



OPEN

## A high-throughput skim-sequencing approach for genotyping, dosage estimation and identifying translocations

Laxman Adhikari<sup>1,6,7</sup>, Sandesh Shrestha<sup>1,7</sup>, Shuangye Wu<sup>1,7</sup>, Jared Crain<sup>1</sup>, Liangliang Gao<sup>1</sup>, Byron Evers<sup>1</sup>, Duane Wilson<sup>1</sup>, Yoonha Ju<sup>1</sup>, Dal-Hoe Koo<sup>1</sup>, Pierre Hucl<sup>2</sup>, Curtis Pozniak<sup>2</sup>, Sean Walkowiak<sup>2,3</sup>, Xiaoyun Wang<sup>4</sup>, Jing Wu<sup>4</sup>, Jeffrey C. Glaubitz<sup>4</sup>, Lee DeHaan<sup>5</sup>, Bernd Friebe<sup>1</sup> & Jesse Poland<sup>1,6</sup>✉

The development of next-generation sequencing (NGS) enabled a shift from array-based genotyping to directly sequencing genomic libraries for high-throughput genotyping. Even though whole-genome sequencing was initially too costly for routine analysis in large populations such as breeding or genetic studies, continued advancements in genome sequencing and bioinformatics have provided the opportunity to capitalize on whole-genome information. As new sequencing platforms can routinely provide high-quality sequencing data for sufficient genome coverage to genotype various breeding populations, a limitation comes in the time and 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 the amount of data required, and could be extended to 3,072 samples or more. Panels of doubled 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 1 × down to 0.01 × per sample. Using reference genomes, we detected chromosome dosage, identified aneuploidy, and karyotyped introgression lines from the 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.

### Abbreviations

DH	Doubled haploid
GBS	Genotyping-by-sequencing
NGS	Next-generation sequencing
Skim-seq	Skim-sequencing
IT	Interstitial translocation

<sup>1</sup>Department of Plant Pathology, Kansas State University, Manhattan Kansas, USA. <sup>2</sup>Crop Development Centre (CDC), University of Saskatchewan, Saskatoon, SK, Canada. <sup>3</sup>Grain Research Laboratory, Canadian Grain Commission, Winnipeg, MB, Canada. <sup>4</sup>Institute of Biotechnology, Cornell University, Ithaca, NY, USA. <sup>5</sup>The Land Institute, Salina, KS, USA. <sup>6</sup>Center for Desert Agriculture, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. <sup>7</sup>These authors contributed equally: Laxman Adhikari, Sandesh Shrestha and Shuangye Wu. ✉email: jesse.poland@kaust.edu.sa

Genotyping is essential to quantitative and population genetic studies, as well as genomics-assisted breeding in crops and animals. Innovations in DNA sequencing technology over the past decades have enabled these disciplines to move from information-limited to data-rich domains. As costs fall, and sequencing becomes adopted more widely, greater focus has been placed on how best to use these methods and technologies in breeding pipelines and genetic studies<sup>1</sup>. The advancement and adoption of sequencing technologies can have a huge impact on accelerating the development of elite crop cultivars<sup>1–3</sup>. In addition to sequencing technologies, efficient library preparation can also drive advancements in genetic and molecular sciences<sup>4</sup>. Molecular markers have played an imperative role in microbial, animal and plant genetic studies. However, until the advent of next-generation sequencing (NGS), marker development was slow and laborious. Genotyping has historically been a time-consuming, laborious task that resulted in tens or possibly hundreds of markers. Some of the initial DNA markers, such as AFLP, RFLP, RAPD, SSR, and DArT<sup>5</sup> require significant upfront discovery, development and validation. NGS has altered the overall genotyping approach, making variant discovery and genotyping a one-step process. Whole-genome sequencing (WGS) is now becoming commonplace for genotyping, being used for both identifying and typing genetic variants<sup>6</sup>. Whole-genome resequencing has been successfully explored in wheat (*Triticum aestivum*)<sup>7</sup>, rice (*Oryza sativa* L.)<sup>8</sup>, chickpea (*Cicer arietinum* L.)<sup>9</sup>, sesame (*Sesamum indicum* L.)<sup>10</sup>, and capsicum (*Capsicum annum* L.)<sup>11</sup>, leading to the discovery of millions of SNPs, used to dissect agronomic traits.

Whole-genome resequencing is an ideal genotyping method, yet the excessive costs for library generation and sequencing restrict its application in larger populations. To overcome these constraints, a variety of targeted sequencing methods have been developed, including RNA-seq, sequence capture, and amplicon sequencing. RNA-seq is primarily used to study the transcriptome, but from a genotyping perspective it is essentially a complexity reduction technique that targets only the gene space, which is a very small portion of the genome<sup>6</sup>. 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. Sequence capture and amplicon sequencing (AmpliSeq) focus on reducing sequencing cost as an alternative to whole-genome sequencing to generate higher coverage of targeted regions with less total sequencing<sup>12</sup>. 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<sup>13</sup>. Similarly, sequence capture uses oligo probe sets to bind and enrich targeted regions of the genome, generating a reduced proportion for higher coverage sequencing<sup>14</sup>. These targeted approaches, however, still necessitate upfront variant discovery with the design and synthesis of oligo sets<sup>15</sup>. Depending on the scope of the genotyping operation, the cost of probe sets may present a barrier to adoption.

To address the need for targeted sequencing without probe sets, genotyping-by-sequencing (GBS) and restriction-site-associated DNA sequencing (RAD-seq) were developed as complexity reduction methods through the use of restriction enzymes<sup>16</sup>. These methods have been useful in genotyping a large range of model and non-model organisms without a reference genome, as they do not require prior genomic information like sequence capture or amplicon sequencing. In particular, the overall low cost of GBS has been a breakthrough for applying genomic selection in breeding programs<sup>17</sup>. Library preparation for GBS involves digestion of the genomic DNA with restriction enzymes followed by ligation of barcoded adapters to the fragments<sup>18</sup>. Multiplexing samples with unique barcodes provides a way to increase throughput and reduce the cost<sup>18</sup>. Numerous modifications have been made to the GBS protocol to bring about a reduction in genome complexity such as the use of two-enzyme systems<sup>19</sup> or the use of restriction enzymes that target low copy regions of the genome<sup>17</sup>. These methods have been helpful to reproducibly sequence a small fraction of the genome from species with large genomes, including wheat and barley<sup>19</sup>. Some of the applications of GBS have included genome-wide association studies<sup>20</sup>, marker-assisted and genomic selection<sup>21</sup>, and haplotype demarcation<sup>22</sup>. Past studies have shown that GBS is an effective genotyping method for population structure and diversity studies<sup>23–25</sup>, selection sweep identification<sup>26</sup> and curation of wild accessions in the gene banks<sup>27</sup>. Further applications of GBS include genotyping the specific population for genetic linkage and association mapping in plants<sup>28,29</sup> and animals<sup>30,31</sup>.

One area in which NGS could greatly reduce time and labor while increasing throughput is in genotyping populations for alien translocations. Introgression of 'alien' segments from wild relatives are common in crop species and play a vital role in increasing genetic diversity and, thus, adaptability of plants<sup>32</sup>. Wide-crossing and introgression of novel haplotypes provides a way to access genetic diversity that is not found in the primary gene pool of crop cultivars<sup>33</sup>. For instance, 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<sup>34</sup>. These alien translocations and introgressions from distant wheat relatives are ubiquitous across wheat breeding programs and wheat germplasm.

Even though alien introgression breeding is valuable for crop improvement, it poses a challenge for marker development and molecular breeding. Initial 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 time consuming and low throughput, and limited in the lower size of detectable alien segments. Although the limits of detection vary between species, presumably reflecting chromosome size and levels of condensation, introgressed segments smaller than 30 Mb are not detectable in wheat<sup>35</sup>. While there are limitations for detection and genotyping of introgressions using cytology, as well as challenges in the development of molecular markers<sup>36</sup>, these segments are readily detected using whole-genome sequencing<sup>37</sup>. However, for high-throughput screening of these introgression lines, the previously mentioned limitations and costs of whole-genome sequencing become a constraining factor.

With the improvement of DNA sequencing technologies, simplified library preparation methods have been developed, such as Nextera, which are enzyme-based but randomly sample a genome-wide uniform distribution of sequences<sup>38</sup>. 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 proceeds with a modified transposition reaction

and is called tagmentation<sup>39</sup>. These libraries can then be sequenced to varying levels of whole genome coverage for genomic analysis.

In this study, we optimized a low-concentration, low-volume Illumina Nextera DNA library preparation that can be used for whole genome characterization in breeding and genetic studies, and give multiple case studies for applying skim-sequencing. Leveraging the increasing availability of reference genomes, we show multiple applications of skim-seq for genomics-assisted breeding, including: (1) genotyping of segregating populations, (2) identification and genotyping of translocations, and (3) assessment of chromosome dosage, deletions and aneuploidy. These applications were evaluated in wheat doubled haploid populations, various introgression and aneuploid addition lines including wheat-barley translocations and *Thinopyrum-durum* wheat introgression lines, and monosomic wheat genetic stocks. Using variations on a single bioinformatics pipeline, 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 doubled haploid population.* We tested a 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<sup>40</sup>. Initially, F<sub>1</sub> hybrids were developed by crossing CDC Stanley and CDC Landmark and followed by planting of F<sub>1</sub> seeds. Spikelets from F<sub>1</sub> plants were emasculated and pollinated with maize pollen to induce haploid embryo development. Embryos were rescued and cultured in media to plantlets. The haploid plants were treated with colchicine to bring about chromosome doubling and generate primary DH seeds/plants. The primary DH seeds were germinated, and plants were self-pollinated to produce the DH<sub>0.1</sub> generation. For this study, 48 unique DH lines were used.

*Wheat 5D monosomic group.* A 5D monosomic line (TA3059), derived in the background of variety “Chinese Spring” (TA3008) 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, named CS M5D, included 839 self-pollinated progenies from TA3059, 16 standard Chinese Spring (TA3008) lines as internal controls and 9 blank samples for negative controls. These genetic stocks 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<sup>41,42</sup> 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)<sup>42</sup> where group 7 translocations including 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<sub>1</sub> hybrids. The F<sub>1</sub> was backcrossed with the respective recurrent parent to form BC<sub>1</sub> progenies for each cross combination. The final population included 335 BC<sub>1</sub> 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.* For *Thinopyrum intermedium*, a panel of 285 genets was evaluated, where genet refers to an individual with a unique genetic makeup<sup>43</sup>. The panel included 141 *Th. intermedium* genets, and 144 amphiploid genets derived from crossing *Th. intermedium x Triticum durum*. 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 induced by treating the young plants with colchicine. Plants with successful doubling of chromosomes were male-fertile and produced self-pollinated 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 by emasculating *Th. intermedium* plants as females followed by embryo rescue of the hybrid. The subsequent progenies were male sterile and were crossed again using *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 the males-sterile plants were crossed as female parents to *Th. intermedium*. 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<sup>44–48</sup>.

*Library construction.* Genomic DNA was extracted from leaf tissue collected from seedlings at the two- to three-leaf stage. The leaf tissues 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 manufacturer’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 incubated at 55 °C for 15 min, and then cooled to room temperature.

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 s), 55 °C (20 s), 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). The library was then cleaned, and the fragment size distribution was verified with an Experion™ DNA 1 K Reagents kit (#7,007,164) 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 X Ten.

**Bioinformatics pipeline.** The analysis pipeline described in this study (Fig. 1) can be used for a range of different genomics applications, including variant discovery and genotype calling, dosage estimation and identifying chromosome segments from different genomes. Processing pipelines for each case include the following steps:

**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 Ten 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/SkimSeq\\_Method](https://github.com/sandeshsth/SkimSeq_Method)). Based on the sequencing machine, the i5 index could also be reverse complement, which should be identified and the barcode file processed 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 bbdut 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 and cleaned the reads after demultiplexing using fastp (<https://github.com/OpenGene/fastp>).

For project data integrity, 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 confirmed 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 sample genome coverage per individual for each population using the following equation:

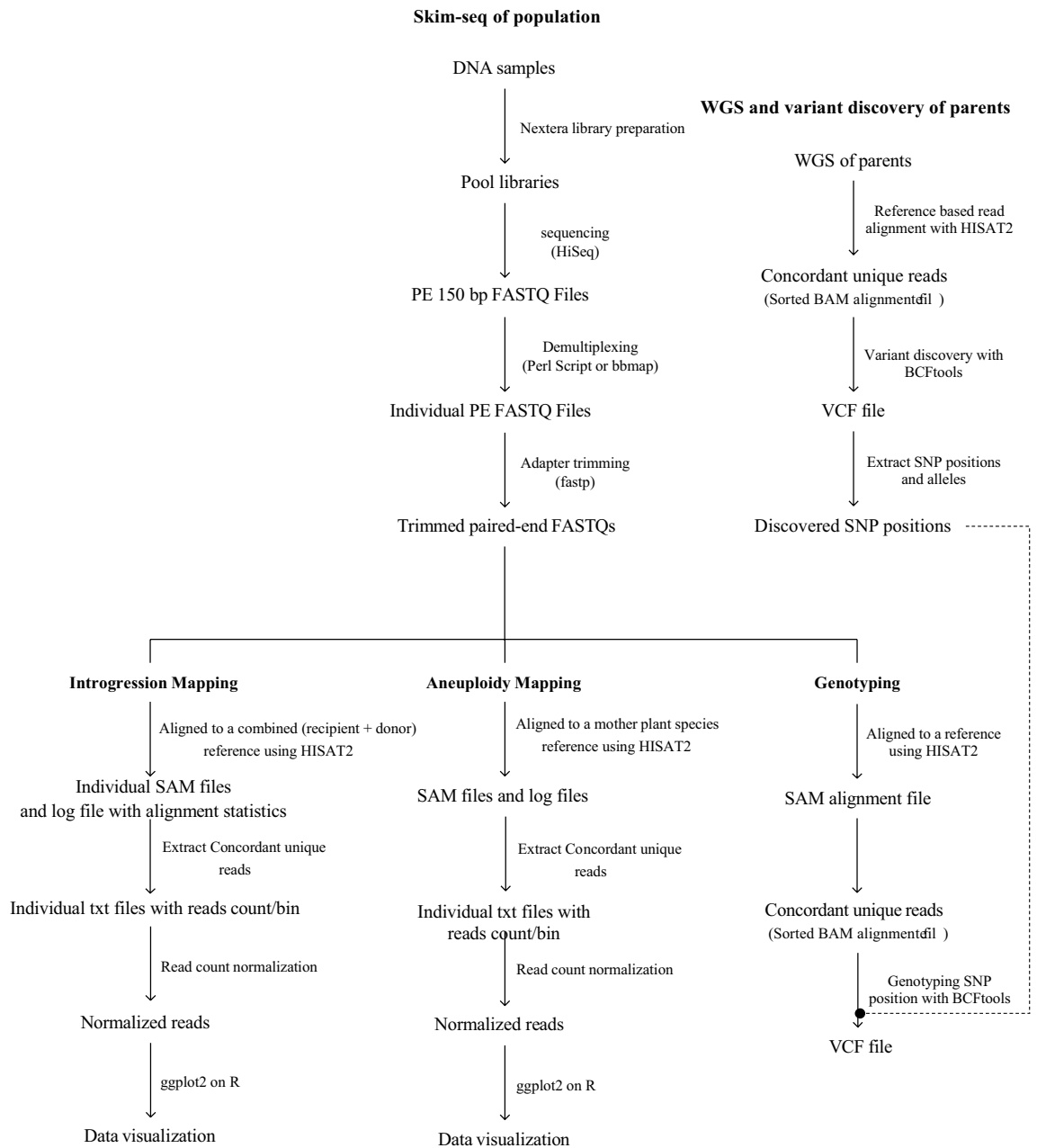
$$\text{genome coverage} = \frac{(\text{read count} * \text{read length} * 2)}{(\text{total genome size} * \text{total number of samples})} \quad (1)$$

**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, we used the ‘Chinese Spring’ RefSeq v1 assembly<sup>49</sup>.

For interspecific introgression mapping, a reference assembly was generated by concatenating the reference sequences of a donor and a recipient species as an “in silico interspecific hybrid”. For identification of wheat-barley group 7 introgressions, we combined the Chinese Spring reference genome v1.0<sup>49</sup> and barley pseudomolecule assembly of barley cv. Morex<sup>50</sup>. 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](https://phytozome-next.jgi.doe.gov/info/Tintermedium_v2_1)) developed from accession C4-5353T1. When combining reference genomes, all chromosomes or pseudomolecule names were specified to be unique.

HISAT2 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 line tools to filter for uniquely mapped concordant reads ([https://github.com/sandeshsth/SkimSeq\\_Method](https://github.com/sandeshsth/SkimSeq_Method)).

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



**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.

$$\frac{\text{normalized reads}}{\text{Mb}} = \frac{\text{sum of reads in Mb bin}}{\text{total number of reads per sample}} \times \text{Normalization Factor} \quad (2)$$

The normalization factor can be specified, where we used reads per 10 million or raw-read-counts/average read count in all bins. The script ([https://github.com/sandeshsth/SkimSeq\\_Method](https://github.com/sandeshsth/SkimSeq_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/SkimSeq\\_Method](https://github.com/sandeshsth/SkimSeq_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/SkimSeq\\_Method](https://github.com/sandeshsth/SkimSeq_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. Marking the centromere position with read depth 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 SNPs between the two parents followed by genotyping the discovered variants 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<sup>37,51</sup>. Alignment was done 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<sup>52</sup> 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*<sup>53</sup>. 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. 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  $\geq 6$  and  $\leq 100$  respectively and reference and alternate allele read depths of  $\geq 3$ . 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<sup>54</sup>. 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  $0.6\times$  to  $1.2\times$ . As the cost for sequencing to the higher depth for a genome with 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  $0.1\times$ ,  $0.05\times$ , and  $0.01\times$ . Sampling was completed using seqtk (<https://github.com/lh3/seqtk>), and the low-coverage samples were mapped and filtered as described above. The DH lines were then genotyped at the positions identified as variants between the parents with option *-T* using BCFtools.

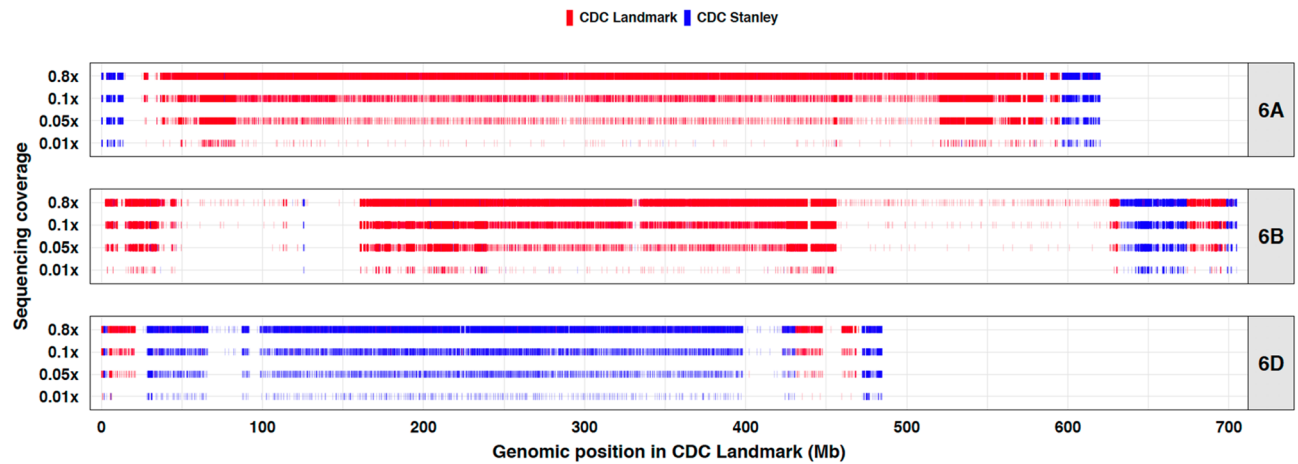
## Results

**Skim-seq pipeline.** To affordably genotype thousands of samples and effectively utilize the extremely high output of the latest sequencing platforms, we developed a modified low-volume Nextera library preparation 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 to genotype and characterize hexaploid wheat, we multiplexed from 48 samples up to 960 samples, giving raw genome coverage from  $\sim 1\times$  down to  $0.01\times$  of the very large,  $\sim 16$  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 (Fig. 1) were applied to several different use cases including SNP discovery and genotyping, introgression mapping, and aneuploidy analysis.

**SNP discovery and genotyping.** Nearly 26 million putative SNPs were identified from approximately  $8\times$  coverage of CDC Stanley and CDC Landmark. As CDC Landmark has a reference genome, the SNP variants were filtered for positions where CDC Stanley had the alternate allele compared to CDC Landmark. After filtering, a total of 12.5 million high-quality genome-wide SNPs were identified between the two parents and then used to genotype the same loci with the skim-seq data.

The average raw sequencing of 48 DH lines was  $0.88\times$  coverage with a range from  $0.61\times$  to  $1.23\times$ . From the total high-quality variants, 10.9 million unique SNPs were genotyped across the population (Supplementary Tables S2 and S3). The variants were assigned as parental alleles to either the CDC Stanley or CDC Landmark for genotyping the DH lines (Fig. 2). To simulate applications with higher plexing levels that would result in lower coverage, we decreased sample coverage through random down sampling to  $0.01\times$  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 (Fig. 2, Supplementary Figures S3). We did observe regions of the genome with low marker density between the two parents (e.g. 450–650 Mb on Chr. 6B) that are likely due to identity by descent with the closely related breeding germplasm.

Our ability to identify the genomic segments contributed by each of the two parents in the DH lines were evaluated by comparing the sequencing depths. The original sequencing depth was close to  $1\times$  coverage which was able to clearly identify the recombination breakpoints (Fig. 2). As we down-sampled, the density of markers



**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 0.8× followed by simulated 0.1 ×, 0.05 ×, and 0.01 × from the original).

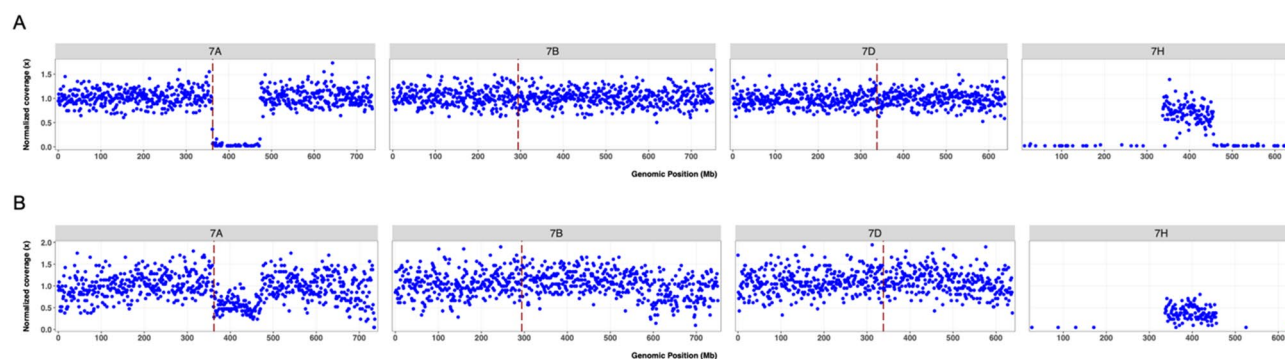
Population	Sample size (n)	Total reads in file	Average coverage*	Total reads in sample	Trimmed reads in samples	Total reads in overall alignment (%)	Total unique concordant reads and alignment (%)	Mapped paired-end reads per 1 Mb bin (mean)
Wheat-barley Group 7	384	485,575,828	0.025X	410,205,551	296,867,400	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	359,405,323	0.03X	302,850,841	258,410,843	185,564,827 (71.81)	103,832,640 (40.2)	61

**Table 1.** Different skim-seq populations, their genome coverage and related information. \*Average genome coverage per sample computed as (read count × read length (× 2))/(Genome size × n), where read length = 150 bp. Wheat genome size = 15 Gb. Intermediate wheat grass genome size = 12 Gb.

decreased but was still sufficient to distinguish haplotype blocks from each parent. At a sequencing depth of 0.05x, the boundaries of recombination intervals becomes less precise but still clearly defined. 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 (Fig. 2, Supplementary Figures S3). We also observed regions of high similarity with few variants (e.g. Figure 2: 460–620 Mb on Chr 6B) between CDC Landmark and CDC Stanley, likely due to identity-by-descent (IBD) between the two breeding lines from the same breeding program.

**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 for each sample, we were able to delimit both the size and the number of copies (dosage) of the barley translocation into the group 7 chromosomes of wheat (Fig. 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 (Fig. 3A).

The translocation lines are known to carry a group 7 translocation between wheat and barley on each of the three homoeologous chromosomes<sup>42</sup>. 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 94 Mb region from barley (337–431 Mb). We also observed a likely mispositioned scaffold in the Chinese Spring v1 reference at 327–337 Mb on Chr 7B as the presence of a region of wheat chromatin despite being in the middle of the translocation (Supplementary Figure S4). On chromosome 7D translocation, a larger wheat segment of 218 Mb (340–559 Mb) was replaced by a barley segment of 273 Mb (337–610 Mb). Skim-seq provided the physical position and size of translocations



**Figure 3.** Normalized read counts of a wheat-barley group 7 translocation (7AS.7HL-7AL) for (a) homozygous parent TA5798 [tissue id: DNA191014P04\_B11] and (b) heterozygous back-cross derived wheat-barley progeny [tissue id: DNA191014P01\_G11]. The dashed vertical lines indicate the position of wheat centromeres based on the IWGSC RefSeq v1 assembly<sup>49</sup>.

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	340–559	337–610		
7DS.7HL-7DL	2019–216-33_X_KS Silverado	26	16				
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	–	362–473	337–456		
7AS.7HL-7AL	TA5798	6	Homozygous				
7BS.7HL-7BL	TA5797	7	Homozygous			296–394	337–431
7DS.7HL-7DL	TA5799	6	Homozygous			340–559	337–610
	Chinese Spring	6	–				
	Blank	4	–				
Total		384	169				

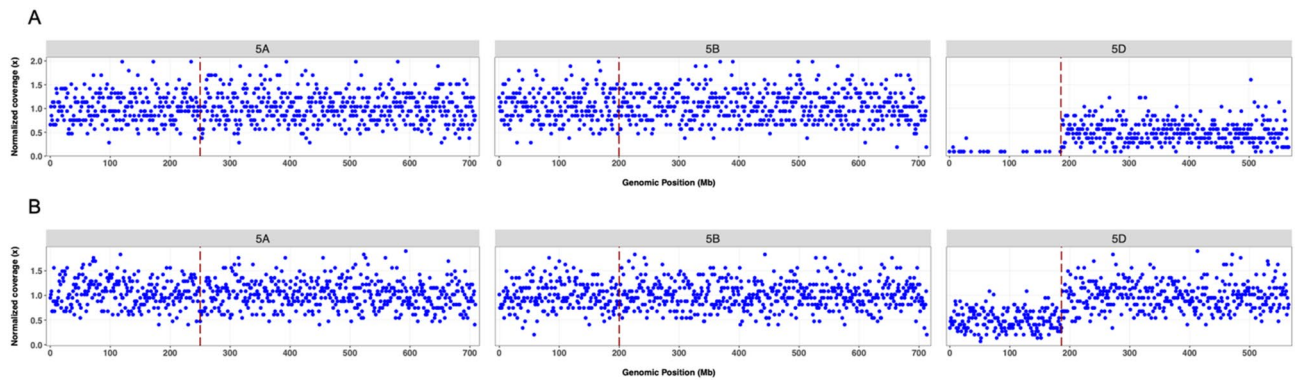
**Table 2.** Wheat-barley group 7 recombinants pedigree, number of samples in different groups, and translocation position information.

in introgression that could be easily used for further breeding work and very high-throughput genotyping of introgression lines.

For backcross-derived progeny, we observed the expected heterozygous translocation, as evidenced by read depth of approximately half the normalized read coverage compared to chromosomes with no translocations (Fig. 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 is a 1:1 ratio of carrier to non-carrier with  $X^2$  test ( $df = 1$ ,  $n = 335$ ) of 0.026 ( $P$ -value = 0.86), confirming typical Mendelian segregation.

**Aneuploidy mapping.** We sequenced the Chinese Spring chromosome 5D monosomic lines (CS M5D) to a target depth of 0.01x. Aligning the reads to the reference genome resulted in 30.6 reads on average per 1 Mb bin for an effective depth of  $0.0092 \times$  coverage (Table 1). The read depth was uniform across the genome except for chromosome 5D as expected for segregating chromosome dosage from the monosomic parent (Supplementary Table S4). As expected for dosage segregation from a monosomic individual, we observed four primary karyotypes in the progeny of the wheat 5D monosomic: euploid, monosomic, nullisomic and various telosomic plants. This enabled rapid identification of the rare telosomic lines, which are only a few percent, that result from breakage of the monosomic chromosome during meiosis (Fig. 4, Supplementary Table S5, Supplementary Figure S5). The mono-telosomic wheat lines have 20 chromosome pairs and a telosomic chromosome consist-





**Figure 4.** Normalized read counts for example individual samples from Chinese Spring monosomic 5D (CS-M5D; 20'' + 1'5D) populations showing telosomic 5DL [tissue id: DNA200317P02\_C03]. **(A)** monotelosomic 5DL line (20'' + t'5DL), carrying only one copy of the 5D chromosome long arm and **(B)** monotelosomic 5DL [tissue id: DNA200317P02\_G01] with one telosomic 5D chromosome (long arm) and one complete 5D chromosome (20'' + 1'5D + t'5DL). The dashed vertical lines indicate the centromere positions based on the IWGSC RefSeq v1 assembly<sup>49</sup>.

ing of one of the chromosome arms. Among the 864 samples, 674 (78%) were 5D monosomic, 130 (15%) were euploid, 35 (4%) were 5DL telosomic, 1 (<1%) was 5DS telosomic, and 7 (1%) were 5D nullisomic. Three other lines were 5D nullisomic and included other structural changes. Less than 0.6% (n = 5) of the samples failed to 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 0.03 × 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 in the 144 potential amphiploid *Th. Intermedium* × *T. durum* plants, skim-seq identified 108 (75%) individuals that had one or more wheat chromosomes. The wheat chromosome presence was variable with wheat chromosome 2A found only in three genets, whereas chromosome 3A was found in 77 genets. Within individuals, the number of alien chromosome 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 S6).

## Discussion

**Skim-seq: Cost and time effective genotyping approach.** The skim-sequencing approach presented in this study is broadly applicable for different genomics studies and molecular breeding that necessitate profiling a large number of samples in a timely and cost-effective manner. For example, for 5D monosomic lines we sequenced over 800 samples within a single lane of Illumina HiSeq X Ten, resulting in an average of 0.01 × coverage for a cost of approximately \$1.2 per sample. Although skim sequencing generates low-coverage data, this is sufficient for many applications. For instance, we showed that 0.01 × to 0.03 × coverage (Table 1) is sufficient to identify the size of introgressed segments from the alien species and to determine chromosome dosage. In addition, coverage as low as 0.01 × 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, and in the case of genotyping, high-coverage sequencing data on the parents. These resources are largely available, particularly for crop species, while continued advancements in highly accurate long-read sequencing are making the needed genome assemblies and genomic resources available 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.

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 which uses in-line barcodes, is limited to the number of barcodes that can be effectively combined as well as the upfront costs of synthesizing the adapters. However, the dual indexing for these skim-seq libraries utilizes 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 into 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–10 × coverage) for a few dollars. This makes the library construction costs and throughput an even larger consideration to keep the per sample costs low. The per sample library costs for skim-seq are in the range of \$1 per sample. Thus, the combined cost of

DNA extraction, library preparation and sequencing are less than \$3 per sample and suitable to provide sufficient sequencing data for many applications in most species. By example, the wheat genomes sequenced here are larger than other important crop species such as rice and maize<sup>55</sup>. The 0.03 × coverage obtained for the 16 Gb hexaploid wheat genomes in this study would be equivalent to over 1 × 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 lines and amphiploids. Compared to low-throughput, time-intensive cytological methods, skim-seq enabled the characterization of very large populations of amphiploids and lines carrying introgressions. Determining chromosome dosage in aneuploid lines is straightforward and could be used routinely to replace cytological approaches. While cytology will be necessary to confirm the exact composition of both addition lines and potentially translocated material, skim-seq provides a very effective way to rapidly screen for candidates that are most likely to have the desired chromosome composition 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, chromosome addition or deletion.

The generation of markers representing the whole-genome is essential for genetic studies. The skim-seq method presented here can generate markers with uniform 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. Alignment of the down sampled sequences to the reference genome showed that the skim-seq generated uniform distribution and sampling along the chromosomes and across the genome, even with extremely low coverage. 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. Likewise, low-coverage 0.01 × sequencing showed uniform density across the genome for determining the dosage of chromosome segments, with easy differentiation of zero, one or two copies. These lower levels of coverage can provide adequate data for routine genotyping, genomic selection, or progeny testing.

## Conclusions

In the 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, skim-seq can be a powerful method to identify translocations and introgressions, evaluate chromosomal dosage in aneuploidy stocks, and genotype segregating populations. 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 growing role in future plant breeding and genetic studies.

## 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 demultiplexed 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](https://datadryad.org/stash/share/v20dkVsSTj3toGn-CHG92eUSgre17uMT5AH_6LE2GDM).

Received: 20 September 2021; Accepted: 6 September 2022

Published online: 20 October 2022

## References

- Rasheed, A. *et al.* Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. *Mol. Plant* **10**, 1047–1064 (2017).
- 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).
- Poland, J. Breeding-assisted genomics. *Curr. Opin. Plant Biol.* **24**, 119–124 (2015).
- Davey, J. W. *et al.* Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**, 499–510 (2011).
- 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).
- 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).
- 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).
- 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).
- Varshney, R. K. *et al.* Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.* **31**, 240–246. <https://doi.org/10.1038/nbt.2491> (2013).
- Wang, L. *et al.* Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. *Genome Bio* **15**, R39. <https://doi.org/10.1186/gb-2014-15-2-r39> (2014).
- Qin, C. *et al.* Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proc. Natl. Acad. Sci.* **111**, 5135–5140. <https://doi.org/10.1073/pnas.1400975111> (2014).

12. 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).
13. Onda, Y., Takahagi, K., Shimizu, M., Inoue, K. & Mochida, K. Multiplex PCR targeted amplicon sequencing (MTA-Seq): Simple, flexible, and versatile SNP genotyping by highly multiplexed PCR amplicon sequencing. *Front. Plant Sci.* <https://doi.org/10.3389/fpls.2018.00201> (2018).
14. Saintenac, C., Jiang, D. & Akhunov, E. D. Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biol.* **12**, R88. <https://doi.org/10.1186/gb-2011-12-9-r88> (2011).
15. Shirasawa, K. *et al.* Target amplicon sequencing for genotyping genome-wide single nucleotide polymorphisms identified by whole-genome resequencing in peanut. *Plant Genome* <https://doi.org/10.3835/plantgenome2016.06.0052> (2016).
16. Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G. & Hohenlohe, P. A. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* **17**, 81–92. <https://doi.org/10.1038/nrg.2015.28> (2016).
17. Poland, J. A. & Rife, T. W. Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome* **5**, 92–102 (2012).
18. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**, e19379 (2011).
19. 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).
20. 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).
21. Juliana, P. *et al.* Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. *Nat. Genet.* **51**, 1530–1539 (2019).
22. Sehgal, D. *et al.* Haplotype-based, genome-wide association study reveals stable genomic regions for grain yield in CIMMYT spring bread wheat. *Front. Genet.* **11**, 1427 (2020).
23. Singh, N. *et al.* Genomic analysis confirms population structure and identifies inter-lineage hybrids in *Aegilops tauschii*. *Front. Plant Sci.* <https://doi.org/10.3389/fpls.2019.00009> (2019).
24. 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 Res.* <https://doi.org/10.1038/s41438-019-0132-8> (2019).
25. Kumar, A. *et al.* Genotyping-by-sequencing analysis for determining population structure of finger millet germplasm of diverse origins. *Plant Genome* <https://doi.org/10.3835/plantgenome2015.07.0058> (2016).
26. Wang, K. *et al.* Detection of selection signatures in chinese landrace and yorkshire pigs based on genotyping-by-sequencing data. *Front. Genet.* <https://doi.org/10.3389/fgenet.2018.00119> (2018).
27. Adhikari, L. *et al.* Genetic characterization and curation of diploid A-genome wheat species. *Plant Physiol.* **188**, 2101–2114. <https://doi.org/10.1093/plphys/kiac006> (2022).
28. Adhikari, L., Lindstrom, O. M., Markham, J. & Missaoui, A. M. Dissecting key adaptation traits in the polyploid perennial *Medicago sativa* using GBS-SNP mapping. *Front. Plant Sci.* <https://doi.org/10.3389/fpls.2018.00934> (2018).
29. 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).
30. 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. <https://doi.org/10.1534/g3.120.401728> (2020).
31. 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).
32. Jauhar, P. P. Modern biotechnology as an integral supplement to conventional plant breeding: The prospects and challenges. *Crop Sci.* **46**, 1841–1859 (2006).
33. Dempewolf, H. *et al.* Past and future use of wild relatives in crop breeding. *Crop Sci.* **57**, 1070–1082 (2017).
34. Kishii, M. An update of recent use of *Aegilops* species in wheat breeding. *Front. Plant Sci.* **10**, 585 (2019).
35. Gao, L. *et al.* The *Aegilops ventricosa* 2N'S segment in bread wheat: Cytology, genomics and breeding. *Theor. Appl. Genet.* **134**, 1–14 (2020).
36. Danilova, T. V., Poland, J. & Friebe, B. Production of a complete set of wheat–barley group-7 chromosome recombinants with increased grain  $\beta$ -glucan content. *Theor. Appl. Genet.* <https://doi.org/10.1007/s00122-019-03411-3> (2019).
37. 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).
38. Caruccio, N. Preparation of next-generation sequencing libraries using Nextera™ technology: simultaneous DNA fragmentation and adaptor tagging by *in-vitro* transposition, in *High-Throughput Next Generation Sequencing*, p. 241–255 (Springer, 2011).
39. 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).
40. Santra, M., Wang, H., Seifert, S. & Haley, S. Doubled haploid laboratory protocol for wheat using wheat–maize wide hybridization. In *Wheat Biotechnology* 235–249 (Springer, 2017).
41. 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  $\beta$ -glucan. *Theor. Appl. Genet.* **131**, 377–388. <https://doi.org/10.1007/s00122-017-3008-z> (2018).
42. Danilova, T. V., Poland, J. & Friebe, B. Production of a complete set of wheat–barley group-7 chromosome recombinants with increased grain  $\beta$ -glucan content. *Theor. Appl. Genet.* **132**, 3129–3141. <https://doi.org/10.1007/s00122-019-03411-3> (2019).
43. Zhang, X. *et al.* Establishment and optimization of genomic selection to accelerate the domestication and improvement of intermediate wheatgrass. *Plant Genome* <https://doi.org/10.3835/plantgenome2015.07.0059> (2016).
44. Fedak, G. & Han, F. Characterization of derivatives from wheat–*Thinopyrum* wide crosses. *Cytogenet. Genome Res.* **109**, 360–367. <https://doi.org/10.1159/000082420> (2005).
45. 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. *Theor. Appl. Genet.* **84**, 899–905 (1992).
46. 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. *Theor. Appl. Genet.* **109**, 1070–1076 (2004).
47. Hayes, R. *et al.* Perennial cereal crops: An initial evaluation of wheat derivatives. *Field Crop Res.* **133**, 68–89 (2012).
48. 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).
49. Appels, R. *et al.* Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* <https://doi.org/10.1126/science.aar7191> (2018).
50. Mascher, M. *et al.* A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427–433 (2017).
51. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
52. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

53. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
54. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890. <https://doi.org/10.1093/bioinformatics/bty560> (2018).
55. Varshney, R. K. *et al.* Designing future crops: Genomics-assisted breeding comes of age. *Trends Plant Sci* **26**, 631–649. <https://doi.org/10.1016/j.tplants.2021.03.010> (2021).

## Acknowledgements

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

## Author contributions

L. A.: Methodology, Software, Formal analysis, Investigation, Data Curation, Writing—Original Draft, Visualization; S. S.: Methodology, Software, Formal analysis, Investigation, Data Curation, Writing—Original Draft, Visualization; S. W.: Methodology, Investigation, Validation; J. C.: Investigation, Validation, Formal analysis, Data Curation, Visualization; L. G.: Software, Investigation, Data Curation; B. E.: Investigation, Resources, Data Curation; D. W.: Investigation, Resources; Y. J.: Investigation, Validation, Formal analysis; D. H. K.: Investigation, Validation, Formal analysis, Resources; P. H.: Resources, Funding acquisition; C. P.: Resources, Funding acquisition, Writing—Review & Editing; S. W.: Resources, Writing—Review & Editing; X. W.: Methodology, Investigation; J. W.: Methodology, Investigation; J. C. G.: Conceptualization, Methodology, Investigation, Project administration, Funding acquisition, Writing—Review & Editing; L. D.: Conceptualization, Resources, Project administration, Funding acquisition; B. F.: Conceptualization, Resources, Supervision; J. P.: Conceptualization, Methodology, Validation Investigation, Resources, Writing—Original Draft, Writing—Review & Editing, Visualization, Supervision, Project administration, Funding acquisition; All authors: Writing—Review & Editing.

## 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. We thank the Department of Energy Joint Genome Institute, the Perennial Agriculture Project, and The Land Institute for prepublication access to the *Thinopyrum intermedium* genome sequence.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-19858-2>.

**Correspondence** and requests for materials should be addressed to J.P.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022