Published November 25, 2015

RESEARCH

Comparing Genotyping-by-Sequencing and Single Nucleotide Polymorphism Chip Genotyping for Quantitative Trait Loci Mapping in Wheat

Prabin Bajgain,* Matthew N. Rouse, and James A. Anderson*

ABSTRACT

Array- or chip-based single nucleotide polymorphism (SNP) markers are widely used in genomic studies because of their abundance in a genome and lower cost per data point than older marker technologies. Genotyping-by-sequencing (GBS), a relatively newer approach of genotyping, suggests equal appeal because of its lesser cost per data point and the avoidance of ascertainment bias during genotyping. In this study, we compared the results from quantitative trait loci (QTL) mapping, marker distribution on linkage maps, genome size, recombination sites covered by the markers, and cost per polymorphic marker, as well as the methodology and workflow between the Illumina Infinium 9000 SNP-chip genotyping with GBS. Results indicate that while GBS offers similar genome coverage at almost one-fourth the cost of SNP chip, the SNP-chip method is less demanding of computational skills and resources. Eight and nine QTL were detected in the GBS and SNP-chip datasets, respectively, with one QTL common between the systems. Additionally, imputation accuracy of the GBS dataset was examined by introducing missing values randomly and imputing the missing alleles using a probabilistic principal components algorithm. Imputation results suggest recovery of the missing alleles with reasonable accuracy in datasets with low (up to 40%) amount of missing data is possible and can provide acceptable accuracy in gene mapping. Overall, the comparative results indicate that both approaches provide good genome coverage and similar mapping results. The choice of the genotyping platform is decided by the nature of the study and available resources.

P. Bajgain and J. A. Anderson, Dep. of Agronomy and Plant Genetics, Univ. of Minnesota, St. Paul, MN 55108, USA; P. Bajgain, Dep. of Agronomy, Purdue Univ., 915 West State Street, West Lafayette, IN 5 47907, USA; and M. N. Rouse, USDA–ARS, Cereal Disease Lab., and Dep. of Plant Pathology, Univ. of Minnesota, St. Paul, MN 55108, USA. Received 22 June 2015. Accepted 22 Sept. 2015. *Corresponding author (bajga002@umn.edu; ander319@umn.edu).

Abbreviations: 9K, 9000 Infinium iSelect SNP assay; CIM, composite interval mapping; CSS, chromosome survey sequence; gb, gigabytes; GBS, genotyping-by-sequencing; LOD, logarithm of odds; PCA, principal component analysis; PIC, polymorphism information content; PVE, phenotypic variance explained; QTL, quantitative trait loci; RIL, recombinant inbred line; SNP, single nucleotide polymorphism.

The era of molecular markers, which began in earnest with restriction fragment length polymorphism markers in the late 1980s, has evolved remarkably. Restriction fragment length polymorphism markers were followed by polymerase-chain-reaction-based marker technologies such as randomly amplified polymorphic DNA, amplified fragment length polymorphism, and simple-sequence repeat markers in the late 1990s (Gupta et al., 2010). The discovery and implementation of these marker systems facilitated the mapping of genes controlling many traits, including several stem rust resistance genes in wheat (*Triticum aestivum* L.). Recently, SNP markers are increasingly being used in gene and QTL mapping approaches, primarily because of the lower cost per data point and the relative ease in assay design as well as in scoring and interpretation of the results. The SNP-chip-based genotyping is often preferred, as it is adaptable to high-throughput systems.

The SNP markers on SNP chips or arrays are discovered on a diversity panel and selected to be included on the genotyping chip based on allele frequency, polymorphism information content,

Published in Crop Sci. 56:1–17 (2016).
doi: 10.2135/cropsci2015.06.0389

and marker segregation. However, they introduce ascertainment bias in downstream applications (Albrechtsen et al., 2010; Li and Kimmel, 2013). This bias can impact the study of genetic relationships among individuals that may not be easily corrected (Moragues et al., 2010; Frascaroli et al., 2013). Genotyping-by-sequencing, conducted directly on the population of interest, is one such method free from ascertainment bias (Poland et al., 2012).

The GBS procedure allows the discovery of population-specific SNPs from sequencing of DNA libraries obtained after restriction digest of samples. The usability of this approach in genotyping a crop species was first demonstrated in maize (*Zea mays* L.) (Elshire et al., 2011), and has been successfully used in other crop species including wheat (Saintenac et al., 2013), and barley (*Hordeum vulgare* L.) (Liu et al., 2014). In addition to the high-quality polymorphism data free of ascertainment bias produced by GBS, the lower cost per sample (and also cost per polymorphic marker) gives this approach greater appeal for genomic studies in crops and noncrop species. Because the GBS approach yields sequence data as opposed to only the allele calls in other existing genotyping approaches, the likelihood of using the sequences for in silico annotation for functional polymorphism adds to the functionality of this genotyping approach.

One potential problem of the GBS approach, however, is the generation of missing allele calls when a large number of lines are multiplexed during sequencing (Williams et al., 2010; Poland et al., 2012; Fu et al., 2014). As more samples are multiplexed for sequencing, the number of reads per sample decreases, leading to a higher frequency of missing allele calls during SNP detection in the population. Construction of accurate linkage maps may not be possible with a genotype matrix with a large proportion of missing allele calls. There are two ways of solving this problem: (i) increasing the sequencing depth either by multiplexing fewer individuals to generate DNA libraries or by sequencing the libraries multiple times and (ii) by imputation of missing genotyped calls using relationship among lines in a population. Of the two, the former requires additional time and resources to sequence and analyze the DNA libraries multiple times and may be less desired. The latter is increasingly preferred because of the richness in data mining offered by the GBS procedure expands to imputation of missing haplotypes in related individuals with low read coverage during sequencing. Recent studies in several crop species, including wheat, have shown that imputing markers to construct missing haplotypes can be done with relatively high accuracy (Rutkoski et al., 2013; Fu, 2014). Thus, in populations where the individuals are either linked in families (F_2, recombinant inbred line [RIL] populations) or via population structure (association mapping panels), imputation offers promising results. Imputation of missing genotypes might be mostly beneficial in populations where the allelic origin based on the parents is known or if coancestry information can be used. However, caution must be applied in datasets with too much missing data or in panels with unrelated individuals.

The main objective of this study was to compare the GBS approach with the 9000 Infinium iSelect SNP assay (9K)-based (Cavanagh et al., 2013) genotyping method in mapping QTL associated with resistance to stem rust of wheat. We also investigated the effect of missing alleles in GBS dataset on QTL mapping. Attributes such as recombination frequency and genome coverage between these two high-throughput genotyping methods were compared. Additionally, difference in the power of QTL detection on simulated trait data and the practicality of using one approach over another were explored.

MATERIALS AND METHODS

Molecular Marker Assay

A RIL population comprising 141 $F_{6:7}$ RILs derived from the cross RB07/MN06113-8 was used in this study (Bajgain et al., 2015). Genomic DNA was extracted from ground seeds of the parents and all RILs using a modified cetyltrimethylammonium bromide protocol (Kidwell and Osborn, 1992). The extracted DNA was quantified using an ND 1000 Spectrophotometer (NanoDrop Technologies). The population was genotyped using two approaches: (i) 9K (Cavanagh et al., 2013) and (ii) SNP markers obtained from GBS (Elshire et al., 2011).

For genotyping using the 9K, DNA suspended in ddH_2O at approximately 80 ng μL^{-1} was submitted to the USDA–ARS Small Grain Genotyping Center in Fargo, ND. The data generated was manually called using Illumina's GenomeStudio 2011.1 (Illumina Inc.). Briefly, each SNP call across the RIL population was manually analyzed, curated, and exported in diploid format (AA/AB/BB). Monomorphic markers, markers with the same calls for the entire population, markers with more than 10% missing data, and markers that deviated from 1:1 segregation ratio were discarded. Markers with 5% or less heterozygous calls were retained to avoid false purging of heterozygous loci. This resulted in 1050 high-quality markers that were retained for linkage mapping.

In the GBS approach, a double-digested library was created using the restriction enzymes *PstI* and *MspI* on 200 ng of DNA per sample following Poland et al. (2012). Each library was 76-plexed with the parents repeated six times each. The libraries were sequenced in two lanes of Illumina HiSeq 2000, generating 100-bp paired-end sequences. The sequences were processed using the UNEAK pipeline (Lu et al., 2013) using the parameters -c 10 -e 0.025 to obtain de novo SNPs. Reads containing SNPs were used as query sequences and BLASTn searched against the wheat chromosome survey sequences (CSS) to assign SNPs to unique chromosomes. The wheat CSS were obtained by assembling reads obtained from sequencing flow-sorted wheat chromosomes from the 'Chinese Spring' variety (International Wheat Genome Sequencing Consortium, http://wheaturgi.versailles.inra.fr/Seq-Repository/). To ensure that correct SNPs were obtained, only the full-length alignment of a query sequence with the survey sequences allowing either

one base mismatch or one gap was permitted. To circumvent retaining redundant SNPs on paralog sequences and duplicated regions among the A, B, and D subgenomes, SNPs thus obtained were filtered to remove SNPs that mapped more than once to multiple chromosomes. The SNPs that were monomorphic, had no allele calls for >10 individuals (>7% missing data), or were heterozygous in >10 individuals (7% heterozygosity) were also discarded. The SNPs obtained after these steps were converted to diploid format (AA/AB/BB) from allelic phases (A/C/G/T). This process resulted in 932 high-quality SNP loci that were retained for linkage mapping.

Linkage Map Construction and Quantitative Trait Loci Mapping

Linkage groups were constructed from SNPs obtained from both genotyping approaches (9K and GBS) using Mapdisto version 1.7.7.0.1 (Lorieux, 2012) and using a minimum logarithm of odds (LOD) values of 3.0. Genetic distances between the markers were calculated based on the Kosambi mapping function (Kosambi, 1943). Phenotypic data (stem rust severity) were collected on the RB07/MN06113-8 RIL population (Bajgain et al., 2015). The program Windows QTL Cartographer 2.5_011, which implements composite interval mapping (CIM) method to identify QTL, was used to analyze marker–trait associations (Wang et al., 2012). A walk speed of 1 cM was used for QTL detection on linkage groups, and a QTL was declared to be present if the LOD threshold was calculated by 1000 permutations at $\alpha = 0.05$. The QTL effects were estimated as the proportion of phenotypic variance explained (PVE) by the QTL.

Imputation of Genotyping-by-Sequencing Single Nucleotide Polymorphisms

Construction of linkage maps for genome mapping can be a difficult task if a dataset is missing a significant portion of allele calls. In this dataset, the samples were 76-plexed, and therefore the issue of missing data was not egregious (<7% missing data; Bajgain et al., 2015). However, to simulate scenarios where missing data could be a problem, the genotype matrix comprising 932 GBS SNPs for the RIL population was modified to introduce missing allele calls. Missing values were introduced randomly in the GBS dataset using R 3.0.2 (R Development Core Team, 2013) to simulate the genotype matrix with 20, 30, 40, 50, 60, 75, and 90% missing data. These datasets are hereafter referred as GBS20, GBS30, GBS40, GBS50, GBS60, GBS75, and GBS90, respectively.

Imputation of missing SNPs on the simulated data was done using principal component analysis (PCA)-based imputation using the probabilistic PCA (ppca) algorithm in the R package pcaMethods (Stacklies et al., 2007). The ppca algorithm first assigns row average values to the missing values and then uses the singular value decomposition of the SNP matrix to create orthogonal principal components. In turn, the principal component values corresponding to the largest eigenvalues are used to reconstruct the missing SNP genotypes in the genotype matrix. The algorithm ppca was chosen for its high imputation accuracy and efficiency in regards to the use of computational resources compared with other imputation algorithms of similar caliber (Moser et al., 2009; Fu, 2014). In the algorithm, 25 PCA values were used to reconstruct all genotype matrices with missing data. Prediction values <2 were assigned the homozygous genotype, 1, to represent alleles originating from the first parent; values equal to 2 were assigned the heterozygous genotype, 2; and values >2 were assigned the homozygous genotype, 3, to represent alleles originating from the second parent. These values were assigned after careful and replicated manual scans of the imputed data to predict the true genotypes as accurately as possible.

Summary Statistics of Genotype Matrices

The number of recombinations in each RIL on all datasets (9K, GBS nonimputed, and GBS imputed) was estimated using the R package hsphase (Ferdosi et al., 2014). Both imputed and nonimputed GBS datasets were analyzed using PowerMarker V3.25 (Liu and Muse, 2005) to estimate population attributes. The method of moments estimator was used to estimate the within-population inbreeding coefficient and relatedness between the individuals (Ritland, 1996). Polymorphism information content (PIC), a diversity measure among the individuals in a population, was calculated according to Botstein et al. (1980). Estimation of these population parameters was done with 10,000 nonparametric bootstraps across different loci at a confidence interval of 95% ($\alpha = 0.05$).

Methodology and Workflow Comparison

The cost, time, and resources required to genotype the RIL population using both (9K and GBS) methods were compared so that some characteristics between the two methods could be understood to determine their usability in programs and projects that are similar to ours. The discussed cost estimate strictly pertains to the genotyping cost and does not include the cost of the manual labor involved. The amount of time required to genotype the population is the active time used in preparation of the DNA samples, sequencing, and data analysis and does not include the latent time between procedures. Computational resources and skills needed to analyze the data obtained from both procedures are also briefly discussed.

RESULTS AND DISCUSSION

Genotype Properties

Genotype calls obtained for both the 9K SNPs and GBS SNPs were first analyzed for deviation from expectation segregation ratio of 1:1. However, as described in the Materials and Methods section, the 9K dataset (1050 SNPs) was devoid of markers deviating from the 1:1 segregation ratio, as markers that deviated from the ratio were discarded. Of the 932 GBS SNP markers, 164 were found to be skewed toward either parental genotype. Of these 164 markers, 48 were overrepresentative of the MN06113-8 genotype, whereas 116 were overrepresentative of the RB07 genotype. Overall, 49.5% of marker genotypes were inherited from MN06113-8 in the 9K dataset, 50% from RB07, and 0.5% were heterozygous. In the GBS dataset, 46% of the marker genotypes originated from MN06113-8, 49% from RB07, and 3% were heterozygotes, with 2% missing data. The RILs used for genotyping were inbred to

Table 1. Results of linkage group formation using the 9000 Infinium iSelect single nucleotide polymorphism (SNP) assay (9K) and genotyping-by-sequencing (GBS) SNPs.

	Chromosome																					
	1A	2A	3A	4A	5A	6A	7A	1B	2B	3B	4B	5B	6B	7B	1D	2D	3D	4D	5D	6D	7D	Total
(A) 9K method																						
Linkage groups†	2	4	1	1	3	2	1	2	1	1	1	1	4	1	1	0	1	0	1	2	0	30
SNPs‡	76	32	138	26	23	121	43	17	110	7	82	145	12	66	23	0	2	0	21	18	0	
Size (cM)§	75	38	156	61	31	76	105	57	141	47	94	168	34	95	73	0	1	0	24	18	0	
SNPs per subgenome			459							439								64				962
Size per subgenome			542							636								116				1294
(B) GBS method																						
Linkage groups†	1	2	1	2	1	3	1	2	1	1	1	1	2	1	1	1	1	2	2	2	2	31
SNPs‡	61	29	56	63	15	87	80	62	131	40	61	117	11	46	10	16	3	11	13	9	4	
Size (cM)§	95	45	146	58	23	36	98	80	111	43	90	147	25	83	29	44	23	35	64	12	18	
SNPs per subgenome			391							468								66				925
Size per subgenome			501							579								225				1305

† Number of linkage groups formed for each wheat chromosome.

‡ Number of SNPs that mapped to all linkage groups representing each chromosome.

§ Size of the linkage groups combined if more than two linkage groups were observed for a chromosome.

$F_{6:7}$ generation, and as such, only 1.6% of genotype calls in the population were expected to be heterozygous, and these numbers observed in both genotyping methods are expected of a highly inbred population. The results also indicated a higher proportion of heterozygous and missing allele calls in the GBS dataset, as permitted during the filtering of genotype calls.

Linkage Groups Construction

Construction of linkage groups for both 9K and GBS datasets was done using the same parameters in Mapdisto version 1.7.7.0.1 (Lorieux, 2012). Of 1050 9K SNP markers, 964 were placed in 30 linkage groups representing 18 wheat chromosomes (Table 1). Chromosomes 2D, 4D, and 7D were not represented by any marker. The number of markers per linkage group ranged from two (chromosomes 2A, 3D, and 6B) to 145 (chromosome 5B) with an average of 32 markers per linkage group. The 964 markers distributed over the 30 linkage groups covered a total of 1294 cM of the wheat genome. The sizes of the smallest and the largest linkage groups were 0.4 and 168 cM, respectively, with an average size of 43 cM. The average centromeric range (distance between the last SNP marker on the short arm and the first SNP marker on the long arm of the linkage groups) was 21 cM and varied from 12 to 28 cM. Several chromosomes were represented by two or more linkage groups that constituted of markers from either arm only, and therefore, it was not possible to determine the position of the centromere in these chromosomes.

Similarly, 925 of the 932 GBS SNP markers were assigned to 31 linkage groups that represented all 21 wheat chromosomes (Table 1, 2). The marker number in these linkage groups ranged from two (chromosomes 4A, 6B, 6D, and 7D) to 131 (chromosome 2B) with an average of 30 markers per linkage group. The smallest linkage group was 0.8 cM, the largest group was 147 cM long, and the average size of all linkage groups was 42 cM. The size of the wheat genome covered by the 31 linkage groups constructed using 925 GBS SNP markers was comparable with that of the 9K dataset at 1305 cM. The average centromeric range was 20 cM and varied from 6 to 31 cM. Similar to linkage groups constructed using 9K dataset, position of the centromere was not determined for several chromosomes as the linkage groups consisted of markers from either arm only.

Table 2: Comparison of linkage mapping results among the 9000 Infinium iSelect single nucleotide polymorphism (SNP) assay (9K) dataset, nonimputed genotyping-by-sequencing (GBS) dataset, and datasets with 40 and 75% missing allele calls.

Method†	SNPs‡	Linkage groups	SNPs in linkage groups	Unlinked SNPs	Genome size	QTL
			%		cM	
9K	1050	30	91.8	8.2	1295	9
GBS	932	31	99.2	0.8	1306	8
GBS40	932	29	98.8	1.2	4785	8
GBS75	932	30	89.2	10.8	14,879	7

† GBS40, imputed GBS dataset with 40% missing allele calls; GBS75, imputed GBS dataset with 75% missing allele calls.

‡ Number of polymorphic markers used in linkage mapping and imputation of missing genotype data.

One immediately obvious advantage of GBS over the SNP-chip genotyping was the slightly better coverage of the wheat D genome. The D genome of wheat is often the least represented in genotyping platforms, owing to its lower frequency of polymorphic sequences (Chao et al., 2009; Allen et al., 2011). While the number of markers mapped to the D genome in our GBS dataset is not as large as the A and B genomes, the retrieval of all seven D chromosomes during construction of linkage groups was a good indication that the GBS approach can be manipulated to obtain more SNP markers from the D genome. One

possible way to achieve this is to use less stringent filtering parameters than the ones used in our study. However, a more reliable approach may be to map the reads obtained from sequencing of GBS libraries to the respective subgenomes of wheat to obtain subgenome-specific polymorphic markers. With two of three of wheat's subgenomes already sequenced—A genome from wheat's diploid ancestor *Triticum urartu* Tumanian ex Gandilyan (Ling et al., 2013) and D genome from *Aegilops tauschii* Coss. (Jia et al., 2013)—subgenome-specific mapping of sequence reads would help to retain reads that would otherwise be discarded during SNP calling as a result of mismatches among the reads originating from the A, B, and D subgenomes.

Recombinations and Genome Coverage

The highest number of recombinations among all three subgenomes in the 9K dataset was observed in the A genome (1791) whereas the B genome recorded the most recombinations in the GBS dataset (1608). In both datasets, the D genome had the fewest recombination events with 369 recombinations in the 9K dataset and 691 in the GBS dataset.

The average number of recombinations per RIL in both 9K and GBS genotype matrices was found to be 27. Sixty-five individuals (46.1%) had more recombinations than the average in 9K dataset, compared with 78 (55.3%) in the GBS dataset. Only two individuals had the same number of recombinations in both datasets. The total number of recombinations in the 9K matrix was 3746, slightly lower than that in the GBS matrix at 3790.

To visualize these results, recombination blocks per line observed in both genotype matrices were plotted against the SNPs in each chromosome (Fig. 1). As seen in the figure, both congruous and incongruous patterns of recombination blocks exist in the two genotype matrices. The difference in reported recombination breakpoints likely arises from the difference in assay design between the two genotyping methods as different parts of the genome might have been sampled, which alters the genome coverage in these two genotyping methods. This can potentially lead to representation of different haplotype matrices, which results in the detection of different sites and number of recombination events. The difference in genome coverage is also corroborated by the difference in properties of the linkage groups constructed using SNPs obtained from the two methods. This is illustrated in Fig. 1C where the portions of the genome captured by the two methods are compared. The figure shows the differences in genome sampling between the two genotyping methods in one RIL (MN06_01) on all 21 chromosomes. Such difference is also observed in a larger set of RILs (randomly chosen and inspected) in the population in general (data not shown). We believe that the accumulation of such differences over the whole genome across the population is the reason behind the observation of incongruous recombination patterns.

To uncover the genetic architecture controlling the traits of interest, the use of molecular markers representative of the genome is important in gene mapping studies. Markers that are significantly linked to the trait can provide remarkable improvements in breeding for allele enrichment and trait improvement. Since the choice of genotyping platform can impact the quantity of molecular markers and their distribution in different genomic regions, understanding the differences in marker-related genome properties can assist in such choice among the different available genotyping approaches. In our investigation of such properties between the two genotyping methods discussed here, the difference in number of polymorphic markers identified and used in creating linkage groups was not strikingly different; whereas, the distribution of those markers in different genomic regions as shown by the difference in sites of recombination was noteworthy. The difference in genome coverage is also supported by the observed differences in properties of the linkage groups constructed using SNPs obtained from the two methods and the failure to detect the same QTL between the two methods, as described in upcoming sections.

Quantitative Trait Loci Mapping with Field Data

Mapping of QTL associated with resistance to stem rust of wheat was performed in the RIL population using stem rust severity data collected at three locations over four seasons. This was done to assess the impact on QTL mapping using linkage groups constructed from markers obtained from the two genotyping approaches. For detailed information on disease phenotyping and data statistics, see Materials and Methods in the recently published study by Bajgain et al. (2015).

The CIM method of QTL mapping detected nine QTL in the 9K dataset on linkage groups representing the chromosomes 2B, 3A, 4A, 4B, 5B, and 6D (Table 3). Similarly, eight QTL distributed on chromosomes 1A, 2A, 2B, 2D, 4A, 4B, and 7A were detected in the GBS dataset (Table 4). One QTL was common in both datasets, that is, the QTL detected on the short arm of chromosome 2B (Fig. 2). This QTL was detected in all environments in both datasets and also explained the largest amount of explained phenotypic variation in both datasets. The additive effects of the parental alleles toward disease resistance were estimated quite accurately by markers in both datasets. Taking the example of the common QTL (2B.2), the range of difference in estimation of allelic effects between the two datasets was found to be within ±1% (−0.8 to +0.6%). This suggests that the adoption of either marker system for QTL mapping should not significantly influence prediction of allelic effects of the detected QTL.

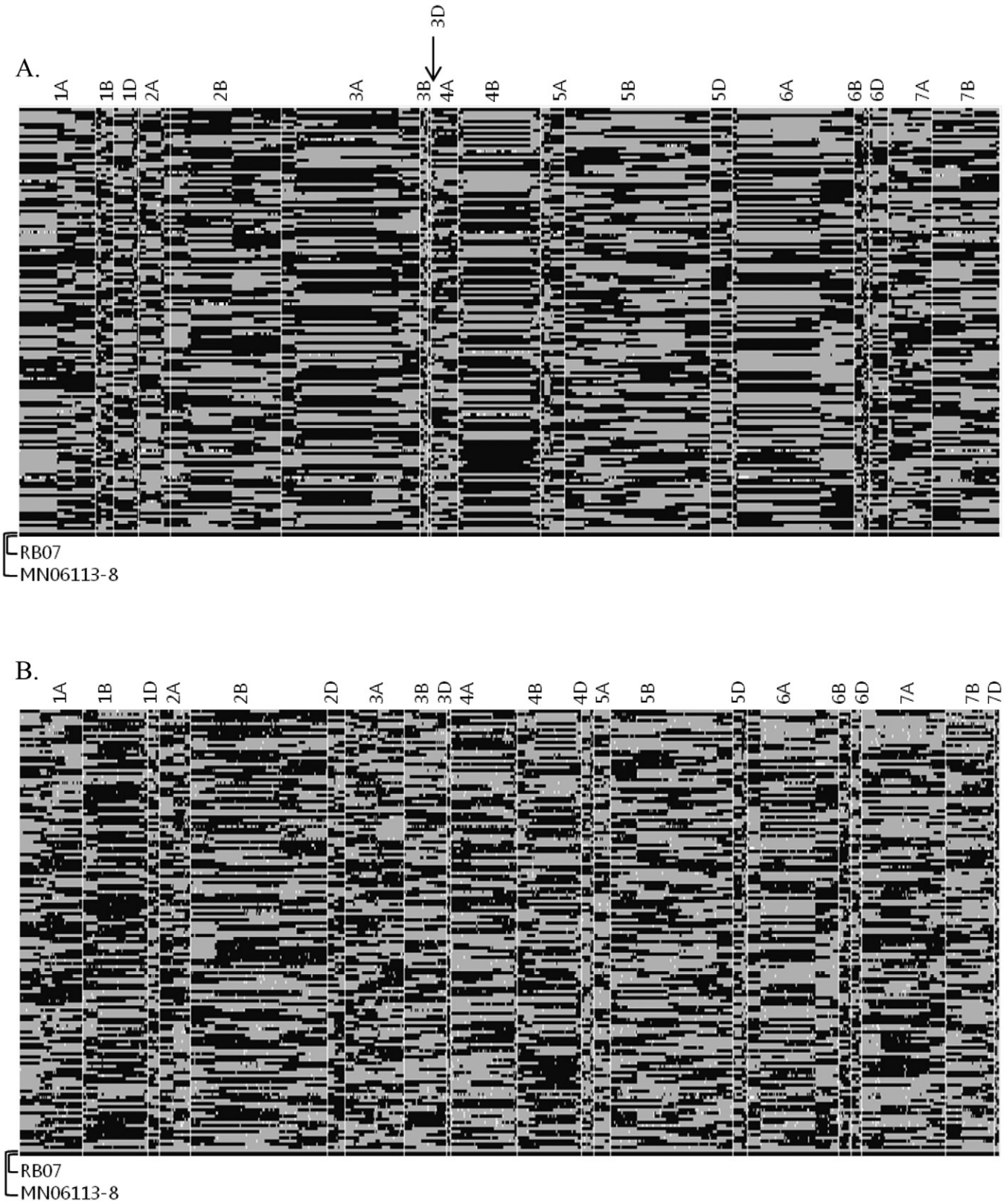

Figure 1. (continued on next page) Recombination blocks observed in (A) 9000 Infinium iSelect single nucleotide polymorphism (SNP) assay (9K) and (B) genotyping-by-sequencing (GBS) genotype datasets. The population is arranged in descending order on the *y*-axis and SNPs are arranged by the chromosomes they belong to on the *x*-axis separated by the white vertical bar. Panel (C) represents an example of the difference in genome coverage between the 9K and GBS methods in the recombinant inbred line (RIL) 'MN06_1' across all 21 chromosomes. The size difference within each chromosome (for example, within 1A_9K and 1A_GBS) is due to the differences in number of SNP markers between the two methods that are distributed along the chromosome. No linkage groups were obtained for chromosomes 2D, 4D, and 7D using the 9K SNPs. In all panels, the colors gray and black represent MN06113-8 and RB07 haplotype blocks, respectively, whereas the white dots (white vertical lines on panel C) indicate missing data.

C.

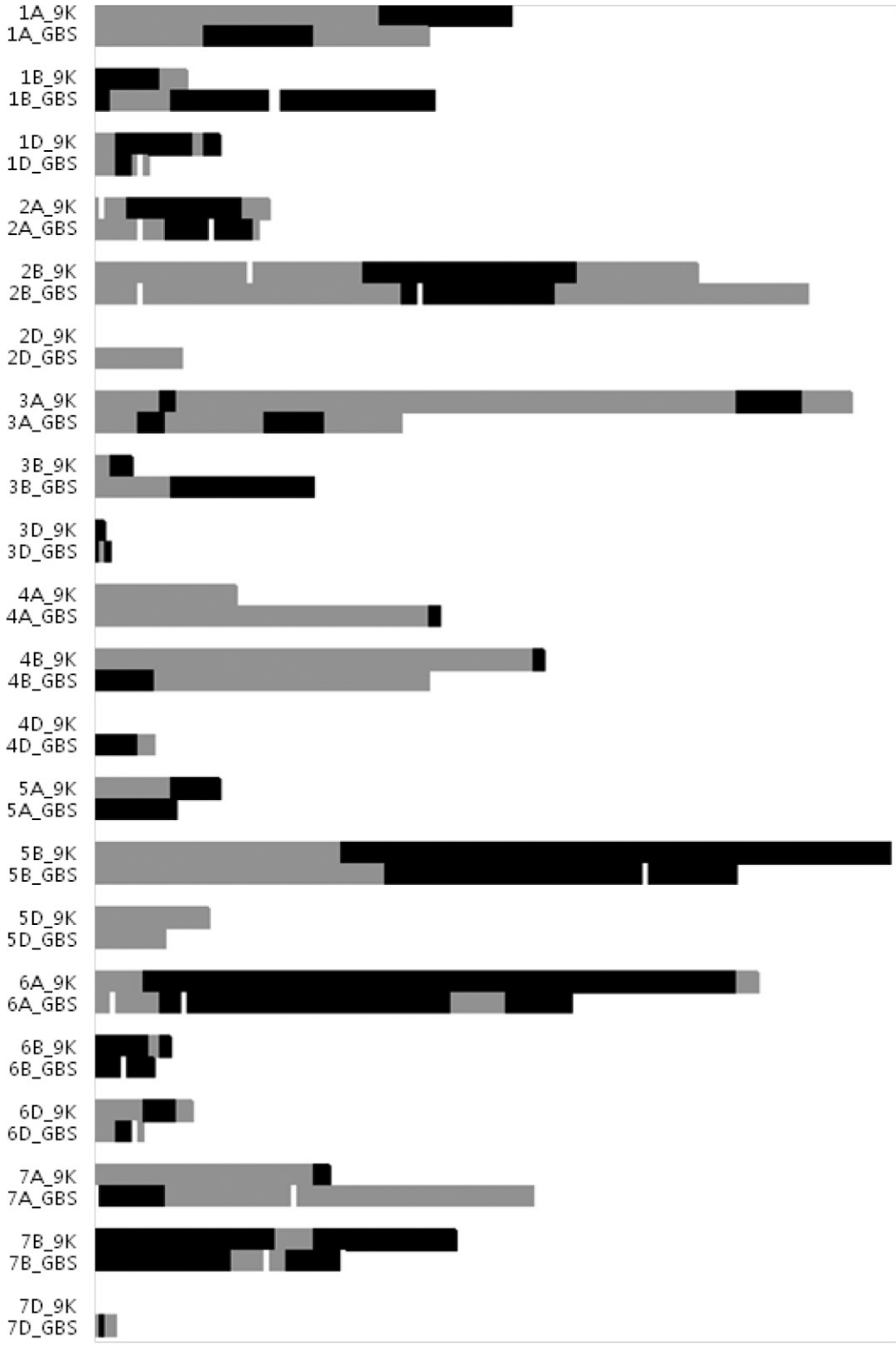

Figure 1. Continued.

Table 3. Quantitative trait loci (QTL) conferring resistance to stem rust detected in the RB07/MN06113-8 population by composite interval mapping in four environments in the 9000 Infinium iSelect single nucleotide polymorphism (SNP) assay (9K) dataset.

Environment	QTL†	Chromosome	Marker	Position	LOD‡	PVE§	Add¶
				cM		%	
Kenya 2012	2B.2	2B	wsnp_Ex_rep_c68623_67474935	24.7	15.7	32.6	6.2
	2B.1	2B	wsnp_Ku_c48694_54811376	122.5	3.1	6.0	–2.6
	4A	4A	wsnp_Ex_c30695_39579408	9.2	3.4	6.6	2.8
	6D	6D	wsnp_Ex_c37749_45436366	2.4	2.9	5.1	–2.4
Kenya 2013	2B.2	2B	wsnp_Ex_c19371_28311667	14.2	4.8	23.7	5.0
	4B.1	4B	wsnp_Ex_c30695_39579408	25.8	2.6	11.0	3.3
	5B	5B	wsnp_RFL_Contig2791_2558632	0.4	4.0	18.5	–5.3
Ethiopia 2013	2B.2	2B	wsnp_Ex_rep_c68623_67474935	24.7	17.0	57.6	11.2
	3A	3A	wsnp_Ex_c742_1458743	60.3	2.8	6.1	3.7
St. Paul, MN 2013	2B.2	2B	wsnp_Ex_c2388_4476302	21.4	3.8	9.8	2.6
	4B.1	4B	wsnp_Ex_c30695_39579408	24.8	2.6	5.7	–2.1
	4B.2	4B	wsnp_Ku_c5502_9765942	49.0	4.7	10.4	3.0
	4B.3	4B	wsnp_Ex_c48922_53681502	62.1	3.0	12.1	3.0

† For simplicity, QTL have been named by the chromosome where detected; the chromosome name is followed by a number if more than one QTL were detected on the same chromosome.

‡ LOD, logarithm of odds; values are the peak LOD score for the given QTL.

§ Value indicates the phenotypic variation explained (PVE) by the QTL.

¶ Value indicates the estimated additive effect of the QTL; negative value means that the allele was contributed by RB07.

Table 4: Quantitative trait loci (QTL) conferring resistance to stem rust detected in the RB07/MN06113-8 population by composite interval mapping in four environments, in the GBS dataset.

Environment	QTL†	Chromosome	Marker	Position	LOD‡	PVE§	Add¶
				cM		%	
Kenya 2012	2B.1	2B	TP46799	15.0	4.1	7.1	–2.9
	2B.2	2B	TP24441	93.9	15.4	32.1	6.2
	7A	7A	TP27831	0.8	4.1	6.7	–2.8
Kenya 2013	1A	1A	TP21885	92.9	3.4	14.3	3.8
	2B.2	2B	TP17690	95.7	5.2	29.6	5.6
	2D	2D	TP8148	32.2	3.3	14.7	–3.8
Ethiopia 2013	2A	2A	TP29711	30.6	2.8	7.4	3.9
	2B.2	2B	TP48796	94.5	13.8	49.6	10.4
	4B	4B	TP12718	33.6	2.7	5.7	3.6
St. Paul, MN 2013	2B.2	2B	TP23420	107.1	3.8	8.8	2.5
	4A	4A	TP49560	21.8	2.6	6.0	2.1

† For simplicity, QTL have been named by the chromosome where detected; the chromosome name is followed by a number if more than one QTL were detected on the same chromosome.

‡ LOD, logarithm of odds; values are the peak LOD score for the given QTL.

§ Value indicates the phenotypic variation explained (PVE) by the QTL.

¶ Value indicates the estimated additive effect of the QTL; negative value means that the allele was contributed by RB07.

Nonetheless, as most QTL were different between the two datasets, each QTL was studied to understand the underlying differences in QTL detection. Based on their presence or absence, the QTL have been divided into three groups, as discussed below. For better elucidation behind this discrepancy, the results were also compared with QTL detected on the combined map reported by Bajgain et al. (2015) that used both 9K and GBS markers from this study. In general, marker density and distribution of markers across the linkage groups are considered to be the causes for the discrepancy in QTL detection.

Group 1: Quantitative Trait Loci Present in the 9K Dataset but Absent in the Genotyping-by-Sequencing Dataset

The 9K marker wsnp_Ex_c742_1458743 was significantly associated with a QTL located at 60 cM on chromosome 3A in Ethiopia 2013 environment. No QTL was detected on 3A in the GBS dataset, but the LOD curve was observed slightly below the threshold (Fig. 3A). The LOD values for the 9K marker wsnp_Ex_c742_1458743 and the GBS marker TP1862, which are colocalized in the combined map, are 2.8 and 1.6, respectively. The 9K linkage map representing 3A is 156 cM long with marker density of one SNP per 1.1 cM, whereas the 146 cM long GBS 3A

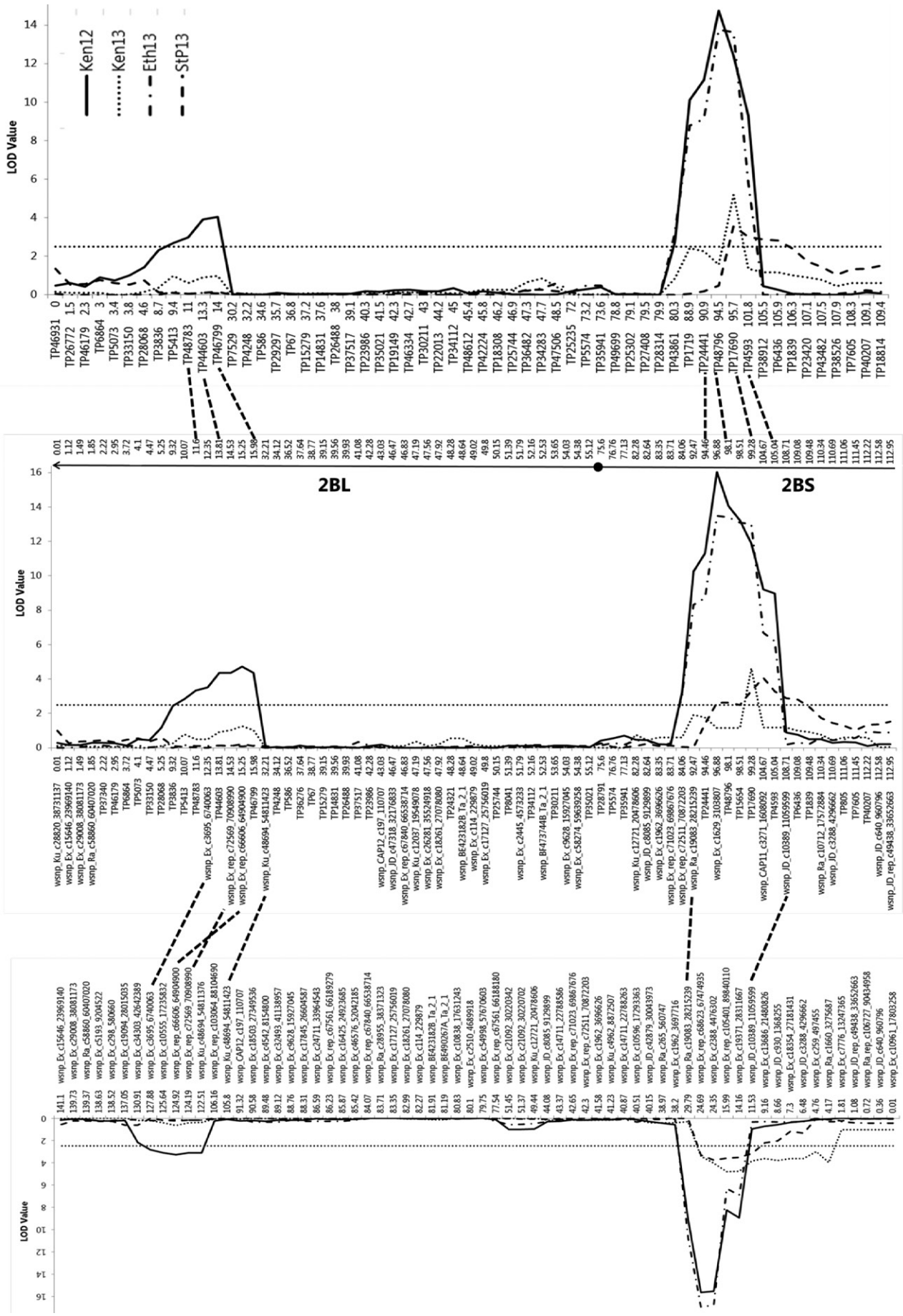

Figure 2. Common quantitative trait loci (QTL) detected between the 9000 Infinium iSelect single nucleotide polymorphism (SNP) assay (9K) and genotyping-by-sequencing (GBS) datasets. The logarithm of odds (LOD) curves on the left and right correspond to QTL detected for stem rust resistance in the RB07/MN06113-8 recombinant inbred line population in the 9K and GBS datasets, respectively. The LOD curves in the middle correspond to QTL detected in a combined (9K plus GBS) dataset in the same population (Bajgain et al., 2015). The values next to the marker names indicate marker positions in centimorgans (cM). Ken12, Kenya 2012; Ken13, Kenya 2013; Eth13, Ethiopia 2013; StP13, St. Paul 2013.

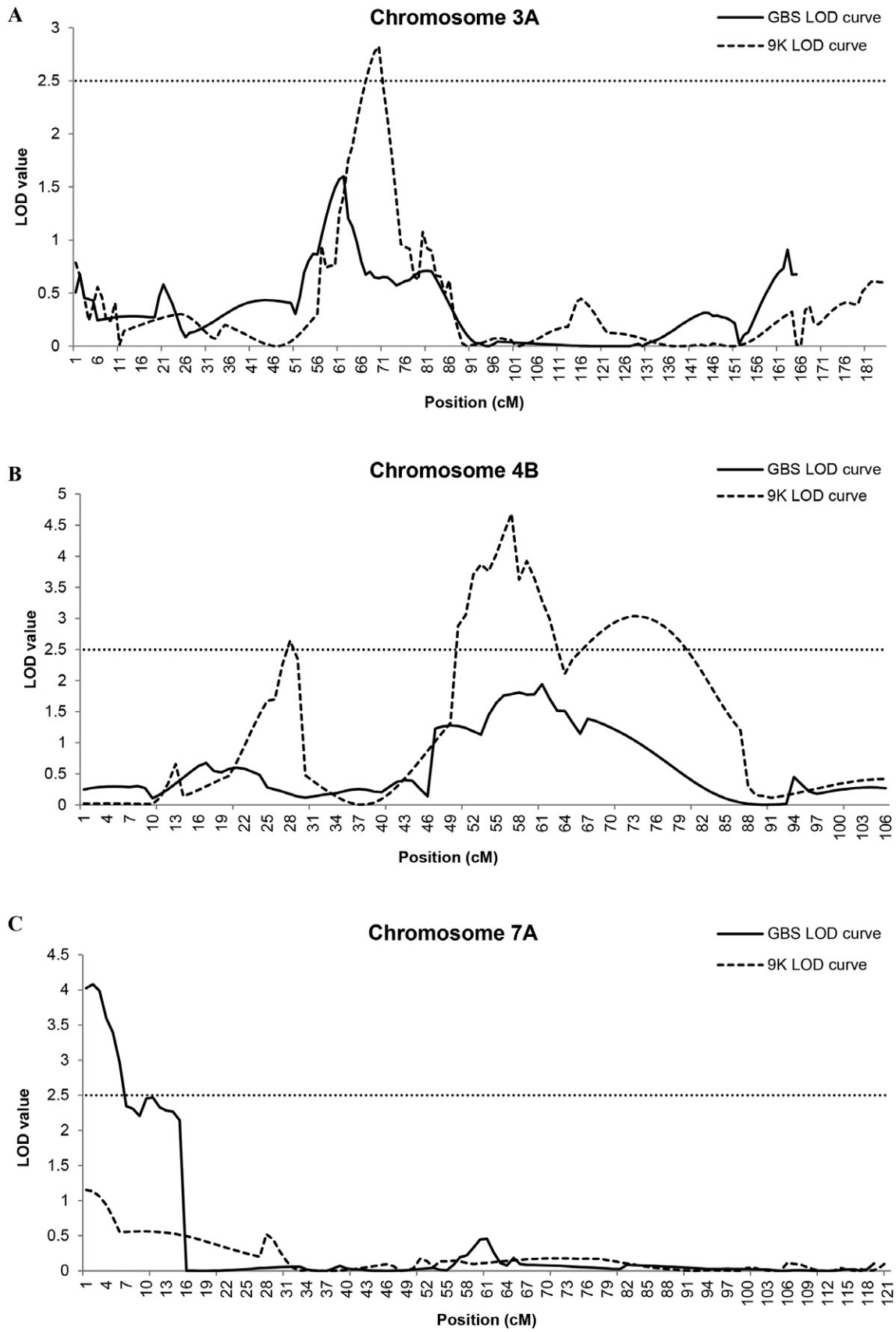

Figure 3. Logarithm of odds (LOD) curves of quantitative trait loci (QTL) detected on a given chromosome in a dataset but absent in another dataset because of lack of enough QTL detection power. (A) QTL detected on 3A in Ethiopia 2013 environment in the 9000 Infinium iSelect single nucleotide polymorphism assay (9K) dataset but absent in the genotyping-by-sequencing (GBS) dataset. (B) QTL detected on 4B in St. Paul 2013 environment in the 9K dataset but absent in the GBS dataset. (C) QTL detected on 7A in Kenya 2012 environment in the GBS dataset but absent in the 9K dataset. In all three panels, the dotted horizontal line represents the threshold LOD score of 2.5.

map has marker density of a SNP marker every 2.6 cM, on average. In addition, the proportion of genotypes with the RB07 allele (the parent contributing the QTL) was found to be 55% for the 9K marker, whereas it was 45% for the GBS marker. Therefore, disproportionate representation of the parental alleles as well as the lack of a denser linkage map could have led to a lower power in QTL detection in the GBS map relative to that in the 9K map.

Similarly, an LOD curve in the GBS dataset similar to that of the 9K QTL detected at 49 cM (marker wsnp_Ku_c5502_9765942) with LOD value of 4.7 on 4B in St. Paul 2013 environment was found below the threshold (Fig. 3B). This 9K marker colocalizes with several GBS markers at position 44 cM in the combined map, and both marker systems represent the parental alleles almost equally: RB07 by 55 and 54% markers from 9K and GBS dataset, respectively, and MN06113-8 by 44 and 42% markers from 9K and GBS dataset, respectively. The marker density in the 9K map (94 cM long) is one SNP per 1.1 cM, whereas the marker density in the GBS map (90 cM long) is a marker every 1.5 cM. While the difference in marker density between the two maps is not very large, perhaps a better marker distribution similar to that of the 9K map could have provided enough power for the LOD curve to rise above the threshold in the GBS dataset.

On chromosome 6D (linkage group 6D_2), the 9K marker wsnp_Ex_c37749_45436366 was found to be associated with a QTL at 2 cM in Kenya 2012 environment with an LOD value of 2.9. The QTL was absent in the GBS dataset, yet was detected in the combined dataset albeit no GBS marker colocalized with this 9K marker. The marker density is quite similar between the two datasets: one SNP marker per 1 and 1.3 cM in the 9K and GBS datasets, respectively. We suspect that the difference in QTL detection occurred from lack of enough markers on the 6D GBS linkage group, which has only nine markers compared with 18 markers in the 9K dataset.

Group 2: Quantitative Trait Loci Present in the Genotyping-by-Sequencing Dataset but Absent in the 9K Dataset

On chromosome 1A, a QTL linked to the GBS marker TP21885 was detected, but no QTL was observed in the 9K dataset. In the combined map, TP21885 is colocalized with 9K markers wsnp_Ex_c38203_45790396 and wsnp_Ex_c1137_2182795. Both these 9K markers are located at 0 cM on the linkage group 1A_2 in the 9K dataset. It is known that the CIM algorithm uses a prespecified number of markers as cofactors to account for background marker noise while detecting a QTL. Hence, the algorithm cannot always ensure that the QTL at the current testing interval is not absorbed by the background marker variables and may result in biased estimation (Zeng, 1994; Li et al., 2007). To test this hypothesis, we used wsnp_Ex_c38203_45790396 as a background marker and reran the CIM algorithm. As expected, a QTL was detected on 1A_2 at 0 cM (location of wsnp_Ex_c1137_2182795) with LOD score of 3.2, PVE of 10.04%, and additive effect of 3.9. These values are similar to those of the GBS 1A QTL (Table 3).

A QTL associated with the GBS marker TP49560 was detected on 4A (linkage group 4A_1) with LOD value of 2.6 in St. Paul 2013 environment. The QTL was also detected in the combined dataset but was absent in the 9K dataset. The LOD curve in the 9K dataset did not resemble that of the GBS QTL curve, and no 9K markers colocalize with TP49560 in the combined map. Lack of enough markers is likely the primary reason behind why the QTL went undetected, as only 26 markers are present on the 9K 4A map with marker density of a marker every 2.4 cM. Compared with the 9K dataset, the GBS dataset has 63 markers on 4A, with a marker at an average distance of every 0.9 cM.

Though the QTL detected at 0.8 cM (GBS marker TP27831) on 7A in Kenya 2012 environment was not detected in the 9K dataset, a comparison of the LOD curves suggests that lack of enough power failed to detect the QTL (Fig. 3C). The highest LOD value observed in the 9K dataset was 1.5 at the marker wsnp_Ex_c35_77935, also located at the proximal end of the linkage group. The QTL was contributed by RB07, yet the difference in the representation of RB07 alleles was not very large: 41% in 9K and 38% in GBS. The marker density in the 105-cM-long 9K map is 2.4 cM on average, whereas that in the 98 cM long GBS map is one SNP per 1.2 cM. A denser map, especially with more markers on the proximal end of the linkage map could provide enough power to detect the QTL in the 9K map.

Group 3: False Quantitative Trait Loci in both datasets

The QTL detected at the 9K marker wsnp_Ex_c57209_59016692 with LOD value of 4.0 on 5B was detected neither in GBS dataset nor in the combined dataset. Bootstrapping of the trait data with 10,000 data points around the detected QTL to estimate the confidence interval showed that the highest LOD score peaked out at 2.1. Similarly, the QTL detected at the GBS marker TP19076 with LOD value of 2.8 on 2A (linkage group 2A_2) was undetected in both 9K and combined datasets. We performed 10,000 bootstraps to get an estimated confidence interval on the QTL location but were unable to validate the presence of this QTL as the highest LOD value observed after bootstrapping was only 0.9. As we were unable to validate both these QTL in their respective datasets, we suspect that these QTL likely are false QTL.

Quantitative Trait Loci Mapping with Altered and Simulated Data

We also investigated if the large-effect QTL 2B (i) had a masking effect on other small-effect QTL and (ii) played a role in difference behind QTL detection between the two datasets. Composite interval mapping performed with SNP markers associated with the 2B locus as cofactors in both datasets led to identification of two additional QTL in the 9K dataset: one each in Kenya 2012 and St. Paul 2013 environments (Table 1; Supplemental File S1). The QTL in Kenya 2012 environment was detected on 7B at position 90 cM and was not observed during regular mapping before. The QTL on 4B (marker wsnp_Ku_c8128_13866660 at position 50 cM) is located 1.3 cM away from the QTL detected at wsnp_Ku_c5502_9765942 during regular mapping and is likely the same QTL as detected before.

Similarly, in the GBS dataset, three additional QTL were detected: one in Kenya 2012, two in St. Paul 2013, and none in remaining environments (Table 1; Supplemental File S1). The QTL in Kenya 2012 was detected at 0 cM on 7A at a distance of 0.75 cM from the QTL detected during regular mapping. The QTL detected on linkage group 4A_1 in St. Paul 2013 is located 4.6 cM away from the 4A QTL detected during regular mapping. The second QTL detected in St. Paul 2013 environment was on 4B and was the same marker (TP48810) as detected during regular mapping; however, the peak shifted by 0.6 cM. As the PVE and additive values are quite similar to that from regular mapping, these QTL are likely identical. Also, based on the combined map, TP48810 is located only 1.5 cM away from the 9K marker wsnp_Ku_c8128_13866660 (discussed above) and, therefore, might be the same QTL.

Except for the detection of a new QTL on 7B in the 9K dataset, no differences were observed between regular mapping and mapping with 2B SNP markers used as cofactors. The masking effect of 2B was therefore not as influential on QTL detection. This also suggests that both datasets can detect QTL of varying effects in the presence of a large QTL. Despite this observation, we simulated trait data with known QTL positions and effects to understand how each dataset would detect QTL of different effects. For this purpose, six QTL located on six different chromosomes were simulated on a phenotypic dataset of six traits (Table 1; Supplemental File S2). Quantitative trait loci were simulated on the same chromosomes for both genotypic datasets, if possible. Each simulated QTL was assigned a different additive effect: 0.25, 0.5, 1, 2, 5, and 10.

In the 9K dataset, all but the smallest-effect QTL (additive value of 0.25) were detected (Supplemental Table S2A within Supplemental File S2). For each SNP marker associated with the QTL, the estimated additive effects were very close to the values assigned during QTL simulation. Two likely false QTL were detected on chromosomes 3A and 4A where no QTL were simulated. A similar trend was observed in the GBS dataset where all simulated QTL were detected except for the QTL with additive effect of 0.25 (Supplemental Table S2B within Supplemental File S2). As a nonsimulated QTL was detected on 7B, this QTL is likely false.

In general, both genotype matrices were able to detect QTL with large and small effects in simulated datasets. Taking together the mapping results using real and simulated data, it can be concluded that our genotype matrices and the linkage groups constituted thereof provide sufficient power to detect small effect QTL without interference from large QTL.

Imputation of Genotyping-by-Sequencing Markers

To our knowledge, an imputation study using GBS markers has not yet been conducted in an RIL population; therefore, we looked at the effect and accuracy of genotype imputation in the GBS dataset after simulating datasets with several successive missing proportions.

The highest imputation accuracy was observed when the proportion of missing data was the lowest (Fig. 4). The missing genotypes were predicted with accuracy of 96% when the dataset was missing 20% of the genotypes. The accuracy of genotype imputation was reduced as datasets had higher proportion of missing data, with a major drop-off in datasets with more than 60% missing data, which had imputation accuracies of less than 84%. The effect of missing data on PIC, and the amount of heterozygosity was not significant. An overall decreasing trend for each of these population characteristics can, however, be observed with increasing missing values in the dataset. The inbreeding coefficient had an overall increasing trend as datasets had more missing proportion of genotype calls. There was no significant change in the allele type from imputation in the imputed datasets, except in the GBS90 dataset where the proportion of RB07 alleles increased by 3%.

Our GBS dataset contained 932 high-quality SNP markers, which is sufficient for gene mapping studies given the high linkage disequilibrium of hexaploid bread wheat (Chao et al., 2007, 2009). However, the marker imputation results presented here show that more markers are necessary to perform high-resolution mapping studies. With as much as 40% missing data, the genotype matrix can be predicted with imputation accuracy of 90% or higher. While this may potentially introduce some biases in the study, the level of PIC and heterozygosity were not significantly altered, implying that the population does not deviate significantly from the expected levels of inbreeding. This is illustrated in Fig. 4, where the inbreeding coefficient increases negligibly from nonimputed GBS dataset to GBS90 dataset. The observed slight increase most likely is due to introduction of false genotype calls in imputed datasets with higher proportion of missing data.

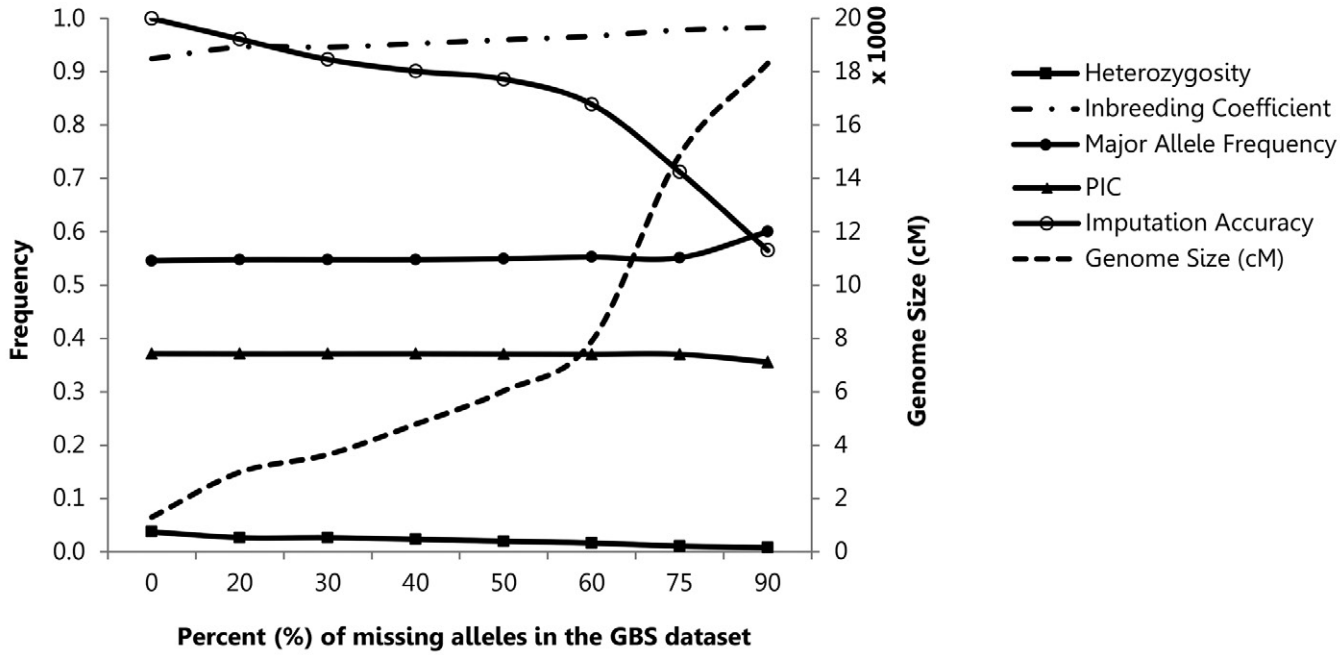

Figure 4. Characteristics of genotype matrices in imputed and nonimputed datasets. Zero represents the original genotyping-by-sequencing (GBS) dataset with no missing allele calls introduced. The genome sizes (sizes of linkage groups summed together) of imputed and nonimputed GBS datasets are shown on the secondary *y*-axis to the right.

Higher level of inbreeding is desired in most crops so that desired traits can be preserved in the breeding program. Yet, the consistency in inbreeding levels observed among the imputed datasets should not be the only determinant behind the use of a dataset with large amount of missing allele calls. Datasets with large amount of missing data introduce severe problems such as inaccurate and inflated linkage groups and erroneous QTL detection.

Quantitative Trait Loci Mapping Using Imputed Genotyping-by-Sequencing Datasets

The effect of marker imputation in construction of linkage groups and QTL mapping in a biparental population has not been investigated yet. Thus, we used imputed GBS datasets, as described in the Materials and Methods section, to investigate QTL mapping. Construction of linkage groups and QTL mapping, however, were performed using the imputed datasets with 40 and 75% missing data only, which had imputation accuracies of 90 and 71%, respectively.

Twenty-nine linkage groups were formed in the GBS40 dataset and covered 4785 cM of the genome (Table 2). Similarly, 30 linkage groups were formed in the GBS75 dataset, covering 14,879 cM of the genome. Genome sizes represented by these linkage groups are approximately 3.5 and 11 times larger than the size covered by the original, nonimputed GBS dataset. This increase in genome size most likely is due to the introduction of inaccurate genotype calls with a large proportion of missing values. Inflation of linkage groups from introduction of false genotype calls is also supported by the average marker interval distance of 5 cM in the GBS40 dataset and 19 cM in the GBS75 dataset. Therefore, marker imputation in datasets with large proportion of missing data appears to introduce errors, and as such, avoiding the use of such datasets is pragmatic. As the imputation accuracy dropped, more markers were also unlinked to any linkage group (Table 2). Two linkage groups (chromosomes 3D and 7D) were not detected in the GBS75 dataset.

The number of QTL discovered in both GBS40 and GBS75 datasets using the CIM approach was eight and seven, respectively (Table 2). However, most of the QTL detected in these datasets were different relative to the nonimputed dataset, with only one consistent QTL between all datasets (Supplemental Table S1C, S1D within Supplemental File S3). The large-effect QTL observed on chromosome 2B in the nonimputed GBS datasets was the only consistent QTL in GBS40 and GBS75 datasets, with lower accuracy of the QTL positions in the imputed datasets (Fig. 5A–C). Only the GBS40 dataset correctly predicted the same SNPs (TP24441 and TP17690) linked to the large-effect QTL as predicted in the nonimputed GBS dataset. Both imputed datasets predicted the percentage of phenotypic variation and allelic effect similarly to the nonimputed dataset. The small-effect QTL observed in the Kenya 2012 environment was detected in both imputed datasets, although their positions and the markers they are linked with were different than that observed in the nonimputed dataset. It is likely, given the inflation in size of linkage groups experienced with imputing, that most of the QTL detected in the imputed datasets are inaccurate. While validation of the detected QTL would provide a definitive answer, the

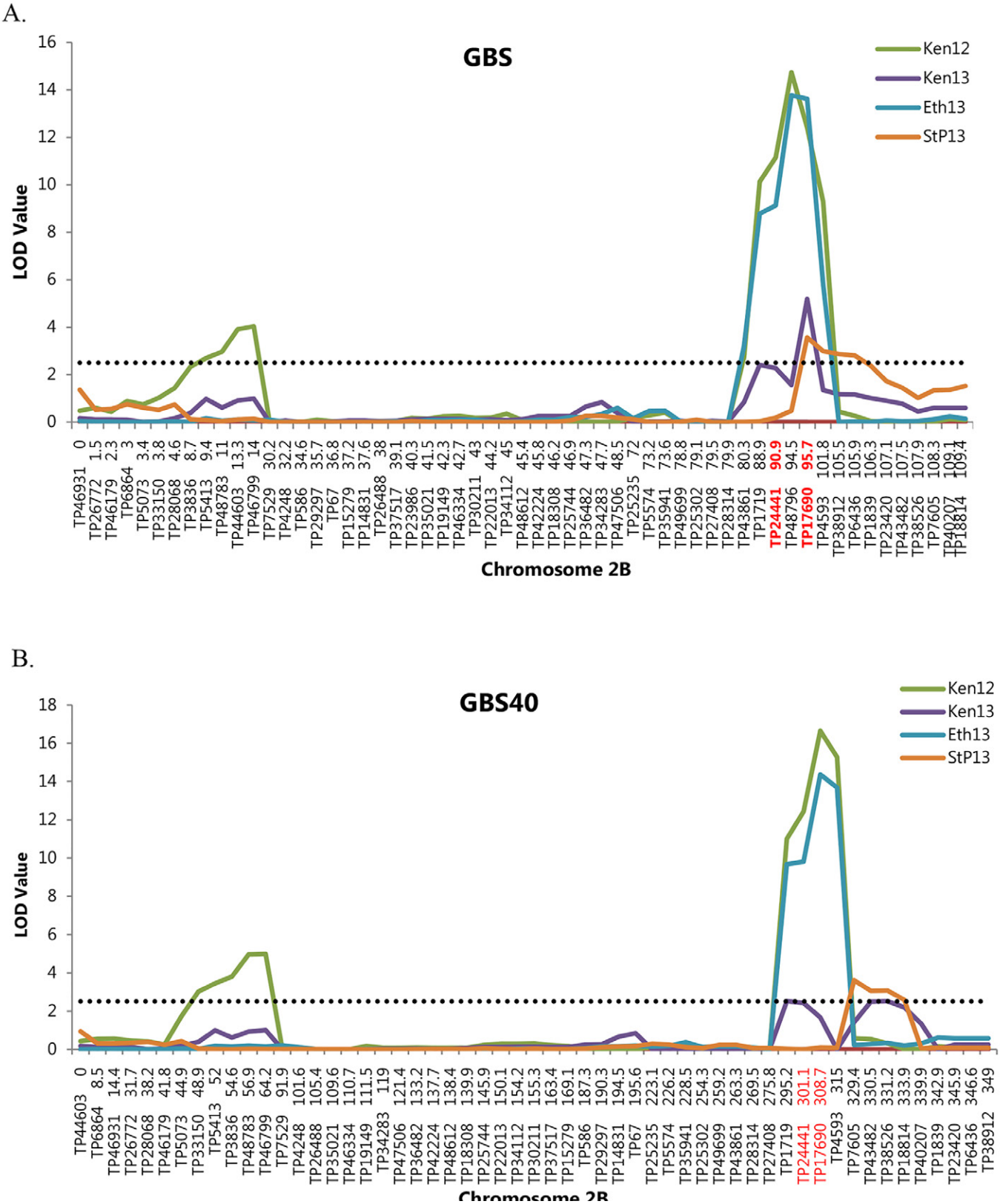

Figure 5. (continued on next page) A comparison of the logarithm of odds (LOD) peaks for the quantitative trait loci (QTL) on 2B among the three genotyping-by-sequencing (GBS) datasets. (A) The nonimputed dataset is labeled as GBS; (B) the imputed dataset with 40% missing allele calls as GBS40; and (C) the imputed dataset with 75% missing allele calls as GBS75. The significant single nucleotide polymorphism markers (TP24441 and TP17690) in nonimputed GBS dataset are labeled in red color in all three panels to indicate their positions relative to the QTL peaks. Note the increase in sizes of the GBS40 and GBS75 linkage groups compared with that of nonimputed GBS dataset. In comparison with the nonimputed GBS dataset, the order of the SNP markers also changes in linkage groups constructed using genotype information in imputed datasets with higher proportion of missing alleles. Ken12, Kenya 2012; Ken13, Kenya 2013; Eth13, Ethiopia 2013; StP13, St. Paul 2013.

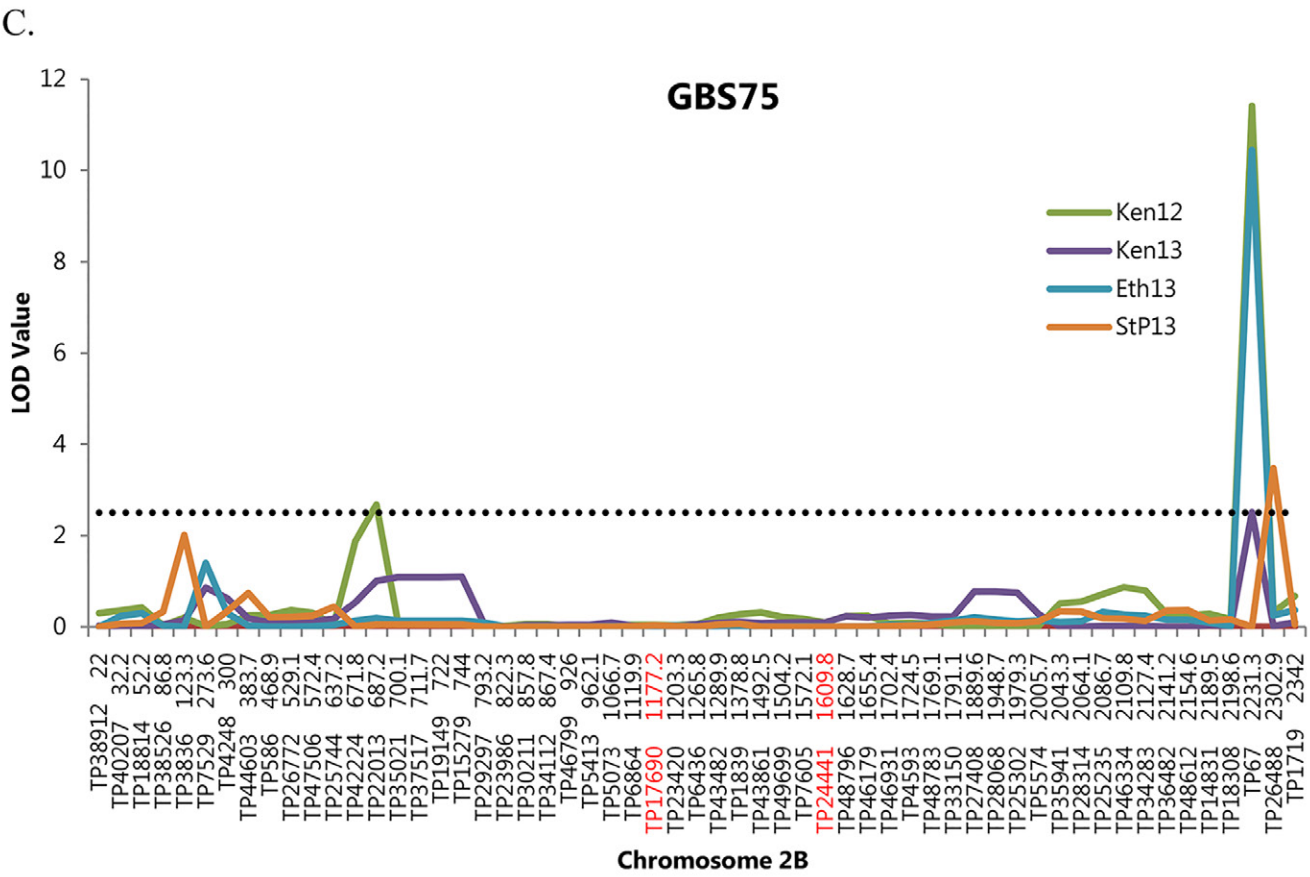

Figure 5. Continued.

results indicate that large-effect QTL can be detected if the dataset comprises large proportion of imputed genotypes. The small-effect QTL may also be detected, but such prediction might not necessarily be accurate.

Comparison of Methodology and Workflow

With regard to the methodologies and workflow, chip-based genotyping is relatively easier than the GBS approach, though the latter had a faster turnaround time in our case. The 9K genotype calls were obtained from the USDA Genotyping Facility at Fargo, ND, that needed manual inspection using Illumina's GenomeStudio program version 2011.1 (Illumina Inc.) before use. In the GBS method, DNA libraries for sequencing were generated in-house in <2 d. The sequences were filtered based on barcodes and trimmed before calling SNPs in the population. In addition, sequence alignment to the wheat CSS and further data parsing was required. Thus, the need for high-end computational resources and bioinformatic expertise (ability to work in the UNIX environment as well as programming skills) is essential in the GBS approach to manage and work with the large amount of sequence data generated from parallel sequencing. On average, 256 gigabytes (gb) of memory was requested on any available node with the Minnesota Supercomputing Institute (https://www.msi.umn.edu/) while working with the GBS procedures such as quality control of the sequences, SNP calling, and sequence alignment. Similarly, several hundred gigabytes of hard disk space was needed to store the sequence files and any output files created during the procedures mentioned above. The SNP-chip-based method, however, required less computational resources, but a proprietary program (GenomeStudio) was needed to visualize and analyze the data generated from the genotyping assay. The program was run on a Windows platform with 64 gb of available memory. The hard disk space requirement to store the data files was <50 megabytes.

One particular advantage of the GBS approach is perhaps the economical aspect of this method. In our study, we obtained a comparable number of usable SNP markers for QTL mapping from both genotyping approaches (964 from 9K and 925 from GBS). The cost per 9K SNP marker used for mapping was approximately $8.20, whereas the cost per GBS SNP marker used in QTL mapping was approximately $2.10. These figures are exclusive of the labor cost, in which the GBS method is also advantageous over the chip-based genotyping method. Although several filters can be applied within the program GenomeStudio to parse the genotype calls obtained using the 9K chip, the genotypes still need to be manually inspected for each SNP between the two parents to recluster the individuals to distinct genotype groups.

In our dataset, 2524 polymorphic SNPs were obtained after applying the filter to remove monomorphic SNPs between the two parents. Each of these SNPs had to be inspected to assign the correct genotype calls to the RILs. At the inspection rate of approximately three SNPs every 2 min, the total time required to tag the population with correct genotype calls was approximately 28 h. As the program did not allow for correction of incorrect genotype calls, the exported data had to be edited to assign final genotypes to each individual. On the other hand, the GBS procedure required less than 3 h to obtain the final genotype calls. Creating the input file with each individual labeled with the barcode used during library preparation was essential before running the SNP-calling program UNEAK. This task was completed in about 30 min followed by approximately 1 h to obtain the SNPs from UNEAK. As the allele calls were reported in base format (AA/CC/GG/TT), they were converted to biallelic format (AA/BB), which was accomplished in about 30 min. While the time needed to troubleshoot errors that appeared during both procedures has not been discussed here, the GBS approach was more efficient because of its lower economic burden and advantages in automated data processing. Yet, as we indicated in the results, the genome areas targeted by these approaches are slightly different, which may lead to detection of different QTL between the two methods. If marker coverage was better—and perhaps uniform—between the two methods, these differences should disappear. We would also like to state that the comparative analysis presented here is a product of our own experiences and may be different for other research groups.

CONCLUSION

Highly economical genotyping approaches provide a user with the option of using different types of high-throughput genotyping methods, ranging from simple-sequence-repeat-based genotyping to sequence-based genotyping. In our study, we compared two high-throughput genotyping methods used in genomic studies of wheat. The results showed that both methods are powerful means of studying the genome and provide enough resolution to carry out marker–trait association studies. The key attributes of high interest to a researcher might be the cost and data turnaround time, in which the GBS approach bests the SNP-genotyping method. The GBS approach was also able to provide a broader coverage of the wheat genome including that of the often poorly represented D genome. The SNP-chip-based genotyping, however, requires less computational knowledge and resources to process the data. The choice of the genotyping platform for gene mapping and other genome studies may come down to the question of cost and available resources.

Acknowledgments

We thank the University of Minnesota Genomics Center, Minnesota Supercomputing Institute, the Microbial and Plant Genomics Institute, the University of Minnesota Graduate School, Kenya Agricultural and Livestock Research Organization, Ethiopia Institute of Agricultural Research, Cereal Disease Laboratory personnel, and Anderson Wheat Lab for their help and support during various phases of the project. Funding for this work was provided by the USDA, Agriculture and Food Research Initiative (Triticeae Coordinated Agricultural Project, Grant no. 2011-68002-30029), the Durable Rust Resistance in Wheat project administrated by Cornell University and funded by the Bill and Melinda Gates Foundation, and the United Kingdom Department for International Development.

References

Albrechtsen, A., F.C. Nielsen, and R. Nielsen. 2010. Ascertainment biases in SNP chips affect measures of population divergence. Mol. Biol. Evol. 27:2534–2547. doi:10.1093/molbev/msq148

Allen, A.M., G.L.A. Barker, S.T. Berry, J.A. Coghill, R. Gwilliam, S. Kirby, et al. 2011. Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). Plant Biotechnol. J. 9:1086–1099. doi:10.1111/j.1467-7652.2011.00628.x

Bajgain, P., M. Rouse, S. Bhavani, and J. Anderson. 2015. QTL mapping of adult plant resistance to Ug99 stem rust in the spring wheat population RB07/MN06113-8. Mol. Breed. 35:1–15. doi:10.1007/s11032-015-0362-x

Botstein, D., R.L. White, M. Skolnick, and R.W. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32:314–331.

Cavanagh, C.R., S. Chao, S. Wang, B.E. Huang, S. Stephen, S. Kiani, et al. 2013. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc. Natl. Acad. Sci. U. S. A. 110:8057–8062. doi:10.1073/pnas.1217133110

Chao, S., W. Zhang, E. Akhunov, J. Sherman, Y. Ma, M.C. Luo, et al. 2009. Analysis of gene-derived SNP marker polymorphism in US wheat (*Triticum aestivum* L.) cultivars. Mol. Breed. 23:23–33. doi:10.1007/s11032-008-9210-6

Chao, S., W. Zhang, J. Dubcovsky, and M. Sorrells. 2007. Evaluation of genetic diversity and genome-wide linkage disequilibrium among U.S. wheat (*Triticum aestivum* L.) germplasm representing different market classes. Crop Sci. 47:1018–1030. doi:10.2135/cropsci2006.06.0434

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, et al. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:E19379. doi:10.1371/journal.pone.0019379

Ferdosi, M., B. Kinghorn, J. van der Werf, S. Lee, and C. Gondro. 2014. hsphase: An R package for pedigree reconstruction, detection of recombination events, phasing and imputation of half-sib family groups. BMC Bioinf. 15:172. doi:10.1186/1471-2105-15-172

Frascaroli, E., T.A. Schrag, and A.E. Melchinger. 2013. Genetic diversity analysis of elite European maize (*Zea mays* L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs. Theor. Appl. Genet. 126:133–141. doi:10.1007/s00122-012-1968-6

Fu, Y.B. 2014. Genetic diversity analysis of highly incomplete SNP genotype data with imputations: An empirical assessment. G3: Genes Genomes Genet. 4:891–900. doi:10.1534/g3.114.010942

Fu, Y.B., B. Cheng, and G. Peterson. 2014. Genetic diversity analysis of yellow mustard (*Sinapis alba* L.) germplasm based on genotyping by sequencing. Genet. Resour. Crop Evol. 61:579–594. doi:10.1007/s10722-013-0058-1

Gupta, P.K., J. Kumar, R.R. Mir, and A. Kumar. 2010. Marker-assisted selection as a component of conventional plant breeding. Plant Breeding Reviews. John Wiley & Sons, Hoboken, NJ. p. 145–217.

Jia, J., S. Zhao, X. Kong, Y. Li, G. Zhao, W. He, et al. 2013. *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 496:91–95. doi:10.1038/nature12028

Kidwell, K., and T. Osborn. 1992. Simple plant DNA isolation procedures. In: J.S. Beckmann and T.C. Osborn, editors, Plant genomes: Methods for genetic and physical mapping. Springer, Netherlands, p. 1–13.

Kosambi, D.D. 1943. The estimation of map distance from recombination values. Ann. Eugen. 12:172–175. doi:10.1111/j.1469-1809.1943.tb02321.x

Li, B., and M. Kimmel. 2013. Factors influencing ascertainment bias of microsatellite allele sizes: Impact on estimates of mutation rates. Genetics 195:563–572. doi:10.1534/genetics.113.154161

Li, H., G. Ye, and J. Wang. 2007. A modified algorithm for the improvement of composite interval mapping. Genetics 175:361–374. doi:10.1534/genetics.106.066811

Ling, H.Q., S. Zhao, D. Liu, J. Wang, H. Sun, C. Zhang, et al. 2013. Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature 496:87–90. doi:10.1038/nature11997

Liu, H., M. Bayer, A. Druka, J. Russell, C. Hackett, J. Poland, et al. 2014. An evaluation of genotyping by sequencing (GBS) to map the *Breviaristatum-e* (*ari-e*) locus in cultivated barley. BMC Genomics 15:104. doi:10.1186/1471-2164-15-104

Liu, K., and S.V. Muse. 2005. PowerMarker: An integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129. doi:10.1093/bioinformatics/bti282

Lorieux, M. 2012. MapDisto: Fast and efficient computation of genetic linkage maps. Mol. Breed. 30:1231–1235. doi:10.1007/s11032-012-9706-y

Lu, F., A.E. Lipka, J. Glaubitz, R. Elshire, J.H. Cherney, M.D. Casler, et al. 2013. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. PLoS Genet. 9:E1003215. doi:10.1371/journal.pgen.1003215

Moragues, M., J. Comadran, R. Waugh, I. Milne, A.J. Flavell, and J.R. Russell. 2010. Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. Theor. Appl. Genet. 120:1525–1534. doi:10.1007/s00122-010-1273-1

Moser, G., M.S. Khatkar, and H.W. Raadsma. 2009. Imputation of missing genotypes in high density SNP data. Proc. Adv. Anim. Breed. Gen. 18:612–615.

Poland, J.A., P.J. Brown, M.E. Sorrells, and J.-L. Jannink. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7:E32253. doi:10.1371/journal.pone.0032253

R Development Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Stat. Comput., Vienna, Austria.

Ritland, K. 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. Genet. Res. 67:175–185. doi:10.1017/S0016672300033620

Rutkoski, J.E., J. Poland, J.L. Jannink, and M.E. Sorrells. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3: Genes Genomes Genet. 3:427–439.

Saintenac, C., D. Jiang, S. Wang, and E. Akhunov. 2013. Sequence-based mapping of the polyploid wheat genome. G3: Genes Genomes Genet. 3:1105–1114. doi:10.1534/g3.113.005819

Stacklies, W., H. Redestig, M. Scholz, D. Walther, and J. Selbig. 2007. pcaMethods: A bioconductor package providing PCA methods for incomplete data. Bioinformatics 23:1164–1167. doi:10.1093/bioinformatics/btm069

Wang, S., C.J. Basten, and Z.B. Zeng. 2012. Windows QTL cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC.

Williams, L., X. Ma, A. Boyko, C. Bustamante, and M. Oleksiak. 2010. SNP identification, verification, and utility for population genetics in a non-model genus. BMC Genet. 11:32. doi:10.1186/1471-2156-11-32

Zeng, Z.B. 1994. Precision mapping of quantitative trait loci. Genetics 136:1457–1468.