BMC Bioinformatics

BioMed Central

Oral presentation

Revealing sequence variation patterns in rice with machine learning methods

Regina Bohnert*[1], Georg Zeller[1,2], Richard M Clark[2,3], Kevin L Childs[4], Victor Ulat[5], Renee Stokowski[6], Dennis Ballinger[6], Kelly Frazer[6], David Cox[6], Richard Bruskiewich[5], C Robin Buell[4], Jan Leach[7], Hei Leung[5], Kenneth L McNally[5], Detlef Weigel[2] and Gunnar Rätsch[1]

Address: [1]Friedrich Miescher Laboratory, Max Planck Society, 72076 Tübingen, Germany, [2]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany, [3]Department of Biology, University of Utah, Salt Lake City, UT 84112, USA, [4]Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA, [5]International Rice Research Institute, Metro Manila, The Philippines, [6]Perlegen Sciences, Inc., Mountain View, California, CA 94043, USA and [7]Bioagricultural Sciences and Pest Management, Colorado State University, Colorado, CO 80523, USA

Email: Regina Bohnert* - Regina.Bohnert@tuebingen.mpg.de

* Corresponding author

from Fourth International Society for Computational Biology (ISCB) Student Council Symposium
Toronto, Canada. 18 July 2008

Published: 30 October 2008

BMC Bioinformatics 2008, **9**(Suppl 10):O8 doi:10.1186/1471-2105-9-S10-O8

This abstract is available from: http://www.biomedcentral.com/1471-2105/9/S10/O8

Motivation

The major breakthrough at the turn of the millennium was the completion of genome sequences for individuals from many species, including human, worm and rice. More recently, it has also been important to describe sequence variation within one species, providing the first step towards the linkage of genetic variation to traits.

Today, rice is the most important source for human caloric intake, making up 20% of the calorie supply and feeding millions of people daily. The more detailed understanding and findings on the molecular assembly of phenotypic rice varieties will therefore be essential for future improvement in rice cultivation and breeding. In order to reveal patterns of sequence variation in *Oryza sativa* (rice), the non-repetitive portion of the genomes of 20 diverse rice cultivars was resequenced, in collaboration with Perlegen Sciences, Inc., using a high-density oligonucleotide microarray technology.

Methods

Based on experience gained in polymorphism studies for *Arabidopsis thaliana* [1] we developed a method for identifying single nucleotide polymorphisms (SNPs) from the array data using Support Vector Machines (SVMs). In a two-layered approach we trained SVMs to discriminate between SNP and non-SNP positions using information from each cultivar and, in a second step, across all cultivars.

Wherever several SNPs or deletion/insertion polymorphisms occur in close vicinity, the hybridisation is suppressed and SNP calling in these regions becomes infeasible. We therefore adapted a machine learning method for sequence segmentation [2,3] to predict *highly polymorphic* regions in *O. sativa* (cf. Figure 1). These regions can then be analysed in more detail using alternative experimental techniques.

For training and evaluation we compiled a set of reference polymorphisms obtained by dideoxy sequencing of more than 3,500 fragments from the 20 cultivars.

Results

Across all cultivars, we discovered 1,349,341 SNPs with the machine learning (ML) method at 316,373 non-

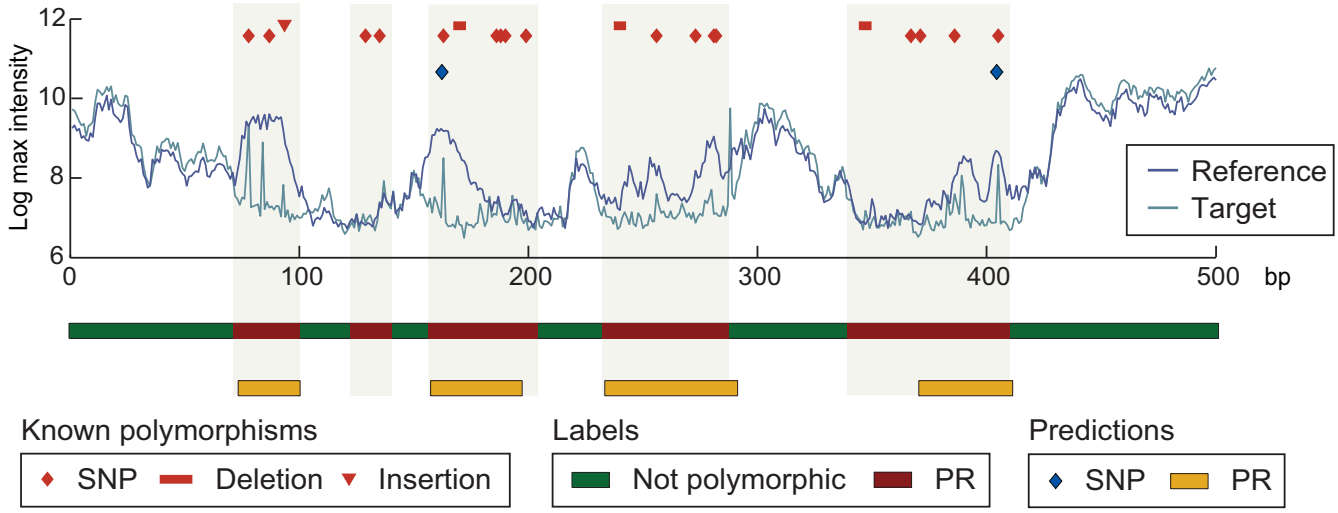

Figure 1
Log_2 intensities for the maximally hybridising oligonucleotide at each tiled position are shown for the reference and a target cultivar together with the known and predicted polymorphisms for the target cultivar (based on data from the *Arabidopsis* project [1,2]). Most of the 21 known polymorphisms in the target cultivar could not be predicted with the SNP calling methods.

redundant positions. In comparison to a model based (MB) SNP calling approach implemented by Perlegen Sciences, Inc. [4], the ML method was found to be much more sensitive by recovering 20.9% of all known SNPs at a precision of 91.7%, compared to 14.4% and 90.9%, respectively, for the MB approach (cf. Figure 2A). The intersection of MB and ML predictions contained 761,606 SNPs predictions at 159,879 non-repetitive positions constituting a set of markedly higher quality with a precision of 97.1%.

In addition to SNP predictions, our polymorphic region predictor discovered a substantial additional proportion of polymorphism regions, resulting in between ~65,000

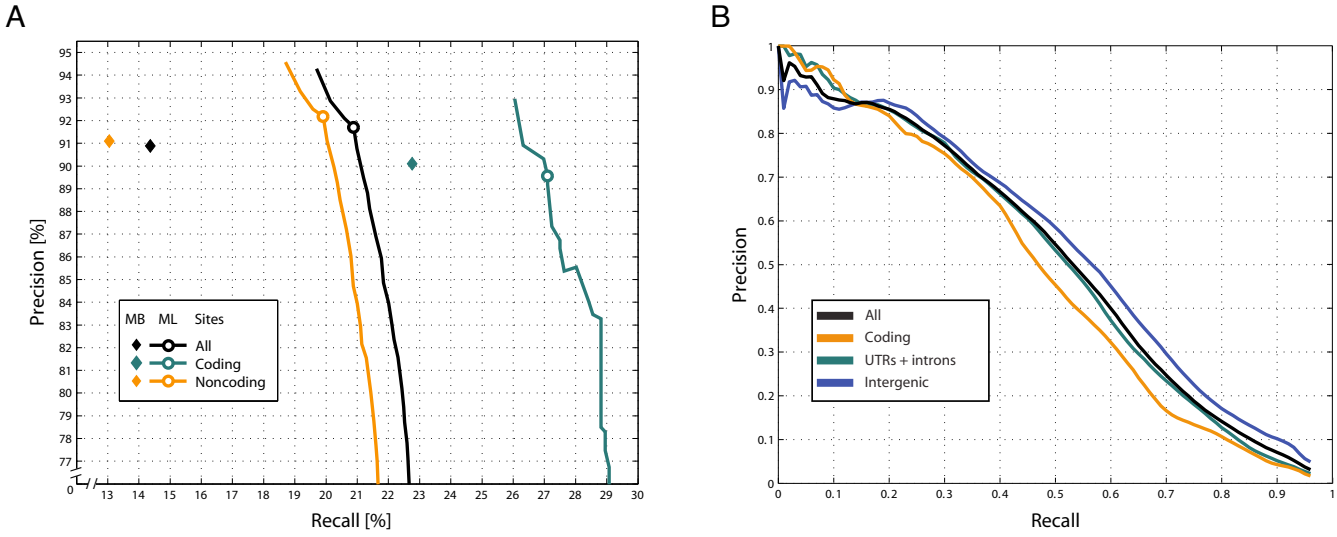

Figure 2
A. Comparison of the proposed SNP prediction (ML) method based on array intensities and additional information with a previously proposed one (MB). **B**. Accuracy of the polymorphic region predictions using the machine learning based segmentation algorithm.

and ~203,000 polymorphic regions per cultivar (cf. Figure 2B).

Conclusion

We identified hundreds of thousands polymorphisms on a genome-wide scale, providing the first whole genome set of polymorphisms for the world's most important crop plant. This polymorphism data represents a valuable resource for further functional studies and modern breeding of rice.

Based on the SNP data, high-density genotyping arrays will be designed to investigate genomic variation in many more rice cultivars. The PR predictions will e.g. be helpful to constrain primer design to conserved regions and thus increase PCR success rates.

References

1. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, Chen H, Frazer KA, Huson DH, Schölkopf B, Nordborg M, Rätsch G, Ecker JR, Weigel D: **Common Sequence Polymorphisms Shaping Genetic Diversity in *Arabidopsis thaliana*.** *Science* 2007, **317:**338-42.
2. Zeller G, Clark RM, Schneeberger K, Bohlen A, Weigel D, Rätsch G: **Detecting Polymorphic Regions in the *Arabidopsis thaliana* Genome with Resequencing Microarrays.** *Genome Research* 2008, **18:**918-29.
3. Tsochantaridis I, Joachims T, Hofmann T, Altun Y: **Large Margin Methods for Structured and Interdependent Output Variables.** *Journal of Machine Learning Research* 2005, **6:**1453-1484.
4. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR: **Whole-genome Patterns of Common DNA Variation in Three Human Populations.** *Science* 2005, **307:**1072-9.