Nucleic Acids Research, 2018 **1**
doi: 10.1093/nar/gky522

CGT-seq: epigenome-guided *de novo* assembly of the core genome for divergent populations with large genome

Meifang Qi[1,2,†], Zijuan Li[1,2,†], Chunmei Liu[1,2,†], Wenyan Hu[1,2], Luhuan Ye[1,2], Yilin Xie[1,2], Yili Zhuang[1,2], Fei Zhao[1,2], Wan Teng[2,3], Qi Zheng[2,3], Zhenjun Fan[1,4], Lin Xu[1,2], Zhaobo Lang[2,5], Yiping Tong[2,3,*] and Yijing Zhang[1,2,*]

[1]National Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Shanghai Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 300 Fengalin Road, Shanghai 200032, China, [2]University of the Chinese Academy of Sciences, Beijing 100049, China, [3]The State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China, [4]Henan University, school of life science and [5]National Key Laboratory of Plant Molecular Genetics, Shanghai Center for Plant Stress Biology and Center of Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai 200032, China

Received February 07, 2018; Revised April 11, 2018; Editorial Decision May 23, 2018; Accepted May 24, 2018

ABSTRACT

Genetic diversity in plants is remarkably high. Recent whole genome sequencing (WGS) of 67 rice accessions recovered 10,872 novel genes. Comparison of the genetic architecture among divergent populations or between crops and wild relatives is essential for obtaining functional components determining crucial traits. However, many major crops have gigabase-scale genomes, which are not well-suited to WGS. Existing cost-effective sequencing approaches including re-sequencing, exome-sequencing and restriction enzyme-based methods all have difficulty in obtaining long novel genomic sequences from highly divergent population with large genome size. The present study presented a reference-independent core genome targeted sequencing approach, CGT-seq, which employed epigenomic information from both active and repressive epigenetic marks to guide the assembly of the core genome mainly composed of promoter and intragenic regions. This method was relatively easily implemented, and displayed high sensitivity and specificity for capturing the core genome of bread wheat. 95% intragenic and 89% promoter region from wheat were covered by CGT-seq read. We further demonstrated in rice that CGT-seq captured hundreds of novel genes and regulatory sequences from a previously unsequenced ecotype. Together, with specific enrichment and sequencing of regions within and nearby genes, CGT-seq is a time- and resource-effective approach to profiling functionally relevant regions in sequenced and non-sequenced populations with large genomes.

INTRODUCTION

Comparing to the relatively low polymorphism in humans, which is one single nucleotide polymorphisms (SNP) per 1000–2000 bases on average (1), the genetic diversity in plants is remarkably high (2). Whole genome sequencing (WGS) of 67 rice accessions lead to the identification of 10 872 novel genes different from the reference genome (3); read-depth analysis of WGS data from 103 inbred maize lines revealed that 90% of 10-kb windows showed at least 2-fold variation in read depth (4). This high genetic diversity in plants provides valuable resources for genetic manipulation to achieve desired traits, and it is essential to pinpoint the genetic component determining the traits. The fundamental approach is to make comprehensive comparison of the genetic architecture among divergent population with varied phenotypes or between cultivated species with wild relatives. WGS accelerated profiling of a full spectrum of genetic variation, and revolutionized approaches to molecular genetics, plant breeding, population and evolutionary genetics (5,6). However, both whole-genome doubling (WGD) (i.e. polyploidy) and TE-driven genomic expansion are prevalent in plant, resulting in some extremely

*To whom correspondence should be addressed. Tel: +86 21 54924206; Fax: +86 21 54924015; Email: zhangyijing@sibs.ac.cn
Correspondence may also be addressed to Yiping Tong. Tel: +86 10 64806566; Fax: +86 10 64806537; Email: yptong@genetics.ac.cn
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

large and complex genomes (7). Bread wheat (*Triticum aestivum*) has an allohexaploid (AABBDD) genome around 17 Gb in size (8–10). Genome sequences of the most widely cultivated barley, cotton, maize, peanut and soybean varieties are all on a gigabase-scale and not well-suited to WGS of individuals at the population scale. Given that large proportion of the genomes are composed of repeat-rich and non-informative sequences, not every base needs to be sequenced. Multiple cost-effective alternatives were developed, but generally have limitations in identifying novel gene and regulatory sequences from divergent population or unsequenced species.

For organisms without a reference genome, restriction-enzyme based methods—for example, restriction site associated DNA sequencing (RAD-seq)—are typically used to identify variations flanking restriction sites. However, only short fragments in the vicinity of restriction site are recovered. And when the genome is large, sequencing reads from this method are mostly in repetitive regions (11). Also, the genetic diversity associated with polymorphism in restriction sites tends to be underestimated. This becomes more pronounced when polymorphism is high (12). In some cases, RNA-seq was applied for genotyping of unsequenced organisms (13). One typical example is the bulk segregant RNA-seq (BSR) developed for gene mapping in populations without genetic markers (14). However, comparison of RNA-seq data in unsequenced population is challenging as differences in splicing, gene expression and allele-specific expression add another source of variation to the allele counts (11,15). For well-sequenced species, whole-exome sequencing (WES) is a popular cost-effective alternative approach. It captures the highly interpretable coding region and reduces the sequencing space (16–18). However, the production of a high-quality probes requires a substantial investment of resource, including the probe design, synthesis, infrastructure and personnel expertise (11). In addition, the markers are specific to the reference genome (19). For population considerably divergent from the reference, WES and even whole genome re-sequencing may miss important genetic variations, novel genes and regulatory sequences determining phenotypic difference. This issue is more apparent when comparing non-sequenced wild species to trace domestication pathways, or for evolutionary and ecological studies. Also, non-coding sequences containing important regulatory information could not be captured by WES.

Thus, an ideal alternative approach for studies in divergent population and wild species would be to directly capture the regulatory and coding sequences. Inspired by the fact that promoter and gene body regions are enriched for specific types of epigenetic marks, we developed core genome targeted sequencing (CGT-seq). The principal goal was to enable specific capturing and re-sequencing of the core genome, composed of gene body and promoter regions, through enrichment of epigenetic marks preferentially located surrounding genes. This method is independent of any prior sequence information, thus avoiding the reference bias, and has the potential to be broadly applicable to both sequenced and unsequenced organisms.

MATERIALS AND METHODS

Plant materials and growth conditions

The bread wheat (*Triticum aestivum*) genotype 'Chinese Spring' ('CS') was used. Seeds were sterilized by 30% H_2O_2 for 10 min followed by 5 times washes with distilled water. The sterilized seeds were germinated in water for 3 days at 22°C. Germinated seeds with residual endosperm were transferred to soil. After 9 days in long-day conditions, the ground and underground parts were harvested.

Two rice cultivars, Huanghuazhan (HHZ, *indica*) and Jizi-1560 (JZ-1560, *japonica*), were used. The rice seeds were germinated in sterilized water, and were then placed in soil in an experimental field of the Institute of Plant Physiology and Ecology (Shanghai, China) under natural growing conditions during kharif season of 2013. Panicles were harvested.

Harvested wheat and rice samples were either frozen in liquid nitrogen for RNA isolation or directly vacuum-infiltrated with formaldehyde cross-linking solution for ChIP assay.

ChIP and RNA sample preparation

For wheat samples, ChIP assay was performed with the antibodies against H3 trimethyl-Lys 27 (Upstate, USA, Cat. 07-449) and H3 trimethyl-Lys 4 (Abcam, Cat. ab8580). For rice samples, ChIP assay was performed with the antibody against H3 trimethyl-Lys 4 (Abcam, Cat. ab8580), H3 monomethyl-Lys 4 (Millipore, 07-436), H3 trimethyl-Lys 27 (Upstate, USA, Cat. 07-449), H3 acetyl-Lys 27 (Upstate, USA, Cat. 07-360) and H3 trimethyl-Lys 36 (Abcam, Cat. ab9050) as previously described (20). More than 10 ng ChIP DNA or 2 μg total RNA from each sample was used. Library construction and deep sequencing were performed by genenergy Biotechnology Co. Ltd. (Shanghai, China). The DSN kit from Illumina was used for DSN library normalization. Libraries were sequenced on the HiSeq2000/2500 (Illumina) to produce 50 bp single-end reads for rice, and 150 bp paired-end reads for wheat.

Public ChIP-seq, RNA-seq and DHS data collection

All sequencing data generated in the present study and from public databases were summarized in Supplemental Table S1. RNA-seq data sets from ten tissues of rice including seedlings, leaf, shoot, endosperm, embryo, inflorescence, anther, pistil, panicle and seed were downloaded from the Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/geo/) under accession numbers SRP002417, SRP001787, SRP004651, SRP008821. 30 808 genes expressed (reads per kilobase per million mapped reads (RPKM) > 1) in at least one of the ten tissues were defined as high confidence rice gene.

38 RNA-seq data sets from seven tissues of bread wheat including shoot, leaf, stem, root, inflorescence, endosperm, grain were downloaded from the SRA under accession numbers SRP028357, SRP041022, ERP004505, ERP004714 and DRP000768. ChIP-seq data of H3K4me3, H3K27me3, H3K27ac, H3K36me3 and H3K4me1 from

Arabidopsis seedlings were from GEO under accessions GSE79259, GSE68370, GSE75071, GSE80056.

Rice DHS data generated from seedlings and callus were collected from GEO under accession number GSE26610. ChIP-seq data of transcription factors OsASR2, bZIP23, OsNAC6, TGAP1, GRF6, NF-YB1, SDG711, OSH1.1 and SRT1 were downloaded from SRP112505, SRP075204, SRP074236, DRP001345, SRP063301, SRP062590, SRP061969, DRP000207 and SRP014840.

Sequencing data processing

Sequencing reads were cleaned with Trim Galore v0.4.4 and sickle, including removing bases with low quality score (<25) and irregular GC content, and cutting sequencing adaptors followed by filtering short reads. The cleaned reads were mapped to genomes of *Arabidopsis thaliana* (TAIR10 release), *japonica* rice (MSU7.0 release), *indica* rice R498 (21), and bread wheat (IWGSC RefSeq v1.0) using BWA 0.7.5a-r405 (22) for ChIP-seq and re-sequencing data, and HISAT2 2.1.0 (23) for RNA-seq data, all with default settings. For saturation assessment to determine sequencing depth, 10, 20, 30, 40 and 50 Gb reads were randomly drawn from the raw reads of H3K4me3 and H3K27me3 ChIP-seq data in wheat. For ChIP-seq data, MACS1.3.7 (24) was used to identify read-enriched regions (peaks) with combined criteria: P value < 1e–50 and fold-change >32. Target genes were defined as genes with a peak within or nearby the gene body (±2 kb).

De novo assembly of the core genome based on ChIP-seq data sets

The cleaned reads from wheat were assembled using ABySS v2.0.2 (25). All paired-end sequencing reads were mapped back to the contigs using BWA 0.7.5a-r405 (22), which were used as input for BESST (version 2.2.7) with insert size 200 bp (26) to join contigs into scaffolds. GapCloser (version 1.12) (27) was applied to further fill the gaps with information from the reads mapped to scaffolds. Finally, in-house script modified from SSPACE (28) was used to extend the assembled scaffolds based on reads mapped to the boundary of the scaffolds, with the fraction of shared nucleotides set to 0.95, and the lowest coverage set to 5.

For assembly in rice, cleaned reads from all modifications were assembled directly using Velvet (version 1.2.10), which is more efficient for short-read assembly, with minimum coverage set to 4, minimum scaffolds length set to 200 bp (29), and kmers set to 27 and 31. Next, both results were merged to get improved results. Given that the read from rice is single end, no further scaffolding or gap filling was performed.

Integrative Genomics Viewer (IGV) was used for illustrating the genomic tracks (30). Circos plot was drawn using Circos (version 0.69-6) (31).

Calculation of mapping accuracy and single nucleotide variation (SNV) calling

For stringent calculation of the accuracy of the scaffolds mapped to the reference genome, we combined the information from the NM tag of SAM file, which represents the edit distance to the reference, including ambiguous bases (i.e. mismatches and INDELs) but excluding clipping, and the soft clip information from CIGAR string of the SAM file. Scaffolds with (NM + soft clipped bases) <1% of the total length were kept.

SNVs were called with SAMtools (version 1.3.1) (32) and BCFtools (version 1.4) (33) with P value cutoff set to 1 for CGT-seq assembled scaffolds, and default setting for re-sequencing data. Next, low-quality variants were filtered out with combined criteria: mapQ score $\geq$ 20, ratio of SNVs $\geq$ 0.8, homozygous, and max alleles = 2 for CGT-seq; an additional criterion of at least five reads covering an allele was set for re-sequencing.

Definition of *indica* rice specific region

The genome sequence of *indica* rice R498 was divided to 100 bp non-overlapping regions. Next, these 100 bp sequences were mapped to *japonica* rice reference genome (MSU7.0). Unmapped regions were defined as *indica* rice specific regions.

Detection of CpG island

The CpG island (CGI) is detected by R package makeCGI (34), which fits two HMMs on GC content and the ratio of observed vs. expected CpG (obsExp) iteratively. CGI is defined using 0.99 as posterior probability threshold by default. All promoter regions were classified to CGI promoters or non-CGI promoters based on whether overlap with the defined CGIs.

RESULTS

Design of CGT-seq

To select appropriate epigenetic marks for capture, we started by characterizing the genomic distribution of different epigenetic marks. ChIP-seq was performed for the major epigenetic marks surrounding genes in monocot model plant rice. Different modifications preferentially locate in different genic regions and mark different sets of genes (Figure 1A). H3K4me3 tends to be present surrounding transcription start site (TSS), while other modifications occur mostly in the gene body. Similar profiles were observed in other monocot (wheat and maize) and dicot (Arabidopsis and tree cotton) plants (Supplemental Figure S1), indicating the binding profiles are relatively conserved in higher plants. H3K27me3 preferentially marks repressed genes, while other modifications are associated with active expression (Figure 1B and Supplemental Figures S2 and S3). Of 30 808 high confidence rice genes, 89% are enriched for at least one of these modifications (Supplemental Figure S2B). Combinations of these marks provide useful resources for capturing both intragenic and regulatory sequences of genes with a broad range of expression levels. Accordingly, we designed an experimental and computational pipeline for reconstruction of the core genome (pipeline shown in Figure 1C). Briefly, ChIP-seq of different marks was performed, sequencing reads of each modification were assembled and further merged. Next, reads from all modifications were mapped to the assembled contigs derived from the previous

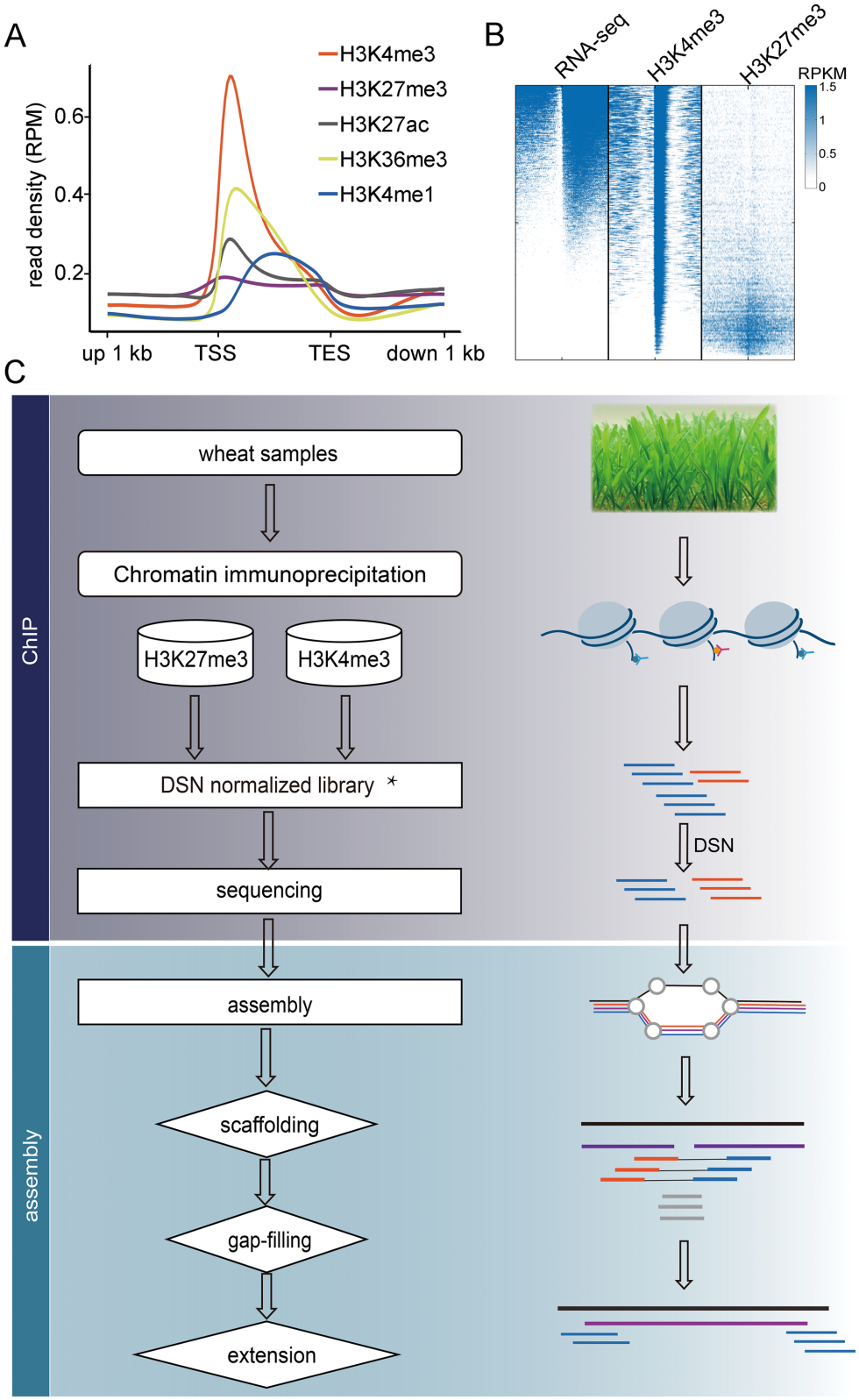

Figure 1. Design of CGT-seq. (**A**) ChIP-seq read distribution of epigenetic marks surrounding genes in rice (*japonica* cultivar JZ-1560). Shown along the y axis is the read density normalized by the sequencing depth (RPM, read per million mapped read). Regions ranging from 1 kb up- to 1 kb downstream of gene body were shown. TSS, transcription start site. TES, transcription end site. (**B**) RNA-seq and ChIP-seq read density of H3K4me3 and H3K27me3 marks surrounding 33 808 genes expressed in at least one tissue. Regions ranging from 6 kb up to 6 kb downstream of TSS was used. (**C**) Workflow for enrichment and de novo assembly of the core genome. ChIP-seq was performed for selected epigenetic marks, followed by DSN normalized library and massively parallel sequencing. The sequencing reads were assembled to contigs based on De Bruijn graph. Contigs from different modifications were merged together and scaffolds were constructed with paired-end information. Further gap filling and extension were guided by reads mapped to the scaffolds. *DSN normalized library was introduced as an improvement of uniformity in discussion.

step, and paired sequencing reads were used to join contigs into scaffolds. All reads were re-mapped to the scaffolds, gaps were filled and scaffolds were extended (see Materials and Methods). All scripts are available on CGT-seq website: http://bioinfo.sibs.ac.cn/zhanglab/cgt-seq/index.htm.

Sensitivity and specificity of CGT-seq

To evaluate the performance of CGT-seq in organisms with large genomes, we generated ChIP-seq data of the active mark H3K4me3 and the repressive mark H3K27me3 in bread wheat (*Triticum aestivum*) genotype 'Chinese Spring' ('CS'), whose genome size is 17 Gb and has been well-sequenced and annotated. A total of 2,874,277 scaffolds were assembled from 100 Gb of sequencing reads (50 Gb from each modification), and 93% of the assembled scaffolds could be mapped with high accuracy to the reference genome of bread wheat (IWGSC RefSeq v1.0) (see Methods and Supplemental Figure S4). The ChIP-seq data and the scaffolds mapped to reference genome could be visualized via the link to genome browser (http://bioinfo.sibs.ac.cn/browser/cgt-seq), and the statistics are summarized in Supplemental Table S1. The Circos plot in Figure 2A and the genomic tracks in Figure 2B and Supplemental Figure S5 illustrate the high coincidence between the assembled sequences and functional important genes in wheat. The NAC transcription factor no apical meristem (NAM) plays crucial role in controlling grain nutrient content (35). Here, CGT-seq successfully assembled both promoter and gene body regions of the homologous genes from A, B and D genomes with 100% identity (Figure 2B and Supplemental Figure S5). The essential parameters for measuring the performance of targeted sequencing include sensitivity and specificity (36). Sensitivity is the percentage of the expected target bases that are represented by at least one of the captured reads. 95% bases from gene body and 89% bases from promoter region could be covered by at least one sequencing reads (Figure 2C). For well-sequenced species, the genetic architecture could be directly profiled from the mapping result. For organisms with genome unknown or plant variety whose sequence is highly divergent from the reference genome, assembly is recommended. The length of each gene covered by assembled scaffolds was calculated. The assembled scaffolds could cover $\geq$ 1 kb genic regions surrounding over 55% of annotated genes, and $\geq$ 500 bp genic regions surrounding over 75% genes (Figure 2D and Supplemental Table S2). We next calculated the sequencing depth of different genomic regions recovered by CGT-seq. The promoter and intragenic regions displayed median coverage of 24–67-fold, while the intergenic regions only have 14-fold coverage (Figure 2E). Specificity is the percentage of scaffolds that map to the intended targets. 70% H3K4me3 scaffolds and 27% H3K27me3 scaffolds localized within or nearby genes, which show 5.4-fold and 2.0-fold enrichment relative to genomic background. For H3K4me3 captured intergenic sequences based on current version of annotation, 16% could be mapped by RNA-seq reads, significantly higher than the genomic background (1%), indicating the coding potential of these regions (Figure 2F). On the other hand, the fraction of intergenic TEs and repetitive sequences, majority of which localized in heterochromatin and are supposed to have no direct role in gene activity, is 2.2-times lower in H3K4me3 captured regions as compared to the genomic background. Together, CGT-seq displayed high sensitivity and specificity for enrichment of core genome regions.

Sequencing read saturation analysis

There is a trade-off between the sequencing depth and target regions detected. We applied random sampling to determine the depth that ensures the assembly performance. 40 Gb sequencing reads (20 Gb from each modification) could reach similar sensitivity as using 100 Gb sequencing reads, with assembled contigs covering $\geq$ 500 bp sequences from 64% genes (Figure 3). The median coverage for captured genic region is $>20\times$ when the sequencing depth is 40 Gb for CGT-seq (Supplemental Figure S6), while the same sequencing depth only ensures 2X coverage for whole-genome re-sequencing, which is far from an accurate identification of polymorphic loci. Thus, CGT-seq can significantly decrease the sequencing space and cost, while maintaining sufficient depth in target regions.

Recovery of novel intragenic and regulatory regions in unsequenced rice variety

To evaluate the performance of CGT-seq in identifying novel functional sequences and variants, we performed CGT-seq in *indica* rice variety Huanghuazhan (HHZ), and mapped to the well-assembled reference genome of *japonica* rice variety *Nipponbare* (MSU version 7.0). There are about 9253 HHZ scaffolds unmapped to MSU7.0. We assume that a large proportion of these unmapped scaffolds are specific to *indica* rice. A relatively high-quality reference genome for *indica* rice variety R498 was recently published (21). We compared the genome sequences between R498 (*indica* variety) and *Nipponbare* (*japonica* variety), and identified 26 965 R498 sequences highly divergent from *Nipponbare*. For the 9253 unmapped scaffolds, 26% were uniquely mapped to R498, located in 352 gene-body regions and 457 promoter regions. The top enriched domains of these genes include NB-ARC and LRR (Figure 4A and Supplemental Table S3), which are actively involved in plant immunity and showed high diversity between different ecotypes. Genomic track in Figure 4B illustrated the CGT-seq recovered LRR gene, which is located in region highly divergent between the two rice ecotypes. Another typical example is *qSW5/GW5* locus, which exerts the greatest effect on rice grain width and weight and contains a 1212-bp deletion in most japonica races (37). CGT-seq assembled contigs covered >3 kb regions surrounding this locus (Figure 4C). Fifty base pairs single-end reads were used, and longer scaffolds are expected to be obtained if using paired-end and longer sequencing reads. Taken together, CGT-seq has the potential to recover long novel sequences near genes in divergent population. This could facilitate fine-mapping in GWAS and QTL analysis for identifying casual regions.

Accurate detection of single nucleotide variations in heterogeneous regions

To evaluate our method in single nucleotide variations (SNV) detection, we performed both re-sequencing and

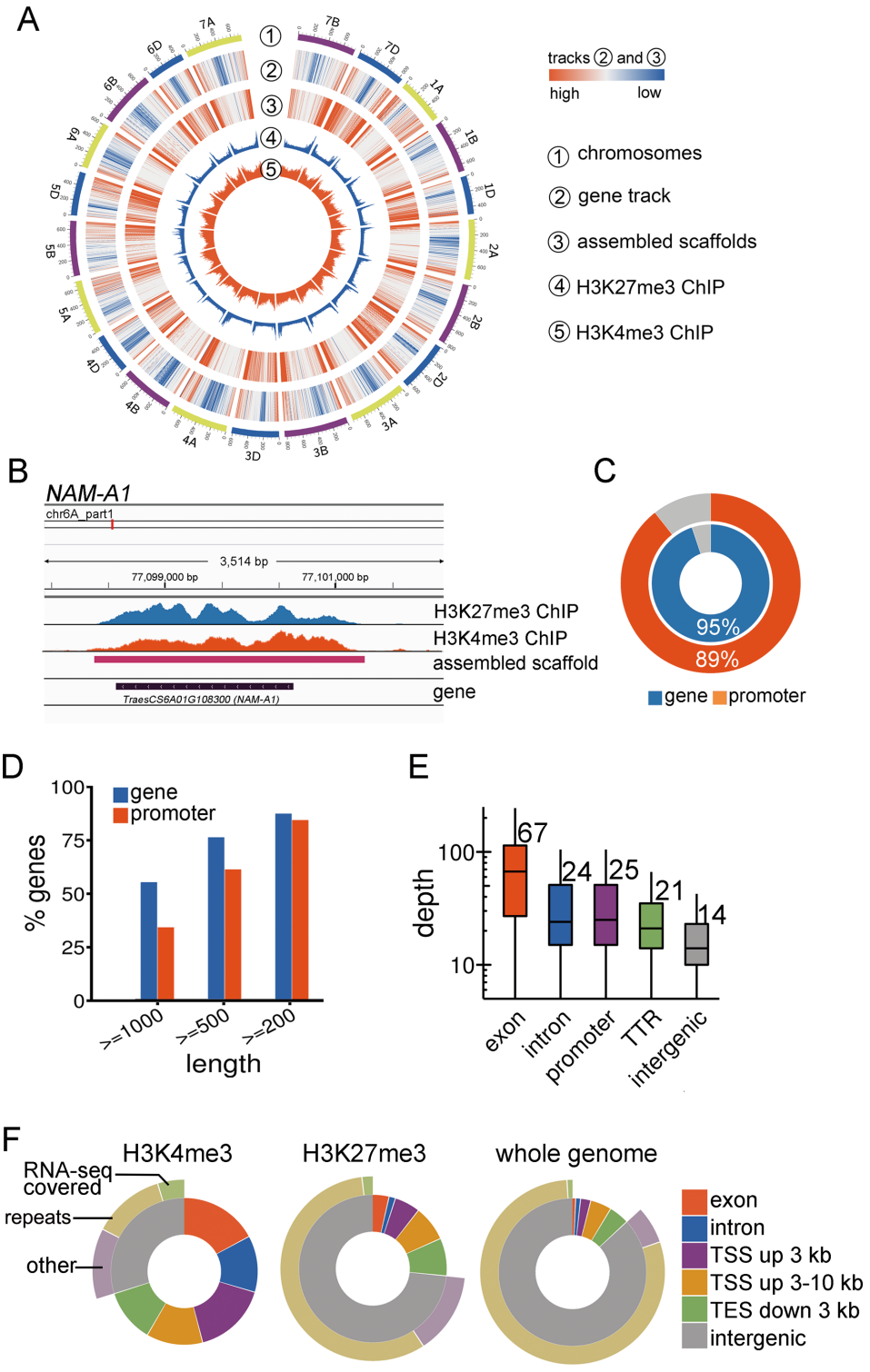

Figure 2. Performance of CGT-seq. (**A**) Circos plot showing the high concordance between wheat genic regions and regions captured by CGT-seq enriched for H3K4me3 and H3K27me3 marks. The outermost circle depicts the ideograms of each chromosome. The rules indicate the length and position of each chromosome. The next two outermost circle represents the density of genes (track 2) and assembled scaffolds (track 3). Orange indicates high density and blue indicates low density. The two internal circles represent H3K27me3 (track 3) and H3K4me3 (track 4) ChIP-seq read density. (**B**) Genomic track illustrates the recovery of promoter and intragenic regions by CGT-seq for a representative gene *NAM-A1*. (**C**) Donut chart showing the fraction of all 110 790 annotated genes and promoter regions (TSS up 3 kb) covered by CGT-seq sequencing reads. (**D**) Fraction of annotated genes whose sequences are recovered by assembled scaffolds. X-axis represents the length of sequences in promoter (3 kb upstream of TSS) or gene body regions covered by the scaffold. Y-axis represents the fraction of annotated genes. (**E**) Box plot showing the distribution of sequencing depth in different genomic regions captured by CGT-seq. Promoter region is defined as above. Transcription termination region (TTR) is defined as 1 kb downstream of TES. The numbers on top of the box represent the median depth. (**F**) Fraction of scaffolds mapped to different genomic regions. Intergenic regions based on current annotation were further divided to those mapped by RNA-seq read, TEs and repetitive sequences, and other regions.

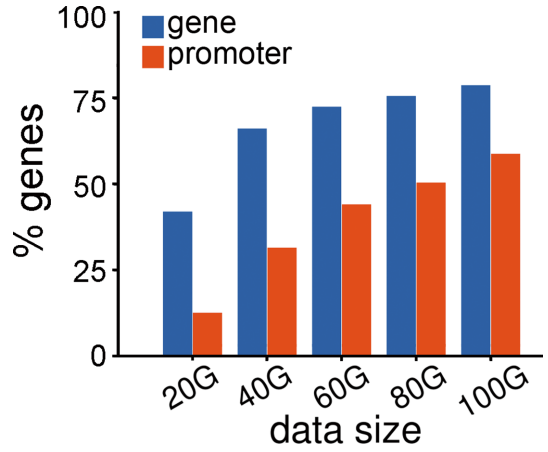

Figure 3. CGT-seq read saturation analysis. Assessment of capturing sensitivity for randomly sampled reads. 20, 40, 60, 80 Gb sequencing reads (half from H3K4me3 and half from H3K27me3) were randomly selected from 100 Gb sequencing reads. Shown is the fraction of annotated genes with ≥500 bp genic (blue) or promoter (orange) regions recovered by assembled scaffolds from sampled reads.

CGT-seq in HHZ (*indica* rice variety), mapped to the well-assembled *japonica* rice reference genome (MSU7.0), called variants and compared the variants identified in CGT-seq captured regions. A total of 847 426 SNVs were detected by re-sequencing in the targeted regions of CGT-seq, 98.3% of which were also identified in CGT-seq captured sequences (Figure 4D). There are 203,043 CGT-seq specific variants, which tended to have lower coverage in re-sequencing (Figure 4E). We hypothesized that re-sequencing reads are poorly mapped to regions with high polymorphism. Plotting the distribution of re-sequencing reads in regions with different levels of polymorphism produced an expected distribution (Figure 4F). CGT-seq specific variants in regions with high level of polymorphism were confirmed through PCR and Sanger sequencing for four randomly selected loci (Figure 4G and Supplemental Figure S7). These results indicate that CGT-seq is capable of capturing specific positions with high polymorphism, which may be missed by reference-dependent approaches, allowing for sensitive variant calling in these regions.

High ratio of allelic regions captured from different rice ecotypes

Besides the accurate identification of novel sequences and variants, successful determination of the causal component requires the polymorphic region to be allelic across samples. Epigenetic marks are relatively stable unless change is triggered by specific developmental or environmental cues, thereby help keeping internal homeostasis (38). We examined the overlap of genomic regions recovered by H3K36me3 in HHZ and JZ-1560, rice lines from rice subspecies *indica* and *japonica* respectively. 91% of captured sequences in HHZ and 78% of sequences in JZ-1560 overlapped with each other (Figure 5), indicating that the epigenetic marks are relatively stable between subspecies, and thus the captured sequences are readily comparable across individuals and across subspecies.

High concordance with chromatin open regions and transcription factor binding loci

It is worth noting that the ChIP-seq reads of major histone modifications are enriched for *qSW5/GW5* locus, indicating this locus is potentially regulated by epigenetic machinery. Indeed, given the essential role of epigenetic modifications in transcriptional regulation, after identification of a large number of trait-associated variants via association studies, employment of epigenomic data can be an efficient way to deduce direct causality (39,40). A systematic survey of human GWAS variants revealed that ∼70% of common disease-associated variation lie within or nearby DNase I hypersensitive site (DHS), predominantly overlapping a transcription factor recognition sequence (41). The regions captured by CGT-seq showed high coincidence with open chromatin regions and transcription-factor binding sites (Figure 6). Thus, although the captured sequences did not cover the whole genic regions for most genes, they contain important functionally relevant information, thereby facilitating subsequent pinpointing of variations and genes directly determining major agronomic traits.

DISCUSSION

CGT-seq is particularly suited to divergent populations with large genome as compared to existing cost-effective methods

In species with gigabase-scale genomes, WGS or re-sequencing is not practical for population-level study, and target enrichment strategy is needed (36). Both WES and CGT-seq aim to enrich genic regions to reduce sequencing space. For studies focusing on coding regions in well-sequenced varieties, the specificity and sensitivity of CGT-seq is not necessarily superior than WES, except that CGT-seq is relatively easily implemented. However, the following features of CGT-seq make it particularly appealing for divergent populations, wild species and non-model organisms lacking reference genomes. Firstly, CGT-seq-based assembly is independent of any prior reference information. Secondly, the regulatory region and gene structure could be recovered, facilitating gene cloning, exploration of regulatory mechanism and identification of casual variation. Thirdly, the longer assembled sequence ensures more accurate identification of SNVs and novel regulatory or genic regions. CGT-seq also overcomes the issues presented in previous restriction-enzyme based approach, which always results in short and functionally irrelevant sequences when applied to large-genome organisms, and is incapable of characterizing loci with high polymorphism.

Other methodologies for removing repetitive sequences in order to enrich for genic regions were proposed previously. Typical example includes the integration of methylation filtering and high C_0t (HC) selection for enrichment of gene-coding sequences in maize by removing highly methylated and repetitive regions before sequencing (42–44). Both reduced representation strategies are highly cost-effective. However, the application of methylation filtration is limited by the fact that in addition to transposon and repetitive regions, genebody of most well-studied plant species are normally hyper-methylated (45). The prevalent usage of high C_0t analysis is restricted by the requirement of deep under-

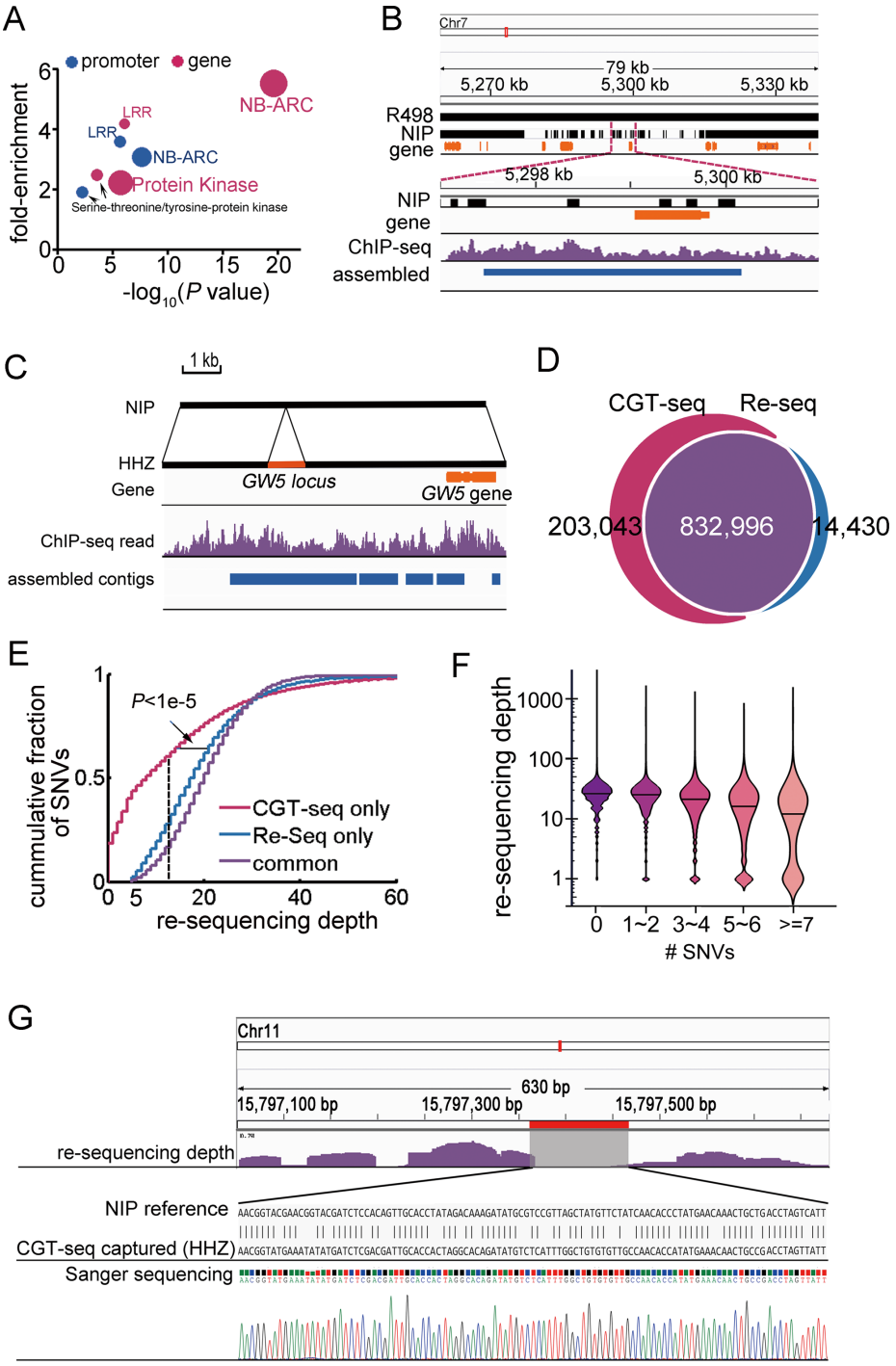

Figure 4. High accurate detection of intragenic and regulatory sequences and variants in heterogeneous regions. (**A**) Enriched protein domains for indica-japonica divergent genes recovered by CGT-seq in HHZ. Shown are enrichment *P* value (x-axis) and fold enrichment (y-axis) for genes with promoter (blue) or intragenic (pink) regions recovered. The size of the circle represents the number of recovered genes contain given domain. (**B**) Genomic tracks illustrate the recovery of the indica specific LRR gene by CGT-seq from HHZ in indica-japonica highly divergent region. Fifty base pairs single-end sequencing reads were used for assembly. (**C**) Genomic tracks illustrate the recovery of the indica rice specific GW5 locus by CGT-seq. Fifty base pairs single-end sequencing reads were used for assembly. (**D**) Concordance of SNVs identified by re-sequencing and identified from CGT-seq captured sequences. (**E**) Cumulative fraction of SNVs (y-axis) with re-sequencing depth less than or equal to the value on the x-axis. The re-sequencing only SNVs, CGT-seq only SNVs and common SNVs are plotted separately. The *P* value reflecting the significance of differential distribution is calculated based on Kolmogorov–Smirnov test. The dashed line indicates that >80% re-sequencing only and common SNVs have re-sequencing depth >10, while only around 45% CGT-seq only SNVs have re-sequencing depth >10. (**F**) Distribution of re-sequencing depth in respect to the density of SNVs. All 20 bp genomic regions harboring SNV(s) were collected and grouped by the number of SNVs. (**G**) PCR validation of CGT-seq only SNV region in HHZ. The grey area represents the region selected for PCR validation. The dark purple bars on the first track represents re-sequencing depth; the second track represents the pair-wise sequence comparison between Nipponbare reference and CGT-seq captured sequence in HHZ; the third track is the Sanger sequencing result. PCR results for other three randomly selected regions harboring CGT-seq only SNVs are shown in Supplemental Figure S7.

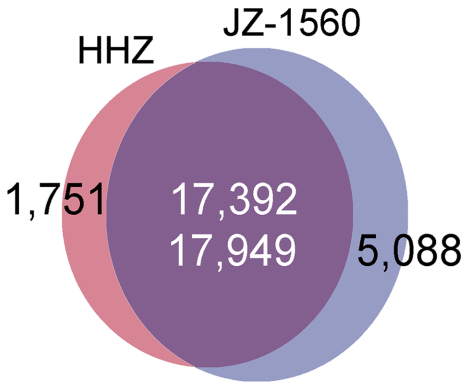

Figure 5. Overlap of assembled contigs between HHZ (indica rice variety) and JZ-1560 (japonica rice variety). H3K36me3 ChIP-seq data were used for assembly.

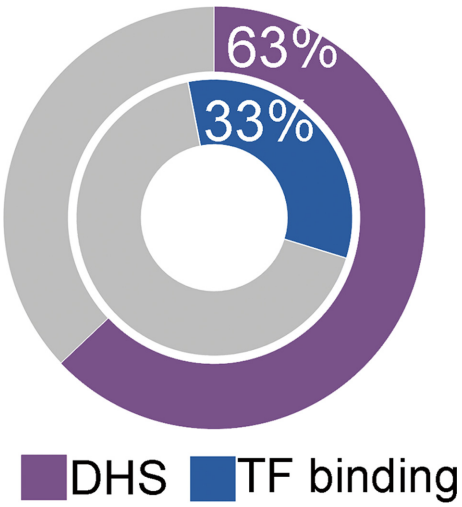

Figure 6. Donut chart showing the high concordance between CGT-seq captured regions, DHS and TF binding regions. The purple region in the outer circle represents the percentage of CGT-seq captured regions overlapping with DHS, and the blue region in the inner circle represents the percentage of CGT-seq captured regions overlapping with TF binding regions collected from nine ChIP-seq data sets (public data summarized in Supplemental Table S1B).

standing of kinetics (46). Supplemental Table S4 summarized the advantages and limitations of CGT-seq and some major reduced representation methods.

Coverage of CGT-seq and potential improvement

We demonstrated in bread wheat that 95% intragenic and 89% promoter regions were covered by CGT-seq read (Figure 2C), and the assembly of CGT-seq read could recover >500 bp regions from >75% genes (Figure 2D). Despite that CGT-seq captured regions did not cover the whole genic regions for most genes, the captured regions are highly functional relevant (Figure 6). In addition, with more genetic markers, longer sequencing reads and mate-pair sequence information, the length and coverage of CGT-seq are expected to be further improved.

Even if the casual region is not recovered, CGT-seq provides high density of long sequences and genetic markers surrounding functional regions with high accuracy. Given that map-based cloning in plant populations with no genetic markers is still time- and labor- consuming, combination of high quality and high density of genetic markers recovered by CGT-seq and bulked segregant analysis would facilitate rapid and efficient gene mapping in populations with no genetic markers.

Cost and uniformity of CGT-seq

Seven parameters were proposed to assess the performance of target-enrichment strategies: (i) sensitivity; (ii) specificity; (iii) uniformity of the read coverage across target regions; (iv) reproducibility of results obtained from replicate experiments; (v) cost; (vi) ease of use and (vii) amount of DNA required per experiment. We showed that CGT-seq had high sensitivity, specificity and reproducibility. The sequencing space and cost is substantially reduced by specific enrichment of the core genome, which account for < 10% of the genome for majority of organisms. For example, to get 10X genomic coverage in bread wheat, 170 Gb re-sequencing reads are required. To ensure a similar coverage surrounding captured genes using CGT-seq, 40 Gb sequencing reads are sufficient (Supplemental Figure S6), saving the cost for 130 Gb sequencing reads, which is approximately $1755 (Supplemental Table S5). The extra ChIP step implemented in CGT-seq only costs about $100 (Supplemental Table S5). Thus, the total cost of CGT-seq is lower than re-sequencing by approximately one order of magnitude. Importantly, re-sequencing is dependent on a well-assembled reference genome, while CGT-seq is reference independent, and has the potential to be broadly applied. Implementation of the method is also practical. ChIP has gradually become a routine technique, and high-efficiency commercial antibodies to the major epigenetic modifications are readily available. In terms of the amount of DNA required, one ChIP experiment requires ∼5–10 g of fresh material. This sample size is easy to obtain from economically important plants, which are generally in good supply.

In terms of uniformity, ChIP-seq read from one mark may suffer from the bias caused by the dynamic range of read distribution surrounding genes. However, by combination of reads from histone modifications preferentially marks both active (e.g. H3K4me3) and repressive (e.g. H3K27me3) genes, this bias could be sufficiently reduced. If further decrease of the prevalence of abundant reads is desired, duplex-specific thermostable nuclease (DSN) enzyme-based normalization is recommended, which is widely used for cDNA library normalization to reduce abundant DNA molecules while preserving less-abundant ones (47). Here, we demonstrated that DSN normalized library help to increase the range of targeted regions recovered (Supplemental Table S6 and Supplemental Figure S8). We further asked whether the markers have preferential localization surrounding CpG island (CGI) promoters, which is closely associated with epigenetic modification, especially DNA methylation (48). The distribution of rice histone markers employed in this study surrounding CGI and non-CGI promoters were calculated, and no apparent preferential localization was observed (Supplemental Figure S9).

Selection of epigenetic markers and generalization of CGT-seq to other species

The epigenetic machineries regulating epigenetic marks are generally ancient (49). The profiles of these modifications surrounding genes are similar in both monocot plants (rice, wheat and maize) and dicot plants (Arabidopsis and tree cotton) (Figure 1A and Supplemental Figure S1). For all these species, the regulatory role of H3K4me3 and H3K27me3 are conserved, with H3K4me3 positively correlated with gene activity, and H3K27me3 negatively correlated with gene activity (Supplemental Figures S2 and S3), which is also the case in mammalian cells (50). Thus, CGT-seq is potentially of wide application. Among these modifications, H3K4me3 is the top enriched marker surrounding TSS for all these species. In addition, given that the antibody to the active H3K4me3 displayed the highest ChIP efficiency in our experiment, i.e. quantity of precipitated DNA by the same amount of antibody, employment of this mark for CGT-seq should be the most cost-effective. If more genes with low expression or no expression need to be captured, antibody to H3K27me3 could be further applied. If longer sequences from genic region are expected, the addition of antibodies to H3K36me3, H3K27ac or H3K4me1 are recommended (Supplemental Figure S10).

CONCLUSION

In summary, the present study developed a reference-independent approach, CGT-seq, which allows enrichment and extensive sequencing of the core genome mainly composed of promoter and intragenic regions. With substantially reduced sequencing space and broad application to both sequenced and non-sequenced species, CGT-seq is a helpful tool for molecular genetic studies in highly divergent population with large genomes and non-model organisms lacking reference genomes.

DATA AVAILABILITY

Data generated in this study were summarized in Supplemental Table S1A, and deposited in GEO under accession number GSE107827 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107827).

Tracks for all sequencing data and assembled scaffolds can be visualized by JBrowse (51) through the following link: http://bioinfo.sibs.ac.cn/browser/cgt-seq

All scripts and description of input and output are available on CGT-seq website: http://bioinfo.sibs.ac.cn/zhanglab/cgt-seq/index.htm.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Prof. Silan Dai from Beijing Forestry University for inspiration discussion. We thank Prof. Hongxuan Lin from Shanghai Institute of Plant Physiology and Ecology for providing rice materials and re-sequencing data. We thank Prof. Jiankang Zhu from Shanghai Center for Plant Stress Biology for the thoughtful discussions.

FUNDING

'Strategic Priority Research Program' of the Chinese Academy of Sciences [XDA08020102]; National Natural Science Foundation of China [31570319, 31770285]; Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences.

Conflict of interest statement. None declared.

REFERENCES

1. Genomes Project, C., Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
2. Zmienko,A., Samelak,A., Kozlowski,P. and Figlerowicz,M. (2014) Copy number polymorphism in plant genomes. *TAG. Theoret. Appl. Genet.*, **127**, 1–18.
3. Zhao,Q., Feng,Q., Lu,H., Li,Y., Wang,A., Tian,Q., Zhan,Q., Lu,Y., Zhang,L., Huang,T. *et al.* (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.*, **50**, 278–284.
4. Chia,J.M., Song,C., Bradbury,P.J., Costich,D., de Leon,N., Doebley,J., Elshire,R.J., Gaut,B., Geller,L., Glaubitz,J.C. *et al.* (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.*, **44**, 803–807.
5. Bevan,M.W. and Uauy,C. (2013) Genomics reveals new landscapes for crop improvement. *Genome Biol.*, **14**, 206.
6. Varshney,R.K., Terauchi,R. and McCouch,S.R. (2014) Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. *PLoS Biol.*, **12**, e1001883.
7. Wendel,J.F., Jackson,S.A., Meyers,B.C. and Wing,R.A. (2016) Evolution of plant genome architecture. *Genome Biol.*, **17**, 37.
8. Luo,M.C., Gu,Y.Q., Puiu,D., Wang,H., Twardziok,S.O., Deal,K.R., Huo,N., Zhu,T., Wang,L., Wang,Y. *et al.* (2017) Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. *Nature*, **551**, 498–502.
9. Brenchley,R., Spannagl,M., Pfeifer,M., Barker,G.L., D'Amore,R., Allen,A.M., McKenzie,N., Kramer,M., Kerhornou,A., Bolser,D. *et al.* (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.
10. International Wheat Genome Sequencing, C. (2014) A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, **345**, 1251788.
11. Davey,J.W., Hohenlohe,P.A., Etter,P.D., Boone,J.Q., Catchen,J.M. and Blaxter,M.L. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.*, **12**, 499–510.
12. Cariou,M., Duret,L. and Charlat,S. (2016) How and how much does RAD-seq bias genetic diversity estimates? *BMC Evol. Biol.*, **16**, 240.
13. Levy,S.E. and Myers,R.M. (2016) Advancements in next-generation sequencing. *Annu. Rev. Genomics Hum. Genet.*, **17**, 95–115.
14. Liu,S., Yeh,C.T., Tang,H.M., Nettleton,D. and Schnable,P.S. (2012) Gene mapping via bulked segregant RNA-Seq (BSR-Seq). *PLoS One*, **7**, e36406.
15. Schneeberger,K. (2014) Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nat. Rev. Genet.*, **15**, 662–676.
16. Clark,M.J., Chen,R., Lam,H.Y., Karczewski,K.J., Chen,R., Euskirchen,G., Butte,A.J. and Snyder,M. (2011) Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.*, **29**, 908–914.
17. Krasileva,K.V., Vasquez-Gross,H.A., Howell,T., Bailey,P., Paraiso,F., Clissold,L., Simmonds,J., Ramirez-Gonzalez,R.H., Wang,X., Borrill,P. *et al.* (2017) Uncovering hidden variation in polyploid wheat. *PNAS*, **114**, E913–E921.
18. Jordan,K.W., Wang,S., Lun,Y., Gardiner,L.J., MacLachlan,R., Hucl,P., Wiebe,K., Wong,D., Forrest,K.L., Consortium,I. *et al.* (2015) A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biol.*, **16**, 48.
19. Asan, Xu,Y., Jiang,H., Tyler-Smith,C., Xue,Y., Jiang,T., Wang,J., Wu,M., Liu,X., Tian,G. *et al.* (2011) Comprehensive comparison of

three commercial human whole-exome capture platforms. *Genome Biol.*, **12**, R95.
20. Wang,H., Liu,C., Cheng,J., Liu,J., Zhang,L., He,C., Shen,W.H., Jin,H., Xu,L. and Zhang,Y. (2016) Arabidopsis flower and embryo developmental genes are repressed in seedlings by different combinations of polycomb group proteins in association with distinct sets of Cis-regulatory elements. *PLos Genet.*, **12**, e1005771.
21. Du,H., Yu,Y., Ma,Y., Gao,Q., Cao,Y., Chen,Z., Ma,B., Qi,M., Li,Y., Zhao,X. *et al.* (2017) Sequencing and de novo assembly of a near complete indica rice genome. *Nat. Commun.*, **8**, 15324.
22. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
23. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
24. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
25. Jackman,S.D., Vandervalk,B.P., Mohamadi,H., Chu,J., Yeo,S., Hammond,S.A., Jahesh,G., Khan,H., Coombe,L., Warren,R.L. *et al.* (2017) ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Res.*, **27**, 768–777.
26. Sahlin,K., Vezzi,F., Nystedt,B., Lundeberg,J. and Arvestad,L. (2014) BESST - Efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics*, **15**, 281.
27. Luo,R., Liu,B., Xie,Y., Li,Z., Huang,W., Yuan,J., He,G., Chen,Y., Pan,Q., Liu,Y. *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, **1**, 18.
28. Boetzer,M., Henkel,C.V., Jansen,H.J., Butler,D. and Pirovano,W. (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, **27**, 578–579.
29. Zerbino,D.R. and Birney,E. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.
30. Thorvaldsdottir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
31. 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.
32. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Proc,G.P.D. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
33. Li,H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.
34. Wu,H., Caffo,B., Jaffee,H.A., Irizarry,R.A. and Feinberg,A.P. (2010) Redefining CpG islands using hidden Markov models. *Biostatistics*, **11**, 499–514.
35. Uauy,C., Distelfeld,A., Fahima,T., Blechl,A. and Dubcovsky,J. (2006) A NAC Gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science*, **314**, 1298–1301.
36. Mamanova,L., Coffey,A.J., Scott,C.E., Kozarewa,I., Turner,E.H., Kumar,A., Howard,E., Shendure,J. and Turner,D.J. (2010) Target-enrichment strategies for next-generation sequencing. *Nat. Methods*, **7**, 111–118.
37. Liu,J., Chen,J., Zheng,X., Wu,F., Lin,Q., Heng,Y., Tian,P., Cheng,Z., Yu,X., Zhou,K. *et al.* (2017) GW5 acts in the brassinosteroid signalling pathway to regulate grain width and weight in rice. *Nat. Plants*, **3**, 17043.
38. Blomen,V.A. and Boonstra,J. (2011) Stable transmission of reversible modifications: maintenance of epigenetic information through the cell cycle. *Cell. Mol. Life Sci.: CMLS*, **68**, 27–44.
39. Stricker,S.H., Koferle,A. and Beck,S. (2017) From profiles to function in epigenomics. *Nat. Rev. Genet.*, **18**, 51–66.
40. Spisak,S., Lawrenson,K., Fu,Y.F., Csabai,I., Cottman,R.T., Seo,J.H., Haiman,C., Han,Y., Lenci,R., Li,Q.Y. *et al.* (2015) CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. *Nat. Med.*, **21**, 1357–1363.
41. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.
42. Palmer,L.E., Rabinowicz,P.D., O'Shaughnessy,A.L., Balija,V.S., Nascimento,L.U., Dike,S., de la Bastide,M., Martienssen,R.A. and McCombie,W.R. (2003) Maize genome sequencing by methylation filtrations. *Science*, **302**, 2115–2117.
43. Rabinowicz,P.D., Schutz,K., Dedhia,N., Yordan,C., Parnell,L.D., Stein,L., McCombie,W.R. and Martienssen,R.A. (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat. Genet.*, **23**, 305–308.
44. Whitelaw,C.A., Barbazuk,W.B., Pertea,G., Chan,A.P., Cheung,F., Lee,Y., Zheng,L., van Heeringen,S., Karamycheva,S., Bennetzen,J.L. *et al.* (2003) Enrichment of gene-coding sequences in maize by genome filtration. *Science*, **302**, 2118–2120.
45. Bewick,A.J. and Schmitz,R.J. (2017) Gene body DNA methylation in plants. *Curr. Opin. Plant Biol.*, **36**, 103–110.
46. Peterson,D.G. (2005) Reduced representation strategies and their application to plant genomes. In: Meksem,K and Kahl,G (eds). *The Handbook of Plant Genome Mapping*. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim , doi:10.1002/3527603514.ch13.
47. Berchtold,M.W. (1989) A simple method for direct cloning and sequencing cDNA by the use of a single specific oligonucleotide and oligo(dT) in a polymerase chain reaction (PCR). *Nucleic Acids Res.*, **17**, 453.
48. Fazzari,M.J. and Greally,J.M. (2004) Epigenomics: beyond CpG islands. *Nat. Rev. Genet.*, **5**, 446–455.
49. Feng,S., Jacobsen,S.E. and Reik,W. (2010) Epigenetic reprogramming in plant and animal development. *Science*, **330**, 622–627.
50. Mikkelsen,T.S., Ku,M.C., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–U552.
51. Skinner,M.E., Uzilov,A.V., Stein,L.D., Mungall,C.J. and Holmes,I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.