A 1260 point genetic linkage map of potato chromosome 1: paving the way for Ultra High Density genetic linkage maps in crop species

D. Milbourne, E. Isidore, H. van Os[1], G. Bryan, H. van Eck[1], F. Rousselle-Bourgeois[2], E. Ritter[3], J. Bakker[4] & R. Waugh

With the paradigm shift towards high throughput characterisation of genes and their expression patterns inherent to the 'genomics revolution', the ability to relate simple and complex phenotypes to their underlying genes is becoming increasingly important. High-density linkage maps are indispensable tools for this task and, in addition, such maps have formed the basis for more in depth, genome-wide characterisation approaches such as physical mapping and whole genome sequencing projects for an array of model organisms. The utility of any genetic linkage map as a platform for all of these purposes is largely a function of both its density and accuracy. A primary goal of our group is marker saturation of the potato genome by the construction of an AFLP-based ultra high density (UHD) map containing approximately 10,000 markers. This represents a tenfold increase in the current coverage of 1000 markers for the combined tomato/potato RFLP-based map developed at Cornell University.

Despite the high throughput nature of AFLPs, time and financial constraints limit the population size upon which this type of experiment can be performed. Population size is important in determining the resolution of a map; the larger the population, the greater the number of meioses upon which the map can be based, and the greater the number of markers which can be ordered. Thus, lower resolution maps have greater numbers of markers that map to the same genetic location. We refer to this as the 'bin' concept (Fig.1), in which groups of co-segregating markers are represented on a map as a single co-segregation bin. This bin is defined by a 'bin signature', which is the common segregation pattern of all markers in that bin.

A second concept central to the creation of UHD maps is the realisation of the disproportionate effect of low levels of error in the segregation dataset used. Inclusion of an erroneous datapoint will result in a difference between the true and calculated position of a marker. The significance of this factor in creating a UHD linkage map can be illustrated by considering the creation of a linkage map of a single chromosome consisting of 1000 markers in a population of 100 individuals, with a marker scoring accuracy of 99%.

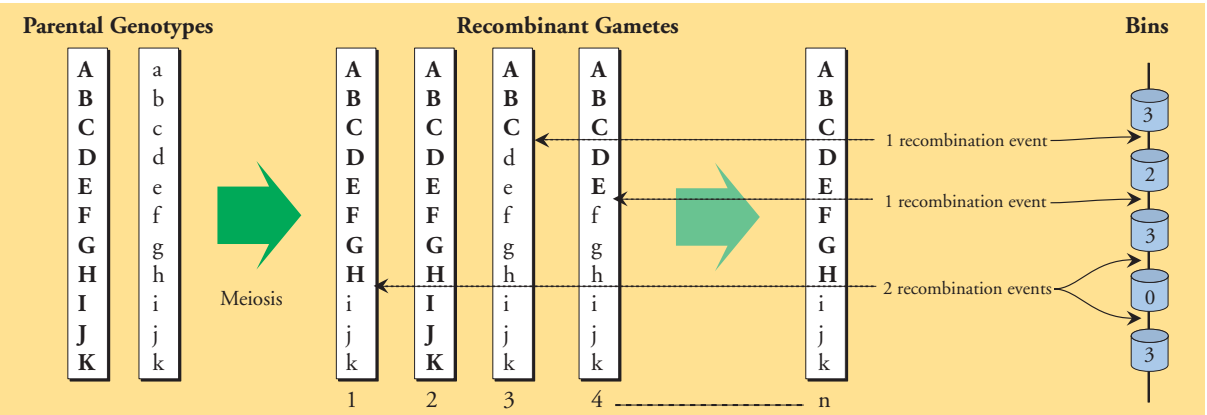

Figure 1 The Bin Map Concept. A completely heterozygous, diploid parental chromosome undergoes meiosis to produce recombinant haploid gametes. Following the segregation of marker alleles (upper and lower-case letters) in progeny derived from these gametes allows the generation of a linkage map. Markers that are never separated by a recombination event (in a population on *n* progeny individuals) co-segregate, and are placed in the same co-segregation bin. When more than one recombination event occurs between two consecutive markers, empty bins (bin 4 above) are placed on the map to represent every recombination event which cannot be visualised.

[1] Laboratory of Plant Breeding, Wageningen University, Lawickse Allee 166, 6709 DB Wageningen, The Netherlands.
[2] Station de Génétique et Amélioration des Fruits et Légumes, Institut National de la Recherche Agronomique, BP 94, 84143 Monfavet Cedex, France.
[3] AZTI- Instituto Pesquero y Alimentario, Aparatado 48, E-01080 Vitoria, Spain.
[4] Labratory of Nematology, Wageningen University, Binnenhaven 10, 6709 PD Wageningen, The Netherlands

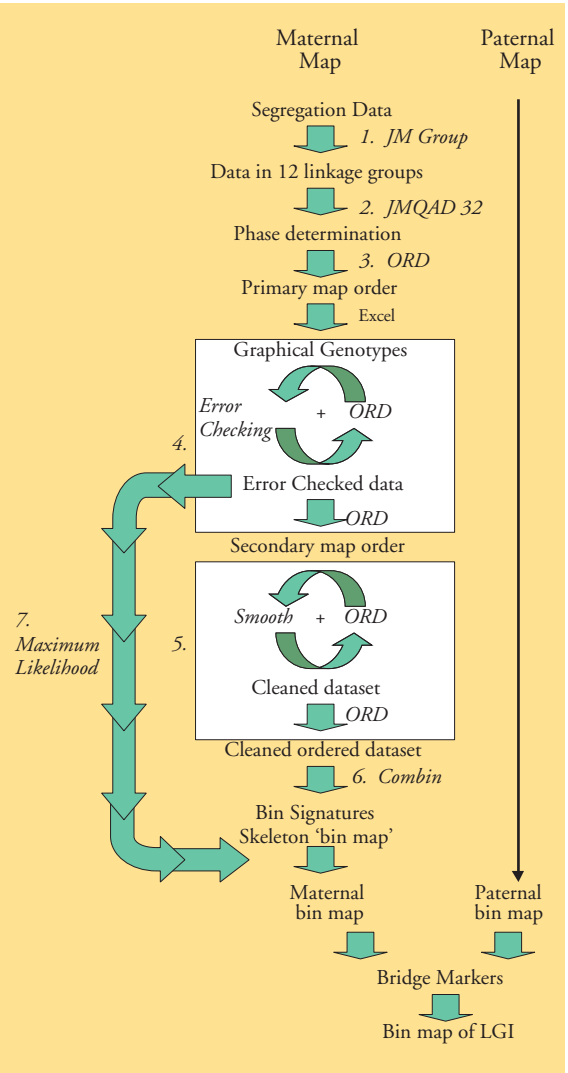

Figure 2 Steps involved in the creation of the 1260 point linkage map of chromosome 1.

The fact that each erroneous datapoint can introduce two false recombination events (a so-called single marker double recombinant) means that there exists the potential for 2000 false recombination events to be introduced into the dataset. This is an order of magnitude greater than the total number of recombination events expected in a population of 100 individuals, assuming 1-2 crossovers per chromosome. The consequence of using such data in currently available mapping software is the generation of vastly inflated maps with nonsensical marker orders.

From the above, we concluded that there are two pivotal requirements for creating UHD genetic linkage maps. The first of these is a system for rigorously and systematically identifying and correcting errors in the marker segregation data. The second is the development of a mapping model that allows the use of the most reliable data to calculate a framework map into which the remaining data can be fitted without upsetting the model. An important factor in this strategy is the fact that, in very large segregation datasets, the most reliable data is easily identified as that for which there is a level of redundancy. Multiple co-segregating markers reinforce the confidence in the accuracy of the shared segregation pattern of those markers. One potential model for a UHD map is the generation of a robust linear map, consisting largely of 'bins' of co-segregating markers and any non-redundant markers which can be incorporated into the linear map without conflict. Anomalous markers not resolved by error checking can be placed subsequently in the bin into which they fit best, without perturbing the overall map order. Thus, markers in a bin fit either perfectly, in complete agreement with the bin signature, or deviate from the bin signature by a number of recombination events. Another advantage of this model is that the quality of the resulting map can be verified by assessing the overall proportion of data that fits well into the model.

In order to achieve the UHD map of potato, we have deployed 400 AFLP primer combinations on 130 individuals of the F_1 progeny of a cross between two highly heterozygous diploid *Solanum tuberosum* genotypes, referred to as SH (maternal parent) and RH (paternal parent). To explore the implementation of the above model, an interim dataset of 6756 segregating markers generated by 234 primer combinations was analysed. Due to concerns about genome coverage, three restriction enzyme combinations, differing in the rare (6bp) cutting enzyme were used. As a result, 1278 of the markers were *PstI/MseI* based AFLPs, 1759 were *SacI/MseI* based, and 3719 were *EcoRI/MseI* based. To facilitate chromosomal identification, segregation of a small set of previously mapped SSR markers was also analysed in the population.

The segregation data were divided into maternal (genotype: abxaa, 2682 markers), paternal (genotype: aaxab, 2223 markers) and biparental (genotype: abxab, 1851 markers) datasets. Step 1 (see Fig.2) in the process was the use of the GROUP function of the mapping package JoinMap v2.0 to split the marker segregation data into 12 linkage groups corresponding to the 12 chromosomes of the potato genome. Chromosomal identities were assigned to the linkage groups on the basis of the SSRs and locus specific AFLP markers. The linkage group identified as Chromosome 1 was chosen to illustrate the principles

outlined above because it was the most extensive linkage group, containing a total of 1260 markers (627 maternal, 420 paternal and 213 biparental). The maternal and paternal datasets of Chromosome 1 were subsequently subjected separately to the process outlined in Figure 2.

Steps 2 and 3 involve determination of marker phase and map order using the JMQAD32 function of JoinMap and a newly developed programme called ORD, which calculates the marker order using an algorithm that minimises the number of recombination events. To identify potential errors, marker segregation data were sorted into a primary map order calculated by ORD and displayed as colour coded graphical genotypes in an Excel spreadsheet (Step 4, Fig. 2). Graphical genotypes are a representation of the recombined parental chromosomes in the progeny, and allow the identification of individual marker datapoints acting as single marker double recombinants (singletons). These singletons are potential marker scoring errors, and once identified, can be rechecked on the original AFLP autoradiograms, and corrected if necessary. This was done once for the entire dataset, and the marker order was recalculated in ORD using this more accurate data. In theory, this process could be repeated several times, but in practice, its time consuming nature allowed only two iterations, producing an improved secondary map order. Removal of remaining singletons was automated using a computer programme called SMOOTH (Step 5, Fig 2) which institutes an algorithm that calculates the probability of each marker datapoint being 'true' on the basis of flanking markers in the secondary map order. Markers that are not supported by observations at flanking markers are replaced by missing values, and a new marker order is again calculated using ORD. This process was repeated several times, gradually decreasing the stringency threshold allowing markers to be nominated as singletons. This cleaned ordered data set was then used to construct maternal and paternal maps of Chromosome I using a programme called ComBin (Step 6, Fig. 2). In ComBin, co-segregating markers are placed in `bins` to remove redundancy in the data set. Subsequently, the bins are 'threaded' like beads on a string (i.e. linearly organised), with adjacent bins differing by a single recombination event. When two adjacent bins were separated by more than one recombination event, a number of empty bins equal to the number of recombination events separating the markers were placed on the 'skeleton bin map'.

The accuracy of the skeleton bin map was verified by fitting the original marker data (after data checking but before cleaning with SMOOTH) into the skeleton bin map on the basis of the highest LOD score between markers and the bin signatures (Step 7, Fig. 2). Chromosome I consists of 90 maternal bins and 93 paternal bins (Fig. 3). The 627 maternal markers fitted into 66 bins, leaving 24 bins empty. The 420 paternal markers fitted into 49 bins, leaving 44 bins empty. The 210 biparental markers and three SSR loci were used to link the two parental maps as bin bridges (again on a maximum likelihood basis), giving a final map of 1260 markers. We estimated a residual singleton rate of 1-3% per marker per primer combination after two rounds of graphical genotype checking. Thus, we chose a threshold of a 3% deviation from the bin signature to determine whether markers fit well into the bins. Overall, 75% and 80% of the maternal and paternal markers respectively fit into bins within a range of 0 to 3% recombination, indicating that the model holds up well for the majority of the data. As the remaining 22.9% of markers outside the threshold do not perturb the map order, they can be retained in the dataset. This is an important aspect

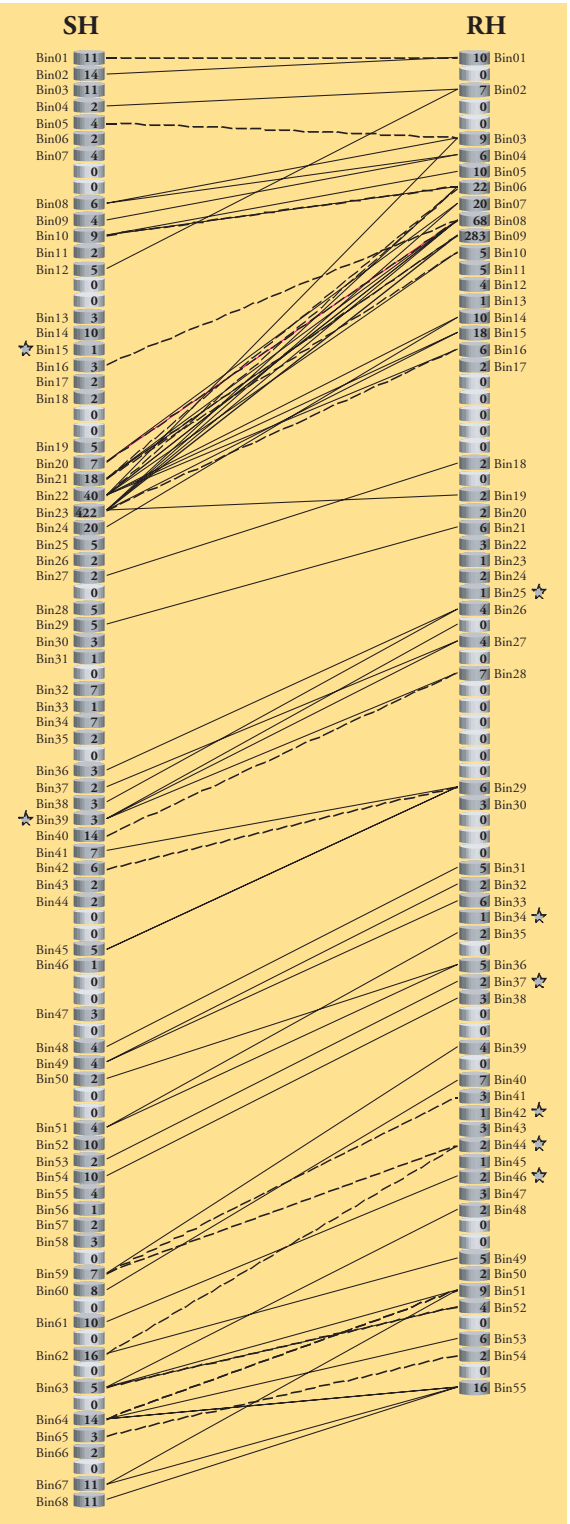

Figure 3 Skeleton bin map of chromosome I of potato. SH and RH are the maternal and paternal maps respectively. Lines between the two maps represent the allelic bridges (<ab x ab> markers and SSRs) between the two individual maps. Number in each bin is the number of 1:1 markers.

of the strategy as subsequent studies may be able to use (or even resolve the true position of) these markers. As well as residual scoring error, technical and biological phenomena such as PCR mispriming, co-migrating independent bands and methylation may all contribute to markers fitting badly into bins. The last of these is a testable hypothesis, since if methylation was responsible for markers not fitting well into bins, we would expect to observe a higher proportion of badly fitting *PstI* markers in comparison to *SacI* or *EcoRI*, due to the methylation sensitivity of *PstI*. We therefore compared markers that deviate from the bin signature at greater than 3% recombination. Approximately double the proportion of *PstI* markers are observed in this class (30%) compared to *EcoRI* and *SacI* markers (15%), indicating that methylation does play a role.

Both gaps (empty bins) and significant clustering (bins containing many markers) are evident on the map (Fig 3). The presence of empty bins might be considered surprising on such a high density map and probably represent regions with high levels of recombination or an absence of polymorphism. The largest bin in both parental maps contains approximately 50% of the markers, and this probably represents an area of suppressed recombination around the centromeric regions observed in many maps. Interestingly, when the distribution of the three enzyme combinations is analysed independently, the centromeric clustering is far less pronounced for *PstI* based markers compared to the *EcoRI* and *SacI* based markers. This is probably due to the fact that the methylation sensitivity of *PstI* favours the targeting of these markers to under-methylated, euchromatic (gene-containing) regions which tend to be located toward the ends of the chromosome.

In conclusion, we have developed a mapping model that will allow the rationalisation of 10,000 segregating markers into an ultra high-density genetic linkage map of the potato genome, resulting in the densest genetic map of any crop plant species to date. Unlike previous genetic linkage maps,the model allows the assessment of the robustness of any marker on the map by virtue of how well it fits the bin in which it has been placed. We are currently developing strategies that will allow the deployment of this map as a generally applicable resource for several uses including rapid local physical mapping, positional cloning, and development of markers for accelerated breeding programmes.