Rice (2010) 3:98–102
DOI 10.1007/s12284-010-9051-x

# SEG-Map: A Novel Software for Genotype Calling and Genetic Map Construction from Next-generation Sequencing

**Qiang Zhao · Xuehui Huang · Zhixin Lin · Bin Han**

Received: 17 August 2010 / Accepted: 2 September 2010 / Published online: 15 September 2010


**Abstract** The advent of next-generation sequencing technologies opens a new era for discovering genome diversity and genetic mapping. A sequencing-based method was recently developed to genotype recombinant populations with considerably improved resolution and reduced time and cost. To effectively implement this method, here we report the development of an analytic pipeline, sequencing enabled genotyping for mapping recombination populations (SEG-Map), for genotype calling and constructing genetic maps from next-generation sequencing data. SEG-Map was designed to interface with the commonly used tools for mapping next-generation short reads. The output data of SEG-Map would then allow constructing genetic maps and subsequent quantitative trait loci analyses directly. The package is available at http://www.ncgr.ac.cn/software/SEG. Moreover, directly sequencing over 500 rice landraces enabled a construction of a high-density rice haplotype map. This data set with an average of 2–3 SNPs per kilobyte between *indica* and *japonica* could be effectively used in the sequencing-based genotyping of their recombinant mapping populations.

**Keywords** SEG-Map · Genotype calling · Next-generation sequencing · Recombination bin

**Electronic supplementary material** The online version of this article (doi:10.1007/s12284-010-9051-x) contains supplementary material, which is available to authorized users.

Q. Zhao · Z. Lin
College of Life Science and Biotechnology,
Shanghai Jiaotong University,
Shanghai 200240, China

Q. Zhao · X. Huang · B. Han (✉)
National Center for Gene Research and Institute of Plant Physiology and Ecology, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences,
Shanghai 200233, China
e-mail: bhan@ncgr.ac.cn

B. Han
Beijing Institute of Genomics, Chinese Academy of Sciences,
Beijing 100029, China

## Introduction

Genetic mapping using various populations has served as the primarily means of gene discovery for important agronomic traits in rice. However, the genotyping processes that are based on PCR markers remain laborious, expensive, and time-consuming. The emergence of next-generation sequencing technologies and multiplexed sequencing methods opened the possibility to develop a new high-throughput genotyping strategy that utilizes single nucleotide polymorphisms (SNPs) generated by whole-genome sequencing (Craig et al. 2008; Cronn et al. 2008).

A sequencing-based method was recently developed to genotype recombinant populations, where all the individuals in the population were sequenced with low genome coverage (Huang et al. 2009). Moreover, through large-scale genome sequencing of hundreds of landraces, a high-density haplotype map containing ~3.6 million SNPs was constructed in rice (http://www.ncgr.ac.cn/RiceHapMap). The haplotype map could then greatly facilitate the SNP identification between parents of recombinant populations, through low-coverage sequencing of parental lines followed by missing genotype imputation. Using the high-density haplotype map in rice, coupled with the continuous increase in the throughput of next-generation sequencing technologies, the sequencing-based genotyping method holds great potential to replace the conventional methods to serve as the primary approach for genetic analysis of recombinant populations.

To facilitate a wide application of this new method, we further developed a pipeline, sequencing enabled genotyping for mapping recombination populations (SEG-Map) that allows the direct analysis of Illumina Genome Analyzer II (GAII) short reads for map construction. In this software package, the program for genotyping calling and breakpoint determination has been modified, to accommodate various types of mapping populations and to interface with programs for SNP identification and recombination bin map construction. With these functions combined, SEG-Map takes the short reads directly from Illumina GAII and outputs recombination bins that can be analyzed by existing programs for linkage map construction and quantitative trait loci analyses. The software package can be freely downloaded from our lab website and can be conveniently used to construct genetic maps from Illumina GAII sequences.

## Results

### A haplotype map of cultivated rice containing 3.6 million SNPs

More than 500 diverse rice landraces, which represented a large collection of rice accessions in China, were sequenced at approximately onefold genome coverage. A total of ~3.6 million SNPs were identified, which captured the largest amount of sequence variation in cultivated rice to date. A high-density rice haplotype map was constructed. The haplotype map of rice genome for the landraces is available in our website (http://www.ncgr.ac.cn/RiceHapMap). The sequencing-based genotype dataset can provide a fundamental platform for rice genetics research and breeding. For classical bi-parental cross-mapping for dissecting traits in rice, constructing recombinant mapping populations from these sequenced landraces will be both cost-effective and powerful, since the availability of high-quality genotype data of the landraces offered an informative resource for the selection of parental lines and facilitate subsequent genotyping work. From the data set of the rice haplotype map, an average of 2–3 SNPs were able to be identified between *indica* and *japonica* lines per kilobyte across the genome, which could be used in the sequencing-based genotyping of their recombinant mapping populations (Fig. 1). The average rate between two *indica* lines and between two *japonica* lines is 0.5–1 SNPs per kb and 0–0.5 SNPs per kilobyte, respectively.

### Single nucleotide polymorphisms identification between parents by low-coverage sequencing

Recently, we developed a strategy for high-throughput genotyping of recombinant inbred lines (RILs) derived from a cross between two sequenced rice accessions (Nipponbare and 93–11) (IRGSP 2005; Yu et al. 2002) by bar-coded resequencing with ~0.02-fold coverage of rice genome for each line (Huang et al. 2009). The applicability in genotyping the mapping population still relies on high-quality sequences of the parents to identify SNPs (Nipponbare and 93–11 had BAC-based sequences and whole-genome shotgun sequences, respectively). Now, we improved a KNN-based imputation approach, taking advantage of the haplotype map of cultivated rice mentioned above. The new approach thus enables SNP identification between parents

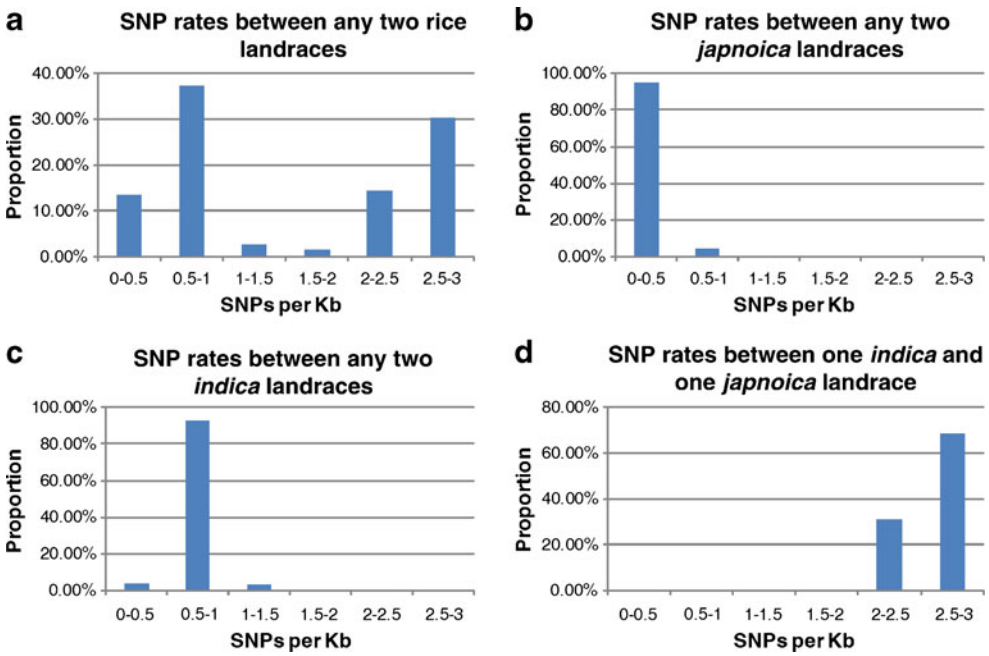


**Fig. 1** The distribution of SNP rates (the average of the rate across the genome for any two lines) between rice landraces based on the data set of rice haplotype map. **a** The distribution of SNP rates between any two rice landraces. **b** The distribution of SNP rates between any two *japonica* landraces. **c** The distribution of SNP rates between any two *indica* landraces. **d** The distribution of SNP rates between *japonica* and *indica* landraces.

with just low-coverage genome sequences of parents. With the highly accurate imputation method, each rice accession that had been sequenced at low coverage (onefold or even lower) was able to impute to over one million SNPs sites, which is much more cost-effective than high-coverage sequencing approach. Therefore, without deep resequencing of parental lines, the rice haplotype map coupled with the imputation method can help us to get the parents' SNP data set much cheaper and faster.

## Principal part of data process in SEG-Map

For a mapping population derived from two parents, genome-wide SNPs need to be identified between the parents prior to the SEG-Map analyses, which can be either the genome sequences already available for the parents in the rice haplotype map or sequences generated from low-coverage genome resequencing followed with missing genotype imputation. Since the identification of SNPs between parental lines can be performed in a fast and cost-efficient way, the sequencing-based genotyping of a recombinant population will mainly rely on subsequent analyses, including genotype calling, recombination breakpoint determination and linkage map construction. As a result, we developed a software, SEG-Map, to implement all the procedures of the work. The functions, procedures, and programs implemented in SEG-Map are illustrated in Fig. 2. The first step included two tasks that can be conducted in parallel. A certain number of individuals from the recombinant population are indexed with short sequence tags and pooled for sequencing in a lane of the Illumina GAII or similar sequencing platform. The number of individual to be pooled is determined by the genome size of mapping organisms, intended genome coverage, and sequencing throughput.

After SNPs are identified between the parents, a perl script, *PseudoMaker*, implemented in SEG-Map, is used to construct the pseudomolecules of each parental genome. Meanwhile, shorts reads generated for the individuals of the mapping population are sorted by indexes and the tags are subsequently trimmed by the perl scripts, *Split* and *Get_paired*, of the SEG-Map package. Then the SEG-Map interface allows the short reads of each individual to be mapped against both parental pseudomolecules with the existing mapping tools, such as Maq (Li et al. 2008), SOAP2 (Li et al. 2009), SSAHA2, and SMALT (Ning et al. 2001). The mapping results are input into script, *Maq2rlt*, *Soap2rlt*, *Ssaha2rlt*, or *Smalt2rlt* in the SEG-Map package for detecting SNPs of the mapping individuals. The SNPs are then examined by the script, *Seq2Bin*, for genotype calling, recombination breakpoint determination, and re-combination map construction. The recombination maps of all individuals are aligned and recombination bins are assigned by the script, *Bin2MCD*.

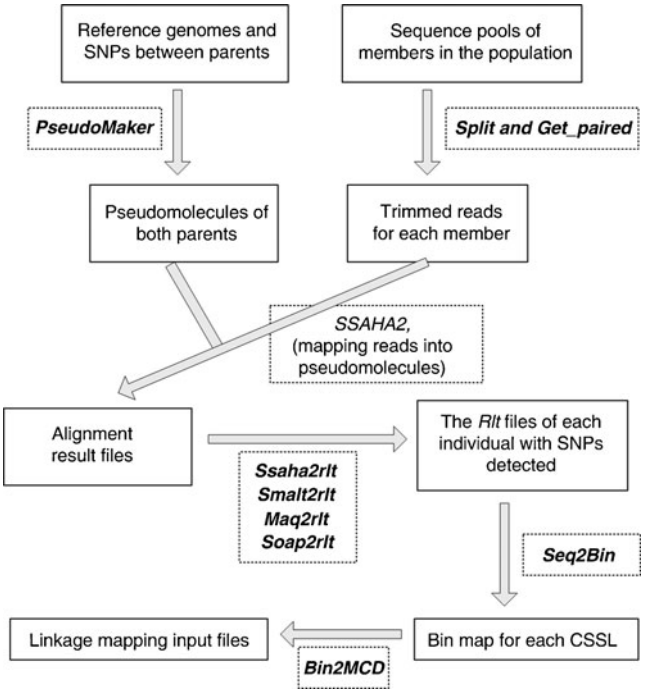


**Fig. 2** Illustration of the analytical framework of the sequencing-based genotyping method which has been implemented in the SEG-Map package. Programs in the SEG-Map package are in italic letters, performing a series of analyses in this study. The final genotype dataset from the SEG-Map package can be then used directly in other programs (including MapMaker and JoinMap) for linkage map construction.

The final output data of SEG-Map are recombination bins, usually with a resolution of a bin per 100 kb or even a bin per 10 kb (Fig. 3). The genotypes of the mapping population are input into programs such as MapMaker (Lincoln and Lander 1993) or JoinMap (Stam 1993) for linkage map construction. The linkage maps can then be used in QTL analysis, which provides a much finer scale than most conventional molecular markers.

## Process of a set of example data by using SEG-Map

We ran the analyses for a total of 270 rice RILs using the SEG-Map package and obtained a genotype map with 3,228 recombination bins (see Electronic supplementary Fig. 1). The entire analysis took approximately 8 min and 120 MB of RAM in a single processor. The software was compatible with multiple platforms (e.g., Unix, Linux and Windows). Only the "GD" module in the perl package is needed.

## Discussion

The work of genotyping a mapping population has long been laborious and time-consuming, including tedious and expensive processes for marker discovery and genotyping of

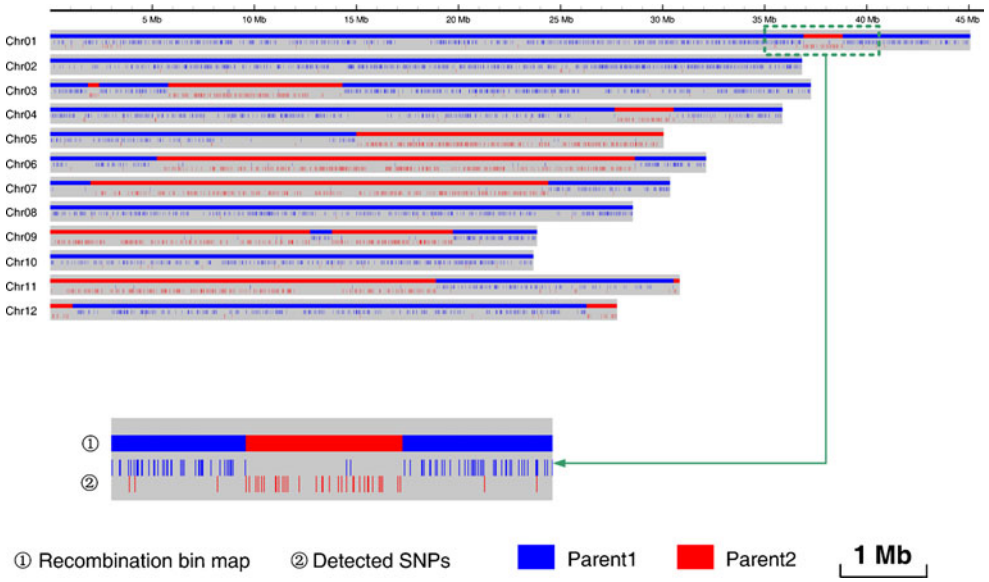


**Fig. 3** The recombination map of RIL #5, whose parents are Parent #1 and Parent #2.

hundreds of lines with hundreds of markers. Nevertheless, the resulting maps were still at relatively low resolution (Eshed and Zamir 1995; Loudet et al. 2002; Simon et al. 2008). Recently, the emergence of next-generation sequencing technology greatly augmented the sequencing throughputs and reduced both time and cost simultaneously, and might make population-scale QTL studies feasible (Charlesworth and Willis 2009). Coupled with bar-coded multiplexed sequencing strategy (Craig et al. 2008; Cronn et al. 2008), it has the potential to accomplish the genotyping work of a population with hundreds of individuals in a single Illumina GAIIx run. And newly available sequencing platforms with up to 10 Gb output per lane and techniques for higher multiplex levels will make this method even faster and more cost-effective. Through this approach, a high-resolution map can be constructed and nearly all recombination events in one population can be identified (Huang et al. 2009; Xie et al. 2010). Then traits are able to be associated with the individuals of the population efficiently (Rounsley and Last 2010).

In order to create a high-density map via resequencing all the members of the population with low coverage, there existed two major issues to be resolved. They were: (1) a single SNP site was no longer a reliable maker due to sequencing error; (2) individuals of a population were no longer scored at a fixed SNP site. To address the new challenges arising with the sequencing-based genotyping, we proposed a sliding window approach for accurate genotype calling, and adopted the "recombination bin" strategy for high-resolution genetic mapping.

SEG-Map was developed for the new genotyping method, implementing the intact analytical framework. SEG-Map was improved to suit different types of mapping populations (e.g., chromosome segment substitution lines, $F_2$, recombination inbred lines and near isolation lines) from various organisms. And it was compatible with several currently common mapping tools for short reads, and could be easily improved to import other sequence alignment tools. A major application for SEG-Map would be to generate a dense map for QTL analysis. One of the standard outputs of SEG-Map is the linkage map constructed with recombination bins. SEG-Map was also capable of outputting the input file (the Mcd format) for Windows QTL Cartographer, one of the most widely used software for QTL mapping. Moreover, SEG-Map can provide the graphs of the recombination map for each member in the population (Fig. 3 and Electronic supplementary Fig. 2), which would help to check or select individuals for further work (e.g., fine mapping and breeding). Scripts in the SEG-Map package may be flexibly modified for use in other studies.

More importantly, large-scale genome sequencing of hundreds of rice germplasm enabled a construction of a high-density rice haplotype map, which captured the largest amount of sequence variation in cultivated rice to date. This data set with a high-density of SNP rate crossing the genome could be greatly used in the sequencing-based genotyping of any rice recombinant mapping populations.

## Methods

### Identification of SNPs between parental lines from the data set of rice haplotype map

To detect SNPs that can be used directly in the sequencing-based genotyping, the genotypes of any pair of lines in the rice haplotype map were retrieved and compared with each other. The average rate between any *indica* line and any *japonica* line, between two *indica* lines, and between two *japonica* lines were counted, respectively.

### Sequence alignment, genotype calling and recombination breakpoint determination

The individuals of a mapping population were sequenced with low genome coverage (0.05×–0.2×). The reads in Illumina FASTQ format were converted to Sanger standard FASTQ format. Reads for each individual were sorted and trimmed by the perl scripts *Split* in the SEG-Map package. Then, 33-mer sequences were obtained after trimming the three-base index for each read.

Pseudomolecule of both parents of the mapping population was generated, by using another perl script *Pseudo-Maker*. The sequences of each individuals were then mapped into the pseudomolecules of both parents, with the short reads mapping tools, including Maq (Li et al. 2008), SOAP2 (Li et al. 2009), SSAHA2, and SMALT (Ning et al. 2001). The script *Ssaha2rlt* in the SEG-Map package then used the raw result of SSAHA2 outputs for SNP detection. Via a sliding window approach, the script *Seq2Bin* in the package collectively examined the detected SNPs for genotype calling based on the Bayes' rule, followed by recombination breakpoint determination and bin map construction (Electronic supplementary note).

### Genetic map construction

The script *Bin2MCD* in SEG-Map generated a dense map constructed with recombination bins (van Os et al. 2006) for quantitative trait loci (QTL) analysis. Once phenotypic evaluation has been performed and trait data are available, the outputs of the SEG-Map can be used to identify QTLs directly, using the QTL analysis software package, including Windows QTL Cartographer V2.5 (Wang et al. 2007).

### Generation of example data of an experimental result

We used the sequence data of a rice RIL population as example data, from a cross between ssp. *indica* cv. 93–11 and ssp. *japonica* cv. Nipponbare (Huang et al. 2009). The population has been expanded, and a total of 270 individuals were sequenced and used in this study. All of the raw data are available in the EBI European Nucleotide Archive with accession numbers ERA000078 (ftp://ftp.era.ebi.ac.uk/). The procedure of plant material, DNA isolation, sequencing and SNP identification were performed as described previously.

**Acknowledgements** We are grateful to Zemin Ning, Yan Zhao, and Tao Huang for computational assistance and helpful discussion.

**Financial support** China Rice Functional Genomics Programs (Ministry of Science and Technology of China to B.H., Grant No. 2006AA10A102); Chinese Academy of Sciences (Grant No. KSCX2-YW-N-024 to B.H.); National Natural Science Foundation of China (Grant No. 30821004 to B.H.).



## References

Charlesworth D, Willis JH. The genetics of inbreeding depression. Nat Rev Genet. 2009;10:783–96.

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, et al. Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods. 2008;5:887–93.

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res. 2008;36:e122.

Eshed Y, Zamir D. An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. Genetics. 1995;141:1147–62.

Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, et al. High-throughput genotyping by whole-genome resequencing. Genome Res. 2009;19:1068–76.

International Rice Genome Sequencing Project (IRGSP). The map-based sequence of the rice genome. Nature. 2005;436:793–800.

Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18:1851–8.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25:1966–7.

Lincoln SE, Lander SL. Mapmaker/exp 3.0 and mapmaker/qtl 1.1. technical report. Cambridge: Whitehead Institute of Medical Research; 1993.

Loudet O, Chaillou S, Camilleri C, Bouchez D, Daniel-Vedele F. Bay-0 x Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. Theor Appl Genet. 2002;104:1173–84.

Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. Genome Res. 2001;11:1725–9.

Rounsley SD, Last RL. Shotguns and SNPs: how fast and cheap sequencing is revolutionizing plant biology. Plant J. 2010;61:922–7.

Simon M, Loudet O, Durand S, Berard A, Brunel D, Sennesal FX, et al. Quantitative trait loci mapping in five new large recombinant inbred line populations of Arabidopsis thaliana genotyped with consensus single-nucleotide polymorphism markers. Genetics. 2008;178:2253–64.

Stam P. Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. Plant J. 1993;3:739–44.

van Os H, Andrzejewski S, Bakker E, Barrena I, Bryan GJ, Caromel B, et al. Construction of a 10, 000-marker ultradense genetic recombination map of potato: Providing a framework for accelerated gene isolation and a genomewide physical map. Genetics. 2006;173:1075–87.

Wang S, Basten CJ, Zeng ZB. Windows QTL Cartographer 2.5. Raleigh: Department of Statistics, North Carolina State University; 2007.

Xie W, Feng Q, Yu H, Huang X, Zhao Q, Xing Y, et al. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. Proc Natl Acad Sci U S A. 2010;107(23):10578–83.

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science. 2002;296:79–92.