the plant journal

SEB Society for Experimental Biology

The Plant Journal (2014) **77**, 430–442 doi: 10.1111/tpj.12370

RESOURCE

Population genomic analyses from low-coverage RAD-Seq data: a case study on the non-model cucurbit bottle gourd

Pei Xu[1,*], Shizhong Xu[2], Xiaohua Wu[1], Ye Tao[3], Baogen Wang[1], Sha Wang[1], Dehui Qin[4], Zhongfu Lu[1] and Guojing Li[1,*]
[1] *Institute of Vegetables, Zhejiang Academy of Agricultural Sciences, Hangzhou 310021, China,*
[2] *Department of Botany and Plant Sciences, University of California, Riverside, CA 92521-0124, USA,*
[3] *Majorbio Pharm Technology Co., Ltd., Shanghai 201203, China, and*
[4] *Institute of Horticulture, Zhejiang Academy of Agricultural Sciences, Hangzhou 310021, China*

Received 2 July 2013; revised 20 October 2013; accepted 22 October 2013; published online 28 October 2013.
*For correspondence (e-mail peixu@mail.zaas.ac.cn, Ligj@mail.zaas.ac.cn).

SUMMARY

Restriction site-associated DNA sequencing (RAD-Seq), a next-generation sequencing-based genome 'complexity reduction' protocol, has been useful in population genomics in species with a reference genome. However, the application of this protocol to natural populations of genomically underinvestigated species, particularly under low-to-medium sequencing depth, has not been well justified. In this study, a Bayesian method was developed for calling genotypes from an F_2 population of bottle gourd [*Lagenaria siceraria* (Mol.) Standl.] to construct a high-density genetic map. Low-depth genome shotgun sequencing allowed the assembly of scaffolds/contigs comprising approximately 50% of the estimated genome, of which 922 were anchored for identifying syntenic regions between species. RAD-Seq genotyping of a natural population comprising 80 accessions identified 3226 single nuclear polymorphisms (SNPs), based on which two sub-gene pools were suggested for association with fruit shape. The two sub-gene pools were moderately differentiated, as reflected by the Hudson's F_{ST} value of 0.14, and they represent regions on LG7 with strikingly elevated F_{ST} values. Seven-fold reduction in heterozygosity and two times increase in LD (r^2) were observed in the same region for the round-fruited sub-gene pool. Outlier test suggested the locus *LX3405* on LG7 to be a candidate site under selection. Comparative genomic analysis revealed that the cucumber genome region syntenic to the high F_{ST} island on LG7 harbors an ortholog of the tomato fruit shape gene *OVATE*. Our results point to a bright future of applying RAD-Seq to population genomic studies for non-model species even under low-to-medium sequencing efforts. The genomic resources provide valuable information for cucurbit genome research.

Keywords: RAD-Seq, population genomics, shotgun sequencing, *Lagenaria siceraria*, cucurbit, comparative genomics, fruit shape.

INTRODUCTION

With the increasing number of reference genomes available, whole genome re-sequencing (WGRS) has been used routinely in discovering DNA sequence variation, based on what historical mutation, drift and selection imprints can be inferred (e.g. Huang *et al.*, 2012; Jiao *et al.*, 2012). However, a considerable proportion of agricultural plants and animals still lack a reference genome and the situation is not expected to change in the foreseeable future. For these species, reduced-representation libraries sequencing (RRLS) or genotyping-by-sequencing (GBS) are considered to be the most promising technical alternatives (Altshuler *et al.*, 2000; Elshire *et al.*, 2011; Helyar *et al.*, 2011).

Restriction site-associated DNA sequencing (RAD-Seq), one of the strategies known as a genome 'complexity reduction' protocol (a form of GBS), has been proved to be particularly useful in non-model species as it combines the advantages of low cost and high throughput (Baird *et al.*, 2008; Rowe *et al.*, 2011). RAD-Seq has been applied to over 20 species without a reference genome in many aspects of genetic/genomic research, including efficient marker development (Scaglione *et al.*, 2012; Pegadaraju *et al.*, 2013;

Pujolar *et al.*, 2013), genetic and comparative map construction (Baxter *et al.*, 2011; Kakioka *et al.*, 2013; Yang *et al.*, 2013), high resolution gene/QTL mapping (Chutimanitsakun *et al.*, 2011; Pfender *et al.*, 2011; Hegarty *et al.*, 2013), phylogenetic/phylogeographic analyses (Emerson *et al.*, 2010; Rubin *et al.*, 2012; Nadeau *et al.*, 2013), genome-wide association studies (Hecht *et al.*, 2013), and conservation genomics (Ogden *et al.*, 2013). In species with available reference genomes or genomes of close relatives, RAD-Seq has been a cost-effective alternative to WGRS for population genomic scans (Hohenlohe *et al.*, 2010; Bruneaux *et al.*, 2013; Varshney *et al.*, 2013). In such practices, a sequenced genome serves not only as a reference for genotype determination from RAD-Seq data but also provides the physical positions of the markers for result interpretation in a genome context. Although intuitively RAD-Seq should work similarly in population genomics of non-model species, case studies are still badly needed to justify its applications. In fact, more technical issues exist in utilizing RAD-Seq for genotyping in a reference-independent manner, particularly under low-to-medium sequencing depth.

First, sequencing errors and heterozygosity present difficulties for genotyping in systems that have no reference genomes. As with other sequencing-by-synthesis techniques on the basis of next-generation sequencing (NGS), raw RAD-Seq data are mingled with sequencing errors (Baird *et al.*, 2008; Nielsen *et al.*, 2011), which can cause erroneous base-pair calls. On a related note, some species or populations (e.g. F_2) are highly heterozygous and thus would lead to additional difficulties in distinguishing true heterozygous alleles from sequencing errors as the risk of having sampled only one of the two alleles (assuming a diploid species) increased (Li *et al.*, 2009; Nielsen *et al.*, 2011). So far, reference-independent analyses of RAD-Seq data in non-inbred systems has been rarely reported in plants. What is more, the read depths in RAD-Seq can vary drastically across loci as well as individuals, and may cause uneven distribution of missing calls that tend to become more severe under limited sequencing depth (Davey *et al.*, 2012; Sharma *et al.*, 2012). This situation usually creates problems for estimating population genomic parameters, as a large number of highly informative (low missing rate) markers are required for this purpose. Thus far, it remains unclear to what extent these issues will affect the feasibility and power of population genomics based on *de novo* RAD-Seq data.

In previous studies, researchers have often adopted a 'fixed threshold' method for *de novo* genotype calling. For example, to discover SNPs from *Cynara cardunculus*, only RAD-tags with ≥5 reads were analyzed and a heterozygous genotype was called only if the proportion of the minor allele was between 25 and 50% (Scaglione *et al.*, 2012). In another example in which QTL mapping for stem rust resistance in *Lolium perenne* was reported, the read number threshold of RAD-tags was set at 8 (Pfender *et al.*, 2011). Apparently, this approach is easy to operate, but it is subjective and sacrifices all sequencing information for loci below the sequencing depth threshold. The optimal strategy should include a statistical control, as done with the maximum likelihood method (Hohenlohe *et al.*, 2010). This method assigns a genotype according to the likelihood ratio test between the two most likely hypotheses for each site, thereby overcoming the arbitrariness of the fixed threshold method and improving data usage.

Bottle gourd [*Lagenaria siceraria* (Mol.) Standl.] ($2n = 2x = 22$), also known as calabash, is an edible, medical, container and grafting stock plant cultivated all over the tropics (Heiser, 1979; Morimoto and Mvere, 2004). This crop belongs to the *Cucurbitaceae* family that contains many other agriculturally important species, including cucumber (*C. sativus* L.), melon (*C. melo* L.), squash (*Cucurbita* spp.) and watermelon (*Citrullus lanatus* (Thunb.) Matsum. & Nakai). Bottle gourd is considered to be one of the first crops to be domesticated (>10000 years ago) (Whitaker, 1971; Erickson *et al.*, 2005). It originated in Africa, with possible independent domestication in Asia (Heiser, 1979; Decker-Walters *et al.*, 2001; Erickson *et al.*, 2005). Amazingly, high genetic variability exists in bottle gourd, especially in fruit size and shape, which can be round, oblate, pyriform, elongated curvilinear, dipper, slender straight, tubby, snake-like and more (Figure 1). China has a long history in cultivating bottle gourd. However, it remains unknown how much the Chinese bottle gourd germplasm is divergent and how natural and human selection has affected its genetic composition. As an initial step towards answering these questions, we developed a set of SSR markers from analysis of a small portion of the bottle gourd genome. Based on 14 of these SSRs, fruit shape was found to be the main factor associated with genetic diversification (Xu *et al.*, 2011).

The objectives of the current study are: (1) to develop an optimized procedure for reference genome-independent genotype calling from RAD-Seq data under low-to-medium depth that is suitable for both homozygous and heterozygous systems; (2) to create genomic resources for bottle gourd/cucurbit research; (3) to gain knowledge of RAD (Restriction Site Associated DNA) marker reproducibility across populations under low-to-medium sequencing depth; and (4) on grounds of the above, to characterize population genomic features of the Chinese bottle gourd germplasm to help pinpoint genomic regions under selection during the process of crop improvement.

RESULTS

Sequencing depth and data quality

RAD sequencing of the F_2 population, including the parents, generated 334 167 874 reads with expected overhangs and barcodes. For the natural population (NP), the total read number generated was 209 743 655. Number of reads for

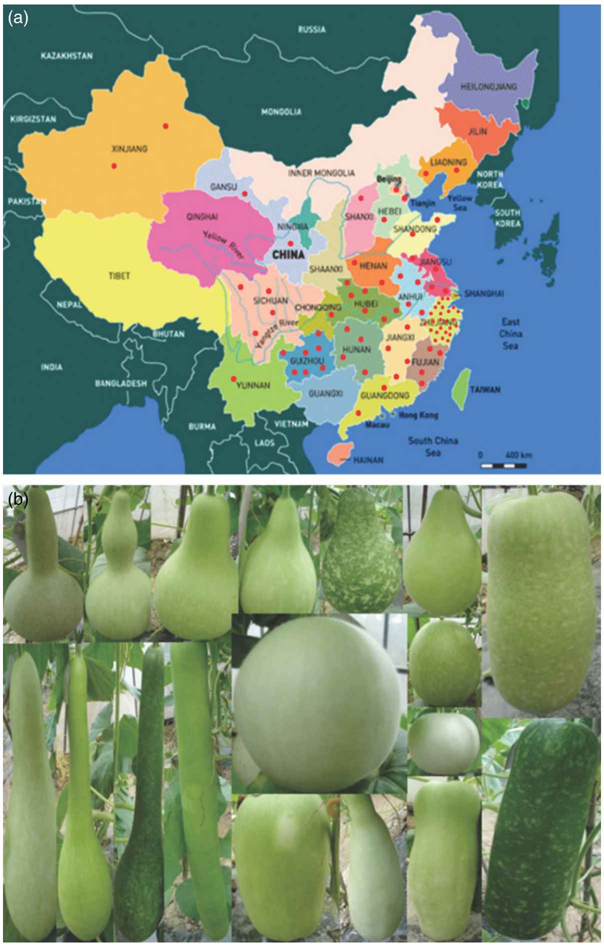

Figure 1. Plant materials used in the current study highlighting the different fruit shapes.
(a) Geographic distribution of the plant materials. Solid red dots indicate accessions that were collected from a certain region.
(b) Photos of representative fruit shape types of the plant materials.

each individual varied drastically, from 30 335 to 11 747 107. This variation was most likely due to uneven quantity and quality of the input DNA, but could also be explained biologically (Davey *et al*., 2012). Fifteen and 14 individuals with very low sequencing volume were removed, respectively, from the F_2 and the NP dataset for further analyses (Table S1). After grouping RAD reads into RAD-tags, the sequencing depth was found to vary significantly across loci, as has been noted in many other RAD-based studies (e.g. Pfender *et al*., 2011; Bus *et al*., 2012). The majority of the loci was at a coverage between 2×and 8× (Figure S1). The mean coverages of polymorphic loci in the two datasets were 5.2×and 4.9×, respectively. Therefore, the overall sequencing depth in the current study was at a lower medium level.

Genotype calling and comparison of the calling methods

A 'majority vote' approach, which assigns a genotype simply according to the base calls in majority, was used to call SNP genotypes between the parents of the F_2 population. With this approach, we discovered 3913 putative SNPs. We then re-examined genotypes of these loci in the F_2 population using a Bayesian method to reduce the SNP number down to 3753. After further filtering of loci showing high missing call rates (≥45%) or severely skewed segregations (χ^2 test, $P = 0.05$), a final set of 2098 SNPs was retained. Regarding the NP, where genotypes were called using solely the Bayesian method, 8778 and 3226 bi-allelic SNP loci with a minor allele frequency (MAF) ≥1.5% were determined under the missing call rate cutoff points of 50 and 20%, respectively. The set of 3226 SNPs was eventually used in the genetic diversity analysis.

A simulation experiment was conducted to compare the performance of the Bayesian method with previously published methods. The Bayesian method behaved similarly to the maximum likelihood method in assigning genotypes for loci sequenced at high depth (e.g. read number $n \geq 8$; Figure 2) in comparison with the fixed threshold method. The Bayesian and maximum likelihood methods can also assign genotypes to some loci at low read depth (e.g. $n \leq 4$) where sequencing information was either discarded (at high stringency) or flagged with high error (at low stringency) under the fixed threshold method, as indicated by the unusually high rate of skewed markers (>60%). The Bayesian method seemed to be more conservative in calling genotypes from loci with four or less reads than the maximum likelihood method, but the two gave comparable results from loci with five or more reads.

A high-density genetic map and the cross-species synteny

Regression model mapping with the 2098 high quality SNPs produced a genetic map (hereafter 'JJ' map) harboring 2084 loci with only 14 SNPs (0.7%) unmapped, which provides indirect evidence for the high reliability of the genotypes called. The mapped loci were distributed on 11 linkage groups presumably corresponding to the 11 chromosomes of the haploid genome of bottle gourd (Figure S2). The total length of the map was 1361 cM with an average marker spacing at a sub-cM level (0.6 cM). Individual linkage groups (hereafter, LGs) range from 75.5 cM to 190 cM in length and 0.3 cM to 1.4 cM in mean marker spacing (Table 1). There are several high-density blocks throughout the map, but no gaps are larger than 5 cM except for one on LG11. To more directly evaluate the reliability of the SNPs called from low-coverage reads and their impact on mapping, we also constructed a genetic map using only 784 SNPs at a high sequencing depth (read number ≥8 in ≥55% segregants). Comparison of the two maps revealed that the vast majority of the markers shared the same order/structure while only two major inversions were noted in LG1 and LG6. Besides these, there were only a few minor marker order rearrangements (Figure S3). As both inversion events occurred in a region with high mar-

Figure 2. Comparison of three different genotype calling methods.
The values of Y-axis represent the total number of loci that were assigned a discrete genotype across the F_2 population. Panels on the first four lines, results from using the fixed threshold method with the read depth cutoff varying from 3 to 10; panel on the bottom left, result from using the Bayesian method at the significance level of 0.05; panel on the bottom right, result from using the maximum likelihood method at the significance level 0.05.

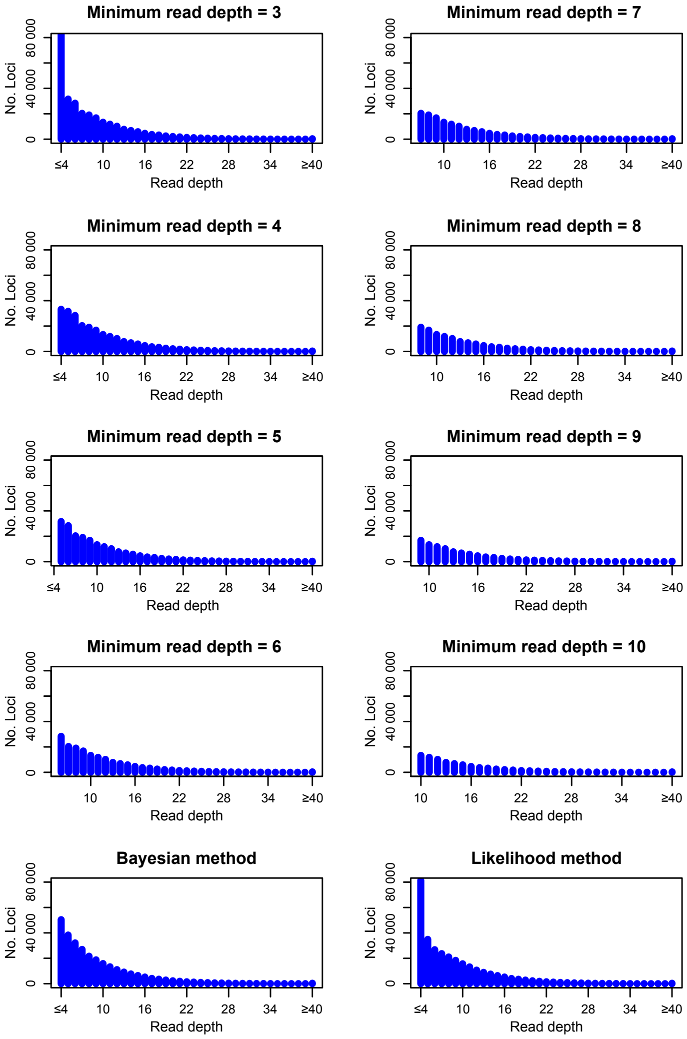

Table 1 Summary of the 11 linkage groups

LG	No. loci	Length (cM)	Mean marker space (cM)
1	320	108.3	0.3
2	300	177.5	0.6
3	235	149.7	0.6
4	210	102.2	0.5
5	178	190.1	1.1
6	176	100.2	0.6
7	165	75.5	0.5
8	162	111.9	0.7
9	139	115.8	0.8
10	115	115.7	1.0
11	84	114.5	1.4
Total	2084	1361.4	0.65

ker density, a phenomenon that is quite normal with changing marker density or flanking markers (Castiglioni *et al*., 1998; Zhang *et al*., 2012), the aforementioned results taken together support the high quality of the SNPs and hence the 'JJ' map.

To provide anchored genome scaffolds for constructing sequence-based syntenic map between bottle gourd and its relatives, a low-depth genome shotgun sequencing was conducted, which generated 5.7 Gb of DNA sequence data from 'Hangzhou gourd' after excluding bacteria/fungi contaminations, approximately 8.9× of the 640 Mb genome according to (Achigan-Dako *et al*., 2008). A total of 309 756 scaffolds longer than 200 bp were obtained after assembly (N_{50} = 325 bp), with a maximum length of 8164 bp (Figure S4). The total length of the scaffolds is ~306 Mb (298 Mb after subtracting 'N' calls). Gene model prediction hit 23 113 putative gene models, of which 15 426 were computationally annotated for hypothesized functions (Table S2). We then aligned the scaffolds/contigs to the genetic map by BLASTN. In total, 922 scaffolds/contigs were successfully anchored to the genetic map and were used to build the syntenic relationships across related cucurbit species (Figure 3, Data S1). Among the 922 scaffolds, 433, 460 and 725 have orthologs in the cucumber, melon and watermelon genomes, respectively, fitting the

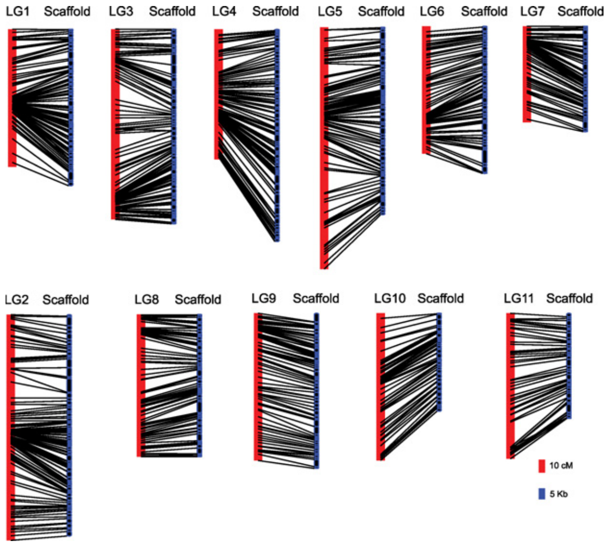

Figure 3. Genome scaffolds anchored onto the 11 linkage groups. The lengths of the bars are proportional to their genetic (LGs, in red) or physical (genome scaffolds, in blue) lengths.

known phylogeny of cucurbits (Renner *et al*., 2007). Macrocollinearity was apparent between bottle gourd and each of the three genomes (Figure 4). Each bottle gourd LG (LsLG) matched one to three chromosomes of cucumber and melon, and up to four chromosomes of watermelon (Data S1). LsLG8 and LsLG10 seemed to have undergone the least chromosome break/fusion events, as they are collinear to only a single or at most two chromosomes of the other genomes. In contrast, LsLG2 and LsLG6 had their syntenic segments in three or more chromosomes. We noted that even though watermelon is phylogenetically closest to bottle gourd and shares the same haploid chromosome number of 11, their syntenic relationship does not appear to be simple. No one-to-one chromosomal correspondence is present between the two genomes and both intra- and inter-chromosomal rearrangements have been observed.

Effective marker reproducibility between F_2 and the NP

Before conducting genetic diversity and population genomic analyses using the above-established resources, we examined the 'effective' reproducibility of SNP markers between the F_2 and the NP. The effective reproducibility is represented by the proportion of SNPs anchored to the reference genetic map that were also discovered in the NP with a high rate of successful genotype calls. We found that 49% (1023) of the SNPs mapped in the F_2 population were rediscovered in the NP at a missing call rate threshold of 0.7, and the reproducibility decreased as the allowed missing call rate decreased (Table 2). At the missing call rate cutoffs of 0.2 and 0.1, the rates of reproducibility dropped to 16% (325) and 8% (157), respectively. To balance marker number and data quality, the 325 SNPs with a maximum missing call rate of 0.2 were adopted in the following population genomic analyses. These markers were distributed on all of the 11 LGs, covering 1151 cM (85% of the total map length) and having an average marker spacing of 3.5 cM.

Population genetic and genomic analyses

We first investigated the genetic diversity among the natural collections using 3226 SNPs with ≥80% successful calls (anchored and unanchored). As an internal control of data quality, the genetic distances calculated between the two replicated samples ('Xiaogan gourd' and 'Early gourd 12', see Experimental procedures for details) based on the 3226 SNPs were found to be marginal (<0.005), which was mainly caused by asymmetric missing data. After excluding redundant data, the overall estimated heterozygosity of the population was 3.6%, consistent with the common nature of inbred plant materials (albeit naturally monoecious). The 66 accessions showed a low Nei's distance (GD) of 0.112 on average, with the greatest value (0.263) found between 'Super early' (NP37) and 'Flame gourd' (NP65). 'Super early' is a landrace from central China with slender straight fruit and 'Flame gourd' is a landrace from the southeast with a pear-shaped fruit. A dendrogram of hypothesized phylogeny generated via the neighbor-joining algorithm suggested two main branches of the 66 accessions (Figure 5), which is in agreement with previous findings that the bottle gourd germplasm was generally differentiated into two clusters by fruit shape rather than collection site (Yetisir *et al*., 2008; Xu *et al*., 2011). The larger branch consists of 46 accessions, most of which have slender straight or elongated tubby fruits; the smaller branch has 20 accessions that mostly have round or pyriform fruits. The two sub-gene pools were thereafter renamed as SubL referring to the elongated fruits and SubR referring to the rounded fruits, respectively. SubR is genetically more diverse than SubL, as reflected by the higher average (0.15) and maximum (0.26) Nei's distances than those of SubL (0.08 and 0.14, respectively).

Estimated with the aforementioned 325 mapped SNPs, the F_{ST} between SubL and SubR was 0.14 ($P < 0.001$), indicating a moderate population divergence (Table 3). Inspection of the nucleotide diversity and heterozygosity across the genome revealed variations among and fluctuations across chromosomes, which is expected. There is a decreased heterozygosity (seven-fold reduction) in SubR on LG7, which is a signal of selection (Figure 6). Congruently, F_{ST} on LG7 calculated with both the Hudson's method and the generalized linear model (GLM) method was exceptionally high compared with other chromosomes. A closer look revealed a small region spanning ca.

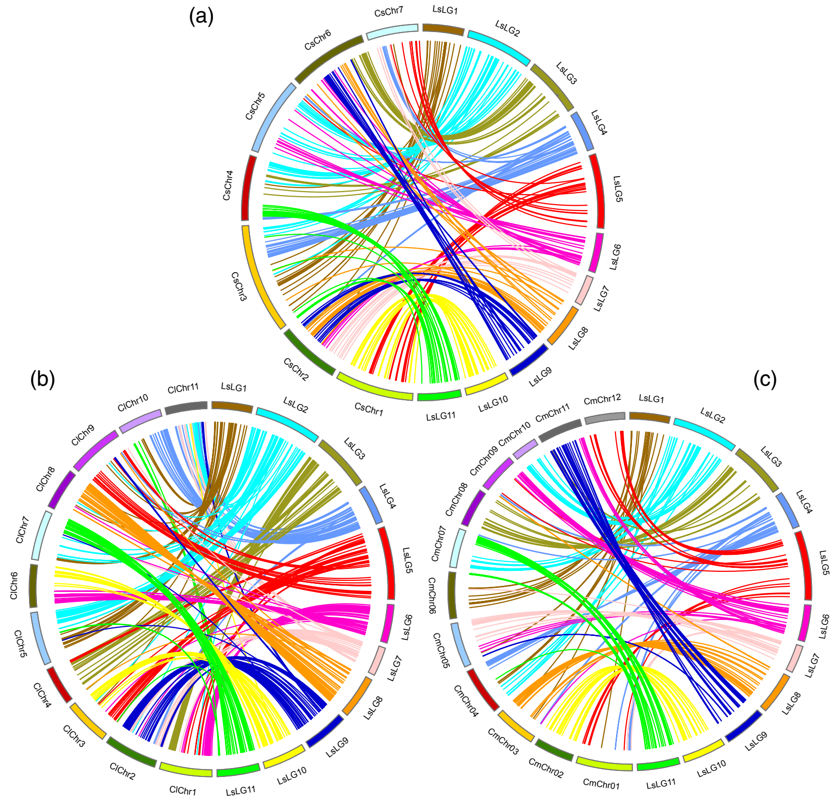

Figure 4. Circos illustration of the genome synteny between bottle gourd and its relatives.
(a) Bottle gourd – cucumber.
(b) Bottle gourd – watermelon.
(c) Bottle gourd - melon.
The bottle gourd linkage groups are denoted as LsLGs and the pseudomolecules of cucumber, melon and watermelon are represented as CsChrs, CmChrs and ClChrs, respectively.

Table 2 Marker reproducibility between the F_2 and the natural populations

	No. of SNPs		
Allowed missing call rate (%)	In NP	Shared BP	Shared and mapped
70	13 432	1324	1023
50	8778	963	870
20	3226	337	325
10	1545	163	157

NP, natural population; BP, between populations.

1.4 cM on LG7 that contributed the most to the high F_{ST} (Table 3 and Figure 6). Outlier test for SNP loci under selection suggested that *LX3405*, which resides in the high F_{ST} island, was under diversifying selection (F_{ST} = 0.89, $P < 0.05$) (Figure S5). In addition to LG7, there are also some regions that show elevated F_{ST} on other chromosomes (Table 3). We noted that the GLM method has the ability to narrow down the candidate intervals inferred to be influenced by selection, as were observed on LG5 and LG6.

The estimated intra-chromosomal LD levels (r^2) in the two sub-gene pools are 0.37 and 0.41, respectively. LG1, LG9 and LG11 showed fairly higher LD in SubS than in SubR, whereas LG2, LG5, LG6, LG7 and LG11 were just the opposite. LG7 held the most elevated r^2 in SubS, which is more than two times higher than that in SubR, supporting the effect of selection (Figures 7 and S6).

Syntenic relationship between the LG7 candidate selection sites and the tomato fruit shape gene orthologs

Because the sub-division of the NP is strongly associated with fruit shape, we wondered whether the chromosome region on LG7 presumably under selection is linked to any known fruit shape genes. The four cloned fruit shape genes (*SUN*, *OVATE*, *FAS* and *LC*) from tomato, a model in fruit shape research (Rodríguez *et al.*, 2011), were used for this purpose. Using the established bottle gourd-cucumber syntenic map as a 'bridge', we found that *SUN*, *FAS* and *LC* each was collinear to a region in the cucumber genome unrelated to LsLG7, and thus are not likely to be the candidate gene. *OVATE*, a gene that is known to control elongated fruit growth in tomato and pepper (Liu *et al.*, 2002; Tsaballa *et al.*, 2011), was found to reside in a region syntenic to the bottle gourd high F_{ST} island (8.7 Mb next to the ortholog of *LX2975*, which is 0.4 cM apart from the outlier site *LX3405*) (Figure 8). These findings provide interesting clues for investigation of the relationship between the known fruit shape genes and the determination of fruit shape in bottle gourd.

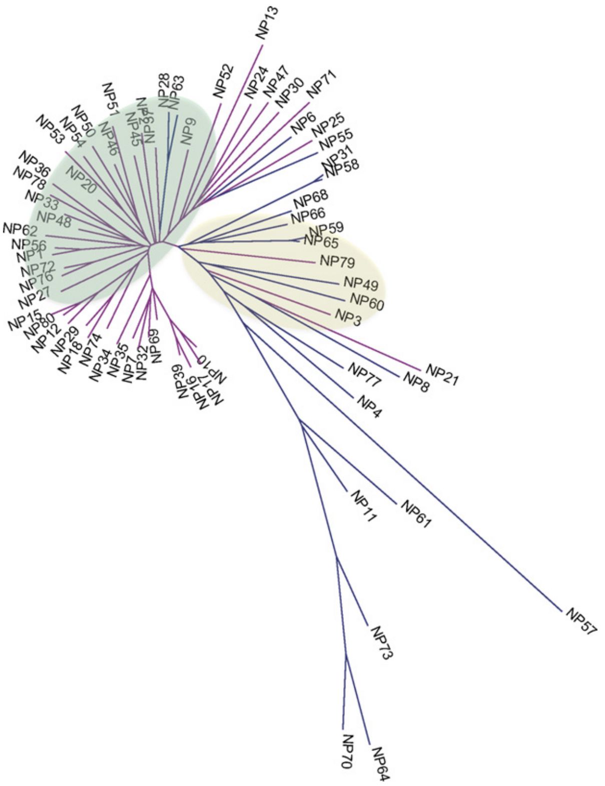

Figure 5. Neighbor-joining tree showing the genetic relatedness of the 66 bottle gourd accessions.
Lines in purple indicate plants with straight or tubby fruits while lines in blue indicate plants with round, pyriform or bulb-shaped plants. The branching of the two sub-gene pools was delineated by green (SubL) and yellow (SubR) shades, respectively.

Table 3 Regions or loci showing elevated F_{ST} as calculated by two different methods ($P < 0.01$)

	Hudson's method		GLM method	
LG	Interval or locus (cM)	Mean F_{ST}	Interval or locus (cM)	Mean F_{ST}
2	25.4	0.183	25.4	0.127
2	90.3	0.170	90.3	0.209
5	64.1–68.3	0.291	68.2	0.126
5	72.4	0.225	72.4	0.227
7	9.5–10.9	0.671	9.5–10.9	0.749
8	12.8–13.9	0.220	13.9	0.212
9	58.5	0.303	58.5	0.210

DISCUSSION

Strategies for effective population genomic analyses of non-model species from low-to-medium depth RAD-Seq data

Taking advantage of NGS to unravel population genomic features in non-model species is still in its infancy. Some researchers achieved the goal by assembling reference genomes at first, followed by WGRS, GBS or RAD-Seq; in this regard, bioinformatic pipelines for model species can be used in data processing (Wang *et al*., 2012; Varshney *et al*., 2013). This approach, however, is still prohibitively expensive or difficult (due to large genome sizes, high genomic complexity etc.) for the majority of non-model species. Thus, studies on *de novo* application of complexity-reduced techniques including RAD-Seq to population genomics are of high importance and urgency. The strategy presented here is straightforward. First, a high-density genetic map is constructed to enable genome scans at the population level. Secondly, the NP is genotyped via the same RAD-Seq protocol to identify regions that show signatures of selection. Similar strategies have been used in other plant species with different marker technologies. For example, an oligo pool assay (OPA) harboring 1536 SNPs for wheat and a 10 k SNPs array for sunflower were applied successfully to population genomic analyses recently. Some known QTLs have been verified and novel sites of selection have been discovered (Chao *et al*., 2010; Mandel *et al*., 2013).

Incorporating shotgun sequencing to allow for sequence-based comparative genomic analysis toward explaining the findings is another important aspect of the strategy we used. As shotgun sequencing is only done on a single selected accession and the genome synteny can be readily established via fairly low-coverage genome sequence data, the cost added to the total budget is minor (less than 1/10 in our case). This approach provided essential genomic resources and greatly facilitated candidate gene analyses. Benefiting from these two NGS-based and complementary approaches (RAD-Seq and genome shotgun), we were not only able to efficiently localize candidate regions under selection but also shed light on the putative candidate genes behind.

Important technical considerations for *de novo* population genomics from low-to-medium depth RAD-Seq data

Sequencing depth and quality control. Several reports have postulated that SNP genotypes can be accurately determined from RAD-Seq data without a reference genome (e.g. Pfender *et al*., 2011; Sharma *et al*., 2012). This conclusion, however, applies to situations under relatively high sequencing depths. An open question at present is how low a sequencing depth can be for accurate genotype calling. Based on our results and those from the pioneer maximum likelihood method (Hohenlohe *et al*., 2010), at least three and four reads, respectively, are required for calling genotypes from homozygous and heterozygous loci to ensure 95% confidence. In this regard, the sequencing depth in the current study (~5× on average) is just above the lower limit. Hence, more or less, we may have provided the most conserved results in using RAD-Seq. Nonetheless, increasing sequencing coverage is not always

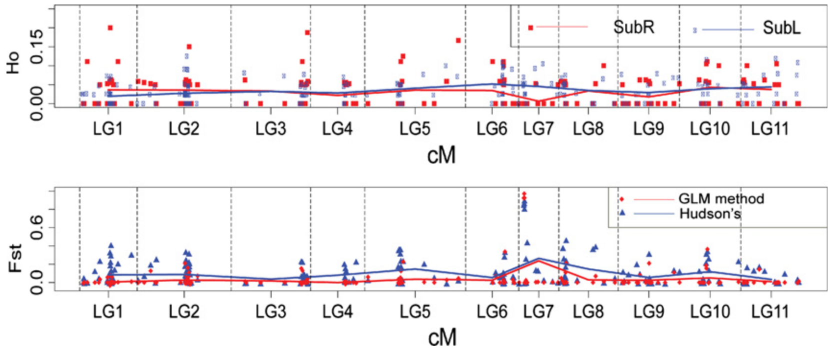

Figure 6. H_o and the F_{ST} calculated with two different methods and plotted against genetic map position.

(a) Observed heterozygosity (H_o) in SubL (in red) and SubR (in blue).

(b) Population differentiation index (F_{ST}) between the two sub-gene pools estimated with the Hudson's and the GLM methods.

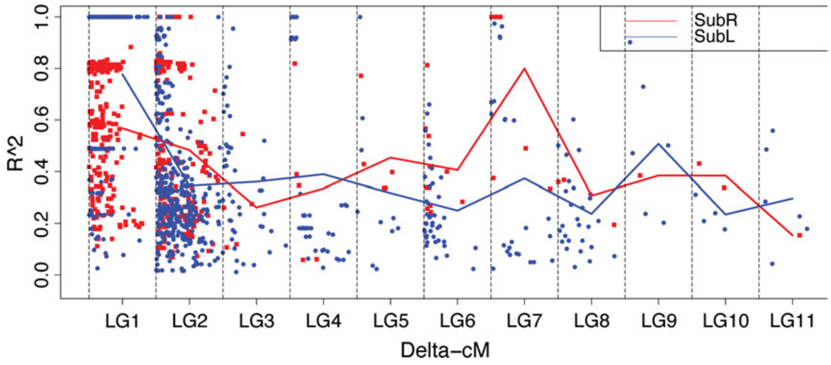

Figure 7. Comparison of the intra-chromosome LD level (r^2) between the two sub-groups along each linkage group.

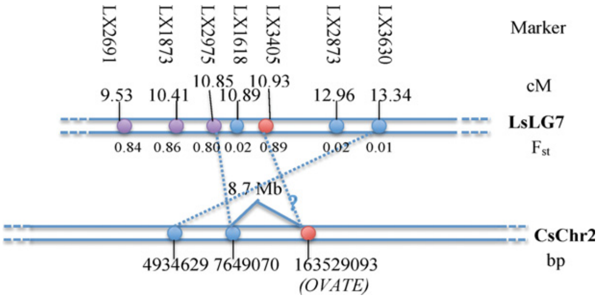

Figure 8. Illustration of syntenic relationship between *OVATE* and the high F_{ST} island on LG7. The candidate selection site *LX3405* by outlier test was denoted in red.

effective and necessary because the read depths at RAD loci are long known to be highly variable (Baird *et al.*, 2008; Davey *et al.*, 2012). Restriction fragment length bias has been a main factor accounting for this. We recommend an average sequencing depth of 6–8× for most *de novo* RAD-Seq-based studies to balance the cost and the quality of data. Even though, we still suggest including internal or external quality controls in experiments as RAD-Seq results can vary by species or restriction enzyme.

High missing call rate is the main factor reducing the resolution of population genome scan. The resolution of genetic map-based population genomic scan on a given population depends on two factors: the density of the genetic map and the availability of the mapped markers in the NP. RAD-Seq, due to the high coverage provided by current short-read sequencing technologies, has the merit in creating a high-density genetic map in most cases (Baxter *et al.*, 2011; Chutimanitsakun *et al.*, 2011; Pfender *et al.*, 2011; this study). Because only markers mapped in the segregating population and re-discoverable in the NP at high successful calls rates are considered fully useful, the metric 'effective reproducibility' is proposed to measure the acquisition of markers qualified for population genomic scan. We found that under a low (5×) average sequencing depth, a large portion of the reproducible SNPs would be filtered out from population genomic analyses because of high missing data rates, which in turn causes lower resolution than expected. Compared with other factors that can alter resolution such as restriction enzyme selection, the impact of missing calls is more common and profound. Nonetheless, the final 325 qualified SNPs dispersed along the 640 Mb bottle gourd genome still provided a reasonable marker density (equivalent to 1 SNP in every 1.9 Mb) as compared with non-NGS-based methods, and were able to identify several genome regions showing selective signals. Chao *et al.* (2010) identified several regions showing altered F_{ST} and co-located with domestication related QTLs in wheat using 849 informative SNPs. Given the huge genome size (16 000 Mb) of wheat, the marker density was only 1/10 (1 SNP in every 19 Mb) of ours. Therefore, even low-depth RAD-Seq can be very useful in dissecting population genomics of non-model species, as long as the SNP markers used are in general evenly distributed.

In many NGS-based studies on model species, missing data imputation has been a solution to high missing call rates (Huang *et al.*, 2012; Jiao *et al.*, 2012), but bioinformatics for reference-independent imputation is still in its infancy. Recently a rigorous evaluation of the suitability of different algorithms for reference-independent imputation of missing values from GBS data was reported (Rutkoski *et al.*, 2013). Algorithms such as random forest regression imputation (RFI) and k-nearest neighbors imputation (kNNI) were found to show high imputation accuracy which in turn led to greater accuracy in the simulation of genomic selection (GS), a method to use genome-wide markers to predict the genetic value of plants in breeding programs. Even though the usefulness of these methods for RAD-Seq data as well as their impact on population genomic parameters have not been justified, they hold great promises for gaining more information from less sequencing effort.

Problems to be solved. Bottle gourd is a diploid with a relatively small genome; however, the possibility of misassembly of RAD reads due to paralogous sequences still exists. Although actions like removing RAD-tags with >500 reads have been taken to reduce the risk of erroneous genotype calling, distinguishing misassembles from true assemblies without referring to prior genome sequence information is always difficult. Increasing read length or using paired-end sequencing may partially solve the problem (Willing *et al.*, 2011; Chong *et al.*, 2012). More recently, UNEAK, a network-based pipeline tailored for filtering out paralogs, repeats and error tags from typically 64 bp-long GBS reads on highly heterozygous species was developed (Lu *et al.*, 2013) and it can be useful for analyzing RAD-Seq data also.

Ascertainment bias can be problematic in population genomic studies using SNP markers (Helyar *et al.*, 2011). RAD-Seq, combining high throughput sequencing and large-scale genotyping in one step, has been shown to be, in principle, unbiased regarding many population genomic statistics (Baird *et al.*, 2008; Hohenlohe *et al.*, 2010). As the informative markers we used were from a sub-set of the total, the potential ascertainment bias remains to be tested in the future. In addition, the use of mapped SNPs from a single bi-parental population would have limited the rare SNPs and caused some biases in population genomic scan. Using SNPs mapped on a consensus genetic map that integrates multiple populations, if available, will be more representative.

Implications for cucurbits speciation

Genome comparison revealed a high level of sequence similarity between bottle gourd and watermelon; this result together with previous molecular evidence based on SSR transferability (Xu *et al.*, 2011) supports current hypotheses of the phylogeny of cultivated cucurbits. However, high similarity at the DNA sequence level does not necessarily mean high similarity at the chromosome structure level. On the contrary, our results showed even more complicated syntenic relationship between bottle gourd and watermelon, despite their common haploid chromosome number of 11. It has been suggested that today's 11 chromosomes of watermelon were from 7-chromosome ancestors through 21-chromosome paleohexaploid intermediates, during which at least 81 fissions and 91 fusions occurred (Guo *et al.*, 2013). We thereby postulated that such multi-step fission–fusion scenario might have happened differentially at multiple sites/stages before the speciation of the genus *Lagenaria* and *citrullus*, eventually resulting in high variability in chromosomal structure in spite of the DNA sequence conservation. Melon and cucumber might have undergone rather independent speciation episode that includes more recent, within-genus fission/fusion events (Garcia-Mas *et al.*, 2012). Indeed, the basal chromosome number of 7 in cucumber is hypothesized to be derived from a 12-chromosome progenitor (Trivedi and Roy, 1970; Singh, 1990; Li *et al.*, 2011).

Implications for bottle gourd improvement after domestication in China

According to Jeffrey (1980), bottle gourd was domesticated independently in Asia, Africa and the Americas. More recent evidence suggest that the bottle gourds in the Americas may be derived from one from Asia (Clarke *et al.*, 2006). The presence of bottle gourd in China and Japan as a domesticated species can be traced back to 7000 years ago (Chang, 1968; Imamura, 1996; Habu *et al.*, 2001). Compared with the African bottle gourd, many fruit types, e.g. 'UFO', 'gooseneck' and 'snake' are now rarely or no longer seen in China, suggesting a genetic bottleneck during domestication. Improvement of bottle gourd in China after domestication was historically toward two contrasting directions: for vegetable use and for household use (container, instrument, crafts). Generally, straight-fruited accessions are more favorable for vegetable use whereas round, oval or pear-shaped ones are more popular for household use. This bias may have caused differential gene flows into the sub-gene pools differing in fruit shape, as supported by our genetic dissimilarity analysis. The current notion that round-fruited trait is more native to the West African progenitors (Heiser, 1973) raises an alternative possibility that the round- and straight-fruited bottle gourds may have been separately improved, and introgressed with each other only recently.

Selection for fruit shape genes is also critical to subpopulation differentiation within cultivated tomato (Rodríguez *et al*., 2011). Among the four tomato genes that have been characterized for their functions in fruit shape control, *SUN* and *OVATE* are known to control elongated shape whereas *FAS* and *LC* control fruit locule number and flat shape (Liu *et al*., 2002; Cong *et al*., 2008; Xiao *et al*., 2008; Muños *et al*., 2011). From the perspective of functional specialty, the putative link of *OVATE* to the selection site in our bottle gourd population is reasonable and may suggest a conserved mechanism underlying fruit shape determination in plants. Nevertheless, we must acknowledge the fact that due to the indirect nature of candidate gene analysis (on the basis of bottle gourd-cucumber synteny) and the presence of a low F_{ST} locus in the high F_{ST} island on LG7, uncertainty still exists with regard to whether *OVATE* is the true selection target. Fruit shape is a complicated trait controlled by multiple genes. Its variation in bottle gourd is much higher than that in tomato and any other cucurbit species (Heiser, 1973). Therefore, the involvement of other loci in shaping the trait cannot be excluded.

CONCLUSION

De novo population genomic analyses from RAD-Seq data under even low-to-medium depth hold great promise for most plant and animal species that lack a reference genome and have a limited research budget. The proposed Bayesian method for genotype calling and the GLM method for F_{ST} calculation will be of broad interest for studies of this kind. The discovery of genome regions that show signals of selection, particularly those associated with fruit shape provides important insights into the impact of human selection on bottle gourd improvement. For cucurbit research, in general, the genomic resources released, together with previous information, will accelerate the characterization of many important traits common to the cucurbits, such as sex determination, high water conductance, metabolism of bitter cucurbit compounds and the vining growing habit.

EXPERIMENTAL PROCEDURES

Plant materials

Plant materials included a 139-individual F_2 population derived from the cross of 'Hangzhou gourd' (hereafter 'HZ') and 'J129' and a natural collection of bottle gourd germplasm totaling 80 accessions (Figure 1 and Table S3). HZ is a commercial cultivar in Southern China with slender strait fruit and J129 is a landrace with near round fruit. Accessions that comprised the NP were sampled from the germplasm repository of Zhejiang Academy of Agricultural Sciences (e-mail peixu@mail.zaas.ac.cn for germplasm queries) and all of them are inbred lines. Two accessions, 'Xiaogan gourd' and 'Early gourd 12', were sampled twice in order to provide internal control of data quality, leading to a total sample number of 82 from 80 accessions.

RAD sequencing, reads clustering and SNP calling

Genomic DNA was extracted from leaves of 2-week-old seedlings using a DNA extraction kit (Aidlab Co. Ltd, Beijing) following the manufacturer's instructions. DNA of the mapping parents of the F_2 population was prepared at two-fold higher concentrations and each was sampled twice with the expectation to achieve higher sequencing depths. RAD library construction, sample indexing and pooling followed Baird *et al.* (2008) for the F_2 and the NPs. Based on a DNA digestion experiment testing various restriction enzymes, *Eco*RI was selected to cut the DNA. In total, 11 multiplexed sequencing libraries were constructed, in which each DNA sample was assigned a unique nucleotide multiplex identifier (MID) for barcoding (Table S1). Single-end (101-bp) sequencing was performed using Illumina HiSeq2000 in a total throughput of four lanes. The raw sequence data have been deposited in the National Center for Biotechnology Information (NCBI) Sequence-Read Archive (SRA) database under the accession number SRA100255.

Raw sequence reads were trimmed to 85 nucleotides from the 3′ end that ensured more than 90% of the nucleotides have a quality value above Q30 (equals 0.1% sequencing error) and more than 99% above Q20 (equals 1% sequencing error). The trimmed reads were clustered into read tags (hereafter RAD-tags) by sequence similarity using *ustacks* (Catchen *et al*., 2011) to produce unique candidate alleles for each RAD locus. A maximum base-pair mismatch of two and only one was allowed in this step for the natural and the F_2 populations, respectively. However, in cases in which more than one SNP was detected, only those with two haplotypes were retained because of the expected high linkage disequilibrium in the short RAD reads. RAD-tags were then collapsed into clusters using *cstacks* under default parameters for SNP calling. RAD clusters with very high read depth (>500) were excluded from SNP calling.

Genotype calling and marker reproducibility analysis

For nomenclature consistency, genotype calling refers to the process of determining genotypes of SNP loci of each individual after SNP calling has been done, in accordance with Nielsen *et al.* (2011). As an optimized procedure for genotype calling, we used a combination of 'majority vote' (for the parents of F_2 only) and a new Bayesian method (for the rest of the lines) modified from the maximum likelihood method (Hohenlohe *et al*., 2010). Under the 'majority vote' mode, a genotype was assigned simply according to the base calls in majority. For instance, at any locus that is bi-allelic, more A calls than B would result in an AA genotype and vice versa, whereas an equal number of A and B would result in a heterozygous genotype call (very rare in inbred lines), irrespective of the total read depth. This method thereby could provide genotypic information for the maximum number of loci, some of which would otherwise be assigned a missing genotype in one or both parents. We then called genotypes from the F_2 offspring (theoretically 50% heterozygous) by using a strict Bayesian method and only SNPs that were consistently discovered in parents and the progenies were retained. In other words, this special procedure reduced the risk of discarding true SNPs due to missing calls in the parents if based solely on Bayesian calling without introducing additional errors. The Bayesian method followed the main body of the formula Hohenlohe *et al.* (2010) proposed in the maximum likelihood method:

$$P(\text{Hemo}) = \frac{n!}{n1!n2!n3!n4!}\left(1 - \frac{3\varepsilon}{4}\right)^{n1}\left(\frac{\varepsilon}{4}\right)^{n2+n3+n4}$$

$$P(\text{Hetero}) = \frac{n!}{n1!n2!n3!n4!}\left(\frac{1}{2} - \frac{\varepsilon}{4}\right)^{n1+n2}\left(\frac{\varepsilon}{4}\right)^{n3+n4}$$

where *P*(Homo) and *P*(Hetero) refer to the probability of a site to be homozygous or heterozygous, respectively; n_1, n_2, n_3, n_4 are the read counts for each of the four possible bases and ε is a sequencing error rate. Instead of calculating a site-specific ε as done in the maximum likelihood method, we calculated the distributions of *P*(Homo) and *P*(Hetero) for each site by testing ε varying from 0 to 0.5 incremented by 0.001. A discrete genotype was assigned according to the hypothesis with the highest probability exceeding 0.95; otherwise a missing genotype call was returned. The in-house Perl script for the Bayesian method can be accessed from Method S1.

To allow for comparison of the Bayesian method with existing methods, we also simulated the 'fixed threshold' method against the F_2 data by taking the read depth cutoff from 3 to 10. The criteria for determining a heterozygous or homozygous genotype are given below. If $1/3 \leq$ minor allele frequency $\leq 1/2$, then the allele is considered heterozygous; if minor allele frequency $\leq 1/10$ then the allele is considered homozygous; otherwise, a missing value will be assigned. Simulation of the maximum likelihood method followed the procedures described in Hohenlohe *et al.* (2010).

Cross-population reproducibility of RAD markers was estimated by calculating the proportion of SNPs that are common between the F_2 and the NPs. To identify common SNPs, a pairwise comparison was made for the polymorphic RAD-tags between the two populations by BLASTN with the criterion for a match being defined as ≥83 bp identical with no gap.

Linkage mapping

The genetic linkage map was constructed using JoinMap 4.0 (Kyazma, Wageningen, The Netherlands). A logarithm of the odds (LOD) score of 8.0 with a maximum recombination of 45% was used to assign markers to linkage groups. The LOD threshold for marker ordering and genetic distance calculation was 5.0. Other important mapping parameters included: Kosambi's mapping function, goodness-of-fit Jump threshold for removal of loci = 5.0, number of added loci after which to perform a ripple = 1 and third round = Yes. Only markers with ≥55% successful calls (equivalent to a minimum effective population size of 67 individuals at any locus) and that passed the $P = 0.05$ level of significance for the χ^2 test of segregation distortion were used for mapping.

Genome shotgun sequencing, assembly, gene model prediction and annotation

Shotgun sequencing was conducted as described elsewhere (e.g. Li *et al.*, 2010a,b). Briefly, two sequencing libraries with the insert size of 200 bp and 500 bp, respectively, were constructed and sequenced on the HiSeq2000 platform. The sequence reads were assembled into scaffolds using the computer program SOAPdenovo with a K-mer of 17 (Li *et al.*, 2010a). To predict coding sequences from the genome scaffolds, gene model parameters were first trained by AUGUSTUS with the cucumber genome data (Huang *et al.*, 2009) and the new trained dataset was applied in predicting bottle gourd gene models. The putative gene models were BLASTX searched against the NCBI NR database under an e-value significance cutoff of e^{-10} to assign functional annotations. The genome shotgun project has been deposited in GenBank under the accession ATBX00000000. The version described in this paper is version ATBX01000000.

Scaffold anchoring and genome comparative mapping

The genome scaffolds were anchored to the genetic map by aligning their sequences with the mapped RAD-tags using BLASTN. Criteria for match were 100% sequence overlap (with regard to RAD-tags) and no more than a 2-bp mismatch. To detect cross-species synteny, each anchored scaffold was BLASTN searched against the genome sequences of cucumber (Huang *et al.*, 2009), watermelon (Guo *et al.*, 2013) and melon (Garcia-Mas *et al.*, 2012) and the sequences were considered orthologous if sharing ≥90% sequence identity with an *e*-value $\leq e^{-10}$. In cases where multiple hits occurred, only the best hits were used. The software Circos (Krzywinski *et al.*, 2009) was employed to visualize the genome syntenic relationships.

Diversity, linkage disequilibrium (LD) and population genomic parameters calculation

Pairwise genetic distances were calculated using the software PowerMarker 3.25 under the Nei 1973 model (Liu and Muse, 2005). The neighbor-joining tree was generated by MEGA5 based on genetic distance data (Tamura *et al.*, 2011). Nucleotide diversity (π) and observed heterozygosity (H_o) were calculated using Arlequin 3.5.1.3 (Excoffier and Lischer, 2010) with the marker missing data rate cutoff of 0.2. Pairwise population differentiation index (F_{ST}) was calculated by two methods, i.e. the widely used Hudson's F_{ST} as implemented in Arlequin and a novel GLM approach as suggested in Xu (2012). The GLM approach was initially developed to detect Hardy–Weinberg disequilibrium with the advantage of reducing false positives caused by LD, but can also be used for calculating population differentiation. Candidate loci under selection were determined using an F_{ST} -based outlier test controlled for heterozygosity as implemented in Arlequin. LD as measured by the squared correlation coefficient (r^2) between each SNP pair was estimated by TASSEL 2.1 (Bradbury *et al.*, 2007). The *P*-values for each r^2 estimate were determined with the two-sided Fisher's exact test implemented in TASSEL. Only r^2 with $P < 0.05$ and between loci within 10 cM genetic distance were included in further analyses.

Accession numbers for sequence data

ATBX00000000, SRA100255.

ACKNOWLEDGEMENTS

This study was partially supported by the Natural Science Foundation of Zhejiang Province (No. Y3090146), the Key Science Project of Plant Breeding in Zhejiang (2012C12903) and Hangzhou (20122512A06), the National Natural Science Foundation of China (No. 30900985) and the 'Benteng' program for outstanding youth (to PX) funded by Zhejiang Academy of Agricultural Sciences. We thank Bao Lam Huynh for advices on genetic diversity analyses and Mark E. Sorrells for revision of the manuscript. The authors declare that they have no conflict of interest.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Frequency distribution of the RAD clusters varying in read depth.

Figure S2. A genetic map (the 'JJ' map) of bottle gourd based on SNP markers from RAD-Seq data.

Figure S3. Comparison of the 'JJ' map and the genetic map constructed using only the high depth SNPs.

Figure S4. Frequency distribution of the genome scaffolds varying in length.

Figure S5. Detection of outlier SNPs from genome scans based on F_{ST}.

Figure S6. Heat map of linkage disequilibrium on LG7 in the two subgene pools.

Data S1. The 922 genome scaffolds anchored onto the genetic map and their syntenic regions in the genomes of closely related species.

Table S1. Code, index sequence and total read number for each individual.

Table S2. Genome shotgun sequencing statistics summary.

Table S3. Accession or cultivar, origin, morphological characteristics and type of the plant materials used in the current study.

Methods S1. The Perl script of the Bayesian algorithm for genome type calling.

REFERENCES

Achigan-Dako, E.G., Fuchs, J., Ahanchede, A. and Blattnerv, F.R. (2008) Flow cytometric analysis in *Lagenaria siceraria* (*Cucurbitaceae*) indicates correlation of genome size with usage types and growing elevation. *Plant Syst. Evol.* **276**, 9–19.

Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L. and Lander, E.S. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. and Johnson, E.A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Baxter, S.W., Davey, J.W., Johnston, J.S., Shelton, A.M., Heckel, D.G., Jiggins, C.D. and Blaxter, M.L. (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE*, **6**, e19315.

Bradbury, P.J., Zhang, Z.W., Kroon, D.E., Casstevens, T.M., Ramdoss, Y. and Buckler, E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23**, 2633–2635.

Bruneaux, M., Johnston, S.E., Herczeg, G., Merilä, J., Primmer, C.R. and Vasemägi, A. (2013) Molecular evolutionary and population genomic analysis of the nine-spined stickleback using a modified restriction-site-associated DNA tag approach. *Mol. Ecol.* **22**, 5065–5082.

Bus, A., Hecht, J., Huettel, B., Reinhardt, R. and Stich, B. (2012) High-throughput polymorphism detection and genotyping in *Brassica napus* using next-generation RAD sequencing. *BMC Genomics*, **13**, 281.

Castiglioni, P., Pozzi, C., Heun, M., Terzi, V., Müller, K.J., Rohde, W. and Salamini, F. (1998) An AFLP-based procedure for the efficient mapping of mutations and DNA probes in barley. *Genetics*, **149**, 2039–2056.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W. and Postlethwait, J.H. (2011) Stacks: building and genotyping loci *de novo* from short-read sequences. *G3 (Bethesda)*, **1**, 171–182.

Chang, K.C. (1968) *The archaeology of ancient China*, 2nd edn. New Haven: Yale University Press.

Chao, S., Dubcovsky, J., Dvorak, J. *et al.* (2010) Population- and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (*Triticum aestivum* L.). *BMC Genomics*, **11**, 727.

Chong, Z., Ruan, J. and Wu, C.I. (2012) Rainbow: an integrated tool for efficient clustering and assembling RAD-seq reads. *Bioinformatics*, **28**, 2732–2737.

Chutimanitsakun, Y., Nipper, R.W., Cuesta-Marcos, A., Cistué, L., Corey, A., Filichkina, T., Johnson, E.A. and Hayes, P.M. (2011) Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. *BMC Genomics*, **12**, 4.

Clarke, A.C., Burtenshaw, M.K., McLenachan, P.A., Erickson, D.L., Penny, D. and SMBE Tri-National Young Investigators. (2006) Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. Reconstructing the origins and dispersal of the Polynesian bottle gourd (*Lagenaria siceraria*). *Mol. Biol. Evol.* **23**, 893–900.

Cong, B., Barrero, L.S. and Tanksley, S.D. (2008) Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication. *Nat. Genet.* **40**, 800–804.

Davey, J.W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K. and Blaxter, M.L. (2012) Special features of RAD Sequencing data: implications for genotyping. *Mol. Ecol.* **22**, 3151–3164.

Decker-Walters, D., Staub, J., Lopez-Sese, A. and Nakata, E. (2001) Diversity in landraces and cultivars of bottle gourd (*Lagenaria siceraria*; *Cucurbitaceae*) as assessed by random amplified polymorphic DNA. *Genet. Resour. Crop Evol.* **48**, 369–380.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Emerson, K.J., Merz, C.R., Catchen, J.M., Hohenlohe, P.A., Cresko, W.A., Bradshaw, W.E. and Holzapfel, C.M. (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proc. Natl Acad. Sci. USA* **107**, 16196–16200.

Erickson, D.L., Smith, B.D., Clarke, A.C., Sandweiss, D.H. and Tuross, N. (2005) An Asian origin for a 10, 000-year-old domesticated plant in the Americas. *Proc. Natl Acad. Sci. USA* **102**, 18315–18320.

Excoffier, L. and Lischer, H.E. (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Biol. Evol.* **10**, 564–567.

Garcia-Mas, J., Benjak, A., Sanseverino, W. *et al.* (2012) The genome of melon (*Cucumis melo* L.). *Proc. Natl Acad. Sci. USA* **109**, 11872–11877.

Guo, S., Zhang, J., Sun, H. *et al.* (2013) The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat. Genet.* **45**, 51–58.

Habu, J., Kim, M., Katayama, M. and Komiya, H. (2001) Jomon subsistence-settlement systems at the Sannai Maruyama site. In: *Indo-Pacific prehistory: the Melaka papers* (Bellwood, P., Bowdery, D., Glover, I., Hudson, M. and Keates, S., eds). Canberra: Australian National University, pp. 9–21.

Hecht, B.C., Campbell, N.R., Holecek, D.E. and Narum, S.R. (2013) Genome-wide association reveals genetic basis for the propensity to migrate in wild populations of rainbow and steelhead trout. *Mol. Ecol.* **22**, 3061–3076.

Hegarty, M., Yadav, R., Lee, M., Armstead, I., Sanderson, R., Scollan, N., Powell, W. and Skøt, L. (2013) Genotyping by RAD sequencing enables mapping of fatty acid composition traits in perennial ryegrass (*Lolium perenne* (L.)). *Plant Biotechnol. J.* **11**, 572–581.

Heiser, C.B. (1973) Variation in the bottle gourd. In: *Tropical forest ecosystems in Africa and South America: a comparative review* (Meggers, B.J., Ayensu, E.S. and Duckworth, W.D., eds). Washington DC: Smithsonian Institute Press, pp. 121–128.

Heiser, C.B. (1979) *The gourd book: A thorough and fascinating account of gourds from throughout the world*. Norman: University of Oklahoma Press, Oklahoma.

Helyar, S.J., Hemmer-Hansen, J., Bekkevold, D. *et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Mol Ecol Resour.* **1**, 123–136.

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A. and Cresko, W.A. (2010) Population genomics of parallel adaptation in three-spine stickleback using sequenced RAD tags. *PLoS Genet.* **6**, e1000862.

Huang, S.W., Li, R.Q., Zhang, Z.H. *et al.* (2009) The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* **41**, 1275–1281.

Huang, X., Kurata, N., Wei, X. *et al.* (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature*, **490**, 497–501.

Imamura, K. (1996) *Prehistoric Japan: new perspectives on insular East Asia*. London: UCL Press.

Jeffrey, C. (1980) A review of the *Cucurbitaceae. Bot. J. Linn. Soc.* **81**, 233–247.

Jiao, Y., Zhao, H., Ren, L. *et al.* (2012) Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* **44**, 812–815.

Kakioka, R., Kokita, T., Kumada, H., Watanabe, K. and Okuda, N. (2013) A RAD-based linkage map and comparative genomics in the gudgeons (genus *Gnathopogon, Cyprinidae*). *BMC Genomics*, **14**, 32.

Krzywinski, M., Schein, J., Birol, I., 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.

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K. and Wang, J. (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124–1132.

Li, R., Fan, W., Tian, G. *et al.* (2010a) The sequence and *de novo* assembly of the giant panda genome. *Nature*, **463**, 311–317.

Li, R., Zhu, H., Ruan, J. *et al.* (2010b) *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272.

Li, D., Cuevas, H.E., Yang, L. *et al.* (2011) Syntenic relationships between cucumber (*Cucumis sativus* L.) an melon (*C. melo* L.) chromosomes as revealed by comparative genetic mapping. *BMC Genomics*, **12**, 396.

Liu, K.J. and Muse, S.V. (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics*, **21**, 2128–2129.

Liu, J., Van Eck, J., Cong, B. and Tanksley, S.D. (2002) A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proc. Natl Acad. Sci. USA* **99**, 13302–13306.

Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., Casler, M.D., Buckler, E.S. and Costich, D.E. (2013) Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet.* **9**, e1003215.

Mandel, J.R., Nambeesan, S., Bowers, J.E., Marek, L.F., Ebert, D., Rieseberg, L.H., Knapp, S.J. and Burke, J.M. (2013) Association mapping and the genomic consequences of selection in sunflower. *PLoS Genet.* **9**, e1003378.

Morimoto, Y. and Mvere, B. (2004) *Lagenaria siceraria*. In: *Vegetables Plant Resources of Tropical Africa 2* (Grubben, G.J.H. and Denton, O.A., eds). Wageningen/Leiden: Backhuys Publishers/CTA, pp. 353–358.

Muños, S., Ranc, N., Botton, E. *et al.* (2011) Increase in tomato locule number is controlled by two single-nucleotide polymorphisms located near *WUSCHEL*. *Plant Physiol.* **156**, 2244–2254.

Nadeau, N.J., Martin, S.H., Kozak, K.M., Salazar, C., Dasmahapatra, K.K., Davey, J.W., Baxter, S.W., Blaxter, M.L., Mallet, J. and Jiggins, C.D. (2013) Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Mol. Ecol.* **22**, 814–826.

Nielsen, R., Paul, J.S., Albrechtsen, A. *et al.* (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443–451.

Ogden, R., Gharbi, K., Mugue, N. *et al.* (2013) Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. *Mol. Ecol.* **22**, 3112–3123.

Pegadaraju, V., Nipper, R., Hulke, B., Qi, L. and Schultz, Q. (2013) *De novo* sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach. *BMC Genomics*, **14**, 556.

Pfender, W.F., Saha, M.C., Johnson, E.A. and Slabaugh, M.B. (2011) Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*. *Theor. Appl. Genet.* **122**, 1467–1480.

Pujolar, J.M., Jacobsen, M.W., Frydenberg, J., Als, T.D., Larsen, P.F., Maes, G.E., Zane, L., Jian, J.B., Cheng, L. and Hansen, M.M. (2013) A resource of genome-wide single-nucleotide polymorphisms generated by RAD tag sequencing in the critically endangered European eel. *Mol. Ecol. Resour.* **13**, 706–714.

Renner, S.S., Schaefer, H. and Kocyan, A. (2007) Phylogenetics of *Cucumis* (*Cucurbitaceae*): cucumber (*C. sativus*) belongs in an Asian/Australian clade far from melon (*C. melo*). *BMC Evol. Biol.* **7**, 58.

Rodríguez, G.R., Muños, S., Anderson, C., Sim, S.C., Michel, A., Causse, M., Gardener, B.B., Francis, D. and vander Knaap, E. , (2011) Distribution of *SUN., OVATE., LC.*, and *FAS* in the tomato germplasm and the relationship to fruit shape diversity. *Plant Physiol.* **156**, 275–285.

Rowe, H.C., Renaut, S. and Guggisberg, A. (2011) RAD in the realm of next-generation sequencing technologies. *Mol. Ecol.* **20**, 3499–34502.

Rubin, B.E., Ree, R.H. and Moreau, C.S. (2012) Inferring phylogenies from RAD sequence data. *PLoS ONE*, **7**, e33394.

Rutkoski, J.E., Poland, J., Jannink, J.L. and Sorrells, M.E. (2013) Imputation of unordered markers and the impact on genomic selection accuracy. *G3 (Bethesda)*, **2**, 427–439.

Scaglione, D., Acquadro, A., Portis, E., Tirone, M., Knapp, S.J. and Lanteri, S. (2012) RAD tag sequencing as a source of SNP markers in *Cynara cardunculus* L. *BMC Genomics*, **13**, 3.

Sharma, R., Goossens, B., Kun-Rodrigues, C., Teixeira, T., Othman, N., Boone, J.Q., Jue, N.K., Obergfell, C., O'Neill, R.J. and Chikhi, L. (2012) Two different high throughput sequencing approaches identify thousands of *de novo* genomic markers for the genetically depleted Bornean elephant. *PLoS ONE*, **7**, e49533.

Singh, A.K. (1990) Cytogenetics and evolution in the *Cucurbitaceae*. In: *Biology and utilization of the Cucurbitaceae* (Bates, D.M., Robinson, R.W. and Jeffrey, C., eds). Ithaca, New York: Cornell University Press, pp. 10–28.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, Evolutionary Distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739.

Trivedi, R.N. and Roy, R.P. (1970) Cytological studies in *Cucumis* and *Citrullus*. *Cytologia*, **35**, 561–569.

Tsaballa, A., Pasentsis, K., Darzentas, N. and Tsaftaris, A.S. (2011) Multiple evidence for the role of an ovate-like gene in determining fruit shape in pepper. *BMC Plant Biol.* **11**, 46.

Varshney, R.K., Song, C., Saxena, R.K. *et al.* (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.* **31**, 240–246.

Wang, N., Thomson, M., Bodles, W.J., Crawford, R.M., Hunt, H.V., Featherstone, A.W., Pellicer, J. and Buggs, R.J. (2012) Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Mol. Ecol.* **22**, 3098–3111.

Whitaker, T.W. (1971) Endemism and pre-Columbian migration of bottle gourd, *Lagenaria siceraria* (Mol.) Standl. In *Man across the sea* (Riley, C.L., Kelley, J.C., Pennington, C.W. and Runds, R.L., eds). Austin: University of Texas Press, pp. 78–218.

Willing, E.M., Hoffmann, M., Klein, J.D., Weigel, D. and Dreyer, C. (2011) Paired-end RAD seq for *de novo* assembly and marker design without available reference. *Bioinformatics*, **27**, 2187–2193.

Xiao, H., Jiang, N., Schaffner, E.K., Stockinger, E.J. and vander Knaap, E. (2008) A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science*, **319**, 1527–1530.

Xu, S. (2012) Testing Hardy-Weinberg disequilibrium using the generalized linear model. *Genet. Res. (Camb)* **94**, 319–330.

Xu, P., Wu, X., Luo, J., Wang, B., Liu, Y., Ehlers, J.D., Wang, S., Lu, Z. and Li, G. (2011) Partial sequencing of the bottle gourd genome reveals markers useful for phylogenetic analysis and breeding. *BMC Genomics*, **12**, 467.

Yang, H., Tao, Y., Zheng, Z., Zhang, Q., Zhou, G., Sweetingham, M.W., Howieson, J.G. and Li, C. (2013) Draft genome sequence, and a sequence--defined genetic linkage map of the legume crop species *Lupinus angustifolius* L. *PLoS ONE*, **8**, e64799.

Yetisir, H., Sakar, M. and Serce, S. (2008) Collection and morphological characterization of *Lagenaria siceraria* germplasm from the Mediterranean region of Turkey. *Genet. Resour. Crop Evol.* **55**, 1257–1266.

Zhang, L., Luo, J.T., Hao, M. *et al.* (2012) Genetic map of Triticum turgidum based on a hexaploid wheat population without genetic recombination for D genome. *BMC Genet.* **13**, 69.