Published November, 2011

ORIGINAL RESEARCH

Genetic Diversity Analysis with 454 Pyrosequencing and Genomic Reduction Confirmed the Eastern and Western Division in the Cultivated Barley Gene Pool

Yong-Bi Fu* and Gregory W. Peterson

Abstract

Next-generation DNA sequencing (NGS) technologies can survey sequence variation on a genome-wide scale, but their utility for crop genetic diversity analysis is poorly known. Many challenges remain in their applications, including sampling complex genomes, identifying single nucleotide polymorphisms (SNPs), and analyzing missing data. This study presented a practical application of the Roche 454 GS FLX Titanium technology in combination with genomic reduction and an advanced bioinformatics tool to analyze the genetic relationships of 16 diverse barley (*Hordeum vulgare* L.) landraces. A full 454 run generated roughly 1.7 million sequence reads with a total length of 612 Mbp. Application of the computational pipeline called DIAL (de novo identification of alleles) identified 2578 contigs and 3980 SNPs. Sanger sequencing of four barley samples confirmed 85 of the 100 selected contigs and 288 of the 620 putative SNPs and identified 735 new SNPs and 39 new indels. Several diversity analyses revealed the eastern and western division in the barley samples. The division is compatible with those inferred with 156 microsatellite alleles of the same 16 samples and consistent with our current knowledge about cultivated barley. These results help to illustrate the utility of NGS technologies for crop diversity studies. The NGS application also provides a new informative set of genomic resources for barley research.

NEXT-GENERATION DNA sequencing (NGS) technologies can generate unprecedented amounts of genomic data, even in non-model organisms (Nordborg and Weigel, 2008; Bräutigam and Gowik, 2010; Metzker, 2010; Seeb et al., 2011). In recent years, these technologies have become an affordable means to survey sequence variation on a genome-wide scale (Seeb et al., 2011; You et al., 2011). In principle, such capability should revolutionize the genetic diversity analysis with high resolution on crop plants with large and complex genomes. However, applications of NGS technologies to assess crop genetic diversity are still full of challenges and feasibility assessments are warranted to inform diversity analysis. Many plants, unlike those model plants such as rice (*Oryza sativa* L.) and maize (*Zea mays* L.) with sequenced genomes, have large and complex genomes with variable ploidy and an abundance of repeated sequences (Wicker et al., 2006; Novaes et al., 2008; You et al., 2011). Also, limitations are not lacking in the application of bioinformatics tools to identify single nucleotide polymorphisms (SNPs) without a reference genome (Imelfort et al., 2009; You et al., 2011). Moreover, a diversity analysis of NGS genomic data with sequencing error, assembly error, and missing data may not be always informative (Rokas and Abbot, 2009; Pool et al., 2010).

Plant Gene Resources of Canada, Saskatoon Research Centre, Agriculture and Agri-Food Canada, 107 Science Pl., Saskatoon, SK S7N 0X2, Canada. Received 1 Aug. 2011. *Corresponding author (yong-bi.fu@agr.gc.ca).

Abbreviations: AMOVA, analysis of molecular variance; DIAL, de novo identification of alleles; dNTP, deoxyribonucleotide triphosphate; EDTA, ethylenediaminetetraacetic acid; MCC, maximum clade credibility; NCBI, National Center for Biotechnology Information; NGS, next-generation DNA sequencing; PCA, principle component analysis; PCR, polymerase chain reaction; RLMID, Rapid Library Multiplex Identifier; RRL, reduced representation library; SFF, standard flowgram format; SNP, single nucleotide polymorphism; SSR, simple sequence repeat; T_a, annealing temperature.

Published in The Plant Genome 4:226–237. Published 2 Nov. 2011.
doi: 10.3835/plantgenome2011.08.0022

5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication

Genomic reduction, also known as the reduced representation library (RRL), is one of the widely applied strategies to reduce the complexity of large genomes. The RRLs are constructed with a restriction digest followed by size selection and allow for sampling diverse but identical genomic regions from several individuals (Altshuler et al., 2000). The RRLs have been used for resequencing, sequence alignment, assembly, and SNP discovery (Altshuler et al., 2000; Barbazuk et al., 2007; Van Tassell et al., 2008; Wiedmann et al., 2008; Gore et al., 2009; Deschamps et al., 2010). Specifically, applications of RRLs for SNP discovery have been made in some non-model plants without sequenced genomes such as amaranth (*Amaranthus caudatus* L.) (Maughan et al., 2009) and common bean (*Phaseolus vulgaris* L.) (Hyten et al., 2010). These applications not only generated large volumes of SNPs in studied plants with scarce genomic resource but also revealed high validation rates of putative SNPs (e.g., 86% in common bean, 97% in amaranth).

Technically, it is more difficult to identify SNPs for a species without a sequenced genome (Cannon et al., 2010; Ratan et al., 2010; Peterlongo et al., 2010). The common practice is to create a consensus sequence using sequence reads from one individual, and these can then be used as a reference sequence to call SNPs in other individuals (Li et al., 2008). This approach may include multiple bioinformatics tools that require considerable knowledge of bioinformatics to operate (Flicek and Birney, 2009). Some integrated tools are sophisticated and the user-friendly ones are largely commercialized (Imelfort et al., 2009). Also, bias could be introduced from developing a consensus sequence from multiple individuals with unequal read coverage (Ratan et al., 2010). Fortunately, alternative tools are available such as the free computational pipeline DIAL (de novo identification of alleles) (Ratan et al., 2010). This tool does not require a reference sequence, masks repetitive sequences, clusters sequence reads from multiple individuals, and performs assembly for clusters with a de novo assembler to identify variants. More importantly, it is well suited for the clustering of sequence reads of multiple individuals from Roche 454 (454 Life Sciences, Branford, CT) and Illumina GA (Illumina Inc. San Diego, CA) platforms, even with a shallow depth of genome coverage and variable read coverage of multiple samples. Interestingly, this promising tool has been tested only in moderately complex genomes and does not handle SNP calls in repeats (You et al., 2011).

Next-generation DNA sequencing genomic data are known to be full of sequencing errors, assembly errors, and missing data (Pool et al., 2010). Errors in the sequences could arise from DNA damage, polymerase chain reaction (PCR) amplification, and sequencing (Rokas and Abbot, 2009). Assembly of short sequence reads with low coverage may not be always accurate and could be more challenging in the repetitive or highly polymorphic genomic regions (e.g., Li et al., 2008). Next-generation DNA sequencing data are typically unbalanced for each sample, due to the stochastic sampling across the genome, and such unbalance will increase with lower coverage of sequence reads. Efforts have been made to minimize the sequencing and assembly errors (e.g., Jiang et al., 2009; Long et al., 2009), but it is difficult to minimize missing data, given more samples are required for an informative diversity analysis with limited sequencing resources (Luca et al., 2011). All of these issues can add uncertainty to crop genetic diversity analyses (Liu et al., 2009).

Barley (*Hordeum vulgare* L.; $2n = 2x = 14$) is an important crop and a diploid model for classical genetics (Harlan, 1976; van Bothmer et al., 2003). Thus, its genomic resources such as linkage map (e.g., Close et al., 2009), expressed sequence tag sequence (Druka et al., 2006), and gene map (e.g., Sato et al., 2009) are not lacking. However, it has a large (5.1 Gbp) and complex genome with over 80% highly repetitive DNA sequences (Doležel et al., 1998; Bennett and Smith, 1976). These genomic features pose a significant challenge to sequencing the barley genome (Mayer et al., 2011) and practical difficulty in applications of NGS technology to survey sequence variation (e.g., see Wicker et al., 2006, 2009). Barley has been intensively studied using advanced molecular markers on its genetic diversity and gene pool (van Bothmer et al., 2003; Malysheva-Otto et al., 2006). A major division of its gene pool following its wild progenitor along the Zagros Mountains has been long recognized (e.g., see Takahashi, 1955; Zohary and Hopf, 2000; Saisho and Purugganan, 2007) and is thought to largely reflect the consequence of multiple barley domestications (Morrell and Clegg, 2007). However, the previous inferences of the gene pool structure may have suffered from limited sampling of the barley genome (Saisho and Purugganan, 2007). These unique features would allow for a better assessment on the informativeness of some NGS technologies in sampling complex crop genomes.

In this study, we specifically explored the utility of high-throughput NGS techniques and bioinformatics tools for crop genetic diversity studies and performed a practical application of the Roche 454 GS FLX Titanium technology (454 Life Sciences) in combination with genomic reduction and an advanced bioinformatics tool to analyze the genetic relationships of 16 diverse barley landraces. The specific objectives of the application were to identify contigs and SNPs from 16 diverse barley landraces using 454 pyrosequencing via genomic reduction and the computational pipeline DIAL (Ratan et al., 2010), to validate a subset of identified contigs and SNPs with Sanger sequencing, to infer the genetic relationships of the assayed landraces based on 454 SNP data, and to assess the deviation of the inferred genetic relationships from those obtained with barley simple sequence repeat (SSR) data.

Materials and Methods

Plant Materials and DNA Extraction

Sixteen genetically diverse barley landrace accessions originating from 16 countries (Table 1) were selected for this study. The selected accessions represent the

eastern and western samples of the barley landrace gene pool, based on the distinct distribution of wild barley described by Zohary and Hopf (2000) and Morrell and Clegg (2007). The eastern region represents cultivated barley from the Zagros Mountains and further east, while the western region includes cultivated barley from the Fertile Crescent and further west. A few seeds were randomly chosen from each selected accession maintained at the Plant Gene Resources of Canada (Saskatoon, SK, Canada). Plants were grown from seed for 2 to 3 wk in a greenhouse at the Saskatoon Research Centre, Agriculture and Agri-Food Canada. Young leaf tissue from individual plants of each accession was collected, freeze dried, and stored at −20°C. Deoxyribonucleic acid was extracted from 15 mg of freeze-dried tissue using the DNEasy Plant Mini kit (Qiagen, Mississauga, ON, Canada) following the manufacturer's instructions, quantified using the Thermo Scientific Nanodrop 8000 spectrometer (Fisher Scientific Canada, Toronto, ON, Canada), and adjusted using Qiagen AE buffer (10 mM Tris-HCl and 0.5 mM ethylenediaminetetraacetic acid [EDTA], pH 9.0) to 25 ng μL^{-1} for SSR analysis and 100 ng μL^{-1} for 454 pyrosequencing.

Genome Reduction and Barcoding

Genomic reduction and multiplex identifiers barcoding of the barley samples were conducted following the method of Maughan et al. (2009) using the same sourced reagents and supplies where possible. *Eco*RI and BfaI adaptors and barcoded PCR primers were synthesized by Integrated DNA Technologies (Coralville, IA) using the Roche 454 Sequencing Rapid Library Multiplex Identifier (RLMID) barcode sequences (Roche 454 Sequencing, 2010). The 16 samples were randomly divided into two pools (Table 1). All samples were digested with *Eco*RI and BfaI. BfaI- and biotin-modified *Eco*RI adaptors were ligated onto the digested fragments. The ligation reactions were cleaned using the Chroma Spin +TE-400 columns (Clontech, Mountain View, CA) following the manufacturer's instructions. Fragments with the biotin-modified EcoR1 adaptor were selected using streptavidin coated paramagnetic beads (Dynabeads M-280; Invitrogen, Burlington, ON, Canada) according to the manufacturer's instructions.

Seven barcodes were shared between the two pools and one unique barcode was used in each pool to differentiate between the two pools and identify any potential cross-contamination between pools. Paramagnetic beads with bound digested DNA fragments were used as a template for PCR using primers specific to the *Eco*RI and BfaI adaptors and containing a specific RLMID barcode for each sample in each pool. The PCR method was followed from Maughan et al. (2009) using the Clontech HF2 chemistry and the C1000 thermocycler (BioRad, Mississauga, ON, Canada). Four replicates of each PCR reaction were performed and a 3-µL sample from each was separated on a 1.5% agarose gel to confirm amplification. Amplifications for each sample were bulked together and concentrated by evaporation in a vacuum centrifuge to approximately 35 µL. Individual samples were separated on a 1.5% agarose gel for 5 h at 60 V. A gel fragment from each sample between 400 and 600 bp based on the New England Biolabs 2-Log ladder (Pickering, ON, Canada) was excised and cleaned using the QIAquick Gel Extraction kit (Qiagen). Samples were eluted in 35 µL of one-third concentration Qiagen EB (3.33 mM Tris, pH 8.5) and quantified using the Thermo Scientific Nanodrop 8000 spectrometer (Fisher Scientific Canada, Toronto, ON, Canada). Individual samples were concentrated by evaporation using a vacuum centrifuge, requantified, and adjusted to 50 ng μL^{-1} with water and 1 mM EDTA pH 8.0 so that the final salt concentration did not exceed 10 mM Tris and 1 mM EDTA. Each pool was prepared consisting of 200 ng of each of eight individual accessions for a total of 1600 ng at 50 ng μL^{-1}.

Pools were submitted to the DNA Technologies Laboratory at the Canadian National Research Council's Plant Biotechnology Institute (Saskatoon, SK, Canada) and sequenced using the Roche 454 GS FLX instrument with Titanium chemistry. All sequences were deposited in the National Center for Biotechnology Information (NCBI) Short Read Archive (NCBI, 2011) under accession number SRA045777.

Generation of Clusters and Single Nucleotide Polymorphisms

Deoxyribonucleic acid reads were separated into sample-specific SFF (standard flowgram format) files according to RLMID barcode using the Roche Newbler SFF tools (454 Life Sciences, 2010) followed by a removal of the forward and reverse adaptor sequences. Cluster generation and SNP detection were performed using the DIAL (Ratan et al., 2010) pipeline. The pipeline adds the SFF file of each sample and performs a completely automatic call of SNPs from all added SFF files in a Linux system. However, it requires the inputs on the expected length of target genome to identify clusters from all added SFF files and the version of Roche Newbler, as it is dependent on the Newbler's gsAssembler to assemble the reads into the identified clusters. Thus, a training of DIAL was made for different versions of Newbler and variable lengths of target genome from 10 to 0.5 Mbp. The final analysis was made using the Newbler v2.0.01.14 and an expected genome size of 3 Mbp to generate the numbers of clusters and SNPs as large as possible for the 16 samples. As highly stringent filters are applied for SNP calling, the pipeline usually generated an unrealistically low yield of two to three SNPs in the output file *snps.txt*. However, the pipeline also generated an output file *report.txt* collecting all the assembled contigs with sequence length and supporting reads, the position of the variant alleles, the number of reads supporting the allele, and the quality value of the reads at that position. Thus, several specific Perl scripts were written to extract contigs and SNPs from *report.txt* into separate files for validation and for data report and analysis, and these custom-built Perl scripts are available on request.

Table 1. List of 16 barley accessions representing the western and eastern divisions of the barley gene pool, 454 pyrosequencing information, and identified single nucleotide polymorphisms (SNPs).

CN†	Origin	Region‡	Label	454 pool	RLMID§ sequence (5′ to 3′)	No. of reads	NC¶	NSP¶	Het%¶	MisSNP%¶
91147	China	E	ChinaE	1a	ACACGACGAC	45,239	191	420	1.9	89.4
81387	India	E	IndiaE	1b	ACACGTAGTA	145,057	930	1,834	2.0	53.9
59909	Iran	E	IranE	1c	ACACTACTCG	142,437	986	1,862	1.8	53.2
92006	Japan	E	JapanE	1d	ACGACACGTA	54,999	395	750	1.3	81.1
58854	Mongolia	E	MongoliaE	1e	ACGAGTAGAC	81,011	621	1,242	1.0	68.8
81307	Nepal	E	NepalE	1f	ACTATACGAG	199,860	1316	2,593	2.8	34.8
80536	Pakistan	E	PakistanE	1g	ACGTACACAC	70,910	604	1,193	0.3	70.0
29118	Kyrgyzstan	E	KyrgyzstanE	1h	ACGTACTGTG	22,540	95	180	0.0	95.5
68637	Turkey	W	TurkeyW	2a	ACTACGTCTC	95,021	567	1,099	0.8	72.4
51205	Syria	W	SyriaW	2b	ACACGTAGTA	182,678	1010	2,079	2.6	47.7
50729	Jordan	W	JordanW	2c	ACACTACTCG	180,606	1055	2,129	3.7	46.5
29233	Iraq	W	IraqW	2d	ACGACACGTA	41,578	197	410	0.5	89.7
77017	Ethiopia	W	EthiopiaW	2e	ACGAGTAGAC	65,108	461	899	0.8	77.4
58689	Egypt	W	EgyptW	2f	ACTATACGAG	108,165	721	1,429	2.0	64.1
68925	Greece	W	GreeceW	2g	ACGTACACAC	36,607	161	295	0.0	92.6
94239	Lebanon	W	LebanonW	2h	ACGTACTGTG	32,411	121	225	9.3	94.3
Average for eastern region						95,257	642	1,259	1.4	68.3
Average for western region						92,772	537	1,071	2.5	73.1

†CN, Canadian National accession number at the Plant Gene Resources of Canada, Saskatoon, SK, Canada.

‡The cultivated barley gene pool was divided based on the distinct distribution of wild barley following Zohary and Hopf (2000) and Morrell and Clegg (2007). The eastern (E) region represents cultivated barley from the Zagros Mountains and further east, while the western (W) region includes cultivated barley from the Fertile Crescent and further west.

§RLMID, Rapid Library Multiplex Identifier.

¶NC, the number of contigs with SNP; NSP, the number of SNPs predicted; Het%, the percentage of the heterozygous SNPs over the predicted SNPs; MisSNP%, the percentage of the predicted SNPs that were missing for the sample due to the lack of sequence reads.

Contig and Single Nucleotide Polymorphism Validation

One hundred contigs with variable SNP count and sequence length were selected for validation with Sanger sequencing based on four randomly selected samples (ChinaE, PakistanE, JordanW, and EthiopiaW). The PCR primers for the selected 100 contigs were designed using the Primer3 v0.4.0 online tool (Rozen and Skaletsky, 2000). The conditions for PCR were 1x KAPA 2G Buffer A containing 1.5 mM $MgCl_2$ (KAPA Biosystems, Woburn, MA), 1x KAPA Enhancer 1, 0.2 mM each deoxyribonucleotide triphosphate (dNTP), 0.4 pmol μL^{-1} each forward and reverse primers, 100 ng of the same genomic DNA template samples as used above for next generation sequencing, and 0.5 U KAPA 2G Robust polymerase in a final volume of 25 µL; touchdown PCR was cycled at 95°C for 3 min followed by 10 cycles of 95°C for 10 s, 60°C decreasing 0.5°C per cycle for 15 s, 72°C for 30 s, and then 25 cycles of 95°C for 10 s, 55°C for 15 s, and 72°C for 20 s with a final extension of 72°C for 30 s. A 3-µL sample of each PCR product was separated on 1.5% agarose for 2 h at 120 V. For the primer set generating single fragment of expected size, its PCR product was cleaned following the method outlined by Rosenthal et al. (1993) and submitted for Sanger sequencing at the DNA Technologies Laboratory, Canadian National Research Council's Plant Biotechnology Institute (Saskatoon, SK, Canada). Polymerase chain reaction samples with multiple bands were first run on a 2% agarose gel for 17 h at 50 V. The band of expected size based on the contig consensus sequence was then cut from the gel and spun for 5 min at 16,000 × *g*, and 1 µL of the resulting liquid was used as a template for PCR re-amplification. The resulting PCR product was checked on agarose and cleaned for sequencing as outlined above for the single band samples. An effort was also made to redesign PCR primers for questionable contigs and then rescreen them.

Forward and reverse Sanger sequences from each sample were assembled using Sequencher v.4.10.1 (GeneCodes Corporation, 2010), aligned using MUSCLE v.3.6 (Edgar, 2004) against the consensus sequence generated via NGS for each contig and proofread by hand. All working primer sets generated the Sanger sequences that were successfully aligned with the contig consensus sequences. The putative SNPs identified via NGS were checked with the Sanger sequences, where sample data was available, and additional SNPs and indels from the Sanger sequencing were also identified, if any. Blasting some set of identified contigs was also made in ACPFG Bioinformatics barley autoSNPdb database v1.4 (ACPFG Bioinformatics, 2011) and the NCBI *Hordeum vulgare* subsp. *vulgare* nucleotide databases (NCBI, 2011) using megablast with E-value of at least 1×10^{-5}.

Simple Sequence Repeat Analysis

Thirty informative barley SSR primers were selected based on the published marker information and genomic coverage from the consensus map (Varshney et al., 2007)

(Supplemental Table S1). All selected primers were synthesized by Integrated DNA Technologies and resuspended in water at 50 pmol μL^{-1}. The PCR reactions were set up as 1x New England Biolabs (Pickering, ON) Standard Buffer containing 1.5 mM $MgCl_2$, 0.2 mM each dNTP (Promega/Fisher Scientific, Nepean, ON, Canada), 0.4 pmol μL^{-1} each forward and reverse primer, 0.5 U *Taq* polymerase (New England Biolabs), and 50 ng template genomic DNA in a final volume of 25 µL. Reaction conditions were as follows: 94°C for 3 min, five touchdown cycles of 94°C for 10 s, 55°C decreasing 1°C per cycle for 20 s, and 68°C for 1 min followed by 25 cycles of 94°C for 10 s, 50°C for 20 s, and 68°C for 1 min with a final extension of 72°C for 5 min. Several primers (HVM62, GBM1323, HVM03, Bmac0018, GBM1419, WCM1E8, and GBM1405) were run with a touchdown PCR starting with an annealing temperature (T_a) of 60°C decreasing 1°C per cycle followed by 25 cycles at 55°C. All T_a's were determined based on the melting temperatures (T_m's) provided by Integrated DNA Technologies. For each amplified sample, 5 µL of glycerol loading buffer III (Sambrook et al., 1989) were added. Polyacrylamide (5%, 19:1) Mega Gels (CBS Scientific, Del Mar, CA) were prepared according to Wang et al. (2003). Gels were pre-run for 1 h with 0.5 µg l^{-1} ethidium bromide in the bottom reservoir and then 15 µL of each sample were loaded and run at 300 V for 2 to 3 h depending on the predicted microsatellite size. Invitrogen 50 bp and New England Biolabs 10 bp ladders were included as size standards. Gels were photographed digitally with ultraviolet light. Deoxyribonucleic acid fragments amplified by SSR primer pairs were manually scored based on their sizes in base pairs measured with DNA ladders and compared with the sizes reported in the literature.

Diversity Analysis

The resulting 454-SNP data were analyzed for each sample by counting the total number of the putative SNPs, the number of the putative heterozygous SNPs, and the number of the putative SNPs that were undetected in the sample due to insufficient sequence reads. An analysis of molecular variance (AMOVA) was performed using Arlequin version 3.01 (Excoffier et al., 2005) to assess genetic variation within and between the eastern and western barley samples. The significance of variance components was tested with 10,010 random permutations.

The genetic relationships of the 16 barley samples were analyzed with three commonly applied approaches. First, the principal component analysis was performed using NTSYS-PC 2.01 (Rohlf, 1997) based on the dissimilarity matrix of the available putative SNPs. Plots of the first three resulting principal components were made to assess the accession associations. Second, a distance-based NeighborNet (Bryant and Moulton, 2004) of the 16 samples was generated using the SplitsTree4 (Huson and Bryant, 2006) with the options of Uncorrected_P and EqualAngle. This tree displayed detailed reticulations where recombination may occur. Third, the maximum clade credibility (MCC) phylogenies were generated using BEAST v1.4 (Drummond and Rambaut, 2007) with a relaxed uncorrelated lognormal clock and with tree prior as constant size, expansion, or exponential growth. The substitution model was under a Hasegawa, Kishino, and Yano (HYK) model with γ distribution for site heterogeneity. The rest of the options were applied with default values. The Bayesian Markov chain Monte Carlo approach applied in BEAST should yield more informative phylogeny, as it directly calculates ultrametric phylogenies based only on observed data and model parameters and incorporates both the branch length errors and the topological uncertainties (Rutschmann, 2006).

The optimal genetic structure of the 16 samples was also inferred with two model-based Bayesian methods available in the BAPS software (Corander et al., 2008) and the program STRUCTURE version 2.2.3 (Pritchard et al., 2000; Falush et al., 2007). For BAPS, individual samples were clustered using the model for nonlinked markers and 20 replicate runs of the algorithm with the upper-bound values (K) for the number of clusters ranging between 2 and 16. The STRUCTURE program was run 15 times for each subpopulation (K) value, ranging from 2 to 10, using the admixture model with 10,000 replicates for burn-in and 10,000 replicates during analysis.

The SSR data were analyzed first for polymorphism with respect to primer and sample, then using AMOVA to quantify the known genetic structure, and last following the same approaches described above for 454-SNP data to assess the genetic relationships of the barley samples.

Results

454 Pyrosequencing

Following the workflow outlined in Fig. 1, this study generated 1,729,435 passed reads with about 612 Mbp of DNA sequence from a full Roche 454 GS FLX Titanium run of the 16 barley samples (Table 2). Two regions of the 454 run generated compatible reads, but the mean of a read length was slightly longer in the region 2 (358 bp) than that in the region 1 (351 bp) (Supplemental Fig. S1). Generally, the read length ranged from 40 to 1195 bp and averaged 354 bp. The number of reads per sample ranged from 22,540 for the KyrgyzstanE sample to 199,860 for the NepalE sample and averaged 94,014 (Table 1).

Training DIAL (Ratan et al., 2010) with the original Newbler v2.0.01.14 (454 Life Sciences, 2010) revealed that the numbers of identified contigs and SNPs increased with a target genome size decreasing from 10 Mbp and reached a maximum number of 2578 contigs and 3980 SNPs with the target genome size of 3 Mbp or smaller. For example, with 10 Mbp, 2547 contigs and 3889 SNPs were identified, and with 5 Mbp, 2573 contigs and 3969 SNPs were found. Training with the newest Newbler version (454 Life Sciences, 2010) showed that the maximum numbers of both contigs and SNPs obtained become much smaller. For example, the training with the newest version revealed only 2116 contigs and 3384 SNPs at the target genome size of 3 Mbp.

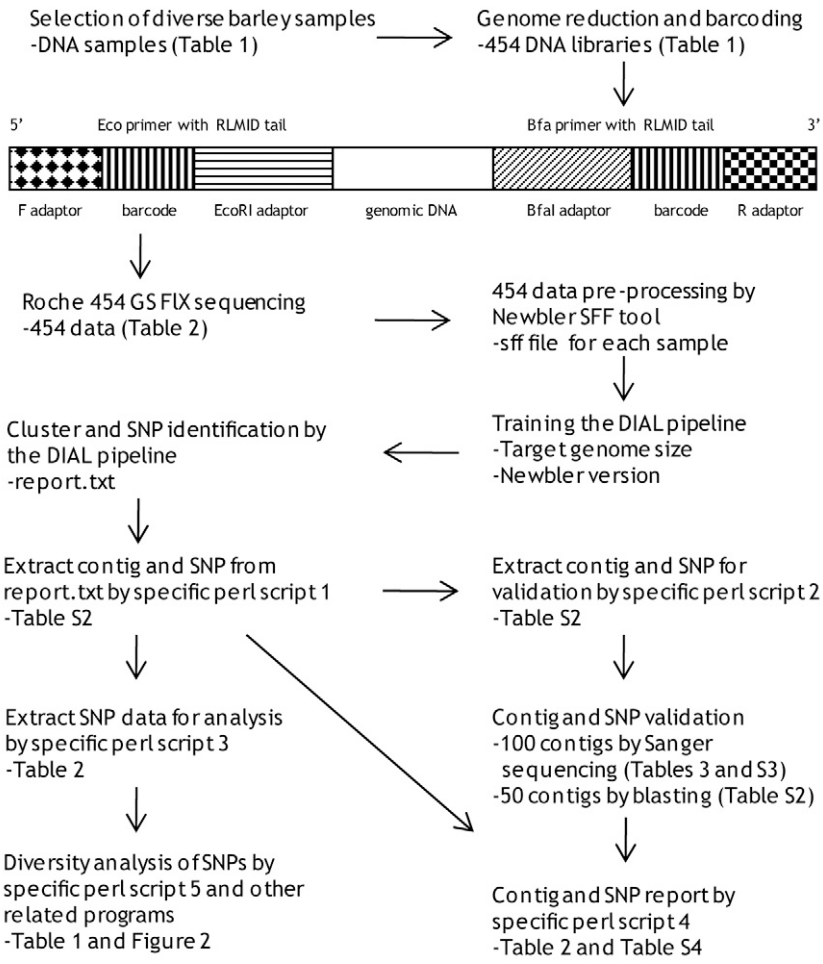

Figure 1. Workflow for the application of Roche 454 pyrosequencing in combination with genomic reduction and the computational pipeline DIAL (de novo identification of alleles) (Ratan et al., 2010) to identify contigs and single nucleotide polymorphisms (SNPs) from 16 barley samples. F, forward; R, reverse; RLMID, Rapid Library Multiplex Identifier; SFF, standard flowgram format.

Table 2. Summary statistics of 454 pyrosequencing reads and identified contigs.

	Read		Length				
Parameter	Raw	Passed	Total	Mean	Median	Minimum	Maximum
454 read and length							
Region1	1,097,247	862,828	302,653,823	350.8	387	40	703
Region2	1,141,029	866,607	310,078,946	357.8	389	40	1,195
All	2,238,276	1,729,435	612,732,769	354.3	388	40	1,195
2578 contigs							
Read			22,588	8.7	8	1	34
Length			1,057,077	410.0	423	97	540
SNP† for 2578 contigs			3,980	1.5	1	0	21
SNP for 1914 contigs with SNP			3,980	2.1	1	1	21
100 contigs for validation							
Read			993	9.9	9	5	24
Length			41,614	416.1	420.5	197	503
SNP predicted			701	7	7	0	21

†SNP, single nucleotide polymorphism.

A total of 2578 contigs were detected from 22,588 (1.31%) passed reads (Table 2; Supplemental Table S2). Effectively, only 41% of the sequence bases generated were used to identify the contigs and 0.173% bases contributed to the 2578 contigs identified. The number of reads per contig ranged from 1 to 34 and averaged 8.7. One contig had one read, 25 contigs had two reads, and accumulative 61 (2.4%) contigs were weakly supported

with only one to four reads. The contig length ranged from 97 to 540 bp with an average of 410 bp and with a median of 423 bp. There were 664 contigs without any putative SNPs and 1914 contigs with up to 21 SNPs. A total of 3980 putative SNPs on the 1914 contigs were identified from the 16 samples. For those 1914 contigs with SNPs, the average and median numbers of putative SNPs per contig were 2.1 and 1, respectively.

Contig and Single Nucleotide Polymorphism Validation

One hundred (3.9%) identified contigs with variable SNPs were selected for validation via Sanger sequencing (Table 2). These selected contigs were derived from 993 passed reads with a total length of 41,614 bp. The average and median lengths of the selected contigs were 416 bp and 221 bp, respectively. A total of 701 putative SNPs were predicted on the selected contigs. One hundred primer sets based on the contig consensus sequences were designed for the Sanger sequencing, and related primer information, including their performance in contig and SNP validation, is given in Supplemental Table S3. Seventy primer sets worked in both directions, 15 worked only in one direction, and 15 amplified no fragments, single fragments with failed Sanger sequencing, or multiple DNA fragments without Sanger sequencing (Supplemental Table S3).

The 100 primer sets positively confirmed 85 (85%) contigs but cannot completely exclude the other 15 contigs with 9.7 average sequence reads (Table 3; Supplemental Table S3). The Sanger sequencing of the four barley samples by 85 primer sets positively confirmed 288 of the 620 putative SNPs identified from DIAL (Ratan et al., 2010) and cannot completely invalidate the other 332 putative SNPs (Supplemental Table S4). Among those 332 SNPs, 49 on 29 contigs resided outside of the flanking primers and 48 on 13 contigs cannot be confirmed due to the lack of Sanger sequences for some samples. Thus, the effective SNP validation rate with the four samples was only 55.1%. The putative SNP bases matched between 454 reads and Sanger sequences available for a sample on the 85 validated contigs ranged from 74.4 to 95.7% and the overall SNP base match was 84.6% (Table 3).

Interestingly, the Sanger sequencing by 85 primer sets revealed 735 new SNPs and 39 indels with lengths ranging from 1 to 26 bp (Supplemental Table S4). There were 135 new SNPs on the two contigs (2502 and 2574) that may reflect the impact of chromosomal duplications or repetitive sequences. Among the total 1023 (288 + 735) SNPs, 185 were heterozygous in some assayed samples. The effective SNP and indel discovery rates for Sanger sequencing were 2.9 SNPs per 100 bp (1023 SNPs per 35,504 bp) and 1.1 indels per 1000 bp, respectively. The effective SNP discovery rate for 454 pyrosequencing was 1.7 putative SNPs per 100 bp or a minimum 0.8 validated SNPs per 100 bp.

Blasting the 15 nonconfirmed contigs revealed that 10 contigs were matched in either NCBI (NCBI, 2011) or autoSNPdb barley data (ACPFG Bioinformatics, 2011) and the matched contigs were largely associated with some transposons. Thus, 95 of the 100 selected contigs were validated. Further blasting of another 50 contigs randomly selected from the 2578 contigs revealed 33 (66%) contigs matched in either NCBI or autoSNPdb databases (Supplemental Table S2). A few matched contigs were associated with wheat (*Triticum aestivum* L.) 3B Bacterial Artificial Chromosome (BAC) library (O'Connor et al., 1989) contigs.

Genetic Diversity

The SNP and contig identification with respect to sample was biased toward the output of the Roche 454 run for a sample (Table 1). A linear regression analysis revealed that the number of passed reads per sample was significantly ($p < 10^{-8}$) associated with the number of contigs and SNPs identified on the sample but not with the percentage of heterozygous SNPs (Table 1). Overall, the samples representing the eastern region had more passed reads, identified contigs, and putative SNPs but lower percentages of putative SNPs that were undetected for a sample due to the lack of related sequence reads than those samples from the western region (Table 1). The samples with the highest and lowest numbers of SNPs (NepalE and KyrgyzstanE, respectively) were from the eastern region. However, the mean percentage of heterozygous SNPs observed for the eastern samples was much lower (1.4%) than that for the western samples (2.5%) (Table 1).

The AMOVA revealed 34.02% of SNP variation residing between the eastern and western samples and a larger average number of pairwise nucleotide differences within the eastern than the western group of samples (151.5 and 96.7, respectively). Removing the KyrgyzstanE sample generated 33.18% of SNP variation residing between the seven eastern and eight western samples. All

Table 3. Validation results on identified contigs, putative single nucleotide polymorphisms (SNPs), and related SNP bases with Sanger sequencing (SS).

Feature	All samples	ChinaE	PakistanE	JordanW	EthiopiaW
85 validated contigs					
SNP predicted	620				
SNP confirmed by SS	288				
SNP not confirmed by SS	332				
New SNP from SS	735				
New indel from SS	39				
All SNP-indels from SS	1060				
SNP base match (%)	83.6%				
Bases missing from 454 reads		510	458	248	463
Bases missing from SS		17	42	68	35
Total valid bases to match		94	121	305	123
Mismatched bases (%)		4 (4.3%)	31 (25.6%)	67 (22.0%)	17 (13.8%)
Matched bases (%)		90 (95.7%)	90 (74.4%)	238 (78.0%)	106 (86.2%)

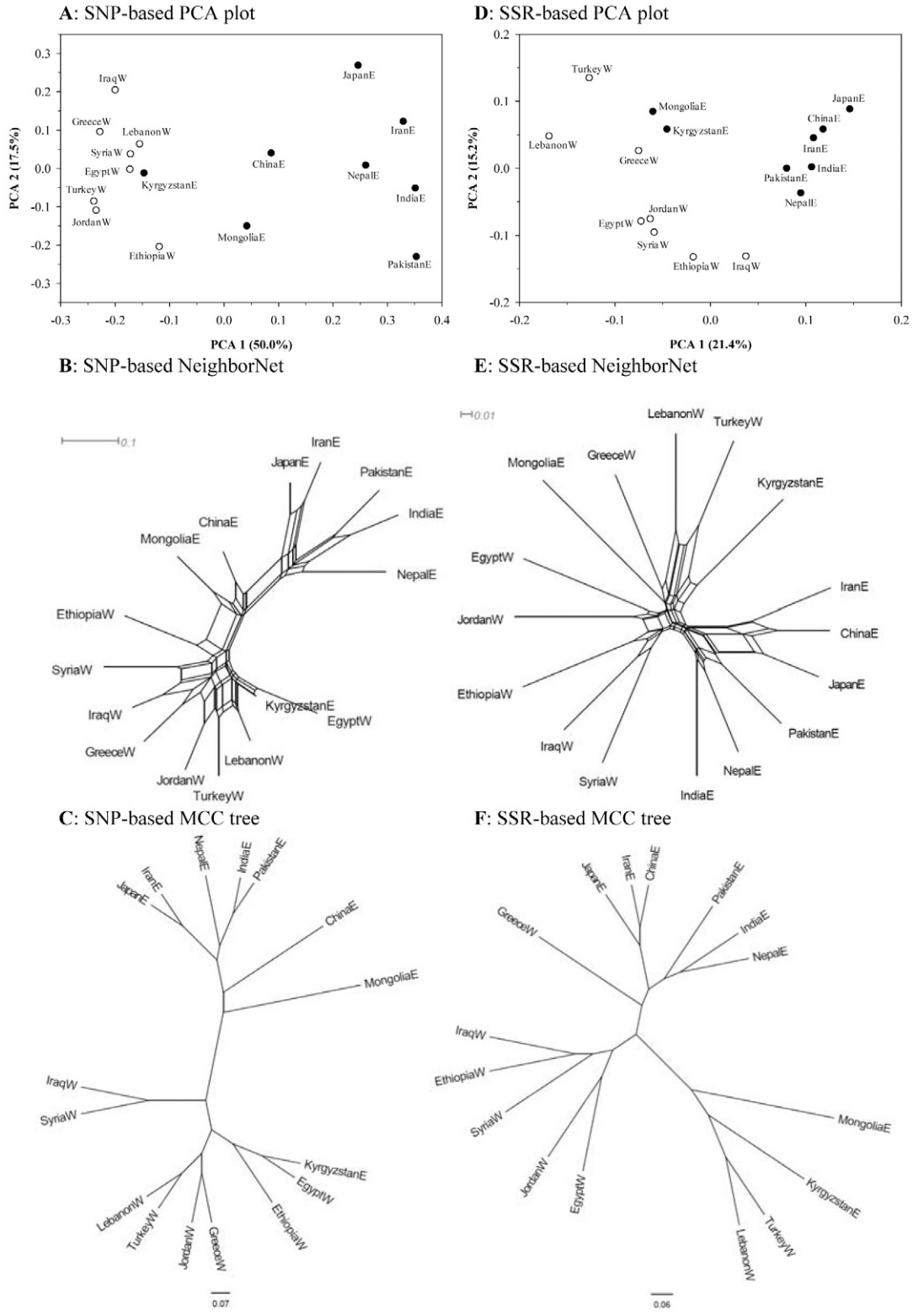

Figure 2. Single nucleotide polymorphism (SNP)- and simple sequence repeat (SSR)-based genetic relationships of the 16 barley samples inferred with the principle component analysis (PCA) (A and D), the NeighborNet (Bryant and Moulton, 2004) of SplitsTree4 (Huson and Bryant, 2006) (B and E) and the maximum clade credibility (MCC) tree of BEAST (Drummond and Rambaut, 2007) with tree prior as exponential growth (C and F). The sample labels are given in Table 1.

the AMOVA components reported here were statistically significant at $p < 0.0001$.

Three different analyses of 3980 putative SNPs were performed to assess genetic relationships among the 16 barley samples and revealed essentially the same patterns of sample clustering. The biplot of the first and second principal components (Fig. 2A) showed that the eight western samples were clustered together with the

eastern sample KyrgyzstanE, while the other seven eastern samples formed another diverse group. Also, the four most distinct samples were JapanE, PakistanE, IraqW, and EithopiaW. The NeighborNet (Bryant and Moulton, 2004) of SplitsTree4 (Huson and Bryant, 2006) (Fig. 2B) shared the same pattern of clustering as the principle component analysis (PCA) clustering but revealed considerable information on reticulations for various samples. For example, condense reticulations were observed in the western samples. The MCC trees obtained with tree prior as constant size, expansion, or exponential growth were topologically the same, although branch lengths varied among the trees. The MCC tree with tree prior as exponential growth (Fig. 2C) mostly mirrored with the NeighborNet but should be more informative to reveal the genetic relationships than the last two analyses. It showed MongoliaE and ChinaE samples were most closely related to the western samples than PakistanE and IndiaE samples. In contrast, the IraqW and SyriaW samples were most closely related to the eastern samples.

The analysis of optimal genetic structures through the STRUCTURE program (Pritchard et al., 2000; Falush et al., 2007) revealed five to six optimal groups of the 16 samples with the highest log-likelihood values of −12,837.2 and −12,837.6, respectively. However, through the BAPS program (Corander et al., 2008), 16 optimal groups were found for these 16 samples with the highest partition log likelihood of −43,065.7 (detailed results not shown).

The analysis of 156 SSR alleles from 30 SSR loci revealed 9.3% SSR variation residing between the eastern and western samples but a smaller average number of pairwise differences within the eastern than the western group of samples (35.6 and 40.4, respectively). The PCA plot (Fig. 2D) showed that the western barley samples were more diverse than the eastern ones and two eastern samples (MongoliaE and KyrgyzstanE) were more close to the western samples. Also, the other six eastern samples were separately formed as a small group. Such grouping was essentially the same as the NeighborNet (Bryant and Moulton, 2004) (Fig. 2E) showed. However, the MCC tree (Fig. 2F) seems to display three major groups with the GreeceW sample close to the eastern sample group. Overall, these SSR-based results are compatible with those of 454-SNP data.

Discussion

As expected, a full 454 run of the 16 barley samples generated a considerably large set of barley genomic resources with 2578 identified contigs and 3980 putative SNPs. The Sanger sequencing of four barley samples with identified contigs not only confirmed most of the putative contigs and SNPs but also revealed a large number of new SNPs and indels. Different diversity analyses of resulting 454-SNP data revealed the eastern and western division in the barley samples. The division is compatible with those inferred from barley SSR data and is consistent with our current knowledge of cultivated barley. These results help to illustrate the utility of 454 pyrosequencing for crop genetic diversity studies, particularly in sampling complex crop genomes.

Application of 454 Pyrosequencing

The NGS application presented here is straightforward, rapid, and cost effective (Fig. 1). This study effectively lasted roughly 2 mo and operationally cost under US$15,000 in 2010. The major features of this study include the use of genetically diverse samples to maximize the identification of genetic variants, the application of the DIAL (Ratan et al., 2010) pipeline requiring no a priori sequence information for the contig and SNP identification, and different diversity analyses of resulting 454-SNP data in comparison with barley SSR data. However, many issues were also found. Considerable unbalance both in sequence reads and putative SNPs was observed for these barley samples. Variable efficiency was also detected with respect to sample barcoding and pooling. Low SNP and SNP base validation rates were observed. Extra efforts were required to apply the DIAL pipeline with respect to target genome size and output delivery.

The application of the DIAL (Ratan et al., 2010) pipeline is fairly easy for those with some knowledge of the Linux operating system. It does not require a reference sequence and is fully automatic in contig and SNP identification. The complete analysis of 16 SFF files took about 2 h on a Linux server, after training for target genome size and Newbler version (454 Life Sciences, 2010). The effective target genome size for this study was 3 Mbp for the effective sequence length of about 600 Mbp. The reduced numbers of contigs and SNPs identified under the newest Newbler version were due to the increased stringency in the updated routines as defaults for longer reads. As the DIAL pipeline employed highly stringent filters and usually reported only a few highly confident SNPs, we developed extra tools with less stringent filters to extract contigs and SNPs for SNP validation and for data report and analysis. Clearly, further improvement is still desirable for its independence to Newbler, automatically defined target genome size, and user-friendly output delivery. Also, further assessments on its performance relative to other commonly used approaches, including the performance of Newbler Mapping Assembler (454 Life Sciences, 2010), would be helpful.

Contig and Single Nucleotide Polymorphism Validation

As the barley genome is not completely sequenced (Mayer et al., 2011), it is important to assess the reliability of the identified contigs and SNPs. The Sanger sequencing with four barley samples confirmed 85% of the selected contigs, 55% of the putative SNPs on the validated contigs, and 85% SNP bases matched on the four samples (Table 3). Blasting in related databases helped to confirm 95 of the selected contigs. However, the SNP validation rate was considerably low when compared with the validation rates of 85 to 95% reported in other plant species (Deschamps et al., 2010; Hyten et al., 2010; Y.B. Fu and G.W. Peterson,

unpublished data, 2011). The low validation rate probably reflects the limited number of barley samples used and increasing barley samples may validate more putative SNPs on the validated contigs (Duran et al., 2009). Also, the duplication chromosome segments or repetitive sequences (Wicker et al., 2008) may have increased the false SNPs identified by the DIAL (Ratan et al., 2010) pipeline. Such an effect was also reflected in the low (85%) SNP base match between 454 pyrosequencing and Sanger sequencing available on the four samples (Table 3). Interestingly, the Sanger sequencing revealed 125% more new SNPs and indels on these validated contigs. It is highly possible that a considerable proportion of such new SNPs and indels may reflect the effects of chromosomal duplication or repetitive segments, as the developed primer pairs cannot distinguish among these types of fragments with same lengths. As extensive barley genomic resources are available, blasting identified contigs in related databases can also provide a useful means to validate the contigs. For example, blasting a random set of 50 contigs revealed that 33 (66%) contigs were matched with high E-values (Supplemental Table S2).

The effective SNP discovery rates of 1.7 putative SNPs per 100 bp for 454 pyrosequencing and of 2.9 SNPs per 100 bp for Sanger sequencing are relatively high when compared with those reported in barley ranging from 1 per 31 bp to 1 per 240 bp (e.g., see Duran et al., 2009). They also are rather high when compared with the average SNP density observed in other plant genomes (or 1 SNP per 200 to 500 bp; Weising et al., 2005). The observed high SNP density may reflect the use of diverse genetic samples and/or the impact of chromosomal duplications or repetitive sequences as discussed above.

Diversity Analysis

As demonstrated above, the resulting 454-SNP data are full of sequencing and assembly errors and also highly unbalanced for each sample (Table 1). The large unbalance was related to the unbalance in effective sequence reads per sample. This may arise from sequencing errors by the 454 machine, developing reduced genomic libraries, and barcoding samples. A similar level of unbalance was also observed in a companion study of flax (*Linum usitatissimum* L.) germplasm (Y.B. Fu and G.W. Peterson, unpublished data, 2011). Thus, extra efforts are needed in these steps to minimize the unbalance. The observed unbalance and errors can complicate and/or invalidate some diversity analyses (Williams et al., 2010; Pool et al., 2010; Nielsen et al., 2011).

Clearly, a direct comparison of allelic richness among the assayed samples with large SNPs data missing was not informative. For example, there were 1259 and 1071 SNPs observed among the samples of eastern and western regions (Table 1), respectively, but the former samples were less unbalanced than the latter samples (68.3 and 73.1%, respectively). Similarly, there were 199 and 139 private SNPs for the samples of eastern and western regions, respectively. All of these results suggested more diversity in the eastern barley samples. However, the average percentage of heterozygous SNPs for the eastern barley samples was 1.4% while it was 2.5% for the western barley samples (Table 1). Also, with only 16 samples, a frequency-based diversity analysis would become more biased with such a large unbalance and thus not informative (Lynch, 2009; Jiang et al., 2009).

However, a direct inference of genetic relationships among the 16 samples based on such unbalanced 454-SNP data seems to be still informative (Fig. 2). Several distance-based analyses revealed the eastern and western division in the barley samples. This division was compatible with those inferred with a smaller scale of SSR data but had the higher resolution from a larger genomic sampling at 1914 contigs (or loci) with 3980 SNPs, thus supporting the argument of multiple barley domestication (Morrell and Clegg, 2007). One possible explanation is that these relationship inferences are based on genetic distance measures rather than specific genetic models or allele frequency spectrums; the former is knowingly less sensitive to unbalance in SNP and sample size than the latter. With a large number of SNPs identified even with considerable unbalance for each sample, the measures of genetic distances among individuals were approaching the true genetic distances. Also, many of the phylogenetic inferences such as the BEAST program (Drummond and Rambaut, 2007) were specifically built to address various uncertainties including data unbalance. In contrast, the model-based BAPS (Corander et al., 2008) and STRUCTURE programs (Pritchard et al., 2000; Falush et al., 2007) are expected to be underperformed because of the sensitivity to such large data unbalance.

In perspective, it is clear that the application of 454 pyrosequencing in a single step to discover and genotype SNPs for individual samples has its limits for crop genetic diversity analysis as demonstrated and discussed above. A more informative approach may be to discover SNPs first as we did here from a diverse set of samples followed by individual genotyping with identified contigs, as illustrated by the swine group (Wiedmann et al., 2008). The added effort and cost may be compensated by reduced errors and unbalance in one step approach and by enlarged sample sizes needed for diversity analysis (Williams et al., 2010). We are currently exploring this two-steps NGS approach for crop diversity analyses. Specific efforts may also be needed to develop proper designs with variable sample size and sequencing depth (Williams et al., 2010; Luca et al., 2011) and effective analytical tools to account for different sources of errors in NGS SNP data for accurate estimation of allelic frequencies (Lynch, 2009; Liu et al., 2010; Pool et al., 2010; Haubold, 2011).

Conclusions

The 454 pyrosequencing effort has generated a new informative set of genomic resources consisting of 2578 contigs and 4754 putative SNPs and indels for barley genomic research. The presented NGS application to identify contigs and SNPs is straightforward, rapid, and cost effective,

does not require a reference sequence, and can provide a good alternative to sample the complex genomes of non-model species. Diversity analyses of resulting 454-SNP data, although highly unbalanced and full of errors, still enhanced our understanding about the eastern and western division in the cultivated barley gene pool.

Supplemental Information Available

Supplemental material is available free of charge at http://www.crops.org/publications/tpg.

Figure S1. The sequence read distribution with respect to sequence read length (bp) for two regions of the full Roche 454 GS FLX Titanium run on the 16 barley samples.

Table S1. Variation of 30 simple sequence repeat (SSR) markers assayed in 16 barley accessions.

Table S2. List of 2578 contigs and 3980 putative single nucleotide polymorphisms (SNPs).

Table S3. List of primer sets for validation of 100 contigs and related putative single nucleotide polymorphisms (SNPs).

Table S4. List of 1393 single nucleotide polymorphisms (SNPs) and indels identified on 85 contigs using contig-specific primer sets.

Acknowledgments

The authors would like to thank Carolee Horbach for assistance in SNP validation and data processing, Matthew Links for assistance with access to a Linux server, Ken Richards for support and encouragement on the barley research, and two anonymous reviewers for their helpful comments on the early version of the manuscript.

References

454 Life Sciences. 2010. 454 sequencing system software manual, v 2.5p1. 454 Life Sciences, Branford, CT.

Altshuler, D., V.J. Pollara, C.R. Cowles, W.J. Van Etten, J. Baldwin, L. Linton, and E.S. Lander. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407:513–516. doi:10.1038/35035083

Australian Centre for Plant Functional Genomics (ACPFG) Bioinformatics. 2011. AutoSNPdb version 1.4. Available at http://autosnpdb.qfab.org.au/ (verified 15 Oct. 2011). Australian Centre for Plant Functional Genomics (ACPFG), School of Agriculture & Food Sciences, University of Queensland, Brisbane, QLD, Australia.

Barbazuk, W.B., S.J. Emrich, C.H. Chen, L. Li, and P.S. Schnable. 2007. SNP discovery via 454 transcriptome sequencing. Plant J. 51:910–918. doi:10.1111/j.1365-313X.2007.03193.x

Bennett, M.D., and J.B. Smith. 1976. Nuclear DNA amounts in angiosperms. Philos. Trans. R. Soc. London B Biol. Sci. 274:227–274. doi:10.1098/rstb.1976.0044

Bräutigam, A., and U. Gowik. 2010. What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. Plant Biol. 12:831–841. doi:10.1111/j.1438-8677.2010.00373.x

Bryant, D., and V. Moulton. 2004. NeighborNet: An agglomerative algorithm for the construction of planar phylogenetic networks. Mol. Biol. Evol. 21:255–265. doi:10.1093/molbev/msh018

Cannon, C., C.-S. Kua, D. Zhang, and J. Harting. 2010. Assembly free comparative genomics of short-read sequence data discovers the needles in the haystack. Mol. Ecol. 19(Suppl. 1):147–161. doi:10.1111/j.1365-294X.2009.04484.x

Close, T.J., P.R. Bhat, S. Lonardi, Y. Wu, N. Rostoks, L. Ramsay, A. Druka, et al. 2009. Development and implementation of high-throughput SNP genotyping in barley. BMC Genomics 10:582. doi:10.1186/1471-2164-10-582

Corander, J., P. Marttinen, J. Sirén, and J. Tang. 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics 9:539. doi:10.1186/1471-2105-9-539

Deschamps, S., M.L. Rota, J.P. Ratashak, P. Biddle, D. Thureen, A. Farmer, S. Luck, M. Beatty, et al. 2010. Rapid genome-wide single nucleotide polymorphism discovery in soybean and rice via deep resequencing of reduced representation libraries with the Illumina genome analyzer. Plant Gen. 3:53–68. doi:10.3835/plantgenome2009.09.0026

Doležel, J., J. Greilhuber, S. Lucretti, A. Meister, M.A. Lysák, L. Nardi, and R. Obermayer. 1998. Plant genome size estimation by flow cytometry: Inter-laboratory comparison. Ann. Bot. (London) 82:17–26. doi:10.1006/anbo.1998.0730

Druka, A., G. Muehlbauer, I. Druka, R. Caldo, U. Baumann, N. Rostoks, et al. 2006. An atlas of gene expression from seed to seed through barley development. Funct. Integr. Genomics 6:202–211. doi:10.1007/s10142-006-0025-4

Drummond, A.J., and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214. doi:10.1186/1471-2148-7-214

Duran, C., N. Appleby, M. Vardy, M. Imelfort, D. Edwards, and J. Batley. 2009. Single nucleotide polymorphism discovery in barley using autoSNPdb. Plant Biotechnol. J. 7:326–333. doi:10.1111/j.1467-7652.2009.00407.x

Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797. doi:10.1093/nar/gkh340

Excoffier, L., G. Laval, and S. Schneider. 2005. Arlequin ver. 3.01: An integrated software package for population genetics data analysis. Evol. Bioinform. Online 1:7–50.

Falush, D., M. Stephens, and J.K. Pritchard. 2007. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. Mol. Ecol. Notes 7:574–578. doi:10.1111/j.1471-8286.2007.01758.x

Flicek, P., and E. Birney. 2009. Sense from sequence reads: Methods for alignment and assembly. Nat. Methods 6:S6–S12. doi:10.1038/nmeth.1376

GeneCodes Corporation. 2010. Sequencher program v 4.10.1. GeneCodes Corporation, Ann Arbor, MI.

Gore, M.A., J.M. Chia, R.J. Elshire, Q. Sun, E.S. Ersoz, B.L. Hurwitz, J.A. Peiffer, et al. 2009. A first-generation haplotype map of maize. Science 326:1115–1117. doi:10.1126/science.1177837

Harlan, J.R. 1976. Barley. p. 93–98. *In* N.W. Simmonds (ed.) Evolution of crop plants. Longman, London, UK.

Haubold, B. 2011. Alignment-free estimation of nucleotide diversity. Bioinformatics 17:449–455. doi:10.1093/bioinformatics/btq689

Huson, D.H., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267. doi:10.1093/molbev/msj030

Hyten, D.L., Q. Song, E.W. Fickus, C.V. Quigley, J.S. Lim, I.Y. Choi, E.Y. Hwang, M. Pastor-Corrales, and P.B. Cregan. 2010. High-throughput SNP discovery and assay development in common bean. BMC Genomics 11:475. doi:10.1186/1471-2164-11-475

Imelfort, M., C. Duran, J. Batley, and D. Edwards. 2009. Discovering genetic polymorphisms in next-generation sequencing data. Plant Biotechnol. J. 7:312–317. doi:10.1111/j.1467-7652.2009.00406.x

Jiang, R., S. Tavare, and P. Marjoram. 2009. Population genetic inference from resequencing data. Genetics 181:187–197. doi:10.1534/genetics.107.080630

Li, H., J. Ruan, and R. Durbin. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18:1851–1858. doi:10.1101/gr.078212.108

Liu, X., Y.-X. Fu, J. Taylor, T.J. Maxwell, and E. Boerwinkle. 2010. Estimating population genetic parameters and comparing model goodness-of-fit using DNA sequences with error. Genome Res. 20:101–109. doi:10.1101/gr.097543.109

Liu, X., T.J. Maxwell, E. Boerwinkle, and Y.-X. Fu. 2009. Inferring population mutation rate and sequencing error rate using the SNP frequency spectrum in a sample of DNA sequences. Mol. Biol. Evol. 26:1479–1490. doi:10.1093/molbev/msp059

Long, Q., D. MacArthur, Z. Ning, and C. Tyler-Smith. 2009. HI: Haplotype improver using paired-end short reads. Bioinformatics 25:2436–2437. doi:10.1093/bioinformatics/btp412

Luca, F., R.R. Hudson, D.B. Witonsky, and A. Di Rienzo. 2011. A reduced representation approach to population genetic analysis and applications to human evolution. Genome Res. 21:1087–1098. doi:10.1101/gr.119792.110

Lynch, M. 2009. Estimation of allele frequencies from high-coverage genome-sequencing projects. Genetics 182:295–301. doi:10.1534/genetics.109.100479

Malysheva-Otto, L.V., M.W. Ganal, and M.S. Röder. 2006. Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.). BMC Genet. 7:6. doi:10.1186/1471-2156-7-6

Maughan, P.J., S.M. Yourstone, E.N. Jellen, and J.A. Udall. 2009. SNP discovery via genomic reduction, barcoding, and 454-pyrosequencing in amaranth. Plant Gen. 2:260–270. doi:10.3835/plantgenome2009.08.0022

Mayer, K.F.X., M. Martis, P.E. Hedley, H. Simkova, H. Liu, J.A. Morris, et al. 2011. Unlocking the barley genome by chromosomal and comparative genomics. Plant Cell 23:1249–1263. doi:10.1105/tpc.110.082537

Metzker, M.L. 2010. Sequencing technologies – The next generation. Nat. Rev. Genet. 11:31–46. doi:10.1038/nrg2626

Morrell, P.L., and M.T. Clegg. 2007. Genetic evidence for a second domestication of barley (*Hordeum vulgare*) east of the Fertile Crescent. Proc. Natl. Acad. Sci. USA 104:3289–3294. doi:10.1073/pnas.0611377104

National Center for Biotechnology Information (NCBI). 2011. Entrez cross-database search. Available at http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi (accessed 14 Oct. 2011). NCBI, Bethesda, MD.

Nielsen, R., J.S. Paul, A. Albrechtsen, and Y.S. Song. 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12:443–451. doi:10.1038/nrg2986

Nordborg, M., and D. Weigel. 2008. Next-generation genetics in plants. Nature 456:720–723. doi:10.1038/nature07629

Novaes, E., D.R. Drost, W.G. Farmerie, G.J. Pappas, D. Grattapaglia, R.R. Sederoff, and M. Kirst. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics 9:14. doi:10.1186/1471-2164-9-312

O'Connor, M., M. Peifer, and W. Bender. 1989. Construction of large DNA segments in *Escherichia coli*. Science 244:1307–1312.

Peterlongo, P., N. Schnel, N. Pisanti, M.-F. Sagot, and V. Lacroix. 2010. Identifying SNPs without a reference genome by comparing raw reads. p.147–158. Proc. String Processing and Information Retrieval (SPIRE), Lecture Notes in Computer Science, vol. 6393.

Pool, J.E., I. Hellmann, J.D. Jensen, and R. Nielsen. 2010. Population genetic inference from genomic sequence variation. Genome Res. 20:291–300. doi:10.1101/gr.079509.108

Pritchard, J., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Ratan, A., Y. Zhang, V.M. Hayes, S.C. Schuster, and W. Miller. 2010. Calling SNPs without a reference sequence. BMC Bioinformatics 11:130. doi:10.1186/1471-2105-11-130

Roche 454 Sequencing. 2010. Multiplex identifier (MID) adaptors for rapid library preparations. Technical bulletin TBC 2010-010, August 2010. 454 Life Sciences, Branford, CT.

Rohlf, F.J. 1997. NTSYS-PC 2.1. Numerical taxonomy and multivariate analysis system. Exeter Software, Setauket, NY.

Rokas, A., and P. Abbot. 2009. Harnessing genomics for evolutionary insights. Trends Ecol. Evol. 24:192–200. doi:10.1016/j.tree.2008.11.004

Rosenthal, A., O. Coutelle, and M. Craxton. 1993. Large-scale production of DNA sequencing templates by microtitre format PCR. Nucleic Acids Res. 21:173–174. doi:10.1093/nar/21.1.173

Rozen, S., and H.J. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. p. 365–386. *In* S. Krawetz and S. Misener (ed.) Bioinformatics methods and protocols: Methods in molecular biology. Humana Press, Totowa, NJ.

Rutschmann, F. 2006. Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times. Diversity Distrib. 12:35–48. doi:10.1111/j.1366-9516.2006.00210.x

Saisho, D., and M.D. Purugganan. 2007. Molecular phylogeography of domesticated barley traces expansion of agriculture in the old world. Genetics 177:1765–1776. doi:10.1534/genetics.107.079491

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. Molecular cloning: A laboratory manual. Cold Spring Harbor Lab., Cold Spring Harbor, NY.

Sato, K., N. Nankaku, and K. Takeda. 2009. A high-density transcript linkage map of barley derived from a single population. Heredity 103:110–117. doi:10.1038/hdy.2009.57

Seeb, J.E., G. Carvalho, L. Hauser, K. Naish, S. Roberts, and L.W. Seeb. 2011. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Mol. Ecol. Resour. 11(suppl. 1):1–8. doi:10.1111/j.1755-0998.2010.02979.x

Takahashi, R. 1955. The origin and evolution of cultivated barley. Adv. Genet. 7:227–266. doi:10.1016/S0065-2660(08)60097-8

Van Tassell, C.P., T.P. Smith, L.K. Matukumalli, J.F. Taylor, R.D. Schnabel, C.T. Lawley, C.D. Haudenschild, S.S. Moore, W.C. Warren, and T.S. Sonstegard. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods 5:247–252. doi:10.1038/nmeth.1185

Varshney, R.K., T.C. Marcel, L. Ramsay, J. Russell, M.S. Röder, N. Stein, et al. 2007. A high density barley microsatellite consensus map with 775 SSR loci. Theor. Appl. Genet. 114:1091–1103. doi:10.1007/s00122-007-0503-7

von Bothmer, R., T. van Hintum, H. Knüpffer, and K. Sato. 2003. Diversity in barley (*Hordeum vulgare*). Elsevier B.V., Amsterdam, The Netherlands.

Wang, D., J. Shi, S.R. Carlson, P.B. Cregan, R.W. Ward, and B.W. Diers. 2003. A low-cost, high-throughput polyacrylamide gel electrophoresis system for genotyping with microsatellite DNA markers. Crop Sci. 43:1828–1832. doi:10.2135/cropsci2003.1828

Weising, K., H. Nybom, K. Wolff, and G. Kahl. 2005. DNA fingerprinting in plants: Principles, methods, and applications, 2nd ed. CRC Press, Boca Raton, FL.

Wicker, T., A. Narechania, F. Sabot, J. Stein, G.T.H. Vu, A. Graner, D. Ware, and N. Stein. 2008. Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. BMC Genomics 9:518. doi:10.1186/1471-2164-9-518

Wicker, T., E. Schlagenhauf, A. Graner, T.J. Close, B. Keller, and N. Stein. 2006. 454 sequencing put to the test using the complex genome of barley. BMC Genomics 7:275. doi:10.1186/1471-2164-7-275

Wicker, T., S. Taudien, A. Houben, B. Keller, A. Graner, M. Platzer, and N. Stein. 2009. A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. Plant J. 59:712–722. doi:10.1111/j.1365-313X.2009.03911.x

Wiedmann, R.T., T.P. Smith, and D.J. Nonneman. 2008. SNP discovery in swine by reduced representation and high throughput pyrosequencing. BMC Genet. 9:81. doi:10.1186/1471-2156-9-81

Williams, L.M., X. Ma, A.R. Boyko, C.D. Bustamante, and M.F. Oleksiak. 2010. SNP identification, verification, and utility for population genetics in a non-model genus. BMC Genomics 11:32. doi:10.1186/1471-2164-11-720

You, F.M., N. Huo, K.R. Deal, Y.Q. Gu, M.-C. Luo, P.E. McGuire, J. Dvorak, and O.D. Anderson. 2011. Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. BMC Genomics 12:59. doi:10.1186/1471-2164-12-59

Zohary, D., and M. Hopf. 2000. Domestication of plants in the Old World, 3rd ed. Oxford Univ. Press, Oxford, UK.