Theor Appl Genet (2013) 126:2699–2716
DOI 10.1007/s00122-013-2166-x

ORIGINAL PAPER

Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations

Jennifer Spindel · Mark Wright · Charles Chen · Joshua Cobb · Joseph Gage · Sandra Harrington · Mathias Lorieux · Nourollah Ahmadi · Susan McCouch

Received: 20 November 2012 / Accepted: 12 July 2013 / Published online: 6 August 2013

Abstract Genotyping by sequencing (GBS) is the latest application of next-generation sequencing protocols for the purposes of discovering and genotyping SNPs in a variety of crop species and populations. Unlike other high-density genotyping technologies which have mainly been applied to general interest "reference" genomes, the low cost of GBS makes it an attractive means of saturating mapping and breeding populations with a high density of SNP markers. One barrier to the widespread use of GBS has been the difficulty of the bioinformatics analysis as the approach is accompanied by a high number of erroneous SNP calls which are not easily diagnosed or corrected. In this study, we use a 384-plex GBS protocol to add 30,984 markers to an *indica* (IR64) × *japonica* (Azucena) mapping population consisting of 176 recombinant inbred lines of rice (*Oryza sativa*) and we release our imputation and error correction pipeline to address initial GBS data sparsity and error, and streamline the process of adding SNPs to RIL populations. Using the final imputed and corrected dataset of 30,984 markers, we were able to map recombination hot and cold spots and regions of segregation distortion across the genome with a high degree of accuracy, thus identifying regions of the genome containing putative sterility loci. We mapped QTL for leaf width and aluminum tolerance, and were able to identify additional QTL for both phenotypes when using the full set of 30,984 SNPs that were not identified using a subset of only 1,464 SNPs, including a previously unreported QTL for aluminum tolerance located directly within a recombination hotspot on chromosome 1. These results suggest that adding a high density of SNP markers to a mapping or breeding population through GBS has a great value for numerous applications in rice breeding and genetics research.

Communicated by R. Snowdon.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-013-2166-x) contains supplementary material, which is available to authorized users.

J. Spindel · M. Wright · C. Chen · J. Cobb · J. Gage ·
S. Harrington · S. McCouch (✉)
Department of Plant Breeding and Genetics, Cornell University,
162 Emerson Hall, Ithaca, NY 14853-1901, USA
e-mail: srm4@cornell.edu

J. Spindel
e-mail: jes462@cornell.edu

M. Wright
e-mail: mhw6@cornell.edu

M. Lorieux
UMR DIADE, Institut de Recherche pour le Développement
(IRD), 34394 Montpellier Cedex 5, France

M. Lorieux
Rice Genetics and Genomics Laboratory, International Center
for Tropical Agriculture (CIAT), AA6713 Cali, Colombia

N. Ahmadi
Centre de Coopération Internationale en Recherche Agronomique
pour le Développement (CIRAD), TA06/01 Avenue Agropolis,
34398 Montpellier Cedex 05, France

Introduction

Plant breeding and genetics research is transitioning from a data-poor to a data-rich environment. Next-generation sequencing of crop plant genomes, including that of rice (*Oryza sativa*), is revolutionizing the field as newly abundant data enable and facilitate the discovery and use of millions of single nucleotide polymorphisms (SNPs) in diverse genomes (Huang et al. 2012; Xu et al. 2012). Yet, at the same time, traditional bi-parental mapping populations continue to play an important role in gene discovery, and both bi-parental and multi-parental breeding populations remain the foundation of many plant breeding programs (Almeida et al. 2013; Famoso et al. 2011; Rosyara et al. 2009). While new "reference genomes" are being sequenced every day, many plant breeders and geneticists using traditional mapping and breeding populations continue to work with sparse molecular marker data, or in cases of extremely resource-limited programs (such as those often found in developing countries) no marker data at all, despite the abundance of public data on select lines (Rosyara et al. 2009). A recent development in genotyping technology is genotyping by sequencing (GBS), i.e., the adaptation of next-gen sequencing protocols to simultaneously discover and score segregating markers in populations of interest. GBS holds the potential to close the genotyping gap between references of broad interest and mapping/breeding populations of local or specific interest. The multiplexing of samples in GBS protocols keeps molecular biology costs low while the resultant next-generation sequencing data has immediate applications to many different research areas, ranging from gene discovery to genomic-assisted breeding (Thomson et al. 2012).

Many GBS-like protocols have been used in recent years, providing a range of methodological options for adding large numbers of markers to new or existing mapping or breeding populations. All methods seek solutions to the same essential problem—how to efficiently sort through millions of short read sequences to identify molecular polymorphisms that segregate among individuals, varieties, or populations, while at the same time, identifying and discarding sequencing and alignment errors, repetitive, and non-informative segments of the genome, and multiplexing DNA samples to optimize throughput and minimize cost (Baird et al. 2008; Davey et al. 2011; Elshire et al. 2011a, b; Huang et al. 2009). One current and popular strategy to achieve these goals is to develop a bar-coded library for each sample by digesting genomic DNA with a restriction enzyme and attaching molecular bar codes and primer annealing sites to the ends of each fragment prior to sequencing. Sequencing is then performed using a next-gen platform (i.e., Illumina HiSeq 2000) that generates short reads (<100 bp long), such that the sequenced library is enriched for regions of the genome located within 100 bp of the selected restriction sites.

Methylation-sensitive restriction enzymes are often employed to help reduce the complexity of the genome and specifically to avoid sequencing through repetitive (methylated) DNA. This strategy is particularly important for large genome plant species such as maize and wheat where the objective is to bias the sequencing towards unmethylated, single copy regions of the genome. In small genome species such as rice, peach, or *Arabidopsis*, complexity reduction is neither necessary nor particularly desirable; so in these cases, the restriction enzyme digestion serves primarily to provide sites for barcode attachment and primer annealing. Regardless of the need to reduce complexity in a given genome, the desire to maximize efficiency and reduce cost has led to the widespread use of GBS protocols that use multiplexing based on barcoding at restriction enzyme sites. Restriction Site Associated DNA (RAD) tags, Diversity Arrays Technology (DArT), reduced-representation sequencing, and low-coverage genotyping all implement restriction enzyme digestion for the dual goals of complexity reduction and creating barcode/primer attachment sites (Baird et al. 2008; Davey et al. 2011; Wenzl et al. 2004).

To evaluate the capacity of GBS to bridge the genotyping gap for rice mapping and breeding populations, we applied the low-coverage (384-plex) GBS protocol described by Elshire et al. (2011a, b) to a population of recombinant inbred lines (RILs) resulting from the cross of IR64 (*indica*) × Azucena (*tropical japonica*). This population represents an ideal test case for using GBS to add high-density SNP markers to a mapping population due to the wide variety of segregating traits present in the RIL progeny as a result of genetic divergence between the *indica* and *tropical japonica* parents, as well as the immortality of the RIL lines. The population consists of 176 F_{10}-F_{12} lines developed by single seed descent and like many classic mapping populations, has been previously genotyped with only sparse SSR markers, 200 in the case of this population (This et al. 2010). The population, or a doubled haploid population derived from the same parents, has already been used to dissect the genetic basis of several complex traits, including aluminum tolerance, root architecture, leaf width, plant ion concentration, and many other morphological and agronomic characteristics (Clark et al. 2011; This et al. 2010; Famoso et al. 2011; Hemamalini et al. 2000; Hittalmani et al. 2003; Li et al. 2003; Prasad et al. 2000; Sallaud et al. 2003; Stangoulis et al. 2007). It is our hypothesis that by saturating the RIL population with dense SNP markers, we will be able to further capture additional QTL for agronomic traits of interest and better resolve the genetic architecture of the population, including regions of segregation distortion and recombination hot and cold spots.

The necessarily intense bioinformatics effort required to analyze sparse GBS data resulting from low-coverage protocols is an obstacle for many poorly resourced programs. We therefore developed a pipeline to streamline the process of adding SNPs to RIL populations such as the IR64 × Azucena population tested here. This pipeline includes alignment of rice GBS data to the reference genome, SNP calling and imputation, and identification and elimination of error, typically 1 % of SNP calls post-imputation, or approximately 50,000 errors in our dataset.

We report our results aligning our GBS data to the rice reference genome using three different algorithms: BWA, Bowtie2, and PANATI. These are just three of many possible sequence aligners. BWA and Bowtie2 are perhaps the two most widespread alignment methods, and both are widely used for aligning plant sequencing data to a reference genome, however both were developed for analyzing the human genome, and are thus optimized for aligning low-diversity genomes. PANATI is an alignment and SNP discovery/genotyping algorithm originally developed by Wright (2011–2013) for high-diversity populations such as rice and/or populations that have no proper reference genome but can be analyzed against the reference assembly of a closely related species. Such populations may be significantly diverged from this proxy reference sequence, even if diversity within the population is low. Ilut et al. (2012) demonstrated PANATI's utility in this case by aligning RNA-seq reads from different *Glycine* diploid and allotetraploid species against the reference sequence for *Glycine max* (cultivated soybean), finding that in simulations, PANATI could align RNA-seq reads to the correct homeolog of the polyploid reference sequence 98.6 % of the time when 5 % divergence from the reference for the simulated reads was assumed.

Using the pipeline developed here in conjunction with a 384-plex low-coverage GBS protocol (Elshire et al. 2011a, b) we successfully mapped more than 30,000 high-quality SNP markers onto the IR64 × Azucena RIL population. Indeed, it is hoped that the efficiency, low cost, and availability of a high-quality analysis pipeline, as outlined in this paper, will make GBS accessible and useful to a greater number of breeders and geneticists. With the availability of next-gen sequencing, low marker coverage should no longer limit the resolution of genetics experiments or genomic-assisted breeding efforts.

Materials and methods

The population

A population of 176 F_{10}-F_{12} RILs was developed by single seed descent (SSD) from a cross between IR64 × Azucena under greenhouse conditions at IRD, Montpellier, France. During the first 7 SSD generations, selfing was controlled by bagging the panicles. IR64 and Azucena belong to the two most distant varietal groups found within *O. sativa*—*indica* and *japonica*, respectively—and have very contrasting morpho-physiological and adaptive characteristics. IR64 is an improved semi-dwarf variety bred by the International Rice Research Institute (IRRI) in the 1960s for favorable irrigated ecosystems, while Azucena is a traditional, tall, aromatic landrace from the Philippines cultivated in upland ecosystems. Mapped with some 200 SSR markers prior to this publication (This et al. 2010), the IR64 × Azucena RILs population represents an important immortal mapping resource for rice.

Plant material

Young leaf tissue was collected from each of the 176 IR64 × Azucnea RILs and the two parents (IR64 and Azucena) and DNA was extracted using the Qiagen 96-plex DNeasy kit as per the Qiagen fresh leaf tissue 96-plex protocol (http://www.qiagen.com/HB/DNeasy96Plant).

Library preparation

384-plex libraries were prepared as described in the protocol by Elshire et al. (2011a, b). *Ape*KI was selected for use with the protocol due to its methylation sensitivity and uniform distribution of cut sites across the rice genome (Online resource 1). 12 μl of 384-plex adapters were obtained from the Cornell Institute for Genomic Diversity (sequences available at http://www.maizegenetics.edu) and were used for the ligation reaction along with 100 ng of high-quality DNA. Post-ligation reactions, 5 μl of each of the 384 reactions were pooled in a total of 10 mL Qiagen PCR clean-up kit binding buffer. The pooled solution was then divided evenly among, and bound to, four Qiagen spin columns. PCR cleanup then proceeded as per the Qiagen PCR clean-up protocol for each of the four columns, producing four tubes of "pre-PCR" GBS library. Library preparation then proceeded as per the published 96-plex protocol (Elshire et al. 2011a, b). Eight replicates of IR64 and 10 replicates of Azucena were included in the 384-plex library.

Upon initial analysis, it was clear that 16 reactions failed sequencing, likely as a result of low-quality DNA samples. New DNA was extracted from frozen tissue collected from individuals 8, 16, 22, 33, 35, 72, 102, 107, 130, 131, 140, 158, 164, 165, 188, 270 using the Qiagen DNeasy kit. The new samples were then analyzed using the 96-plex GBS protocol with 12 μl 96-plex adapters and 100 ng DNA. Another four replicates of each parent were included on the 96-plex library. The rest of the 96-plex library was filled with samples from another project—the data from these samples were separated and removed from the IR64 × Azucena data via de-multiplexing prior to data analysis.

This combined approach produced an average of 6,788,434.00 reads for the two parents, and an average of 607,918.44 reads for the 176 RILs. The average number of reads per individual sequenced at 384-plex coverage was 452,427.13, while the average number of reads per individual sequenced at 96-plex was 967,296.10, for an average of 5,429,125.51 and 11,607,553.20 base pairs

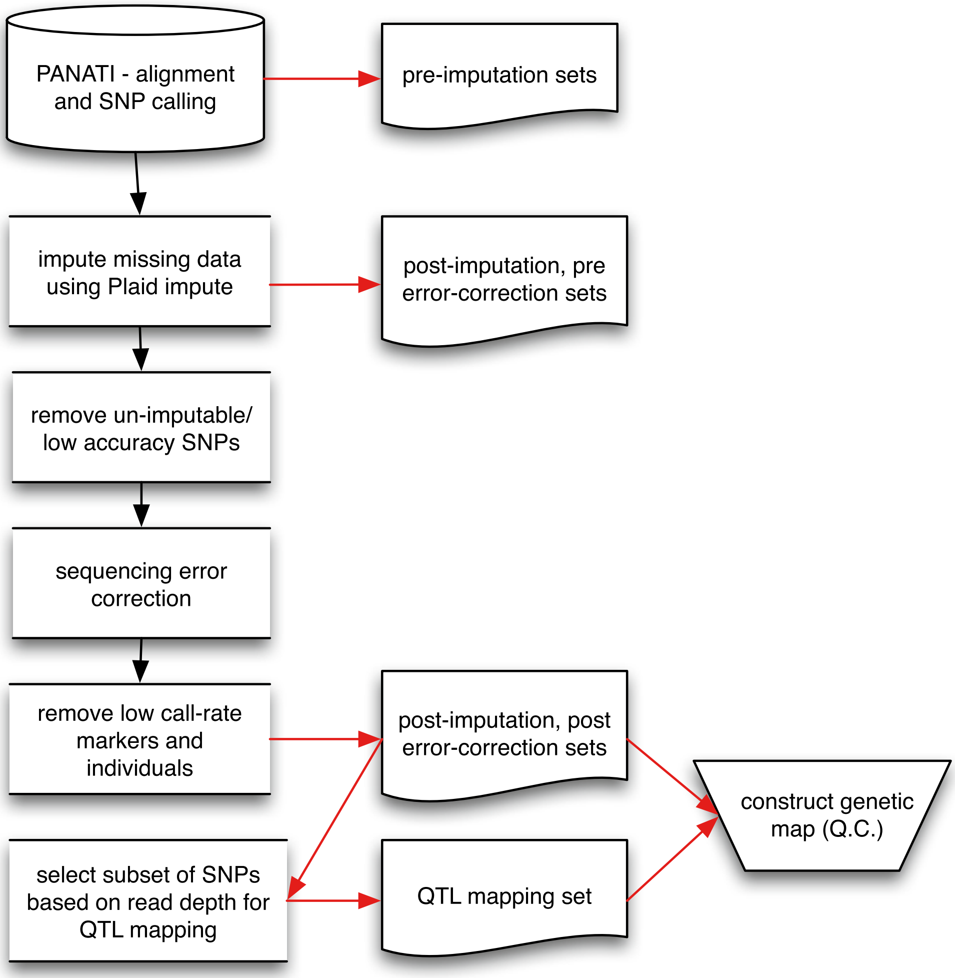
PANATI - alignment and SNP calling
pre-imputation sets
impute missing data using Plaid impute
post-imputation, pre error-correction sets
remove un-imputable/ low accuracy SNPs
sequencing error correction
remove low call-rate markers and individuals
post-imputation, post error-correction sets
construct genetic map (Q.C.)
select subset of SNPs based on read depth for QTL mapping
QTL mapping set

Fig. 1 Flowchart outlining our final, custom rice GBS pipeline. First raw GBS data is aligned to the rice reference genome and SNPs called using our custom aligner/SNP caller PANATI. Missing data is then imputed using Plaid-impute. PLUMAGE, a collection of custom python scripts, is then used to remove unimputable and low-accuracy imputable SNPs from the dataset. A sequencing error correction is then implemented to change likely sequencing errors to "NA", and any markers with 25 % or more missing data are removed. If desirable, a subset of this dataset can be selected for QTL mapping. Regardless, a genetic map is constructed as a means of quality control using the R/qtl package kosambi mapping function

sequenced/individual at the two different multiplexing levels, respectively.

Data analysis

A custom-designed pipeline combining a novel alignment algorithm and SNP caller (PANATI), imputation script (GBS-PLAID), and error correction and quality control (PLUMAGE) was developed for streamlined data analysis (Fig. 1).

Short read alignment and SNP calling

Three different alignment and SNP calling methods were used to produce three pre-imputation GBS datasets: (1) BWA sequence alignment in conjunction with the TASSEL GBS SNP discovery pipeline, available publicly at maizegenetics.net (BWA-TASSEL) (Bradbury et al. 2007), (2) Bowtie2 sequence alignment in conjunction with the TASSEL GBS pipeline (Bowtie2-TASSEL), and (3) PANATI, our in-house combination sequence aligner and SNP caller (available on request). For all three datasets, data from both the 384-plex and 96-plex libraries were analyzed together as one joint library, providing GBS data for all 176 RILs plus two parents.

BWA-TASSEL

For the BWA-TASSEL dataset, a single key file containing all IR64 × Azucena individuals and parent replicates from both the 384 and 96-plex libraries was used with the TASSEL GBS pipeline to identify good quality, unique, sequence reads with barcodes (termed "tags" by the pipeline developers). These sequence tags were aligned to the MSU v 6.0 Nipponbare rice reference genome using the Burrows-Wheeler Aligner (BWA)(Li and Durbin 2010), the SNPs were then called using the TASSEL quantitative SNP caller. Identical SNPs and parent replicates were merged using the MergeDuplicateSNPs and MergeIdenticalTaxa plugins. SNPs monomorphic to the two parents were removed as part of the following imputation step. Online resource 2 contains the exact commands and parameters used to generate the dataset. Details and directions for implementing the TASSEL GBS pipeline including details of key file creation are available online in the TASSEL 3.0 genotyping by sequencing pipeline documentation at http://www.maizegenetics.net. Details and directions for implementing BWA alignment are available online at the BWA sourceforge page (http://bio-bwa.sourceforge.net/bwa.shtml).

Bowtie2-TASSEL

The Bowtie2-TASSEL dataset was obtained exactly as the BWA-TASSEL dataset, however instead of aligning sequence tags to the rice reference genome using BWA, tags were aligned using Bowtie2 v2.0.0-beta7 (Langmead and Salzberg 2012). Online resource 3 contains the exact commands and parameters used to generate the dataset. Details and directions for implementing Bowtie2 can be found online at the Bowtie sourceforge page (http://bowtie-bio.sourceforge.net/index.shtml).

PANATI

PANATI is an independent map-to-reference alignment/mapping tool for short read sequences with integrated population sample SNP and small in/del (<20 bp) discovery and simultaneous genotyping. PANATI was originally designed with specific attention to the characteristics of *O. sativa* populations and the related wild species *Oryza rufipogon* for the analysis of population samples with genome-wide high coverage (10X or greater) and is known to be accurate and sensitive in these settings. For use with GBS data, PANATI was modified and extended to include sample extraction from bar-coded multiplexed FASTQ files using key files similar or identical to those used by the TASSEL-based pipelines above (see TASSEL documentation for details on key file creation), reference index construction restricted to GBS enzyme recognition site(s), and improved performance for low-coverage samples.

The SNP discovery and simultaneous genotyping step in the PANATI pipeline works the same as for deep coverage population samples with unrelated individuals, but specific options can be set to take advantage of the fact that the sample collection here is a RIL mapping population with the parents sampled to higher coverage than progeny. Namely, the PANATI "combine-samples" program that performs this step can be instructed to treat all progeny samples as outgroup samples, so that only polymorphisms between the two parent samples are discovered but the discovered polymorphisms are genotyped at all samples. Combine-samples can be further instructed to only output polymorphisms that segregate between the parent samples and therefore only those polymorphisms for which both parent samples have a confident genotype call.

Alternatively, the opposite approach can be used where information is pooled across progeny to discover polymorphisms at a high stringency even though the low coverage in any individual sample might prevent a high-confidence polymorphism call on the basis of the individual samples alone. Using combine-samples in this mode is appropriate if parent samples were not sequenced or not sequenced deeply enough. PANATI combine-samples output genotypes in standard VCF format with phred-scale polymorphism call confidence scores and individual genotype call confidence scores. Unlike the outputs of the other two pipelines, polymorphisms and genotypes can be filtered on the basis of these confidence scores.

PANATI v3.10 source code as well as a UNIX makefile for automating PANATI execution on this dataset is available on request. Default PANATI v3.10 options were used except for specifying the *Ape*KI recognition site for index generation.

Imputation (GBS-PLAID)

Following short read alignment and SNP calling using one of the three methods described above, missing genotype calls as a result of too few or no reads observed at a locus were imputed using a program ("GBS-PLAID") developed for this work and designed for GBS on bi-parental mapping or breeding populations. The method employed works by resolving phase of two-locus haplotypes using a Bayesian framework where the prior reflects the relative expectation of coupling versus repulsion haplotypes and any preference for either parent's haplotype given the breeding scheme of the population. Posterior haplotype probabilities are then computed using the observed data from all samples where both loci have a genotype call. For samples which are missing data at the locus to be imputed but have a genotype call at the reference locus, posterior probabilities of the diploid genotype at the missing locus are computed based on the probability of the necessary two-locus haplotypes for each possible genotype combined with a prior for the genotype reflecting any expected bias for a parental allele and bias for or against heterozygote genotypes. In the case of RIL populations with homozygous parents, the genotype prior reflects equal expectation for either parent allele as a homozygote and bias against observation of a heterozygote genotype.

This simple framework is then naturally extended such that adjacent markers both 5′ and 3′ of the imputed locus are used as reference loci. The number of markers on either side can be selected by the user. A larger number of markers results in a larger fraction of missing data having an imputed genotype but at the expense of potentially lower confidence in these genotypes as more distal markers have a higher fraction of recombinants. For mapping populations with known parental genotypes, linkage is extensive and in rice the density of GBS markers is high; most genotypes can be imputed confidently.

As a measure of imputation accuracy, GBS-PLAID also calculates imputed genotypes and their posterior probabilities for genotypes that are already observed in the output of any of the three pipelines to which GBS-PLAID is applied. The accuracy of imputation is estimated as the fraction of observed genotypes that match the imputed genotypes that met the minimum confidence threshold. These values are calculated for each locus and can be used downstream to filter out markers with lower accuracy estimates. GBS-PLAID reads VCF genotype data and currently outputs HapMap format with missing data replaced by imputed genotypes along with marker summary information such as the number of missing genotypes remaining and the accuracy estimate for the marker. To connect the TASSEL-based pipelines to GBS-PLAID, TASSEL's HapMap format is converted to an interim VCF format without confidence scores (equivalents are not provided in TASSEL's output) which is then used as input to GBS-PLAID. Output of the input VCF except with missing genotypes filled by their imputed values along with confidence scores corresponding to the posterior probability of the imputed genotype is planned for the next version of GBS-PLAID. This could be used to estimate a genotype confidence value for genotypes observed in TASSEL outputs simply by inserting the phred-scale confidence score corresponding to the posterior of the observed genotype as if it was imputed.

For this analysis, GBS-PLAID command line options were set such that at least 15 minor allele observations (-m 15) and at least 60 samples with observed genotypes (-n 60) were required to accept a marker on input for imputation and use as a reference locus. Any marker not satisfying these constraints are dropped from the input and excluded from output. We used five flanking markers both 5′ and 3′ as reference markers for imputing genotypes (-w 5). Other settings give similar imputation results. GBS-PLAID is available as part of our GBS data analysis pipeline as Online resource 7 (also available online at http://www.ricediveristy.org/data).

Post-imputation error correction and filtering (PLUMAGE)

All post-imputation data filtering and error correction were performed using PLUMAGE, a streamlined pipeline consisting of custom Python scripts for GBS data analysis, now publicly available as part of our GBS data analysis pipeline as Online resource 7 (also available online at http://www.ricediversity.org/data) (Fig. 1). Our first step post-imputation was to remove all SNPs that were either unimputable or had imputation accuracy scores lower than 95 % (see previous section on imputation for details on why SNPs can be unimputable or low accuracy). The next step was to implement a basic sequencing error correction. For every individual and for each chromosome, recombination breakpoints were tested for errors. If a breakpoint was followed by at least four SNP calls on different tags without reverting to the previous parent allele, the breakpoint was considered true. Otherwise, the breakpoint call was considered an error, and changed to "NA", to represent "missing data". Following the sequencing error correction, markers with 25 % or more missing data were removed from the dataset. Individuals with >8 % missing data (user-defined threshold) can also be identified and removed at this juncture in the pipeline via an optional flag, however this was not done for the dataset reported here for the sake of completion. The data prior to running them through the three steps described above are referred to as the "post-imputation, pre-error correction" data. The data after they are run through these three steps are referred to as the "post-imputation, post-error correction" data (Fig. 1).

As a final, important quality control step, for all datasets generated including the pre-imputation, post-imputation, pre-error correction, and post-imputation, post-error correction, a genetic linkage map was calculated using the R/qtl Kosambi mapping function (R version 2.15.1, R/qtl package 1.24.9). Specifically, to calculate the genetic map, the complete dataset-of-interest (either pre-imputation, post-imputation, pre-error correction, or post-imputation, post-error correction) including linkage groups based on the physical map (i.e. chromosome numbers) was loaded into R/qtl in the: "csvr" format (A PLUMAGE script is available to convert the default hapmap-formatted data into the R/qtl "csvr" format). The data was then coded within R/qtl as an RIL population, after which, the genetic map for the population was calculated using the R/qtl est.map() function with the map.function parameter set to "kosambi". The Kosambi mapping function calculates map distance (m) between two markers on the same chromosome as $\frac{1}{4}\ln\left(\frac{1+2c}{1-2c}\right)$, where c is the observed recombination frequency between the two markers. The order of the markers

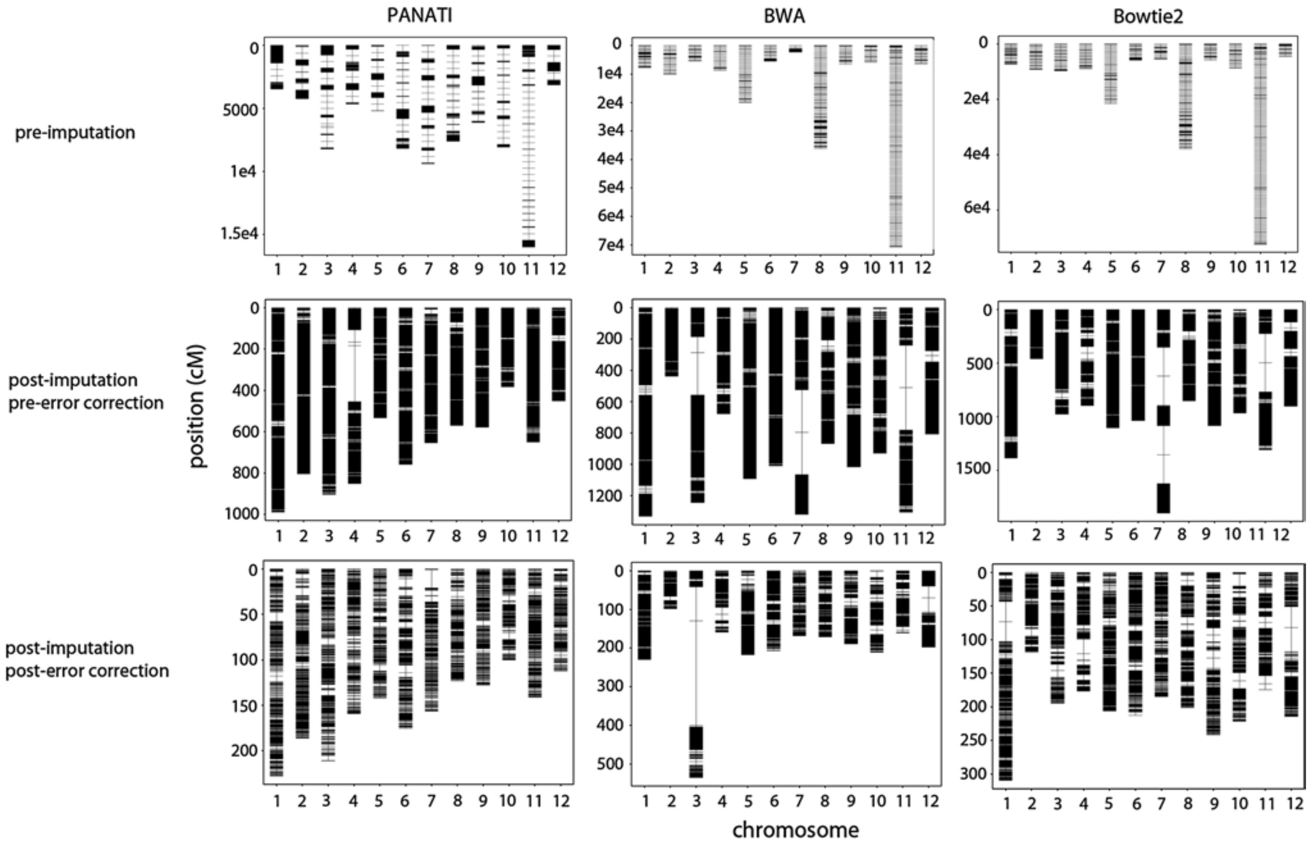

Fig. 2 Pre-imputation, post-imputation-pre-error correction, and post-imputation-post-error correction genetic maps resulting from PANATI, BWA-TASSEL, and Bowtie2-TASSEL IR64 × Azucena GBS datasets. Note that the *y*-axis scales of the nine maps are not the same. As the amount of error decreases in the dataset (at each stage of pipeline) the degree of map distention also decreases. Only the PANATI post-imputation, post-error correction dataset, however, produces the "expected" genetic map for rice. Note also that expansions such as those seen on chromosome 3 of the BWA-TASSEL or chromosome 1 of the Bowtie2-TASSEL post-imputation, post-error correction genetic maps cannot be fixed by removing markers in the expanded regions. This is because expansion is not the result of double recombination between one or two pairs of markers but rather results from a series of errors that results in two essentially independent linkage groups on the same chromosome

along the chromosome was fixed using the SNP physical map positions. The Kosambi function was selected over other mapping functions because it allows for modest interference among double cross-over events and is therefore thought to be a more accurate representation of true map distances than, for example, the Haldane mapping function which does not account for interference (Walsh 1998) (see our publicly available R/qtl mapping code for exact commands). The genetic maps were converted to visual representations where vertical lines represent the chromosomes and short horizontal lines represent the markers using R/qtl. The spaces between the horizontal lines are proportional to the map distances between markers (Fig. 2). The number of breakpoints per RIL per chromosome was counted using a custom Python script. All counted breakpoints were then summed to obtain the total number of breakpoints for the population. Per chromosome averages were obtained for the final PANATI post-imputation, post-error correction dataset by averaging the number of breakpoints per chromosome for all lines in the population. Standard deviations are reported for these averages (Table 2).

In some cases, users of GBS data may wish to choose subsets of a large dataset that are uniformly distributed across the genome. To facilitate these analyses, we developed an algorithm (included in PLUMAGE) for choosing subsets of SNPs evenly spaced across the genome. Interval size between selected SNPs is determined via a bin parameter. First, the total SNP set is binned according to the desired spacing of SNPs, then the SNP with the deepest sequencing coverage is selected from each bin to form the subset. For this study, a QTL mapping subset was developed by selecting 1 SNP every 240 Kb (approximately 1 cM) from the final post-imputation, post-error correction dataset (no genetic map is shown for this QTL mapping subset as a quality map is shown for the superset). Another PLUMAGE script allows the user to go back and select additional SNPs in specific regions of interest, if desired. This allows a user to increase SNP density in one or more target regions to facilitate fine mapping and/or marker-assisted selection.

Analysis of coverage, segregation distortion, recombination frequency, and call rate by dataset

Call rates were calculated per SNP as the percent of individuals that had a non-"missing data" call in any given dataset, and read number was calculated as the number of sequencing reads that covered a given SNP. Call rate distributions were calculated using the JMP® Pro 10.0.0 statistical program by SAS. SNPs were put into 1,550 250 Kb bins to assess the genome-wide coverage of each SNP set (86/1,550 bins were empty), and the number of SNPs in each bin was charted using JMP. The average number of sequence reads, calculated as the average of the number of reads covering the SNPs in a particular bin, was then overlaid on the distribution of SNP counts as a line. Segregation ratios were calculated for every SNP in the final post-imputation, post-error correction PANATI dataset, as well as for the 200 SSRs already placed on this IR64 × Azucena population, and the results plotted against physical position using JMP. The ratio of genetic:physical position of SNPs was obtained by dividing a SNP's genetic position (cM) by its physical position (Mb). The results were plotted by physical position using JMP.

QTL mapping

Aluminum tolerance

The QTL mapping was performed using both the full, post-imputation, post-error correction PANATI marker set (30,984 markers) and the QTL mapping subset (1,464 markers) on all 171 genotyped RILs. Previously published aluminum tolerance phenotype data (Famoso et al. 2011) were used to validate the mapping of new marker sets and demonstrate the value to QTL mapping of saturating a mapping population with markers. For details on phenotype data collection, see Famoso et al. (2011). QTL mapping was performed using the R/QTL package (R version 2.15.1, R/qtl package 1.24.9), and the same code was used for both the full 30,984-marker set and the 1,464-marker subset. The datasets were loaded into R/qtl and genetic maps calculated as described previously in the methods section on post-imputation error correction and filtering. After calculating the genetic map, the genetic marker positions were "jitter-mapped", i.e. adjusted very slightly, to avoid identical positions for markers on different chromosomes, after which the underlying genotype probabilities were calculated using the R/qtl calc.genoprob() function and the Kosambi mapping function (see previous section for details on the Kosambi function). An initial single-marker QTL scan was then performed using the scanone() function with Haley-Knott Regression, under the assumption that the phenotype data were normally distributed. 1,000 permutations were used to determine the LOD threshold for significance. After scanning for initial QTL, the QTL model was refined by scanning for additional linked QTL, still using Haley-Knott Regression and assuming the phenotypes were normally distributed, but conditioning on the QTL already detected. The model was finalized by using stepwise forward selection and backward elimination to probe the model space for the best fit QTL model for the data. An ANOVA analysis was run on the final model to determine the percentage of variance explained by each QTL and the estimated effect sizes. The peak QTL positions are reported along with the right and left flanking markers, which correspond to the nearest flanking marker within 1.5 LOD units of the peak marker. Together, the interval constructed by the two flanking markers roughly represents the 95 % confidence interval for the QTL (Dupuis and Siegmund 1999; Mangin et al. 1994). Given the high density of markers on the population, this procedure is equivalent to composite interval mapping methods (Darvasi et al. 1993). The QTL mapping code used in this study is available publicly as Online resource 7 and online at http://www.ricediversity.org/data and is generalized for convenience of use.

Leaf width

The same 30,984 and 1,464 marker datasets used to map QTL for Aluminum tolerance were used to map QTL for leaf width using data generated as part of this study. The RIL population was planted in Guterman Greenhouse 160 at Cornell University in Ithaca, NY, in late September 2010 and was phenotyped at maturity in January 2011. Three replicates of each RIL were planted in a randomized complete block design. Three mature leaves from each replicate were measured at the widest point and leaf width per plant was calculated as the mean of the three measurements. The grand mean of the three replicates was calculated for each RIL and used for QTL mapping. The same QTL mapping procedure and code used to map the aluminum tolerance QTL (described above) was also used to map the leaf width QTL.

Results

The GBS sequencing reads were aligned to the rice reference genome using either BWA (Li and Durbin 2010), Bowtie2 (Langmead and Salzberg 2012), or PANATI (Ilut et al. 2012) (see methods for details). SNPs aligned using BWA or Bowtie2 were called using the TASSEL GBS pipeline (http://www.maizegenetics.net/

Table 1 Statistics for post-imputation, post-error correction, and post-imputation, pre-error correction, and pre-imputation datasets

Dataset	Total markers	Genetic map length (171 RILs)	Total number of breakpoints for 171 RILs	Median marker call rate
PANATI pre-imputation	66,780	84,388.95	76,167	33.53
PANATI post-imputation, pre-error correction	30,991	8,128.98	35,856	100
PANATI post-imputation, post-error correction	30,984	1,862.96	6,160	99.4
BWA pre-imputation	56,422	185,274.49	86,314	47.37
BWA post-imputation, pre-error correction	43,438	12,032.54	48,358	100
BWA post-imputation, post-error correction	43,280	2,546.42	7,310	98.86
Bowtie2 pre-imputation	60,245	197,457.92	91,733	47.98 %
Bowtie2 post-imputation, pre-error correction	46,111	12,863.11	50,254	100
Bowtie2 post-imputation, post-error correction	45,946	2,454.23	7,620	98.86

index.php?option=com_content&task=view&id=89&Itemid=119), while SNPs aligned with PANATI were called with PANATI, our in-house alignment and SNP calling algorithm. Any of the three methods produced initial pre-imputation GBS datasets that contained between 56,400 and 66,800 polymorphic SNPs, with the PANATI dataset containing the most SNPs (Table 1). All initial data, however, were very sparse with median call rates of 47.4, 48.0, and 33.5 percent for the BWA-TASSEL, Bowtie2-TASSEL, and PANATI datasets, respectively (Table 1, Online resource 4). Furthermore, pre-imputation data were subject to high error, as evidenced by massive expansions in the genetic map (Table 1; Fig. 2). The pre-imputation BWA-TASSEL and Bowtie2-TASSEL datasets had total genetic map sizes of 184,275 and 197,458 cM, respectively, 120–130 times the expected size of 1,520 cM for the rice genetic map (Harushima et al. 1998). The PANATI SNP set produced a genetic map of 84,389 cM, or 55 times the expected size (Table 1).

To address both sparseness and error rate, all three data sets were imputed (see Methods for details) and all non-imputable SNPs or SNPs with imputation accuracies lower than 95 % were discarded. As a result, in all post-imputation (but pre-error correction) SNP sets, median call rates were equal to 100 % (Table 1, Online resource 4). Removal of unimputable and low-imputation accuracy SNPs also decreased genetic map expansion, although all three maps remained elongated (Table 1; Fig. 2). The PANATI set produced a genetic map that was 8,129 cM long, while the BWA and Bowtie2 genetic maps were 12,032 and 12,863 cM long, respectively (Table 1).

Remaining map distention was thought to result from a combination of sequencing errors and tag misalignments, so a simple sequence error correction algorithm was implemented (see "Materials and methods"). While median call rates remained high for all three datasets post-error correction (between 98 and 99.5 %), only the final post-imputation, post-error correction PANATI dataset produced a genetic map with zero distended chromosomes, a reasonable genetic map length of 1,862 cM, and a total of 6,160 breakpoints across all 171 RILs, or ~36.02 breakpoints per RIL (Fig. 2; Table 1, Online resource 4). Upon removal of three individuals with missing data greater than or equal to 8.0 % (individuals 153, 206, and 293), the number of breakpoints on the 168 RILs further drops to 5,348, for an average of 31.83 breakpoints per RIL. The average number of breakpoints per chromosome, along with the standard deviations from the mean is given for both the full 171 RILS and the 168 RILs in Table 2. Removal of the three individuals with large degrees of missing data significantly lowered the standard deviations on all chromosomes, in addition to adjusting the mean values, but did not significantly change the distribution of markers on the genetic map (data not shown).

By contrast to the PANATI dataset, in the BWA-TASSEL dataset, chromosome three remained elongated, while in the Bowtie2-TASSEL dataset, chromosomes 1 and 12 were slightly distended (Fig. 2), unless more stringent imputation parameters were used (data not shown). Breakpoint counts were higher in both the BWA-TASSEL and Bowtie2-TASSEL datasets as well, with 7,310 and 7,620 breakpoints on 171 RILs for the BWA-TASSEL and Bowtie2-TASSEL datasets, respectively (Table 1). It is important to note that in both cases the map distensions did not result from one or two "bad" markers which could hypothetically be removed from the datasets, but from distinct

Table 2 Mean and standard deviation of the number of breakpoints per chromosome for the final PANATI post-imputation, post-error correction dataset calculated using either all 171 RILs (all RILs) or 168 RILs (168 RILs) after removal of three individuals with missing data ≥8.0 %

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	Grand
All RILs mean breakpoints/chrom	4.40	3.61	4.10	3.10	2.76	3.36	2.92	2.39	2.50	1.94	2.77	2.19	3.00
All RILs standard deviation	3.85	3.21	3.67	3.79	2.36	4.77	3.65	2.81	3.65	1.78	4.72	2.77	3.42
168 RILs mean #breakpoints/chrom	4.07	3.30	3.70	2.71	2.54	2.84	2.57	2.14	2.06	1.80	2.23	1.88	2.65
168 RILs standard deviation	2.42	1.79	1.89	1.69	1.55	1.88	1.47	1.40	1.58	1.42	1.60	1.30	1.67

sets of markers at both ends of the chromosome in question (e.g. 5 or 11) that were essentially unlinked. In other words, removing the markers that appear to lie between these groups does not change the picture of the map, suggesting that a high degree of stochastic error remains within the BWA-TASSEL and Bowtie2-TASSEL datasets; error that is detected when a genetic map is calculated.

The final post-imputation, post-error correction PANATI dataset thus contained 30,984 high-quality markers (Table 1) on 171 RILs. Publicly available dataset for 168 RILs with individuals 153, 206, and 293 removed, is available online at http://www.ricediversity.org/data. SNPs were well distributed across the genome, with an average of 21.16 SNPs per cM (240 Kb) (Wu et al. 2003). While SNPs were well distributed, they were not uniformly distributed. Some 250 Kb regions contained as many as 77 SNPs, while a very few contained none. Figure 3 shows this distribution for chromosome 1, along with the average number of sequence reads covering the SNPs in each bin (see Online resource 5 for all other chromosomes). In some cases, for example at 39 Mb on chromosome 1, a low number of SNPs/bin correlated with lower read coverage for the bin. However, in other cases, the opposite correlation was observed. For example, the bin beginning at 21.14 Mb on chromosome one contained only four SNPs, but those four SNPs were covered by an average of 539 sequence reads (Fig. 3). Overall, this suggests that micro regions of low-SNP detection were not necessarily the result of low-sequence coverage, but were due to the discarding of repetitive or methylated DNA, or resulted from low polymorphism between the parents. An example supporting this explanation can be seen in the region between 9 and 13.5 Mb on chromosome 5, a known SNP desert (Wang et al. 2009; Feltus et al. 2004; Nasu et al. 2002) that is well covered by sequencing reads in this dataset, but contains few SNPs (Online resource 5).

Segregation distortion

Segregation distortion is to be expected in any *indica* × *japonica* rice intercross due to the sterility barriers that exist between the two varietal groups. Identifying these regions has always been of interest to geneticists and breeders, however, with only 200 SSRs on a population such as the IR64 × Azucena RILs, it was not previously feasible to map more than the grossest trends in segregation distortion (Fig. 4). The high resolution of our final GBS marker dataset, however, greatly enhanced our ability to define the regions showing segregation distortion across the genome in this population. By graphing the segregation ratio (number of IR64 calls/Azucena calls at a given locus), we are able to visualize solid curves that range above and below the neutral segregation ratio of 1:1 in this RIL population (indicated by the red line in Fig. 4). Valleys below the red line represent regions of the genome favoring Azucena alleles, while peaks above the red line represent regions favoring IR64 alleles.

Recombination frequency

Numerous groups have found recombination frequency to vary substantially across the rice genome (Chen et al. 2002; Wu et al. 2003; Zhao et al. 2002). The resolution of our new data also made it possible to map recombination hot and cold spots across the genome in this population. The ratio of a SNP's genetic:physical position (cM/Mb) was plotted versus the SNP's physical (Mb) position (Fig. 5). One cM in rice is approximately equal to 0.24 Mb (Wu et al. 2003); therefore, the expected ratio between the two units is approximately 4, represented on the graphs in Fig. 4 as a horizontal red line. With only 200 SSRs, it was not possible to accurately map recombination hot and cold spots, just as it was not possible to adequately map segregation distortion. However, by saturating the population with ~31,000 SNPs, we were able to clearly identify both regions of heightened recombination (peaks above the red line) and regions of depressed recombination (valleys below the line) (Fig. 5). Centromeres and pericentromeric regions, delineated on the graphs as vertical blue lines, correlated with regions of decreasing recombination frequency, although not necessarily with recombination cold spots, per se.

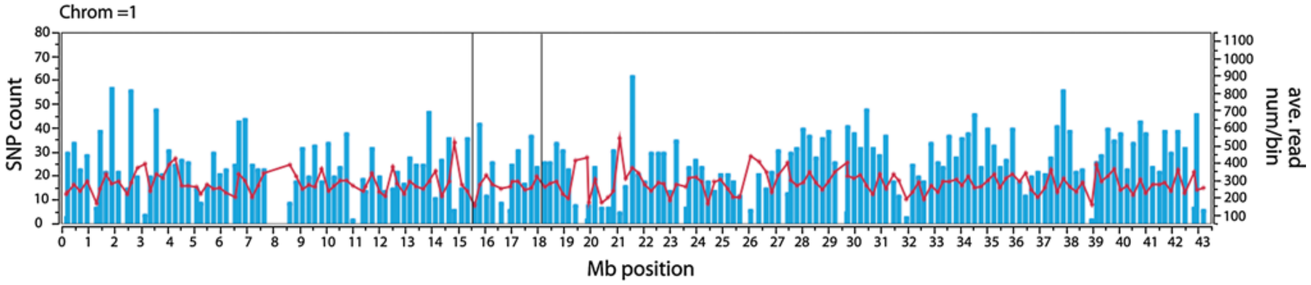

Fig. 3 Count of SNPs (*left y*-axis, *blue bars*) by 250 Kb bin (bin start position in Mb on *x*-axis) on chromosome 1, with the average number of reads/SNP/bin (*right y*-axis) overlaid as a *red line*. The number of reads is defined here as the number of sequence reads covering a given SNP. The average number of reads per bin (plotted here) was calculated by taking the mean of the read counts for the SNPs in each bin. *Vertical gray lines* delineate the pericentromeric regions. For other chromosomes, see supplement (color figure online)

QTL mapping

To demonstrate the quality of our final post-imputation, post-error correction PANATI dataset and the value to QTL mapping of saturating a mapping population with SNPs over using more sparsely distributed markers, we used both the entire 30,984-SNP post-imputation, post-error correction set, as well as a 1,464-SNP subset, selected by choosing the SNPs covered by the highest number of reads every 240 Kb (cM), to re-map QTL for aluminum tolerance using publicly available phenotype data (Famoso et al. 2011), and to identify QTL for leaf width using previously unpublished phenotype data.

Aluminum tolerance

In Famoso et al. (2011), four QTL were identified as segregating for aluminum tolerance in the IR64 × Azucena RIL population based on an underlying marker dataset consisting of ~200 SSR markers. Using either the 1,464-SNP subset or 30,984-SNP full set, we were able to identify three out of the four previously mapped QTL (Table 3, Online resource 6). The fourth QTL, at 27.61 Mb on chromosome 2, which had the lowest LOD-score in the previously published analysis, was registered as a peak in our analysis, but did not pass our significance threshold. LOD scores used to determine significance of QTL are calculated empirically and thus the larger number of markers and higher probability of false positives (Type 1 error) in our dataset required an elevated LOD significance threshold. In addition to those QTL already identified by Famoso et al. (2011), when using the saturated map of 30,984 SNP markers, we also identified two additional significant QTL on chromosome 1 at 11.01 and 11.43 Mb. With LOD scores of 6.86 and 8.07, respectively, these data support the existence of a previously unidentified QTL in this region of chromosome 1, a region which, according to Fig. 5, also corresponds to a recombination hot spot. Together, in a multi-QTL model, the four Al tol (LRG) QTL identified using the full marker set explained 48.68 % of the variance (Table 3), while the two Al tol (LRG) QTL identified using only the subset of 1,464 markers explained only 27.96 % of the variance (Online resource 6). LOD scores associated with QTL identified using both the subset and full SNP set were very similar, although not identical (Online resource 6). Confidence intervals of all identified QTL are reported as the nearest right and left flanking markers within 1.5 LOD units of the peak marker in Table 3 and Online resource 6.

Leaf width

The results of mapping QTL for leaf width were also dependent on which SNP dataset was used. Using either the 1,464 SNP subset or the 30,984 SNP full set, we were able to identify two significant QTL for leaf width in the IR64 × Azucena RILs. Both QTL were located on chromosome 1, one at either 2.20 or 4.69 Mb (for the subset or full SNP set, respectively), and one at approximately 34.23 Mb (Table 3, Online resource 6). Both QTL have been previously identified in other studies of rice leaf width, further confirming the quality of our new SNP marker dataset. The QTL on chromosome 1 at 4.69 Mb corresponds to Qflw1, identified by Mei et al. (2003) in an (*indica* × *japonica* RIL) × *indica* F2 testcross population while the QTL at 34.2 Mb was identified by Yan et al. (2003) in another *indica* × *japonica* population (Gramene ID AQEJ025). In addition, using the full SNP set, we identified another four significant QTL: one on chromosome 1 at 41.34 Mb, one on chromosome 4 at 19.73 Mb, one on chromosome 5 at 21.08 Mb, and one on chromosome 8 at 26.79 Mb (Table 3). These QTL have also been identified in previous studies. The additional QTL on chromosome 1 was identified in the study by Yan et al. (2003), while the remaining additional QTL on chromosomes 4, 5, and 8 were identified in a third *indica* × *japonica* RIL population also by Mei et al. (Mei et al. 2005), further suggesting the value to QTL

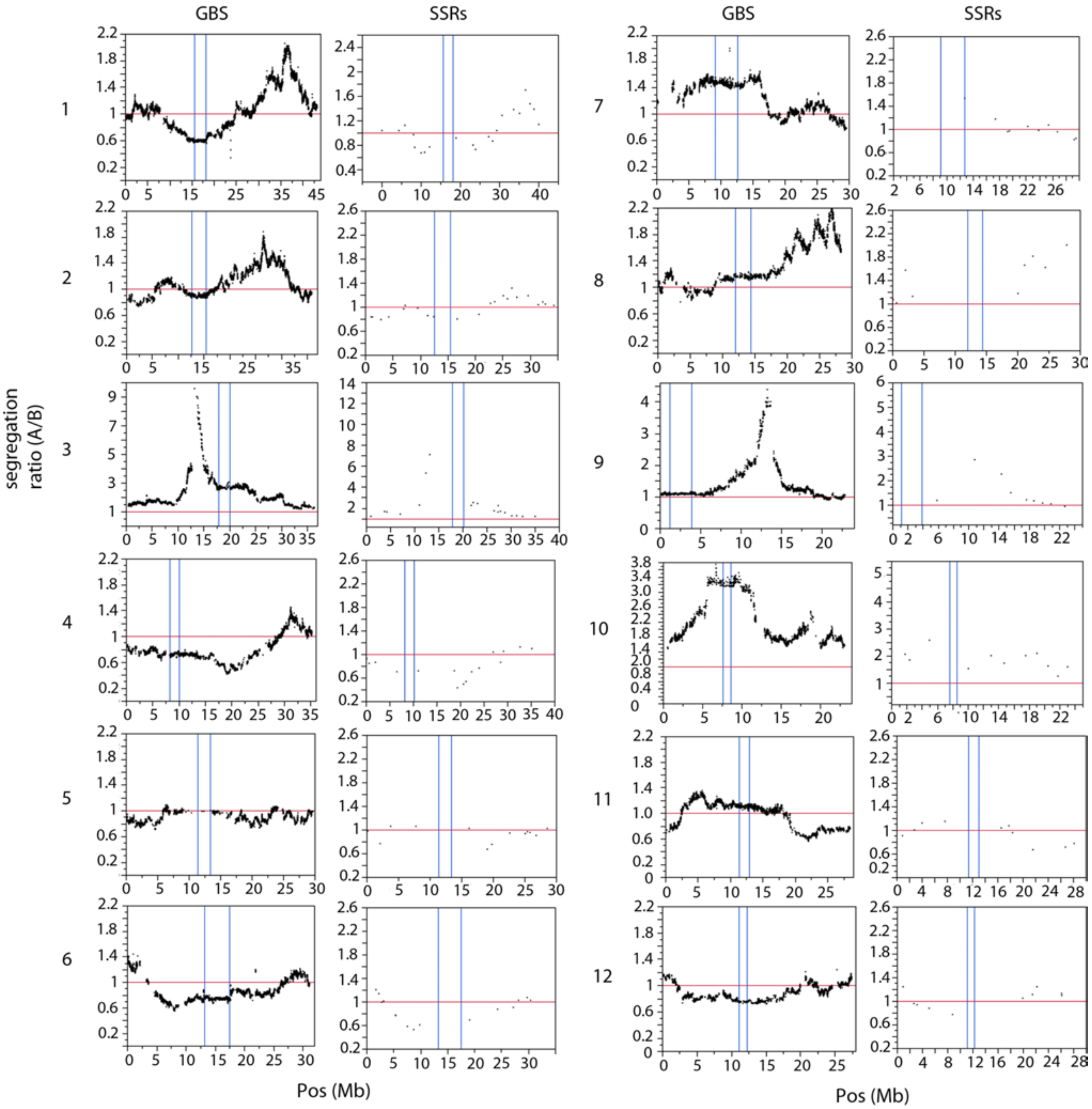

Fig. 4 Ratio of IR64 genotype (A) to Azucena genotype (B) by physical position (Kb) by chromosome for the 200 SSR markers originally mapped on this population (*right*) and the final PANATI post-error correction dataset. A *red line* is marked at the expected segregation ration of 1:1. *Blue lines* delineate the pericentromeric regions. In some cases the physical position of the SSR markers could not be determined. These SNPs were excluded from the analysis (color figure online)

mapping of saturating the mapping population with SNP markers.

In a multi-QTL model generated using the full marker dataset, these five QTL accounted for 53.1 % of the variation in mean leaf width (Table 3). By contrast, in a multi-QTL model generated using the 1,464 SNP subset, the two LW QTL identified only accounted for 27.6 % of the variation. As was the case for aluminum tolerance, the positions and LOD scores of the QTL identified by both the full SNP set and the SNP subset were very similar (Online resource 6). Confidence intervals of all identified QTL are reported as the nearest right and left flanking markers within 1.5 LOD units of the peak marker in Table 3 and Online resource 6.

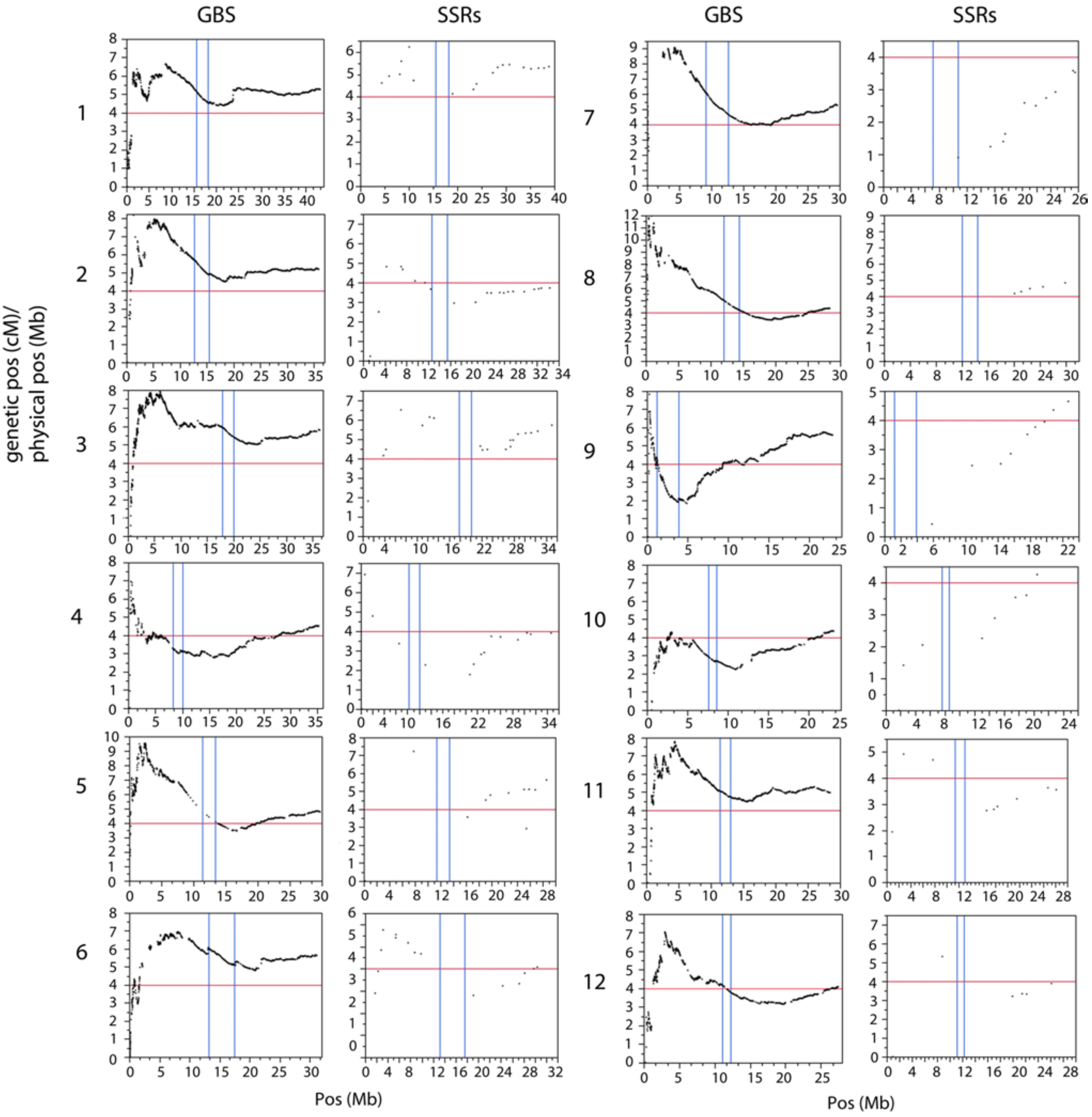

Fig. 5 Ratio of genetic position: physical position by chromosome for the 200 SSR markers originally mapped on this population (*right*) and the final PANATI post-error correction dataset. A *red line* is marked at the expected genetic:physical position ratio at 0.004. *Blue lines* delineate the pericentromeric regions. In some cases the physical position of the SSR markers could not be determined. These SNPs were excluded from the analysis (color figure online)

Discussion

Genotyping by sequencing (GBS) has generated high levels of interest within the plant breeding and genetics community. The low up-front cost of approximately $9.00/sample for 384-plex (Elshire et al. 2011a, b) and simple and straightforward library preparation protocol promises the ability to put thousands of markers on any population of interest—breeding, mapping or otherwise, thus bridging the genotyping gap between reference and non-reference lines and removing low marker coverage as a barrier to any genetics experiment or marker-assisted breeding effort. Our results suggest that under the right circumstances GBS can fulfill this hope, however, they also advise caution, as raw

Table 3 Significant QTL for Aluminum tolerance (Al tol) and leaf width (LW) using final post-imputation, post-error correction PANATI dataset (30,984 markers) on all 171 genotyped RILs

Trait	Chrom	Peak Mb position	Peak marker	Left flanking marker (Mb Pos)	Right flanking marker (Mb pos)	LOD	PVE[a]	Additive effect (IR64)	Published QTL ($^{+}$Gramene accession ID)
Al tol (LRG)	1	39.97	S1_39901238	S1_39467067 (39.47)	S1_40600105 (40.60)	4.68	9.62	−7.25 %	Alt_{TRG} 1.1/Alt_{PRG} 1.1
Al tol (LRG)	1	11.01	S1_11007213	S1_10687413 (10.69)	S1_11037843 (11.04)	6.86	14.67	−0.850 %	None
Al tol (LRG)	1	(11.43)	S1_11428700	S1_11044487 (11.04)	S1_11465139 (11.46)	8.07	17.64	−8.19 %	None
Al tol (LRG)	9	18.49	S9_18550466	S9_18000425 (18.00)	S9_18943615 (18.94)	5.82	12.20	−8.33 %	Alt_{LRG} 9.1
% Variance explained by multi-QTL model (Al tol-LRG) = 48.68 %									
Al tol (TRG)	12	3.49	S12_3487279	S12_2934532 (2.93)	S12_3780644 (3.78)	7.76	24.04	−4.02 %	Alt_{TRG} 12.1
Al tol (PRG)	12	3.72	S12_3724276	S12_2913082 (2.29)	S12_4615131 (4.62)	3.88	12.84	−4.72 %	Alt_{PRG} 12.1
LW	1	34.23	S1_34186208	S1_33820964 (33.82)	S1_34873874 (34.87)	12.04	18.86	1.02	AQEJ025^{+}
LW	1	4.69	S1_4689187	S1_2218565 (22.19)	S1_4826553 (48.27)	7.53	11.05	−0.61	Qflw1
LW	1	41.34	S1_41336661	S1_41172542 (41.17)	S1_42108167 (42.11)	5.25	7.45	−0.40	AQEJ031^{+}
LW	4	19.73	S4_19731255	S4_19311806 (19.31)	S4_20479063 (20.48)	4.42	6.20	0.63	AQFW156^{+}
LW	5	21.08	S5_21080332	S5_20448724 (20.41)	S5_21677168 (21.68)	4.32	6.05	0.79	AQCU145
LW	8	26.79	S8_26791035	S8_26385163 (26.39)	S8_27419806 (27.42)	4.22	5.90	0.50	Qflw8
% Variance explained by multi-QTL model (LW) = 53.1 %									

[a] Percent variance explained (by single QTL)

GBS data are sparse and prone to error, and the costs of the bioinformatics analysis necessary to address these two deficiencies are not factored into the “$9.00/sample” sticker price.

We therefore developed here a streamlined bioinformatics pipeline for adding markers to RIL populations to help lower the barrier posed by bioinformatics analysis to groups looking to use GBS to add markers to their mapping or breeding populations. In developing our pipeline, we experimented with three sequencing data alignment algorithm-SNP calling combinations: BWA-TASSEL, Bowtie2-TASSEL, and PANATI. In all three cases, construction of a genetic map, a once-standard practice that is now falling to the wayside with the increased prevalence of physical maps, was calculated as a means of obtaining a visual indication of and quantifying the error within the GBS dataset. Prior to imputation, all three datasets produced genetic maps that were 50–130 times the expected size of a rice genetic map. This extreme elongation of chromosomes occurred because the prevalence of error within the unimputed and unfiltered GBS datasets makes it “appear” as though many more double recombination events have occurred between markers than have, in reality, occurred (Lincoln and Lander 1992). In fact, these presumed double cross-overs result from incorrect SNP calls. In a smaller marker dataset, the effect of such an error rate might be relatively limited. However, as can be seen in Fig. 2, in a GBS dataset containing more than 50,000 markers, the effect of the SNP call error rate is multiplied by many orders of magnitude.

Interestingly, such genetic map expansion has been seen before in the rice genetic maps built using AFLP (Amplified Restriction Fragment Length Polymorphism) markers in the 1990s. As in restriction enzyme based GBS, in AFLP analysis samples are digested with restriction enzymes and the restriction fragments are ligated to adapters and pooled. The key difference is that in AFLP, the fragments are then size separated using polyacrylamide gel electrophoresis (PAGE) as a means of identifying size variants while GBS uses next-gen sequencing to identify SNP variants (Vos et al. 1995). Two different groups working with an IR64 × Azucena double haploid population (developed using the same IR64 × Azucena parents as in this RIL population) noted that chromosomes were “stretched” with the integration of AFLP markers into RFLP genetic maps (Maheswaran et al. 1997; Virk et al. 1998). In 1996, Maheswaran et al. (1997) specifically noted a correlation between genetic map size and the number of AFLP markers and hypothesized that these expansions had to be the result of map function error, possibly as a result of segregation distortion. Virk et al. (1998) followed up on this hypothesis in 1997 by trying to reduce the size of their genetic map by controlling for segregation distortion, without success. Not coincidentally, AFLPs in rice were quickly replaced with other more reliable marker systems, such as microsatellites/SSRs (McCouch et al. 1997). Now, with the growing popularity of GBS, we have stumbled back into the old set of problems associated with AFLPs—error, sparcity, and stretching of the genetic map. Fortunately, it is now possible to address both the SNP calling error and data sparcity present in the GBS data through a reasonable degree of data imputation and filtering.

The GBS data sparseness can be attributed mainly to the high degree of multiplexing per lane during sequencing, though it is also affected by the distribution of restriction enzyme cut sites and the filtering out of sequence reads that align to multiple locations in the genome. This data sparseness can be addressed by either lowering the degree of multiplexing (from 384-plex to 96-plex, or 96-plex to 48-plex), by running multiple lanes of a library (i.e., two lanes of a 384-plex library will generate twice the read number without having to make a new library) and/or by imputing missing data. As cost is a prime motivation for choosing to use GBS for genotyping in the first place, we focused on imputation as the solution to our data sparsity problem, and designed GBS-PLAID to impute missing data calls on RIL populations, specifically, using a Bayesian framework (see “Materials and methods” for details). After imputation, we removed all non-imputable and low-accuracy SNPs as a quality control measure. While this step reduced our total number of SNPs by a little more than 50 %, it also greatly reduced the size of all three genetic maps while boosting the median SNP call rates to 100 % (Table 1).

While the post-imputation reduction in genetic map size was dramatic, removing unimputable or low-quality imputable SNPs alone was not enough to bring the genetic map sizes down to a reasonable size. Post-imputation, pre-error correction maps were still approximately 5–8 times larger than the expected genetic map size. These data suggested that sequencing errors still remained, so a simple sequencing error correction algorithm was introduced to the pipeline to change calls that are likely errors to “missing data”. This error correction was then followed by removal of any SNPs with call rates lower than 75 %. Under lax imputation parameters the implementation of this error correction on the PANATI resulted in a genetic map containing no elongated chromosomes. Under more stringent imputation parameters, the Bowtie2 and BWA datasets also produced genetic maps with no elongated chromosomes. The final PANATI dataset contained 30,984 markers, and had a total genetic map size of 1862.96 cM, a size comparable to the 1803 cM genetic map created from the 237 SSR markers and to the expected size for a rice genetic map (Harushima et al. 1998).

Similarly, Huang et al. (2009) identified an average of 33.83 breakpoints per RIL using 1,493,461 markers generated via whole-genome re-sequencing on a population of 150 rice RILs (Huang et al. 2009). The number for our final dataset was comparatively higher, at 36.02 breakpoints per RIL, until we removed the three RILs (individuals 153, 206, and 293) with more than 8.0 % missing data from the population. This reduced the total number of breakpoints on the 168 RILs to 5,348, for an average of 31.83 breakpoints per RIL—highlighting the ability of individual outliers to distort population averages. It is reasonable to expect that we might detect slightly fewer breakpoints than Huang et al. as our dataset contains only 30,984 markers; however, the fact that our number is so close to theirs indicates that ~31,000 markers provides essentially equivalent information as ~1.5 million markers for a rice RIL population of this size.

Notably, BWA-TASSEL and Bowtie2-TASSEL both still had at least one stretched chromosome after the sequencing error correction—chromosome 3 in the case of BWA, and chromosomes 1 and 12 in the case of Bowtie2—when the more lenient GBS-PLAID parameters applied here were used (see "Materials and methods" for details). These distorted chromosomes proved to be somewhat enigmatic. Removing the markers found in the stretched middle of these chromosomes did not decrease the genetic map size because the problem was not simply double recombination between one or two pairs of markers, but rather a series of errors that resulted in the calculated presence of two essentially independent linkage groups on one chromosome. The application of more stringent GBS-PLAID parameters, however, solved the problem first for Bowtie2, and then, upon applying even more stringent GBS-PLAID parameters, for BWA, producing non-distended, BWA or Bowtie2 post-imputation, post-error correction maps (data not shown). The greater room for imputation leniency within the PANATI dataset, however, underscores the importance and utility of using a species-appropriate alignment algorithm. PANATI was designed and programmed specifically to optimize alignments for species with levels of genetic diversity similar to those found in rice. The genetic map produced by the final PANATI dataset under the imputation parameters used in this study is evidence that it is better suited for GBS data alignment in rice than either BWA or Bowtie2, both of which were designed for low-diversity species such as humans.

While appropriately rigorous methods for addressing GBS data errors and sparsity were necessary to produce our final dataset, the results of our QTL analyses and our analysis of the genetic architecture of the RIL population using our final dataset strongly suggest that via our streamlined pipeline we were able to produce a high-quality dataset that adds great value to the IR64 × Azucena RIL mapping population. By saturating the population with 30,984 SNPs, we were able to define regions of segregation distortion down to 0.24 Mb—the recombinational limits in an RIL population of this kind, thus identifying regions of candidate sterility genes. The majority of these regions, including those on chromosomes 1, 3, 4, 6, 8, and 11, correspond to previously identified putative sterility loci, lending further validation to the value of our dataset for both mapping segregation distortion and identifying putative sterility loci (Harushima et al. 2001; Harushima et al. 2002; Wu et al. 2010; Xu et al. 1997; Garavito et al. 2010; Matsubara et al. 2011). The saturation of the population with markers also allowed us to map recombination hot and cold spots across the genome with a similar high degree of precision.

Furthermore, when the full 30,984 SNPs were used to re-map QTL for aluminum tolerance using previously published phenotype data (Famoso et al. 2011), two new QTL were discovered in a region of high recombination that went undetected when either the 200 SSRs were used by Famoso et al. or when the 1,464 SNP subset was used. Likewise, when the full set of 30,984 SNPs was used to map QTL for leaf width, four more QTL were identified than when the 1,464 subset was used. These results strongly indicate that fully saturating a mapping population with SNP markers can enhance the ability to detect QTL, particularly in regions of heightened recombination, and subsequently lower linkage disequilibrium, that, specifically, the large number of markers now available on the IR64 × Azucena RIL population should serve as a valuable genetic resource for the rice community.

Overall, our results suggest that GBS can help fill the genotyping gap between reference lines of broad general interest and non-reference lines of more specific interest by providing an inexpensive means of adding SNP markers to mapping and breeding populations. RIL populations such as the one explored here are particularly well suited to this new technology as line immortality means that genotyping is a one-time investment and results can be utilized for many years and by many research groups, to evaluate many traits or genetic characteristics. Just as importantly, the high degree of homozygosity in an RIL population simplifies the bioinformatic analysis and error correction, as it eliminates the difficulty of distinguishing heterozygotes from sequencing errors. While outside the scope of this paper, bioinformatics tools such as those contained in TASSEL also exist for the treatment of other types of bi-parental mapping populations, although we advise caution and careful quantification of error when using highly multiplexed GBS (i.e., 384-plex or greater) for larger, more complex genomes, or for populations where a significant degree of heterozygosity is expected, particularly if allele frequencies for some heterozygous classes are low. As demonstrated here, calculating a genetic map, when possible, is a good way to assess error contained within GBS datasets. Finally, we conclude by noting that the data sparsity and error inherent in raw GBS data requires a significant investment in bioinformatics that is often not factored into the low up-front cost of generating GBS data. New computational pipelines, such as the one described here, are being developed to address these problems.

Acknowledgments We thank Sharon Mitchell, Charlotte Acharya, and Wenyan Zhu with the Cornell Institute of Genomic Diversity for assistance with GBS library prep, Ed Buckler, Jeff Glaubitz, Rob Elshire, Peter Bradbury, and James Harriman at Cornell University for assistance and advice on GBS data analysis and using the TASSEL GBS pipeline, Gen Onishi for greenhouse support, Cheryl Utter for helping format the manuscript, Francisco Agosto-Perez, Genevieve DeClerck, and Chih-Wei Tung for bioinformatics support, and Mike Spindel for Python consulting and troubleshooting support.

References

Almeida GD, Makumbi D, Magorokosho C, Nair S, Borem A, Ribaut JM, Banziger M, Prasanna BM, Crossa J, Babu R (2013) QTL mapping in three tropical maize populations reveals a set of constitutive and adaptive genomic regions for drought tolerance. Theor Appl Genet 126(3):583–600. doi:10.1007/s00122-012-2003-7

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3(10):e3376. doi:10.1371/journal.pone.0003376

Bradbury PJ, Zhang Z, Dallas EK, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635

Chen MS, Presting G, Barbazuk WB, Goicoechea JL, Blackmon B, Fang FC, Kim H, Frisch D, Yu YS, Sun SH, Higingbottom S, Phimphilai J, Phimphilai D, Thurmond S, Gaudette B, Li P, Liu JD, Hatfield J, Main D, Farrar K, Henderson C, Barnett L, Costa R, Williams B, Walser S, Atkins M, Hall C, Budiman MA, Tomkins JP, Luo MZ, Bancroft I, Salse J, Regad F, Mohapatra T, Singh NK, Tyagi AK, Soderlund C, Dean RA, Wing RA (2002) An integrated physical and genetic map of the rice genome. Plant Cell 14(3):537–545. doi:10.1105/tpc.010485

Clark RT, MacCurdy RB, Jung JK, Shaff JE, McCouch SR, Aneshansley DJ, Kochian LV (2011) Three-dimensional root phenotyping with a novel imaging and software platform. Plant Physiol 156(2):455–465. doi:10.1104/pp.110.169102

Darvasi A, Weinreb A, Minke V, Weller JI, Soller M (1993) Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics 134(3):943–951

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12(7):499–510

Dupuis J, Siegmund D (1999) Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 151(1):373–386

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011a) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6(5):e19379. doi:10.1371/journal.pone.0019379

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011b) Powerpoint presentation: reduced representation sequencing for rapidly genotyping highly diverse species

Famoso AN, Zhao K, Clark RT, Tung C-W, Wright MH, Bustamante C, Kochian LV, McCouch SR (2011) Genetic architecture of aluminum tolerance in rice (*Oryza sativa*) determined through genome-wide association analysis and QTL mapping. PLoS Genet 7(8):e1002221. doi:10.1371/journal.pgen.1002221

Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH (2004) An SNP resource for rice genetics and breeding based on subspecies Indica and Japonica genome alignments. Genome Res 14(9):1812–1819. doi:10.1101/gr.2479404

Garavito A, Guyot R, Lozano J, Gavory F, Samain S, Panaud O, Tohme J, Ghesquière A, Lorieux M (2010) A genetic model for the female sterility barrier between Asian and African cultivated rice species. Genetics 185(4):1425–1440. doi:10.1534/genetics.110.116772

Harushima Y, Yano M, Shomura A, Sato M, Shimano T, Kuboki Y, Yamamoto T, Lin SY, Antonio BA, Parco A, Kajiya H, Huang N, Yamamoto K, Nagamura Y, Kurata N, Khush GS, Sasaki T (1998) A high-density rice genetic linkage map with 2275 markers using a single F2 population. Genetics 148(1):479–494

Harushima Y, Nakagahra M, Yano M, Sasaki T, Kurata N (2001) A genome-wide survey of reproductive barriers in an intraspecific hybrid. Genetics 159(2):883–892

Harushima Y, Nakagahra M, Yano M, Sasaki T, Kurata N (2002) Diverse variation of reproductive barriers in three intraspecific rice crosses. Genetics 160(1):313–322

Hemamalini GS, Shashidhar HE, Hittalmani S (2000) Molecular marker assisted tagging of morphological and physiological traits under two contrasting moisture regimes at peak vegetative stage in rice (*Oryza sativa* L.). Euphytica 112(1):69–78. doi:10.1023/a:1003854224905

Hittalmani S, Huang N, Courtois B, Venuprasad R, Shashidhar HE, Zhuang JY, Zheng KL, Liu GF, Wang GC, Sidhu JS, Srivantaneeyakul S, Singh VP, Bagali PG, Prasanna HC, McLaren G, Khush GS (2003) Identification of QTL for growth- and grain yield-related traits in rice across nine locations of Asia. Theor Appl Genet 107(4):679–690. doi:10.1007/s00122-003-1269-1

Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, Guan J, Fan D, Weng Q, Huang T, Dong G, Sang T, Han B (2009) High-throughput genotyping by whole-genome resequencing. Genome Res 19(6):1068–1076. doi:10.1101/gr.089516.108

Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, Guo Y, Lu Y, Zhou C, Fan D, Weng Q, Zhu C, Huang T, Zhang L, Wang Y, Feng L, Furuumi H, Kubo T, Miyabayashi T, Yuan X, Xu Q, Dong G, Zhan Q, Li C, Fujiyama A, Toyoda A, Lu T, Feng Q, Qian Q, Li J, Han B (2012) A map of rice genome variation reveals the origin of cultivated rice. Nature 490(7421):497–501. http://www.nature.com/nature/journal/v490/n7421/abs/nature11532.html (supplementary information)

Ilut DC, Coate JE, Luciano AK, Owens TG, May GD, Farmer A, Doyle JJ (2012) A comparative transcriptomic study of an allotetraploid and its diploid progenitors illustrates the unique advantages and challenges of RNA-seq in plant species. Am J Bot 99(2):383–396. doi:10.3732/ajb.1100312

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Meth 9(4):357–359. http://www.nature.com/nmeth/journal/v9/n4/abs/nmeth.1923.html (supplementary information)

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26(5):589–595. doi:10.1093/bioinformatics/btp698

Li ZK, Yu SB, Lafitte HR, Huang N, Courtois B, Hittalmani S, Vijayakumar CHM, Liu GF, Wang GC, Shashidhar HE, Zhuang JY, Zheng KL, Singh VP, Sidhu JS, Srivantaneeyakul S, Khush GS (2003) QTL $\sqrt{6}$ environment interactions in rice. I. Heading date and plant height. Theor Appl Genet 108(1):141–153. doi:10.1007/s00122-003-1401-2

Lincoln SE, Lander ES (1992) Systematic detection of errors in genetic linkage data. Genomics 14(3):604–610

Maheswaran M, Subudhi PK, Nandi S, Xu JC, Parco A, Yang DC, Huang N (1997) Polymorphism, distribution, and segregation of

AFLP markers in a doubled haploid rice population. Theor Appl Genet 94(1):39–45. doi:10.1007/s001220050379

Mangin B, Goffinet B, Rebai A (1994) Constructing confidence intervals for QTL location. Genetics 138(4):1301–1308

Matsubara K, Ebana K, Mizubayashi T, Itoh S, Ando T, Nonoue Y, Ono N, Shibaya T, Ogiso E, Hori K, Fukuoka S, Yano M (2011) Relationship between transmission ratio distortion and genetic divergence in intraspecific rice crosses. Mol Genet Genomics 286(5–6):307–319. doi:10.1007/s00438-011-0648-6

McCouch SR, Chen X, Panaud O, Temnykh S, Xu Y, Cho YG, Huang N, Ishii T, Blair M (1997) Microsatellite marker development, mapping and applications in rice genetics and breeding. Plant Mol Biol 35(1–2):89–99

Mei HW, Luo LJ, Ying CS, Wang YP, Yu XQ, Guo LB, Paterson AH, Li ZK (2003) Gene actions of QTLs affecting several agronomic traits resolved in a recombinant inbred rice population and two testcross populations. Theor Appl Genet 107(1):89–101. doi:10.1007/s00122-003-1192-5

Mei HW, Li ZK, Shu QY, Guo LB, Wang YP, Yu XQ, Ying CS, Luo LJ (2005) Gene actions of QTLs affecting several agronomic traits resolved in a recombinant inbred rice population and two backcross populations. Theor Appl Genet 110(4):649–659. doi:10.1007/s00122-004-1890-7

Nasu S, Suzuki J, Ohta R, Hasegawa K, Yui R, Kitazawa N, Monna L, Minobe Y (2002) Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (*Oryza sativa*, *Oryza rufipogon*) and establishment of SNP markers. DNA Res 9(5):163–171. doi:10.1093/dnares/9.5.163

Prasad SR, Bagali PG, Hittalmani S, Shashidhar HE (2000) Molecular mapping of quantitative trait loci associated with seedling tolerance to salt stress in rice (*Oryza sativa* L.). Curr Sci 78(2):162–164

Rosyara UR, Gonzalez-Hernandez JL, Glover KD, Gedye KR, Stein JM (2009) Family-based mapping of quantitative trait loci in plant breeding populations with resistance to Fusarium head blight in wheat as an illustration. Theor Appl Genet 118(8):1617–1631. doi:10.1007/s00122-009-1010-9

Sallaud C, Lorieux M, Roumen E, Tharreau D, Berruyer R, Svestasrani P, Garsmeur O, Ghesquiere A, Notteghem JL (2003) Identification of five new blast resistance genes in the highly blast-resistant rice variety IR64 using a QTL mapping strategy. Theor Appl Genet 106(5):794–803. doi:10.1007/s00122-002-1088-9

Stangoulis JR, Huynh B-L, Welch R, Choi E-Y, Graham R (2007) Quantitative trait loci for phytate in rice grain and their relationship with grain micronutrient content. Euphytica 154(3):289–294. doi:10.1007/s10681-006-9211-7

This D, Comstock J, Courtois B, Xu YB, Ahmadi N, Vonhof WM, Fleet C, Setter T, McCouch S (2010) Genetic analysis of water use efficiency in rice (*Oryza sativa* L.) at the leaf level. Rice 3(1):72–86. doi:10.1007/s12284-010-9036-9

Thomson M, Zhao K, Wright M, McNally K, Rey J, Tung C-W, Reynolds A, Scheffler B, Eizenga G, McClung A, Kim H, Ismail A, de Ocampo M, Mojica C, Reveche M, Dilla-Ermita C, Mauleon R, Leung H, Bustamante C, McCouch S (2012) High-throughput single nucleotide polymorphism genotyping for breeding applications in rice using the BeadXpress platform. Mol Breed 29(4):875–886. doi:10.1007/s11032-011-9663-x

Virk PS, Ford-Lloyd BV, Newbury HJ (1998) Mapping AFLP markers associated with subspecific differentiation of *Oryza sativa* (rice) and an investigation of segregation distortion. Heredity 81:613–620. doi:10.1046/j.1365-2540.1998.00441.x

Vos P, Hogers R, Bleeker M, Reijans M, Vandelee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP—a new technique for DNA-fingerprinting. Nucleic Acids Res 23(21):4407–4414. doi:10.1093/nar/23.21.4407

Walsh MLaB (1998) Genetics and analysis of quantitative traits. Sinauer Associates Inc., Sunderland

Wang L, Hao L, Li X, Hu S, Ge S, Yu J (2009) SNP deserts of Asian cultivated rice: genomic regions under domestication. J Evol Biol 22(4):751–761. doi:10.1111/j.1420-9101.2009.01698.x

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A (2004) Diversity arrays technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci USA 101(26):9915–9920. doi:10.1073/pnas.0401076101

Wu JZ, Mizuno H, Hayashi-Tsugane M, Ito Y, Chiden Y, Fujisawa M, Katagiri S, Saji S, Yoshiki S, Karasawa W, Yoshihara R, Hayashi A, Kobayashi H, Ito K, Hamada M, Okamoto M, Ikeno M, Ichikawa Y, Katayose Y, Yano M, Matsumoto T, Sasaki T (2003) Physical maps and recombination frequency of six rice chromosomes. Plant J 36(5):720–730. doi:10.1046/j.1365-313X.2003.01903.x

Wu YP, Ko PY, Lee WC, Wei FJ, Kuo SC, Ho SW, Hour AL, Hsing YI, Lin YR (2010) Comparative analyses of linkage maps and segregation distortion of two F-2 populations derived from japonica crossed with indica rice. Hereditas 147(5):225–236. doi:10.1111/j.1601-5223.2010.02120.x

Xu Y, Zhu L, Xiao J, Huang N, McCouch SR (1997) Chromosomal regions associated with segregation distortion of molecular markers in F-2, backcross, doubled haploid, and recombinant inbred populations in rice (*Oryza sativa* L.). Mol Gen Genet 253(5):535–545

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M, McCouch S, Nielsen R, Wang W (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30(1):105–111. doi:10.1038/nbt.2050

Yan CJ, Liang GH, Chen F, Li X, Tang SZ, Yi CD, Tian S, Lu JF, Gu MH (2003) Mapping quantitative trait loci associated with rice grain shape based on an indica/japonica backcross population. Yi Chuan Xue Bao 30(8):711–716

Zhao Q, Zhang Y, Cheng ZK, Chen MS, Wang SY, Feng Q, Huang YC, Li Y, Tang YS, Zhou B, Chen ZH, Yu SL, Zhu JJ, Hu X, Mu J, Ying K, Hao P, Zhang L, Lu YQ, Zhang LS, Liu YL, Yu Z, Fan DL, Weng QJ, Chen L, Lu TT, Liu XH, Jia PX, Sun TG, Wu YR, Zhang YJ, Lu Y, Li C, Wang R, Lei HY, Li T, Hu H, Wu M, Zhang RQ, Guan JP, Zhu J, Fu G, Gu MH, Hong GF, Xue YB, Wing R, Jiang JM, Han B (2002) A fine physical map of the rice chromosome 4. Genome Res 12(5):817–823. doi:10.1101/gr.48902