A High-Throughput Skim-sequencing Approach for Genotyping, Dosage Estimation and Identifying Translocations

Laxman Adhikari
Kansas State University
Sandesh Shrestha
Kansas State University
Shuanyge Wu
Kansas State University
Jared Crain
Kansas State University
Liangliang Gao
Kansas State University
Byron Evers
Kansas State University
Duane Wilson
Kansas State University
Yoonha Ju
Kansas State University
Dal-Hoe Koo
Kansas State University
Pierre Hucl
University of Saskatchewan
Curtis Pozniak
University of Saskatchewan
Sean Walkowiak
University of Saskatchewan
Xiaoyun Wang
Cornell University
Jing Wu
Cornell University
Jeffrey C. Glaubitz
Cornell University
Lee DeHaan
The Land Institute
Bernd Friebe
Kansas State University
Jesse Poland (jesse.poland@kaust.edu.sa)
King Abdullah University of Science and Technology

Research Article

Keywords: genome sequencing, introgressions, read mapping, translocations

Posted Date: October 14th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-923020/v1

Abstract

The development of next generation sequencing (NGS) enabled a shift from array-based genotyping to high-throughput genotyping by directly sequencing genomic libraries. Even though whole genome sequencing was initially too costly for routine analysis in large populations, such as those utilized for breeding or genetic studies, continued advancements in genome sequencing and bioinformatics have provided the opportunity to utilize whole-genome information. As new sequencing platforms can routinely provide high-quality sequencing data for sufficient genome coverage, a limitation comes in the time and high cost of library construction when multiplexing a large number of samples. Here we describe a high-throughput whole-genome skim-sequencing (skim-seq) approach that can be utilized for a broad range of genotyping and genomic characterization. Using optimized low-volume Illumina Nextera chemistry, we developed a skim-seq method and combined up to 960 samples in one multiplex library using dual index barcoding. With the dual-index barcoding, the number of samples for multiplexing can be adjusted depending on amount of data required and extended to 3,072 samples or more. Panels of double haploid wheat lines (*Triticum aestivum*, CDC Stanley x CDC Landmark), wheat-barley (*T. aestivum* x *Hordeum vulgare*) and wheat-wheatgrass (*Triticum durum x Thinopyrum intermedium*) introgression lines as well as known monosomic wheat stocks were genotyped using the skim-seq approach. Bioinformatics pipelines were developed for various applications where sequencing coverage ranged from 1x down to 0.01x per sample. Using reference genomes, we detected chromosome dosage, identified aneuploidy, and karyotyped introgression lines from the low coverage skim-seq data. Leveraging the recent advancements in genome sequencing, skim-seq provides an effective and low-cost tool for routine genotyping and genetic analysis, which can track and identify introgressions and genomic regions of interest in genetics research and applied breeding programs.

Introduction

Genotyping is essential to quantitative and population genetic studies, as well as genomics-assisted breeding in crops and animals. Thus, the innovation in DNA sequencing technology over the past decades has enabled these disciplines to move from information limited to data rich domains. As sequencing costs continue to become more affordable, the challenge has become determining how to best implement these methods and technologies in breeding pipelines and genetic studies[1]. The advancement and adoption of sequencing technologies can have a huge impact on accelerating the development of elite crop cultivars[1-3]. In addition to sequencing technologies, efficient library preparation can also drive advancements in genetic and molecular sciences[4]. Although molecular markers have played an essential role in genetic studies for microbial, animal and plant genetic studies, genotyping has historically been a time-consuming, laborious task that resulted in tens or possibly hundreds of markers. Traditional DNA markers commonly used in molecular plant breeding comprise AFLP, RFLP, RAPD, SSR, and DArT[5]. These markers require significant upfront discovery, development and validatoin. Next-generation sequencing (NGS) has fundamentally altered the overall genotyping approach, making variant discovery in concert with genotyping. Whole genome sequencing (WGS) is now becoming commonplace for genotyping being used for both identifying and typing genetic variants[6]. Whole genome resequencing has been successfully explored in wheat[7] rice[8], chickpea[9], sesame (*Sesamum indicum* L.)[10], Capsicum[11] and several other species identifying millions of SNPs used to dissect agronomic traits.

Whole genome resequencing is the ideal genotyping method yet technological limitations, genome size and complexity, typically lead to constraints for WGS of large populations due to excessive sequencing cost. To overcome these issues, a variety of targeted sequencing methods have been developed that target regions on the genome, including the exome, particular genes of interest or subsets. Exome sequencing comprises deep sequencing of the protein coding region of a genome[12] and has effectively been used in a range of studies such as identifying copy number variation in maize, identifying the causal gene affecting flowering under long days in barley[6], and dissecting the contribution of wild relatives to the diversity of modern wheat[13]. Similarly, RNA-seq is a popular method to study the transcriptome, providing an attractive method to study coding regions of the genome while sequencing a greatly reduced portion of the genome[6]. However, the complexity of RNA extraction, the challenge of library construction and variability of libraries do not make RNA-seq a readily useful approach for most high-throughput genotyping applications.

Amplicon sequencing (AmpliSeq) is another genotyping method that can leverage the high sequencing output of current machines and is an alternative to whole genome sequencing and exome capture since it allows much higher coverage of the targeted regions with a lower cost[14]. The ampliseq approaches utilize multiplexed PCR amplification and can be used for very high levels of

multiplexing samples while targeting up to thousands of loci. These targeted approaches, however, still necessitate upfront variant discovery with the design and synthesis of oligo sets. Depending on the scope of the genotyping operation, the cost of probe sets can present a formable barrier.

To address the need for targeted sequencing without probe sets, genotyping-by-sequencing and restriction-site-associated DNA sequencing (RAD-seq) were developed as methods to reduce the genome complexity through the use of restriction enzymes[15]. These methods have been very useful in genotyping a large range of model and non-model organisms without a reference genome, as they do not require prior genomic information and have been an important breakthrough for applying genomic selection in due to the very low sample costs for genotyping (reviewed by[16]). GBS is a library preparation method in which the genomic DNA is digested with restriction enzymes and adapters are ligated for sequencing[17]. Multiplexing samples with unique barcodes provides a way to increase throughput while still maintaining sufficient coverage for genetic mapping and genomics-assisted breeding[17]. Many adaptations have been made to the GBS protocol to decrease genome complexity from two enzyme GBS[18] to choosing restriction enzymes covering low copy regions[16]. These methods have been helpful to reproducibly sequence a small fraction of the genome from species with large genomes, such as wheat and barley, as well as genotype species without prior genomic information[18]. Some of the applications of GBS have included genome-wide association studies[19], marker-assisted and genomic selection[20], and haplotype demarcation[21]. Past studies have shown that GBS is an effective genotyping method for population structure and diversity studies[22-24], selection sweep identification[25] and curation of wild accessions in the gene banks[26]. Further applications of GBS include genotyping the specific population for genetic linkage and QTL mapping in arrays of plants[27,28] and animals[29,30]

One area in which NGS could greatly reduce time and labor while simultaneously increasing throughput is in genotyping alien translocations. Alien translocations have proven to be a source of genetic variation across a wide variety of crop species[31], and have played a vital role in increasing the genetic variation and adaptability of plants. Wide-crossing and introgression of novel haplotypes provides a way to access higher levels of genetic diversity that are not available in elite cultivars[32], as many crops have gone through domestication and breeding bottlenecks[33]. For instance, successful alien introgression of various genes has been reported in rice[34], in cotton (*Gossypium hirsutum*)[35], and in maize[36]. In wheat, there are multiple instances of agronomically very important alien introgressions/translocations. For example, the 1BL-1RS translocation, where the short arm of chromosome 1B was replaced with the short arm of chromosome 1R of rye (*Secale cereale*), provides resistance to stripe rust (*Yr9*), leaf rust (*Lr26*), stem rust (*Sr31*), and powdery mildew (*Pm8*)[37]. Secondary and tertiary gene pools have been utilized in hexaploid wheat improvement, as wheat can tolerate significant introduction of alien chromosomal segments through genetic buffering[38]. Successful translocations of chromosome segments from *Aegilops* species have provided wheat with resistance to the devastating stem rust Ug99 by incorporating effective genes such as *Sr33*, *Sr32*, *Sr51*, *Sr47*, and *Sr53* into elite wheat lines[39]. These alien translocations and introgressions from distant wheat relatives, are ubiquitous across wheat breeding programs and wheat germplasm. Furthermore, there is considerable effort to develop and utilize new translocations from a range of different species.

Even though alien introgression breeding can be quite useful, it poses a challenge to the breeding germplasm as alien segments are often large and can influence trait(s) other than targeted traits through linkage drag. In practice, traditional phenotypic selection of introgressed lines has been used to evaluate for negative effects, yet genomic characterization of translocations requires determination of the number of copies of the alien chromosome segments, their length, and physical position. This characterization has mainly been conducted using cytogenetic and molecular marker analysis. However, cytogenetic approaches such as fluorescence *in situ* hybridization (FISH) and genomic *in situ* hybridization (GISH) are often tedious and unable to detect small alien segments. Detectable alien segment lengths by cytology vary from species to species, but usually a 30 Mb segment size is the minimum that can be detectable by GISH in wheat[40]. Additionally, examining the introgression lines using cytology requires extensive work and usually does not scale to large breeding populations.

Given the challenges of characterizing introgression lines, novel methods are needed to characterize introgression breeding. One important consideration for breeding with alien introgressions in wheat is the use of molecular markers that effectively tag the introgression segment and allow differentiation of the respective locus. While translocations and introgressions generally occur on homoeologous chromosomes, there can be considerable sequence divergence between haplotypes such that they can be

considered presence, absence, or hemizygous genotypes. This divergence between wheat and the respective translocations, creates challenges for developing and utilizing molecular markers to tag the respective segments.

With the improvement of DNA sequencing technologies, simplified library preparation methods have been developed, such as Nextera, which randomly targets and generates a genome-wide uniform distribution of sequences[41]. Compared to GBS where restriction digestion and adapter ligations are two-step processes, Nextera uses a transposome complex (transposase plus transposon) to make random double stranded breaks and ligate adapters in genomic DNA in a single step. This method leverages a modified transposition reaction and is called tagmentation[42]. These libraries can then be sequenced to varying levels of whole genome coverage for genomic analysis.

In this study, we report an optimized low-volume method of Illumina Nextera DNA library preparation that can be used for whole genome characterization. Leveraging the power of reference genomes with high-throughput sequencing, we show multiple applications of skim-seq for use in various genomic studies and genomics-assisted breeding, including: (1) genotyping of segregating populations, (2) identification of translocations and genotyping of the translocations in segregating populations, and (3) assessment of chromosome dosage, deletions and aneuploidy. These applications were evaluated in wheat double haploid populations, various introgression and aneuploid addition lines including wheat-barley translocations and *Thinopyrum-durum* wheat introgression lines, and monosomic wheat genetic stocks. Using bioinformatics pipelines, all three approaches for genomic characterization are tractable using the same skim-seq library preparations, which enables the use of a single high-throughput laboratory technique for diverse genetics and breeding applications. The implementation of whole-genome low-coverage sequencing as presented here opens new opportunities for leveraging whole-genome variant information in a range of genomics studies as well as crop and animal breeding.

Materials And Methods

Plant Material and Germplasm

CDC Stanley x CDC Landmark Double Haploid Population

We tested the doubled haploid (DH) population from the cross of spring wheat cultivars ‘CDC Stanley’ and ‘CDC Landmark’ developed by the Crop Development Centre at the University of Saskatchewan, and hence termed the “StanMark-DH” population. The development of DH lines was performed with the wheat-maize wide hybridization method[43]. Initially, F_1 hybrids were developed by crossing CDC Stanley and CDC Landmark and followed by planting of F_1 seeds. The spikelets from the F_1 plants were emasculated and pollinated with maize pollen to induces haploid embryo development. The developing embryos were then excised and cultured in media, and developed into plantlets. The haploid plants were treated with colchicine to doubled the chromosomes and generated primary DH plants. The primary DH plants were self-pollinated to produce the $DH_{0:1}$ generation, from which 48 unique DH lines were used in this study.

Wheat 5D monosomic group

A 5D monosomic line (TA3059), derived in the Chinese Spring (TA3008) background and maintained by the Wheat Genetics Resource Center (WGRC), Manhattan, KS, USA, was self-pollinated to produce progenies segregating for the dosage of the 5D chromosome. This population of 864 samples, named CS M5D, included 839 self-pollinated progeny from TA3059, 16 standard Chinese Spring (TA3008) lines as internal controls and 9 blank samples with no DNA. These genetic stocks were developed by and are available through the WGRC.

Wheat-barley introgressions

Two advanced backcross populations of wheat-barley translocation lines were made by crossing wheat-barley recombinants with group 7 translocations [44,45] to the elite breeding lines, KS090616K-1 and ‘KS Silverado’ developed by the Kansas State University winter wheat breeding program. The wheat-barley recombinants were developed and described previously by Danilova et al. (2019) [45]where group 7 translocations (7AS.7HL-7AL(TA5798), 7BS.7HL-7BL(TA5797), and 7DS.7HL-7AL(TA5799)) were cytologically verified. The wheat-barley homozygous recombinant lines in the ‘Chinese Spring’ background were independently crossed with the

two elite lines to generate F_1 hybrids. The F_1 was backcrossed with the respective recurrent parent to form BC_1 progenies for each cross combination. The final population included 335 BC_1 lines, in addition to the homozygous wheat-barley recombinant lines, the elite recurrent parent lines, and Chinese Spring as internal checks.

Thinopyrum intermedium—Wheat Amphiploid Mapping

A panel of 285 *Thinopyrum intermedium* and *Th. intermedium*—wheat (*Triticum durum*) addition lines were evaluated which included 141 *Th. intermedium* genets and 144 amphiploid genets derived from crossing *Th. intermedium x Triticum durum* lines. The amphiploids were developed by crossing winter *T. durum* as females to *Th. intermedium* as the males. Embryos were rescued and germinated on a modified MS medium, and chromosome doubling was completed with colchicine in young plants. Plants with successful doubling of chromosomes were male-fertile and produced self-progeny that had the complete set of 28 wheat-derived chromosomes and 42 chromosomes for *Th. intermedium*. These amphiploids were then used as male parents and crossed to *Th. intermedium.* Crosses were made using emasculating *Th. intermedium* plants as females followed by embryo rescue of the hybrid. The subsequent progeny were male sterile and were crossed again to *Th. intermedium* as the male parent. A small number of viable seeds were obtained from these crosses, with the resulting progeny including both male-fertile and male-sterile plants. The male-fertile plants were crossed as male parents to *Th. intermedium* and as female parents to *Th. intermedium* in the case of male-sterile plants. The resulting seed was germinated, and young leaf tissue was collected for DNA extraction, genotyping and evaluating the chromosome constitution. Previous research has shown that crosses of *Th.intermedium* to wheat can have variable chromosome composition[46-50].

Library construction

Genomic DNA was extracted from leaf tissue collected from seedlings at the two- to three-leaf stage. Approximately 1.5-inch-long leaves were collected, lyophilized for 3 days and ground using a Retsch mixer mill MM400. Genomic DNA was extracted in 96 well plates using BioSprint DNA kit (Qiagen Inc.) following the manufacture's protocol. In each plate, a random blank well was left as a negative control.

An optimized, low-volume high-throughput library preparation was developed using Illumina Tagment DNA TDE1 Enzyme and Buffer Kits (Illumina Tagment DNA TDE1 Enzyme and Buffer Kits, Illumina, Inc., San Diego, CA, USA), (Supplementary Text S1). This library preparation method provides a high level of multiplexing into a single library that can be sequenced in a single flow cell lane. First, the DNA samples were diluted to ~20 ng/µl and quantified using a Quant-iT™ PicoGreen™ dsDNA Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA). The quantified DNA was then normalized to a target volume of 40 µl at 0.75 ng/µl. Next, a tagmentation reaction consisting of 1 µl normalized to 0.75 ng/µl of the genomic DNA, 0.9966 µl TDE1 Tagment DNA Enzyme, 0.504 µl Tagment DNA Buffer, and 3.3964 µl water was fragmented was incubated at 55°C for 15 minutes, and then cooled to room temperature.

Next, the libraries were PCR amplified to add dual indexes with a unique i5 index for each plate and a unique i7 index for each sample to the tagmented DNA (Supplementary Table S1). For each sample, 5.0 µl of tagmented DNA, 12.5 µl of Taq 2X Master Mix (New England Biolabs Inc., Ipswich, MA, USA), 2 µl of combined i7 and i5 index adapters at 2.5 µM each, and 5.5 µl water were added to make a final reaction volume of 25 µl. The PCR amplification was completed as follows:72°C (3 min), 95°C (1 min), 18 cycles consisting of 95°C (10 sec), 55°C (20 sec), 72°C (3 min), and a final cycle of 72°C (5 min).

For multiplexing, all barcoded and amplified samples were quantified using the Quant-iT™ PicoGreen™ dsDNA Assay Kit. The samples were normalized to 15 µl at 6 ng/µl and then pooled into a single tube. This library was purified using a QIAquick PCR Purification Kit (QIAGEN, Hilden, Germany) and then size-selected from 600 to 800 bp using BluePippin (Sage Science, Inc., Beverly, MA, USA). This library was then cleaned, and the fragment size distribution was verified with an Experion™ DNA 1K Reagents kit (#7007164) using Experion™ Automated Electrophoresis Station (Bio-Rad Laboratories, Inc., Hercules, CA, USA). Finally, the libraries were quantified using the Quant-iT™ PicoGreen™ dsDNA Assay Kit before paired-end sequencing. Paired-end library sequencing was performed by Psomagen (Rockville, MD, USA) with Illumina NovaSeq 6000 or HiSeq X10.

Bioinformatics Pipeline

The analysis pipeline described in this study (Figure 1) can be used for a range of different genomics applications, including variant calling, dosage estimation and identifying chromosome segments from different genomes. This analysis is readily adapted to evaluate introgressions, aneuploidy (dosage), and SNP discovery and genotyping. Efficient processing pipelines for each use case include the following step:

Demultiplexing

The first step in the skim-seq approach demultiplexes the combined sequence library into individual samples. Depending on the sequencing machine, e.g., HiSeq X and NexSeq 2000, the returned sequence files could require varying levels of processing. If sequence data includes separate fastq files for the index reads, (R1.fq, R2.fq, and separate index files I1.fq and I2.fq), a custom Perl script as used here provides easy demultiplexing (https://github.com/sandeshsth/Skim-seq_Method). Based on the sequencing machine,the i5 index could also be reverse complement, which should be identified and the barcode file read accordingly. If the i7 and i5 barcodes are present in the header of the raw fastq file, trimming raw reads to remove the Nextera adapters and primers before demultiplexing can be done using the bbduk program of BBTools (BBMap) suite (https://jgi.doe.gov/data-and-tools/bbtools/). When the i7 and i5 barcodes were provided in separate fastq files than the sequence files, we trimmed the reads after demultiplexing using fastp (https://github.com/OpenGene/fastp).

For project data integrity, we always include a random blank well in each plate to identify any potential plate mix-ups. Blank wells in each 96-well plate were used to assess data quality, as these wells should have little if any sequence data which we confirm as a negative control as less than 0.01% of the average reads per sample.

After a quality check of the sequencing data, we estimated the average sample genome coverage per individual for each population using the following equation:

$$genomecoverage = \frac{(readcount * readlength * 2)}{(totalgenomesize * totalnumberofsamples)}$$ Equation 1.

Sequence Alignment and Concordant Read Selection

We used HISAT2 v2.1.0 (Kim et al., 2019) for read alignment of the skim-seq data to relevant reference sequences. For each genome, index files were generated using HISAT2. For aneuploidy, SNP discovery, and genotyping, a single species reference was utilized. For these analyses, we utilized the Chinese Spring RefSeq v1 assembly (Apples et al., 2018)

For interspecific introgression mapping, a reference assembly was generated by concatenating the reference sequences of a donor and a recipient species. We combined the Chinese Spring reference genome V1.0[51] and barley pseudomolecule assembly of barley cv. Morex[52] for identification of wheat-barley group 7 introgressions. An additional combined reference was generated to map *Th. intermedium* – wheat introgression lines using the Chinese Spring (CS) wheat reference and *T. intermedium* draft genome assembly (provided by *Thinopyrum intermedium* Genome Sequencing Consortium https://phytozome-next.jgi.doe.gov/info/Tintermedium_v2_1) developed from accession C4-5353T1. When combining reference genomes, all chromosomes or pseudomolecule names must be unique.

The HISAT2 pipeline was run with the default parameters for paired-end reads in a multithreaded environment. We disabled the spliced alignment option and suppressed the sequencing alignment map (SAM) records for reads that failed to align. The output SAM files were then filtered using command lines tools to filter for uniquely mapped concordant reads (https://github.com/sandeshsth/Skim-seq_Method).

Normalized read counts were computed using the AWK programming language. Information about chromosome and physical location written to a bed file was used as the input to calculate normalized read counts per one Mb bin. The normalized read counts were computed as:

$$\frac{normalizedreads}{Mb} = \frac{sumofreadsinMbbin}{totalnumberofreadspersample} xNormalizationFactor$$ Equation 2

The normalization factor can specified, where we used reads per 10 million or 100/average read count in all bins. The script (https://github.com/sandeshsth/Skim-seq_Method) also added sample names to the text file. To efficiently process hundreds of

samples, we ran array jobs on a high-performance cluster. The resulting text files included read count in bins, with chromosome and physical locations.

Data Filtering and Visualization for Introgressions and Aneuploidy

Once each sample had been processed to obtain normalized read counts, unknown chromosomes were removed using the UNIX sed command (https://github.com/sandeshsth/Skim-seq_Method), and a final file for all samples was made by concatenating all sample files together. Graphical displays to visualize karyotypes of introgression and aneuploid lines, were plotted using ggplot2 (Wickham, 2009) in R (R programming language). The R scripts for data visualization (https://github.com/sandeshsth/Skim-seq_Method) also allowed us to easily generate read counts per bin and view read depth. For the *Th. intermedium*—wheat lines, read depth provided an efficient way to determine which chromosome additions were present. Leveraging the centromere position with this information also allowed for visualization of Robertsonian Translocations and aneuploidy.

SNP Discovery and Genotyping in StanMark-DH

The genotyping of the DH population was accomplished in two bioinformatics steps by discovering SNP between the two parents and later genotyping the discovered SNPs in the population. To discover SNPs between the two parents, the high-coverage paired-end raw reads of CDC Stanley and CDC Landmark were mapped to the CDC Landmark reference genome (available through the Sequence Read Archive PRJNA544491) using HISAT2[53,54]. The alignment was completed with default parameters except for turning off the spliced alignment function and preventing the unaligned reads from being output in the SAM files. In preparation for variant calling, the alignment files were sorted by chromosome and position. The alignments were filtered using samtools v1.10[55] to keep reads with unique and concordant alignment based on the SAM tags *NH:i:1* and *YT:Z:CP* respectively. The filtered output BAM files were *csi* indexed using SAMtools to generate index files needed for variant calling. Variant discovery was performed with BCFtools commands: *bcftools mpileup* followed by *bcftools call*[56]. The output VCF was annotated with the *–annotate AD,DP,INFO/AD* option with *mpileup* in BCFtools. Variants were discovered on an individual sample basis instead of a population level with option *-G* - in *bcftools call*. The SNP discovery process was run in parallel for each chromosome individually with *-regions*. Output VCF files were filtered and merged together. Each SNP position was filtered based on read depth to keep the SNPs when the following criteria were met: minimum and maximum filtered read depths of ≥ 6 and ≤ 100 respectively and reference and alternate allele read depths of ≥ 3. High-quality SNPs discovered between the parents, CDC Stanley and CDC Landmark, were then called (genotyped) in the 48 DH lines. To genotype the StanMark-DH population, the skim-seq data was filtered using fastp to remove any reads containing adapters while maintaining the final read length of 150 bp[57]. The paired-end fastq files of each sample were processed to generate alignment files with the same pipeline used for the two parents. The alignment files of 48 DH lines were used in genotyping the SNP positions discovered between the two parents using the *-T* option in BCFtools.

Down sampling for low-coverage samples

While most target applications for genotyping in breeding programs such as genomic selection will utilize very low-coverage sequencing to reduce costs, the StanMark-DH population was sequenced at relatively higher depth with raw coverage ranging from 0.6-1.2x. As the cost for sequencing to this depth for a genome the size of wheat would be untenable within a breeding program for large populations, we mimicked low coverage empirical data by randomly sampling three different low coverage levels of 0.1x, 0.05x, and 0.01x. Sampling was completed using seqtk (https://github.com/lh3/seqtk), and the low-coverage samples were mapped and filtered as described earlier and genotyped the SNP positions identified between the parents with option *-T* using BCFtools.

Results And Discussion

Skim-Seq Pipeline

To affordably genotype thousands of samples and leverage the extremely high output of the latest sequencing platforms, we developed a modified low-volume Nextera library preparation method for whole genome sequencing. A high-level of multiplexing enables sequencing of ten or more 96-well plates together. Depending on the species and genome size, the level of multiplexing

can be adjusted up to several thousand, resulting in the target genome coverage of the individual samples. For our applications in genotyping and characterizing hexaploid wheat, we multiplexed from just 48 samples up to 960 samples, giving genome coverage from ~1x down to 0.01x of the very large, 16 Gb wheat genome. To efficiently process the sequence data, we also developed automated scripts that demultiplexed sequence files, aligned samples to reference genomes, and provided efficient ways to visually karyotype samples. The different skim-seq analysis pipelines (Figure 1) were applied to several different use cases including SNP discovery and genotyping, introgression mapping, and aneuploidy analysis.

SNP Discovery and Genotyping

High-quality SNPs were generated from approximately 8x coverage of CDC Stanley and CDC Landmark and then used to genotype the same loci within the skim-seq data. Nearly 26 million raw SNPs were generated from the WGS of CDC Stanley and CDC Landmark. As CDC Landmark has a reference genome, the SNP variants were filtered for position where CDC Stanley had the alternate allele compared to CDC Landmark. After filtering, a total of 12.5 million genome-wide SNPs were identified between the two parents.

The average raw sequencing of 48 DH lines was 0.88x (coverage ranged from 0.61x to 1.23x) and 10.9 million unique SNPs were genotyped across the population. The reads also had very high concordant unique alignment which was used in genotyping SNPs identified between the two parents (Supplementary Tables S2 and S3). The variants discovered in the parents were genotyped on the progeny and assigned to either the CDC Stanley or CDC Landmark parent (Figure 2). To simulate applications with higher plexing levels that would result in lower coverage, we decreased sample coverage through random down sampling to 0.01x coverage. As the coverage was decreased, the number of SNPs genotyped also decreased and simultaneously increased the missing data in each sample (Supplementary Figure S1 and S2). However, the extremely large number of genome-wide variants present along the chromosome provided sufficient markers to genotype haplotype blocks inherited from the respective parents even at 95% to 99% missing data in the DH lines (Figure 2, Supplementary Figures S3). We did observe regions of the genome with low marker density between the two parents (e.g. 450Mb to 650Mb on Chr. 6B) that are likely due to identity by decent with the closely related breeding germplasm.

Our ability to identify genomic segments contributed by either of the two parents in the DH lines were evaluated by comparing the sequencing depths. The original sequencing depth was close to 1x coverage which was able to clearly identify the recombination breakpoints (Figure 2). As we down sampled to low coverage depths from the original sequencing, the density of markers decreased but still clearly distinguished the genomic segments from the two parents to a level of 0.05x. At the lowest sequencing depth, some genomic regions became ambiguous due to low marker density but overall the genotyping of the DH lines and assignment of parental alleles was possible (Figure 2, Supplementary Figures S3).

Wheat-Barley Introgression Mapping

We evaluated a panel of 384 wheat-barley introgression lines using skim-seq with a mean sample genome coverage in the population of 0.025x (Table 1). Using the skim-seq pipeline, demultiplexing followed by trimming using fastp resulted in nearly 90% of the filtered reads being retained for alignment. Even at this low coverage, we observed approximately 70 reads per 1 Mb bin for both the 21 wheat chromosomes and the 7 barley chromosomes when mapped onto the combined reference genome (Table 1). There was some variation in read density across different chromosomes with a minimum of 64.4 reads per Mb in chromosome 2A to 76.8 reads per Mb on chromosome 5D (Supplementary Table S4). Using the normalized read count per Mb, we were able to delimit both the size and the number copies (dosage) of the barley translocation into the group 7 chromosomes of wheat (Figure 3). For example, parental chromosomes with no translocations had very consistent read coverage across the genome. Parental chromosomes with translocations showed minimal read mapping to the wheat genome, and similar coverage mapping to the barley genome (Figure 3A).

The translocation lines are known to carry a group 7 translocation between wheat and barley on each of the homologous wheat groups [45]. Using the skim-seq, we were able to precisely delimit each of the translocations on the physical map (Table 2). Within this population, a 111 Mb segment on chromosome 7A (362-473) was replaced with a 119 Mb segment from barley chromosome 7 (337-456 Mb). On chromosome 7B, the translocation spanned 98 Mb (296-394 Mb) with translocation of a 94Mb region from barley (337-431 Mb). We also observed a likely mispositioned scaffold in the Chinese Spring v1 reference at 326-338 Mb which

was showing presence of wheat chromatin despite being in the middle of the translocation segment. On chromosome 7D translocation, a larger wheat segment of 218 Mb (340-558 Mb) was replaced by a barley segment of 273 Mb (337-610 Mb). Skim-seq provided the physical position and size of translocations in introgression that could be easily used for further breeding work and very high-throughput genotyping of these translocations.

For backcross-derived progeny, we observed the expected heterozygous translocation, as evidenced by read depth at approximately half the normalized read coverage compared to chromosomes with no translocations (Figure 3B). Of the total 335 BC1 progeny potentially carrying the wheat-barley translocation, 169 and 166 were observed with and without the translocation, respectively. This resulted in a 1:1 (carrier vs non-carrier) segregation ratio $X^2(1, N=335) = 0.026$, $P = 0.86$, confirming typical Mendelian segregation.

Aneuploidy mapping

Within the Chinese Spring monosomic Chr5D genetic stocks (CS M5D) the skim-seq resulted in a mean 0.01x sample sequence coverage for samples in the larger population (Table 1). Aligning the samples to the Chinese Spring reference genome, the average number of reads mapped per 1 Mb bin was 30.6 (Table 1). The read depth was uniform across the genome except for chromosome 5D as expected for segregating dosage from the monosomic parental stock (Supplementary Table S4). Reflecting the dosage segregation from a monosomic individual, we observed four primary karyotypes in the progeny of the wheat 5D monosomic: monosomic, euploid, nullisomic and various telosomic plants. This enabled rapid identification of the rare telosomic lines that result at only a few percent from chromosome breakage of the monosomic chromosome during meiosis (Figure 4, Supplementary Table S5, Supplementary Figure S4). Among the 864 samples, 674 (78%) were 5D monosomic, 130 (15%) were euploid, 35 (4%) were 5DL telosomic, 1 (<1%) was 5DS telosomic, 7 (1%) were 5D nullisomic, and 3 other lines were 5D nullisomic and included other structural changes. Less than 0.6% (n=5) of the samples did not produce enough reads for analysis, while the negative control blanks (n=9) were observed as expected with less than 0.01% of average sample reads.

Thinopyrum – wheat Introgression Mapping

Skim-seq was used to evaluate a panel of *Th. intermedium* and *Th. intermedium*—durum wheat amphiploid lines with an average coverage of 0.03x of the *Th. intermedium* genome (Table 1). Within the *Th. intermedium* lines, we used skim-seq to verify the presence of all chromosomes, and then the *Th. intermedium*—durum amphiploid lines were evaluated for additional wheat chromosomes in the *Th. intermedium* background. These crosses are known to harbor a variable number of chromosomes, and the skim-seq pipeline was used to quickly identify which wheat chromosomes were added to the *Th. intermedium* genome. In the 144 potential amphiploid *Th. Intermedium x T. durum* plants, skim-seq identified 108 (75%) individuals that had one or more wheat chromosomes. The wheat chromosome presence was variable with chromosome 2A found only in three genets, whereas chromosome 3A was found in 77 genets. Within individuals, alien chromosome number ranged from 0 to 11, with a median of 3 wheat chromosomes per individual. There was also some evidence of partial chromosomes that could represent translocations between *Th. intermedium* and wheat or chromosome fragments that had been disrupted during meiosis (Supplementary Figure S5). While cytology will be necessary to confirm the exact composition of both addition lines and potentially translocated material, skim-seq provides a very effective pipeline to rapidly screen candidates that are most likely to have desired chromosome additions for further testing and characterization. This provides an efficient way to quickly process large numbers of progeny that may be needed to obtain a desired translocation.

Discussion

Skim-seq: Cost and Time Effective Genotyping Approach

The skim-sequencing approach presented in this study with the data processing pipeline is broadly applicable for different genetics and genomics studies that necessitate profiling a large number of samples in a cost-effective manner. For example, for 5D monosomic lines we sequenced over 800 samples within a single lane of Illumina HiSeq, resulting in an average of 0.01x coverage for a cost of approximately $1.2 per sample. Although skim sequencing generates low-coverage data, this is sufficient for most applications. For instance, we show that 0.01 to 0.03x coverage (Table 1) is sufficient to identify the size of introgressed segments from the alien species and to determine the dosages of chromosomes in the samples. In addition, coverage at this low level of

0.01x to 0.05x was sufficient to identify parentage of and genotyping of double haploid (or recombinant inbred line, RIL) populations.

It is important to note that these various applications of skim-seq leverage available genomic resources, including a genome assembly of the species and in the case of genotyping, high-coverage sequencing data on the parents. These resources are largely available, particularly for crop species. In addition, continued advancements in sequencing including long-read genome assemblies are quickly making the needed genomic resources possible for any species. When combined with the various flexible data processing pipelines there are many straightforward, fast and applicable implementations that can utilize skim-seq datasets.

The important focus of skim-seq is the rapid, low-cost library preparation that can be scaled to extremely high multiplexing. Previous reduced representation sequencing, such as GBS using in-line barcodes, is limited on the number of barcodes that can be effectively combined as well as the upfront costs of synthesizing the adapters. However, the dual indexing possible with these skim-seq Nextera libraries gives combinatorial barcoding to reach much higher levels of multiplexing. This is an important consideration as the sequencing output of new machines continues to increase. To continue generating low-cost genomic profiles on a per sample basis, an increasing number of samples should be sequenced together a single sequencing run.

As the cost of sequencing has dropped below $10 per gigabase and is quickly approaching $1 per gigabase, many species can now be sequenced to relatively high coverage (e.g. 1x-10x coverage) for a few dollars. This makes the library construction costs and throughput an even larger consideration to keep the per sample costs low. As such, the per sample library costs using this skim-seq approach are in the range of $1 per sample. Thus, the combined cost of DNA extraction, library preparation and sequencing is less than $3 per sample and suitable to provide sufficient sequencing data for many applications in most any species. By example, the wheat genomes sequenced here are orders of magnitude larger than other plant and animal genomes. The 0.03x coverage obtained for the 16 Gb hexaploid wheat genomes in this study would be over 1x coverage for a ~400 Mb rice genome.

Application to Genomic Studies and Plant Breeding

The skim-seq approach offers a tractable method to evaluate introgression and amphiploid lines. Compared to low-throughput, labor-intensive cytological methods, skim-seq enabled characterization of very large populations of amphiploids and introgressions. Identifying missing or extra chromosome(s) and its dosage in aneuploidy stocks using skim-seq is straightforward and applicable to speed up the breeding program by replacing tedious cytological methods to identify aneuploidy.

The generation of markers representing the whole genome is essential for genetic studies. The skim-seq method that we presented in this study can generate markers with genome-wide coverage (Supplementary Figure S1). From the down sampled low coverage sequencing, we observed that the marker density decreased commensurate with the decreasing sequence coverage but continued to provide full genome-wide coverage. This highlights that the skim-seq created random reads and even with low coverage sequencing the distribution and sampling remained uniform along the chromosomes. The uniform distribution of markers even with low sequencing coverage will have advantages for various applications such as distinguishing parental segments in the progenies. We were able to clearly identify segments from CDC Stanley and CDC Landmark in the DH lines even at very low coverage of 0.05x. Even though the recombination breakpoints were less precisely observed at 0.01x coverage, this or lower levels of coverage can provide adequate data for routine genotyping, genomic selection, or progeny testing.

Conclusions

In this study, we presented an optimized protocol and bioinformatics pipeline to identify the origin and structural changes of genomic segments in multiple wheat populations using high-throughput low-cost Skim-seq. Using reference genomes, our approach shows that this can be a powerful method to identify translocations and introgression, evaluate chromosomal dosage in aneuploidy stocks, and serve as routine genotyping methods. Moreover, the streamlined skim-seq library preparations, when combined with flexible bioinformatics, can provide a single laboratory method to handle a range of different studies and genomic profiling, greatly simplifying the overall lab operations. As sequencing output continues to increase with commensurate decreasing costs, we anticipate that skim-seq will play a large role in future plant breeding and genetic studies.

Declarations

Acknowledgments

We would like to thank Steve Larson for critical review of the manuscript draft.

Data Availability

The DH population developed from CDC Stanley x CDC Landmark is deposited in sequence read archive (SRA) accession SRS8963504 with BioProject accession PRJNA729723.

The sequence data for each of the demutiplexed samples of the 5D monosomics line are available at NCBI SRA under BioProject accession number PRJNA742385. The sequence data of wheat-barley translocation lines are available at NCBI SRA under BioProject accession number PRJNA738484. IWG sequence data are available at NCBI SRA under BioProject accession PRJNA736976.

All scripts to perform the skim-seq methods have been placed in the Dryad digital data repository:

https://datadryad.org/stash/share/v20dkVsSTj3toGn-CHG92eUSgre17uMT5AH_6LE2GDM

FUNDING

This material is based upon work supported by Feed the Future through the U.S. Agency for International Development, under the terms of Contract No AID-OAA-A-13-00051 and the U.S. National Science Foundation under Grant No. (1339389). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Agency for International Development or the National Science Foundation. This work was funded in part by the Perennial Agriculture Project in conjunction with the Malone Family Land Preservation Foundation and The Land Institute. The Thinopyrum intermedium Genome Sequencing Consortium provided pre-publication access to the IWG genome sequence.

ETHICS DECLARATIONS

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

Laxman Adhikari: Methodology, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization; **Sandesh Shrestha:** Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization; **Shuanyge Wu:** Methodology, Investigation, Validation; **Jared Crain:** Investigation, Validation, Formal analysis, Data Curation, Visualization; **Liangliang Gao:** Software, Investigation, Data Curation; **Byron Evers:** Investigation, Resources, Data Curation; **Duane Wilson:** Investigation , Resources; **Yoonha Ju:** Investigation, Validation, Formal analysis; **Dal-Hoe Koo:** Investigation, Validation, Formal analysis, Resources; **Pierre Hucl:** Resources, Funding acquisition; **Curtis Pozniak:** Resources, Funding acquisition, Writing - Review & Editing; **Sean Walkowiak:** Resources, Writing - Review & Editing; **Xiaoyun Wang:** Methodology, Investigation; **Jing Wu:** Methodology, Investigation; **Jeffrey C. Glaubitz:** Conceptualization, Methodology, Investigation, Project administration, Funding acquisition, Writing - Review & Editing; **Lee DeHaan:** Conceptualization, Resources, Project administration, Funding acquisition; **Bernd Friebe:** Conceptualization, Resources, Supervision; **Jesse Poland:** Conceptualization, Methodology, Validation Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition; **All authors:** Writing - Review & Editing

References

1. Rasheed, A. *et al.* Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. *Mol. Plant*, **10**, 1047–1064 (2017).
2. Varshney, R. K., Terauchi, R. & McCouch, S. R. Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. *PLoS Biol*, **12**, e1001883 (2014).

3. Poland, J. Breeding-assisted genomics. *Current opinion in plant biology*, **24**, 119–124 (2015).
4. Davey, J. W. *et al.* Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510 (2011).
5. Yang, H. *et al.* Application of next-generation sequencing for rapid marker development in molecular plant breeding: a case study on anthracnose disease resistance in Lupinus angustifolius L. *BMC Genomics*, **13**, 318 https://doi.org/10.1186/1471-2164-13-318 (2012).
6. Onda, Y. & Mochida, K. Exploring Genetic Diversity in Plants Using High-Throughput Sequencing Techniques. *Curr Genomics*, **17**, 358–367 https://doi.org/10.2174/1389202917666160331202742 (2016).
7. Rimbert, H. *et al.* High throughput SNP discovery and genotyping in hexaploid wheat. *PLoS One*, **13**, e0186329 https://doi.org/10.1371/journal.pone.0186329 (2018).
8. Xu, X. *et al.* Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol*, **30**, 105–111 https://doi.org/10.1038/nbt.2050 (2011).
9. Varshney, R. K. *et al.* Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. *Nature Biotechnology*, **31**, 240–246 https://doi.org/10.1038/nbt.2491 (2013).
10. Wang, L. *et al.* Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. *Genome Biol*, **15**, R39 https://doi.org/10.1186/gb-2014-15-2-r39 (2014).
11. Qin, C. *et al.* Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proceedings of the National Academy of Sciences* **111**, 5135-5140, doi:10.1073/pnas.1400975111 (2014).
12. Chilamakuri, C. S. R. *et al.* Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics*, **15**, 449 https://doi.org/10.1186/1471-2164-15-449 (2014).
13. He, F. *et al.* Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nature Genetics*, **51**, 896–904 https://doi.org/10.1038/s41588-019-0382-2 (2019).
14. Hawliczek, A. *et al.* Deep sampling and pooled amplicon sequencing reveals hidden genic variation in heterogeneous rye accessions. *BMC Genomics*, **21**, 845 https://doi.org/10.1186/s12864-020-07240-3 (2020).
15. Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G. & Hohenlohe, P. A. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, **17**, 81–92 https://doi.org/10.1038/nrg.2015.28 (2016).
16. Poland, J. A. & Rife, T. W. Genotyping-by-sequencing for plant breeding and genetics. *The Plant Genome*, **5**, 92–102 (2012).
17. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379 (2011).
18. Poland, J. A., Brown, P. J., Sorrells, M. E. & Jannink, J. L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One*, **7**, e32253 https://doi.org/10.1371/journal.pone.0032253 (2012).
19. Juliana, P. *et al.* Genome-wide association mapping for wheat blast resistance in CIMMYT's international screening nurseries evaluated in Bolivia and Bangladesh. *Sci. Rep*, **10**, 1–14 (2020).
20. Juliana, P. *et al.* Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. *Nature Genetics*, **51**, 1530–1539 (2019).
21. Sehgal, D. *et al.* Haplotype-Based, Genome-Wide Association Study Reveals Stable Genomic Regions for Grain Yield in CIMMYT Spring Bread Wheat. *Frontiers in Genetics*, **11**, 1427 (2020).
22. Singh, N. *et al.* Genomic Analysis Confirms Population Structure and Identifies Inter-Lineage Hybrids in Aegilops tauschii. *Frontiers in Plant Science*, **10**, https://doi.org/10.3389/fpls.2019.00009 (2019).
23. Pereira-Dias, L., Vilanova, S., Fita, A., Prohens, J. & Rodríguez-Burruezo, A. Genetic diversity, population structure, and relationships in a collection of pepper (Capsicum spp.) landraces from the Spanish centre of diversity revealed by genotyping-by-sequencing (GBS). *Horticulture Research*, **6**, 54 https://doi.org/10.1038/s41438-019-0132-8 (2019).
24. Kumar, A. *et al.* Genotyping-by-Sequencing Analysis for Determining Population Structure of Finger Millet Germplasm of Diverse Origins. *Plant Genome*, **9**, https://doi.org/10.3835/plantgenome2015.07.0058 (2016).

25. Wang, K. *et al.* Detection of Selection Signatures in Chinese Landrace and Yorkshire Pigs Based on Genotyping-by-Sequencing Data. *Frontiers in Genetics*, **9**, https://doi.org/10.3389/fgene.2018.00119 (2018).
26. Singh, N. *et al.* Efficient curation of genebanks using next generation sequencing reveals substantial duplication of germplasm accessions. *Sci. Rep*, **9**, 650 https://doi.org/10.1038/s41598-018-37269-0 (2019).
27. Adhikari, L., Lindstrom, O. M., Markham, J. & Missaoui, A. M. Dissecting Key Adaptation Traits in the Polyploid Perennial Medicago sativa Using GBS-SNP Mapping. *Frontiers in Plant Science*, **9**, https://doi.org/10.3389/fpls.2018.00934 (2018).
28. Carrasco, B. *et al.* Construction of a highly saturated linkage map in Japanese plum (Prunus salicina L.) using GBS for SNP marker calling. *PLoS One*, **13**, e0208032 https://doi.org/10.1371/journal.pone.0208032 (2018).
29. Yin, X., Arias-Pérez, A., Kitapci, T. H. & Hedgecock, D. High-Density Linkage Maps Based on Genotyping-by-Sequencing (GBS) Confirm a Chromosome-Level Genome Assembly and Reveal Variation in Recombination Rate for the Pacific Oyster *Crassostrea gigas*. *G3: Genes|Genomes|Genetics* **10**, 4691-4705, doi:10.1534/g3.120.401728 (2020).
30. Everett, M. V. & Seeb, J. E. Detection and mapping of QTL for temperature tolerance and body size in Chinook salmon (Oncorhynchus tshawytscha) using genotyping by sequencing. *Evol. Appl*, **7**, 480–492 https://doi.org/10.1111/eva.12147 (2014).
31. Jauhar, P. P. Modern biotechnology as an integral supplement to conventional plant breeding: the prospects and challenges. *Crop Sci*, **46**, 1841–1859 (2006).
32. Dempewolf, H. *et al.* Past and future use of wild relatives in crop breeding. *Crop Sci*, **57**, 1070–1082 (2017).
33. Tanksley, S. D. & McCouch, S. R. Seed banks and molecular maps: unlocking genetic potential from the wild., **277**, 1063–1066 (1997).
34. Khush, G. S. & Brar, D. S. Alien introgression in rice. *The Nucleus*, **60**, 251–261 https://doi.org/10.1007/s13237-017-0222-7 (2017).
35. Wang, Y. *et al.* Inducement and identification of chromosome introgression and translocation of Gossypium australe on Gossypium hirsutum. *BMC genomics*, **19**, 15–15 https://doi.org/10.1186/s12864-017-4398-7 (2018).
36. Wang, L., Yang, A., He, C., Qu, M. & Zhang, J. Creation of new maize germplasm using alien introgression from Zea mays ssp. mexicana., **164**, 789–801 https://doi.org/10.1007/s10681-008-9730-5 (2008).
37. Friebe, B., Jiang, J., Raupp, W., McIntosh, R. & Gill, B. Characterization of wheat-alien translocations conferring resistance to diseases and pests: current status., **91**, 59–87 (1996).
38. Hao, M. *et al.* The resurgence of introgression breeding, as exemplified in wheat improvement. *Frontiers in Plant Science*, **11**, 252 (2020).
39. Kishii, M. An update of recent use of Aegilops species in wheat breeding. *Frontiers in Plant Science*, **10**, 585 (2019).
40. Gao, L. *et al.* The *Aegilops ventricosa* 2N^{v}S segment in bread wheat: cytology, genomics and breeding.*Theoretical and Applied Genetics*,1–14(2020).
41. Caruccio, N., Grunenwald, H. & Syed, F. Nextera^{TM} technology for NGS DNA library preparation: simultaneous fragmentation and tagging by *in vitro* transposition.*Nature Methods***16** (2009).
42. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol*, **11**, R119 https://doi.org/10.1186/gb-2010-11-12-r119 (2010).
43. Santra, M., Wang, H., Seifert, S. & Haley, S. *in Wheat Biotechnology*235–249(Springer, 2017).
44. Danilova, T. V., Friebe, B., Gill, B. S., Poland, J. & Jackson, E. Development of a complete set of wheat–barley group-7 Robertsonian translocation chromosomes conferring an increased content of β-glucan. *Theoretical and Applied Genetics*, **131**, 377–388 https://doi.org/10.1007/s00122-017-3008-z (2018).
45. Danilova, T. V., Poland, J. & Friebe, B. Production of a complete set of wheat-barley group-7 chromosome recombinants with increased grain β-glucan content. *Theor Appl Genet*, **132**, 3129–3141 https://doi.org/10.1007/s00122-019-03411-3 (2019).
46. Fedak, G. & Han, F. Characterization of derivatives from wheat-Thinopyrum wide crosses. *Cytogenetic and Genome Research*, **109**, 360–367 https://doi.org/10.1159/000082420 (2005).
47. Friebe, B., Mukai, Y., Gill, B. & Cauderon, Y. C-banding and in-situ hybridization analyses of Agropyron intermedium, a partial wheat x Ag. intermedium amphiploid, and six derived chromosome addition lines. *Theoretical and Applied Genetics*, **84**, 899–

905 (1992).

48. Han, F., Liu, B., Fedak, G. & Liu, Z. Genomic constitution and variation in five partial amphiploids of wheat–Thinopyrum intermedium as revealed by GISH, multicolor GISH and seed storage protein analysis. *Theoretical and applied genetics*, **109**, 1070–1076 (2004).
49. Hayes, R. *et al.* Perennial cereal crops: An initial evaluation of wheat derivatives. *Field Crops Research*, **133**, 68–89 (2012).
50. Turner, M. K., DeHaan, L., Jin, Y. & Anderson, J. A. Wheatgrass–wheat partial amphiploids as a novel source of stem rust and Fusarium head blight resistance. *Crop Sci*, **53**, 1994–2005 (2013).
51. Appels, R. *et al.* Shifting the limits in wheat research and breeding using a fully annotated reference genome., **361**, https://doi.org/10.1126/science.aar7191 (2018).
52. Mascher, M. *et al.* A chromosome conformation capture ordered sequence of the barley genome. *Nature*, **544**, 427–433 (2017).
53. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, **37**, 907–915 (2019).
54. Walkowiak, S. *et al.* Multiple wheat genomes reveal global variation in modern breeding. *Nature*, **588**, 277–283 https://doi.org/10.1038/s41586-020-2961-x (2020).
55. Li, H. *et al.* The sequence alignment/map format and SAMtools., **25**, 2078–2079 (2009).
56. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data., **27**, 2987–2993 (2011).
57. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor., **34**, i884–i890 https://doi.org/10.1093/bioinformatics/bty560 (2018).

Tables

Table 1. Different skim-seq populations, their genome coverage and related information

Population	sample size (n)	Total Reads in file	Average Coverage*	Total reads in Sample	Trimmed reads in samples	Total reads in overall alignment (%)	Total unique concordant reads and alignment (%)	Mapped paired-end reads per 1 Mb bin (mean)
wheat-barley Group 7	384	485,575,828	0.025X	410,205,551	296,867,400	266,992,743 (89.9)	192,128,852 (64.7)	71
Wheat 5D Monosomic	864	403,673,248	0.01X	337,742,288	249,616,176	234,389,589 (93.9)	188,373,518 (75.4)	31
IWG-Wheat and IWG	288	359405323	0.03X	302850841	258410843	185564827 (71.81)	103832640 (40.2)	61

* average genome coverage in sample computed as (read count x read length (x 2))/(Genome size x n), where

read length = 150 bp

wheat genome size = 15 Gb

intermedium wheat grass genome size = 12 Gb

Table 2. Wheat-barley group 7 recombinants pedigree, number of samples in different groups, and translocation position information.

Translocation Designation	Pedigree	No. samples	No. of Samples carrying translocation	Translocation breakpoints in Wheat (Mb)	Translocation breakpoints in Barley (Mb)
7AS.7HL-7AL	2019-219-57_X_KS Silverado	27	11	362 - 473	337 - 456
7AS.7HL-7AL	2019-219-36_X_KS090616K-1	34	16		
7AS.7HL-7AL	2019-219-57_X_KS090616K-1	35	20		
7BS.7HL-7BL	2019-215-6_X_KS Silverado	28	16	296 - 394	337 - 431
7BS.7HL-7BL	2019-215-34_X_KS Silverado	25	13		
7BS.7HL-7BL	2019-215-6_X_KS090616K-1	36	16		
7BS.7HL-7BL	2019-215-34_X_KS090616K-1	30	16		
7BS.7HL-7BL	KS090616K-1_X_2019-215-26	14	6		
7DS.7HL-7DL	2019-216-33_X_KS Silverado	26	16	340 - 555	337 - 610
7DS.7HL-7DL	2019-216-36_X_KS Silverado	26	13		
7DS.7HL-7DL	2019-216-33_X_KS090616K-1	32	12		
7DS.7HL-7DL	2019-216-36_X_KS090616K-1	22	14		
-	KS Silverado (PARENT)	10	-	-	-
-	KS090616K-1 (PARENT)	10	-		
7AS.7HL-7AL	TA5798	6	Homozygous	362 - 473	337 - 456
7BS.7HL-7BL	TA5797	7	Homozygous	296 - 394	337 - 431
7DS.7HL-7DL	TA5799	6	Homozygous	340 - 555	337 - 610
	Chinese Spring	6	-		
	Blank	4	-		

Total	384	169

Figures

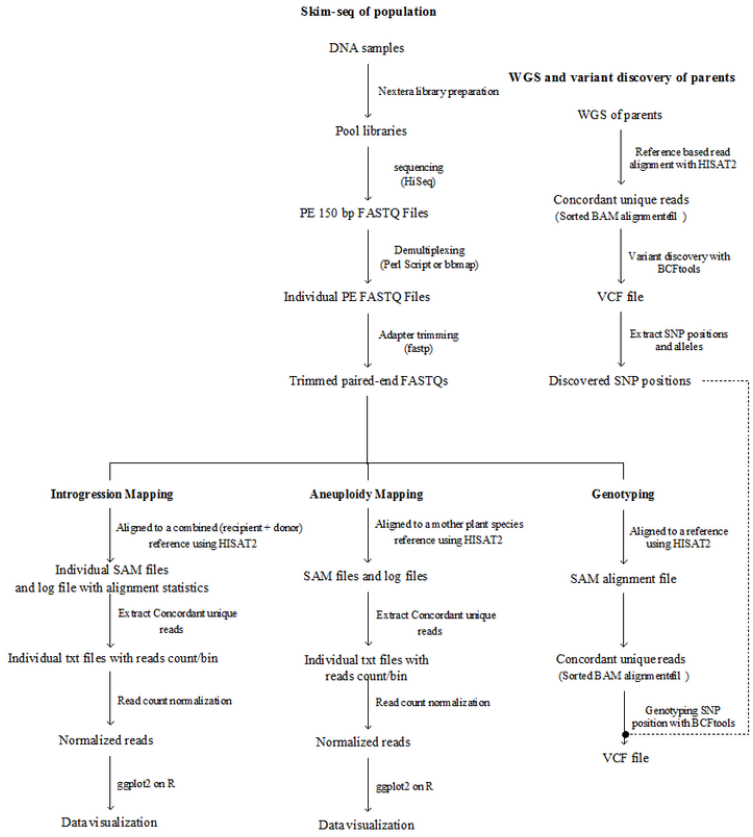

Figure 1

Skim-seq processing pipelines using sequence data generated from optimized Nextera library preparation followed by applications including introgression mapping, aneuploidy determination, and single nucleotide polymorphism (SNP) discovery and genotyping.

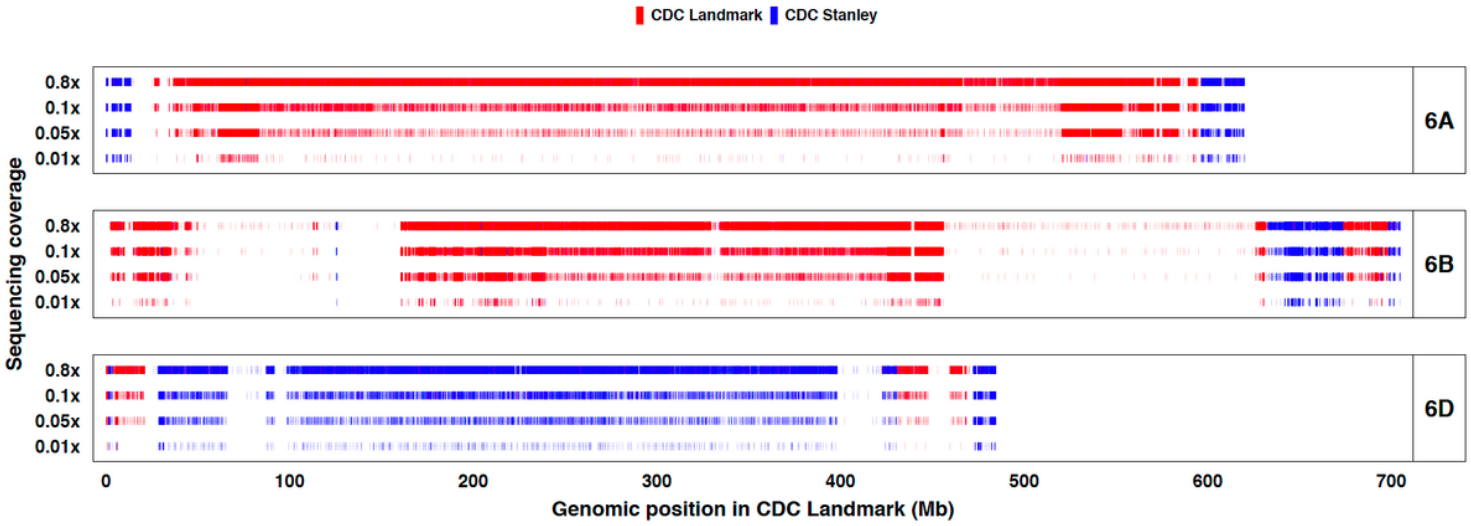

Figure 2

Genomic segments of CDC Landmark and CDC Stanley observed on chromosomes 6A, 6B and 6D of a doubled haploid line (DH01029-0) using various sequencing depths (original 0.8x followed by simulated 0.1x, 0.05x, and 0.01x from the original).

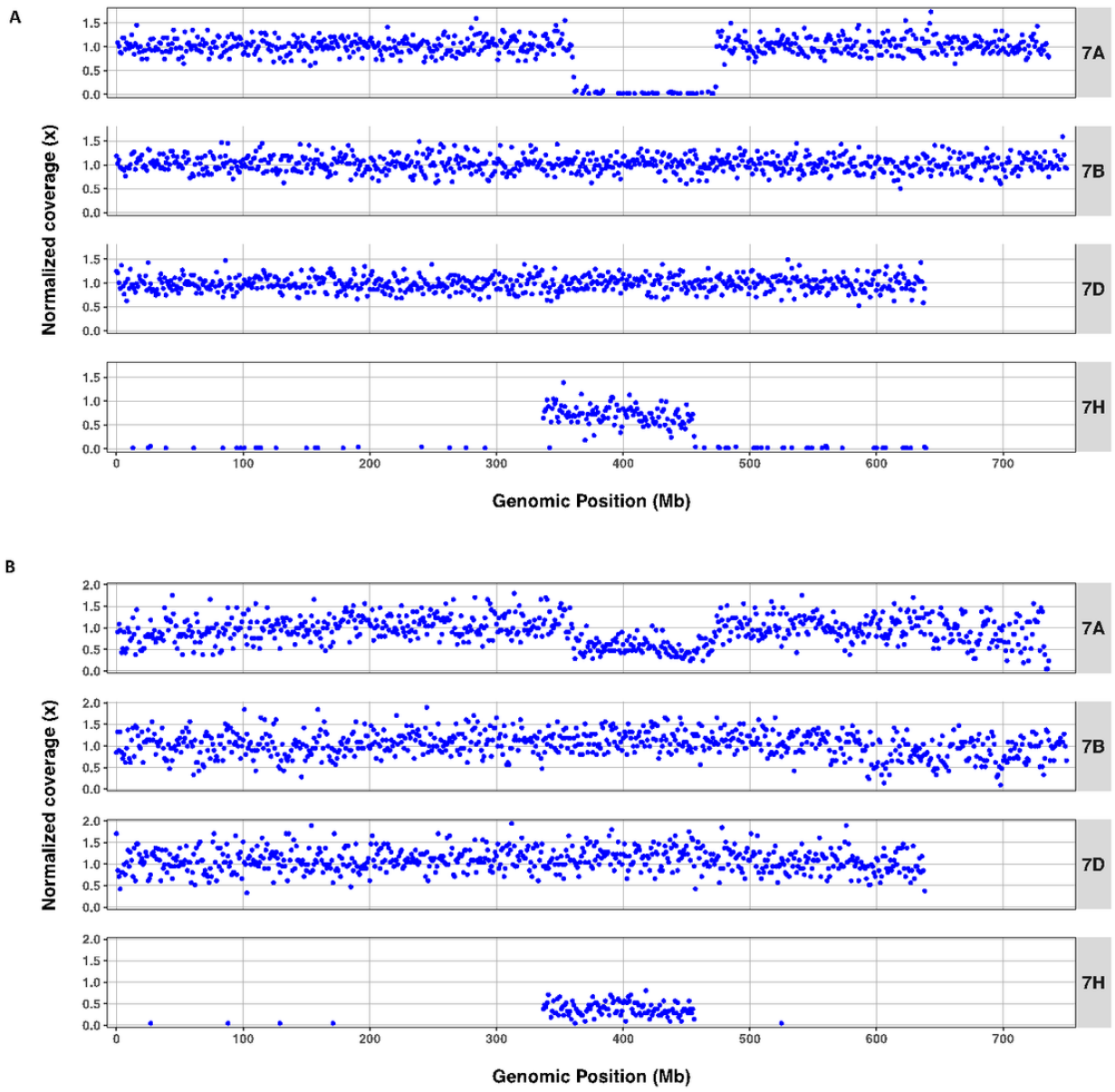

Figure 3

Normalized read counts of a wheat-barley group 7 translocation (7DS.7HL-7DL) for (a) homozygous parent TA5799 and (b) heterozygous back-cross derived progeny.

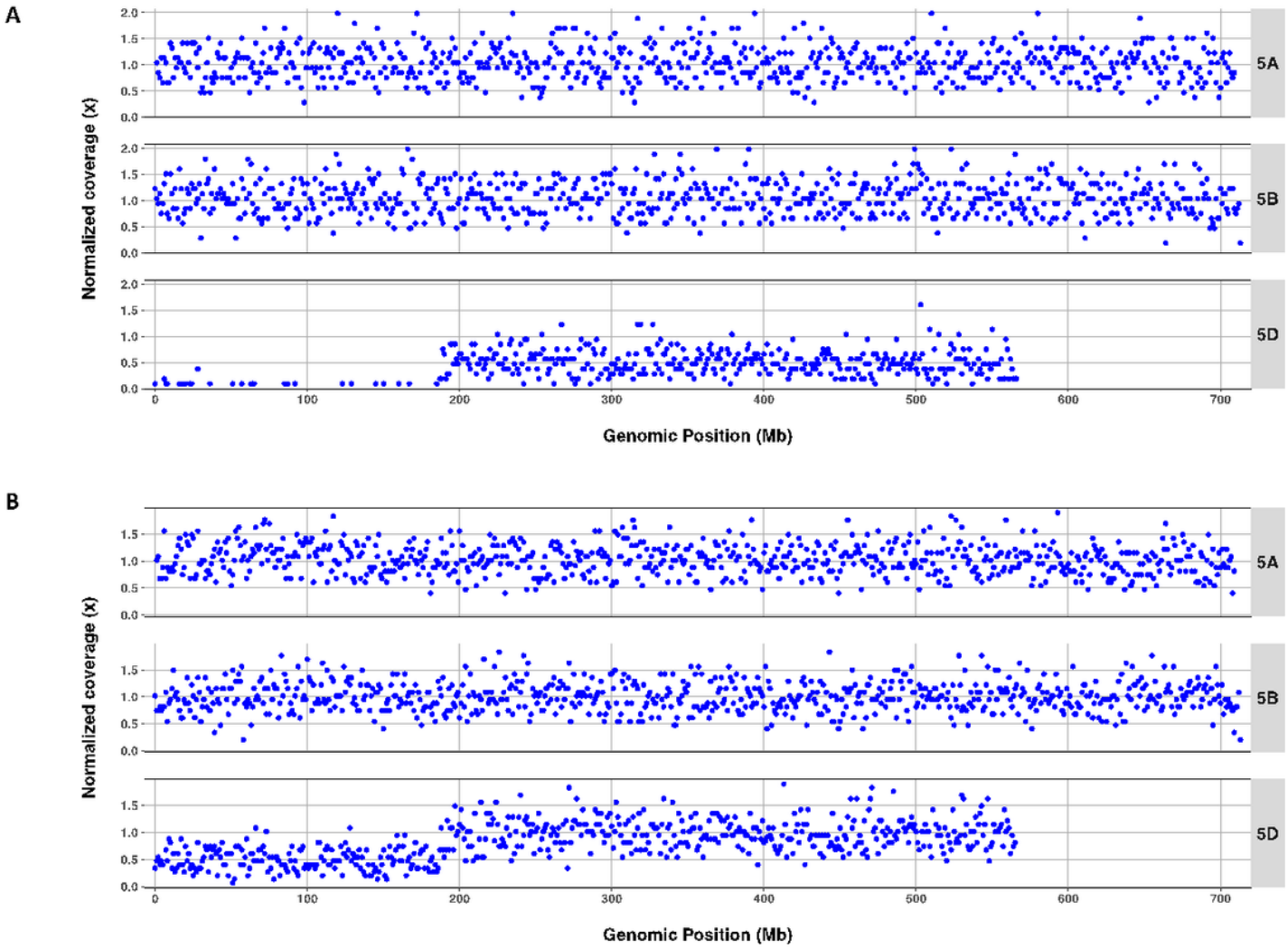

Figure 4

Normalized read counts for example individual samples from CS-M5D populations showing telosomic 5DL. Panel A shows a mono-telosomic 5DL line without additional copy of 5D, and panel B shows mono-telosomic 5DL with additional full 5D chromosome.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalMaterial.docx
- SupplementaryFigureS1SNPdistribution.pdf
- SupplementaryFigureS2DistributionSNPsDH.pdf
- SupplementaryFigureS3StanleyLandmarkDH.zip
- SupplementaryFigureS4CS5DmonoteloDNA200317P01C04.pdf
- SupplementaryFigureS5IWGwheataneuploid.png
- SupplementaryTableS1barcodesequence.xlsx
- SupplementaryTableS2StanMarkalignmentsummary.xlsx
- SupplementaryTableS3StanMarkalignmentsummarylowcoveragesamples.xlsx
- SupplementaryTableS4readsmappedCSv1.docx

- SupplementaryTableS5readcountchrarm.docx