Biology **2012**, *1*, 370-382; doi:10.3390/biology1020370

OPEN ACCESS

biology

ISSN 2079-7737

www.mdpi.com/journal/biology

Article

Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP

Michał T. Lorenc [1], Satomi Hayashi [2], Jiri Stiller [3], Hong Lee [1], Sahana Manoli [1], Pradeep Ruperao [1,4], Paul Visendi [1], Paul J. Berkman [3], Kaitao Lai [1], Jacqueline Batley [2] and David Edwards [1,*]

[1] Australian Centre for Plant Functional Genomics, School of Agriculture and Food Science, University of Queensland, Brisbane, QLD 4072, Australia; E-Mails: m.lorenc@uq.edu.au (M.T.L.); leehongching@gmail.com (H.L.); m.sahana@uq.edu.au (S.M.); p.ruperao@uq.edu.au (P.R.); paul.muhindira@uqconnect.edu.au (P.V.); k.lai1@uq.edu.au (K.L.)

[2] Centre for Integrative Legume Research, School of Agriculture and Food Science, University of Queensland, Brisbane, QLD 4072, Australia; E-Mails: s.hayashi@uq.edu.au (S.H.); j.batley@uq.edu.au (J.B.)

[3] CSIRO Plant Industry, Brisbane, QLD 4072, Australia; E-Mails: Jiri.Stiller@csiro.au (J.S.); Paul.Berkman@csiro.au (P.J.B.)

[4] Centre of Excellence in Genomics (CEG), International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502324, Andhra Pradesh, India

* Author to whom correspondence should be addressed; E-Mail: Dave.Edwards@uq.edu.au; Tel.: +61-7-3346-7084; Fax: +61-7-3365-1176.

Received: 12 July 2012; in revised form: 9 August 2012 / Accepted: 10 August 2012 / Published: 27 August 2012

Abstract: Single nucleotide polymorphisms (SNPs) are becoming the dominant form of molecular marker for genetic and genomic analysis. The advances in second generation DNA sequencing provide opportunities to identify very large numbers of SNPs in a range of species. However, SNP identification remains a challenge for large and polyploid genomes due to their size and complexity. We have developed a pipeline for the robust identification of SNPs in large and complex genomes using Illumina second generation DNA sequence data and demonstrated this by the discovery of SNPs in the hexaploid wheat genome. We have developed a SNP discovery pipeline called SGSautoSNP (Second-Generation Sequencing AutoSNP) and applied this to discover more than 800,000 SNPs between four hexaploid wheat cultivars across chromosomes 7A, 7B and 7D. All SNPs are presented for download and viewing within a public GBrowse database.

Validation suggests an accuracy of greater than 93% of SNPs represent polymorphisms between wheat cultivars and hence are valuable for detailed diversity analysis, marker assisted selection and genotyping by sequencing. The pipeline produces output in GFF3, VCF, Flapjack or Illumina Infinium design format for further genotyping diverse populations. As well as providing an unprecedented resource for wheat diversity analysis, the method establishes a foundation for high resolution SNP discovery in other large and complex genomes.

Keywords: single nucleotide polymorphisms; wheat; autoSNP; genome diversity; genotyping by sequencing; haplotype

1. Introduction

Molecular genetic markers describe genetic variations and provide a link between observed phenotypes and the underlying genotype [1–3]. The development of high-throughput methods for the detection of single nucleotide polymorphisms (SNPs) has led to a revolution in their use as molecular markers [4–7]. SNPs may be considered the ultimate genetic marker as they represent the finest resolution of a DNA sequence, are generally abundant in populations and have a low mutation rate [8]. The principal challenge in SNP discovery remains the discrimination between true genetic polymorphisms and the often more abundant sequence or mapping errors. SNP discovery is further confounded in polyploid species where multiple related genomes are present within each nucleus. The identification of sequence errors can be based on three methods: sequence quality score, redundancy of the polymorphism in a sequence alignment and presence of multiple haplotypes at a locus [9,10]. SNP redundancy provides an effective means for estimating confidence in the validity of SNPs independently of sequence quality scores and has been demonstrated to be an accurate method for SNP discovery in a range of species [11–13].

Many plant genomes are large and complex. The bread wheat (*Triticum aestivum*) genome is 17 Gbp in size, around 6 times larger than the human genome [14], consists of 75%–90% repeats [15,16] and is hexaploid, containing of the A, B and D genomes, each with 7 homoeologous chromosomes. The presence of multiple genomes, large size and abundant repeats make any genetic and genomic analysis a challenge [17]. The recent shotgun sequence assembly of isolated chromosome arms for the group 7 chromosomes provides a reference for SNP discovery across these chromosomes [18–20]. When combined with the recently produced whole genome shotgun sequence data for several wheat cultivars [21], there is a potential to identify large numbers of cultivar specific SNPs.

The rapidly expanding genome datasets, driven by advances in second generation DNA sequencing, present a challenge for their management and application [22]. The majority of SNP discovery software is designed for human or simple bacterial genomes, and they are not well adapted for polyploid plant genomes, especially crop genomes which are often highly homozygous [11,23–25]. Because of this, we have established a novel method for SNP discovery from complex genomes called SGSautoSNP, extended from original concepts developed in autoSNP, SNPServer and autoSNPdb [11,12,26]. Rather that attempting to identify all possible SNPs across a genome,

SGSautoSNP aims to identify as many highly confident SNPs as possible with the acknowledgement that not all biologically present SNPs will be identified. Here, we present the application of SGSautoSNP to wheat chromosomes 7A, 7B and 7D to identify more than 800,000 SNPs with an accuracy of greater than 93%, and present these polymorphisms within a GBrowse genome viewer at the wheatgenome.info web site [27].

2. Results and Discussion

We have developed a novel pipeline for the discovery of SNP polymorphisms in complex genomes. The SGSautoSNP (Second-Generation Sequencing AutoSNP) pipeline calls single nucleotide polymorphisms (SNPs) between different individuals using Illumina paired read data aligned to a reference. SGSautoSNP uses BAM (Binary Alignment/Map) format in order to save memory and space. These SNPs can be viewed using a broad range of visualization tools using GFF3, VCF and Flapjack output files. There is often a requirement to generate a consensus sequence based on the reads mapped to the reference and so SGSautoSNP can generate a consensus sequence as well as marker design files Illumina GoldenGate of Infinium assay designs.

There are many SNP discovery tools available. The main difference between our approach and most other SNP callers such as ACCUSA [28], AGSNP [29], NGS-SNP [30] and Atlas-SNP2 [31] is that the SGSautoSNP method does not consider the reference for SNP discovery. Instead, the reference is used to assemble the reads, and SNPs are then called between these assembled reads. Another difference is that other SNP callers are designed for SNP discovery from heterozygous populations. However, crop plants are frequently homozygous. In SGSautoSNP, mismapped reads produce a heterozygous genotype call at a locus, allowing their distinction from true homozygous SNPs. The SGSautoSNP method does not consider read quality score because these scores are not very reliable, with erroneous nucleotide calls often having high quality scores.

We demonstrate the potential of SGSautoSNP by searching for SNPs between four wheat cultivars using whole genome shotgun paired read Illumina sequence data. The reference consisted of bread wheat chromosome arm shotgun assemblies representing chromosomes 7A, 7B and 7D, as well as 4AL [32]. All three of the group 7 chromosome homoelogs were included in the mapping reference to ensure reads mapped to their correct homoeologous location. In the absence of one of the homoeologs, cultivar specific reads from the missing homoeolog would likely map to one of the other homoeologous genomes, confounding SNP discovery. An assembly from chromosome arm 4AL was included as this arm contains a reciprocal translocation with 7BS [18]. If 4AL was absent from the reference, 4AL specific reads which correspond to the 7BS translocation may map to the orthologous region on 7AS or 7BS, again confounding SNP discovery.

Cultivar specific reads were mapped to the reference sequences using SOAP [33]. The −r 0 option was applied which removes reads where they match multiple positions equally well. This option aims to increase SNP calling accuracy by ignoring read pairs that cannot be accurately positioned on the reference. Similarly, only reads that mapped as a pair were used for SNP discovery. Due to the short length of the reads, one read could match at many positions, but two reads separated by a gap of defined insert size provides a greater confidence of specific and accurate read mapping. The calling of SNPs between reads aligned to a reference while ignoring the reference allele allows this pipeline to be

applied to accurately call SNPs between individuals using a reference from a divergent species. While this pipeline does not attempt to call all biological SNPs, the very large numbers of SNPs identified are valuable for genetic studies and the association of traits with candidate agronomic genes.

Between 661,600,411 and 899,700,085 genome sequence read pairs where generated for each of the four wheat cultivars (Table 1). Of these, between 4.70% and 3.10% mapped to the group 7/4AL reference as read pairs. As the reference is estimated to cover 14% of the complete genome, the number of mapped reads is less than predicted. This is likely due to many read pairs mapping to multiple locations in this highly repetitive genome and subsequently being ignored due to the SOAP –r 0 option.

SGSautoSNP identified a total of 881,289 SNPs across the group 7 chromosomes. These consisted of 68% transitions and 32% transversions. This bias in transition/transversion ratio is commonly observed in SNP discovery and reflects the high degree of methlyl C to U mutation in genomes [34]. It may be expected that the bread wheat genome is highly methylated due to the two rounds of polyploidy and high repeat content. The observed transition/transversion bias provides a level of confidence in SNP prediction accuracy since erroneously called SNPs caused by sequence read errors or mismapping would be unlikely to display such a bias.

Validating individual SNPs in a hexaploid species is a challenge as the amplification of loci requires the design of homoeolog specific PCR primers. Of 40 SNPs selected for validation, 12 did not produce clean PCR amplification products or Sanger sequence. This reflects inefficiency in validation rather than SNP calling errors and so these SNPs were ignored. Of the 28 SNPs that did produce clean Sanger sequence data, 26 (93%) produced the expected genotype. One SNPs was homozygous across cultivars and not a true SNP, while one appeared to be heterozygous, suggesting a SNP between the homoeologous genomes rather than between cultivars.

All predicted SNPs have been included in a public wheat genome GBrowse database hosted at the wheatgenome.info web site [27]. This resource represents a substantial expansion over other recent wheat SNP discovery projects and provides a much greater density of SNPs than recently described methods for wheat. Allen *et al.* [35] recently identified 14,078 putative SNPs in 6,255 distinct reference sequences with Illumina GAIIx data from wheat lines Avalon, Cadenza, Rialto, Savannah and Recital. The validation rate from a subset of 1,659 was 67%. In a separate project, Lai *et al.* [36] identified a total of 38,928 candidate SNPs from bread wheat Roche 454 transcriptome data with an accuracy of 78% and presented these in an online database [37]. A pipeline package called AGSNP has also been applied to identify SNPs between two accessions of one of the diploid progenitors of bread wheat, *Aegilops tauschii*. Roche 454 sequencing of *A. tauschii* accession AL8/78 was combined with Applied Biosystems SOLiD sequencing of genomic DNA and cDNA from *A. tauschii* accession AS75 using AGSNP to identify a total of 497,118 candidate *A. tauschii* SNPs [29]. Currently, genome wide identification of hexaploid bread wheat SNPs using our pipeline is limited by the lack of publically available chromosome sequences. However, the identification of 881,289 SNPs across the group 7 chromosomes suggests that genome wide discovery would identify a total of more than 6 million SNPs.

The SNPs identified using SGSautoSNP can be used for genotyping by sequencing by low coverage skim sequencing of segregating wheat populations and calling genotypes where the low coverage sequence data aligns to a predicted polymorphic position. The successful application of the SGSautoSNP method to hexaploid wheat demonstrates that this approach should work for SNP discovery in other large and complex genomes.

Table 1. Summary of wheat cultivar data and mapping.

Wheat variety	Data generated	Data mapped to reference	% read pairs mapped
Drysdale	168 Gbp	8.65 Gbp	5.14%
Excalibur	146 Gbp	5.36 Gbp	3.66%
Gladius	180 Gbp	8.47 Gbp	4.70%
RAC875	132 Gbp	4.1 Gbp	3.10%

Table 2. Summary of Single nucleotide polymorphism (SNP) validation.

SNP Primer Name	Forward Primer	Reverse Primer	SNP score	Validation
UQ7A27	TAACATAAGCAAAGTTCTATTA	TTTGGAACACAATCGGAACTT	6	Failed
UQ7A1397	TCTATTGGATTCTTTCCGAT	TCACCCTGTGGAATGAAAGA	5	Failed
UQ7A5622	TTAGCCAAAATGGACCCAAA	CCTCTTTATTCAATCTGGAAACG	2	True SNP
UQ7A129835	TTCTTACTGTGGCTGCATCA	GCCATCCTAAACGACCTTCA	5	True SNP
UQ7A9400	GCCCATATGCAGTTCATGGT	AGAGCCAAACCTTCCCTGAT	2	Failed
UQ7A7915	CATGCCAACCCAAGTAGACC	GAAGCGTGAAAATTTCGTGA	6	True SNP
UQ7A6107	TGGTGTTTACGCTGAAGTTACC	CTGGCCTGGGCACTACATA	6	True SNP
UQ7A2603	GTCACCAACCAGCTCGAAAT	TTGTAGCTTTGCCTCTGTGAA	2	Failed
UQ7A3491	AGTCGCCGGCAGTAAAAATA	CCGAAGAAAATGTGGTGGAG	4	True SNP
UQ7A4532	TTTCCTCTAGATCTGTGCAAAATG	CATCCAGGACTGCATAAGCTC	6	True SNP
UQ7A100138	TCCCTGGTCCACGAGTTATT	AAATGGTTTGAGCCTTGTGC	7	Failed
UQ7A136305	CATCATCTTTGAAAAATCCTAGCC	TGTTCTGCAAGCTTCGTCTG	5	True SNP
UQ7A155877	AAGCTGTTGTGCCAGTGTTG	GAGCTAGCGTCGCTGACATA	4	True SNP
UQ7A180868	GACCGTCATCGAATGTAGCA	TCGTCCACCCAGACCTTATC	3	True SNP
UQ7A287189	GGCGATCATCACTTAAGAAACC	CAGTAATGAGGTTTCTGCTTGG	2	Failed
UQ7A322716	TCTGTTCGCAAACCAACG	GTGCGTTATCAGGGGAACAT	11	True SNP
UQ7A57227	ATGGGTGAAGGGAATACAGC	TGCATGCACATACAACCAAA	5	True SNP
UQ7A87191	TCAGTTCGGTAAGGATGAAGA	GAAGCAGTATGCATCTAAACTTTG	6	Heterozygous

Table 2. *Cont.*

SNP Primer Name	Forward Primer	Reverse Primer	SNP score	Validation
UQ7B21	GCAGGGTTAATTTCTAGCAAGC	GCCTTTTATCCAAAGCCATC	8	Failed
UQ7B484	CTCAACCTCCCAAGCATGA	GCTATCCAGCTACCCTGTGC	11	Failed
UQ7B3940	GCCAGAGGCACTAGCATCAC	GGTAATTGTGGAGCAAGCAA	6	True SNP
UQ7B4960	GCATGGCATTTCAAGATCAG	GGAGGAGGACAAAGCCAGAT	5	True SNP
UQ7B5991	CCAAGCCACCACCCTTTAT	TAATCCCCGTCATCTCGAAG	4	True SNP
UQ7B120997	CTCCTCAGATGACCAATTTGC	CACCAAAATATGCTGTACAATTCTATG	7	Failed
UQ7B256895	GCAGCAGAGGTAGGCACTTC	GAAATGCTTCGAGTGTGGTG	11	True SNP
UQ7B64318	GGGTCCAGACTTCCACGTTA	CCCACATTAATTTGTACGACCTC	6	Failed
UQ7B97303	TGATTCGAGCCCATATAGGAA	AGCCATGCGGAAATATTGAG	8	True SNP
UQ7D283	TGAGTAAGACAACAATCAGAGCA	CAATGCGAGCAAAAAGATCA	5	True SNP
UQ7D429	TGTGCTGACGTGGCATCTAT	GCATGTGGAAAACGAGTGTG	3	True SNP
UQ7D689	CATCTGGCCTCAACATCAAA	TGTTGGTAGTGAGGCACTTCTT	9	Failed
UQ7D948	GGCGATACTCGATGAAAGAAA	TTGGAAACTACAATTGCACAAC	9	True SNP
UQ7D1189	GCGTGGAGTAGAGGGACAAG	TCCAAAAAGCAAAACAAATGC	4	True SNP
UQ7D1491	AGCGCAAGGAGGAGGTTAGT	GAGCCAAGTCCTTGTCAATTT	7	True SNP
UQ7D1846	AATGTGTTCCATCCAAGACG	GCCAAGGTCGACATGTGATA	10	True SNP
UQ7D2314	AAACAAGTCTGTGTTGCGTCA	TGCAGATACATGGCTCCAGA	2	Monomorphic
UQ7D20375	CTGCCACCAAACGGATTAAC	AATGCATTGGCAGTCACAAG	6	True SNP
UQ7D27168	TAATGCTATGCCGTGTCAGC	GCCACCTATTATTGAAGGCATC	2	True SNP
UQ7D38754	GAGCGAGCAATGCTAGTGTG	GAACCCATTTGATAACCGTGA	3	Failed
UQ7D59683	CGTCCACATTGTTGCAAATC	TTGACCCTGAAGGAAGGATG	6	True SNP
UQ7D68910	TTGCTTTATGCCACTGGAGA	TAGGCCGTGAAACATCAACA	3	True SNP

3. Experimental Section

3.1. Data and Dependencies

Assemblies for each of the wheat 7A, B and D chromosomes, including the syntenic builds and extra contigs, were generated as described by Berkman *et al.* [18] and used as references for variety specific read mapping. The latest versions of the syntenic builds are accessible at the wheatgenome.info web site [20,27]. The assembly for chromosome 4AL was kindly provided by Dr Pilar Hernandez [32]. Whole genome Illumina sequence data was downloaded from the bioplatforms web site [21,38].

The SGSautoSNP pipeline is implemented in Python 2.7 and runs from the command line on any operating system where Python is available. Most of the SGSautoSNP scripts are multithreaded to improve performance with large genomes. Other requirements are pysam [39], a Python module for SAM/BAM formats; Biopython [40,41]; SAMtools version [42] and soap2sam.pl [43] to covert SOAP results to SAM format.

3.2. Read Mapping

All Illumina paired-end reads from each cultivar were aligned to the combined assemblies representing 4AL, 7A, 7B and 7D references using SOAPaligner 2 [33] with the –r 0 option and soap_wrapper.py. Depending on data volumes and compute infrastructure, read mapping generally takes between 3 and 48 hours. SOAP generates three results files for each cultivar: paired-end; single mapped reads; and unmapped reads. Only mapped paired reads were used for further analysis. Each of paired mapped read files was converted to sorted and indexed BAM files using SOAP2BAM.py. In order to allow SGSautoSNP.py to detect different cultivars, each read ID in the BAM file was modified to include a cultivar reference tag using generate_subset_BAM.py. Finally, BAM files for each cultivar were merged using SAMtools [42] to produce three BAM files representing 7A, 7B and 7D.

3.3. SNP Discovery

The reference is used to assemble the reads, and SNPs are then called between these assembled reads. Depending on data volumes and compute infrastructure, SNP discovery generally takes between 1 and 10 hours. The SGSautoSNP algorithm uses two steps to call a SNP at each locus. Primary SNP calling requires a SNP redundancy score of at least 2. The SNP redundancy score is the minimum number of reads calling the SNP allele at the locus. As at least 2 reads are required for at least 2 cultivars to call a SNP, the minimum coverage at a locus to call a SNP is 4. After this initial SNP call, the algorithm asks if all bases within each cultivar at a locus are the same, which would be expected for homozygous genomes. This process identifies erroneously called SNPs that are due to mis-mapping of reads.

SGSautoSNP.py produces five output types. A statistics file with the file extension ‘.stat’ contains SNP calling statistics including: (i) scaffold name (ii) SNP number (iii) SNP types (transitions and transversions) (iv) scaffold length. The end of this file contains a summary of all scaffolds. The first results file with the extension ‘.snp’ contains human readable SNP information in text format which

can be easily parsed to other formats. Information includes: (i) scaffold name (ii) SNP position on the scaffold (iii) SNP position on the chromosome (iv) SNP score (v) genotypes (which base and how many appear in a particular cultivar) (vi) allele. Three further results formats are produced. VCF [44] files are created to allow the user to view the SNPs in MagicViewer [45]. GFF3 format results are produced for viewing in the GBrowse generic genome browser [46] and Tablet [47], while Flapjack format files (.map and .extension) enable visualisation in Flapjack [48].

3.4. SNP Filtering

While SNP calling may use many individuals, SNPs that differentiate between two specific individuals are frequently required for downstream analysis. The filter_SNPs.py script parses the text '.snp' file to retrieve SNPs between specific individuals of interest. This script generates the same format output files as SGSautoSNP.py, but specifically for SNPs between two defined individuals. This script also produces a .matrix file which details the number of SNPs between all combinations of cultivars.

3.5. Generating a Consensus Sequence

The Bam2ConsensusSequence.py script accepts the alignment in BAM format and generates a consensus sequence for each scaffold. Using multiple CPU cores, the script goes through all nucleotide positions and generates a consensus sequence using the following rules: (i) if no base exists at the position then a N will be inserted; (ii) if only a single read covers the locus then the algorithm uses this read sequence (iii) if more than one read covers the position, and all nucleotides are the same, this nucleotide will be inserted; (iv) if more than one read covers the position, and one single read conflicts with the others, this single read is assumed to be an error and ignored, the majority base is inserted; (iv) if more than one read covers the position, and more than one read conflicts with the others, a degenerate base is inserted using IUPAC notation. The output file is one multiple FASTA file which include all contigs in the original BAM file.

3.6. Generating Illumina Marker Assay Files

The SGSautoSNP method can generate Illumina marker assay files for the design of Illumina Infinium and GoldenGate genotyping arrays. The SNP2Markers.py script requires as an input file the consensus sequence in FASTA format generated by Bam2ConsensusSequence.py and the text SNP file with a '.snp' extension generated by SGSautoSNP.py. Additional parameters include (i) species (ii) number of cultivars (iii) SNP library name (iv) version number (v) chromosome name (vi) output directory for the results files.

The script extracts the 5' and 3' sequence surrounding each predicted SNP in the following way: (i) the nucleotide sequence 150 bases each side of the SNP is extracted together with the SNP position in the format [A/C]; (ii) as the Illumina GoldenGate and Infinium assays design probes up to 60 bp adjacent to the SNP, assays are discarded if this region contains any N characters within the consensus sequence. Output files include a file of summary statistics '*_marker.stat' and a marker assay file for input into the Illumina SNP assay design '*_GoldenDB.csv'.

3.7. Validation

A total of 40 SNPs were randomly selected from the three group 7 reference genomes for validation, representing 18, 9 and 13 SNPs from the A, B and D genomes respectively. These SNPs had a range of redundancy scores. Genomic DNA was isolated from the four wheat cultivars Drysdale, Gladius, Excalibur and RAC875, according to a protocol adapted from [49]. PCR amplification of the 40 loci was performed using primers designed to conserved sequence surrounding the SNPs (Table 2) in a 20 μL reaction volume containing 1 × iTaq PCR buffer (containing 100 mM Tris-HCl and 500 mM KCl, pH 8.3) (Bio-Rad), 200 μM each dNTP (Bio-Rad), 0.5 μM each primer, 1.5 U iTaq DNA polymerase (Bio-Rad), RNase and DNase free water (Gibco) and 60 ng DNA. Thermocycling conditions for the reaction were 94 °C for 2 min, followed by 35 cycles of: 94 °C for 30 s, annealing for 1 min at 60 °C and extension for 1 min at 72 °C. Final extension was performed at 72 °C for 10 min. Gel electrophoresis on a 1% (w/v) agarose gel in 1 × TAE buffer [50] containing ethidium bromide resolved products, which were excised and purified using a silica method based on [51].

The purified PCR products were Sanger sequenced using BigDye 3.1, using forward PCR primers, and analysed using an ABI3730xl. The sequences for each locus and cultivar were aligned and compared using Geneious Pro v5.4.6 [52] with a cost matrix of 65%, a gap open penalty of 6, and a gap extension penalty of 3, and each of the SNPs assessed.

4. Conclusions

We have developed a pipeline for the identification of single nucleotide polymorphisms in complex genomes using second generation DNA sequence data. The application of this pipeline to wheat identified more than 800,000 SNPs across the three homoeologous group 7 chromosomes 7A, 7B and 7D with a validation rate of greater than 93%. This provides an unprecedented resource for wheat genomics and diversity analysis and demonstrates the potential to expand this application for SNP discovery in other large and complex genomes. SGSautoSNP is freely available on request for academic use.

Acknowledgments

This work was supported by the Australian Research Council [Projects LP0882095, LP0883462 and DP0985953] awarded to DE and JB. Support from the Queensland Cyber Infrastructure Foundation (QCIF), the Australian Genome Research Facility (AGRF), and the Australian Partnership for Advanced Computing (APAC) is gratefully acknowledged. Second generation DNA sequence data was provided by Bioplatforms Australia through the Australian Government's Education Investment Fund (EIF) Super Science Initiative.

References

1. Batley, J.; Edwards, D. SNP applications in plants. In *Association Mapping in Plants*; Oraguzie, N., Rikkerink, E., Gardiner, S., de Silva, H., Eds.; Springer: New York, NY, USA, 2007; pp. 95–102.

2. Duran, C.; Appleby, N.; Edwards, D.; Batley, J. Molecular genetic markers: Discovery, applications, data storage and visualisation. *Curr. Bioinformatics* **2009**, *4*, 16–27.
3. Edwards, D.; Batley, J. Plant bioinformatics: From genome to phenome. *Trends Biotechnol.* **2004**, *22*, 232–237.
4. Duran, C.; Eales, D.; Marshall, D.; Imelfort, M.; Stiller, J.; Berkman, P.J.; Clark, T.; McKenzie, M.; Appleby, N.; Batley, J.; *et al.* Future tools for association mapping in crop plants. *Genome* **2010**, *53*, 1017–1023.
5. Gupta, P.K. Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol.* **2008**, *26*, 602–611.
6. Rafalski, A. Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* **2002**, *5*, 94–100.
7. Varshney, R.K.; Nayak, S.N.; May, G.D.; Jackson, S.A. Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* **2009**, *27*, 522–530.
8. Edwards, D.; Forster, J.W.; Chagné, D.; Batley, J. What are SNPs? In *Association Mapping in Plants*; Oraguzie, N.C., Rikkerink, E.H.A., Gardiner, S.E., de Silva, H.N., Eds.; Springer: New York, NY, USA, 2007; pp. 41–52.
9. Barker, G.; Batley, J.; O'Sullivan, H.; Edwards, K.J.; Edwards, D. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics* **2003**, *19*, 421–422.
10. Edwards, D.; Forster, J.W.; Cogan, N.O.I.; Batley, J.; Chagné, D. Single nucleotide polymorphism discovery. In *Association Mapping in Plants*; Oraguzie, N., Rikkerink, E., Gardiner, S., de Silva, H., Eds.; Springer: New York, NY, USA, 2007; pp. 53–76.
11. Batley, J.; Edwards, D. Mining for Single Nucleotide Polymorphism (SNP) and Simple Sequence Repeat (SSR) molecular genetic markers. In *Bioinformatics for DNA Sequence Analysis*; Posada, D., Ed.; Humana Press: New York, NY, USA, 2009; pp. 303–322.
12. Duran, C.; Appleby, N.; Clark, T.; Wood, D.; Imelfort, M.; Batley, J.; Edwards, D. AutoSNPdb: An annotated single nucleotide polymorphism database for crop plants. *Nucleic Acids Res.* **2009**, *37*, D951–D953.
13. Duran, C.; Appleby, N.; Vardy, M.; Imelfort, M.; Edwards, D.; Batley, J. Single nucleotide polymorphism discovery in barley using autoSNPdb. *Plant Biotechnol. J.* **2009**, *7*, 326–333.
14. Paux, E.; Sourdille, P.; Salse, J.; Saintenac, C.; Choulet, F.; Leroy, P.; Korol, A.; Michalak, M.; Kianian, S.; Spielmeyer, W.; *et al.* A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* **2008**, *322*, 101–104.
15. Flavell, R.B.; Rimpau, J.; Smith, D.B. Repeated sequence DNA relationships in four cereal genomes. *Chromosoma* **1977**, *63*, 205–222.
16. Wanjugi, H.; Coleman-Derr, D.; Huo, N.X.; Kianian, S.F.; Luo, M.C.; Wu, J.J.; Anderson, O.; Gu, Y.Q. Rapid development of PCR-based genome-specific repetitive DNA junction markers in wheat. *Genome* **2009**, *52*, 576–587.
17. Berkman, P.J.; Lai, K.; Lorenc, M.T.; Edwards, D. Next generation sequencing applications for wheat crop improvement. *Am. J. Bot.* **2012**, *99*, 365–371.

18. Berkman, P.J.; Skarshewski, A.; Lorenc, M.T.; Lai, K.; Duran, C.; Ling, E.Y.S.; Stiller, J.; Smits, L.; Imelfort, M.; Manoli, S.; *et al.* Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotechnol. J.* **2011**, *9*, 768–775.
19. Berkman, P.J.; Skarshewski, A.; Manoli, S.; Lorenc, M.T.; Stiller, J.; Lars; Smits, L.; Lai, K.; Campbell, E.; Kubalakova, M.; *et al.* Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theor. Appl. Genet.* **2012**, *124*, 423–432.
20. Lai, K.; Berkman, P.J.; Lorenc, M.T.; Duran, C.; Smits, L.; Manoli, S.; Stiller, J.; Edwards, D. WheatGenome.info: An integrated database and portal for wheat genome information. *Plant Cell Physiol.* **2012**, *53*, 1–7.
21. Edwards, D.; Wilcox, S.; Barrero, R.A.; Fleury, D.; Cavanagh, C.R.; Forrest, K.L.; Hayden, M.J.; Moolhuijzen, P.; Keeble-Gagnère, G.; Bellgard, M.I.; *et al.* Bread matters: A national initiative to profile the genetic diversity of Australian wheat. *Plant Biotechnol. J.* **2012**, in press.
22. Batley, J.; Edwards, D. Genome sequence data: Management, storage, and visualization. *Biotechniques* **2009**, *46*, 333–336.
23. Duran, C.; Edwards, D.; Batley, J. Molecular marker discovery and genetic map visualisation. In *Applied Bioinformatics*; Edwards, D., Hanson, D., Stajich, J., Eds.; Springer: New York, NY, USA, 2009.
24. Imelfort, M.; Duran, C.; Batley, J.; Edwards, D. Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnol. J.* **2009**, *7*, 312–317.
25. Lee, H.; Lai, K.; Lorenc, M.T.; Imelfort, M.; Duran, C.; Edwards, D. Bioinformatics tools and databases for analysis of next generation sequence data. *Brief. Funct. Genomics* **2012**, *2*, 12–24.
26. Savage, D.; Batley, J.; Erwin, T.; Logan, E.; Love, C.G.; Lim, G.A.C.; Mongin, E.; Barker, G.; Spangenberg, G.C.; Edwards, D. SNPServer: A real-time SNP discovery tool. *Nucleic Acids Res.* **2005**, *33*, W493–W495.
27. Edwards, D. Wheatgenome.info. Available online: http://www.wheatgenome.info (accessed on 17 August 2012).
28. Fröhler, S.; Dieterich, C. ACCUSA—Accurate SNP calling on draft genomes. *Bioinformatics* **2010**, *26*, 1364–1365.
29. You, F.; Huo, N.; Deal, K.; Gu, Y.; Luo, M.-C.; McGuire, P.; Dvorak, J.; Anderson, O. 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* **2011**, *12*, 59.
30. Grant, J.R.; Arantes, A.S.; Liao, X.; Stothard, P. In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. *Bioinformatics* **2011**, *27*, 2300–2301.
31. Shen, Y.; Wan, Z.; Coarfa, C.; Drabek, R.; Chen, L.; Ostrowski, E.A.; Liu, Y.; Weinstock, G.M.; Wheeler, D.A.; Gibbs, R.A.; *et al.* A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res.* **2010**, *20*, 273–280.
32. Hernandez, P.; Martis, M.; Dorado, G.; Pfeifer, M.; Gálvez, S.; Schaaf, S.; Jouve, N.; Šimková, H.; Valárik, M.; Doležel, J.; *et al.* NGS and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant J.* **2011**, *69*, 377–386.

33. Li, R.; Yu, C.; Li, Y.; Lam, T.W.; Yiu, S.M.; Kristiansen, K.; Wang, J. SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **2009**, *25*, 1966–1967.
34. Coulondre, C.; Miller, J.H.; Farabaugh, P.J.; Gilbert, W. Molecular-Basis of Base Substitution Hotspots in *Escherichia coli*. *Nature* **1978**, *274*, 775–780.
35. Allen, A.M.; Barker, G.L.A.; Berry, S.T.; Coghill, J.A.; Gwilliam, R.; Kirby, S.; Robinson, P.; Brenchley, R.C.; D'Amore, R.; McKenzie, N.; *et al.* Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (Triticum aestivum L.). *Plant Biotechnol. J.* **2011**, *9*, 1086–1099.
36. Lai, K.; Duran, C.; Berkman, P.J.; Lorenc, M.T.; Stiller, J.; Manoli, S.; Hayden, M.; Forrest, K.L.; Fleury, D.; Baumann, U.; *et al.* Single nucleotide polymorphism discovery from wheat next generation sequence data. *Plant Biotechnol. J.* **2012**, in press.
37. Edwards, D. AutoSNPdb. Available online: http://autosnpdb.appliedbioinformatics.com.au/ (accessed on 17 August 2012).
38. Bioplatforms. Bioplatforms datasets. Available online: http://www.bioplatforms.com.au/datasets/wheat (accessed on 17 August 2012).
39. Heger, A. *Pysam*, 0.5+, 2012. Available online: http://code.google.com/p/pysam (accessed on 17 August 2012).
40. Foundation, P.S. *Biopython*, 1.58+; Python Software Foundation: Wolfeboro Falls, NH, USA, 2012.
41. Cock, P.J.A.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; *et al.* Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423.
42. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Subgroup, G.P.D.P. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079.
43. Institute, B.G. *soap2sam.pl*, 2010. Available online: http://soap.genomics.org.cn/down/soap2sam.tar.gz (accessed on 17 August 2012).
44. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T.; *et al.* The variant call format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158.
45. Hou, H.; Zhao, F.; Zhou, L.; Zhu, E.; Teng, H.; Li, X.; Bao, Q.; Wu, J.; Sun, Z. MagicViewer: Integrated solution for next-generation sequencing data visualization and genetic variation detection and annotation. *Nucleic Acids Res.* **2010**, *38*, W732–W736.
46. Donlin, M. Using the Generic Genome Browser (GBrowse). *Curr. Protoc. Bioinformatics* **2007**, *Chapter 9*, Unit 9.9.
47. Milne, I.; Bayer, M.; Cardle, L.; Shaw, P.; Stephen, G.; Wright, F.; Marshall, D. Tablet—Next generation sequence assembly visualization. *Bioinformatics* **2010**, *26*, 401–402.
48. Milne, I.; Shaw, P.; Stephen, G.; Bayer, M.; Cardle, L.; Thomas, W.T.B.; Flavell, A.J.; Marshall, D. Flapjack-graphical genotype visualization. *Bioinformatics* **2010**, *26*, 3133–3134.
49. Fulton, T.; Chunwongse, J.; Tanksley, S. Microprep protocol for extraction of DNA from tomato and other herbaceous plants. *Plant Mol. Biol. Rep.* **1995**, *13*, 207–209.

50. Sambrook, J.; Russel, D.W. *Molecular Cloning: A Laboratory Manual*, 3rd ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2001.
51. Boyle, J.S.; Lew, A.M. An inexpensive alternative to glassmilk for DNA purification. *Trends Genet.* **1995**, *11*, 8.
52. Drummond, A.J.; Ashton, B.; S, B.; Cheung, M.; Cooper, A.; Duran, C.; Field, M.; Heled, J.; Kearse, M.; Markowitz, S.; *et al.* Geneious, v5.4. Available online: http://www.geneious.com/ (accessed on 17 August 2012).