Toward a Marker-Dense Meiotic Map of the Potato Genome: Lessons From Linkage Group I

Edwige Isidore,*,[1] Hans van Os,†,[1] Sandra Andrzejewski,‡ Jaap Bakker,§ Imanol Barrena, Glenn J. Bryan,* Bernard Caromel,‡ Herman van Eck,† Bilal Ghareeb,‡,[2] Walter de Jong,*,[3] Paul van Koert,§,[4] Véronique Lefebvre,‡ Dan Milbourne,*,[5] Enrique Ritter,** Jeroen Rouppe van der Voort,§,[4] Françoise Rousselle-Bourgeois,‡ Joke van Vliet†,§ and Robbie Waugh*,[6]**

**Genome Dynamics Programme, Scottish Crop Research Institute, Dundee DD2 5DA, United Kingdom, †Laboratory of Plant Breeding, Wageningen University, 6700 AJ Wageningen, The Netherlands, ‡Station de Génétique et Amélioration des Fruits et Légumes, Institut National de la Recherche Agronomique, 84143 Monfavet Cedex, France, §Laboratory of Nematology, Wageningen University, 6709 PD Wageningen, The Netherlands and **NEIKER, E-01080 Vitoria, Spain*

Manuscript received June 27, 2003
Accepted for publication August 20, 2003

ABSTRACT

Segregation data were obtained for 1260 potato linkage group I-specific AFLP loci from a heterozygous diploid potato population. Analytical tools that identified potential typing errors and/or inconsistencies in the data and that assembled cosegregating markers into bins were applied. Bins contain multiple-marker data sets with an identical segregation pattern, which is defined as the bin signature. The bin signatures were used to construct a skeleton bin map that was based solely on observed recombination events. Markers that did not match any of the bin signatures exactly (and that were excluded from the calculation of the skeleton bin map) were placed on the map by maximum likelihood. The resulting maternal and paternal maps consisted of 95 and 101 bins, respectively. Markers derived from *Eco*RI/*Mse*I, *Pst*I/*Mse*I, and *Sac*I/*Mse*I primer combinations showed different genetic distributions. Approximately three-fourths of the markers placed into a bin were considered to fit well on the basis of an estimated residual "error rate" of 0–3%. However, twice as many *Pst*I-based markers fit badly, suggesting that parental *Pst*I-site methylation patterns had changed in the population. Recombination frequencies were highly variable across the map. Inert, presumably centromeric, regions caused extensive marker clustering while recombination hotspots (or regions identical by descent) resulted in empty bins, despite the level of marker saturation.

MARKER-dense meiotic linkage maps are valuable tools in fundamental and applied genetic research. They serve multiple purposes ranging from the dissection of simple and complex phenotypes to the isolation of genes by map-based cloning (Tanksley *et al.* 1995). Marker-dense maps provide ordered frameworks for the construction of physical maps onto which yeast artificial chromosome or bacterial artificial chromosome (BAC) contigs can be anchored (Klein *et al.* 2000). Thus, construction of a high-density genetic map was one of the first goals of the human (Murray *et al.* 1994; Dib *et al.* 1996) and mouse (Dietrich *et al.* 1996) genome mapping projects. In crop plant species such as rice (Harushima *et al.* 1998: 2275 markers), maize (Vuylsteke *et al.* 1999: 1539 and 1355 markers mapped in two populations), and potato and tomato [Tanksley *et al.* 1992 (~1000 markers) and Haanstra *et al.* 1999 (1175 markers), respectively], high-density genetic linkage maps have already been constructed. The combined maps of the tomato and potato genomes are composed of ~1000 restriction fragment length polymorphism (RFLP) markers assembled from several populations and together they represent an average spacing of ~1.2 cM (Gebhardt *et al.* 1991; Tanksley *et al.* 1992).

With the objective of constructing a 10,000-point marker-dense meiotic map of the potato genome as a platform for map-based gene isolation and for the construction of a genetically anchored whole-genome physical map, we have assembled an interim data set composed of >6500 independent PCR-based segregating markers from a diploid mapping population. Interpreting this data set in the context of linkage analysis proved problematic because, as the number of markers included in the experiment increased above a given threshold, computation-

[1]These authors contributed equally to this work.

[2]*Present address:* Arab American University, P.O. Box 240, Jenin, Palestine.

[3]*Present address:* Department of Plant Breeding, Cornell University, Ithaca, NY 14853.

[4]*Present address:* Keygene N.V., 6700 AE Wageningen, The Netherlands.

[5]*Present address:* Plant Biotechnology Centre, TEAGASC, Carlow, Ireland.

[6]*Corresponding author:* Genome Dynamics, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, United Kingdom. E-mail: rwaugh@scri.sari.ac.uk

ally intensive mapping algorithms, based on the use of pairwise distances between loci to derive marker order, became slow and eventually failed. Here we present the results and the challenges that we encountered when analyzing data from the largest single linkage group in our experiment, linkage group I (LG I), which contains 1260 markers.

Meiotic linkage mapping uses the frequency of recombination events that occur during meiosis as a basis for calculating genetic distances between loci. The observed recombination frequencies are commonly converted into map units (centimorgans) by applying a mapping function, which imposes certain assumptions on the data (*e.g.*, the presence or absence of "interference"; Kosambi 1944). On the basis of several populations (*e.g.*, Bonierbale *et al.* 1988; Gebhardt *et al.* 1991; van Eck *et al.* 1995; Collins *et al.* 1999), the cumulative length of the potato genetic map is ~600–1100 cM, with the 12 individual chromosomes ranging from ~40 to >100 cM. These map lengths are consistent with cytological observations that indicate the formation of, on average, less than one chiasma per bivalent during meiosis. Thus, we anticipate that during meiosis a given potato chromosome will generally be engaged in a single recombination event, with none or more than one occurring less frequently.

By following the inheritance of genetic markers in a meiotic mapping population, recombination events can be linearly ordered along each chromosome. This linear order defines intervening segments of chromosomes, which vary in both physical and genetic size. These variables are largely defined by the number of descendants in the mapping population and by the average number of recombination events that occur during meiosis. Clearly, as the number of markers scored in the population exceeds the number of recombination-defined chromosomal segments, some segments will be identified by multiple cosegregating markers. When a very large number of markers have been followed, this will occur frequently, resulting in many chromosomal segments being multiply marked (Figure 1). We call these chromosomal segments cosegregation bins. A cosegregation bin has a bin signature, that is, the consensus segregation pattern of all markers in that bin. It is the number of recombination events in the population, not the number of markers, that defines the maximum number of bins in a chromosome in a given experiment. Adjacent bins should be separated by a single recombination event. However, in practice, multiple recombination events occur frequently between adjacent bins and as a result all theoretical bins cannot be identified directly from the data. This situation could arise from, for example, chromosomal segments being either "identical by descent" or simply physically small. Here, segregation data from the adjacent filled bins are sufficient to calculate the minimum number of intervening recombination events. Once established, empty bins can be inserted between filled ones until the chromosome is represented as a linear string of bins, each separated by a single recombination event.

While achievable in principle, one overriding practical reality—error—complicates the construction of a marker-dense bin map. Erroneous data introduce conflict between the true and the observed number of recombination events. The significance of this can be illustrated by considering the creation of a meiotic linkage map of a single chromosome consisting of 1000 markers in a population of 100 individuals and a marker scoring accuracy of 99%. Because each erroneous data point can introduce two false recombination events (a single-marker double recombinant), the potential exists for 2000 false recombination events to be introduced into the data set. This is an order of magnitude greater than the total number of recombination events expected in a population of 100 individuals, assuming one to two crossovers per chromosome. The consequence of analyzing such data with any mapping software is the generation of inflated maps with tenuous and potentially erroneous marker orders.

We conclude that there are two pivotal requirements for creating marker-dense meiotic maps. The first is a system for rigorously and systematically identifying and correcting errors in the marker segregation data. While this will make improvements, identification of all errors in a large data set will be impossible. The second requirement, therefore, is the development of a mapping model that identifies and makes use of the most reliable data to calculate a framework map into which the remaining data can be placed. The most reliable data are likely to be those for which redundancy, revealed as multiple cosegregating markers from independent experiments, improves confidence and provides support for the hypothesis that the shared segregation pattern is in fact "true," assuming random, not systematic, error. We explore a model that generates a robust linear map consisting of bins of cosegregating markers and nonredundant markers if they are incorporated without conflict. All other markers are subsequently placed in the bin into which they best fit by statistical procedures without perturbing the overall map order.

MATERIALS AND METHODS

Plant material: A diploid F_1 potato population of 130 individuals was used for the construction of the genetic map. This mapping population was generated from a cross between two diploid heterozygous parents: SH83-92-488 (hereafter denoted SH) × RH89-039-16 (hereafter denoted RH) (Rouppe van der Voort *et al.* 1997a). Genomic DNA isolation was performed on frozen leaf tissue as described by van der Beek *et al.* (1992).

Marker assays: The amplified fragment length polymorphism (AFLP) procedure of Vos *et al.* (1995) was used with minor modifications. Three restriction enzyme combinations were used to prepare template DNA: *Eco*RI/*Mse*I, *Pst*I/*Mse*I, and *Sac*I/*Mse*I. After digestion, adapters corresponding to each

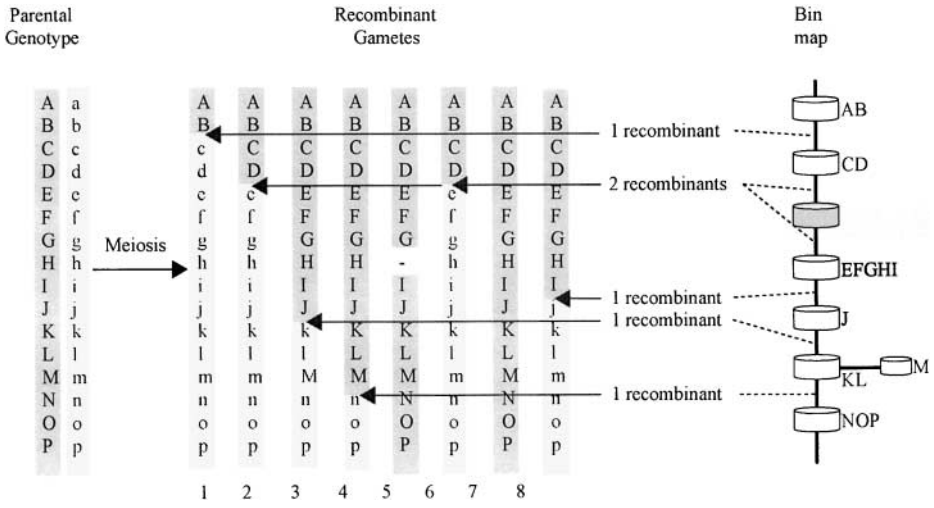

FIGURE 1.—The recombination bin-mapping concept. For simplicity, only one heterozygous parental chromosome pair and eight potential recombinant gametes are shown. Allelic marker loci, Aa to Pp, are represented as upper- or lowercase letters on a white or shaded background. During meiosis, recombination breaks and rejoins the parental chromosomes, which segregate into gametes that are a mosaic of the parental chromosomes. The diagram illustrates how the position of the recombination events can be visualized as a linearly ordered set of bins, each separated by a single recombination event. In this example, six of the eight parental gametes have undergone recombination. In the marker data set, gamete 3 contains a singleton (M), which on analysis is hypothesized as being unlikely on the basis of the genotype of the flanking markers because it introduces two additional recombination events. In the final map, marker M, however, is placed on the map in bin 6 at a distance of one apparent recombination event from the core. In gamete 5, a missing data point (–) is hypothesized as being H on the basis of the flanking marker data and, as a result, fits into bin 4. In gametes 2 and 6, recombination has occurred between the same two marker loci (Dd and Ee), resulting in the insertion of an empty bin (shaded) in the map. The resulting bin map is therefore composed of seven linear bins with a side branch from bin 6, which contains marker M at an apparent recombination distance of 1. At high marker density (*i.e.*, when the number of markers is much greater than the number of recombination events), individual bins will contain multiple-marker loci (as illustrated for five of the seven bins). All marker loci in a bin either have identical segregation patterns (*i.e.*, the bin signature) or deviate by a defined number of apparent recombination events (*e.g.*, M).

enzyme cleavage site were ligated to the restricted DNA. Their sequences are as follows: *Eco*RI (5′-CTCGTAGACTGCGTACC-3′/3′-CTGACGCATGGTTAA-5′), *Pst*I (5′-CTCGTAGACTGCGTACATGCA-3′/3′-CATCTGACGCATGT-5′), *Sac*I (5′-CTCGTAGACTGCTACAAGCT-3′/3′-CATCTGACGCATGT-5′), and *Mse*I (5′-GACGATGAGTCCTGAG-3′/3′-TACTCAGGACTCAT-5′). Preamplification of the restricted-ligated fragments was performed using primers strictly complementary to the adapters. For selective amplification, the primers had the common adapter sequence plus a 2- or 3-bp extension at their 3′ end. The 234 selective AFLP primer combinations used in this study are tabulated at http://www.dpw.wageningen-ur.nl/uhd/index.html. *Eco*RI (E), *Pst*I (P), or *Sac*I (C) primers were 5′ end labeled with [γ-^{33}P]ATP as described by Vos *et al.* (1995) prior to selective amplification. Amplification products were separated on 5% polyacrylamide, 1× TBE sequencing gels. Buffer at the anodal side was supplemented with 0.5 M NaOAc to create an ionic gradient, which allowed better separation of the larger fragments. Gels were run at 110 W (constant power) for 3 hr. After drying the gels, amplification products were visualized by autoradiography. Three chromosome I-specific microsatellites (STM1049, STM1029, and STM2020) were used to test the integrity of the population under study and to align the two parental maps. Simple sequence repeat (SSR) primer sequences and assay conditions were as described in MILBOURNE *et al.* (1998). Autoradiograms were scanned at a resolution of 150 dpi and scored using the computer package Cross-Checker (available at http://www.spg.wageningen-ur.nl/pv/pub/CrossCheck/), and the scores were manually checked by comparing them with the primary autoradiographs.

Marker nomenclature: Band nomenclature was assigned from reference autoradiograms, which were provided by Keygene NV, Wageningen, The Netherlands. The marker names indicate the enzyme used, the primer combination, and the mobility of the fragment as defined by a size marker (Sequamark 10-bp ladder; Research Genetics, Huntsville, AL). Decimal points in the mobility values (*e.g.*, PAC/MAGA: 120.5) are due to interpolation of band sizes between 10-bp markers by the proprietary software used.

Mapping algorithms: A combination of existing JoinMap V2.0 modules (JMGRP32 and JMQAD32), new algorithms (RECORD and SMOOTH), and recently developed software (ComBin) were used to analyze the segregation data.

JMGRP32: This module within the JoinMap V2.0 software package (STAM 1993) allows the grouping of markers that belong to the same linkage group. The largest group of markers, significantly distinct from other marker groups (at LOD 6) representing LG I, was exported and analyzed with JMQAD32.

JMQAD32: This *q*uick *a*nd *d*irty module within the JoinMap package calculates recombination frequencies between marker loci. The best map is selected from all possible orders on the basis of minimization of the sum of adjacent recombination frequencies. In general, these maps are inflated, and the extra length is best understood by assuming double recombination events or scoring errors (STAM 1993; STAM and VAN OOIJEN 1995).

RECORD: RECORD finds the best possible marker order by minimization of the number of recombination events as counted in a data set of marker segregation data. In contrast to JoinMap or MapMaker, this algorithm does not make use of many pairwise distance estimates, but it uses the much simpler raw segregation data. Simulations showed that the performance of RECORD is particularly good in marker-dense regions, as well as with any level of missing values and scoring errors (up to 20%) where software packages based on pairwise distance estimates encounter severe difficulty (VAN OS *et al.* 2000).

SMOOTH: SMOOTH identifies and removes singletons from genetic mapping data sets. Once a preliminary marker order has been proposed (*e.g.*, by RECORD), SMOOTH calculates the probability that each data point of a segregating marker locus is true on the basis of the genotype of flanking markers. The probability calculation is based on 15 flanking

data points on either side, with the nearest data points being given a higher weighting. SMOOTH is applied in conjunction with RECORD by cyclically reiterating the process of marker ordering and singleton removal. Initially, a strict probability threshold of $P < 0.01$ is used to eliminate the least-well-supported data points. The marker order is then recalculated (with RECORD) and further weakly supported data points are removed by SMOOTH by releasing the threshold by $P = 0.01$ over 30 cycles until a threshold of $P = 0.3$ is reached. The process of removing conflicting data points and recalculating the marker order is continued until no further poorly supported inconsistent data points (*i.e.*, singletons) can be identified. Simulation studies have demonstrated that a significant increase in the accuracy of marker order is obtained with the combined use of RECORD and SMOOTH without the risk of introducing artifactual marker orders (H. van Os and H. van Eck, unpublished results). The software is relatively insensitive to high levels of noise, as observed in extensive marker data sets as used here.

ComBin: ComBin differs from existing mapping software as maps are built by placing markers (or bins of cosegregating markers) next to each other, separated by a single recombination event (Buntjer *et al.* 2000; available at http://www.dpw.wau.nl/pv/pub/combin/index.htm). This process resembles threading beads on a string. The marker bins within the developing string are used to identify the next marker (or bin) at a distance of one recombination event. The software allows the formation of side branches when adding the next marker to the developing string and as a result facilitates the visualization of singletons or other ambiguities in the data set. Here, ComBin was used to inspect the data for secondary structures in the linkage groups, while calculating the skeleton bin map.

RESULTS

Genome-wide segregation data: Using a population of 130 individuals, 234 AFLP primer combinations were used for selective amplifications. This generated a total of 6756 clear and scorable segregating bands composed of 1759 *Sac*I/*Mse*I, 3719 *Eco*RI/*Mse*I, and 1278 *Pst*I/*Mse*I AFLP markers. As the population was derived from a cross between two noninbred parental lines, the 6756 markers (and three multiallelic SSR markers) were first separated into maternal, paternal, and biparental data sets according to the parental profiles of each band scored in the population. A total of 2682 (39.7%) were heterozygous in the female parent (coded ab × aa for analysis), 2223 (32.9%) in the male parent (coded aa × ab), and 1851 (27.4%) were heterozygous in both parents (coded ab × ab and from here on referred to as bridge markers).

The GROUP function of JoinMap V2.0 split the maternal data into the expected 12 linkage groups at LOD 6.0. For the paternal data, at LOD 6.0 the markers in linkage groups corresponding to chromosomes II–XI were separated. However, one linkage group was obtained, which contained markers from LGs I and XII and was split only when the LOD was raised to 12. At these thresholds, a group of 11 highly skewed markers remained unassigned. Assignment of parental linkage groups to chromosomes and chromosome orientation was achieved unequivocally on the basis of common AFLP markers mapped previously in the same population (Rouppe van der Voort *et al.* 1997a,b), which form part of a catalog of locus-specific AFLP markers (Rouppe van der Voort *et al.* 1997c). Finally, the bridge markers were assigned to linkage groups at LOD 8.0 by analysis with the maternal and paternal data sets separately. Being less informative, some of the bridge markers exhibited spurious (multiple) linkages to different maternal and paternal linkage groups and were therefore excluded from the data set. After this analysis, 282 markers (4.17% of the 6756) remained unassigned. LG I was the largest linkage group, containing a total of 1260 markers (627 maternal, 420 paternal, and 213 bridge). The identity and correspondence of the maternal LG I and paternal LG I were confirmed by use of three genetically characterized multiallelic LG I-specific SSRs (Milbourne *et al.* 1998). The remaining analysis focuses on only this linkage group with the objective of deriving an optimal marker order.

Map construction: In populations derived from non inbred parents, a necessary step after grouping the marker data into linkage groups is to determine marker phase. Phase information is required to convert data from non inbred parents into BC1 format for further analysis. This was achieved using the JoinMap V2.0 module JMQAD32 (Stam and van Ooijen 1995). However, attempts to use the standard modules in JoinMap V2.0 to subsequently order the markers were unsuccessful (the program crashed). We therefore applied the following map construction process.

Primary marker ordering and error checking: The raw data were analyzed initially with RECORD. As RECORD is input order dependent, the stepwise map construction process was repeated 10 times and the shortest resulting map was assumed to be the most correct. Generally, the shortest map will be one from a number of equally likely potential solutions (*i.e.*, it is not perfect). However, simulation studies show that RECORD is computationally less demanding, faster, and less sensitive to missing observations and scoring errors than JMMAP, especially in small populations and in regions with high marker density (H. van Os and H. van Eck, personal communication).

On the basis of the output order from analysis with RECORD, singletons and other potential errors in the marker segregation data were identified by visual inspection of graphical genotypes of each of the progeny and then rechecked on the original AFLP autoradiograms and corrected when necessary. This was performed once on the complete data set after which a new map order was calculated using RECORD. This whole process was considered too time consuming to repeat fully, so in a subsequent round, only markers containing two singletons or more (on the basis of graphical genotypes derived from the new map order) were checked manually again, corrected if necessary, and a new order was calculated. These two rounds of data checking allowed a significant

improvement of the data quality as the singleton rate for each primer combination decreased from >5% to <3% on the basis of inspection of graphical genotypes. As a general observation, for a given restriction enzyme digest, primer combinations that generated complicated fingerprints (*i.e.*, >80 bands per lane) on analysis tended to reveal a higher frequency of singletons.

Automated singleton removal: Remaining singletons were removed and replaced automatically with missing values through an iterative process of repeatedly calculating the marker order with RECORD and replacing potential errors with "missing data" using SMOOTH, starting with a strict probability threshold for singleton removal of $P < 0.01$ and slowly releasing it over 30 cycles to $P < 0.3$. A final order was then calculated with RECORD. Such iterative use of SMOOTH is not harmful to the map order although, occasionally, rejecting the hypothesis that a singleton was "true" may cause adjacent bins to merge (the equivalent of removing a recombination event from the population). No singletons remained in our data set when the threshold was relaxed to $P < 0.3$.

Production of the skeleton bin map: The cleaned data set was then used to construct maternal and paternal maps of LG I using ComBin (Buntjer *et al.* 2000). ComBin complements SMOOTH by identifying certain data ambiguities, such as multiple markers containing an identical singleton. These would not be identified by SMOOTH, as the shared singletons jointly support each other. Visual inspection of our data indicated that this was the case for many of the markers placed in side branches. These shared singletons were then replaced by missing values until a linear string of bins was obtained. We call the resulting linear map the skeleton bin map. 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 flanking filled bins were placed in the skeleton bin map. Bin signatures were derived from the most complete marker (in terms of genotypic information) incorporated in the bin.

Populating the skeleton bin map: The skeleton bin map is effectively a minimum tiling path of recombination events along a chromosome. It was populated retrospectively by fitting the original marker data (*i.e.*, error-checked data before the removal of singletons by SMOOTH) on the basis of the highest LOD score between individual markers and bin signatures. Inspection of markers in a bin confirmed that the apparent recombination distance between markers and their bin signature was mainly due to singletons. Populating the skeleton bin map did not result in a change in the order of the bins and allowed discrimination between distance due to true recombination and to potential error. After populating the skeleton bin map of both parents, the bridge markers were mapped. All possible putative bridge bins of this linkage group were generated by superimposing all maternal and all paternal bin signatures in coupling and repulsion phase (cc, cr, rc, rr). Subsequently, the observed bridge marker data were analyzed against the postulated bridge bin signatures. The bridge markers were then placed into the putative bridge bins on the basis of the highest LOD score.

Bin map of potato linkage group I: LG I consists of 95 maternal bins and 101 paternal bins. The 627 maternal markers fitted into 72 bins, leaving 23 bins empty. The 420 paternal markers fitted into 48 bins, leaving 53 bins empty. The smaller number of segregating markers from RH indicates that it is more homozygous. As a result, the higher proportion of empty bins was not unexpected. The 210 markers segregating in both parents and the three SSR loci were used to link the two parental maps as bin bridges, giving a final map of 1260 markers. In Figure 2 both parental skeleton bin maps are represented, showing the number and type of markers in each bin. Figure 2 does not display distance between markers in map units (centimorgans) or recombination values that are independent of population size, but shows the actual number of recombination events between two markers as observed in these 130 genotypes. The bridge markers reveal minor discontinuities in the order of the parental bins into which they best fit (data not shown). We consider this to be a direct consequence of our inability to clean the biparentally inherited data of errors based on graphical genotypes or SMOOTH and the highly skewed nature of the loci on the top third of the parental map. The detailed map, including complete names of all the markers in each bin, is available at http://www.dpw.wageningen-ur.nl/uhd/index.html.

Surveying graphical genotype images from the skeleton bin map revealed that 55/130 SH and 44/130 RH parental chromatids had not recombined, 57/130 SH and 72/130 RH parental chromatids had undergone a single recombination event, and 18/130 SH and 14/130 RH had undergone two recombination events, respectively, during meiosis. No chromosome had more than two recombination events and no singletons remained. There was significant segregation distortion from a 1:1 ratio in the paternal map from bins 1–27 up to a chi-square value of 27.7. No segregation distortion was observed in the maternal map.

Marker distribution: The AFLP markers are not evenly distributed along the genetic map of LG I. On the paternal bin map, there are two gaps of seven recombination events (*i.e.*, six empty bins) and two gaps of six recombination events. This is surprising, given the number of markers on this paternal chromosome, but may reflect either a high level of meiotic recombination in these regions (recombination hotspots) or an absence of polymorphism. There is also significant clustering of markers in single bins for each parental map. For instance, the biggest bins, no. SH032 of the maternal map and no. RH013 of the paternal map, contain 353 and 265 markers, respectively!

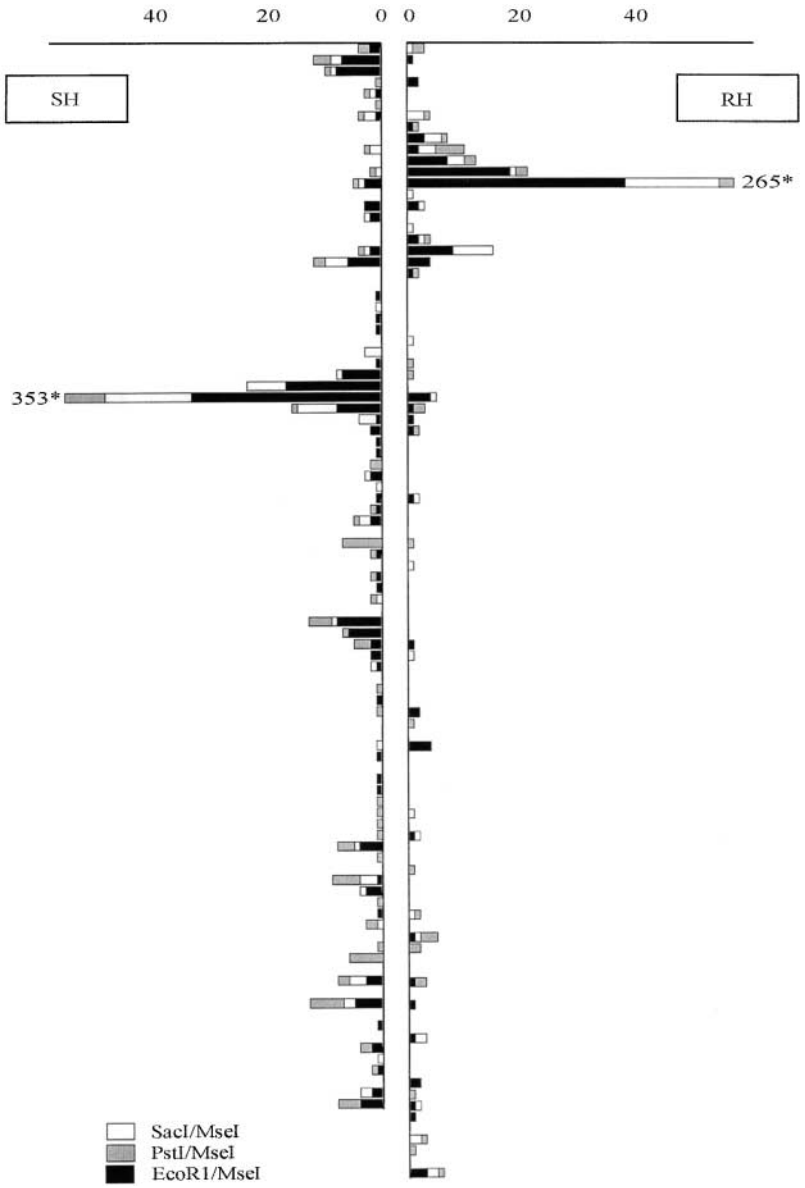

FIGURE 2.—Final bin maps of SH and RH showing marker number and composition of each bin. The SH and RH maps are composed of 95 and 101 bins, respectively. The histograms with asterisks representing bins SH032 and RH013 have been scaled to fit on the page with the total number of markers indicated. *Eco*RI-, *Pst*I-, and *Sac*I-derived markers in these bins are proportionally scaled.

The distribution of the three different types of AFLP markers is shown in Figure 2. The graphs show clustering of markers for all enzyme combinations in a short interval around the maternal bin SH032 and the paternal bin RH013. The biggest clusters are observed for *Eco*RI/*Mse*I and *Sac*I/*Mse*I, where 61–69% of the markers are located in a single bin of the maternal or paternal map. *Pst*I/*Mse*I AFLP markers are more evenly distributed along the chromosome, with 36 and 23% of the markers clustered in SH032 and RH013, respectively.

Map quality: Our original hypothesis was that a skeleton bin map would provide a high-confidence framework for the production of a marker-dense genetic linkage map. To check the quality of the skeleton bin map, we first examined how well the original marker segregation data fit into each of the bins. After placing markers by maximum likelihood, the apparent recombination value between the bin signature and the segregation data of each marker in the bin was graphically summarized. The apparent recombination value does not represent genetic distance, but rather represents a distance we describe as "perpendicular" to the linear axis of the map, caused by potentially erroneous or inconsistent data. The data incorporated into the final map are displayed in Figure 3, which summarizes the apparent recombination value of each marker in terms of the number of observed singletons, relative to its bin signature. A threshold value of 0.03 was chosen to discriminate between good and poorly fitting markers because, after two rounds of error checking using graphical genotypes, a residual singleton rate of 0–3% per marker per primer combination was estimated to remain. Overall, 74.8% of the maternal markers and 80.4% of the paternal markers fit into bins within an apparent recombination distance range from 0 to 0.03, effectively equivalent to markers scored with 0–3% error. Bins SH032 and RH013 are

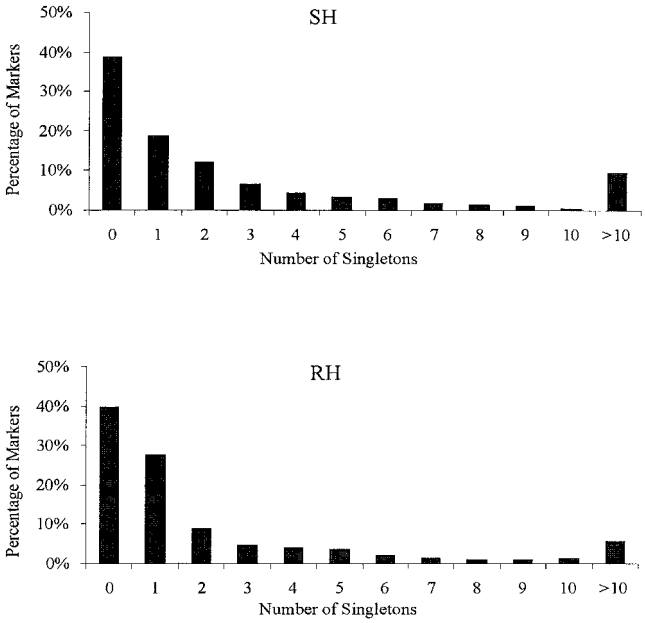

FIGURE 3.—The percentage of error-checked 1:1 markers (*y*-axis) that fit into the skeleton bin maps of SH and RH either exactly or by the indicated number of singletons (*x*-axis) from a bin signature is shown. Deviation in the marker segregation pattern from the bin signature is expressed as the actual number of inconsistent data points (*i.e.*, 1/130–>10/130).

shown in detail in Figure 4 because they provide good examples of marker behavior in a bin and because of the extremely high number of markers that they contain. For both, approximately half of the markers have a recombination value of 0, which means that their segregation pattern is identical to their bin signature. A total of 18.9% of the markers had an apparent recombination value >0.03 and are considered not to fit well in the bin into which they are placed (they are, however, retained in the total data set on the website listed above because they may be of some use in subsequent studies).

Second, a subset of the marker data was analyzed separately by JoinMap V2.0 and marker order and map length were compared to the bin map (data not shown). Overall the order was remarkably consistent between maps. Significant inflation was restricted to SH032 where the 30 markers chosen for analysis by JoinMap V2.0 were distributed over a 17-cM interval. The length of the maternal map was 88 cM *vs.* 95 bins and the paternal map 101 cM *vs.* 101 bins.

DNA methylation and singleton frequency: For many years *Pst*I has been used to isolate single- and low-copy genomic clones to use as probes for RFLP analysis (BURR *et al.* 1988) and we considered that a similar approach could be transferred to AFLP analysis. *Pst*I is effective for this because of its sensitivity to CpNpG methylation, which focuses its activity on hypomethylated regions of the genome, such as transcriptionally or biologically active euchromatic DNA. In contrast, *Eco*RI and *Sac*I are much less sensitive to cytosine methylation (*Sac*I is sensitive to GAGmCTC but not to GAGCTmC methylation). We therefore asked whether the origin of a proportion of the singletons in the data set was likely to be the result of the changing methylation status (at certain loci) of the DNA in different individuals in the population. Our hypothesis was that if methylation changes were responsible for markers not fitting well into bins, then the proportion of badly fitting *Pst*I/*Mse*I markers would be higher than that from other enzyme combinations. The relative frequencies of markers that deviate from the bin signature with an apparent recombination value of >0.03 therefore were compared for the different enzyme combinations employed. We found that approximately double the proportion of *Pst*I markers was observed in this category (30%) compared to *Eco*RI and *Sac*I markers (15%), suggesting that changing patterns of methylation are contributing to the "error" frequencies observed in the data. This finding prompted us to reexamine the *Pst*I autoradiographs because, if methylation were playing a role in error frequencies of segregating loci, we would also expect to see novel bands appearing in the population at low frequency, as previously methylated regions became susceptible to digestion with *Pst*I. By definition, these bands would not appear in the parental tracks and would not have been included in the data set used for linkage analysis. Such markers were

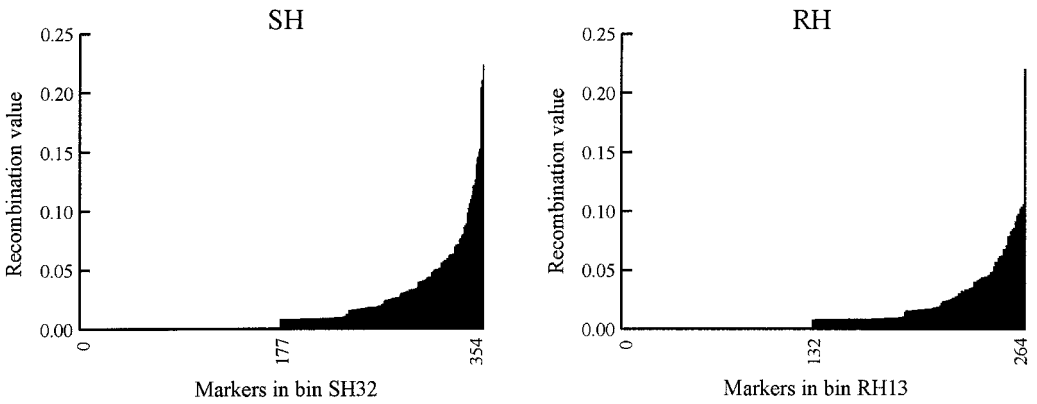

FIGURE 4.—Distance (represented as a recombination fraction) between the actual marker segregation pattern and the bin signature for the markers in the largest maternal (SH032) and paternal (RH013) bins. The markers are ordered from left to right according to their goodness of fit in the bin.

found in almost every *Pst*I primer combination in the population. They were not found on the *Eco*RI/*Mse*I or *Sac*I/*Mse*I autorads (data not shown).

DISCUSSION

In this report, we have presented the principles and approaches that we adopted to analyze 1260 segregating loci from potato LG I, the outcomes of these analyses, and their implications for our ultimate objective of accurately mapping ~10,000 AFLP markers across the entire potato genome. Our major challenge was to obtain an accurate marker order using a data set that contained errors, inconsistencies, and missing data (like all mapping studies). We initially considered that a logical strategy for map construction would be to identify cosegregating markers with complete data sets (*i.e.*, no missing data) and use this data to calculate an optimal bin map. The bin map would have a high degree of confidence attached to it because each of the marker scores would be effectively verified by the multiple representations in a bin. We could then fit incomplete or singly represented marker data sets into this robust framework. However, while in theory bins of cosegregating markers are easily definable, in practice a mixture of data error and, we hypothesize, biological phenomena, *e.g.*, methylation and demethylation, confound bin fitting. Such inconsistencies were revealed as individual marker data points that produce artifactual double recombinants in conflict with both the concept of interference and the flanking marker data (*i.e.*, singletons). Inconsistencies can be incorporated into lower-density maps without great impact. However, in a saturation-mapping scenario the result will be additional apparent recombinants and a loss of map linearity. Therefore, we applied an iterative process based on calculating marker order and replacing singletons with missing values on the basis of the flanking marker genotypes. The output was an ordered set of filled and empty bins, the latter inserted when adjacent filled bins were separated by greater than a single recombination event. Together, the filled and empty bins represent what we have termed the skeleton bin map. Under the assumption that the skeleton bin map was correct, its "accuracy" was then evaluated by assessing how well the error-checked raw marker data fit into the model (by maximum likelihood) and by comparing the map order of a subset of the data to an order obtained using JoinMap. The first assessment confirmed that the identification and replacement of singletons with missing values was a valid and effective approach that does not create artifacts in marker order. The second assessment revealed overall similarity between marker orders calculated using each approach. However, visual inspection of LG I graphical genotypes on the basis of the JoinMap order revealed a high incidence of multiple recombinants, which was at odds with our biological expectations. In contrast, in the bin map we found that 12.3% (32/260), 49.6% (129/260), and 38.1% (99/260) of the chromosomes had experienced 2, 1, and 0 recombination events, consistent with cytological observations of one or two chiasma per bivalent during meiosis (Sherman and Stack 1995).

It is impossible to distinguish between singletons that are scoring errors and singletons that are rare but true observations caused by biological phenomena such as double recombination, local DNA inversions, or methylation polymorphism. Initially, the finding of a higher percentage of singletons among *Pst*I-derived markers was surprising. *Pst*I cleaves plant DNA much less frequently than *Eco*RI and *Sac*I do, and as a result, AFLP profiles have fewer bands and greater clarity, making data collection easier and less prone to scoring error. A different genetic distribution of *Pst*I- and *Eco*RI-derived AFLPs has been documented previously (Young *et al.* 1999), but probably because of the marker density, combined with the way linkage maps have been constructed, there has been little direct evidence to suggest that methylation status has a significant impact on marker analysis in sexually derived segregating populations. However, such epigenetic variation would be relevant both in a high-density mapping scenario and when considering the link between genotype and phenotype, as shown in animals (de Koning *et al.* 2000), humans (Morison *et al.* 2001), Drosophila (Lloyd *et al.* 1999), and plants (Alleman and Doctor 2000). The population used here has a wide range of morphological and developmental variation, including dormancy break and time to maturity. Consequently, leaf material for DNA isolation was harvested from physiologically and developmentally contrasting carbohydrate "sink" or "source" leaves. If changes in methylation occur during this switch, it is possible that analysis of the DNA with a methylation-sensitive enzyme will result in the appearance or disappearance of marker bands used in genetic linkage experiments. This is not without precedent. Epigenetic differences have been detected by AFLP analysis of somatically regenerated plants from a number of species, including Arabidopsis (Polanco and Ruiz 2002), oilpalm (Matthes *et al.* 2001), and, of particular relevance here, somatically regenerated potato microplants exhibiting mature *vs.* juvenile leaf morphologies (Joyce and Cassells 2002). Furthermore, naturally occurring, heritable, differentially methylated epialleles at the *P1* locus have been shown to be responsible for conditioning altered kernel pigmentation in maize (Das and Messing 1994). It is therefore tempting to speculate that in populations such as those utilized in this study, epigenetic variation—revealed as changing methylation status at *Pst*I sites across the genome—contributes to the observed frequency of singletons and to other potential data inconsistencies.

Both gaps and severe clustering of markers were observed in the map. In Arabidopsis, clustering of *Eco*RI AFLP markers occurs around the centromeric regions of the chromosomes (Alonso-Blanco *et al.* 1998). Similar clustering of *Eco*RI markers around centromeres has

been observed in potato (van Eck *et al.* 1995) as well as in other plant species such as barley (Becker *et al.* 1995; Powell *et al.* 1997), soybean (Keim *et al.* 1997; Young *et al.* 1999), maize (Vuylsteke *et al.* 1999), and tomato (Haanstra *et al.* 1999). This clustering might reflect the low content of single-copy sequences present in pericentromeric regions. In Arabidopsis, these regions contain mainly repeated sequences of unknown function. An extra enrichment of AFLP markers in this region could be due to the use of *Eco*RI or *Sac*I combined with *Mse*I, which recognizes 5′-TTAA-3′ and therefore will cut more frequently in A + T-rich regions, such as pericentromeric heterochromatin [although this reason is not the case in soybean (Young *et al.* 1999)]. More likely, centromeric clustering is related to suppression of recombination because the markers based on *Eco*RI and *Sac*I differ in the CG content of their recognition site but target similar genomic regions.

Due to the population size, the map developed here may be marker dense, but it remains low resolution because the number of individuals effectively defines the total number of recombination events upon which the map can be based. It is further limited by the finding that over half of the markers fall into two bins: one on the maternal and one on the paternal map. The remainder of the map is represented by a combination of filled and empty bins. As a result, the utility of the information to address our original objective of linking genetic and physical maps using an approach broadly similar to that described recently for sorghum by Klein *et al.* (2000) is somewhat compromised, but nonetheless remains an overall valid strategy. In parallel with the development of a marker-dense genetic map, we have constructed BAC libraries of both parental clones and developed a pooling strategy, which allows the identification of individual BACs by screening with AFLPs. This approach currently allows the identification of BACs and BAC contigs while it simultaneously assigns their chromosomal location (G. Bryan, personal communication). However, it should be stated that the logistical problems of adopting this approach for a whole genome are considerable. In the current experiment, 33,000 lanes of AFLP products (254 combinations × 130 individuals) were run to collect the segregation data to construct the marker-dense linkage map and a similar or greater number would have to be run on BAC pools (depending on library size and pooling strategy) to connect the physical and genetic maps. This is equivalent to the number of lanes required to obtain individual clone fingerprint information of a more than sixfold genome coverage BAC library, assuming an average insert size of 150 kb and a potato genome size of 800 Mbp, which, it could be argued, would be more robust and provide an archive of genomic information. Thus, while the approach advocated by Klein *et al.* (2000) for linking physical and genetic maps is feasible in principle, it will require a massive effort that will be compromised by the types of data errors and inconsistencies described in this report. Even if the inconsistencies were discounted, assuming the LG I information extends to other chromosomes, we would expect the majority of BACs to fall into the centromeric bins on each of the 24 chromosomes. As a result, we will fail in our objective of determining an order, which will *de facto* require a complementary approach such as high-throughput individual BAC clone fingerprinting. Adopting a combination of approaches would therefore appear a sensible conclusion.

At present, potato is not considered a target species for full-genome sequencing. This marker-dense map represents a vast amount of sequence information contained by the AFLP markers, which can be readily exploited in subsequent genetical studies. We have found that up to 50% of the markers segregating in the SH × RH population also segregate in other *Solanum tuberosum* populations (E. Isidore and B. Pande, unpublished results). As comigrating AFLP fragments have been demonstrated to map to the same location in different crosses, a catalog of mapped AFLPs forms the basis of transferability. A previously developed catalog (Rouppe van der Voort *et al.* 1997c) is currently being extended to incorporate the data summarized here and to allow the transfer of marker information from the marker-dense bin map to any other potato population.

The volume of genotypic data generated in this experiment makes it difficult to provide the information in a single publication. Thus, an important facet of this study was presentation of the data in electronic format. The website http://www.dpw.wageningen-ur.nl/uhd/index.html will facilitate communication of these results. It provides the detailed parental bin maps and the bridges between the maps, including all the marker information for LG I. In future versions, the complete marker-dense map of potato will be available on this site as well as all the segregation data and gel images. In the era of RFLP mapping, the dissemination of mapping results was obtained by distributing RFLP probes among research groups. In the PCR era, dissemination was achieved by sharing primers or primer sequences. For AFLP, the electronic availability of annotated gel images is necessary to compare results among labs. We have found that within the context of an internationally collaborative project well-annotated AFLP gel images provide an efficient way of aligning linkage maps constructed from other potato populations.

In conclusion, this experiment represents the first steps toward our goal of developing a 10,000-point genetic map that will form a framework for both genetic studies and the construction of an integrated physical/genetic mapping resource of potato. Our results highlight the issues of data errors and inconsistencies and provide potential analytical solutions to overcoming them. The data suggest that epigenetic variation may be a significant feature of potato populations, although this conclusion should be treated with caution as we have not definitively proved this to be the case. However, this area does

warrant further investigation—particularly given the phenotypic parallels between progeny from methylation mutants in Arabidopsis (Vongs *et al.* 1993) and the acute inbreeding depression apparent in potato populations.

The authors thank David Marshall, Luke Ramsay, Christine Hackett, and John Bradshaw for reading the manuscript and providing valuable comments and ideas. This work was carried out under the European Union FAIR (Agriculture and Fisheries) program grant FAIR5-PL97-3565.

LITERATURE CITED

Alleman, M., and J. Doctor, 2000 Genomic imprinting in plants: observations and evolutionary implications. Plant Mol. Biol. **43:** 147–161.

Alonso-Blanco, C., A. J. M. Peeters, M. Koornneef, C. Lister, C. Dean *et al.*, 1998 Development of an AFLP based linkage map of Ler, Col and Cvi *Arabidopsis thaliana* ecotypes and construction of a Ler/Cvi recombinant inbred line population. Plant J. **14** (2): 259–271.

Becker, J., P. Vos, M. Kuiper, F. Salamini and M. Heun, 1995 Combined mapping of AFLP and RFLP markers in barley. Mol. Gen. Genet. **249:** 65–73.

Bonierbale, M. W., R. L. Plaisted and S. D. Tanksley, 1988 RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. Genetics **120:** 1095–1103.

Buntjer, J., H. van Os and H. J. van Eck, 2000 ComBin: software for ultra-dense mapping. International Conference on Plant Animal Genome Research, PAG VIII, January 9–12, 2000, San Diego (http://www.intl-pag.org/pag/8/abstracts/pag8038.html).

Burr, B., F. A. Burr, K. H. Thompson, M. C. Albertson and C. W. Stuber, 1988 Gene mapping with recombinant inbreds in maize. Genetics **118:** 519–526.

Collins, A., D. Milbourne, L. Ramsay, R. Meyer, C. Chatot-Balandra *et al.*, 1999 QTL for field resistance to late blight in potato are strongly correlated with maturity and vigour. Mol. Breed. **5:** 387–398.

Das, O. P., and J. Messing, 1994 Variegated phenotype and developmental methylation changes of a maize allele originating from epimutation. Genetics **136:** 1121–1141.

de Koning, D-J., H. Bovenhuis and J. A. M. van Arendonk, 2000 On the detection of imprinted quantitative trait loci in experimental crosses of outbred species. Genetics **161:** 931–938.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot *et al.*, 1996 A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380:** 152–154.

Dietrich, W. F., J. Miller, R. Steen, M. A. Merchant, D. Damron-Boles *et al.*, 1996 A comprehensive genetic map of the mouse genome. Nature **380:** 149–152.

Gebhardt, C., E. Ritter, A. Barone, T. Debener, B. Walkemeier *et al.*, 1991 RFLP maps of potato and their alignment with the homeologous tomato genome. Theor. Appl. Genet. **83:** 9–57.

Haanstra, J. P. W., C. Wye, H. Verbakel, F. Meijer-Dekens, P. Van Den Berg *et al.*, 1999 An integrated high-density RFLP-AFLP map of tomato based on two *Lycopersicon esculentum* × *L. pennellii* F_2 populations. Theor. Appl. Genet. **99:** 254–271.

Harushima, Y., M. Yano, A. Shomura, M. Sato, T. Shimano *et al.*, 1998 A high-density rice genetic linkage map with 2275 markers using a single F_2 population. Genetics **148:** 479–494.

Joyce, S. M., and A. C. Cassells, 2002 Variation in potato microplant morphology *in vitro* and DNA methylation. Plant Cell Tissue Org. **70:** 125–137.

Keim, P., J. Schupp, S. Travis, K. Clayton, T. Zhu *et al.*, 1997 A high density soybean genetic map based on AFLP markers. Crop Sci. **37:** 537–543.

Klein, P. E., R. R. Klein, S. W. Cartinhour, P. E. Ulanch, J. Dong *et al.*, 2000 A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress towards a sorghum genome map. Genome Res. **10:** 789–807.

Kosambi, D. D., 1944 The estimation of map distances from recombination values. Ann. Eugen. **12:** 172–175.

Lloyd, V. K., D. A. Sinclair and T. A. Grigliatti, 1999 Genomic imprinting and position effect variegation in *Drosophila melanogaster.* Genetics **151:** 1503–1516.

Matthes, M., R. Singh, S. C. Cheah and A. Karp, 2001 Variation in oil palm (Elaeis guineensis Jacq.) tissue culture-derived regenerants revealed by AFLPs with methylation-sensitive enzymes. Theor. Appl. Genet. **102:** 971–979.

Milbourne, D., R. Meyer, A. Collins, L. Ramsay, C. Gebhardt *et al.*, 1998 Isolation, characterisation and mapping of simple sequence repeat loci in potato. Mol. Gen. Genet. **259:** 233–245.

Morison, I. M., C. J. Paton and S. D. Cleverley, 2001 The imprinted gene and parent-of-origin effect database. Nucleic Acids Res. **29:** 275–276.

Murray, J. C., K. H. Buetow, J. L. Weber, S. Ludwigsen, T. Scherpbier-Heddema *et al.*, 1994 A comprehensive human linkage map with centimorgan density. Science **265:** 2049–2054.

Polanco, C., and M. L. Ruiz, 2002 AFLP analysis of somaclonal variation in *Arabidopsis thaliana* regenerated plants. Plant Sci. **162:** 817–824.

Powell, W., W. Thomas, E. Baird, P. Lawrence, A. Booth *et al.*, 1997 Analysis of quantitative traits in barley by the use of amplified fragment length polymorphisms. Heredity **79:** 48–59.

Rouppe van der Voort, J., P. Wolters, R. Folkertsma, R. Hutten, P. van Zandvoort *et al.*, 1997a Mapping the cyst nematode locus *Gpa2* in potato using a strategy based on comigrating AFLP markers. Theor. Appl. Genet. **95:** 874–880.

Rouppe van der Voort, J., P. van Zandvoort, H. van Eck, R. Folkertsma, R. Hutten *et al.*, 1997b Use of allele specificity of comigrating AFLP markers to align genetic maps from different potato genotypes. Mol. Gen. Genet. **255:** 438–447.

Rouppe van der Voort, J. N. A. M., H. J. van Eck, J. Draaistra, P. M. van Zandvoort, E. Jacobsen *et al.*, 1997c An online catalogue of AFLP markers covering the potato genome. Mol. Breed. **4:** 73–77.

Sherman, J. D., and S. M. Stack, 1995 Two-dimensional spreads of synaptonemal complexes from solanaceous plants. VI. High-resolution recombination nodule map for tomato (*Lycopersicon esculentum*). Genetics **141:** 683–708.

Stam, P., 1993 Construction of integrated genetic-linkage maps by means of a new computer package—JoinMap. Plant J. **3:** 739–744.

Stam, P., and J. van Ooijen, 1995 *JoinMap Version 2.0: Software for the Calculation of Genetic Linkage Maps.* CPRO-DLO, Wageningen, The Netherlands.

Tanksley, S., M. Ganal, J. Prince, M. De Vicente, M. Bonierbale *et al.*, 1992 High density molecular linkage maps of the tomato and potato genomes. Genetics **132:** 1141–1160.

Tanksley, S. D., M. W. Ganal and G. B. Martin, 1995 Chromosome landing: a paradigm for map-based gene cloning in plants with large genomes. Trends Genet. **11:** 63–68.

van der Beek, J. G., R. Verkerk, P. Zabel and P. Lindhout, 1992 Mapping strategy for resistance genes in tomato based on RFLPs between cultivars: Cf-9 (resistance to *Cladosporium fulvum*) on chromosome 1. Theor. Appl. Genet. **84:** 106–112.

van Eck, H., J. Rouppe van der Voort, J. Draaistra, P. van Zandvoort, E. Van Enckevort *et al.*, 1995 The inheritance and chromosomal localization of AFLP markers in a non-inbred potato offspring. Mol. Breed. **1:** 397–410.

van Os, H., J. Buntjer, P. Stam and H. J. van Eck, 2000 Evaluation of two algorithms for the construction of genetic linkage maps. International Conference on Plant Animal Genome Research, PAG VIII, January 9–12, 2000, San Diego (http://www.intl-pag.org/pag/8/abstracts/pag8621.html).

Vongs, A., T. Kakutani, R. A. Martienssen and E. J. Richards, 1993 *Arabidopsis thaliana* DNA methylation mutants. Science **260:** 1926–1928.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee *et al.*, 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. **23:** 4407–4414.

Vuylsteke, M., R. Mank, R. Antonise, E. Bastiaans, M. L. Senior *et al.*, 1999 Two high-density AFLP linkage maps of *Zea mays* L.: analysis of distribution of AFLP markers. Theor. Appl. Genet. **99:** 921–935.

Young, J. P., J. M. Schupp and P. Keim, 1999 DNA methylation and AFLP marker distribution in the soybean genome. Theor. Appl. Genet. **99:** 785–790.

Communicating editor: T. Brown