An improved high-quality genome assembly and annotation of Qingke, Tibetan hulless barley

Xingquan Zeng[1,2,3,] †, Xiangfeng Li[5,]†, Lijun Bai[5,]†, Yulin Wang[1,2,3,]†, Tong Xu[5], Hongjun Yuan[1,2,3] , Qijun Xu[1,2,3], Sang Zha[1,2,3], Zexiu Wei[1,3,4], Wangmu Qimei[1,2,3], Yuzhen Basang[1,2,3], Jiabu Dunzhu[1,2,3],Mingzai Yu[1,2,3] & Tashi Nyima[1,3,]*.

Affiliations

1 State Key Laboratory of Barley and Yak Germplasm Resources and Genetic Improvement, Lhasa 850002, China.

2 Research Institute of Agriculture, Tibet Academy of Agricultural and Animal Husbandry Sciences Lhasa, Tibet 850002, China.

3 Tibet Academy of Agricultural and Animal Husbandry Sciences, Lhasa, Tibet 850002, China.

4 Agricultural Resources and Environment Research Institute, Tibet Academy of Agricultural and Animal Husbandry Sciences, Lhasa, Tibet 850002, China.

5. Chengdu Life Baseline Technology Co., LTD, Chengdu 610041, China

† These authors contributed equally to this work.

* Correspondence and requests for materials should be addressed to T.N. (email: nima_zhaxi@sina.com)

Abstract

Background

The Tibetan hulless barley (Hordeum vulgare L. var. nudum), also called “Qingke” in Chinese and “Ne” in Tibetan, is the staple food for Tibetans and an important livestock feed in the Tibetan Plateau. The Qingke in China has about 3500 years of cultivation history, mainly produced in Tibet, Qinghai, Sichuan, Yunnan and other areas. In addition, Qingke has rich nutritional value and outstanding health effects, including the beta glucan, dietary fiber, amylopectin, the contents of trace elements , which are higher than any other cereal crops.

Findings

Here, we reported an improved high-quality assembly of Qingke genome with 4.0 Gb in size. We employed the falcon assembly package ,scaffolding and error correction tools to finish improvement using PacBio long reads sequencing technology, with contig and scaffold N50 lengths of 1.563Mb and 4.006Mb, respectively, representing more continuous than the original Qingke genome nearly two orders of magnitude. We also re-annotated the new assembly, and reported 61,303 stringent confident putative protein-coding genes, of which 40,457 is HC genes. We have developed a new Qingke genome database (QGD) to download and use friendly, as well as to better manage the information of the Qingke genetic resources.

Conclusions

The availability of new Qingke genome and annotations will take the genetics of Qingke to a new level and will greatly simplify the breeders effort. It will also enrich the granary of the Tibetan people.

Keyword: Qingke; Hulless barley; Genome assembly; full-length cDNA; Qingke Genome Database

Introduction

Qingke (Hordeum vulgare L. var. nudum) is a cereal crop, a variant of barley, alias called hulless barley, yuan wheat, rice barley(**Fig.1**). Qingke has a long history of cultivation in the Qinghai-Tibet Plateau. The adaptation to extreme environmental conditions in high altitudes made it the staple food for Tibetans beginning at least 3,500–4,000 years ago [1]. It continues to be the predominant crop in Tibet, occupying ~70% of crop lands.

Genome sequences provide a substantial basis for understanding the biological essences of crops associated with biologically and economically essential traits[2]. Therefore, it is necessary and crucial to construct the genome of a species. In recent years, with the development of NGS sequencing technology, many crops have published the genome draft or near-finished genome sequences, especially the larger genome of the wheat family[3-13]. At the same time, two studies have published two versions of the Qingke genome in 2015[2] and 2017[14], respectively. And the former assembly strain is Lasa Goumang , the latter is Zangqing320, all excellent cultivars of Qingke. The estimated genome size of the Qingke is 4.48Gb[2], which is much smaller than the 5.1Gb[4, 10] of barley. Although the former is a subspecies of barley, considering that the barley genome is rich in repeats(about 80.8%-84%[2, 4, 10]), the difference in genome size may come from the numbers and length of transposons.

Although the publication of the Qingke genome has attracted a new upsurge in Qingke genetics research[15, 16], due to the current poor genome assembly, such as the particularly short contig, the very low contig N50, and the poor genome annotation, the functional genome studying of Qingke has caused some vital problems. In addition, with the development of single-molecule sequencing platforms, such as those by Pacific Biosciences and Oxford Nanopore Technologies, have become available to researchers and are currently being increasingly applied in *de novo* genome assembly and genome update , and other scaffolding techniques(such as BioNano optical mapping, Chicago, Hi-C and 10X genomics)[17-19], it provides a new opportunity for improving Qingke genome assembly. Therefore, we *de novo* assembled a high-quality genome of Qingke using high-coverage Pacbio long reads, combining published Illumina paired-ends and mate-pairs

reads. And then, we applied PacBio Iso-seq to get high confidence full-length transcripts and re-annotation Qingke protein-coding genes. This new assembly, annotations and newly online Qingke genomics database discussed below will serve as valuable resources for investigating the economic and genetic of Qingke and will also help the molecular breeders to continually improve Qingke or barley germplasm resources.

DNA and RNA sequencing

DNA isolation, libraries construction and sequencing

The seedling of Qingke (NCBI BioSample ID **SAMN09914874**) were collected from Tibet, China. After collection, tissues were immediately transferred into liquid nitrogen and stored until DNA extraction. DNA was extracted using the Cetyltrimethyl Ammonium Bromide (CTAB) method according to the protocal " Preparing Arabidopsis Genomic DNA for Size-Selected ~20 kb SMRTbell™ Libraries "(https://www.pacb.com/wp-content/uploads/2015/09/Shared-Protocol-Preparing-Arabidopsis-DNA-for-20-kb-SMRTbell-Libraries.pdf). The quality of the extracted genomic DNA was checked using 1% agarose gel electrophoresis, and the concentration was quantified using a Qubit fluorimeter (Invitrogen, Carlsbad, CA, USA).

Single-molecule Real-time long reads sequencing was performed at NextOmics Technology Corporation (Wuhan, China) with a PacBio Sequel Sequencer (Pacific Biosciences, Menlo Park, CA,USA). The SMRT Bell library was prepared using a DNA Template Prep Kit 1.0, and Six 20-kb SMRT Bell libraries were constructed(**Fig. S1**). Genomic DNA（10ug）was mechanically sheared using a Covaris g-Tube with a goal of DNA fragments of approximately 20kb. A Bioanalyzer 2100 12K DNA Chip assay was used to assess the fragment size distribution. Sheared genomics DNA(5ug) was DNA-damage repaired and end-repaired using polishing enzymes. A blunt-end ligation reaction followed by exonuclease treatment was conducted to generate the SMRT Bell template. A Blue Pippin device(Sage Science) was used to size select the SMRT Bell template and enrich large fragments(>10kb).The size-selected library was quality inspected and quantified on an Agilent Bioanalyzer 12Kb DNA Chip (Agilent) and a Qubit fluorimeter (Life Technologies). A ready-to-sequence SMRT Bell Polymerase Complex was created

using a Binding Kit 2.0 (PacBio p/n 100-862-200), according to the manufacturer's instructions. The PacBio Sequel instrument was programmed to load and sequenced the sample on PacBio SMRT cells v2.0 (Pacific Biosciences), acquiring one movie of 360 min per SMRT cell. The MagBead loading (Pacific Biosciences) method was used to improve the enrichment of the larger fragments. A total of 64 SMRT cells were processed yielding 300.24G subread sequences, average 4.69Gb per cell, and 9,358bp in length(**Table S1**).

For Illumina sequencing data, we applied previously published paired-ends and mate-pair library data[2](SRA no. SRA201388 , totally 819.37Gb clean data) for PacBio assembly error correction and scaffolding, respectively.

RNA isolation and sequencing

Fresh roots, and leaves of Qingke were sampled in Tibet, China. Totally, 12 samples ,different N condition (Prepared for unpublished paper) were used for Iso-seq sequencing. Total RNA was prepared by grinding tissue in TRIzol reagent (Invitrogen 15596026) on dry ice and processed following the protocol provided by the manufacturer. To remove DNA, an aliquot of total RNA was treated with RQ1 DNase (Promega M6101), followed by phenol/chloroform/isoamyl alcohol extraction, chloroform/isoamyl alcohol extraction using Phase Lock Gel Light tubes (5 PRIME 2302800) and ethanol precipitation. Precipitated RNA was stored at -20℃.

One microgram of total RNA per reaction per tissue was reverse transcribed using the Clontech SMARTer cDNA synthesis kit. Specific barcoded oligo dT in separate PCR tubes, to generate barcoded FL cDNA. Large-scale PCR products were purified with AMPure PB beads and quality control (QC) was performed on a 2100 BioAnalyzer (Agilent). Equimolar ratios of six cDNA libraries were pooled together. A total of 3.8 mg cDNA was subjected to size fractionation using the SageELF system. Size fractions eluted from the run were subjected QC and pooled in equimolar ratios for subsequent re-amplification to yield four libraries (1–2, 2–3, 3–6 and 5–10Kb). The pooled PCR products were purified using AMPure PB beads. One to five micrograms of purified amplicons were subjected to Iso-Seq SMRT Bell library preparation (https://pacbio.secure.force.com/SamplePrep) .A total of 17 SMRT cells (unpublished data) were sequenced on the PacBio RS II platform using P6-C4 chemistry with 3–4 h movies.

Genome assembly

The Sequel raw bam files were converted into subreads in fasta format using the standard Pacbio software package BAM2fastx. Totally, 300.24Gb PacBio data, covered genome ~67X(estimated genome size 4.48Gb),was used to initially assemble.

Falcon assembly

we used falcon software package（https://github.com/PacificBiosciences/falcon ）to construct primary assembly. Firstly, corrected the error reads by the overlap strategy (parameter length_cutoff control is used to correct the wrong seed reads length);Secondly, pre-assembled the corrected reads; Thirdly, Constructed contig with long reads obtained from pre-assembly (parameter length_cutoff_pr controls the pre-assembled reads length used to build contig); After many attempts, we determined that the following two reference settings can get better assembly results: length_cutoff = 5000;length_cutoff_pr = 10000. And, we get the initial assembly, 1.229Mb in length of contig N50.

Arrow polish

There are often errors in the preliminary assembly results of the genome, such as SNP or InDel, which can be fixed by certain algorithms. Arrow is a genomic polishing tool based on hidden Markov models. We use the following steps to polish the genome:

1) Align the PacBio sequencing reads to the genome assembly with blasr (pbalign --algorithm=blasr, the rest of the parameters default);

2) Use Arrow to polish based on the alignment result (variantCaller --algorithm=arrow, the rest of the parameters default).pbalign and variantCaller are from the SMRT Link 4 toolkit (https://www.pacb.com/support/software-downloads/)

Scaffolding

The Scaffolding tool can group, sort, and orient contigs to join contigs into longer sequence fragments, called scaffold. The sequence between some contigs may be unknown but the distance between them can be roughly known, which is also helpful for subsequent related work.

SSPACE[20] is a widely used tool for building scaffold. Firstly aligned the reads to contigs using alignment tools (such as BWA, bowtie), and then built scaffold based on the connection information between the contigs. This project uses Illumina data previously published and PacBio data to build scaffold.

Firstly， using Illumina mate-pair library (20kb and 40Kb fragment long) sequencing data to scaffolding contigs. Aligned the paired-end reads to contigs with bowtie[21]; then, constructed scaffold (parameter: -p 1 -g 2) with SSPACE-STANDARD-3.0.

Secondly, using PacBio data to continuously improving the scaffolds. Aligned PacBio reads to the scaffold version of Illumina data using blasr; and then, built scaffold with SSPACE-LongRead[22].

The final assembly produced 1,856 scaffolds, 1,092 of which greater than or equal to 1Mb, with a contig N50 length of 1.563Mb and scaffold N50 of 4.006Mb. Compared to the preliminary draft of the published two Qingke genomes[2, 14], the contiguity was improved more than ~80 and 243 times, respectively(**Table 1**;**Table S2-S3**).

Assembly evaluation

First, the 682.57Gb Illumina paired-ends reads(library fragment 250,500 and 800,respectively) were mapped to new assembly, and 99.87-99.89% PE reads were mapped concordantly (**Table S4**).

Using these short reads, the estimated quality value (QV) of Qingke genome was calculated according to a previously described method[23]. The erroneous bases in the genome assembly were identified using the variant calling software Genome Analysis Toolkit (GATK) v4[24] with default parameters. The QV of the Qingke genome was estimated to be 52(**Table S5**), which means that the accuracy of the assembly in base level is very accurate after base correction.

We randomly downloaded 4 *Hordeum vulgare subsp. vulgare* (domesticated barley) BAC clone sequencing data (NCBI ID: AC248476.1,AC250041.1,AC253057.1 and AC253422.1) to evaluate assembly quality. Using nucmer[25] mapping BAC data to new assembly, the results manifested very high consistency.(**Fig. S2**).

194 In addition, we also used BUSCO[26] to perform an assessment of genome
195 completeness . BUSCO used a total of 1440 genes to evaluate the genome; the genome has
196 1312 complete BUSCO genes, accounting for 91.1%, of which 1288 have single copies,
197 accounting for 89.4%, 24 have multiple copies, accounting for 1.7%; 34 gene have only a
198 few fragments of the gene; 94 genes are missing from the genome.
199 Together, the results indicated that our dataset represented a genome assembly with
200 a high level of coverage, quality and completeness.

201 Repetitive sequences annotation

202 Repetitive DNA accounts for much of the remarkable variation in plant genome
203 sizes[27], especially some Gramineae crop plants, above 80% [10, 12, 28] of genome for
204 repetitive elements . Repetitive sequences are divided into transposons and simple repeats.
205 According to Tandem repeats finder program[29]，we marked simple repeat elements,
206 155,813,415 in length，3.89% of genome. For transposons, we performed homology-
207 based and de novo prediction, respectively. The software packages involved are
208 RepeatMasker[30], RepeatProteinMask[30] , repeatmodeler
209 (http://www.repeatmasker.org/RepeatModeler/) and ltr_finder[31]. The repetitive
210 database used here is Repbase 21.01[32](**Fig. S3**). Long terminal repeat(LTR)
211 retroelements, belong to class I elements, is accounted for 70.22% of genome. All repeats
212 together accounted for 87.48% of the genome(**Table 2**;**Table S6**).

213 Gene prediction

214 For the annotation of protein-coding gene, we used a combination of multiple data
215 sources, followed by a high confidence (HC) and low confidence (LC) classification of the
216 annotated genes(**Fig. S3-4**).

217 **Combination of multiple annotations**

218 According to Iso-seq methods, we obtained 39,442 high-quality full-length
219 transcripts. Theses high-quality transcripts were aligned to the genome using GMAP
220 (version 2017-05-08)[33], with a total of 14,099 perfect alignments; TransDecoder

221 v4.0.1(https://github.com/TransDecoder) was used to predict the coding region of 14,099
222 transcripts, of which 13,936 contained open reading frame (ORF), these ORF contained
223 transcripts are ultimately assigned to 9,360 genes, called perfect genes. We employed
224 Augustus v3.3[34],which used perfect genes as input train sets, to get 128,400 putative
225 genes. At the same time ,we downloaded representative proteins of 7 related species
226 (*Triticum aestivum A*[12], *Triticum aestivum D*[9], *Brachypodium distachyon*[35],
227 *Hordeum vulgare*[10], *Oryza sativa*[36], *Sorghum bicolor*[37] and *Zea mays*[38]) from
228 public database(**Table S7**). Aligning the protein sequences of each species to the Qingke
229 genome using tblastn v2.6.0+[39]; and then, using Genewise v2.4.1[40] to predict gene
230 structure and obtain homologous gene sequences from various species in the Qingke
231 genome. In addition, using 16 transcriptome data (**Table S8**) from different tissues and
232 different conditions, we mapped the reads to Qingke genome by hisat2 v2.1.0[41] and
233 reference-based assembled the 47,490 transcripts by stringtie v1.2.4[42, 43]. The gene
234 information extracted from PacBio Iso-seq, Augustus, Homolog and RNA-seq was
235 integrated using EVidenceModeler[44], according to different weight parameters (**Table**
236 **S9**), to obtain 129,269 putative genes. The same RNA-seq reads were de novo assembled
237 using Trinity v2.4.0[45] to obtain 722,803 transcripts. Structural optimization of EVM
238 integration results was performed using PASA v2.1.0[46] and Trinity assembly results.
239 129,269 transcripts were annotated with NR, Uniprot, and Pfam database. And filter
240 the putative genes according to the function assignment result: (1) NR, Uniprot comparison
241 identity>=50%, coverage>=50%; (2) Not satisfying (1), but protein length >= 50, and
242 containing the Pfam domain. The tandem duplicated genes were identified using MCScanX.
243 Only the longest protein of each subgroup was kept to represent the corresponding gene.
244 The combine Qingke genome annotation included 65,833 genes with names derived from
245 BLAST analysis and genes with domain names derived from Pfam analysis.
246 **High-confidence Genes**
247 Stringent confidence classification was applied to all predicted genes to
248 discriminate between loci representing high-confidence (HC) protein-coding genes and
249 less reliable low-confidence (LC) genes, which potentially consisted of gene fragments,
250 putative pseudogenes and non-(protein)-coding transcripts. We assigned confidence values
251 to a gene model in a two-step procedure by using the similar criteria and methods described

252 previously[10](**Fig. S4**). Finally, we obtained 61,303 stringent confident putative genes, of
253 which 40,457 is HC genes(**Table 3**; **Fig. S4**).
254 Functional annotation of all 61,303 protein sequences was performed using NR,
255 KEGG, SwissProt, Trembl, GO, PFAM, and InterPro databases. 96.57% of all protein
256 sequences were assigned to at least one database(**Table S10**).
257
258 **Non-coding Gene prediction**
259 In the process of annotation of non-coding RNAs, tRNAscan-SE[47] software is
260 used to search for tRNA sequences in the genome according to the structural characteristics
261 of tRNA; since rRNA is highly conserved, the rRNA sequence of related species can be
262 selected as a reference sequence, through BLASTN Alignment to find rRNA in the genome;
263 in addition, using the Rfam family's covariance model, Rfam's own INFERNAL
264 software[48] can predict miRNA and snRNA sequence information on the genome(**Table**
265 **4**).
266 **Annotation evaluation**
267 The BUSCO[26] assessment showed that 91% of the 1,440 genes were completed
268 annotated. And RNA-seq reads rates mapping to gene set also have a significant
269 improvement, from 60% previously annotation to exceed 70%(**Table S8**).
270 In summary, we also acquired a high-quality genome annotation, and it will greatly
271 benefit future function genomics research.

272 Genome database

273 A new Qingke Genome Database(QGD) has been developed to facilitate studies
274 on Qinghai-Tibet Plateau crop plant genomics. The database currently provides access to
275 Qingke genome assembly V1,de novo assembly using NGS sequencing and V2,de novo
276 assembly improvement using combined TGS and NGS. Meanwhile, it also provides access
277 to genome annotation, including repetitive element annotation, ncRNA prediction and
278 protein-coding genes prediction(**Fig. 2**). The QGD embedded JBrowse[49] visual genomic
279 display module to display genome and gene annotation information. One can click on the
280 gene name on the screen to display specific information about the gene. One also can use

online tools Blast v2.4.0+ to enter the query sequence in fasta format, align and return the visual aligned results. In the future, we will focus on optimizing the interactive operation and visualization of the QGD.

You can access to the QGD by visiting http://show.genebang.com/project/index

Discussion

Qingke is a local species of barley and is of great significance to the life and culture of the Tibetan. Its genome and annotation updates help future research on functional genomics, and also help breeders to simplify their work and enrich the genomics resources of barley. The generation of Qingke high-quality reference genomes also demonstrates the superiority of PacBio long reads sequencing technology for larger genome assembly[50]. Although the genome has a long contig, the scaffold is still a bit short at the state-of-art of the current genome assembly, mainly because of the complexity and data volume of the scaffolding mate-pair library is relatively few. In the future, we will use BioNano optical mapping, Hi-C, and akin weapon to complete the genomic construction at the chromosome level[17]. Having an excellent Qingke genomics resources, it will enable geneticists to focus more on functional genomics itself and breeding research.

Availability of supporting data

Raw genomic sequence reads are available in the NCBI Sequence Read Archive under project number **SRS3725794**. Genome assembly and annotation file are available in Qingke Genome Database (QGD) or contact the correspondence author. Supplementary files can be downloaded from the GigaScience Website.

Abbreviations

BLAST: Basic Local Alignment Search Tool;

BUSCO: Benchmarking Universal Single-Copy Orthologs;

QV:quality value;

PacBio: Pacifc Biosciences;

RNA-seq: RNA sequencing;

NGS: Next generation sequencing

TGS: Third generation sequencing

QGD: Qingke Genome Database

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This research was supported by the following funding sources: the Tibet Autonomous Region Financial Special Fund 2017CZZX001, 2017CZZX002, XZNKY-2018-C-021.

We acknowledge (1) Zuoyi Jian, genome assembly executor; (2) Hao Tang, genome annotation pipeline writer; (3) Zhen Zeng ,repetitive annotation executor and (4) NextOmics and BGI-Shenzhen for sequencing supports.

Shan Li, Zhenhua Zhuang and Jie Li also contributed to this Paper.

Author Contributions

Project coordination: X.Z.,X.L.,Y.W.,L.B. and N.T.(Leader).

Sampling and experimentation: X.Z.,Y.W.,H.Y.,Q.X.,S.Z.,Z.W.,W.Q.,Y.B.,J.D.,M.Y..

Genome assembly and Genome annotation guidance and performance：X.L.(Co-leader) and his team.

Data analysis: T.X., X.Z.,Y.W.,H.Y.,Q.X.,S.Z.,Z.W.,W.Q.,Y.B.,J.D.,M.Y..

Genome database construction: T.X. and his team

Data submission: T.X.

Writing: X.L.(Co-leader).

All authors read and commented on the manuscript.

Reference

1. Fu DX, Ruan RW, Dai XM and Liu YM. Astudy of ancient barley, wheat and millet discovered at Changguo in Tibet. 2000.
2. Zeng X, Long H, Wang Z, Zhao S, Tang Y, Huang Z, et al. The draft genome of Tibetan hulless barley reveals adaptive patterns to the high stressful Tibetan Plateau. Proceedings of the National Academy of Sciences of the United States of America. 2015;112 4:1095-100. doi:10.1073/pnas.1423628112.
3. Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Allen AM, et al. Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature. 2012;491 7426:705-10. doi:10.1038/nature11650.
4. International Barley Genome Sequencing C, Mayer KF, Waugh R, Brown JW, Schulman A, Langridge P, et al. A physical, genetic and functional sequence assembly of the barley genome. Nature. 2012;491 7426:711-6. doi:10.1038/nature11543.
5. Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, et al. Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature. 2013;496 7443:91-5. doi:10.1038/nature12028.
6. Ling HQ, Zhao S, Liu D, Wang J, Sun H, Zhang C, et al. Draft genome of the wheat A-genome progenitor Triticum urartu. Nature. 2013;496 7443:87-90. doi:10.1038/nature11997.
7. International Wheat Genome Sequencing C. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science. 2014;345 6194:1251788. doi:10.1126/science.1251788.
8. Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science. 2017;357 6346:93-7. doi:10.1126/science.aan0032.
9. Luo MC, Gu YQ, Puiu D, Wang H, Twardziok SO, Deal KR, et al. Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. Nature. 2017;551 7681:498-502. doi:10.1038/nature24486.
10. Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, et al. A chromosome conformation capture ordered sequence of the barley genome. Nature. 2017;544 7651:427-33. doi:10.1038/nature22043.
11. Zhao G, Zou C, Li K, Wang K, Li T, Gao L, et al. The Aegilops tauschii genome reveals multiple impacts of transposons. Nat Plants. 2017;3 12:946-55. doi:10.1038/s41477-017-0067-8.
12. Ling HQ, Ma B, Shi X, Liu H, Dong L, Sun H, et al. Genome sequence of the progenitor of wheat A subgenome Triticum urartu. Nature. 2018;557 7705:424-8. doi:10.1038/s41586-018-0108-0.

371 13. Appels R, Eversole K, Feuillet C, Keller B, Rogers J, Stein N, et al. Shifting the
372 limits in wheat research and breeding using a fully annotated reference
373 genome. Science. 2018;361 6403 doi:10.1126/science.aar7191.
374 14. Dai F, Wang X, Zhang XQ, Chen Z, Nevo E, Jin G, et al. Assembly and analysis of
375 a qingke reference genome demonstrate its close genetic relation to modern
376 cultivated barley. Plant Biotechnol J. 2018;16 3:760-70.
377 doi:10.1111/pbi.12826.
378 15. Wei Z, Zeng X, Qin C, Wang Y, Bai L, Xu Q, et al. Comparative transcriptome
379 analysis revealed genes commonly responsive to varied nitrate stress in leaves
380 of Tibetan hulless barley. Frontiers in plant science. 2016;7:1067.
381 16. Zeng X, Bai L, Wei Z, Yuan H, Wang Y, Xu Q, et al. Transcriptome analysis
382 revealed the drought-responsive genes in Tibetan hulless barley. BMC
383 genomics. 2016;17 1:386.
384 17. Sedlazeck FJ, Lee H, Darby CA and Schatz MC. Piercing the dark matter:
385 bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 2018;19
386 6:329-46. doi:10.1038/s41576-018-0003-4.
387 18. Yuan Y, Bayer PE, Batley J and Edwards D. Improvements in Genomic
388 Technologies: Application to Crop Genomics. Trends Biotechnol. 2017;35
389 6:547-58. doi:10.1016/j.tibtech.2017.02.009.
390 19. Paajanen P, Kettleborough G, Lopez-Girona E, Giolai M, Heavens D, Baker D, et
391 al. A critical comparison of technologies for a plant genome sequencing project.
392 bioRxiv. 2017; doi:10.1101/201830.
393 20. Boetzer M, Henkel CV, Jansen HJ, Butler D and Pirovano W. Scaffolding pre-
394 assembled contigs using SSPACE. Bioinformatics. 2011;27 4:578-9.
395 doi:10.1093/bioinformatics/btq683.
396 21. Langmead B. Aligning short sequencing reads with Bowtie. Curr Protoc
397 Bioinformatics. 2010;Chapter 11:Unit 11.7.
398 doi:10.1002/0471250953.bi1107s32.
399 22. Boetzer M and Pirovano W. SSPACE-LongRead: scaffolding bacterial draft
400 genomes using long read sequence information. BMC Bioinformatics. 2014;15
401 1:211. doi:10.1186/1471-2105-15-211.
402 23. Xia M, Han X, He H, Yu R, Zhen G, Jia X, et al. Improved de novo genome
403 assembly and analysis of the Chinese cucurbit Siraitia grosvenorii, also known
404 as monk fruit or luo-han-guo. Gigascience. 2018;7 6
405 doi:10.1093/gigascience/giy067.
406 24. Heldenbrand JR, Baheti S, Bockol MA, Drucker TM, Hart SN, Hudson ME, et al.
407 Performance benchmarking of GATK3.8 and GATK4. bioRxiv. 2018;
408 doi:10.1101/348565.
409 25. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al.
410 Versatile and open software for comparing large genomes. Genome Biology.
411 2004;5 2:R12. doi:10.1186/gb-2004-5-2-r12.
412 26. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM.
413 BUSCO: assessing genome assembly and annotation completeness with single-
414 copy orthologs. Bioinformatics. 2015;31 19:3210-2.
415 doi:10.1093/bioinformatics/btv351.

416 27. Paterson AH. Leafing through the genomes of our major crop plants: strategies
417 for capturing unique information. Nat Rev Genet. 2006;7 3:174-84.
418 doi:10.1038/nrg1806.
419 28. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73
420 maize genome: complexity, diversity, and dynamics. Science. 2009;326
421 5956:1112-5. doi:10.1126/science.1178534.
422 29. Benson G. Tandem repeats finder: a program to analyze DNA sequences.
423 Nucleic Acids Research. 1999;27 2:573-80.
424 30. Tarailo-Graovac M and Chen N. Using RepeatMasker to Identify Repetitive
425 Elements in Genomic Sequences. Current Protocols in Bioinformatics. 2009;25
426 1:4.10.1-4..4. doi:doi:10.1002/0471250953.bi0410s25.
427 31. Xu Z and Wang H. LTR_FINDER: an efficient tool for the prediction of full-
428 length LTR retrotransposons. Nucleic Acids Research. 2007;35 suppl_2:W265-
429 W8. doi:10.1093/nar/gkm286.
430 32. Bao W, Kojima KK and Kohany O. Repbase Update, a database of repetitive
431 elements in eukaryotic genomes. Mobile DNA. 2015;6 1:11.
432 doi:10.1186/s13100-015-0041-9.
433 33. Wu TD and Watanabe CK. GMAP: a genomic mapping and alignment program
434 for mRNA and EST sequences. Bioinformatics. 2005;21 9:1859-75.
435 doi:10.1093/bioinformatics/bti310.
436 34. Stanke M and Waack S. Gene prediction with a hidden Markov model and a
437 new intron submodel. Bioinformatics. 2003;19 suppl 2:ii215-ii25.
438 35. Initiative IB. Genome sequencing and analysis of the model grass
439 Brachypodium distachyon. Nature. 2010;463 7282:763.
440 36. Du H, Yu Y, Ma Y, Gao Q, Cao Y, Chen Z, et al. Sequencing and de novo assembly
441 of a near complete indica rice genome. Nature communications. 2017;8:15324.
442 37. McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, et al. The
443 Sorghum bicolor reference genome: improved assembly, gene annotations, a
444 transcriptome atlas, and signatures of genome organization. The Plant Journal.
445 2018;93 2:338-54.
446 38. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize
447 reference genome with single-molecule technologies. Nature. 2017;546
448 7659:524.
449 39. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. Basic local alignment
450 search tool. Journal of molecular biology. 1990;215 3:403-10.
451 doi:10.1016/s0022-2836(05)80360-2.
452 40. Birney E, Clamp M and Durbin R. GeneWise and Genomewise. Genome
453 Research. 2004;14 5:988-95. doi:10.1101/gr.1865504.
454 41. Kim D, Langmead B and Salzberg SL. HISAT: a fast spliced aligner with low
455 memory requirements. Nature Methods. 2015;12:357.
456 doi:10.1038/nmeth.3317
457 https://www.nature.com/articles/nmeth.3317#supplementary-information.
458 42. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT and Salzberg SL.
459 StringTie enables improved reconstruction of a transcriptome from RNA-seq
460 reads. Nature Biotechnology. 2015;33:290. doi:10.1038/nbt.3122
461 https://www.nature.com/articles/nbt.3122#supplementary-information.

43. Pertea M, Kim D, Pertea GM, Leek JT and Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature Protocols. 2016;11:1650. doi:10.1038/nprot.2016.095 https://www.nature.com/articles/nprot.2016.095#supplementary-information.
44. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology. 2008;9 1:R7. doi:10.1186/gb-2008-9-1-r7.
45. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 2011;29:644. doi:10.1038/nbt.1883 https://www.nature.com/articles/nbt.1883#supplementary-information.
46. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic acids research. 2003;31 19:5654-66.
47. Lowe TM and Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic acids research. 1997;25 5:955-64.
48. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. Nucleic acids research. 2014:gku1063.
49. Skinner ME, Uzilov AV, Stein LD, Mungall CJ and Holmes IH. JBrowse: a next-generation genome browser. Genome Res. 2009;19 9:1630-8. doi:10.1101/gr.094607.109.
50. Li C, Lin F, An D, Wang W and Huang R. Genome Sequencing and Assembly by Long Reads in Plants. Genes. 2017;9 1 doi:10.3390/genes9010006.

492 # Tables

493 **Table 1** Comparison of the new genome with previously published assemblies of the Qingke genome.

	This study	Zeng *et al.*[2]	Dai *et al.*[14]
Sequenced genome size (Gb)	4.00	3.89	3.73
Contig N50 (Kb)	1,563.00	18.07	5.94
Scaffold N50 (Kb)	4,006.00	242.00	171.11
Repeat proportion(%)	87.48	81.39	NA
Annotated protein-coding genes	61,303*	36,151	46,787

494 * Only refer to stringent confident putative protein-coding genes.
495
496
497 **Table 2** Repetitive element annotation statistics.

Type	Repeat Size	% of genome
Trf	155,813,415	3.89
Repeatmasker	2,935,487,965	73.27
Proteinmask	19,934,506	0.50
De novo	3,359,143,070	83.85
Total	3,504,452,012	**87.48**

498

499 **Table 3** Protein-coding genes prediction summary.

Method	Software and gene set	Gene number
PacBio	Smrtlink	39,442
	Gmap	14,099
	TransDecoder	13,936
Ab initio	Augustus	128,400
Homology-based	Triticum aestivum A	42,601
	Brachypodium distachyon	33,911
	Hordeum vulgare	46,312
	Oryza sativa	38,426
	Sorghum bicolor	29,909
	Triticum aestivum D	44,897
	Zea mays	29,872
RNA-seq	StringTie	47,490
Integration	EVidenceModeler	129,269
	PASA	129,269
Filter	NR、SwissProt、PFAM	68,018
	After MCScanX moved	65,833
	After TE moved	**61,303**

500

501 **Table 4** ncRNA genes annotation statistics.

Type		Copy	Average length(bp)	Total length(bp)	% of genome
miRNA		5,530	129.27	714,857	0.0178
tRNA		2,477	74.71	185,064	0.0046
rRNA	**Total**	1,583	151.69	240,131	0.0060
	18S	63	1,076.60	67,826	0.0017
	28S	174	133.20	23,177	0.0006
	5.8S	39	147.33	5,746	0.0001
	5S	1,307	109.70	143,382	0.0036
snRNA	**Total**	1,156	115.14	133,097	0.0033
	CD-box	831	103.44	85,960	0.0021
	HACA-box	132	133.48	17,620	0.0004
	splicing	193	152.94	29,517	0.0007

502

503

Figures

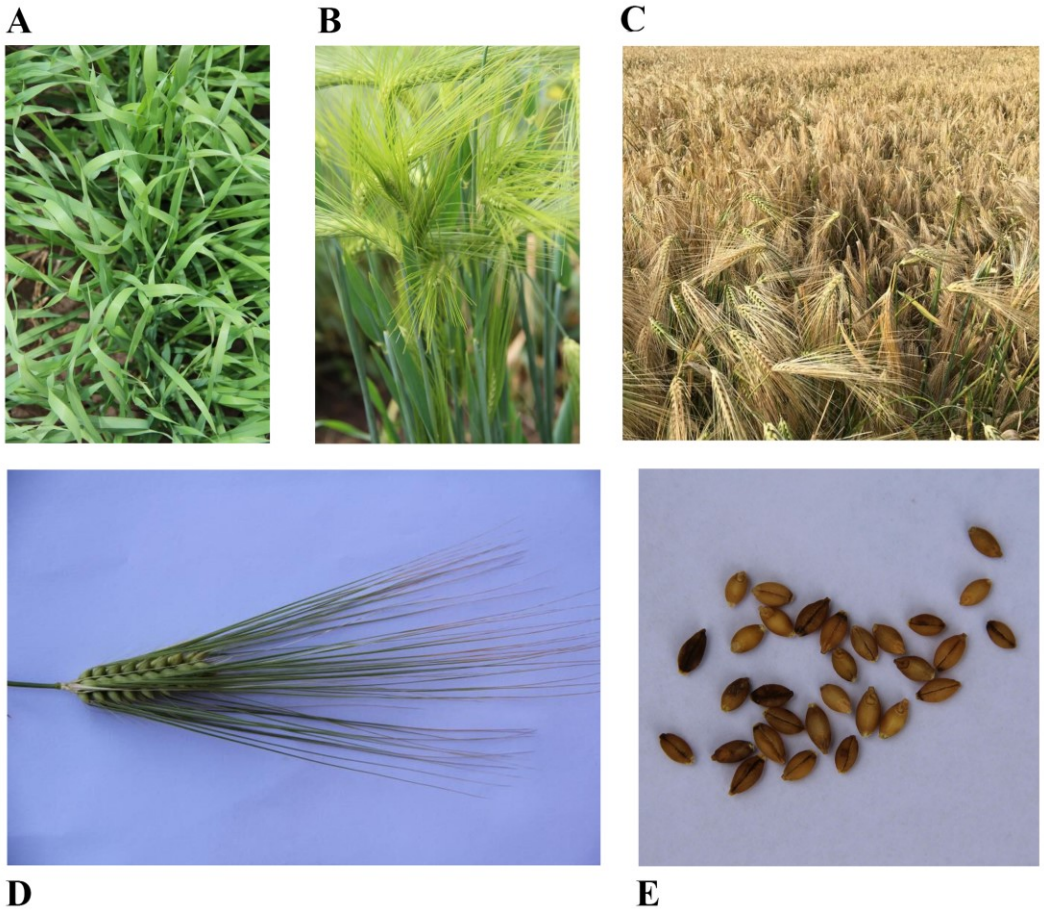


Figure 1 Morphological characteristics of the Qingke, Tibetan hulless barley. These pictures show (A),seedling;(B), heading stage; (C), mature stage;(D), filling stage spike and (E), grain of Qingke.

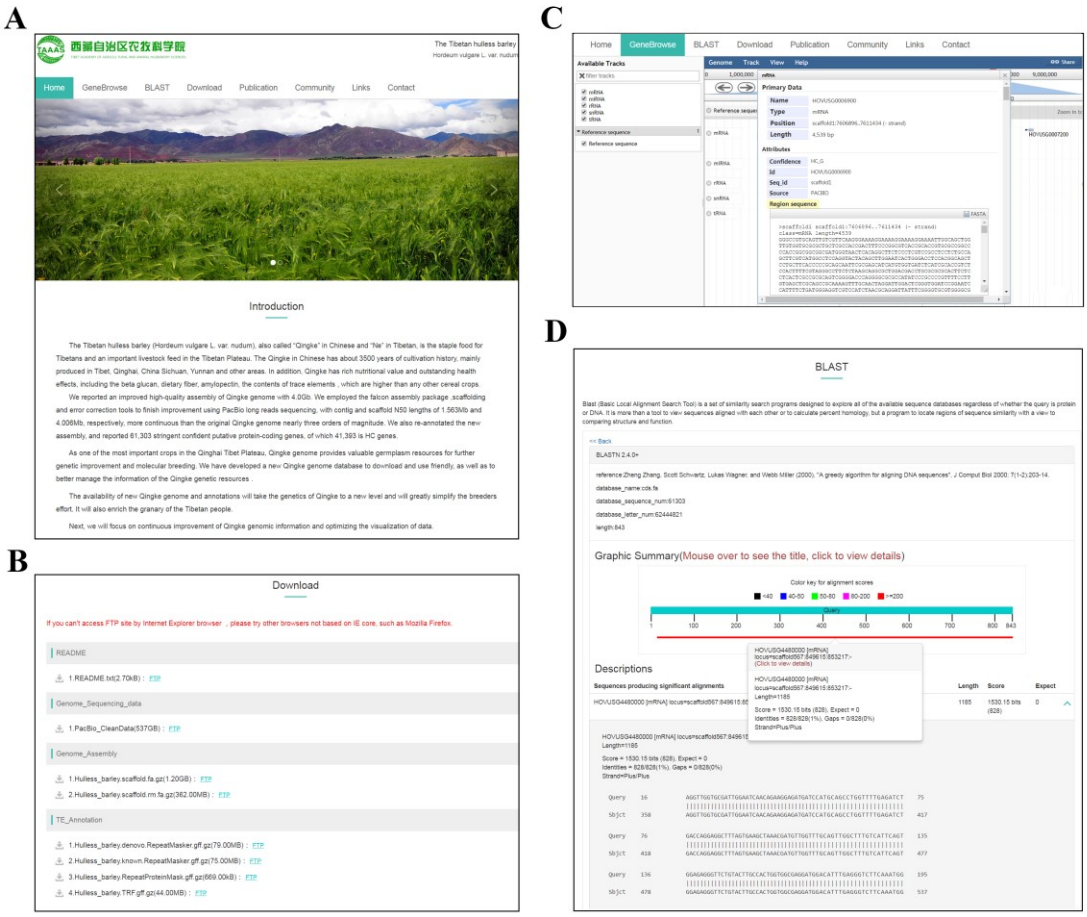
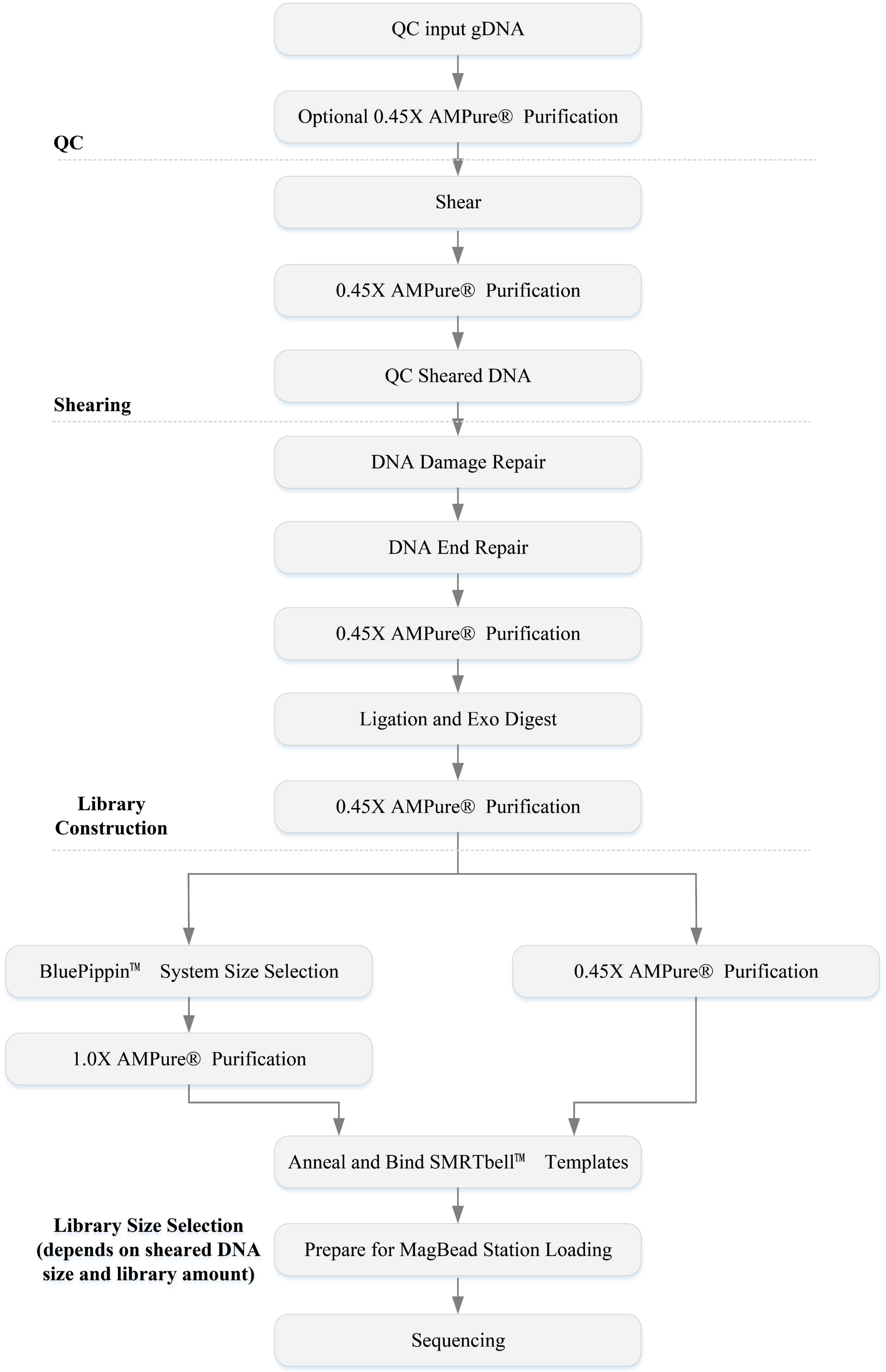
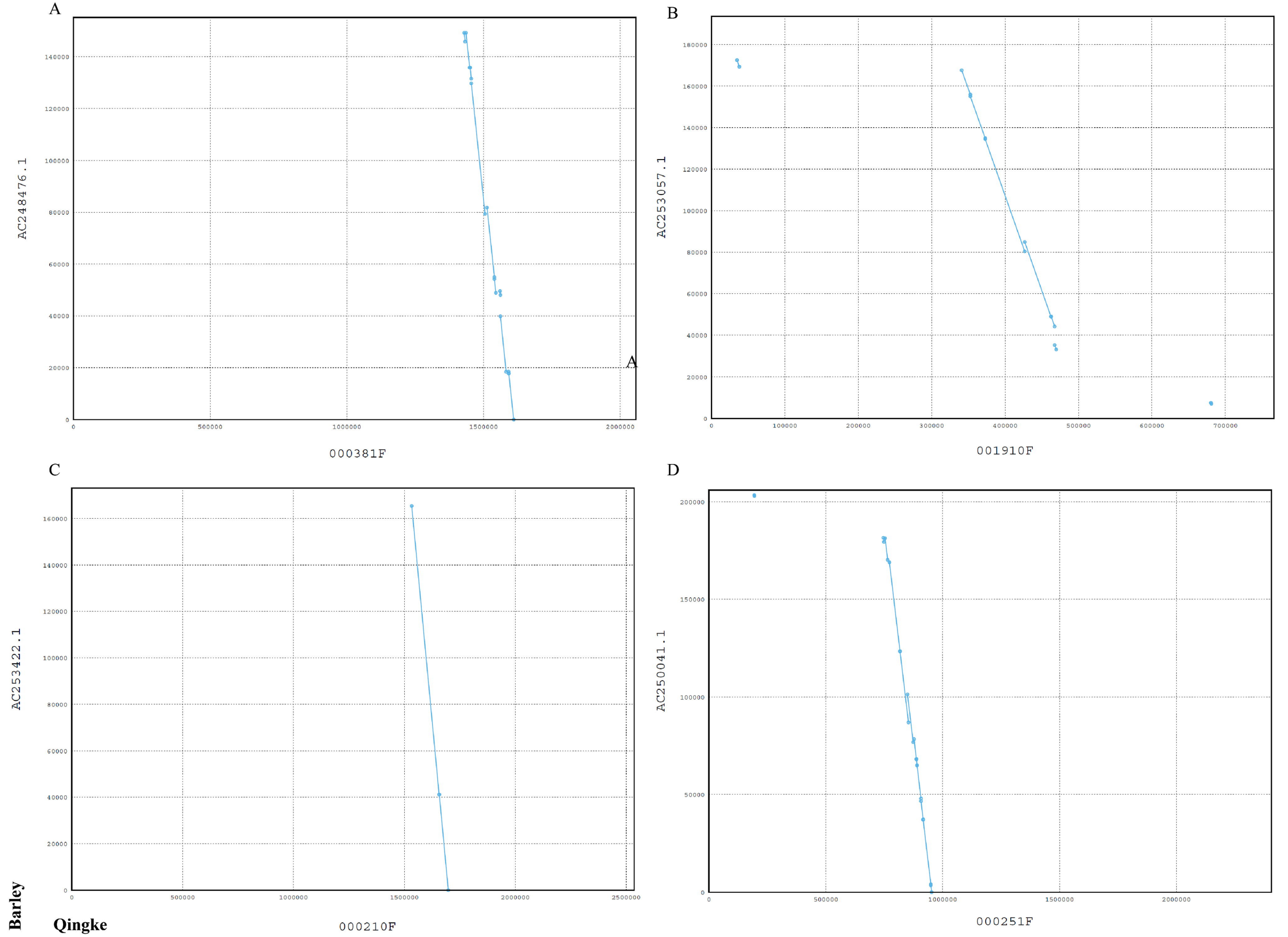
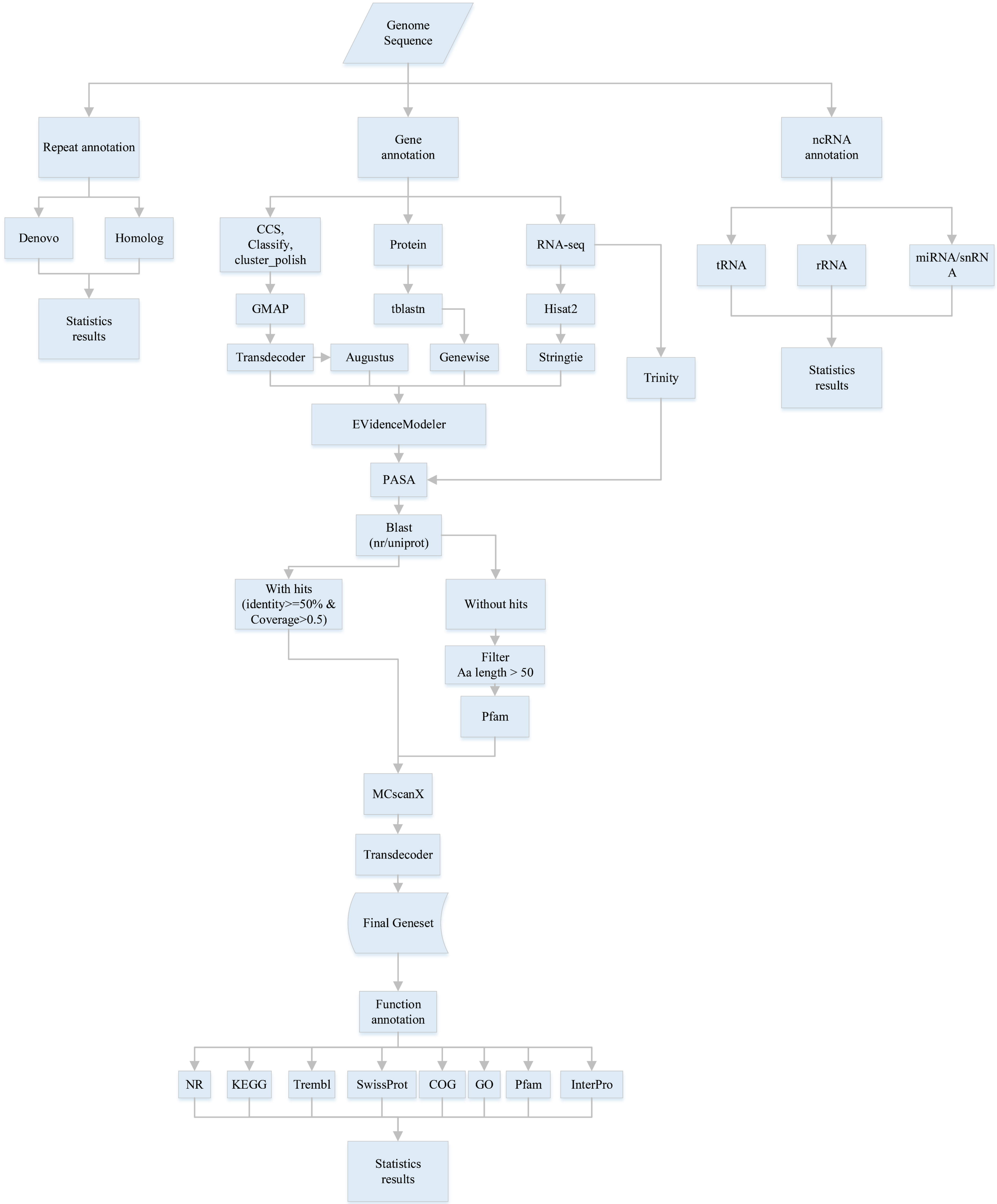
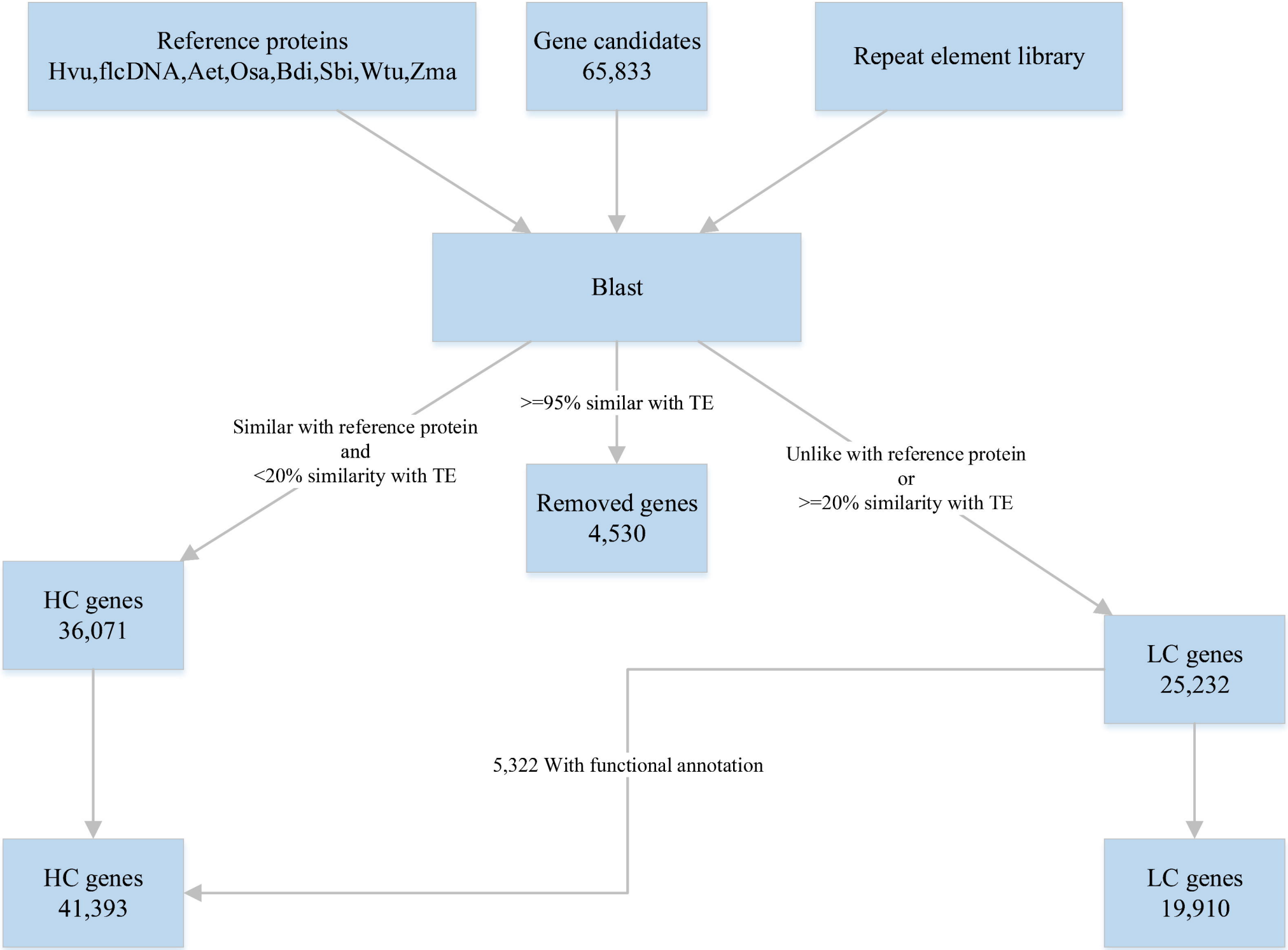


Figure 2 Qingke Genome Database(QGD). (A),the home page of QGD;(B), Data download page;(C),gene browse page and (D),sequences BLAST page.


QC input gDNA
Optional 0.45X AMPure® Purification
QC
Shear
0.45X AMPure® Purification
QC Sheared DNA
Shearing
DNA Damage Repair
DNA End Repair
0.45X AMPure® Purification
Ligation and Exo Digest
0.45X AMPure® Purification
Library Construction
BluePippin™ System Size Selection
0.45X AMPure® Purification
1.0X AMPure® Purification
Anneal and Bind SMRTbell™ Templates
Library Size Selection (depends on sheared DNA size and library amount)
Prepare for MagBead Station Loading
Sequencing


A
AC248476.1
000381F
B
AC253057.1
001910F
C
AC253422.1
000210F
D
AC250041.1
000251F
Barley
Qingke


Genome Sequence
Repeat annotation
Gene annotation
ncRNA annotation
Denovo
Homolog
Statistics results
CCS, Classify, cluster_polish
Protein
RNA-seq
GMAP
tblastn
Hisat2
Transdecoder
Augustus
Genewise
Stringtie
Trinity
EVidenceModeler
PASA
Blast (nr/uniprot)
With hits (identity>=50% & Coverage>0.5)
Without hits
Filter Aa length > 50
Pfam
MCscanX
Transdecoder
Final Geneset
Function annotation
NR
KEGG
Trembl
SwissProt
COG
GO
Pfam
InterPro
Statistics results
tRNA
rRNA
miRNA/snRNA
Statistics results


Reference proteins
Hvu,flcDNA,Aet,Osa,Bdi,Sbi,Wtu,Zma
Gene candidates
65,833
Repeat element library
Blast
Similar with reference protein
and
<20% similarity with TE
>=95% similar with TE
Unlike with reference protein
or
>=20% similarity with TE
Removed genes
4,530
HC genes
36,071
LC genes
25,232
5,322 With functional annotation
HC genes
41,393
LC genes
19,910