Towards whole genome association genetic scans in barley

Varshney R., Kearsey M., Luo Z., Potokina E., Close T.J., Waugh R., Rostoks N., Ramsay L., Marshall D., Thomas B., Druka A., Wannamaker S., Svensson J., Bhat P., Graner A., Stein N.

in

Molina-Cano J.L. (ed.), Christou P. (ed.), Graner A. (ed.), Hammer K. (ed.), Jouve N. (ed.), Keller B. (ed.), Lasa J.M. (ed.), Powell W. (ed.), Royo C. (ed.), Shewry P. (ed.), Stanca A.M. (ed.).
Cereal science and technology for feeding ten billion people: genomics era and beyond

Zaragoza : CIHEAM / IRTA
Options Méditerranéennes : Série A. Séminaires Méditerranéens; n. 81

2008
pages 99-102

Towards whole genome association genetic scans in barley

R. Waugh*, N. Rostoks*, L. Ramsay*, D. Marshall*, B. Thomas*, A. Druka*, S. Wannamaker, J. Svensson**, P. Bhat***, A. Graner***, N. Stein***, R. Varshney***, M. Kearsey****, Z. Luo****, E. Potokina**** and T.J. Close****

*Genetics, SCRI, Invergowrie, Dundee, DD2 5DA, Scotland

**Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

***Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Department Genebank, AG Genome Diversity (GED) Corrensstr. 3, Gatersleben, D-06466, Germany

****School of Biosciences, The University of Birmingham, Birmingham, B15 2TT, UK

Introduction and background

Mapping traits and isolating the underlying genes is largely based on following the inheritance of both the trait and molecular markers in experimental populations from crosses between parents that contrast for the trait under study. However an alternative population-based approach termed "association genetics", or linkage disequilibrium (LD) mapping, is now being routinely used to map disease genes in humans (Ardlie *et al*., 2002, Hirschhorn and Daly, 2005, Tishkoff and Verrelli, 2003) (LD - defined as the non-random association of alleles at distinct loci in a sample population). In crop plants, the potential of association mapping, with the objective of estimating the position of genes conferring a specific trait or phenotype using LD between alleles of genetically mapped markers, has recently also become a focus of considerable interest. One major attraction of association genetics is the potential to locate genes responsible for a wide range of traits in a single sample population using pre-existing phenotypic data that has been collected during crop improvement and cultivar registration programs.

For whole genome association genetics to be successful, the extent of LD must be known in the genepool under study, and based on this information an appropriate number of markers developed and assembled to allow LD studies to proceed. In addition, for effective association mapping any inherent population structure must also must be known, as differences in allele frequencies at different loci among populations may create spurious associations (Lander and Schork 1994). Recent studies have already revealed that the extent of LD varies considerably between different crop species. In outbreeding maize, LD was shown to decay rapidly over approximately 1 – 2 kbp (Remington *et al*., 2001) in a set of inbred lines representing a large proportion of the genetic diversity in maize. In natural populations of *Arabidopsis*, an inbreeding species, LD was shown to extend to approximately 1 cM or 250 kbp, although local populations exhibited a much higher level of LD of up to 50 cM (Nordborg *et al*., 2002). These extremes illustrate a key issue in association studies. In maize, it is necessary to have a very large number of markers for association mapping. However, if the required number were available then the resolution they would afford may allow the identification of individual genes or even the polymorphisms responsible for determining a trait. In contrast, in *Arabidopsis*, fewer markers are needed for association mapping, but the resolution of the map will only allow the identification of large linkage blocks containing multiple genes.

Caldwell *et al*. (2006) recently identified very strong, but population dependent LD in barley. One important output of this study was that genome wide LD mapping in cultivated barley populations was at best likely to identify a region of the genome that controlled a trait. However, subsequent higher-resolution LD mapping may also be possible in landraces or wild barley to help resolve the underlying genes. Association genetics was successfully used to map yield and yield stability QTL in European spring barley using AFLP markers (Kraakman *et al*., 2004). The latter study identified associations between markers as far as 10 cM apart. Unfortunately, the technology used, even though it potentially provided markers for marker assisted selection, was not well suited for cloning the genes underlying the traits.

Development of a gene based genotyping platform for high throughput, low cost genotypic analysis

Several years ago we set out to develop a gene-based SNP genotyping platform to simplify genetic analyses in barley and realise the opportunities afforded by both comparative genetics to the model rice genome and, in relation to this paper, to facilitate whole genome association mapping scans. Despite SNPs being essentially biallelic, they have been shown to be equivalent to multiallelic SSR markers at moderate densities and superior to them at high densities in human linkage mapping. Most importantly, SNPs are the basis of a number of gel-free high throughput genotyping technologies that have been developed alongside the human genome sequencing and HapMap project. We initiated this project by re-sequencing EST-derived unigenes after PCR amplification from a number of genotypes to identify SNPs, and followed this by mapping the discovered SNPs onto the reference barley genetic linkage map (Rostoks *et al*., 2005, Nils Stein *et al*., unpublished results). Latterly, we have exploited electronic SNPs (eSNPs) extracted from EST data derived from different varieties (Close *et al*., in preparation). Using information from both sources we assembled 1524 barley SNPs (in 1524 unigenes) and used them to produce a pilot oligonucleotide pool array (OPA) for use with the Illumina Golden Gate Bead Array platform (Fan *et al*., 2003; Oliphant *et al*., 2002). We then used the Pilot OPA to genotype 3 mapping populations and 102 barley cultivars.

Extending the barley gene map

Our first objectives were to confirm the usefulness of the OPA platform for barley, the success rate of the eSNP predictions and, more importantly, to establish a high-density gene map of barley as the location and order of markers along the genetic map is preferentially required for association genetic analysis. We genotyped the parents and progeny from three reference doubled haploid mapping populations, Steptoe x Morex, OWBD x OWBR and Morex x Barke. Of the 1524 SNPs on the Pilot OPA, 1391 assays (91%) were successful, confirming the utility of the technology for HTP genotypic analysis in barley and eSNP prediction approach. The output of each individual SNP assay was manually inspected using the Illumina genotype calling software to check the quality of the calls and to make any required manual adjustments. A screen capture from the software is given in Fig. 1.

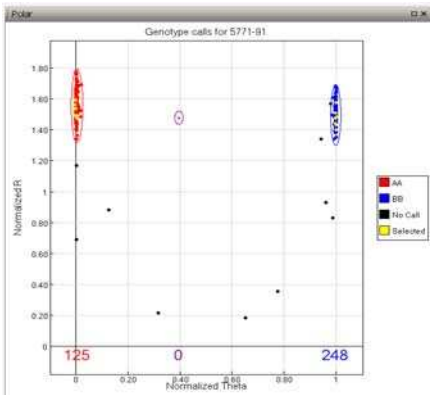

Fig 1. Graphical representation of the raw genotype calls for a single SNP locus using the Illumina software. The "no call" (black) points were DNA from an unrelated species and confirm the specificity of the assays. The tight grouping of the two alleles (dark grey within elipses) and their clear separation provide high quality binary data.

Over 1000 of the 1391 SNPs segregated in at least one of the three populations with approximately 200 segregating in any two cross comparisons and 100 segregating in all three. Initially, we constructed linkage maps of each population and these were subsequently used to develop a high quality integrated map comprised of 1029 SNPs using the markers mapped in more than one population. As part of this analysis we routinely construct graphical genotypes of each of the

individuals in a mapping population to get an idea of potential error in the dataset. An immediate and striking observation of the graphical genotypes was the obvious lack of single marker double recombinants in the dataset and the expected number of 0, 1 or 2 cross over events per chromosome per line suggesting that the data quality was high. The integrated map has good overall genome coverage but some gaps of over 10 cM remain, likely corresponding to highly recombinogenic regions of the barley genome.

The average number of SNP haplotypes in elite cultivated barley's

In order to estimate whether genotyping one SNP per locus would allow us to estimate the overall diversity present in cultivated barley, we re-sequenced alleles from a random set of 88 genes from 24 barley lines that included 19 elite European spring and winter barley varieties. The linkage map locations for most of these genes are known and they are located throughout the barley genome. In total, we obtained ca. 35 kb of aligned sequence and the full data set contained a total of 367 polymorphic sites. However the European subset (19) contained only 193 polymorphic sites. On average, we observed 3.4 haplotypes per locus in the full data set which was reduced to only 2.6 haplotypes per locus (with a maximum of six haplotypes) in the European barley sub-set. Based on this and previous observations relating to the extent of LD and haplotype diversity in cultivated barley (Caldwell *et al*., 2006, Piffanelli *et al*., 2004, Kraakman *et al*., 2004, Russell *et al*., 2004, our unpublished results) we concluded that assaying a single SNP per gene (locus), at approximately one gene per cM across the 1100 cM barley genome would provide suitable coverage for initial LD mapping, with the dual caveats that SNP haplotype information for certain genome regions rather than a single SNP, and that additional SNPs in certain genomic regions, may subsequently have to be factored into our analysis.

Whole genome SNP diversity and LD scans

We genotyped 102 barley accessions including 91 European barley varieties using the Pilot OPA and used the data to study genetic diversity, population structure and the extent of LD in the cultivated genepool. Of the 1391 successful SNP assays, only 0.3% heterozygous genotypes were observed, consistent with the inbred nature of barley. We also observed a surprisingly large amount of diversity in this relatively narrow germplasm selection. However there was a large number of SNPs with a minor allele frequency (MAF) of <0.1 and these were excluded from our subsequent LD analyses which included only 612 informative assays.

One of the primary objectives of our study was to determine whether whole genome association genetic scans would be possible in barley using this number of genetic markers. We used both principal coordinate analysis and the program "Structure" to investigate population substructure within the data. These identified three major subgroups within the germplasm, European spring and winter material (n=91) and more exotic lines. We then derived measures of LD (R^2) in the European germplasm subset set using the 612 markers with a MAF>0.1. As expected, the extent of LD we observed was strongly affected by population structure. Highly significant intra-chromosomal LD ($p>0.001$, $R^2>0.5$) extended over more than 60 cM (mean 3.9 cM, median 1.16 cM) in the combined set of European spring and winter barley with 20.4% of all significant ($p>0.001$) associations ($R^2>0.05$) being inter-chromosomal. However, in the spring 2-row subset (n=53), LD extended only up to 15 cM (mean 1.53 cM, median 0.8 cM) and the proportion of inter-chromosomal associations was reduced to 2%. The extent of LD varied across the chromosomes with an obvious relationship between genetic distance and LD. In contrast, there was no obvious relationship with physical distance: regions with reduced recombination, such as centromeres, showed strong LD even though the physical distances can be hundreds of megabases.

To test whether we could use LD to locate genes we examined whether we could correctly position any of the 362 unmapped genes in our Pilot OPA dataset via an LD mapping approach. 85 of these showed a MAF>0.1 in the 53 European spring barley varieties. We considered these to be the equivalent of 85 simple Mendelian traits that segregated in our population. We first established a quantitative threshold for LD mapping by attempting to re-map 140 loci with known linkage map locations using LD at different R^2 value cut-offs and compared the locations predicted by LD mapping with their positions determined by bi-parental meiotic mapping. This indicated that an $R^2>0.5$ was a

reasonable threshold giving >50% accuracy with ca. 5% false positive calls.

We calculated pair-wise LD (R^2) for each of the 85 unmapped and with mapped loci (MAF>0.1) and assumed that strong LD indicated linkage. When this threshold was applied, we were able to assign a putative map location to 43 of the 85 unmapped loci. We then compared these 43 loci with a putative map location, and the mapped loci with which they were in LD, to TIGR rice pseudomolecules version 4 using a BLAST x homology search (http://www.tigr.org/tdb/e2k1/ osa1/pseudomolecules/ info.shtml) to see if the predicted locations were supported by their known positions in rice based on collinevity between the two species. 81% (34 out of 42) co-located to the same region of the rice genome (with one mapped locus not having significant homology to rice), which supported the LD mapping results. We are in the process of extending our analysis to include a number of traditional and surrogate phenotypic traits.

Conclusions

Both this study and the previous studies of Kraakman *et al*. (2004, 2006) testify to the potential of exploiting whole genome LD-scans to locate genes controlling key biological traits in cultivated barley. We are currently increasing the density of makers, particularly those with a MAF >0.1, by developing two further pilot OPAs, which in due course will be compressed into two commercially available platforms for high throughput low cost genotyping in cultivated barley. In the immediate future these will be used in large association genetic studies in the UK and US involving approximately 4000 barley genotypes with the aim of realising the potential for whole genome association genetic scans in cultivated barley.

References

Ardlie, K.G. *et al.* (2002). *Nat. Rev. Genet.,* 3: 299-309.
Caldwell, K.S. *et al*. (2006). *Genetics,* 172: 557-567.
Fan, J. B. *et al.* (2003). *Cold Spring Harb. Symp. Quant. Biol.,* 68: 69-78.
Hirschhorn, J.N. and Daly, M.J. (2005). *Nat. Rev. Genet.,* 6: 95-108.
Kraakman, A.T. *et al.* (2004). *Genetics,* 168: 435-446.
Kraakman, A.T. *et al.* (2006). *Molecular Breeding*, 17: 41-58
Lander, E.S. and Schork, N.J. (1994). *Science,* 265: 2037-2048
Nordborg, M. *et al.* (2002). *Nat. Genet.,* 30: 190-193.
Oliphant, A. et al. (2002). *Biotechniques,* Suppl.: 56-1.
Piffinelli, P. *et al.* (2004). *Nature,* 430: 887-891
Remington, D.L. *et al.* (2001). *PNAS* 98: 11479-11484.
Rostoks, N. *et al.* (2005). *Mol. Genet. Genomics,* 274: 515-527.
Russell, J. *et al.* (2004). *Genome,* 47: 389-398.
Tishkoff, S.A. and Verrelli, B.C. (2003). *Annu. Rev. Genomics Hum. Genet.,* 4: 293-340.