published reference assemblies. For shotgun sequencing, all chromosomes were flow sorted from the sequenced reference *kabuli* 'CDC Frontier', with chromosomes D and E sorted together as a group, while chromosomes A, B and H were flow sorted from the sequenced reference *desi* 'ICC 4958', (See Appendix 1 for details). DNA from these isolated chromosomes was amplified to produce samples suitable for sequencing using Illumina technology. All chromosome isolates could be sorted at high purity from both genotypes as determined by microscopic observation.

Estimation of molecular sizes of chickpea chromosomes

We estimated the molecular size of individual chromosomes based on relative chromosome lengths at mitotic metaphase. The results indicate differences in size between *desi* and *kabuli* chromosomes as large as 10 Mbp for chromosomes A and B and as small as several hundred Kbp for chromosome F (Table 3). Although the differences between the two types of chickpea may be ascribed in part to differences in chromatin condensation, they correspond well to differences between flow karyotypes of *desi* and *kabuli* and differences in chromosome peak positions (Figure 2). For example, chromosomes F and G of *desi* 'ICC 4958' differ by about 7 Mbp (7%), and their peaks cannot be discriminated based on flow karyotype. In *kabuli* 'CDC Frontier', the two chromosomes differ by about 10 Mbp (11%) and can be discriminated.

Comparison of pseudomolecule assemblies

A much greater portion of the *kabuli* assembly could be placed into pseudomolecules (347 247 Kbp) compared with *desi* (124 386 Kbp). The length of each of the pseudomolecules for *kabuli* was higher than for *desi,* and the pseudomolecules represented 39.37% and 14.33% of the estimated genome size in *kabuli* and *desi*, respectively (Table 2). Individual pseudomolecules differed in size and their representation of their predicted chromosome size (Table 4). Striking discrepancies were observed for *kabuli* chromosomes A, B and H, whose pseudomolecules represented on average only about 26% of their predicted size, compared to an average 50%. The smaller than expected pseudomolecule size of these three chromosomes could be explained by the presence of satellite CaRep2 on chromosomes A and B, satellite CaSat2 on chromosomes A and H, and the 45S rDNA locus on chromosome A (Zatloukalová *et al.*, 2011). These highly repetitive regions are likely to collapse into shorter representative regions during de Bruijn graph-based whole-genome assembly.

The *kabuli* reference contains short defined misassembled regions

Mapping each of the *kabuli* isolated chromosome sequence data sets to the *kabuli* reference genome assembly demonstrated that the majority of the reads matched to their respective pseudomolecule with the exception that chromosome F and G reads map to pseudomolecules Ca2 and Ca1, respectively, the inverse

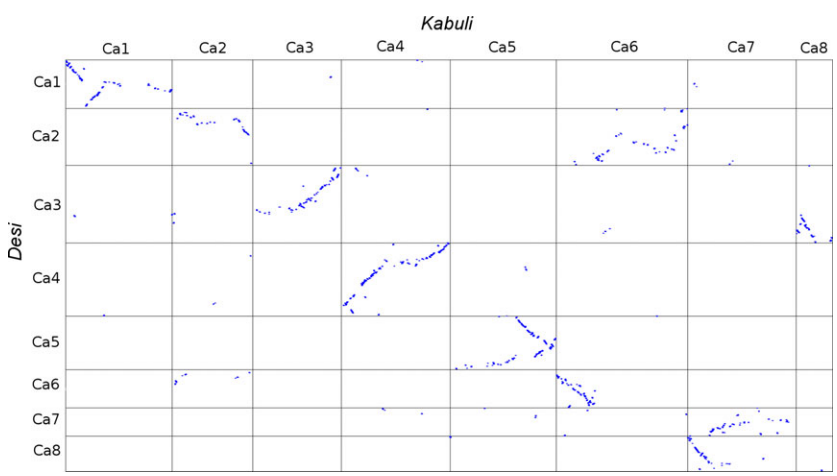

Figure 1 Dot plot matrix of a comparison of the *kabuli* and *desi* and *kabuli* draft chromosome assemblies.

Table 3 Chickpea *desi* and *kabuli* chromosome nomenclature, their assignment to peak on flow karyotypes, linkage groups, corresponding pseudomolecules and molecular chromosome sizes as determined cytologically

Cicer arietinum								
Desi 'ICC 4958'		*Kabuli* 'CDC Frontier'			Relative chromosome length [%]		Molecular chromosome size [Mbp]*	
Peak	Chromosome	Peak	Chromosome	Pseudomolecule	*Desi* '4958'	*Kabuli* 'Frontier'	*Desi* '4958'	*Kabuli* 'Frontier'
I	H	I	H	Ca8	7.2	7.8	62.5	68.80
II	G	II	G	Ca1	9.9	9.4	85.93	82.91
	F	III	F	Ca2	10.7	10.5	92.88	92.61
III	E	IV	E	Ca4	11.5	11.1	99.82	97.90
IV	D		D	Ca7	12.6	11.8	109.37	104.01
	C	V	C	Ca6	13.2	12.8	114.58	112.90
V	B	VI	B	Ca3	15.8	16.7	137.14	147.29
VI	A	VII	A	Ca5	19.0	19.8	164.92	174.64

*Calculated based on nuclear genome size and relative chromosome length.

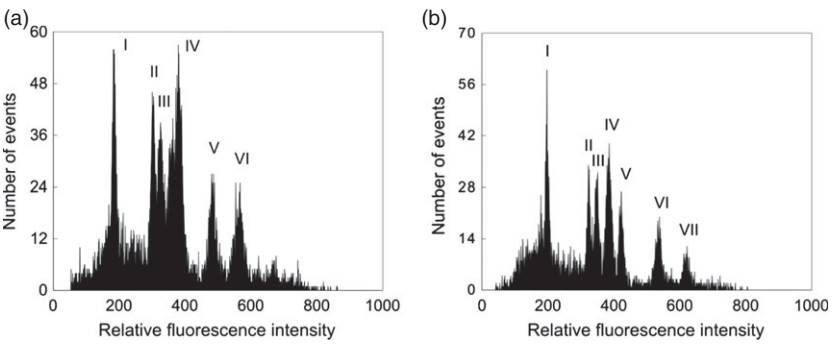

Figure 2 Histograms of relative fluorescence intensity obtained after flow cytometric analysis of DAPI-stained liquid suspensions of mitotic metaphase chromosomes prepared from chickpea *desi* 'ICC 4958' (a) and *kabuli* 'CDC Frontier' (b). For chromosome assignment of peaks on flow karyotypes, please see Table 3.

Table 4 Pseudomolecule size and percentage of predicted chromosome size

Pseudomolecule (chromosome)	*Desi* (%)	*Kabuli* (%)
Ca1 (G)	14 791 696 (15.9)	48 359 943 (52.2)
Ca2 (F)	17 304 114 (20.1)	36 634 854 (44.1)
Ca3 (B)	23 376 002 (17.0)	39 989 001 (27.1)
Ca4 (E)	22 093 647 (22.1)	49 191 682 (50.2)
Ca5 (A)	16 301 343 (9.8)	48 169 137 (27.5)
Ca6 (C)	11 482 212 (10)	59 463 898 (52.6)
Ca7 (D)	8 461 617 (7.7)	48 961 560 (47.0)
Ca8 (H)	10 574 966 (16.9)	16 477 302 (23.9)

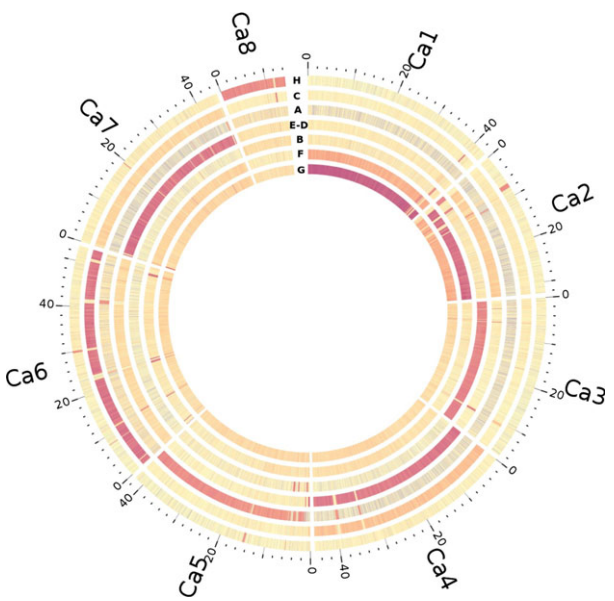

Figure 3 Circos heat map of the *kabuli* reference pseudomolecules demonstrating density of mapped Illumina paired sequence reads (red colour) from isolated *kabuli* chromosomes G, F, B, (E,D), A, C and H.

of the earlier assignments to genetic linkage experiments (Millan *et al.*, 2010; Thudi *et al.*, 2011; Zatloukalová *et al.*, 2011). Inspection of the read mapping density (Figure 3) suggested that chromosome F data included sequences specific for pseudomolecule G and *vice versa*. Chromosome C and the chromosome D/E group also shared contamination, while chromosomes A, B and H demonstrated a greater purity. The proportion of contamination of chromosome isolates with other chromosomes matched what was expected from the isolation method, with contamination between chromosomes from adjacent flow-sorted peaks.

In addition to the cross-mapping of reads due to chromosomal contamination, we observed regions in the reference pseudomolecules where few reads mapped from the respective chromosome sequence data (Figure 3). For example, a region from 40 141 642 to 40 436 753 bp on pseudomolecule Ca1 had very few reads mapping from the corresponding isolated chromosome G. Interestingly, this region had high mapped read coverage from

isolated chromosome C (Ca6). A similar pattern was observed for other gaps across the pseudomolecules and suggests that there are numerous small regions across the *kabuli* pseudomolecule assembly which were misplaced. In total, we observed 46 regions ranging in size from 57 to 1371 Kbp and representing 16 164 Kbp (3.0%) of the pseudomolecule assemblies that were placed into the wrong pseudomolecule (Table 5). Pseudomolecule Ca8 appears to be the most accurate assembly with only a single region of 341 Kbp which should be located on pseudomolecule Ca6 (Figure 4). In contrast, pseudomolecule Ca6 contains 11 blocks of sequence which should be relocated onto other pseudomolecules.

Some misassembled regions appeared to be contigs misplaced during the scaffolding process, while others appeared within contigs suggesting chimeric contig assembly. Many of the misassembled regions were also flanked by highly repetitive retrotransposon sequences, although there was no clear correlation between the presence of these sequences and the type of misassembly.

An advantage of applying chromosomal genomics approaches to identify genome misassembles is the exceptional resolution provided by NGS read mapping. This resolution will greatly facilitate the relocation of these regions into their correct pseudomolecule. One of the limitations of this approach, however, is the inability to identify intrachromosomal misassembly or misassembles between chromosomes which cannot be separated physically by flow sorting. In this case, chromosomes D and E could only be isolated as a pool, and while we identified several regions on these chromosomes which should be placed on other chromosomes, we could not identify chromosome E (Ca4) regions which were misplaced onto pseudomolecule Ca7 (D) and *vice versa*.

Table 5 Positions and sizes of misassembled genome blocks on the pseudomolecules of *kabuli* chickpea CDC Frontier, together with their correct chromosome (pseudomolecule) location

Pseudo-molecule	Start	End	Length (bp)	Chromosome (pseudomolecule)
Ca1	17 709 355	17 768 444	59 089	H (Ca8)
Ca1	39 419 875	39 639 265	219 390	F (Ca2)
Ca1	39 875 495	40 137 184	261 689	F (Ca2)
Ca1	40 141 642	40 436 753	295 111	C (Ca6)
Ca1	40 737 888	41 342 099	604 211	B (Ca3)
Ca2	1	1 370 632	1 370 631	B (Ca3)
Ca2	4 000 604	4 701 466	700 862	B (Ca3)
Ca2	5 875 970	5 981 305	105 335	G (Ca1)
Ca2	6 978 984	7 977 546	998 562	H (Ca8)
Ca2	8 329 465	8 839 803	510 338	D/E (Ca7/Ca4)
Ca2	9 713 165	10 056 176	343 011	A (Ca5)
Ca3	2 147 981	2 291 454	143 473	D/E (Ca7/Ca4)
Ca3	5 222 428	5 315 148	92 720	G (Ca1)
Ca3	5 817 530	5 985 956	168 426	D/E (Ca7/Ca4)
Ca3	16 652 391	16 736 790	84 399	C (Ca6)
Ca4	33 548 471	34 015 816	467 345	A (Ca5)
Ca4	39 258 943	40 092 009	833 066	A (Ca5)
Ca5	1	475 756	475 755	D/E (Ca7/Ca4)
Ca5	479 469	967 381	487 912	B (Ca3)
Ca5	1 056 592	1 302 990	246 398	D/E (Ca7/Ca4)
Ca5	1 443 073	1 575 871	132 798	B (Ca3)
Ca5	3 057 844	3 436 790	378 946	B (Ca3)
Ca5	4 300 510	4 897 203	596 693	B (Ca3)
Ca5	5 459 276	5 583 037	123 761	H (Ca8)
Ca5	14 570 063	14 984 575	414 512	H (Ca8)
Ca6	1	95 085	95 084	F (Ca2)
Ca6	1 273 532	1 471 783	198 251	F (Ca2)
Ca6	8 091 437	8 171 394	79 957	H (Ca8)
Ca6	10 834 481	11 048 871	214 390	F (Ca2)
Ca6	11 049 103	11 286 027	236 924	D/E (Ca7/Ca4)
Ca6	22 607 989	23 187 652	579 663	F (Ca2)
Ca6	23 191 730	23 605 994	414 264	A (Ca5)
Ca6	23 607 585	23 929 615	322 030	F (Ca2)
Ca6	29 686 973	30 164 801	477 828	H (Ca8)
Ca6	40 809 786	41 720 980	911 194	A (Ca5)
Ca6	50 526 024	51 189 107	663 083	F (Ca2)
Ca7	259 820	579 136	319 316	G (Ca1)
Ca7	10 581 321	10 638 294	56 973	G (Ca1)
Ca7	19 431 822	19 639 120	207 298	H (Ca8)
Ca7	24 424 554	24 528 692	104 138	F (Ca2)
Ca7	31 258 312	31 438 623	180 311	B (Ca3)
Ca7	34 418 700	34 570 247	151 547	A (Ca5)
Ca7	37 738 053	37 860 082	122 029	G (Ca1)
Ca7	37 864 185	37 939 907	75 722	A (Ca5)
Ca7	44 316 766	44 615 144	298 378	A (Ca5)
Ca8	11 734 308	12 075 396	341 088	C (Ca6)

Large-scale misassemblies in *desi* reference genome

To determine whether the differences between the two draft genome sequences reflect true structural genome variation or pseudomolecule misassembly, we isolated and sequenced chromosomes A, B and H from *desi* type chickpea and mapped these reads, together with the related *kabuli* chromosome-specific reads to the *desi* reference pseudomolecules (Figure 5) as well as the kabuli pseudomolecules (Figure S1). Sequence reads from both *desi* and *kabuli* isolated chromosomes demonstrated almost identical mapping patterns on the pseudomolecules suggesting that the physical genomes, at least for these three chromosomes, are highly similar between *desi* and *kabuli*. In contrast to the results from mapping *kabuli* chromosome reads to the *kabuli* pseudomolecules, we observed that the chromosome B (Ca3) reads from *kabuli* and *desi* only matched the first portion of *desi* pseudomolecule Ca3. Sequence reads from isolated chromosome H (Ca8) preferably mapped to the remaining portion of pseudomolecule Ca3 and not to pseudomolecule Ca8. This analysis suggested that the observed differences between the *desi* and *kabuli* reference genome assemblies are not due to structural genome differences but are due to misassembly of the *desi* reference genome.

Interestingly, there were regions of the *desi* reference pseudomolecules where no reads mapped. We investigated these regions further by mapping *desi* whole-genome sequence data to the *desi* pseudomolecules (Figure 5). Surprisingly, again no reads mapped to these regions. To assess whether these regions reflect highly rearranged misassembled chickpea sequence data, for example due to concatenation of reads from the Roche 454 sequencing platform used in the assembly of the draft *desi* genome, we remapped the Illumina *desi* whole-genome data and isolated chromosome data to the *desi* pseudomolecules at a low stringency. This again failed to produce specific read mapping, and we therefore concluded that these regions of the *desi* reference pseudomolecules do not reflect the physical content of the *desi* genome. Extraction of the sequence for these regions

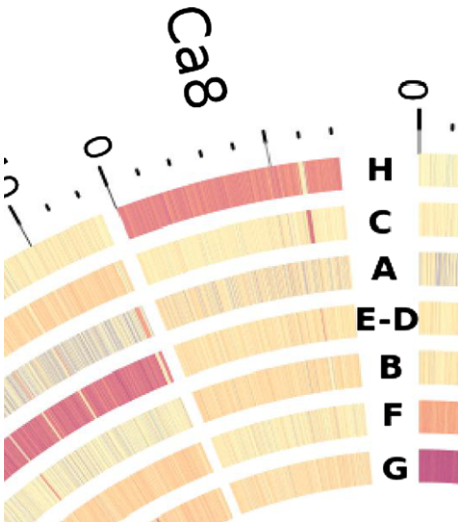

Figure 4 Circos heat map plot of the *kabuli* reference pseudomolecules demonstrating a high density of sequence reads (red colour) from *kabuli* chromosome H mapping to pseudomolecule Ca8. A small region on pseudomolecule Ca8 which lacks chromosome H reads is covered by chromosome C reads.

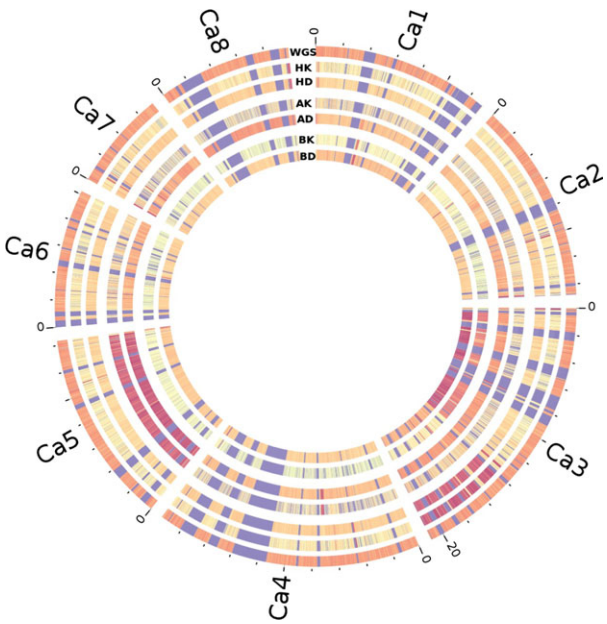

Figure 5 Circos heat map plot of the *desi* reference pseudomolecules demonstrating a high density of sequence reads (red colour) from *kabuli* and *desi* (D=*desi*, K=*kabuli*) chromosomes B, A, H and whole-genome sequence (WGS) reads of *desi*.

and comparison with the swissprot gene database failed to identify a significant number of genes (data not shown), again suggesting that these regions are not true genome sequences.

Conclusions

The expansion of genome sequencing projects and variable quality of published genomes highlights the need for additional approaches to validate and finish high-quality genome assemblies. We have established and assessed a chromosomal genomics approach to validate and compare reference genome assemblies. Overall, the assembly quality of the *kabuli* genome is high, with relatively few regions in the reference pseudomolecules which appear to have been misassembled into scaffolds on the wrong pseudomolecule. The high-resolution identification of these misplaced regions will aid their relocation on their correct pseudomolecule and the production of an improved reference genome assembly. Observed differences between the *kabuli* and *desi* published reference sequences contrast with our previous understanding of the similarity between the genomes. Our chromosomal genomics analysis suggests that the physical genomes of *kabuli* and *desi* chickpea types are in fact very similar and the observed differences in the sequence assemblies are due to major errors in the *desi* genome assembly, including the misplacement of whole chromosomes, portions of chromosomes and the inclusion of a large portion of sequence assembly which does not appear to be from the genome of chickpea. In addition to validating and assessing the genomes of chickpea, chromosomal genomics can be applied to validate and assist in the accurate assembly of other genome references where chromosomes can be isolated using flow sorting and thereby provide more robust genome assemblies that can provide a higher level of value for the many end-users of a particular genome assembly.

Experimental procedures

Estimation of genome size

Nuclear genome size was estimated using flow cytometry according to Doležel *et al.* (2007) (Doležel *et al.*, 2007). Approximately 30 mg of young chickpea leaf and 10 mg of leaf of soybean (*Glycine max* L. cv. Polanka, 2C = 2.5 pg DNA), which served as internal standard (Doležel *et al.*, 1994), were used for sample preparation. Suspensions of cell nuclei were prepared by simultaneous chopping of leaf tissues of chickpea and soybean in a glass Petri dish containing 500 μL Otto I solution (0.1 M citric acid, 0.5% v/v Tween 20). Crude homogenate was filtered through a 50-μm nylon mesh. Nuclei were then pelleted (300 g, 5 min) and resuspended in 300 μL Otto I solution. After 30-min incubation at room temperature, 900 μL Otto II solution (0.4 M Na_2HPO_4) (Otto, 1990) supplemented with 50 μg/mL RNase and 50 μg/mL propidium iodide was added. Samples were analysed using a Partec PAS flow cytometer (Partec GmbH, Münster, Germany) equipped with a 488-nm argon laser. At least 5000 nuclei were analysed per sample. Three individuals were analysed for each chickpea accession, and each individual was measured three times on three different days. Nuclear DNA content was then calculated from individual measurements following the formula: 2C nuclear DNA content [pg] = 2.5 × G_1 peak mean of chickpea / G_1 peak mean of soybean. Mean nuclear DNA content was then calculated for each plant. Genome size (1C value) was then determined considering 1 pg DNA is equal to 0.978×10^9 bp (Doležel *et al.*, 2003).

Molecular sizes of chickpea chromosomes

We determined the relative chromosome lengths in chickpea *desi* 'ICC 4958' and *kabuli* 'CDC Frontier'. Mitotic metaphase plates were prepared using synchronized root tip meristems (Vláčilová *et al.*, 2002). Root tips were fixed in 3:1 fixative (absolute ethanol: glacial acetic acid) for a week at 37°C and stained in 2% acetocarmine solution. Chromosome preparations were made according to Masoudi-Nejad *et al.* (Masoudi-Nejad *et al.*, 2002).

The preparations were counterstained with 4′,6-diamidino-2-phenylindole (DAPI) in Vectashield (Vector Laboratories, Burlingame) and observed under a fluorescence microscope (Olympus AX70, Tokyo, Japan). Chromosome lengths were estimated using the MicroImage software (Olympus) in 15 complete metaphase plates in each genotype, and average values were determined for each chromosome. Molecular chromosome sizes were determined considering relative chromosome lengths and 1C nuclear genome sizes as shown in Table 3.

Flow cytometric chromosome sorting and sequencing

Actively growing roots were used for cell cycle synchronization and preparation of liquid chromosome suspensions according to Vláčilová *et al.* (Vláčilová *et al.*, 2002). Chromosomes in suspension were stained with 2 μg/mL DAPI and sorted using a FACSAria flow cytometer (BD Biosciences, San José). The identification of the sorted chromosomes A and B was performed using fluorescent *in situ* hybridization (FISH) following the protocol of Vláčilová *et al.* (Vláčilová *et al.*, 2002), using tandem repeat probe CaSat1. The purity of the chromosome H fraction was determined based on chromosome morphology without a specific probe. The chromosomal fractions were sorted with the following purities: A: 93.75% (88.8%), B: 93.50% (91%) and H: 96% (92%) for *desi* (and *kabuli*), respectively. For whole-genome amplification, aliquots of 100 000–180 000 chromosomes (corresponding to ~20 ng DNA) were sorted into PCR tubes containing 10 μL of deionized water. Chromosomal DNA was purified as described in Šimková *et al.* (Šimková *et al.*, 2008) using increased proteinase K concentration (300 ng/μL). The purified DNA was amplified using the Illustra GenomiPhi V2 DNA amplification kit (GE Healthcare, New York).

A total of 1 μg of amplified DNA was used to prepare an Illumina TruSeq DNA HT library for each isolated chromosome, according to the manufacturer's instructions, and sequenced on the Illumina Hiseq2000 platform using standard protocols (Table S1). Chromosomes D and E from *kabuli* were isolated and sequenced as a group.

Desi and *kabuli* genome comparison

A pairwise comparison of all *desi* pseudomolecules with all *kabuli* pseudomolecules (Figure 1) was produced using the synteny block and anchor filtering algorithms in SyMap v4.0 (Soderlund *et al.*, 2011). SOAP2.21 was applied to map Illumina sequence data to the draft reference genome assemblies. For high-confidence mapping, only paired reads mapping uniquely to the reference was considered. For low stringency mapping, single and nonunique mappings were permitted. Circos v0.56 (Krzywinski *et al.*, 2009) was used to produce circular heatmaps using modified reference genomes with all 'N' nucleotides removed. Custom perl scripts soap2nc.pl and nc2circos.pl were used to convert SOAP output to Circos format. The boundaries of misassembled regions were determined manually by visual examination of the BAM file of mapped reads.

Acknowledgements

The authors would like to acknowledge funding support from the Australian Research Council (Projects LP0882095, LP0883462, LP110100200 and DP0985953), the Australian India Strategic Research Fund (AISRF) Grand Challenge fund (GCF010013), the Australian Grains Research and Development Corporation (UWA00151), CGIAR Generation Challenge Programme (Theme Leader Discretionary grant), Czech Science Foundation (P501/12/G090) and by the grant award LO1204 from the National Program of Sustainability I, the Australian Genome Research Facility (AGRF), the Queensland Cyber Infrastructure Foundation (QCIF) and the Australian Partnership for Advanced Computing (APAC) and the Center of Excellence in Genomics (CEG) of ICRISAT. The part of this work has been undertaken as part of the CGIAR Research Program on Grain Legumes. ICRISAT is a member of CGIAR Consortium. We thank our colleagues M. Kubaláková, J. Číhalíková, R. Šperková and Z. Dubská from IEB for assistance in chromosome sorting.

References

Arabidopsis_Genome_Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Batley, J. and Edwards, D. (2009) Genome sequence data: management, storage, and visualization. *Biotechniques*, **46**, 333–336.

Bennett, M.D. and Smith, J.B. (1976) Nuclear DNA amounts in angiosperms. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **274**, 228–274.

Berkman, P.J., Skarshewski, A., Lorenc, M.T., Lai, K., Duran, C., Ling, E.Y.S., Stiller, J., Smits, L., Imelfort, M., Manoli, S., McKenzie, M., Kubaláková, M., Šimková, H., Batley, J., Fleury, D., Doležel, J. and Edwards, D. (2011) Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotechnol. J.* **9**, 768–775.

Berkman, P.J., Lai, K., Lorenc, M.T. and Edwards, D. (2012a) Next generation sequencing applications for wheat crop improvement. *Am. J. Bot.* **99**, 365–371.

Berkman, P.J., Skarshewski, A., Manoli, S., Lorenc, M.T., Stiller, J., Smits, L., Lai, K., Campbell, E., Kubaláková, M., Šimková, H., Batley, J., Doležel, J., Hernandez, P. and Edwards, D. (2012b) Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theor. Appl. Genet.* **124**, 423–432.

Berkman, P.J., Visendi, P., Lee, H.C., Stiller, J., Manoli, S., Lorenc, M.T., Lai, K., Batley, J., Fleury, D., Šimková, H., Kubaláková, M., Weining, S., Doležel, J. and Edwards, D. (2013) Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnol. J.* **11**, 564–571.

Doležel, J. and Bartoš, J. (2005) Plant DNA flow cytometry and estimation of nuclear genome size. *Ann. Bot.* **95**, 99–110.

Doležel, J., Doleželová, M. and Novák, F.J. (1994) Flow cytometric estimation of nuclear-DNA amount in diploid bananas (*Musa acuminata* and *M. balbisana*). *Biol. Plant.* **36**, 351–357.

Doležel, J., Bartoš, J., Voglmayr, H. and Greilhuber, J. (2003) Nuclear DNA content and genome size of trout and human. *Cytometry A*, **51A**, 127–128.

Doležel, J., Kubaláková, M., Paux, E., Bartoš, J. and Feuillet, C. (2007) Chromosome-based genomics in the cereals. *Chromosome Res.* **15**, 51–66.

Duran, C., Eales, D., Marshall, D., Imelfort, M., Stiller, J., Berkman, P.J., Clark, T., McKenzie, M., Appleby, N., Batley, J., Basford, K. and Edwards, D. (2010) Future tools for association mapping in crop plants. *Genome*, **53**, 1017–1023.

Edwards, D. and Batley, J. (2010) Plant genome sequencing: applications for crop improvement. *Plant Biotechnol. J.* **7**, 1–8.

Edwards, D. and Wang, X. (2012) Genome Sequencing Initiatives. In: *Genetics, Genomics and Breeding of Oilseed Brassicas* (Edwards, D., Parkin, I.A.P. and Batley, J. eds), pp. 152–157. New Hampshire, USA: Science Publishers Inc.

Edwards, D., Batley, J. and Snowdon, R. (2013) Accessing complex crop genomes with next-generation sequencing. *Theor. Appl. Genet.* **126**, 1–11.

Hernandez, P., Martis, M., Dorado, G., Pfeifer, M., Gálvez, S., Schaaf, S., Jouve, N., Šimková, H., Valárik, M., Doležel, J. and Mayer, K.F.X. (2012) Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant J.* **69**, 377–386.

Imelfort, M. and Edwards, D. (2009) De novo sequencing of plant genomes using second-generation technologies. *Brief. Bioinform.* **10**, 609–618.

Jain, M., Misra, G., Patel, R.K., Priya, P., Jhanwar, S., Khan, A.W., Shah, N., Singh, V.K., Garg, R., Jeena, G., Yadav, M., Kant, C., Sharma, P., Yadav, G.,

Bhatia, S., Tyagi, A.K. and Chattopadhyay, D. (2013) A draft genome sequence of the pulse crop chickpea (*Cicer arietinum* L.). *Plant J.* **74**, 715–729.

Jana, S. and Singh, K.B. (1993) Evidence of geographical divergence in kabuli chickpea from germplasm evaluation data. *Crop Sci.* **33**, 626–632.

Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645.

Lai, K., Berkman, P.J., Lorenc, M.T., Duran, C., Smits, L., Manoli, S., Stiller, J. and Edwards, D. (2012a) WheatGenome.info: an integrated database and portal for wheat genome information. *Plant Cell Physiol.* **53**, 1–7.

Lai, K., Lorenc, M.T. and Edwards, D. (2012b) Genomic databases for crop improvement. *Agronomy*, **2**, 62–73.

Lee, H., Lai, K., Lorenc, M.T., Imelfort, M., Duran, C. and Edwards, D. (2012) Bioinformatics tools and databases for analysis of next generation sequence data. *Brief. Funct. Genomics*, **2**, 12–24.

Maesen, L.J.G.V.D. (1972) Cicer L., A monograph of the genus, with special reference to the chickpea (*Cicer arietinum* L.), its ecology and cultivation. *Maded. Landbouw. Wageningen*, **72**, 1–342.

Marshall, D., Hayward, A., Eales, D., Imelfort, M., Stiller, J., Berkman, P., Clark, T., McKenzie, M., Lai, K., Duran, C., Batley, J. and Edwards, D. (2010) Targeted identification of genomic regions using TAGdb. *Plant Methods*, **6**, 19.

Masoudi-Nejad, A., Nasuda, S., McIntosh, R.A. and Endo, T.R. (2002) Transfer of rye chromosome segments to wheat by a gametocidal system. *Chromosome Res.* **10**, 349–357.

Millan, T., Winter, P., Jungling, R., Gil, J., Rubio, J., Cho, S., Cobos, M.J., Iruela, M., Rajesh, P.N., Tekeoglu, M., Kahl, G. and Muehlbauer, F.J. (2010) A consensus genetic map of chickpea (*Cicer arietinum* L.) based on 10 mapping populations. *Euphytica*, **175**, 175–189.

Moreno, M.T. and Cubero, J.I. (1978) Variation in *Cicer arietinum* L. *Euphytica*, **27**, 465–485.

Ohri, D. and Pal, M. (1991) The origin of chickpea (*Cicer arietinum* L) - karotype and nuclear DNA amount. *Heredity*, **66**, 367–372.

Otto, F. (1990) DAPI staining of fixed cells for high-resolution flow cytometry of nuclear DNA. *Methods Cell Biol.* **33**, 105–110.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J.X., Mitros, T., Nelson, W., Hyten, D.L., Song, Q.J., Thelen, J.J., Cheng, J.L., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S.Q., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J.C., Tian, Z.X., Zhu, L.C., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C. and Jackson, S.A. (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Schnable, J.C., Springer, N.M. and Freeling, M. (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl Acad. Sci.* **108**, 4069–4074.

Šimková, H., Svensson, J.T., Condamine, P., Hřibová, E., Suchánková, P., Bhat, P.R., Bartoš, J., Šafář, J., Close, T.J. and Doležel, J. (2008) Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley. *BMC Genomics*, **9**, 237.

Soderlund, C., Bomhoff, M. and Nelson, W.M. (2011) SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res.* **39**, e68.

Thudi, M., Bohra, A., Nayak, S.N., Varghese, N., Shah, T.M., Penmetsa, R.V., Thirunavukkarasu, N., Gudipati, S., Gaur, P.M., Kulwal, P.L., Upadhyaya, H.D., KaviKishor, P.B., Winter, P., Kahl, G., Town, C.D., Kilian, A., Cook, D.R. and Varshney, R.K. (2011) Novel SSR Markers from BAC-End Sequences, DArT Arrays and a Comprehensive Genetic Map with 1,291 Marker Loci for Chickpea (*Cicer arietinum* L.). *PLoS ONE*, **6**, e27275.

Varshney, R.K., Graner, A. and Sorrells, M.E. (2005) Genomics-assisted breeding for crop improvement. *Trends Plant Sci.* **10**, 621–630.

Varshney, R.K., Nayak, S.N., May, G.D. and Jackson, S.A. (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* **27**, 522–530.

Varshney, R.K., Chen, W.B., Li, Y.P., Bharti, A.K., Saxena, R.K., Schlueter, J.A., Donoghue, M.T.A., Azam, S., Fan, G.Y., Whaley, A.M., Farmer, A.D., Sheridan, J., Iwata, A., Tuteja, R., Penmetsa, R.V., Wu, W., Upadhyaya, H.D., Yang, S.P., Shah, T., Saxena, K.B., Michael, T., McCombie, W.R., Yang, B.C., Zhang, G.Y., Yang, H.M., Wang, J., Spillane, C., Cook, D.R., May, G.D., Xu, X. and Jackson, S.A. (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**, 83–89.

Varshney, R.K., Song, C., Saxena, R.K., Azam, S., Yu, S., Sharpe, A.G., Cannon, S.B., Baek, J., Tar'an, B., Millan, T., Zhang, X., Rosen, B., Ramsay, L.D., Iwata, A., Wang, Y., Nelson, W., Farmer, A.D., Gaur, P.M., Soderlund, C., Penmetsa, R.V., Xu, C., Bharti, A.K., He, W., Winter, P., Zhao, S., Hane, J.K., Carrasquilla-Garcia, N., Condie, J.A., Upadhyaya, H.D., Luo, M., Singh, N.P., Lichtenzveig, J., Gali, K.K., Rubio, J., Nadarajan, N., Thudi, M., Doležel, J., Bansal, K.C., Xu, X., Edwards, D., Zhang, G., Kahl, G., Gil, J., Singh, K.B., Datta, S.K., Jackson, S.A., Wang, J. and Cook, D. (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.* **31**, 240–246.

Vláčilová, K., Ohri, D., Vrána, J., Číhalíková, J., Kubaláková, M., Kahl, G. and Doležel, J. (2002) Development of flow cytogenetics and physical genome mapping in chickpea (*Cicer arietinum* L.). *Chromosome Res.* **10**, 695–706.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.-H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Wang, J., Wang, X., Freeling, M., Pires, J.C., Paterson, A.H., Chalhoub, B., Wang, B., Hayward, A., Sharpe, A.G., Park, B.-S., Weisshaar, B., Liu, B., Li, B., Liu, B., Tong, C., Song, C., Duran, C., Peng, C., Geng, C., Koh, C., Lin, C., Edwards, D., Mu, D., Shen, D., Soumpourou, E., Li, F., Fraser, F., Conant, G., Lassalle, G., King, G.J., Bonnema, G., Tang, H., Wang, H., Belcram, H., Zhou, H., Hirakawa, H., Abe, H., Guo, H., Wang, H., Jin, H., Parkin, I.A.P., Batley, J., Kim, J.-S., Just, J., Li, J., Xu, J., Deng, J., Kim, J.A., Li, J., Yu, J., Meng, J., Wang, J., Min, J., Poulain, J., Hatakeyama, K., Wu, K., Wang, L., Fang, L., Trick, M., Links, M.G., Zhao, M., Jin, M., Ramchiary, N., Drou, N., Berkman, P.J., Cai, Q., Huang, Q., Li, R., Tabata, S., Cheng, S., Zhang, S., Zhang, S., Huang, S., Sato, S., Sun, S., Kwon, S.-J., Choi, S.-R., Lee, T.-H., Fan, W., Zhao, X., Tan, X., Xu, X., Wang, Y., Qiu, Y., Yin, Y., Li, Y., Du, Y., Liao, Y., Lim, Y., Narusaka, Y., Wang, Y., Wang, Z., Li, Z., Wang, Z., Xiong, Z. and Zhang, Z. (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **43**, 1035–1039.

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., Wang, J., Orjeda, G., Guzman, F., Torres, M., Lozano, R., Ponce, O., Martinez, D., De la Cruz, G., Chakrabarti, S.K., Patil, V.U., Skryabin, K.G., Kuznetsov, B.B., Ravin, N.V., Kolganova, T.V., Beletsky, A.V., Mardanov, A.V., Di Genova, A., Bolser, D.M., Martin, D.M.A., Li, G., Yang, Y., Kuang, H., Hu, Q., Xiong, X., Bishop, G.J., Sagredo, B., Mejia, N., Zagorski, W., Gromadka, R., Gawor, J., Szczesny, P., Huang, S., Zhang, Z., Liang, C., He, J., Li, Y., He, Y., Xu, J., Zhang, Y., Xie, B., Du, Y., Qu, D., Bonierbale, M., Ghislain, M., del Rosario Herrera, M., Giuliano, G., Pietrella, M., Perrotta, G., Facella, P., O'Brien, K., Feingold, S.E., Barreiro, L.E., Massa, G.A., Diambra, L., Whitty, B.R., Vaillancourt, B., Lin, H., Massa, A., Geoffroy, M., Lundback, S., DellaPenna, D., Buell, C.R., Sharma, S.K., Marshall, D.F., Waugh, R., Bryan, G.J., Destefanis, M., Nagy, I., Milbourne, D., Thomson, S.J., Fiers, M., Jacobs, J.M.E., Nielsen, K.L., Sonderkaer, M., Iovene, M., Torres, G.A., Jiang, J., Veilleux, R.E., Bachem, C.W.B., de Boer, J., Borm, T., Kloosterman, B., van Eck, H., Datema, E., Hekkert, B.T.L., Goverse, A., van Ham, R.C.H.J., Visser, R.G.F. and Potato Genome Sequencing, C. (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189–195.

Yu, J., Hu, S.N., Wang, J., Wong, G.K.S., Li, S.G., Liu, B., Deng, Y.J., Dai, L., Zhou, Y., Zhang, X.Q., Cao, M.L., Liu, J., Sun, J.D., Tang, J.B., Chen, Y.J., Huang, X.B., Lin, W., Ye, C., Tong, W., Cong, L.J., Geng, J.N., Han, Y.J., Li, L., Li, W., Hu, G.Q., Huang, X.G., Li, W.J., Li, J., Liu, Z.W., Li, L., Liu, J.P., Qi, Q.H., Liu, J.S., Li, L., Li, T., Wang, X.G., Lu, H., Wu, T.T., Zhu, M., Ni, P.X., Han, H., Dong, W., Ren, X.Y., Feng, X.L., Cui, P., Li, X.R., Wang, H., Xu, X., Zhai, W.X., Xu, Z., Zhang, J.S., He, S.J., Zhang, J.G., Xu, J.C., Zhang, K.L., Zheng, X.W., Dong, J.H., Zeng, W.Y., Tao, L., Ye, J., Tan, J., Ren, X.D., Chen, X.W., He, J., Liu, D.F., Tian, W., Tian, C.G., Xia, H.G., Bao, Q.Y., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W.M., Li, P., Chen, W., Wang, X.D., Zhang, Y., Hu, J.F., Wang, J., Liu, S., Yang, J., Zhang, G.Y., Xiong, Y.Q., Li, Z.J., Mao, L., Zhou, C.S., Zhu, Z., Chen, R.S., Hao, B.L., Zheng, W.M., Chen, S.Y., Guo, W., Li, G.J., Liu, S.Q., Tao, M., Wang, J., Zhu, L.H., Yuan, L.P. and Yang, H.M. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). *Science*, **296**, 79–92.

Zatloukalová, P., Hřibová, E., Kubaláková, M., Suchánková, P., Šimková, H., Adoracion, C., Kahl, G., Millan, T. and Doležel, J. (2011) Integration of genetic and physical maps of the chickpea (*Cicer arietinum* L.) genome using flow-sorted chromosomes. *Chromosome Res.* **19**, 729–739.

Supporting information

Additional Supporting information may be found in the online version of this article:

Figure S1 Circos heat map plot of the *kabuli* reference pseudomolecules demonstrating a high density of sequence reads (red colour) from *kabuli* and *desi* (D=*desi*, K=*kabuli*) chromosomes B, A and H.

Table S1 Chromosome sequence data generated.

Appendix S1 Detailed analysis of chickpea chromosome sorting.