Divergent improvement of two cultivated allotetraploid cotton species

Lei Fang1,†, Ting Zhao1,†, Yan Hu1,†, Zhanfeng Si1,†, Xiefei Zhu2,†, Zegang Han1, Guizhen Liu2,3, Sen Wang2,4, Longzhen Ju2, Menglan Guo1, Huan Mei1, Luyao Wang1, Bowen Qi1, Heng Wang1, Xueying Guan1,† and Tianzhen Zhang1,†

1Zhejiang Provincial Key Laboratory of Crop Genetic Resources, Institute of Crop Science, Plant Precision Breeding Academy, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, China
2State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing, China
3Henan Province Seed Station, Zhengzhou, China
4Institute of Food Crops, Jiangsu Academy of Agricultural Sciences, Nanjing, China

Summary
Interspecific genomic variation can provide a genetic basis for local adaptation and domestication. A series of studies have presented its role of interspecific haplotypes and introgressions in adaptive traits, but few studies have addressed their role in improving agronomic character. Two allotetraploid Gossypium species, Gossypium barbadense (Gb) and G. hirsutum (Gh) originating from the Americas, are cultivated independently. Here, through sequencing and the comparison of one GWAS panel in 229 Gb accessions and two GWAS panels in 491 Gh accessions, we found that most associated loci or functional haplotypes for agronomic traits were highly divergent, representing the strong divergent improvement between Gb and Gh. Using a comprehensive interspecific haplotype map, we revealed that six interspecific introgressions from Gh to Gb were significantly associated with the phenotypic performance of Gb, which could explain 5%–40% of phenotypic variation in yield and fibre qualities. In addition, three introgressions overlapped with six associated loci in Gb, indicating that these introgression regions were under further selection and stabilized during improvement. A single interspecific introgression often possessed yield-increasing potential but decreased fibre qualities, or the opposite, making it difficult to simultaneously improve yield and fibre qualities. Our study not only has proved the importance of interspecific functional haplotypes or introgressions in the divergent improvement of Gb and Gh, but also supports their potential value in further human-mediated hybridization or precision breeding.

Keywords: Gossypium species, interspecific haplotypes, interspecific introgression, divergent improvement.

Introduction
Interspecific genomic variation can provide a genetic basis by which beneficial traits come into being. For crop species, cultivar inbreeding gradually excludes genomic variation due to purifying selection, resulting in a bottleneck. Genomic variation in closely related wild species or cultivars provides reservoirs that can be used as genetic sources in future breeding. The construction of interspecific introgressive populations is a customized breeding technique for transferring genetic information, such as in indica-japonica (Zhang et al., 2015b), maize-teosinte (Liu et al., 2016a; Liu et al., 2016b), G. barbadense (Gb) – G. hirsutum (Gh) (Nie et al., 2015; Shi et al., 2015; Si et al., 2017; Wang et al., 2013) and Glycine max-Glycine soja (Swarm et al., 2019). However, the effect of interspecific genomic haplotypes or introgression on phenotypic variation remains to be deeply studied.

Comparative genomics (multiple species) and population genomics (single species) are two strategies for studying the evolutionary and phenotypic impacts of genomic variation. Comparative genomics has the power to detect functional elements maintained over millions of years of evolutionary history, but is limited when detecting recently arisen functional loci that have not yet been lost or fixed (Lawrie and Petrov, 2014). Population genomics can identify genetic variation that occurs in a large number of individuals from the same species (Lawrie and Petrov, 2014), which can help to elucidate the evolutionary history of the species and mine the loci associated with elite traits. Extending these approaches, comparative population genomics is a method of combining polymorphism and divergence data, thereby improving power and precision in the study of genomic function and variation between two species. This strategy has also been employed to compare cultivars and their wild relatives in order to provide novel insights into the genetic origins of haplotypes or genomic regions using methods such as relative identity by descent (rIBD) (Wang et al., 2019), the D statistic (He et al., 2019; Wang et al., 2019) and fd (He et al., 2019). For instance, identity by descent (IBD) can be used to trace variation across genomes and populations (Thompson, 2013). Then, the relative proportions of haplotypes in the compared groups are expressed as rIBD, as has been used to detect introgression in wild and cultivated soybean populations (Wang et al., 2019), Tibetan cattle and yaks (Wu et al., 2018), and European and Asian pigs (Bosse et al., 2014).
The allopolyploid of the genus *Gossypium* occurred 1-1.5 million years ago due to the combination of two diploid genomes, the Old World A genome and the New World D genome (Wendel and Cronn, 2003; Zhang et al., 2015a). Among the seven extant allotetraploid species, the two most important cultivated allotetraploid cottons, *Gb* and *Gh*, diverged approximately one million years ago (Hu et al., 2019) and have the same origin, but were domesticated independently (Fang et al., 2017a). *Gb* has great advantages over *Gh* in terms of fibre quality and disease resistance, and its fibre is suitable for weaving high-grade fabric; together, these species provide an ideal model system for comparative population genomics. Although the biological impact of interspecific genomic diversity on growth and agronomic traits has been described previously (Fang et al., 2017a; Nie et al., 2020), the effective application of interspecific haplotypes or introgressions in mutual improvement of agronomic traits needs to be further evaluated. Here, we performed a genome-wide association study (GWAS) of 229 *Gb* accessions with genomic resequencing and corresponding phenotyping data sets, which were further compared with two GWAS panels from 491 *Gh* accessions. Then, an integrated haplotype map was constructed based on comparative population genomics for the identification of interspecific introgression. Our study provides a comparison of *Gb* and *Gh* improvement history, and also unlocks the potential value of interspecific haplotypes or introgressions for further breeding.

**Results and discussion**

**Genome-wide association analysis in *G. barbadense***

In terms of fibre quality, *Gb* has great advantages over *Gh* (Figure 1a, Tables S1 and S2). To identify genomic loci associated with yield and fibre quality in *Gb*, we performed a GWAS of 229 *Gb* accessions, including Egyptian, American Pima and Central Asia-type extra-long staple (ELS) cotton (Table S3). A total of 3.67 terabases (Tb) of clean data were obtained and then mapped to the reference genome of *Gb* cv. Hai7124 (Hu et al., 2019), yielding 4 476 574 high-quality biallelic single-nucleotide polymorphisms (SNPs) for further analysis. According to phylogenetic and principal component (PCA) analysis, all 229 *Gb* accessions could be classified into three groups (Figures S1 and S2). Group 1 (G1) mainly consisted of the Egyptian, American Pima and Central Asia-type accessions and their derived cultivars or lines. Group 2 (G2) contained the Central Asia-type cotton cultivars derived from founders introduced from the former Soviet Union, which played a significant role in modern *Gb* breeding in China. Group 3 (G3) integrated the Egyptian, American Pima-type and Soviet Union accessions, which may have crossed many times during the breeding process. Using phenotypes observed for the *Gb* population over four years of growth (Table S1), we performed a GWAS that identified 119 loci having strong associations (suggestive threshold of \( P < 9.27 \times 10^{-7} \) in the mixed model) with the yield components of lint yield (LY), seedcotton yield (SY), lint percentage (LP), and seed index (SI) and the fibre qualities of fibre elongation (FE), length (FL), fineness or micronaire (FM), strength (FS) and uniformity (FU). Among associated loci, 64 associated with yield traits, and more than 55 with fibre qualities (Figure 1b and Table S4).

**Divergent associated loci for improvement of *G. barbadense* and *G. hirsutum***

To reveal the genetic basis of improvement in *Gb* and *Gh*, we compared their functional haplotypes as determined from GWAS loci for agronomic traits. Using the same determination strategy based on the whole-genome resequencing of 234 *Gh* commercial cultivars (Table S5), we identified 111 loci associated with yield and fibre quality traits in *Gh* (Figure 1c and Table S6). Owing to the high-quality reference and broad synteny, we now have a much deeper understanding of lineage-specific associated loci from *Gb* and *Gh* accessions. Surprisingly, the 111 associated loci in *Gb* did not align with the 119 loci in *Gh*; loci were independent except for three on Chr. A08, D05 and D12 (Figure 1b-d and Table S7), as determined using BLAST (Altschul et al., 1997; Tables S4 and S6). Furthermore, although these three loci had partly overlapping associations, they are related to different yield and fibre quality traits (Figures S3-S5), indicating that the influence of genetic variation on improved traits altered during species formation and improvement. For instance, the locus A08Gb:106724585 in *Gb* associated with yield traits, but the corresponding locus A08Gh:112329779 in *Gh* related to fibre qualities such as FE, FL and FM (Figure 1e-f). Within that overlapped LD region, seven genes in *Gb* contained significant nonsynonymous SNPs and were expressed during the fibre development stage, but only one gene in *Gh* was identified as having a nonsynonymous SNP (Figure S6). In fact, no candidate gene within the A08 locus was shared between these two cultivated species, indicating that different functional haplotypes developed for different improved agronomic traits. The asynchronous improvement of *Gb* and *Gh* was further validated when we compared the GWAS panel in *Gb* with another GWAS panel of 257 *Gb* cultivars historically released worldwide in the course of crop improvement (Fang et al., 2017b); again, the associated loci were different in each species, except for two (Table S7). These results additionally provided an abundance of *Gb* and *Gh* interspecific associated loci or functional haplotypes, which are likely to be useful for further mutual improvement.

As a representative example, the locus A03Gb:4021204 (\( P < 1.28 \times 10^{-7} \)) in *Gb* showed significant association with FL, FS and uniformity (FU) in six environments, and overlapped with four quantitative trait loci (QTLs) for fibre qualities previously detected with chromosome segment substitution lines (CSSL) (Lacape et al., 2010; Nie et al., 2015; Shi et al., 2015; Si et al., 2017; Yu et al., 2014; Figure 2a). Within this locus, seven genes had nonsynonymous SNPs associated with fibre quality. Combining that information with gene expression during fibre development, two genes were identified as putative causal genes: the B3 domain-containing transcription repressor *GbVAL1* (GB_A03G0317) and CBL-interacting protein kinase *GbCIPK* (GB_A03G0324) (Figure 2b-c). The nonsynonymous SNPs A03Gb:3896896 (CC versus GG 138:75) in *GbVAL1* and A03Gb:3983696 (GG versus TT 138:72) in *GbCIPK* demonstrated significant associations with FL and FS, which were only present in *Gb* accessions (Figure 2d). The *Arabidopsis* orthologous gene of *GbVAL1* was identified as involved in the regulation of seed maturation, specifically initiating the switch from embryonic to postgerminative growth (Yang et al., 2013). Meanwhile, research on CIPKs in *Arabidopsis*, rice, maize and other plants over the past few years has testified to that gene’s function in the regulation of K+ homeostasis (Xu et al., 2006), which is important in fibre elongation (Ruan et al., 2001). However, how *GbVAL1* and *GbCIPK* regulate fibre quality improvement in *Gb* remains to be further explored.

In another example only present in *Gb*, the locus A05Gb:27770745 was responsible for effects on FE, FL and FS in seven environments (Figure 2e). One gene encoding the
MAPK/ERK kinase 1 (GH_A05G2318) was selected as the putative causal gene, as such kinases are the most important components of MAPK cascades and play integral roles in plant development and response to salt and cold stress (Figure 2f). On the contrary, the significant nonsynonymous SNP in GhMAPKK (AA vs TT 114:68) was only present in Gh accessions (Figure 2g-h). This locus overlapped with two QTLs mapped using Gh recombinant inbred lines (RILs) (Liu et al., 2018; Zhang et al., 2020) and with six...
Lei Fang et al.

(a) FL2015(Gb) 
Ga7_ML_3.1_5.47 
F2-3-qFL-c3-1 
LUT_UQLw_3.1_7.62 

(b) GBVAL1 
GbCIPK 
A03 (300 k) 
3.9 4.0 4.1 4.2 Mb 

(c) Fibre (DPA) 

(d) GbVAL1 
GbCIPK 

(e) FE2011XJ(Gh) 
qFL-A5-1 
qFL-A5-4 
Lan-C5 
qFL-Chr05-1 

(f) FS2012NY(Gh) 
qFS-A5-1 
BC1_STR_5_3_3 
q-STR-5-1 
q-FS-chrG-2 

(g) Fiber (DPA) 

(h) GHMAPKK 

© 2021 The Authors. Plant Biotechnology Journal published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 19, 1325–1336
To assess the potential applications of Gh–Gb CSSL populations: qFL-AS-1, qFL-AS-4, qFS-AS-1, BC1_Str_5_3_3, q-STR-5-1 and Len-CS (Lacape et al., 2013; Lacape et al., 2010; Si et al., 2017).

**Figure 2** Functional haplotypes in interspecific associated loci. (a) Manhattan plots for FL, FS and FU on Chromosome A03 in Gb. Interspecific QTLs LTL7_UQw_3_1_7.62, Ga7_Ma_3_1_5.47 and F2_3-qFL-c3-1 overlapped with the FS-associated QTL qFS3-1. (b) Genes with significant nonsynonymous SNPs in GWAS loci. (c) Transcriptomic levels of GbVAL1 and GbCIPK in Gb tissues based on FPKM values. (d) Boxplot for two haplotypes of GbCIPK and GbVAL1. Centre line, median; box limits, upper and lower quartiles; whiskers, 1.5 × the interquartile range; dots, outliers (**P < 0.01, two-tailed t-test). (e) Manhattan plots for FE, FL and FS on Chromosome A05 in Gh. Interspecific FL QTLs qFL-AS-1, qFL-AS-4 and Len-CS overlapped with FS QTLs qFS-AS-1, BC1_Str_5_3_3 and q-STR-5-1. (f) Genes with significant nonsynonymous SNPs in GWAS loci. (g) Transcriptomic levels of GmMAPK in Gh tissues based on FPKM values. (h) Boxplot for two haplotypes of GmMAPK. Centre line, median; box limits, upper and lower quartiles; whiskers, 1.5 × the interquartile range; dots, outliers (**P < 0.01, two-tailed t-test).

**Figure 3** Comprehensive whole-genome characterization of interspecific introgressions. To delineate the genomic characteristics of these interspecific introgressions, we first constructed a map integrating interspecific haplotypes from 229 Gb and 234 Gh accessions (Tables S3 and S5). A total of 4.54 Tb clean data were mapped to the reference genome of Gh (TM-1) (Hu et al., 2019), yielding 5 566 352 high-quality biallelic single-nucleotide polymorphisms (SNPs) for comparative genome polymorphisms analysis. Our analysis obtained 4.63 million pairwise IBDs (IBD length: median 4.7 Mb, mean 7.4 Mb) among the Gb and Gh accessions, covering an average of 80% per genome (Figure 3a). Most of the identified haplotypes were species-specific (37.80% in Gb and 57.92% in Gh), and only 198 148 IBDs (4.27%) were identified in both Gb and Gh populations. The rIBD was then defined as the normalized number of IBDs in Gb minus that in Gh (rIBD = nIBDgb−nIBDbg) (Figure 3b). A common haplotype is derived from Gb (Gb-HAP) when rIBD > 0, or derived from Gh when rIBD < 0 (Bosse et al., 2014; Wang et al., 2019; Wu et al., 2018). Only a small proportion of common IBD blocks were identified (Figure 3a), indicating a strong interspecific genomic divergence consistent with previous findings (Fang et al., 2017a; Hu et al., 2019); this divergence was further confirmed by phylogenetic analysis and principal component analysis (Figures S7 and S8). Using rIBD values, a comprehensive genome-wide introgression map was then constructed from individual Gb and Gh accessions (Figure 3c). Interestingly, all Gb accessions contained putative Gh-to-Gb introgressions (Gh-i), with genomic coverage ranging from 0.20% to 4.50% and averaging 1.40% (Figure 3d). Meanwhile, 74.60% of Gh accessions contained Gb-to-Gh introgressions (Gb-i), with genomic coverage ranging from 0.1% to 0.5% and averaging 0.19% (Figure 3d). The strong divergence of haplotypes between Gb and Gh can partly explain the asynchrony of associated loci and functional haplotypes between the species.

To delineate the genomic characteristics of these interspecific introgressions, we focused on twelve obvious Gh-i regions in the Gb genome, labelled Gh-i1 to Gh-i12; these ranged in length from 2.24 Mb to 46.28 Mb (average > 1 Mb), spanning about 120.83 Mb in all, and covered 6.55%-79.91% of Gh accessions (Figure 3e and Table S8). In these Gh-i regions, we observed a significant drop of Fixation index (Fst) values (0.20 versus 0.79, Wilcoxon test P < 2.2 × 10^-16) accompanied by an increase in genetic diversity (8.16 × 10^-4 versus 4.40 × 10^-4, Wilcoxon test P = 2.2 × 10^-16) relative to other regions (Figure 3f and g), which is consistent with a previous report in wheat (He et al., 2019). In addition, the linkage disequilibrium parameter r2 in these regions was significantly higher (0.65 versus 0.40, Wilcoxon test P < 2.2 × 10^-16; Figure 3h). On the contrary, the recombination was suppressed, which could introduce linkage-breaking mutations and produce different haplotypes. For instance, both Gh-i3 and Gh-i6 spanned centromeres, and those spans had not been disrupted in the course of genomic evolution and domestication. These Gh-i events may occur in Gb due to hybridization, introgression or incomplete lineage sorting of extant ancestral polymorphisms.

**Interspecific introgressions have large effects on agronomic traits.** Of the twelve interspecific Gh-i regions, six were significantly correlated with variation of fibre qualities and yield in Gb (Figure 4a and Table S9). Although Gh-i1, Gh-i2 and Gh-i3 were all located in a region of Chromosome A01 with an extremely low level of polymorphism between species (Hu et al., 2019), we for the first time revealed that only Gh-i3 is closely related to phenotype variation of Gb accessions. Namely, plants harbouring Gh-i3 showed significantly higher yield with increased LP (1.12%; t-test) and decreased SI (2.25%; t-test) (Figure 4b). Meanwhile, fibre qualities such as FL and FS were significantly decreased by 3.70% (t-test, P = 1.40 × 10^-12) and 9.71% (t-test, P = 2.60 × 10^-25), respectively (Figure 4a and Table S9).

As with Gh-i3, Gh-i5 also increased LP (3.18%; t-test P = 0.00028), but decreased FL (7.27%; t-test P = 5.90 × 10^-28), FS (14.10%; t-test P = 2.48 × 10^-42), and SI (5.60%; t-test, P = 2.40 × 10^-12) (Figure 4a). Interestingly, Gh-i5 largely occurred in 44 (96%) members of G1, which group also showed increased LP and LY relative to other groups, but lower fibre qualities, including FL, FS and FU (Figure 4a and c). Within G1, Gh-i5 occurred in 25 XLD, three Pima, and two Egyptian Giza cultivars and so on. Interestingly, other accessions derived from founders introduced from the Soviet Union did not contain this haplotype, including 21 (95.5%) Y14 accessions and 23 (92.0%) Xinhai ELS cultivars (Figure 4c).

In contrast, the other four Gh-i regions significantly increased fibre qualities but partly decreased yield in Gb (Figure 4a). Gh-i9 increased FL by 2.11% (t-test P = 1.06 × 10^-4) and FS by 3.65% of FS (t-test P = 3.4 × 10^-4) and decreased LP by 2.58% (t-test P = 1.40 × 10^-5). Similarly, the adjacent region Gh-i10 increased FL by 3.15% (t-test P = 1.1 × 10^-6), FS by 6.94% (t-test P = 6.80 × 10^-13) and SI by 3.10% (t-test P = 2.47 × 10^-6). Despite being adjacent in the genome, the distributions of these haplotypes in Gb accessions were quite different—especially in G3, which could be further classified into G3i and G3ii subgroups based on the respective presence of Gh-i9 and Gh-i10 (Figure 4c and Table S9). Interestingly, the haplotype with the longest IBD region, Gh-i6, was only found in G2 (Figure S9), revealing the distinct contribution to fibre quality improvement of the Central Asia-type cotton derived from founders introduced from the former Soviet Union. Taken together, the above results...
Figure 3  Interspecific introgressions identified by comparative population genomics. (a) Pairwise IBDs identified in Gb and Gh accessions. Gb-IBDs and Gh-IBDs are coloured yellow, blue and red, respectively. (b) Distribution of the relative proportions of IBD haplotypes (rIBD) in Gb (yellow) and Gh (blue) in bins of 10 kbp. rIBD = nGB/gB-nGHD/gHd, ranging from 1 (all haplotypes are IBD with Gb) to −1 (all haplotypes are IBD with Gh). rIBD_gb ≤ 0.0 are shown in inner box. c. Genome-wide distribution of interspecific common haplotypes. From outside to inside, I: Fat between Gb and Gh; II: Distribution of Gh-HAPs in the Gb population (>5%); III: Distribution of Gh-HAPs in the Gh population (>5%); IV: Chromosomal distribution of Gh-HAPs in each of 12 Gb accessions; V: Chromosomal distribution of Gh-HAPs in each of 9 Gh accessions. Gh-HAPs and Gh-HAPs are shown in orange and blue, respectively. (d) Phylogenetic neighbour-joining tree of a panel of Gb and Gh accessions. Branches labelled in red and blue indicate Gb and Gh, respectively. The heatmap of individual accessions indicates the ratio of interspecies-derived haplotypes. (e) Introgression regions from Gh in Gb populations. Blue, introgression origin from Gh; red, haplotype origin from Gb. Each column is a 10-kb genomic region, and each row is a phased haplotype of Gb. (f) Fat index between Gb and Gh plotted against chromosome position. The x-axis represents sliding windows of 100 kb on each chromosome, and the introgression regions are highlighted in grey. (g) Nucleotide diversity scores of Gb population plotted against chromosome position. The x-axis represents sliding windows of 100 kb, and the introgression regions are highlighted in grey. (h) Average r² of Gb plotted against chromosome position. The introgression regions are highlighted in grey.

underscore the difficulty of simultaneously improving yield and fibre qualities in Gb. Similar result was found in three QTL clusters (qClu-Chr7-3, qClu-Chr7-4, qClu-Chr7-5) in previous research in linkage studies, which increased FS and FL, but decreased LP and BW (Zhang et al., 2020).

To calculate the heritability of these six common haplotypes in terms of their effects on Gb fibre qualities and yield as measured in four-year replicate tests, a variance-component method (Yang et al., 2011b) was applied to construct genetic relationship matrices, with other genomic regions being used as control (Figure 4b). We found that each haplotype could explain more than 5% and even up to 40% of phenotypic variation in fibre qualities and yield (Figure 4b and Table S10). Gh-i5 alone explained ~20% of phenotypic variation in FL, FS, LP and SI. Of the four common haplotypes significantly associated with substantial increases in fibre qualities (FL, FS and FU), only Gh-i9 and Gh-i10 led to a decrease in yield, in contrast with Gh-i5. Gh-i9 explained ~5% of phenotype variation in FS, while Gh-i10 explained ~5% of LP and more than 10% of SI, indicating the different degrees of influence from these haplotypes (Figure 4b and Table S10). All told, Gh-i5 had the greatest contribution to phenotypic variation in fibre qualities and yield. Enhancing gene flow and the combination of elite interspecific haplotypes in further breeding programmes could increase varietal diversification of Gb and improve both yield and fibre quality, for example through introducing Gh-i5 into other ELS-type cultivars in G2/G3I/Gb increased genetic diversity in cultivated Gb, f lower diversity of upland cotton cultivars (Fang et al., 2017b; Janzen et al., 2019—d). In other words, this Gh locus (A06:19969833, P < 3.87 × 10⁻³), accessions with the A/A SNP haplotype have increased LP (yield) and decreased FL relative to other accessions containing the G/G SNP haplotype (Figure 5c and d). In further analysis, we also found that 06NH-16 and Giza29 have the haplotype similar to Gb, but this haplotype cannot be detected in Y14-18 and 03H-1 (Figure 5c and e).

If the 46 Gb accessions containing Gh-i5, 40 were enriched in A/A in Gbizi1, which allele significantly increased yield and was derived from Gh (two-tailed Fisher’s exact test P < 2.2 × 10⁻⁶, Figure 5f). These results highlight that some common haplotypes increased genetic diversity in Gb and then stabilized during the course of domestication and improvement, contributing superior agricultural performance relative to the receiving species. As observed in other crops, the introgression from wild cotton relatives has facilitated adaptation to novel conditions, affecting the targets of selection during domestication (Janzen et al., 2019).

In order to assess the range of improvement and application value of interspecific haplotypes or introgressions, we analysed the yield performance of 229 Gb and 45 Gh cultivars or lines under the same farming conditions for four years at four different locations (Figure S11). Gb cultivars showed superior FL (19%) and FS (34%); however, Gh cultivars had higher yield metrics, such as LP (18%). This indicates a range of improvement in both yield and quality. Most of the cultivars developed in China were derived from three founder landraces—DPL15 and STV2B from the United States, and Uganda Mian from Uganda—representing the narrow diversity of upland cotton cultivars (Fang et al., 2017a). Exchanges of interspecific functional haplotypes could offset the extant bottleneck and counteract the reduction of genetic diversity in cultivated Gb. Our study illustrates the
potential wide applications of these interspecific functional haplotypes in breeding practices.

In conclusion, analysis of a single population is limited to detecting components of intra-species importance. Comparative population genomics represents a new prospect for detecting interspecific functional haplotypes or introgressions during species formation and domestication. Here, most interspecific functional haplotypes identified from associated loci for agricultural traits were specific to Gb or Gh, reflecting the asynchronous improvement of each species during human-mediated hybridization or precision design. In particular, interspecific introgressions affected the agricultural performance of Gb, and some have undergone further natural or human selection and stabilized during the crop’s breeding history. It is very difficult to simultaneously improve yield and fibre qualities by incorporating a single natural hybridization or introgression. Our study not only highlights the strong asynchronous improvement in cotton species but also provides valuable interspecific introgressions from wild or cultivated species and a strategy for their application in further precision breeding.

**Experimental procedures**

**Sampling**

To conduct a comparative population genomics study, we selected 229 Gb and 491 Gh accessions. The 229 Gb accessions included the Egyptian, American Pima and Central Asia-type cotton mainly planted in China and were grown for four years (2014–2017) in Korla, Xinjiang autonomous region, that is the north-western cotton-growing area of China. The 234 Gh accessions from National cotton germplasm Mid-term bank in Institute of Cotton Research of CAAS were grown for three years (2011–2013) in three cotton-growing areas of China: Xinxiang in Henan province, Nanyang in Henan province and Korla in Henan province, Nanyang in Henan province and Korla in Henan province, Nanyang in Henan province and Korla in Henan province, Nanyang in Henan province and Korla in Henan province.
Figure 5 Interspecific introgressions are under further human selection. (a) Manhattan plots of overlapped loci associated with FL and LP on Chr.A06 of Gb, which overlapped a segment of an interspecific introgression. Two interspecific QTLs, F2-qFL-c6-1 and Mp7_uhml_6_1_4.25, are also shown at the top. (b) Candidate genes with significant nonsynonymous SNPs within GWAS loci from a, including GbTTL, GbLURP and GbJAZ1. Candidate genes in the locus A06Gb:19557428 overlapped with interspecific Gh-i5. (c) Haplotype patterns of chromosome A06 in GB population. Blue represents haplotype origin from Gh; orange, haplotype origin from Gb. (d) Boxplot for FL and LP by the associated SNPs of each gene. Centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× the interquartile range; dots, outliers (**P < 0.01, two-tailed t-test). (e) Diagram of two accessions without and two with Gh-i5 based on rIBD. (f) Significant relationships between associated loci and Gh-i5. The frequencies are for GWAS-associated SNPs in GbJAZ1, GbLURP and GbTTL in accessions with and without Gh-i5.
Xinjiang. The remaining 257 Gh accessions were selected in a previous study (Fang et al., 2017b). We analysed key agronomic traits related to yield (lint percentage, lint yield, seed index and seedcotton yield) and fibre quality (fibre elongation, fibre length, fibre micronaire, fibre strength and fibre uniformity). Fibre quality traits were measured with the HVI9000 system (Üster Technologies AG, Charlotte, NC, USA) at the Supervision, Inspection and Test Center of Cotton Quality, Ministry of Agriculture, China.

Library construction and sequencing

We collected young leaf tissues for genomic DNA extraction using a standard cetyltrimethylammonium bromide (CTAB)-based protocol (Murray and Thompson, 1980) and prepared genomic DNA from a single plant of each accession for sequencing. Each library was sequenced on an Illumina HiSeq2500 instrument (Illumina, Inc., San Diego, CA, USA) to generate paired-end reads of about 150 bp.

SNP identification

For GWAS analysis, 229 Gb accessions were aligned to the Hai7124 reference (Hu et al., 2019). Briefly, raw paired-end reads were filtered by Trimomatic (v 0.38) (Bolger et al., 2014) with default parameters and subsequently mapped to the Gb and Gh reference (Hu et al., 2019) using the ‘mem’ algorithm of BWA (v 0.7.17-r1188) (Li and Durbin, 2009). Aligned reads were converted to BAM format using samtools (v 1.9) (Li and Durbin, 2009), and duplicate reads marked with the MarkDuplicates method of Picard (v 1.124) (http://broadinstitute.github.io/picard). Variants were detected using germline (v 1.124) (Li and Durbin, 2009). Aligned reads were converted to BAM format using samtools (v 1.9) (Li and Durbin, 2009), and duplicate reads marked with the MarkDuplicates method of Picard (v 1.124) (http://broadinstitute.github.io/picard). Variants were called using samtools (v 1.9) with parameters ‘-q 20 -Q 15 –ugf’ and bcftools (v 1.8) with parameters ‘vmoZ’ (Li and Durbin, 2009). Biallelic SNPs were selected for further analysis using vcftools (v 0.1.13) (Danecek et al., 2011) with filter parameters of mapping depth > 3, mapping quality > 20, genotyping rate > 90% and minor allele frequency (MAF) > 0.05. To reveal interspecific genetic divergence and IBD identification, all 229 Gb and 185 Gh accessions were aligned to Gh TM-1 (Hu et al., 2019) for SNP calling.

Genotype imputation and accuracy estimation

To improve mapping resolution, imputation of missing genotypes was implemented using Beagle (v 4.1) with 50 iterations in a sliding window of 50 000 SNPs (Browning and Browning, 2016). Imputed genotypes were performed using ANNOVAR (version 2015-12-14) (Wang et al., 2010) based on the reference genome annotation.

Phylogenetic and population structure analysis

Phylib software (v 3.69) was used to generate the neighbour-joining tree, which was then visualized using EvolView (www.evolgenius.info/evolview) (He et al., 2016). PCA was performed with GCTA (v 1.92.1) (Yang et al., 2011a) using the SNP data set from Gb and Gh. For genetic diversity analysis, vcftools (v 0.1.13) was used to calculate \( \pi \) and Fst with a window size of 100k bp and a step size of 100 kb (Danecek et al., 2011).

Detection of interspecific introgressions

We used the rIBD method to detect interspecific introgression without requirement for an outspecies, and the modified rIBD method (Wang et al., 2019) to identify putative introgression regions. First, identity by descent (IBD) was detected using phased SNPs from Gb and Gh accessions, and haplotypes shared between each pair of accessions were detected using GERMLINE (v 1.5.3) (Gusev et al., 2009). We tested different GERMLINE parameters for the detection of IBD (1: -bits 30 -het 0 -hom 0; 2: -bits 30 -hom 3 -het 3; 3: -bits 50 -hom 3 -het 0; 4: -bits 50 -hom 3 -het 3; 5: -bits 50 -hom 3 -het 3; 6: -bits 100 -hom 0 -het 0; 7: -bits 100 -hom 3 -het 3; 8: -bits 128 -hom 0 -het; and 9: -bits 128 -hom 3 -het 3). Next, each chromosome was divided into bins of 10 kb with a sliding window of 10 kb. The number of IBD tracts identified between each accession and the two groups of accessions (i.e. Gb and Gh) was computed per bin using a custom perl script. Relative IBD (\( \text{rIBD} = \text{rIBD}_{\text{in}} - \text{rIBD}_{\text{out}} \)) was used to define putative introgression, with \( \text{rIBD} = 1 \) signifying that all haplotypes originate from Gb and \( \text{rIBD} = -1 \) that all haplotypes originate from Gh.

SNP-based heritability estimation

The phenotypic variation of interspecific common haplotypes was evaluated using the GCTA-GREML method (Yang et al., 2011a). For each common haplotype identified, we built genetic relationship matrices using GCTA (v 1.92.1) with the parameter ‘make-grm’ (Yang et al., 2011a). We then estimated the amount of phenotypic variation in FI, FS, FU and LP that could be explained by seven SNP sets using the parameter ‘mgrp’. We repeated this process 100 times, each time randomly sampling the set of SNPs, and calculated the heritability as \( \text{V(G)/V(p)} \).

Identifying regions of suppressed recombination

The level of linkage disequilibrium (LD) used to demarcate regions was calculated using plink (v 1.90b6.8) with parameters ‘--r2 --ld-window 100’ (Purcell et al., 2007). The \( r^2 \) for all chromosomes was calculated by sliding across the chromosomes in 100-kb windows with a step size of 100 kb. The mean \( r^2 \) value for each window was estimated by calculating the mean for all pairwise combinations of SNPs within a window using a custom script.

Genome-wide association study

We carried out a large-scale GWAS using 4 476 553 common SNPs with MAF > 0.05. The association analysis was done using the program Efficient Mixed-Model Association eXpedited (EMMAX) (http://genetics.cs.ucla.edu/emmax/index.html). The effective number of independent SNPs and the suggestive \( P \)-value were estimated using gec (v 0.2) (Li et al., 2012), with the \( P \)-value being \( 9.27 \times 10^{-7} \) for the \( Gb \) population. Associated loci passing a threshold of 10^{-5} were also selected for subsequent analysis because these loci showed good repeatability in different phenotypic datasets.

Distance- and LD-based clumping

Clumping was used to convert significant SNPs to regions. Within a given genomic window, the SNP with the smallest \( P \)-value was kept as the index SNP, and others in high LD with the index SNP were used to define the left and right ends of the region (Gb SNPs with \( P < 9.27 \times 10^{-7} \) and \( r^2 > 0.6 \) within 1-Mb windows, PLINK v1.90b6.8 with parameters ‘-ld-window-r2 0.6 -ld-window 99999’ (Purcell et al., 2007).

RNA-seq analysis

RNA-seq reads from different tissues were obtained from our previously published data (accession ID, PRJNA490626) (Hu et al., 2019). The raw reads were cleaned using fastq (v 0.12.2) (Chen et al., 2018) and then aligned to the corresponding reference genome using Hisat (v 2.1.0) (Pertea et al., 2016). Gene expression levels were quantified as fragments per kilobase of transcript per million mapped reads (FPKM) using Stringtie (v 2.0) (Pertea et al., 2016).
Acknowledgments
This work was financially supported in part by grants from NSFC (31822036, 31661143016), the Fundamental Research Funds for the Central Universities (2019XZZX004-13, 2020XZZX004-03) and the Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (2019R01002).

Conflict of interest
The authors declare no competing financial interests.

Author contributions
L.F. and T. Zhang conceptualized the research programme. Z.S., X. Z., Z.H., S.W. and G.L. collected the cotton samples and worked on the phenotype. Y.H., S.W. and Z.S. extracted the high-quality DNA, constructed DNA sequencing libraries and performed the genome sequencing. L.F., T. Zhao, H.M., L.W., B.Q., H.W., M.G. and X.G. performed the genotyping and bioinformatics analyses. L.F. T. Zhao and T. Zhang analysed all of the data and wrote the manuscript. All authors discussed results and commented on the manuscript.

Data availability statement
All sequenced genomic data for GWAS analysis and genetic divergence analysis are available in the NCBI Sequence Read Archive under the accession PRJNA613140. The genomic variants, genotype data and phenotype data can be downloaded from the Cotton Omics Database (http://cotton.zju.edu.cn/).

References
Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
Böger, A.M., Lohse, M. and Usadel, B. (2014) Trimomatic: a flexible trimmer for illumina sequence data. Bioinformatics 30, 2114–2120.
Bosse, M., Megens, H.J., Frantz, L.A., Madsen, O., Larson, G., Paudel, Y., Duijvestein, N. et al. (2014) Genomic analysis reveals selection for asian genes in European pigs following human-mediated introgression. Nat. Commun. 5, 4392.
Browning, B.L. and Browning, S.R. (2016) Genotype imputation with millions of reference samples. Am. J. Human Genet. 98, 116–126.
Chen, S., Zhou, Y., Chen, and Gu, J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, 884–890.
Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.e et al. (2011) The variant call format and VCFtools. Bioinformatics 27, 2156–2158.
Fang, L., Gong, H., Hu, Y., Liu, C., Zhou, B., Huang, T., Wang, Y. et al. (2017a) Genomic insights into divergence and dual domestication of cultivated allotetraploid cottons. Genome Biol. 18, 33.
Fang, L., Wang, G., Hu, Y., Jia, Y., Chen, J., Liu, B., Zhang, Z. et al. (2017b) Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. Nat. Genet. 49, 1089–1098.
Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altschuler, D., Breslow, J.L. et al. (2009) Whole population, genome-wide mapping of hidden relatedness. Genome Res. 19, 318–328.
He, F., Pasam, R., Shi, F., Kanti, S., Keeble-Gagnere, G., Kay, P., Forrest, K. et al. (2019) Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat. Genet. 51, 896–904.
He, Z., Zhang, H., Gao, S., Lercher, M.J., Chen, W.H. and Hu, S. (2016) Evolvui v2: an online visualization and management tool for customized and annotated phylogenetic trees. Nucleic Acids Res. 44, 236–241.
Hu, Y., Chen, J., Fang, L., Zhang, Z., Ma, W., Niu, Y., Ju, L. et al. (2019) Gossypium barbadense and Gossypium hirsutum genomes provide insights into the origin and evolution of allotetraploid cotton. Nat. Genet. 51, 739–748.
Hu, H., He, X., Tu, L., Zhu, L., Zhu, S., Ge, Z. and Zhang, X. (2016) GHIAZ2 negatively regulates cotton fiber initiation by interacting with the R2R3-MYB transcription factor GHMYB25-like. Plant J. 88, 921–935.
Janzen, G.M., Wang, L. and Hufford, M.B. (2019) The extent of adaptive wild introgression in crops. New Phytol. 221, 1279–1288.
Lacape, J.M., Gavrishvsi, G., Cao, T.-V., Viot, C., Llewellyn, D., Liu, S., Jacobs, J. et al. (2013) Mapping QTLs for traits related to phenoology, morphology and yield components in an inter-specific Gossypium hirsutum × G. barbadense cotton RIL population. Field Crops Res. 144, 256–267.
Lacape, J.M., Llewellyn, D., Jacobs, J., Arioli, T., Becker, D., Calhoun, S., Al-Ghazy, Y. et al. (2010) Meta-analysis of cotton fiber quality QTLs across diverse environments in a Gossypium hirsutum × G. barbadense RIL population. BMC Plant Biol. 10, 132.
Lawrie, D.S. and Petrov, D.A. (2014) Comparative population genomics: power and principles for the inference of functionality. Trends Genet. 30, 133–139.
Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.
Li, M.X., Yeung, J.M., Cherry, S.S. and Sham, P.C. (2012) Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. Human Genet. 131, 747–756.
Liu, Z., Cook, J., Melia-Hancock, S., Guil, K., Bottoms, C., Garcia, A., Ott, O. et al. (2016a) Expanding maize genetic resources with Predomestication alleles: maize-teosinte introgression populations. Plant Genome 9(1). https://doi.org/10.38335/plantgenome2015.07.0053
Liu, Z., Garcia, A., McMullen, M.D. and Flint-Garcia, S.A. (2016b) Genetic analysis of kernel traits in maize-teosinte introgression populations. G3: Genes - Genetics - Genomic. 6, 2523–2530.
Liu, R.X., Gong, J.W., Xiao, X.H., Zhang, Z., Li, J.W., Liu, A.Y., Liu, Q.W. et al. (2018) GWAS analysis and QTL identification of fiber quality traits and yield components in upland cotton using enriched high-density SNP Markers. Front. Plant Sci. 9, 1067. https://doi.org/10.3389/fpls.2018.01067
Murray, M.G. and Thompson, W.F. (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. 8, 432–435.
Nie, X., Tu, J., Wang, B., Zhou, X. and Lin, Z. (2015) A BIL Population derived from G. hirsutum and G. barbadense provides a resource for cotton genetics and breeding. PLoS One 10, e0141064.
Nie, X., Wen, T., Shao, P., Tang, B., Nuriman-Guli, A., Yu, Y., Du, X. et al. (2020) High-density genetic variation maps reveal the correlation between asymmetric interspecific introgressions and improvement of agronomic traits in Upland and Pima cotton varieties developed in Xinjiang, China. Plant J. 103, 677–689.
Percy, R.G. and Wendel, J.F. (1990) Allozyme evidence for the origin and diversification of Gossypium barbadense. L. Theoret. Appl. Genet. 79, 529–542.
Pertea, M., Kim, D., Pertea, G.M., Leek, J.T. and Saltzberg, S.L. (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, Stringtie and Ballgown. Nat. Protocols 11, 1650–1667.
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, B. et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Human Genet. 81, 559–575.
Qi, T., Song, S., Ren, Q., Wu, D., Huang, H., Chen, Y., Fan, M. et al. (2011) The Jasmonate-ZIM-domain proteins interact with the WD-Repeat/LHL/MBT complexes to regulate Jasmonate-mediated anthocyanin accumulation and trichome initiation in Arabidopsis thaliana. Plant Cell 23, 1795–1814.
Ruan, Y.L., Llewellyn, D.J. and Furbank, R.T. (2001) The control of single-celled cotton fiber elongation by developmentally reversible gating of plasmodesmata and coordinated expression of sucrose and K+ transporters and expansion. Plant Cell 13, 47–60.
Shi, Y., Li, W., Li, A., Ge, R., Zhang, B., Li, J., Liu, G. et al. (2015) Constructing a high-density linkage map for Gossypium hirsutum x Gossypium barbadense and identifying QTLs for lint percentage. J. Integr. Plant Biol. 57, 450–467.
Si, Z.F., Chen, H., Zhu, X.F., Cao, Z.B. and Zhang, T.Z. (2017) Genetic dissection of lint yield and fiber quality traits of \textit{G. hirsutum} in \textit{G. barbadense} background. \textit{Mol. Breed.} \textbf{37}, 9. https://doi.org/10.1007/s11032-016-0607-3

Swarm, S.A., Sun, L.J., Wang, X.T., Wang, W.D., Brown, P.I., Ma, J.X. and Nelson, R.L. (2019) Genetic dissection of domestication-related traits in soybean through genotyping-by-sequencing of two interspecific mapping populations. \textit{Theor. Appl. Genet.} \textbf{132}, 1195–1209.

Thompson, E.A. (2013) Identity by descent: variation in meiosis, across genomes, and in populations. \textit{Genetics} \textbf{194}, 301–326.

Wang, X., Chen, L. and Ma, J. (2019) Genomic introgression through interspecific hybridization counteracts genetic bottleneck during soybean domestication. \textit{Genome Biol.} \textbf{20}, 22.

Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. \textit{Nucleic Acids Res.} \textbf{38}, e164.

Wang, F.R., Xu, Z.Z., Sun, R., Gong, Y.C., Liu, G.D., Zhang, J.X., Wang, L.M. et al. (2013) Genetic dissection of the introgressive genomic components from \textit{Gossypium barbadense L.} that contribute to improved fiber quality in \textit{Gossypium hirsutum L.} \textit{Mol. Breed.} \textbf{32}, 547–562.

Wendel, J.F. and Cronn, R.C. (2003) Polyploidy and the evolutionary history of cotton. \textit{Adv. Agronomy} \textbf{78}, 139–186.

Wu, D.D., Ding, X., Wang, S., Wojcik, J., Wang, Y., Tokarska, M., Li, Y. et al. (2018) Pervasive introgression facilitated domestication and adaptation in the \textit{Bos} species complex. \textit{Nat. Ecol. Evol.} \textbf{2}, 1139–1145.

Xu, J., Li, H.D., Chen, L.Q., Wang, Y., Liu, L.L., He, L. and Wu, W.H. (2006) A protein kinase, interacting with two calciumin B-like proteins, regulates K+- transporter AKT1 in Arabidopsis. \textit{Cell} \textbf{125}, 1347–1360.

Yang, C., Batzel, F., Hohmann, N., Koch, M., Turk, F. and Calonje, M. (2013) VAL- and ATBMI1-mediated H2Aub initiate the switch from embryonic to postgerminative growth in Arabidopsis. \textit{Curr. Biol.} \textbf{23}, 1324–1329.

Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011a) GCTA: a tool for genome-wide complex trait analysis. \textit{Am. J. Human Genet.} \textbf{88}, 76–82.

Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M. et al. (2011b) Genome partitioning of genetic variation for complex traits using common SNPs. \textit{Nat. Genet.} \textbf{43}, 519–544.

Yu, J.Z., Ulloa, M., Hoffman, S.M., Kohel, R.J., Pepper, A.E., Fang, D.D., Percy, R.G. et al. (2014) Mapping genomic loci for cotton plant architecture, yield components, and fiber properties in an interspecific (\textit{Gossypium hirsutum L.} × \textit{G. barbadense L.}) RIL population. \textit{Mol. Gen. Genom.} \textbf{289}, 1347–1367.

Zhang, T., Hu, Y., Jiang, W., Fang, L., Ghan, X., Chen, J., Zhang, J. et al. (2015a) Sequencing of allotetraploid cotton (\textit{Gossypium hirsutum L.} acc. TM-1) provides a resource for fiber improvement. \textit{Nat. Biotechnol.} \textbf{33}, 531–537.

Zhang, Z., Li, J.W., Jamshed, M., Shi, Y.Z., Liu, A.Y., Gong, J.W., Wang, S.F. et al. (2020) Genome-wide quantitative trait loci reveal the genetic basis of cotton fibre quality and yield-related traits in a \textit{Gossypium hirsutum} recombinant inbred line population. \textit{Plant Biotechnol. J.} \textbf{18}, 239–253.

Zhang, Y.D., Zheng, J., Liang, Z.K., Liang, Y.L., Peng, Z.H. and Wang, C.L. (2015b) Verification and evaluation of grain QTLs using RILs from TD70 x Kasalath in rice. \textit{Genet. Mol. Res.} \textbf{14}, 14882–14892.

**Supporting information**

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Figure S1** Phylogenetic tree of all resequenced \textit{Gb} accessions used in GWAS analysis.

**Figure S2** Principal component analysis of resequenced \textit{Gb} accessions using whole-genome SNP data.

**Figure S3** Partial overlap of GWAS loci between \textit{Gb} and \textit{Gh} on A08.

**Figure S4** Partial overlap of GWAS loci between \textit{Gb} and \textit{Gh} on D05.

**Figure S5** Partial overlap of GWAS loci between \textit{Gb} and \textit{Gh} on D12.

**Figure S6** Potential causative gene in the partial overlap of GWAS loci on A08.

**Figure S7** Phylogenetic neighbour-joining tree of all sequenced accession of \textit{Gb} and \textit{Gh} used in comparative population genomics study.

**Figure S8** Principal Component Analysis (PCA) of all sequenced \textit{Gb} and \textit{Gh} accessions used in the comparative population genomics study.

**Figure S9** Phylogenetic relationships among the 229 accessions revealed by introgression regions from \textit{Gh}.

**Figure S10** Distribution of introgressions from wild \textit{Gh} in \textit{Gb} accessions.

**Figure S11** Comparison of phenotypic distributions between 229 \textit{Gh} and 45 \textit{Gb} accessions grown in the same environment.

**Table S1** Yield and fibre quality phenotypes of \textit{Gb} accessions used in the comparative population genomics study.

**Table S2** Yield and fibre quality phenotypes of \textit{Gh} in multiple environments.

**Table S3** Summary of resequenced data from \textit{Gb} accessions.

**Table S4** Associated loci for yield and fibre quality in \textit{Gb} accessions.

**Table S5** Summary of resequenced data from \textit{Gh} accessions.

**Table S6** Associated loci for yield and fibre quality in \textit{Gh} accessions.

**Table S7** Partial overlap of GWAS loci between \textit{Gb} and \textit{Gh} accessions.

**Table S8** Distribution of twelve obvious interspecific introgressions in \textit{Gb} accessions.

**Table S9** Effects of eight introgressions on phenotype variation in \textit{Gb} accessions.

**Table S10** Heiratability for fibre qualities and yield traits from six interspecific introgressions.

**Table S11** Overlap of associated loci with interspecific introgressions in \textit{Gb}.