The Construction of an Ultra-Dense Genetic Linkage Map of Potato

Het maken van een Ultradichte Genetische Koppelingskaart van Aardappel

Promotor:
Prof. dr. R. G. F. Visser, hoogleraar in de plantenveredeling, Wageningen Universiteit

Copromotor:
Dr. ir. H. J. van Eck, universitair docent, laboratorium voor plantenveredeling

Promotiecommissie:
Prof. dr. ir. J. Bakker, Wageningen Universiteit
Prof. dr. C. Heyting, Wageningen Universiteit
Dr. ir. J. Jansen, Plant Research International
Dr. F. A. van Eeuwijk, Wageningen Universiteit

Dit onderzoek is uitgevoerd binnen de onderzoekschool Experimentele Plantenwetenschappen

The Construction of an Ultra-Dense Genetic Linkage Map of Potato

Hans van Os

Proefschrift
ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit,
Prof.dr.ir. L. Speelman,
in het openbaar te verdedigen
op maandag 13 juni 2005
des middags te half twee in de Aula.

CIP-DATA Koninklijke Bibliotheek, Den Haag

The construction of an ultra-dense genetic linkage map of potato
Van Os, H.

Thesis Wageningen University, The Netherlands
With summaries in Dutch and English

ISBN 90-8504-221-6

Contents

Chapter 1 **7**
General Introduction

Chapter 2 **13**
Toward a Marker-Dense Meiotic Map of the Potato Genome: Lessons From Linkage Group I

Chapter 3 **31**
RECORD: a Novel Method for Ordering Loci on a Genetic Linkage Map

Chapter 4 **47**
SMOOTH: a Statistical Method for Successful Removal of Genotyping Errors from High-Density Genetic Linkage Data

Chapter 5 **61**
A 10,000 Marker Ultra-Dense Genetic Recombination Map as a New Tool for Anchoring a Physical Map and Fast Gene Cloning in Potato

Chapter 6 **81**
General Discussion

Summary **89**

Samenvatting **91**

Nawoord **93**

Curriculum Vitae **95**

Chapter 1: General Introduction

PLANT BREEDING

Since the industrial revolution, a huge increase in food production efficiency can be seen over the years. The improvement in agricultural production has been achieved by several factors: a more efficient use of land through redistribution of farm land and the merger of farms; better quality and availability of inputs like fertilizer and pesticides; and the genetic improvement of crops or plant breeding. With a rapidly growing world population, food security becomes an increasingly important issue (FAO). Breeding new varieties in crops may help to fill the needs of our next generation. Plant breeding is an applied science and relies on genetic variation. By crossing existing material or wild material and selecting for favorable genotypes in the progeny, improved cultivated varieties (cultivars) can be produced. With the aid of modern biotechnology, cultivar production can be accelerated and a much broader range of genetic variation can be used.

MOLECULAR MARKERS

To select for a favorable genotype can be very difficult. For instance, traits like disease resistance can only be evaluated by exposing the genotypes to the pathogen and traits like yield inherit quantitatively and can be controlled by several genes. THODAY (1961) proposed to use single gene markers to map and characterize the genes that control these traits. These markers are supposed to have a high genetic linkage with these genes and can therefore be used to indirectly select for the favorable genotype.

Isozyme markers were the first biochemical single gene markers that were employed for genetic analyses of quantitative traits (EDWARDS *et al.* 1987; TANKSLEY *et al.* 1982). These markers were more abundant than morphological markers, but still their number is limited and there are not enough informative markers to cover an entire genome.

The discovery of restriction enzymes (SMITH and WILCOX 1970) and the polymerase chain reaction (PCR; MULLIS and FALOONA 1987) have created the opportunity to visualize the composition of organisms at the DNA level, and obtain a so-called genetic fingerprint (*e.g.* KEARSEY and POONI 1996). The visible differences in these fingerprints among the genotypes are called molecular markers. The number of molecular single gene markers that can be produced is virtually inexhaustible. These markers are DNA based and distributed over the entire genome. Other advantages compared to morphological markers are their phenotypic neutrality and lack of epistatic and pleiotropic effects. Nowadays, several different types of molecular markers are available for genetic studies. The AFLP (amplified fragment length polymorphism) technique in particular, is very fast and yields an enormous amount of data in a relatively short time (VOS *et al.* 1995).

GENETIC LINKAGE MAPS

A genetic linkage map represents the relative order of genetic markers along a chromosome. Recombination frequencies are used to determine the relative distance between the markers. Genetic linkage maps should not be confused with physical genomic maps, which can be obtained by determining the DNA sequence of chromosomes, as is currently being done in several genome mapping projects. Linkage maps and physical maps are related, but this relation is usually not linear (e.g. SCHMIDT *et al.* 1995). Genetic linkage maps are a valuable tool in basic genetical studies and applied breeding programs, especially for the identification and selection of genotypes with specific combinations of favorable traits. Traditionally, only classical genetic markers like morphological traits were available, which required a large number of segregating populations to develop a linkage map, as only a limited number of loci segregated in each cross. With the advent of molecular markers, unlimited amounts of segregating loci became available for a single cross. The parents in these crosses are mostly homozygous and F_2, RIL or BC generations are used to perform linkage analysis. For outbreeding species like potato, heterozygous parents are used to obtain segregating populations and mapping can be performed in the F1 offspring of a single cross. In the past, genetic linkage maps have been constructed for several plant species like maize (HELENTJARIS *et al.* 1987), rice (MCCOUGH *et al.* 1988), Brassica (SLOCUM *et al.* 1990), potato (GEBHARDT *et al.* 1991), tomato (TANKSLEY *et al.* 1992), lettuce (KESSELI *et al.* 1994), sorghum (PEREIRA *et al.* 1994), Arabidopsis (ALONSO-BLANCO *et al.* 1998) and wheat (MESSMER *et al.* 1999).

As an intermediate between conventional linkage maps and sequencing the complete genome of an organism, high density maps have been generated. Marker-dense maps provide ordered frameworks for the construction of physical maps, onto which YAC or BAC contigs can be anchored (KLEIN *et al.* 2000). In crop plant species such as rice (HARUSHIMA *et al.* 1998: 2,275 AFLP markers), maize (VUYLSTEKE *et al.* 1999: 1,539 and 1,355 markers mapped in two populations), potato and tomato (TANKSLEY *et al.* 1992: ca. 1,000 markers; HAANSTRA *et al.* 1999: 1,175 markers), high-density genetic linkage maps have already been constructed. The combined maps of the tomato and potato genomes comprise approximately 1,000 RFLP markers assembled from several populations and together they represent an average spacing of c. 1.2 cM (GEBHARDT *et al.* 1991; TANKSLEY *et al.* 1992).

SOFTWARE

Locus ordering on a linkage map requires a criterion that defines the ‘best’ map and an algorithm to find the optimal sequence of loci. The criteria that have been proposed include the maximum likelihood (LANDER *et al.* 1987; JANSEN *et al.* 2001), the minimum sum of adjacent recombination fractions (SARF), the maximum sum of adjacent LOD scores (SALOD) (LIU and KNAPP 1995), the minimum number of crossovers (THOMPSON 1987) and the ‘least square locus order’ (STAM 1993).

Various computer packages for linkage mapping have implemented these criteria, combined with a certain search algorithm. For example, GMENDEL (LIU and KNAPP 1995) minimizes SARF using simulated annealing. The PGRI package (LU and LIU 1995) can minimize SARF or maximize the likelihood using simulated annealing and/or branch-and-bound. JoinMap (STAM 1993) minimizes the least square locus order using a stepwise search, which is a combination of seriation and branch-and-bound with some additional local reshuffling. For practical purposes the target function should not require intensive calculations and yet be acceptable from a statistical viewpoint. Especially with incomplete data (missing observations and/or incomplete genotype information as is the case with dominance) calculation of the complete likelihood and the least square criterion is time consuming. As a result, the methods that use these criteria are becoming too computing-intensive for constructing linkage maps of over 400 loci, for instance, on a regular basis. The construction of ultra-dense maps in particular requires a time-efficient criterion and a heuristic search algorithm to deal with the amounts of data and error.

LINKAGE MAPPING IN POTATO

Potato (*Solanum tuberosum*) is the world's fourth major crop for food production after wheat, rice and maize. In the Netherlands almost 2.3% of the world production is grown (Table 1). The crop belongs to the family of *Solanaceae*, which also harbors food crops and ornamentals like tomato (*Lycopersicon esculentum*), eggplant (*Solanum melongena*), tobacco (*Nicotiana tabacum*), pepper (*Capsicum* spp.), *Petunia*, *Physalis*, *Atropa* and *Datura*. The genus *Solanum* comprises many species of which over 200 are tuber bearing. The cultivated potato itself is a highly heterozygous, autotetraploid species (2n=4x=48). Potato is an outbreeder and suffers from inbreeding depression when self-fertilized. Classical breeding involves evaluation and selection, based on several traits including yield, disease resistance and quality, on the clonal propagated progeny of a cross between two tetraploid clones. These clones can be existing cultivars or clones with introgressions from wild species. Potato breeding is a time-consuming process; it takes more than 10 years to produce a new cultivar.

Table 1. Potato production in 2004 (Mt) (FAO)

Country	Production
Netherlands	7,435,000
Western Europe	48,251,551
World	328,865,936

Genetic analysis and mapping of loci is also hampered by a high heterozygosity and an autotetraploid genome. Genetical studies of specific loci are complex, with the inheritance of traits often being masked by multiple alleles and even lethal alleles. Despite severe inbreeding depression and self-incompatibility, breeding methods have been developed to produce dihaploid clones (*e.g.* HERMSEN and VERDENIUS 1973). This has made the study of potato genetics more feasible, although the construction of genetic linkage maps is still more complex than in inbreeding species. In a cross between two heterozygous dihaploid clones, in theory four different alleles can segregate at a single locus <abxcd>. The progeny is the result of two independent meioses, which led to the approach of the two-way pseudo-testcross (GRATTAPAGLIA and SEDEROFF 1994). A linkage map is constructed for each parent and the information on loci segregating in both parents is used to align the two maps. The first genetic linkage map in a non-inbred species was developed for potato, using the segregation data from only one of the parents (BONIERBALE *et al.* 1988). Later, the approach with allelic bridges to align the parental maps was used (GEBHARDT *et al.* 1989). Other genetic linkage maps in potato were constructed by JACOBS *et al.* 1995; VAN ECK *et al.* 1995; MILBOURNE *et al.* 1998, and have enabled the localization of resistance genes, quality traits and QTL.

OUTLINE OF THIS THESIS

The aim of this study was to construct an ultra-dense genetic linkage map of potato and saturating the genome with markers for gene cloning via BAC landing. For this purpose an F1 population of 130 individuals from a cross between two diploid potato clones was evaluated for up to 10,000 AFLP markers. During an early stage of data analysis, it was noticed that the available mapping software could not cope with these data quantities and new software had to be developed. The program RECORD proved to be much faster and less sensitive to errors than the existing software in an evaluation experiment with simulated data described in Chapter 3. However, the relatively small portion of erroneous data caused too many ordering ambiguities, especially in the ultra-dense marker clusters. Chapter 4 presents a statistical error detection and removal program: SMOOTH. Simulation experiments provide the evidence that the vast majority of errors can be detected and that a reliable placement of markers can be realized. The successful application of RECORD and SMOOTH resulted in a new mapping concept based on a framework map consisting of bins. A detailed description of the methods and the bin concept is provided in Chapter 2 in a case study with chromosome I. According to the new concept framework maps have been constructed for all chromosomes. Chapter 5 describes the characteristics of the ultra-dense genetic linkage map of potato and its applications. Finally, Chapter 6 contains a general discussion on the results obtained during this study.

LITERATURE CITED

ALONSO-BLANCO, C., A. J. M. PEETERS, M. KOORNNEEF, C. LISTER, C. DEAN, N. VAN DEN BOSCH, J. POT and M. T. R. KUIPER, 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.

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.

EDWARDS M. D., C. W. STUBER and J. F. WENDEL, 1987 Molecular-marker-facilitated investigations of quantitative trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics **116**: 113-125

FAO: Review of the state of food and agriculture, FAO conference, Rome, november-december 2004. www.fao.org

GEBHARDT, C., E. RITTER, A. BARONE, T. DEBENER, B. WALKEMEIER, U. SCHACHTSCHABEL, H. KAUFMANN, R. THOMPSON, M. BONIERBALE, M. GANAL, S. TANKSLEY and F. SALAMINI, 1991 RFLP maps of potato and their alignment with the homeologous tomato genome. Theor. Appl. Genet. **83**: 9-57.

GEBHARDT, C., E. RITTER, T. DEBENER, U. SCHACHTSCHABEL, B. WALKEMEIJER and F. SALAMINI, 1989 RFLP analysis and linkage mapping in Solanum toberosum. Theor. Appl. Genet. **78**: 65-75

GRATTAPAGLIA, D., and R. SEDEROFF, 1994 Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers, Genetics **137**: 1121-1137.

HAANSTRA, J. P. W., C. WYE, H. VERBAKEL, F. MEIJER-DEKENS, P. VAN DEN BERG, P. ODINOT, A. W. VAN HEUSDEN, S. TANKSLEY, P. LINDHOUT and J. PELEMAN, 1999 An integrated high-density RFLP-AFLP map of tomato based on two *Lycopersicon esculentum x L. pennellii* F_2 populations. Theor. Appl. Genet. **99**: 254-271.

HARUSHIMA, Y., M. YANO, A. SHOMURA, M. SATO, T. SHIMANO, Y. KUBOKI, T. YAMAMOTO, S. YANG LIN, B. A. ANTONIO, A. PARCO, H. KAJIYA, N. HUANG, K. YAMAMOTO, Y. NAGAMURA, N. KURATA, G. S. KHUSH and T. SASAKI, 1998 A high-density rice genetic linkage map with 2275 markers using a single F_2 population. Genetics **148**: 479-494.

HELENTJARIS, T., M. SLOCUM, S. WRIGHT, A. SCHAEFER and J. NIENHUIS, 1986 Construction of linkage maps in maize and tomato using restriction fragment length polymorphisms. Theor. Appl. Genet. **87**: 392-401

HERMSEN J. G. TH. and J. VERDENIUS, 1973 Selection from Solanum tuberosum Group phureja of genotypes combining high frequency haploid induction with homozygosity for embryo-spot. Euphytica **22**: 244-259.

JACOBS, J. M. E., H. J. VAN ECK, P. ARENS, B. VERKERK-BAKKER, B. TE LINTEL HEKKERT, H. J. M. BASTIAANSSEN, A. EL-KHARBOTLY, A. PEREIRA, E. JACOBSEN and W. J. STIEKEMA, 1995 A genetic map of potato (Solanum tuberosum) integrating molecular markers, including transposons, and classical markers. Theoretical and Applied Genetics **91**: 289-300.

JANSEN J., A. G. DE JONG and J. W. VAN OOIJEN, 2001 Constructing dense genetic linkage maps. Theor. Appl. Genet. **102**: 1113-1122.

KEARSEY, M. J. and H. S. POONI, 1996 The genetical analysis of quantitative traits. Chapman & Hall, London. 381 pp.

KESSELI, R. V., I. PARAN and R. W. MICHELMORE, 1990 Genetic linkage map of lettuce (Lactuca sativa, 2n=18) in ‘Genetic maps’, ed S. J. O’BRIEN, 5th Ed. Cold Spring Harbor Press, Cold Spring Harbor, New York

KLEIN, P. E., R. R. KLEIN, S. W. CARTINHOUR, P. E. ULANCH, J. DONG, J. A. OBERT, D. T. MORISHIGE, S. D. SCHLUETER, K. L. CHILDS, M. ALE and J. E. MULLET, 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.

LANDER, E. S., P. GREEN, J. ABRAHAMSON, A. BARLOW, M. J. DALY, S. E. LINCOLN and L. NEWBURG, 1987 MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics **1**: 174-181.

LIU, B. H. and S. J. KNAPP, 1990 GMENDEL: A program for Mendelian segregation and linkage analysis of individual or multiple progeny populations using log-likelihood ratio's. J Hered **81**: 407.

LU, Y. Y. and B. H. LIU, 1995 A new computer package for genomic research: PGRI (Plant Genome Research Initiative), Plant Genome Conference III, San Diego. http://www.intl-pag.org/3/abstracts/201pg3.html

MCCOUGH, S. R., G. KOCHERT, Z. H. YU, Z. Y. WANG, G. S. KHUSH, W. R. COFFMAN and S. D. TANKSLEY, 1988 Molecular mapping of rice chromosomes. Theor. Appl. Genet. **76**: 815-829.

MILBOURNE, D., R. MEYER, A. COLLINS, L. RAMSAY, C. GEBHARDT and R. WAUGH, 1998. Isolation, characterisation and mapping of simple sequence repeat loci in potato. Mol. Gen. Genet. **259**: 233-245.

MESSMER, M. M., M. KELLER, S. ZANETTI and B. KELLER, 1999 Genetic linkage map of a wheat X spelt cross. Theor. Appl. Genet. **98**: 6-7, 1163-1170.

MULLIS, K. B. and F. A. FALOONA, 1987 Specific synthesis of DNA in vitro via the polymerase catalysed reaction. Meth. Enzymol. **255**: 335-350.

PEREIRA, M. G., M. LEE, P. BRAMEL-COX, W. WOODMAN, J. DOEBLEY, R. WHITKUS and M. LEE, 1994 Construction of an RFLP map in sorghum and comparative mapping in maize. Genome **37**: 236-243.

SCHMIDT, R., J. WEST, K. LOVE, Z. LENEHAN, C. LISTER, H. THOMPSON, D. BOUCHEZ and C. DEAN, 1995 Physical map and organisation of Arabidopsis thaliana chromosome 4. Science **270**: 480-483.

SLOCUM, M. K., S. S. FIGDORE, W. C. KENNARD, J. Y. SUZUKI and T. C. OSBORN, 1990 RFLP linkage map of Brassica oleracea (2n=18) in 'Genetic maps', ed S.J. O'BRIEN, 5th Ed. Cold Spring Harbor Press, Cold Spring Harbor, New York.

SMITH, H. O. and K. W. WILCOX, 1970 A restriction enzyme from Hemophilus influenzae. 1. Purification and general properties. J. Mol. Biol. **51**: 379-392.

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

TANKSLEY, S., M. GANAL, J. PRINCE, M. DE VICENTE, M. BONIERBALE, P. BROWN, T. FULTON, J. GIOVANNONI, S. GRANDILLO, G. MARTIN, R. MESSEGEUR, J. MILLER, L. MILLER, A. PATERSON, O. PINEDA, M. RÖEDER, R. WING, W. WU and N. YOUNG, 1992 High density molecular linkage maps of the tomato and potato genomes. Genetics **132**: 1141-1160.

TANKSLEY, S. D., R. BERNATZKY, N. L. LAPITAN and J. P. PRINCE, 1982 Use of naturally-occurring enzyme variation to detect and map genes controlling quantitative traits in an interspecific backcross of tomato. Heredity **49**: 11-25.

THODAY, J. M., 1961 Location of polygenes. Nature **191**: 368-370.

THOMPSON, E. A., 1987 Crossover counts and likelihood in multipoint linkage analysis. IMA J. Math. Appl. Med. Biol. **4**: 93-108.

VAN ECK, H. J., J. ROUPPE VAN DER VOORT, J. DRAAISTRA, P. VAN ZANDVOORT, E. VAN ENCKEVORT, B. SEGERS, J. PELEMAN, E. JACOBSEN, J. HELDER and J. BAKKER, 1995 The inheritance and chromosomal localization of AFLP markers in a non-inbred potato offspring. Mol. Breed. **1**: 397–410.

VOS, P., R. HOGERS, M. BLEEKER, M. REIJANS, T. VAN DE LEE, M. HORNES, A. FRIJTERS, J. POT, J. PELEMAN, M. KUIPER and M. ZABEAU, 1995 AFLP: a new technique for DNA fingerprinting. Nucl. Acid Res. **23**: 4407-4414.

VUYLSTEKE, M., R. MANK, R. ANTONISE, E. BASTIAANS, M. L. SENIOR, C. W. STUBER, A. E. MELCHINGER, T. LUBBERSTEDT, X. C. XIA, P. STAM, M. ZABEAU and M. KUIPER, 1999 Two high-density AFLP linkage maps of *Zea mays* L.: analysis of distribution of AFLP markers. Theor. Appl. Genet. **99**: 921-935.

Chapter 2:

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

Edwige Isidore*, Hans van Os*, Sandra Andrzejewski, Jaap Bakker, Imanol Barrena, Glenn J. Bryan, Bernard Caromel, Herman J. van Eck, Bilal Ghareeb, Walter de Jong, Paul van Koert, Véronique Lefebvre, Dan Milbourne, Enrique Ritter, Jeroen N. A. M. Rouppe van der Voort, Françoise Rousselle-Bourgeois, Joke van Vliet and Robbie Waugh

Published in Genetics **165**: 2107-2116

* These authors contributed equally to the work

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 hot spots (or regions identical by descent) resulted in empty bins, despite the level of marker saturation.

INTRODUCTION

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, computationally 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.

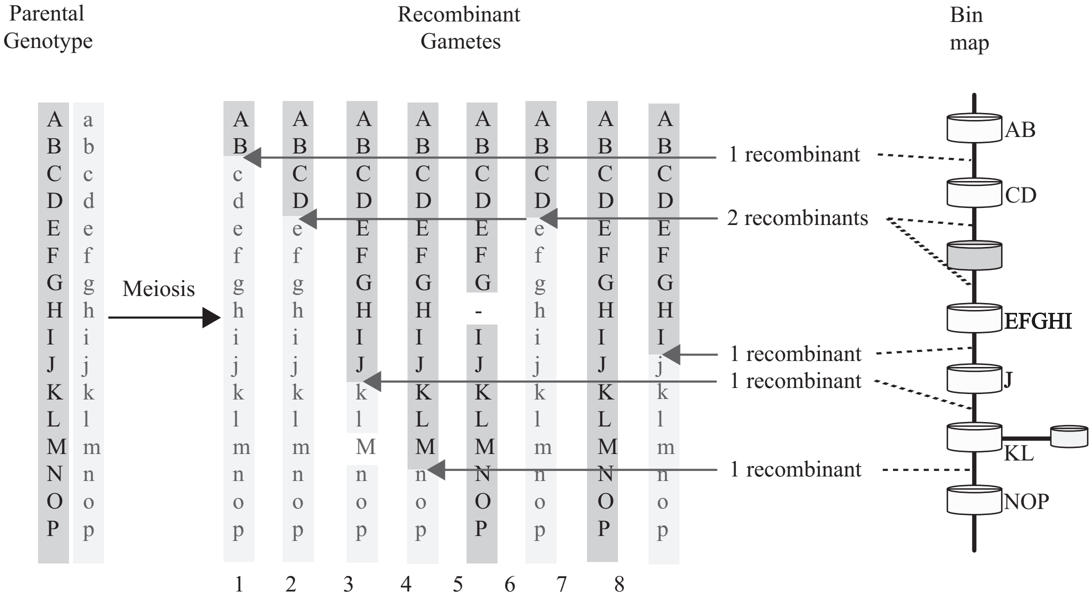


Figure 1. The recombination bin-mapping concept. For simplicity, only one heterozygous parental chromosome pair and eight potential recombinant gametes are shown. Allelic marker loci, Aa to Pp, are represented as upper- or lowercase letters on a white or shaded background. During meiosis, recombination 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).

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) x 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 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, 1x 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 x aa for analysis), 2223 (32.9%) in the male parent (coded aa x ab), and 1851 (27.4%) were heterozygous in both parents (coded ab x 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, ROUPPE VAN DER VOORT *et al.* 1997B), 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 fit into 72 bins, leaving 23 bins empty. The 420 paternal markers fit 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

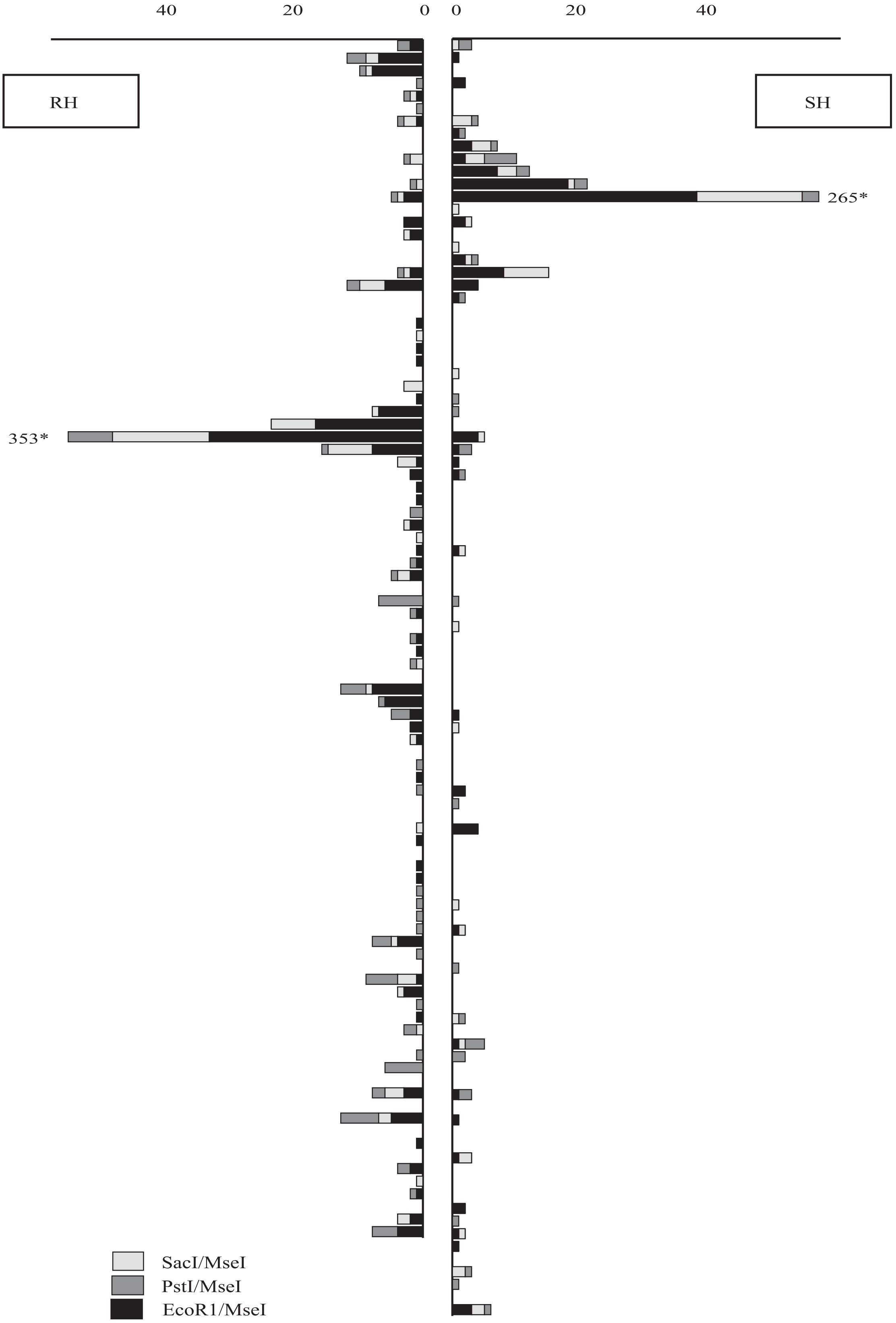


Figure 2. Final bin maps of SH and RH showing marker number and composition of each bin. The SH and RH maps are composed of 95 and 101 bins, respectively. The histograms with asterisks representing bins SH032 and RH013 have been scaled to fit on the page with the total number of markers indicated. *Eco*RI-,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

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 hot spots) 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!

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 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).

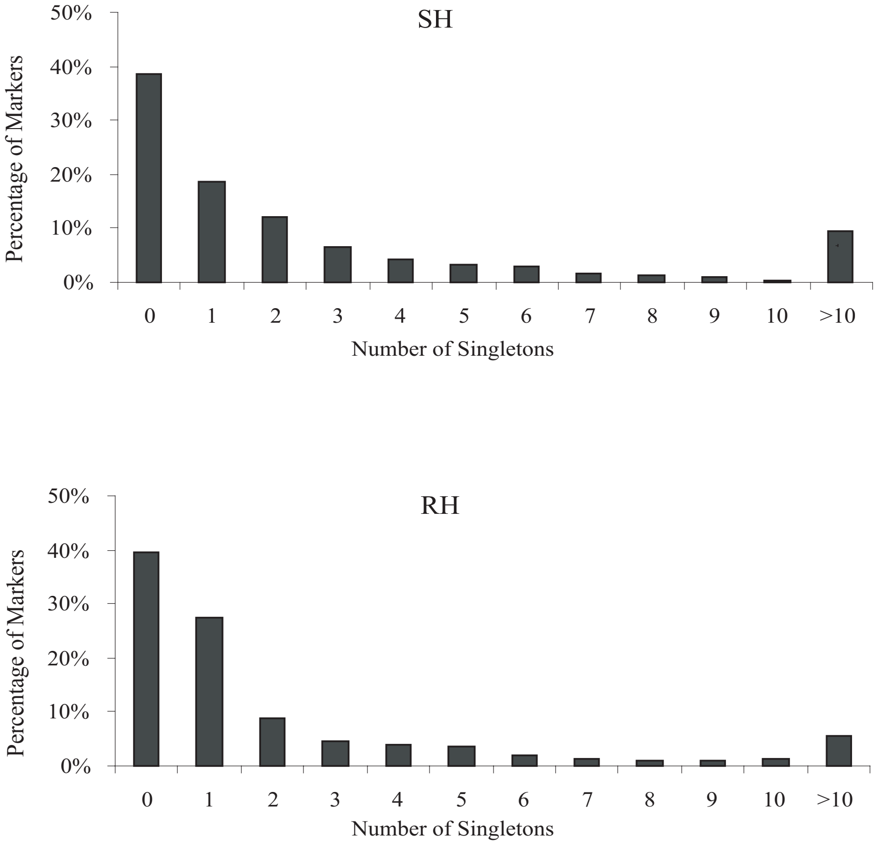


Figure 3. The percentage of error-checked 1:1 markers (*y*-axis) that fit into the skeleton bin maps of SH and RH either exactly or by the indicated number of singletons (*x*-axis) from a bin signature is shown. Deviation in the marker segregation pattern from the bin signature is expressed as the actual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

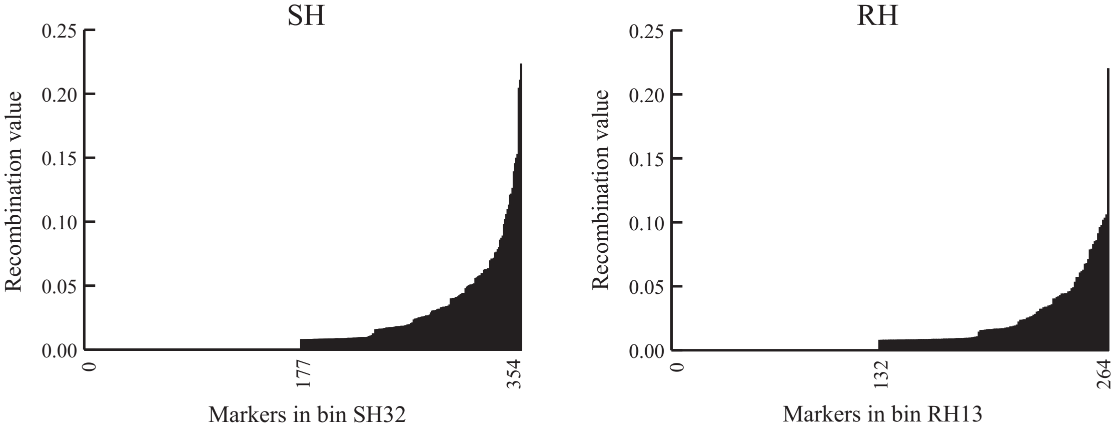


Figure 4. Distance (represented as a recombination fraction) between the actual marker segregation pattern and the bin signature for the markers in the largest maternal (SH032) and paternal (RH013) bins. The markers are ordered from left to right according to their goodness of fit in the bin.

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 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 x 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 x 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.

ACKNOWLEDGEMENTS

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, N. VAN DEN BOSCH, J. POT and M. T. R. KUIPER, 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. Intl. Conf. Plant Animal Genome Research: PAG VIII January 9-12, 2000, San Diego, CA. 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, P. OBERHAGEMANN, W. DE JONG, C. GEBHARDT, E. BONNEL and R. WAUGH, 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, A. VIGNAL, P. MILLASSEAU, S. MARC, J. HAZAN, E. SEBOUN, M. LATHROP, G. GYAPAY, J. MORISSETTE and J. WEISSENBACH, 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, Z. HUSAIN, R. DREDGE, M. J. DALY, K. A. INGALLS, T. J. O'CONNOR, C. A. EVANS, M. M. DEANGELIS, D. M.

LEVINSON, L. KRUGLYAK, N. GOODMAN, N. G. COPELAND, N. A. JENKINS, T. L. HAWKINS, L. STEIN, D. C. PAGE and E. S. LANDER, 1996 A comprehensive genetic map of the mouse genome. Nature **380**: 149-152.

GEBHARDT, C., E. RITTER, A. BARONE, T. DEBENER, B. WALKEMEIER, U. SCHACHTSCHABEL, H. KAUFMANN, R. THOMPSON, M. BONIERBALE, M. GANAL, S. TANKSLEY and F. SALAMINI, 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, P. ODINOT, A. W. VAN HEUSDEN, S. TANKSLEY, P. LINDHOUT and J. PELEMAN, 1999 An integrated high-density RFLP-AFLP map of tomato based on two *Lycopersicon esculentum x L. pennellii* F_2 populations. Theor. Appl. Genet. **99**: 254-271.

HARUSHIMA, Y., M. YANO, A. SHOMURA, M. SATO, T. SHIMANO, Y. KUBOKI, T. YAMAMOTO, S. YANG LIN, B. A. ANTONIO, A. PARCO, H. KAJIYA, N. HUANG, K. YAMAMOTO, Y. NAGAMURA, N. KURATA, G. S. KHUSH and T. SASAKI, 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 Tiss. Org. **70**: 125-137.

KEIM, P., J. SCHUPP, S. TRAVIS, K. CLAYTON, T. ZHU, L. SHI, A. FERREIRA and D. WEBB, 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, J. A. OBERT, D. T. MORISHIGE, S. D. SCHLUETER, K. L. CHILDS, M. ALE and J. E. MULLET, 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**: 172-75.

LLOYD, V. K., D. A. SINCLAIR and T. A. GRIGLIATTI, 1999 Genomic imprinting and position effect variegation in *Drosophila melanogaster*. Genetics **151(4)**: 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 and R. WAUGH, 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, F. MANION, J. QUILLEN, V. C. SHEFFIELD, S. SUNDEN, G. M. DUYK, J. WEISSENBACH, G. GYAPAY, C. DIB, J. MORISSETTE, G. M. LATHROP, A. VIGNAL, R. WHITE, N. MATSUNAMI, S. GERKEN, R. MELIS, H. ALBERTSEN, R. PLAETKE, S. ODELBERG, D. WARD, J. DAUSSET, D. COHEN and H. CANN, 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, B. HARROWER, J. MCNICOL and R. WAUGH, 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, H. VINKE, K. KANYUKA, A. BENDAHMANE, E. JACOBSEN, R. JANSSEN and J. BAKKER, 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, J. DRAAISTRA, F. GOMMERS, E. JACOBSEN, J. HELDER and J. BAKKER, 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 and J. BAKKER, 1997c An online catalogue of AFLP markers covering the potato genome. Mol. Breed. **4**: 73-77.

SHERMAN, J. D., and S. M. STACK, 1995 2-dimensional spreads of synaptonemal complexes from solanaceous plants. 6. high-resolution recombination nodule map for tomato (*Lycopersicon esculentum*). Genetics **141:** 683-708.

STAM, P., and J. VAN OOIJEN, 1995 JoinMap version 2.0: Software for the calculation of genetic linkage maps. CPRO-DLO, Wageningen.

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

TANKSLEY, S., M. GANAL, J. PRINCE, M. DE VICENTE, M. BONIERBALE, P. BROWN, T. FULTON, J. GIOVANNONI, S. GRANDILLO, G. MARTIN, R. MESSEGEUR, J. MILLER, L. MILLER, A. PATERSON, O. PINEDA, M. RÖEDER, R. WING, W. WU and N. YOUNG, 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, B. SEGERS, J. PELEMAN, E. JACOBSEN, J. HELDER and J. BAKKER, 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. Intl. Conf. Plant Animal Genome Research: PAG VIII January 9-12, 2000, San Diego, CA.
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, M. HORNES, A. FRIJTERS, J. POT, J. PELEMAN, M. KUIPER and M. ZABEAU, 1995 AFLP: a new technique for DNA fingerprinting. Nucl. Acid Res. **23**: 4407-4414.

VUYLSTEKE, M., R. MANK, R. ANTONISE, E. BASTIAANS, M. L. SENIOR, C. W. STUBER, A. E. MELCHINGER, T. LUBBERSTEDT, X. C. XIA, P. STAM, M. ZABEAU and M. KUIPER, 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., SCHUPP, J. M. and P. KEIM, 1999 DNA methylation and AFLP marker distribution in the soybean genome. Theor. Appl. Genet. **99**: 785-790.

Chapter 3:
RECORD: a Novel Method for Ordering Loci on a Genetic Linkage Map

Hans van Os, Piet Stam, Richard G. F. Visser and Herman J. van Eck

In press

ABSTRACT

A new method, RECORD (REcombination Counting and ORDering) is presented for the ordering of loci on genetic linkage maps. The method minimizes the total number of recombination events. The search algorithm is a heuristic procedure, combining elements of branch-and-bound with local reshuffling. Since the criterion we propose does not require intensive calculations, the algorithm rapidly produces an optimal ordering as well as a series of near-optimal ones. The latter provides insight into the local certainty of ordering along the map. A simulation study was performed to compare the performance of RECORD and JoinMap. RECORD is much faster and less sensitive to missing observations and scoring errors, since the optimization criterion is less dependent on the position of the erroneous markers. In particular, RECORD performs better in regions of the map with high marker density. The implications of high marker densities on linkage map construction are discussed.

INTRODUCTION

Genetic linkage maps have become an indispensable tool for locating genes or quantitative trait loci (QTL), marker assisted breeding and map based gene cloning. The first linkage maps were based on few loci of morphological characteristics, like the classical *Drosophila* linkage map of chromosome X (STURTEVANT 1913). Sturtevant introduced the concept that the frequency of crossing-over between two genes provides an index of their distance on a linear genetic map. He proposed that one percent of crossing-over should be taken as equal to one map unit. He devised a crucial test of the principles of mapping genes by constructing crosses in which at least two or three genes were segregating simultaneously. These two- or three-point crosses provided the principles and methods for ordering and mapping genes. These principles have enabled geneticists to map genes and markers to the chromosomes of a variety of higher organisms, including man. From this historical perspective it is obvious that mapping methods embarked on pair-wise distance estimates. However, when large numbers of markers segregate in a single mapping population, the analysis of recombination events from marker segregation data is more rewarding. Distance estimates of marker pairs in dense regions are blurred by errors. The segregation data are a more direct reflection of the data ambiguities. Now, with the advent of molecular markers much larger numbers of segregating loci can be mapped within one single mapping population. As an intermediate between conventional linkage maps and sequencing the complete genome of an organism, high density maps are currently being generated (STEEN *et al.* 1999: 4736 SSLP-markers; KONG *et al.* 2002: 5136 microsatellite markers; HARUSHIMA *et al.* 1998: 2275 EST markers; ISIDORE *et al.* 2003: 1260 AFLP markers). These maps sometimes comprise over 500 markers per linkage group. Since the number of possible orders asymptotically increases exponentially with the number of loci to be ordered, the problem of finding the optimal or near-optimal ordering requires a search algorithm that avoids an exhaustive search. For example, with 100 loci in a linkage group the number of orders equals $(100!)/2 = 4.7 \times 10^{157}$, which clearly prohibits an exhaustive search. Another factor that may set limits to the practical application of a search algorithm is the complexity of the target function to be minimized or maximized.

The optimization problem: Locus ordering on a linkage map requires a criterion that defines the 'best' map and an algorithm to find the optimal sequence of loci. The criteria that have been proposed include the maximum likelihood (LANDER *et al.* 1987; JANSEN *et al.* 2001), the minimum sum of adjacent recombination fractions (SARF), the maximum sum of adjacent LOD scores (SALOD) (LIU and KNAPP 1995), the minimum number of crossovers (THOMPSON 1987) and the 'least square locus order' (STAM 1993).

Various computer packages for linkage mapping have implemented these criteria, combined with a certain search algorithm. For example, GMENDEL (LIU and KNAPP 1995) minimizes SARF using simulated annealing. The PGRI package (LU and LIU 1995) can minimize SARF or maximize the likelihood using simulated annealing and/or branch-and-bound. JoinMap (STAM 1993) minimizes the least square locus order using a stepwise search which is a combination of seriation and branch-and-bound with some additional local reshuffling. For practical purposes the target function should not require intensive calculations and yet be acceptable from a statistical viewpoint. Especially with incomplete data (missing observations and/or incomplete genotype information as is the case with dominance) calculation of the complete likelihood and the least square criterion is time consuming. As a result, the methods that use these criteria are becoming too computing-intensive for constructing linkage maps of over 400 loci, for instance, on a regular basis.

In this paper we propose the use of the minimum number of crossovers as the optimization criterion, combined with a heuristic search for the optimum. This combination of target function and search algorithm should enable us to order data sets with more than 500 loci within a reasonable time.

MATERIALS AND METHODS

The optimization criterion we use is COUNT, the number of recombination events. In a BC1 backcross with perfect data (no missing observations) this number is easily obtained by counting the number of recombinants per locus pair, and, for a given sequence of loci, by adding over adjacent loci. Although COUNT and SARF are similar, there is an essential difference: COUNT cannot decrease as more gametes (individuals) are added to the population (cf. THOMPSON 1987). Since the likelihood, as well as COUNT and SARF are monotonic functions of the recombination frequencies between adjacent loci, COUNT, SARF and likelihood will give the same optimal ordering for perfect data (see also JANSEN *et al.* 2001; HACKETT *et al.* 2003). When information is incomplete, due to for example missing observations or dominance in an F_2 mapping population, this counting of observable crossovers is replaced by the expected number of crossovers for any incomplete observation of a pair of loci. This expected number in turn is based on the maximum likelihood (ML) estimate of recombination frequency between the corresponding loci. Table 1 illustrates this calculation for the observation of the genotype *AAB●,* being *AABB* or *AABb* in an F_2,where alleles A and B are in coupling phase. In this situation locus A inherits codominantly and locus B dominantly. For other genotypes of incomplete information, the calculation runs along the same lines, using the ML-estimate of recombination frequency to calculate the conditional probabilities of the hidden genotypes.

Table 1. Calculation of the expected number of recombination events (crossovers) resulting in the genotype *AAB*● in an F_2 derived from the cross *AABB* × *aabb*. The probabilities of the hidden genotypes (*AABB* and *AABb*) are expressed in terms of the recombination frequency, *r*.

Observed genotype	Hidden genotypes	Conditional probability	Number of crossovers
AAB●	***AABB***	$\frac{(1-r)^2}{1-r^2}$	0
	AABb	$\frac{2r(1-r)}{1-r^2}$	1

$$E(x \mid AAB\bullet) = 0 \times \frac{(1-r)^2}{1-r^2} + 1 \times \frac{2r(1-r)}{1-r^2} = \frac{2r}{1+r}$$

In this way a matrix, *Xij*, representing the number of recombination events between marker pairs is constructed. Calculation of the criterion COUNT for a given sequence of *n* loci is done by simple addition of those numbers of recombination events over the proper (adjacent) loci, i.e.

$$COUNT = \sum_{i=1}^{n-1} X_{seq(i),seq(i+1)} \tag{1}$$

where *seq(i)* is the *i*th element of the sequence.

The computational advantage of using COUNT is that for any exchange of two positions or an inversion of a window of certain size in a given sequence, the resulting value of COUNT requires the replacement of only a few terms of the summation in Eqn (1).

In order to prevent an unnecessary computational overload, the population is tested for the presence of 'duplicate markers', that is markers with exactly the same segregation pattern, including missing observations. Groups of markers with identical segregation signature are placed in 'bins', and each bin is represented by one of its members in the subsequent analysis. The order of loci within a bin remains unresolved, unless additional information, not included in the 'current' mapping experiment, is available.

The core of the search algorithm is as follows. First, a sequence is constructed stepwise, starting with a randomly chosen pair of markers, and adding one marker at a time. For each marker to be added the best position is determined (one out of *n+1* positions if the current sequence has *n* elements). This is a branch-and-bound-like procedure. The order in which markers are added to the sequence is random.

Once all markers have been added to the linkage group, thus making a 'sequence', an additional search for improvement is performed, in the following way. A window of given size is moved along the sequence from head to tail and for every position of this window the sub-sequence within the window is inverted, and the resulting COUNT-value calculated. This is repeated for windows of increasing size, starting with size two until the window covers all but one of the loci in the sequence. Every improvement encountered this way is accepted before a larger window of markers is considered. The whole procedure is repeated until no further improvements are encountered. Notice that the strictness of the branch-and-bound method is lifted by the additional final search for local improvements, with the obvious goal to avoid getting trapped in a local minimum. However, also this reshuffling by a moving window of increasing size does not guarantee to find the global minimum. Indeed, experimentation with simulated data sets containing missing observations has shown that the final solution produced by this stepwise assembling and additional search, slightly depends on the order in which markers are added to the sequence. A solution to this input order dependency would be to add markers by the seriation principle (BUETOW and

CHAKRAVARTI 1987), *i.e.* at each step add the marker that is closest to the one at the current head or tail. In the context of the traveling salesman problem this strategy is also known as a ‘greedy’ one: at each step, travel to the nearest city that has not been visited before. It is known that this seriation strategy is not a guarantee to arrive at the global optimum either. For that reason we chose to simply repeat the procedure a number of times and select the best one from these replicate assemblages. With good quality data the replicate solutions produced by RECORD are all identical. Upon experimentation with simulated data we found that for data sets with up to 20% missing observations, increasing the number of replicate assemblages beyond ten is hardly rewarding. So we consider ten replicate build-ups of the sequence as a good compromise between speed and quality of the solution obtained.

Since the producer of a linkage map is not only interested in a single ‘best’ sequence of markers, but also in the certainty of that sequence, we have added the following procedure to the algorithm. Starting from the last and optimal solution, a search is performed for ‘almost equivalent’ solutions. An ‘almost equivalent’ solution is defined as one that induces a pre-set additional number of crossovers. So, a search is done for solutions that fall within this range of ‘admissible’ values of COUNT. The search itself is the same as described above: inversion of the sequence within a moving window, repeated for windows of increasing size. From the set of admissible solutions obtained this way, for each locus its distribution of positions is recorded. Inspection of this distribution provides a quick impression of the local certainty of the sequence. Figure 1 gives a sample of RECORD output, listing the positions taken by each marker in the set of ‘admissible’ sequences. It shows that for approximately 50% of the loci in this example the position is fixed, whereas for ‘islands’ of clustered markers the order within such a cluster is indeterminate.

```
                            000000000011111111112222222222
                            012345678901234567890123456789
  0       g3715           | 00
  1       w121            | 0000
  2       m217            | 0000
  3       g3837           |   00
  4       w174            |     00
  5       CHS             |     00
  6       w322            |       0
  7       g4560           |        0
  8       w138            |         000
  9       w433            |         000
  10      m291            |         000
  11      g4715-b         |            0
  12      w219            |             0
  13      w125            |              0
  14      w291b           |               0
  15      w137            |                0
  16      w323            |                 0
  17      m247            |                  0
  18      g4028           |                   0
  19      w194            |                    0
  20      w423b           |                     00
  21      w61             |                     00
  22      w271            |                       00
  23      w2              |                       00
  24      m435            |                         0 0
  25      w184            |                          00
  26      w69             |                         000
  27      g2368           |                            0
  28      m555            |                             00
  29      w335            |                             00
```

Figure 1. Sample output of RECORD, showing the rank numbers taken by markers in a series of near-optimal solutions. A '0' indicates that the corresponding rank number was given to this marker in one of the near-optimal solutions, for instance: markers w138, w433 and m291 can be found at rank numbers 8, 9 and 10; marker w137 is found only at rank number 15. (Data taken from the *Arabidopsis* genome data base.)

RECORD can deal with the following types of mapping populations: BC1, F_2, F_3, RILs (in fact any generation obtained by repeated selfing of a hybrid between homozygous parents). Mapping populations from non-inbreds should be split into BC1 or HAP data that represent the maternal and paternal gametes, according to the two-way pseudo-testcross method (GRATTAPAGLIA and SEDEROFF 1994).

The algorithm described above has been implemented in a DOS-oriented, C++ written computer program which is available from our web site (http://www.dpw.wageningen-ur.nl/pv/). We have chosen for the DOS platform since it enables running large batch jobs which is convenient for the purpose of the remainder of this study, a comparison of the performance of RECORD and JoinMap using simulated data.

A comparison of JoinMap and RECORD: In JoinMap the stepwise assembling of a locus sequence is essentially the same as in RECORD, *i.e.* a seriation-like procedure with local reshuffling (called 'rippling' in JoinMap) in a search for improvements (STAM 1993; STAM and VAN OOIJEN 1995). The search method of RECORD requires *(1/2) n (n-1)* evaluations of the target function for a sequence of length *n*. In JoinMap a similar number is required. However, evaluation of the JoinMap target function involves the inversion of an *n* by *n* matrix for each sequence of size *n+1*. So, asymptotically the number of operations in RECORD increases as n^2, whereas in JoinMap this increase is approximately by n^4. Moreover, calculation of COUNT, going from a given sequence to one with an inverted segment requires the replacement of only a few terms in the summation of Eqn. 1. This makes the RECORD algorithm extremely fast.

Three different experiments were performed. The first experiment was done to test whether or not the new method of minimizing recombination events as implemented in RECORD can produce maps of the same quality as the approach based on pairwise marker distances as implemented in JoinMap. Both RECORD and JoinMap were tested under a number of varying conditions such as population size, missing observations and error rate. In the second experiment, the two programs were tested for their error-sensitivity under different marker densities. In the third experiment, the speed of the software was evaluated.

Simulated data: We simulated first generation backcross populations (BC1). The simulated data were produced as follows. A given number of loci were randomly positioned (according to a Poisson process) along a single chromosome of specified length in cM. cM values are given as if calculated from an infinite amount of genotypes. Genotypes were generated for a BC1 progeny following standard Mendelian segregation and assuming no crossover interference. The number of crossover events solely depends on the distance as specified by the positions of the loci on the map. Scoring results were generated by assuming that missing observations and errors were independently and randomly distributed. (Note: Throughout this paper we imply that genotyping errors cover both human errors in the lab, scoring errors, typing errors, as well as reproducible although conflicting data points, resulting from biological phenomena as *e.g.* gene conversion.)

In experiment I, 150 independent maps of 50 loci spread along 50 cM were simulated. Next to speed, error-sensitivity is one of the most important factors while coping with high-density data sets. In this study emphasis is put on both error-sensitivity and speed. From each map four populations were simulated consisting of 25, 50, 100 and 250 individuals. In all population data noise was introduced by either 5, 10, 15, 20 and 30 percent errors or missing observations.

Experiment II was based on two data sets of different marker density. One data set was simulated from a map with 100 loci on a 10 cM map, the other data set from 100 loci on a 100 cM map. Both data sets consisted of 100 individuals and three percent scoring errors.

Experiment III was set up to assess the calculation speed of the two algorithms. Data sets were varied in the number of loci (50, 100, 150 and 200 loci) and population size (25, 50, 100, 250). All data sets contained 5 percent scoring errors, because perfect data do not provide a realistic impression of the mapping time in practice. The different settings for the simulations in the three experiments are summarized in Table 2.

Table 2. Values of simulation variables used in the three different experiments

Variables	**Experiment I**	Experiment II	**Experiment III**
Map length (cM)	50	10, 100	50
Number of loci	50	100	50, 100, 150, 200
Population size	25, 50, 100, 250	100	25, 50, 100, 250
Percentage scoring errors	0, 5, 10, 15, 20, 30	3	5
Percentage missing observations	0, 5, 10, 15, 20, 30	0	0

A yardstick for performance: As a measure for the performance of both algorithms we examined two different correlation coefficients between marker positions of the calculated sequence and the true order in the map that was used to generate the data. Since we are not dealing with map positions in centimorgans, but rather with rank numbers, the first correlation coefficient is Spearman's rank correlation (r_s). The second correlation coefficient is Kendall's τ coefficient.

In order to see to what extent local rearrangements of a given sequence of rank numbers affects the correlation coefficients we derived the following equations for local inversion of a segment. Inverting a window of size *k* in a sequence of length *n* leads to

$$r_s = 1 - 2\frac{k(k^2-1)}{n(n^2-1)}, \text{ and } \tau = 1 - 2\frac{k(k-1)}{n(n-1)}$$

Taking *k* as a fraction of *n* and writing $k/n = p$, one obtains, as *n* tends to infinity,

$$r_s(p) = 1 - 2p^3, \text{ and } \tau(p) = 1 - 2p^2 \qquad (2)$$

Figure 2 presents a graph of these relations. It shows that upon inverting 50% (p = 0.5) of a long sequence, r_s is still 0.75, whereas τ is 0.50. Clearly, Kendall's τ is a more sensitive correlation coefficient than Spearman's r_s. Small inversions, of less than 5% of the total length, have a negligible effect on the correlation coefficients. Multiple inversions will, of course, have larger impact. For *m* non-overlapping inversions covering a proportion p_i of the sequence, r_s and τ become

$$r_s = 1 - \sum_{i=1}^{m} 2p_i^3; \quad \tau = 1 - \sum_{i=1}^{m} 2p_i^2 \quad (\sum p_i \le 1)$$

We conclude that for r_s to drop below 0.8, or for τ to drop below 0.6, for instance, a very serious distortion of the sequence is required. In fact, such a distortion would be unacceptable in a real mapping experiment. To correct for possible (almost) complete map inversions, the absolute value of r_s and τ was taken for further calculations.

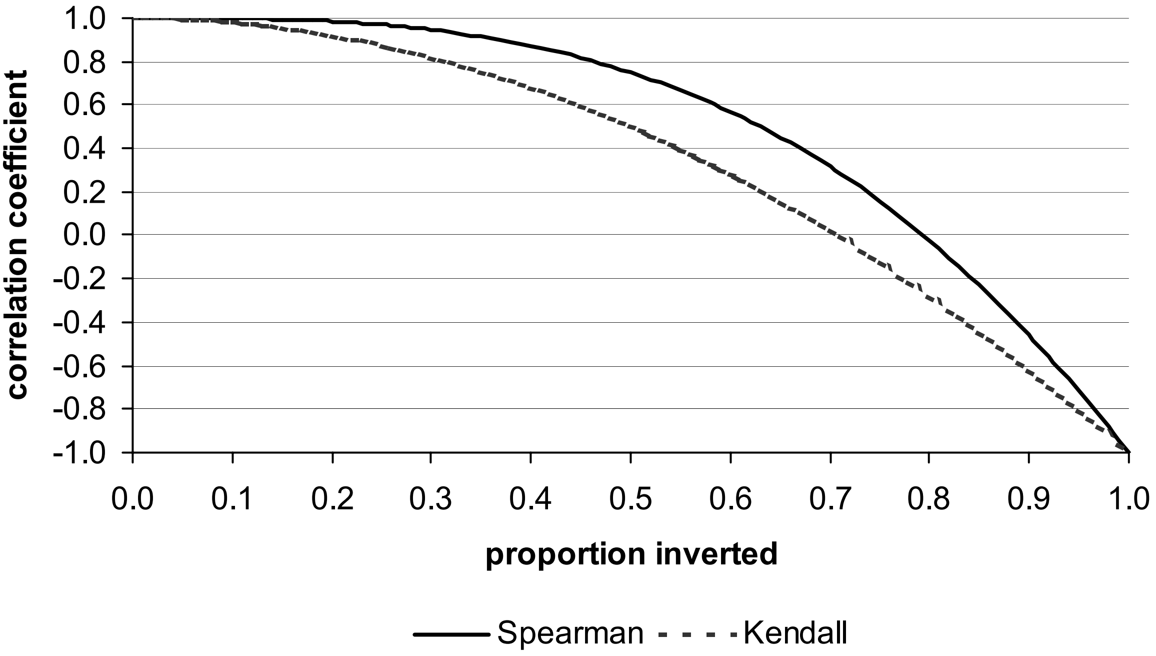


Figure 2. Change of two different correlation coefficients, Spearman's r_s and Kendall's τ, by inverting a window of markers consisting of a proportion p of a long sequence (Eqn. 2).

For testing purposes, rather general program settings were chosen for JoinMap. This means that all pairwise data were used with a LOD score higher than 1.0 and an estimated recombination fraction smaller than 0.45. Before actual mapping starts, JoinMap calculates the likelihood of the three possible orders of every triplet. When one of these exceeds the other two by a user-defined threshold value, this order is inferred as a so called 'fixed order'. (In the subsequent step-wise build-up and search of JoinMap, every order that is in conflict with a 'fixed order' is taboo.) In these experiments, the triplet threshold (logarithm of likelihood ratio) was set to 7.0. Finally, both JoinMap and RECORD have the option to perform a 'ripple' after adding a marker to the map. With a ripple, local marker order changes are systematically considered while improvements are maintained. In these tests, neither program performs ripples.

During this study, JoinMap 3.0 (VAN OOIJEN and VOORRIPS 2001) became available. This version of JoinMap is user-friendlier, because of the graphical user interface. However, for our experiments the MSDOS oriented JoinMap 2.0 was chosen because of its ability to run batch jobs. The results from this study can be extrapolated to JoinMap 3.0, since only minor changes in the algorithm have been introduced (J. W. VAN OOIJEN, personal communication).

RESULTS

Experiment I: In this experiment, both JoinMap and RECORD were tested with simulated data representing 50 marker loci on a 50 cM linkage group. Irrespective of the size of the mapping population (N = 25, 50, 100, 250), perfect marker orders were obtained. This result demonstrates that map construction using perfect data is not really a test case. In addition we tested two more algorithms, *i.e.* ComBin (BUNTJER *et al.* 2000) and JMQAD (the 'Quick-And-Dirty' module within the JoinMap 2.0 package) to recognize again that perfect maps are surely obtained with perfect data (results not shown). Apparently, the real test case for the performance of mapping algorithms is their sensitivity for ambiguities in the data caused by missing observations and/or genotyping errors. In realistic data, the proportion of missing observations and genotyping errors generally does not exceed five percent. However, to get a better view on the sensitivity of the methods for noise, both programs were tested with elevated levels of missing observations (5% up to 30%) and scoring errors (5% up to 30%). The performance of each of the programs, defined as the correlation coefficient between the true marker order and the order inferred by the software, was averaged over the 150 replications for every situation and is shown in Figure 3. It is clear that the accuracy of the marker order produced by the programs decreases with the data quality, reflecting a decrease in the ability of both programs to recover the correct order when data quality gets poor.

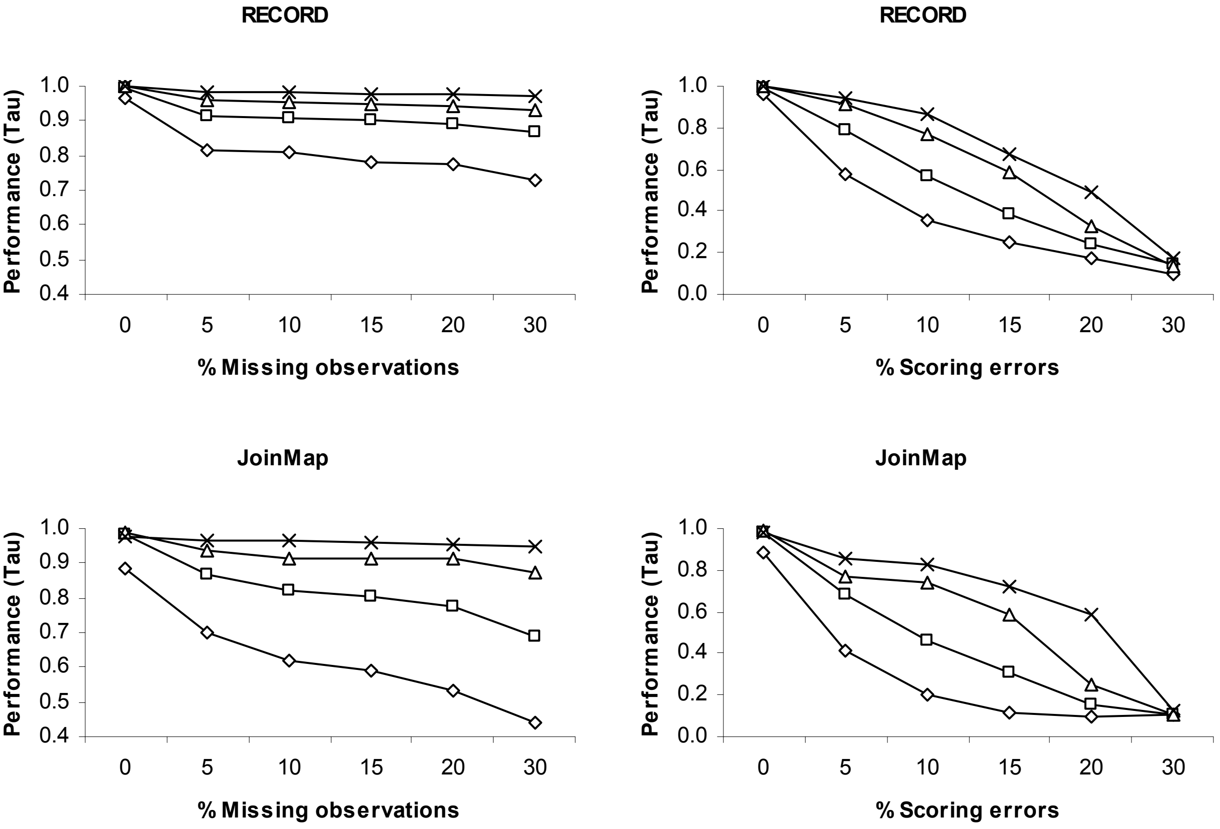


Figure 3. Performance in Kendall's τ of RECORD and JoinMap on data sets differing in population size and noise level. The population size is indicated by: '◊' for 25; '□' for 50; 'Δ' for 100 and '×' for 250 individuals. The results are based on 150 replicate runs. Data were obtained from experiment I.

Missing observations do not severely harm the recovered marker order. Especially in large mapping populations, the number of observations across descendants largely compensates the ambiguities caused by missing observations. Moreover, the vast majority of the missing observations do not induce ambiguities. Only when missing observations occur near recombinations, the placement of the markers with RECORD will be less accurate. Under these circumstances missing observations complicate the separation of markers from neighboring loci and make a pair of co-segregating loci of unspecified order. When more missing observations are present, the chance increases that these occur near recombinations. JoinMap however is more sensitive to missing observations than RECORD. Since in JoinMap not only recombination estimates between adjacent markers, but all pair-wise recombination estimates beyond a certain LOD threshold are used in the target function, and since a single missing observation slightly affects many of these pair-wise estimates, the impact of an increasing proportion of missing observations in JoinMap is greater than in RECORD.

The consequences of scoring errors are much more serious. An error may cause a separation of two co-segregating markers into two different loci. In this respect scoring errors have the same effect as recombinations. While recombinations are generally confirmed by other data points, errors occur on their own and seldom confirm each other.

In figure 4, an example data set is shown containing two forms of genotyping errors.

	1	2	3	4	5	6	7	8	9
MARKER1	A	A	B	B	A	A	B	B	A
MARKER2	A	A	B	B	A	A	B	B	A
MARKER3	B	A	A	B	A	B	B	B	A
MARKER4	A	B	A	B	B	B	A	B	A
MARKER5	B	B	A	A	A	B	A	B	B
MARKER6	B	B	A	A	A	B	A	B	B

Figure 4. Two types of errors in an example data set (individuals 1 to 9 in columns; markers 1 to 6 in rows). The error in marker4 at individual 5 does not cause an ordering ambiguity. However, the order of marker3 and marker4 is based on individual 2 and 7, but is in contrast with individual 1. Individual 1 contains an error close to a recombination event. In this case it is not clear whether marker3 or marker4 contains the error in individual 1.

Marker4 contains an error in individual 5. In reality, individual 5 does not contain any recombination events. Therefore this particular error will not add to the cost function of RECORD, when marker4 is tested on different positions, namely each position of marker4 yields two spurious recombination events to the COUNT-value. Placing marker4 at the end of the linkage group will not improve the order, as it causes a higher increase of the cost function in the other individuals. While RECORD is not sensitive to this kind of errors, JoinMap and other methods based on pairwise distances consider this error as a recombination and include it in the map distance calculation.

A different situation occurs in individual 1, where the error is close to a recombination event. Initially, RECORD will invert marker3 and marker4. This change will decrease the cost function in individual 1. However, this will cause a higher increase in the total cost function due to individuals 2 and 7. This situation remains insolvable as it is not clear whether marker3 or marker4 contains the error. The best order is determined based on the other individuals in the data set. In conclusion, scoring errors provide RECORD with ordering ambiguities only when they occur near recombination events. On the other hand, pairwise distance estimates are always affected by errors, independent of their position.

In general, larger populations have a beneficial effect on the mapping result. As population size increases, more recombination events between a pair of markers can be observed, which adds to the resolution between the markers. The positioning of the markers will be more accurate and the relative impact of missing observations and scoring errors decreases.

T-tests (data not shown) demonstrate that RECORD produces equally good or significantly better results than JoinMap. The T-tests were more significant when using Kendall's τ rather than Spearman's r_S. By exception, on data sets containing 250 individuals with an exceptional high error rate of 15% or 20%, JoinMap has a small advantage over RECORD although neither algorithm produces accurate maps in this situation. The reason for this small advantage for JoinMap is that at larger population sizes, errors have a smaller impact on the distance estimates.

Experiment II: In the second experiment, JoinMap and RECORD were tested for their ability to determine the marker order at higher densities. For this purpose, two data sets were used. The first set was based on a 100-marker map of 100 cM length ('normal' density). The second one was generated from a 'saturated' map, where 100 markers were spread over a distance of only 10 cM. From both maps, a BC1 population was simulated and a realistic amount of 3% errors was introduced. Calculated orders from both programs were compared with the true one and the results are shown by the scatter plot of Figure 5.

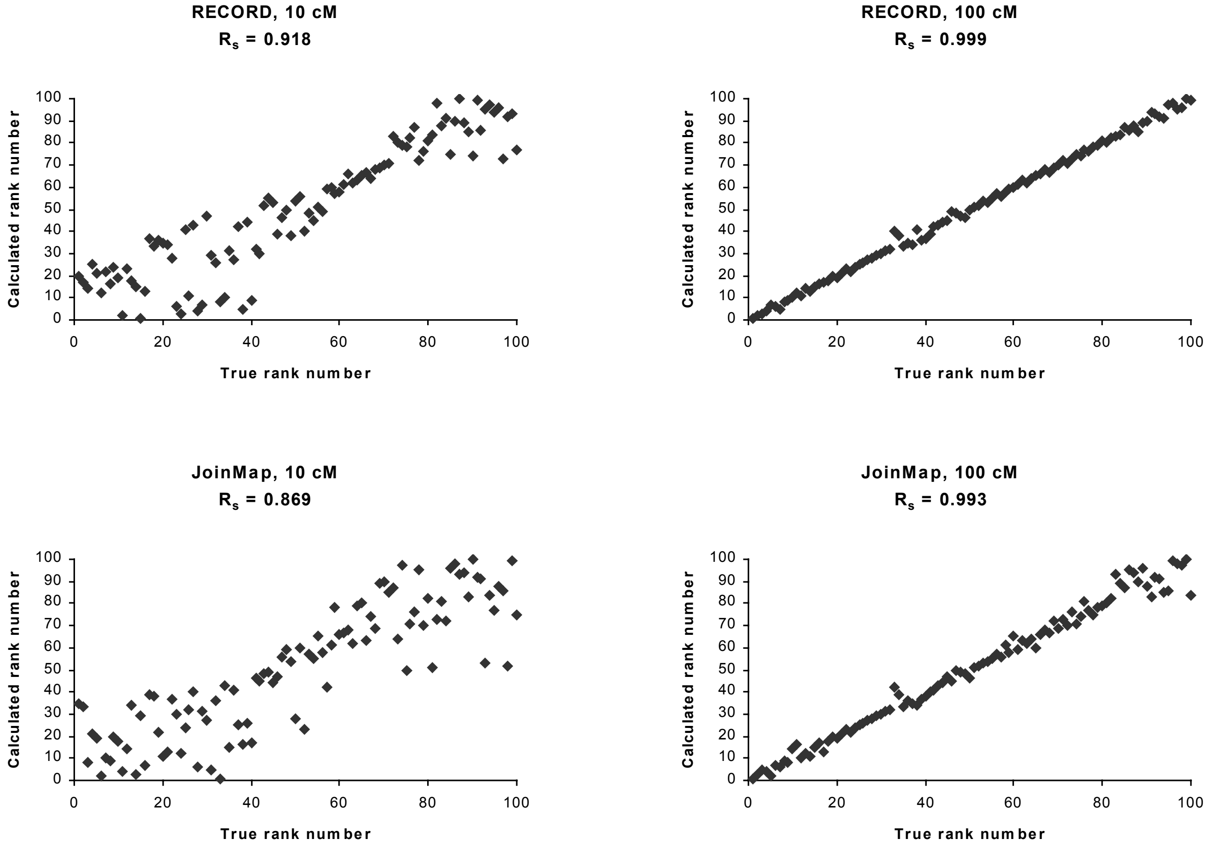


Figure 5. Performance of RECORD and JoinMap in dense maps. The calculated rank number of markers by both RECORD and JoinMap is compared with the true rank number by Spearman's r_s. Data were obtained from experiment II.

The dense map was more challenging to both programs. Although the mean number of errors remains the same, the average number of true crossovers in the dense map is reduced by a factor 10 as compared to the sparse map. This causes the signal/noise ratio to decrease by a factor 10. This explains why mapping in dense regions is more error-sensitive than mapping in less dense regions. The results of experiment II show that, in more dense regions, RECORD performs better than JoinMap.

Experiment III: In the third experiment, RECORD and JoinMap were compared for their speed. Calculation time was measured for a number of data sets varying in the number of loci and population size on a computer with a pentium II MMX processor of 350 MHz. Population size does not have a big effect on JoinMap. Therefore the results were averaged over tests at different offspring sizes with the same number of loci. Figure 6 shows the increase in calculation time for both programs. We fit power curves to these data and as anticipated, computation times for RECORD and JoinMap nicely fit curves of power 2 and 4, respectively. Thus, especially with data sets of over 100 loci, the speed advantage of RECORD over JoinMap is beyond discussion.

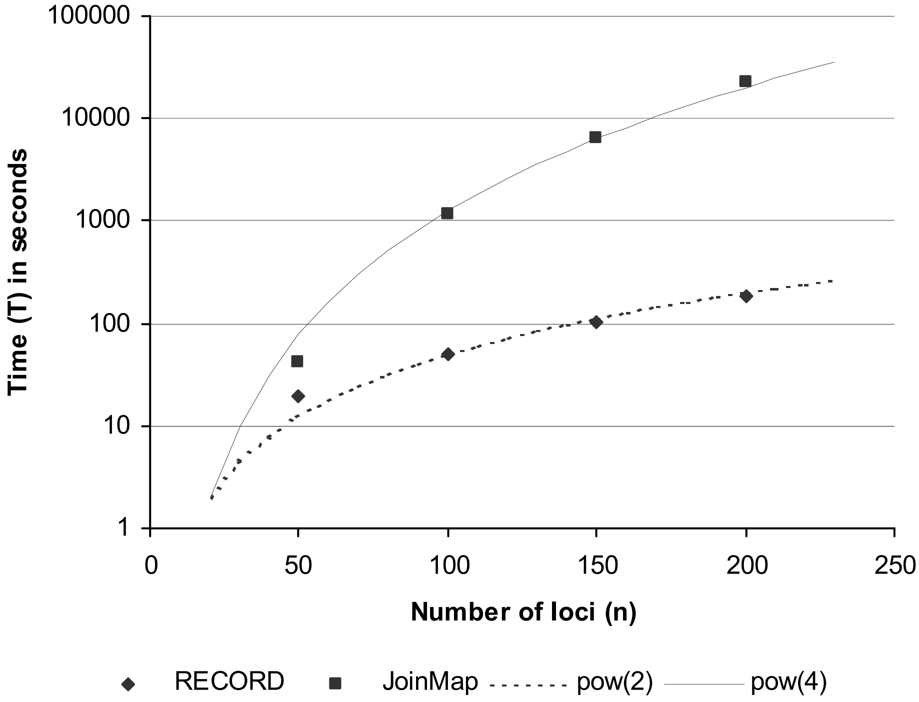


Figure 6. Computation time for RECORD and JoinMap. Power curves were fit by regression of time on number of loci. RECORD: $t = 0.00534n^2$; JoinMap: $t = 0.000011n^4$. Data were obtained from experiment III.

DISCUSSION

There are two major aspects to methods for efficient ordering of gene loci on a linkage map. First, the target function is important. In this paper we propose the total number of observable recombination events between adjacent markers as the target function, with an adaptation for situations in which genotype information is incomplete or missing. From a statistical point of view the full likelihood function would be an attractive alternative. The two criteria are equivalent in case the data are perfect (no missing observations and complete genotype information). In order to investigate the behavior of COUNT and likelihood with realistic data sets, we have compared the two methods using simulated data sets with incomplete information, *i.e.* an F_2 of size 100 with dominant markers and 5% missing observations. The two target functions were calculated for a series of near-optimal sequences (obtained by local inversion of segments) as well as a series of random rearrangements in the correct sequence.

Specifically for the first set of sequences (which corresponds to the part of the parameter space searched by RECORD), the squared correlation between COUNT and likelihood never dropped below 0.90. An example of the results of these calculations, where the correlation is one of the poorest we encountered, is shown in the scatter diagram of Figure 7. So, for practical purposes our heuristic COUNT criterion appears to be a quite acceptable compromise between statistical rigor and common sense.

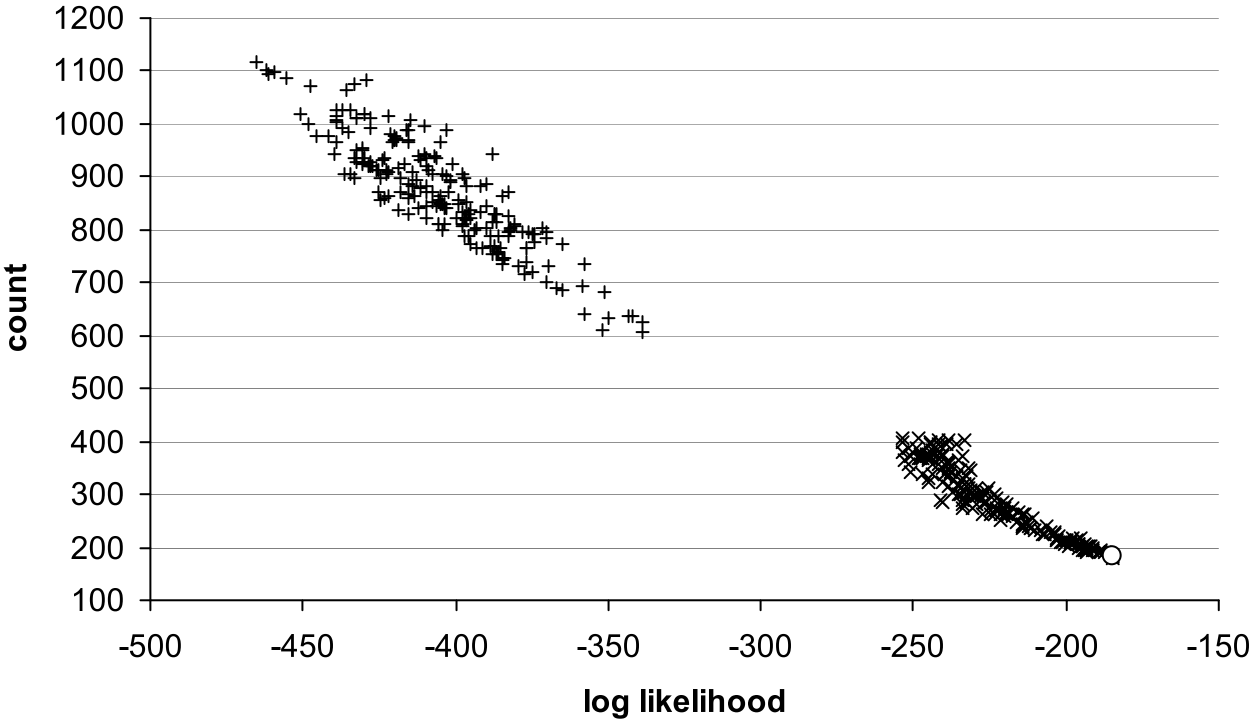


Figure 7. Relation between COUNT and log-likelihood. Data source: a simulated F_2 population of size 100 with dominant markers and 5% missing observations. Inversions (x): result for 80 sub-optimal sequences obtained by inversion of sequence segments. Random (+): result for sequences obtained by random exchange of pairs of loci. Best (O): solution produced by RECORD. Notice the much smaller likelihoods for sequences obtained by random changes, as explored by simulated annealing, in comparison with the likelihoods obtained by local inversion of segments.

Several other easy-to-calculate target functions have been proposed in the past. Among these are sum of adjacent map distances (SAD), sum of adjacent recombination frequencies (SARF) and sum of adjacent LOD scores (SALOD). For perfect data all of these are equivalent, in the sense that they have the same global optimum. However, with incomplete data both SARF and SALOD are inferior to COUNT. This is because SARF does not account for variation in precision of pairwise estimates, whereas SALOD may lead to erroneous results when the number of informative individuals varies between pairs of loci. Contrarily, the COUNT function comes close to the full likelihood since it uses observable recombination events (which is equivalent to likelihood) for that part of the data which has complete information and uses maximum likelihood estimates for the data that are incomplete.

The second aspect of map construction concerns the search algorithm for the optimum. In analogy to the traveling salesman problem, several approaches have been proposed. Among these are branch-and-bound (THOMPSON 1987), seriation (BUETOW and CHAKRAVARTI 1987) and simulated annealing (SA; KIRKPATRICK *et al.* 1983), or combinations thereof. Although SA generally produces optimal or near-optimal solutions, we did not choose it for the following reason. Extensive experience with linkage mapping has shown that most alternative maps that are produced by different computer packages and/or different program settings in JoinMap differ by inverted segments in the locus sequence. This is the result of ambiguities in real data and is in line with what one would expect intuitively. So, rather than the SA-search, which starts from a random sequence and subsequently randomly exchanges two loci or randomly moves a single locus along the sequence, we decided to search that part of the parameter space which most likely represents biological reality, starting from an ‘educated first guess’, obtained by branch-and-bound.

One may, of course, think of heuristic variations to both SA and the RECORD search. For example, to first construct a 'skeleton map' of not-too-closely linked markers and, during the subsequent SA-search involving all loci, consider any exchange of position involving two skeleton markers as a taboo area of the parameter space (J. JANSEN, personal communication).

An additional aspect of linkage mapping, which until recently has received little attention, concerns the (un)certainty of the map produced by a particular algorithm. We have added a feature to RECORD which provides the user with the distribution of rank numbers in a series of near-optimal solutions. Recently JANSEN *et al.* (2001) and HACKETT *et al.* (2003) have described a similar approach by recording the positions of loci in a series of sub-optimal solutions encountered in the SA-search.

In our comparison of the performance of RECORD and JoinMap we did not account for the fact that RECORD only produces orders, whereas JoinMap produces map positions in centimorgans. Therefore, the comparison is not a completely 'fair' one. On the other hand, correct locus ordering is of more importance than having 'exact' map distances, especially when constructing high-density maps. In such high-density maps the resolution that can be attained is primarily dictated by the size of the mapping population, usually not surpassing 1.0-0.25 centimorgans. Estimated 'exact' map distances in this order of magnitude do not make much sense, as their standard error readily exceeds the estimate itself.

Subsequent reasons why we have put emphasis on correct locus ordering and consider distance as relatively insignificant, are based on the unequal distribution of both recombination events and AFLP markers on the physical map. Highly localized hot spots or cold spots for recombination may cause manifold differences in map distance estimates between loci, depending on the sex or genetic background of the parental genotype. As a result, physical to genetic distances can vary from 25 kb/cM (BÜSCHGES *et al.* 1997) to 40 Mb/cM (ZHONG *et al.* 1999). Furthermore, successful application of mapping information in map based cloning or marker assisted selection with flanking markers also depends more on a correct marker order than accurate genetic distance estimates.

Apart from the observed difference in error-sensitivity between the programs, the results of experiment II once more demonstrate the disastrous effect that typing errors will have on the ability to recover the correct locus order, especially for regions of high marker density. Figure 3 indicates that the penalty for a typing error is roughly fivefold the penalty for a missing observation. For this reason we have developed a procedure, 'SMOOTH', for the detection of 'suspect' data points in a mapping population (VAN OS *et al.*, submitted). We have successfully applied this procedure in constructing a high-density linkage map for chromosome I in diploid potato (ISIDORE *et al.* 2003).

At this moment the RECORD-approach is being used for ultra-dense map construction in potato (ISIDORE *et al.* 2003). In these situations, linkage groups may contain more than 500 markers, numbers unthinkable to be analyzed simultaneously by conventional mapping software, as it would take more than nine days to calculate the map. Contrarily, RECORD analyses data sets of 500 markers within 20 minutes.

When RECORD was being developed, there were no alternative programs available that could handle these amounts of data. A new algorithm that can speed up map calculation, based on pairwise distances by using the simulated annealing approach, has been tested, but is not yet available (JANSEN *et al.* 2001; J. W. VAN OOIJEN, personal communication).

RECORD is capable of handling data sets of backcross populations, but to apply RECORD for the construction of the high density map of potato, which is based on a population derived from non-inbred parents, several modifications have to be made to the raw data. First, the observations recorded in the offspring have to be split into the products of male and female meiosis. From there on, the maps from both parents have to be calculated separately. Within the parental data sets, the linkage phase of each marker has to be assessed. This can be done with the 'Quick-And-Dirty' mapping module, which is included in the JoinMap 2.0 software package. This program calculates the best marker order by minimizing the sum of adjacent distances. Although this module does not produce very accurate marker orders, it is accurate enough for linkage phase ascertainment, which can be done by hand based on the neighboring markers. By converting all markers that are in repulsion phase in to coupling phase, the data are comparable with two separate BC1 populations for each parent, also referred to as the two-way pseudo-testcross (GRATTAPAGLIA and SEDEROFF 1994).

The version of RECORD used in this study only produces orders of loci, but no map positions in centimorgans. Currently we are preparing a version which does have this feature, as well as several sophistications, like a choice of target functions, an extended search algorithm for the more ambiguous data sets, a graphical user interface and a variety of output options.

In summary, conventional software has been sufficient in calculating linkage maps of low density. For the construction of high density maps, there is a strong need for faster and error-tolerant methods. The method described in this paper exceeds the currently available software both in speed and accuracy.

ACKNOWLEDGEMENTS

We are very grateful to Dr. Hans Jansen and Dr. Johan van Ooijen for providing valuable comments. This work was carried out under the European Union framework 5 program grant: FAIR5-PL97-3565.

LITERATURE CITED

BUETOW K. H. and A. CHAKRAVARTI, 1987 Multipoint gene mapping using seriation. I General methods. Am J Hum Genet **41**: 180-188.

BUNTJER J. B., H. VAN OS and H. J. VAN ECK, 2000 ComBin: Software for ultra-dense mapping Plant and Animal Genome Conference VIII, San Diego. http://www.intl-pag.org/pag/8/abstracts/pag8038.html

BUNTJER J. B., H. VAN OS and H. J. VAN ECK, 2000 Construction of ultra-dense maps using novel software Plant and Animal Genome Conference VIII, San Diego. http://www.intl-pag.org/pag/8/abstracts/pag8039.html

BÜSCHGES R., K. HOLLRICHER, R. PANSTRUGA, G. SIMONS, M. WOLTER, A. FRIJTERS, R. VAN DAELEN, T. VAN DER LEE, P. DIERGAARDE, J. GROENENDIJK, S. TOPSCH, P. VOS, F. SALAMINI and P. SCHULZE-LEFERT, 1997 The barley *Mlo* gene: a novel control element of plant pathogen resistance. Cell **88**: 695–705.

GRATTAPAGLIA, D., and R. SEDEROFF, 1994 Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers, Genetics **137**: 1121-1137.

HACKETT C. A., B. PANDE and G. J. BRYAN, 2003 Constructing linkage maps in autotetraploid species using simulated annealing. Theor Appl Genet **106**: 1107-115.

HARUSHIMA, Y., M. YANO, A. SHOMURA, M. SATO, T. SHIMANO, Y. KUBOKI, T. YAMAMOTO, S. YANG LIN, B. A. ANTONIO, A. PARCO, H. KAJIYA, N. HUANG, K. YAMAMOTO, Y. NAGAMURA, N. KURATA, G. S. KHUSH and T. SASAKI, 1998 A high-density rice genetic linkage map with 2275 markers using a single F_2 population. Genetics **148**: 479-494.

ISIDORE, E., H. VAN OS, S. ANDRZEJEWSKI, J. BAKKER, I. BARRENA, G. J. BRYAN, B. CAROMEL, H. J. VAN ECK, B. GHAREEB, W. DE JONG, P. VAN KOERT, V. LEFEBVRE, D. MILBOURNE, E. RITTER, J. ROUPPE VAN DER VOORT, F. ROUSSELLE-BOURGEOIS, J. VAN VLIET and R. WAUGH, 2003 Toward a marker-dense meiotic map of the potato genome: Lessons from linkage group I. Genetics **165**: 2107-2116.

JANSEN J., A. G. DE JONG and J. W. VAN OOIJEN, 2001 Constructing dense genetic linkage maps. Theor Appl Genet **102**: 1113-1122.

KIRKPATRICK S., C. D. GELATT and M. P. VECCHI, 1983 Optimization by simulated annealing. Science **220**: 671-680.

KONG A., D. F. GUDBJARTSSON, J. SAINZ, G. M. JONSDOTTIR, S.A. GUDJONSSON, B. RICHARDSSON, S. SIGURDARDOTTIR, J. BARNARD, B. HALLBECK, G. MASSON, A. SHLIEN, S. T. PALSSON, M. L. FRIGGE, T. E. THORGEIRSSON, J. R. GULCHER ans K. STEFANSSON, 2002 A high-resolution recombination map of the human genome. Nature Genetics **31**: 241-247.

LANDER, E. S., P. GREEN, J. ABRAHAMSON, A. BARLOW, M. J. DALY, S. E. LINCOLN and L. NEWBURG, 1987 MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics **1**: 174-181.

LIU, B. H. and S. J. KNAPP, 1990 GMENDEL: A program for Mendelian segregation and linkage analysis of individual or multiple progeny populations using log-likelihood ratio's. J Hered **81**: 407.

LU, Y. Y. and B. H. LIU, 1995 A new computer package for genomic research: PGRI (Plant Genome Research Initiative), Plant Genome Conference III, San Diego. http://www.intl-pag.org/3/abstracts/201pg3.html

STAM, P., and J. VAN OOIJEN, 1995 JoinMap version 2.0: Software for the calculation of genetic linkage maps. CPRO-DLO, Wageningen.

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

STEEN, R.G., A. E. KWITEK-BLACK, C. GLENN, J. GULLINGS-HANDLEY, W. VAN ETTEN, O. S. ATKINSON, D. APPEL, S. TWIGGER, M. MUIR, T. MULL, M. GRANADOS, M. KISSEBAH, K. RUSSO, R. CRANE, M. POPP, M. PEDEN, T. MATISE, D. M. BROWN, J. LU, S. KINGSMORE, P. J. TONELLATO, S. ROZEN, D. SLONIM, P. YOUNG, M. KNOBLAUCH, A. PROVOOST, D. GANTEN, S. D. COLMAN, J. ROTHBERG, E. S. LANDER and H. J. JACOB, 1999 A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. Genome Research **9**: AP1-AP8.

STURTEVANT, A. H., 1913 The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of a association. J Exp Zool **14**: 43-59.

THOMPSON, E. A., 1987 Crossover counts and likelihood in multipoint linkage analysis. IMA J Math Appl Med Biol **4**: 93-108.

VAN OOIJEN, J. W. and R. E. VOORRIPS, 2001 JoinMap® Version 3.0, Software for the calculation of genetic linkage maps. Plant Research International, Wageningen, The Netherlands.

VAN OS, H., R. G. F. VISSER and H. J. VAN ECK, 2005b SMOOTH: a statistical method for successful removal of genotyping errors from high-density genetic linkage data. Theor Appl Genet (accepted for publication)

ZHONG, X. B., J. BODEAU, P. F. FRANSZ, V. M. WILLIAMSON, A. VAN KAMMEN, H. J. DE JONG and P. ZABEL, 1999 FISH to meiotic pachytene chromosomes of tomato locates the root-knot nematode resistance gene Mi-1 and the acid phosphatase gene Aps-1 near the junction of euchromatin and pericentromeric heterochromatin of chromosome arms 6S and 6L, respectively. Theor Appl Genet **98**: 365–370.

Chapter 4:

SMOOTH: a Statistical Method for Successful Removal of Genotyping Errors from High-Density Genetic Linkage Data

Hans van Os, Piet Stam, Richard G. F. Visser and Herman J. van Eck

In press

ABSTRACT

High-density genetic linkage maps can be used for purposes such as fine-scale targeted gene cloning and anchoring of physical maps. However, their construction is significantly complicated by even relatively small amounts of scoring errors. Currently available software is not able to solve the ordering ambiguities in marker clusters, which inhibits the application of high-density maps. A statistical method named SMOOTH was developed to remove genotyping errors from genetic linkage data during the mapping process. The program SMOOTH calculates the difference between the observed and predicted values of data points based on data points of neighboring loci in a given marker order. Highly improbable data points are removed by the program in an iterative process with a mapping algorithm that recalculates the map after cleaning. SMOOTH has been tested with simulated data and experimental mapping data from potato. The simulations prove that this method is able to detect a high amount of scoring errors and demonstrates that the program enables mapping software to successfully construct a very accurate high-density map. In potato the application of the program resulted in a reliable placement of nearly 1000 markers in one linkage group.

INTRODUCTION

Linkage maps based on molecular markers are important tools in genetic analysis. They are useful for the localization of genes underlying quantitative traits, marker assisted breeding and map based gene cloning. Molecular marker systems like AFLP (VOS *et al.* 1995) allow that many markers can be generated in short time. This leads to the construction of highly saturated to enable fine-scale genetic mapping and the anchoring of physical maps (KLEIN *et al.* 2000).

In principle, these highly saturated or high-density maps can be constructed with the same software as genetic linkage maps of normal density. Commonly used programs like Joinmap (STAM 1993; STAM and VAN OOIJEN 1995) and MapMaker (LANDER *et al.* 1987) are very suitable for low-density genetic linkage map construction. However, these methods have difficulty in solving the increasing ordering ambiguities in denser maps (LINCOLN and LANDER 1992; VAN OS *et al.* submitted). Denser maps have more loci than normal maps, but the offspring genotypes contain the same amount of recombinations. With these fixed amounts of recombinations, the increased number of markers in denser maps are separated on average by less recombination events. Moreover, mapping algorithms based on pairwise distances will try to determine the order within clusters of markers; even for co-segregating markers or markers that only differ in a few scoring errors, but in fact share the same genetic position. In high-density maps, errors do not only give problems within marker clusters, but also across recombination events and thus severely complicate the establishment of the true marker order.

An accurate marker order is indispensable for further application of the map like for instance map based cloning. We state that marker order is more important than estimated map distances. Map distance estimates are trivial as they may vary across mapping studies by several cM. For clusters of markers that cosegregate, the order is indeterminate. Therefore it is not correct to suggest non-existing distance between markers caused by scoring errors or missing values.

The troublesome data points in the data are most likely to be caused by inaccurate scoring, but some data points that cause ambiguities in the marker order can also be caused by double recombination events, gene conversions, mutations and other biological phenomena. These various causes of ambiguous data are collectively called singletons. The term 'singleton' in the context of mapping data has first been used to indicate the misclassification of a marker phenotype (NILSSON *et al*. 1993). A singleton is in fact a single locus in one plant that appears to have recombined with both its directly neighboring loci (see figure 1). During map calculation, every singleton has to be treated as the unlikely event of a double recombination. We propose to identify and temporarily remove the singletons from the data. By eliminating these singletons, most ordering ambiguities are solved, including those in marker dense clusters.

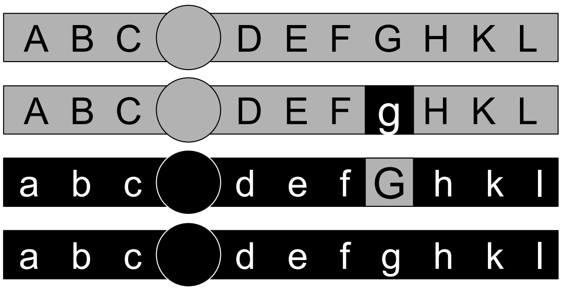


Figure 1. Four meiotic products after a double recombination event involving two non-sister chromatids causing a singleton at locus 'G/g'. The bar indicates a chromatid, whereas the centromere is represented by a circle.

As a consequence, mapping algorithms will not be hampered by these ordering ambiguities and can calculate the best possible map for the data set. This ideal framework map will be used to refit the raw data, supplying the verification of this map and a quality label for each marker that specifies the number of singletons these markers contain.

In this paper a statistical method is presented that can identify and remove the most obvious singletons in genetic linkage data. The method uses the marker order calculated by mapping software before eliminating the singletons. It is implemented in a computer program for the case of a first generation backcross population. This paper illustrates the advantages and potential pitfalls of eliminating singletons and provides concise instruction on how this method should be applied.

The method is tested on simulated data for different aspects like error percentage, population size and marker density. Besides the results of the simulation studies, experiments with real mapping data from potato are discussed as well.

METHODS

Software: The idea behind the identification of a singleton at a particular marker locus i is to compare the observed marker score at locus i, y_i, with a local prediction of the marker score, $\hat{y}_i$. The observed marker score, y_i, takes the value 1 when the allele is identified as coming from one of the sister chromatids, and 0 when coming from the non-sister chromatids.

The local prediction of the marker score $\hat{y}_i$ is calculated as the weighted w_j average of the observed scores y_j within a defined number of loci L flanking locus i on either side:

$$\hat{y}_i = \frac{\sum_{j \in L} w_j y_j}{\sum_{j \in L} w_j}, \text{ with } L = \{j : j \le \delta, j \ne 0\},$$

where δ is the maximum number of flanking loci around locus i that contributes to the local prediction for the marker score at i. Various weighing regimes were tested in combination with different choices for δ, but these parameter settings were rather immaterial to the performance of the procedure. For that reason we only present the results for δ =15 loci and with weights declining in a roughly quadratic fashion ($w_1 = 0.998$; $w_2 = 0.981$; $w_3 = 0.934$; $w_4 = 0.857$; $w_5 = 0.758$; $w_6 = 0.647$; $w_7 = 0.537$; $w_8 = 0.433$; $w_9 = 0.342$; $w_{10} = 0.265$; $w_{11} = 0.202$; $w_{12} = 0.151$; $w_{13} = 0.112$; $w_{14} = 0.082$; $w_{15} = 0.059$).

The absolute difference between observed marker score and predicted marker score, $d = |y_i - \hat{y}_i|$, is proportional to the probability that the marker score at i represents a singleton. Threshold values for d, above which singletons are identified, were adaptively chosen.

Illustration of application: Before SMOOTH can be applied to the data, a preliminary marker order has to be established. This map is the starting point for singleton detection and is still far from ideal. Although a singleton is context dependent, the most obvious singletons are clearly perceptible even in less ideal maps.

In Figure 2, the difference *d* for one female gamete in the ultra-dense AFLP map of one chromosome of potato (ISIDORE *et al.* 2003) is shown. This gamete was the result of two recombination events: one recombination event occurred between locus 76 and locus 77 and one recombination event occurred between locus 944 and locus 954. Around the recombination events, the value of *d* approaches 0.5. The most likely singletons ($d = 1$) can be observed at loci 2, 21, 105, 474, 508, 536, 615, 735, 793, 898 and 918.

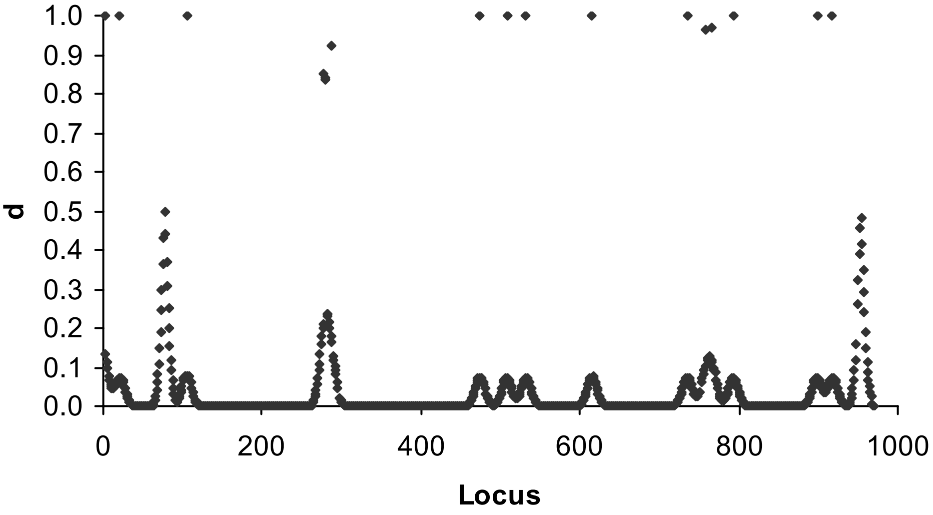


Figure 2. The difference d for all 971 loci in a particular gamete in the ultra-dense potato map. Data from chromosome I (ISIDORE *et al.* 2003).

When the value *d* of each data point is calculated, a threshold for singleton removal can be set. The singletons are removed in an iterative process, alternately using a mapping algorithm and

SMOOTH. In other words, a mapping algorithm like RECORD (VAN OS *et al*. submitted) is used to calculate the marker order, subsequently SMOOTH is used to remove singletons, after which the marker order is recalculated with RECORD, *etcetera*. In principal, all mapping algorithms can be used, but data sets containing 500 markers demand only 20 minutes for analysis with RECORD on a 350 MHz processor. By comparison, JoinMap will take 9 days to calculate a map from a linkage group containing 500 markers on the same type of machine. In the first cycle of the iteration, a high threshold (0.99) is set and the most likely singletons are replaced by missing values. During the following cycles, the threshold is slightly decreased. Using more cycles in the iteration and smaller decreasing steps in the threshold, singleton removal is more accurate. In this experiment, the iteration is continued for 15 cycles while decreasing the threshold with 0.02, until the final threshold of 0.70 is reached. At this point, most singletons are removed from the data. Empirical evidence will be provided below, that at threshold 0.70, the amount of singletons that remain in the data set is in balance with the number of correct data points that are unjustly removed.

After removing all singletons, an unambiguous framework map can be constructed. Subsequently, the original marker data can be fit into the framework map by maximum likelihood, providing a verification of the framework map and also a quality label for each marker.

Simulations are used to demonstrate that the program detects the singletons and that eventually the correct marker order is obtained. The practical applicability is established by the analysis of an experimental data set of potato comprising 971 markers in 130 individuals.

Simulated data: The power of SMOOTH to detect singletons was tested on simulated data. For this purpose several first generation backcross (BC1) populations were generated varying in the number of loci, population size and error percentage as shown in table 1.

Table 1. Values for simulation variables used in the two different simulation experiments

Variables	Experiment I	Experiment II
Map length (cM)	50 (fixed)	50 (fixed)
Number of loci	10, 25, 50, 100, 250, 500	100 (fixed)
Population size	100 (fixed)	50, 100, 150
Percentage scoring errors	1, 2, 3, …, 25%	1, 2, 3, …, 25%
Percentage missing observations	0% (fixed)	0% (fixed)

The simulated data were produced as follows: a given number of loci were randomly positioned (according to a Poisson process) along a single chromosome of 50 cM; genotypes were generated for a BC1 progeny following standard Mendelian segregation (assuming no crossover interference). Errors were randomly introduced in the data set and the positions of these errors were stored in a log-file. The range of error percentages increased from 1% to 25% thus creating 25 data sets for each population. In experiment I emphasis is put on both error percentage and marker density. In experiment II the effect of error percentage and population size is evaluated.

Corrected data sets were obtained from each simulated data set with introduced errors by calculating marker orders with RECORD (VAN OS *et al*. submitted) and removing singletons with SMOOTH. In the mean time, SMOOTH kept track of all the data points that were removed during the mapping and cleaning process. After completion of the process, this list of removed data points was compared with the list of introduced errors. From this comparison the number of errors were counted that were found and missed by SMOOTH. Also the number of correct data points that should not have been removed were counted. The marker order before and after cleaning with SMOOTH was compared with the original simulated map, using Spearman's rank-order correlation coefficient $r_{s,}$ between the expected marker position on the simulated map and the observed marker position on the map calculated by ORD.

Experimental data: Besides the simulations, actual mapping data were analyzed from the ultra-dense genetic map of potato (ISIDORE *et al*. 2003). From the data set of this outbreeding population, the AFLP markers segregating from only one parent were considered. Both parental maps were analyzed separately. In the maternal map, 4187 markers were segregating and 3413 markers segregated in the paternal map. Grouping was done with JoinMap 2.0 and divided the data in 12 groups. A preliminary marker order was used to assign the linkage phase to all markers based on their flanking markers. After linkage phase assignment, the data could be treated as if it were a first generation backcross. This approach, also called two-way pseudo-testcross (GRATTAPAGLIA and SEDEROFF 1994), is commonly applied for map construction in populations descending from non-inbred parents.

Marker ordering was done by RECORD, while SMOOTH cleaned the data from singletons applying the same approach as was used for the simulations.

RESULTS

Simulated data: The utility of SMOOTH in obtaining an accurate marker order was evaluated by simulation experiments. In experiment I, the consequences of error percentage and marker density were assessed. The accuracy of the marker order with and without the application of SMOOTH was examined using the rank correlation coefficient between the calculated marker order and the simulated marker order. The quality of dense genetic maps can be improved considerably by the application of SMOOTH. The value of SMOOTH was most obvious in the data set with the highest marker density in experiment I. The rank correlation coefficients for this data set consisting of 500 loci and 100 individuals are shown in Figure 3. Results are generated for error percentages ranging from 1% to 25%. Rank correlation coefficients are shown for both approaches, i.e. before and after cleaning with SMOOTH. Without SMOOTH, marker orders with intolerable inaccuracy are produced, when more than 5% error is present. However, SMOOTH enables mapping software to calculate accurate maps from data sets with error levels up to 20%. Obviously SMOOTH is able to recognize most of the singletons in the data and enables the mapping software to accurately position the markers.

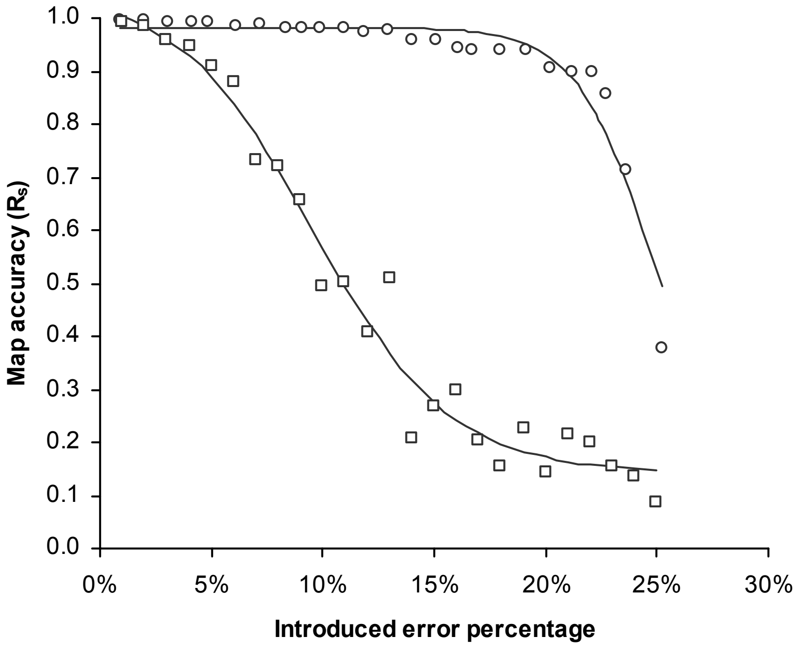


Figure 3. The rank correlation coefficient between the calculated map and the original simulated map before ('□') and after ('○') using SMOOTH for different levels of scoring errors based on simulated data sets with 500 loci on 50 cM and 100 individuals.

To understand the process of singleton removal in detail, the detected singletons were compared with the introduced errors in the data sets. In this comparison we monitored the unjust removal of correct data points and the errors that were not detected by SMOOTH.

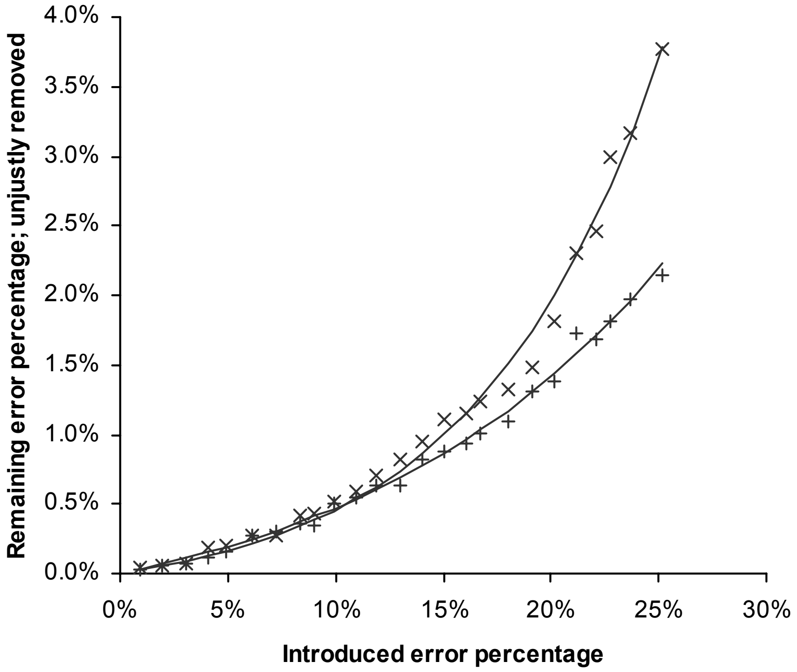


Figure 4. The percentage of remaining errors from the total data set ('×') and the percentage of unjustly removed data points ('+') for different levels of scoring errors based on simulated data sets with 500 loci on 50 cM and 100 individuals.

Figure 4 shows the percentage of errors that were not detected by SMOOTH and the percentage of data points that were unjustly removed from the total amount of data points in the same data set as mentioned in Figure 3. SMOOTH recognizes the vast majority of errors, e.g. at 10% error level, 95% of the errors were detected, reducing the amount of errors to 0.5%. The number of errors that were not detected and the number of data points that were unjustly removed are more or less similar for lower error levels. This indicates that the choice to stop SMOOTH at a final threshold of $d = 0.7$ is justified. By decreasing this threshold even further the number of data points that are unjustly removed would increase and surpass the number of undetected errors.

Close inspection of the position of errors that were not detected or data points that were unjustly removed, revealed that they occurred near recombinations and at the ends of the map. Close to recombination events, the flanking markers at either side of the recombination offer contradicting information. Therefore error detection in the vicinity of recombination events is more complicated. At the ends of the map, the difference between the last recombination or the last singleton can not be determined. Therefore the last recombination event should be confirmed by at least 2 markers distal to that recombination.

Marker density is an important factor to enable error detection, as can be observed from the results of experiment I shown in Figure 5. The percentage of undetected errors is lower in data sets with a higher marker density. This is not surprising because the concept of smoothing genetic linkage data is based on the redundancy in genetic information. In high density data sets, the required amount of 30 neighboring data points at close genetic distance is available, but data sets with 10 markers per linkage group only contain up to 9 neighboring data points over a large distance to predict the marker score.

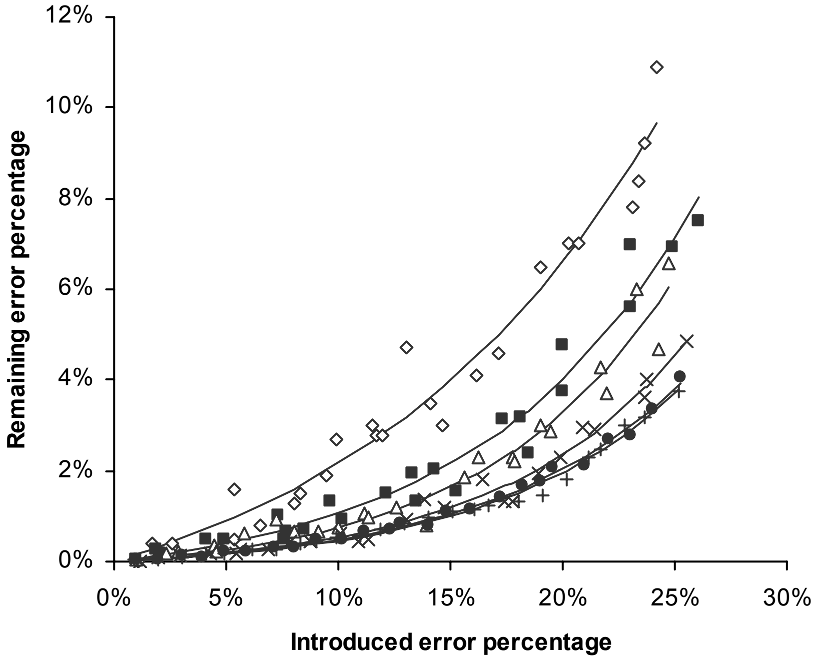


Figure 5. The percentage of remaining errors from the total data set for different levels of scoring errors based on simulated data sets with 100 individuals. The number of loci is indicated by: '◊' for 10; '■' for 25; 'Δ' for 50; '×' for 100; '+' for 250 and '●' for 500 loci.

The effect of population size was analyzed in experiment II. Figure 6 shows that marker ordering is more accurate in larger populations. In fact, this is not the result of applying SMOOTH, but due to the increased performance of the mapping algorithm. As population size increases, more recombination events between a pair of markers can be observed, which adds to the resolution between the markers. The ordering of the markers will be more accurate and the relative impact of missing observations and singletons will decrease. Furthermore, the marker score predictions by SMOOTH will be more precise due to the more accurate order of the markers.

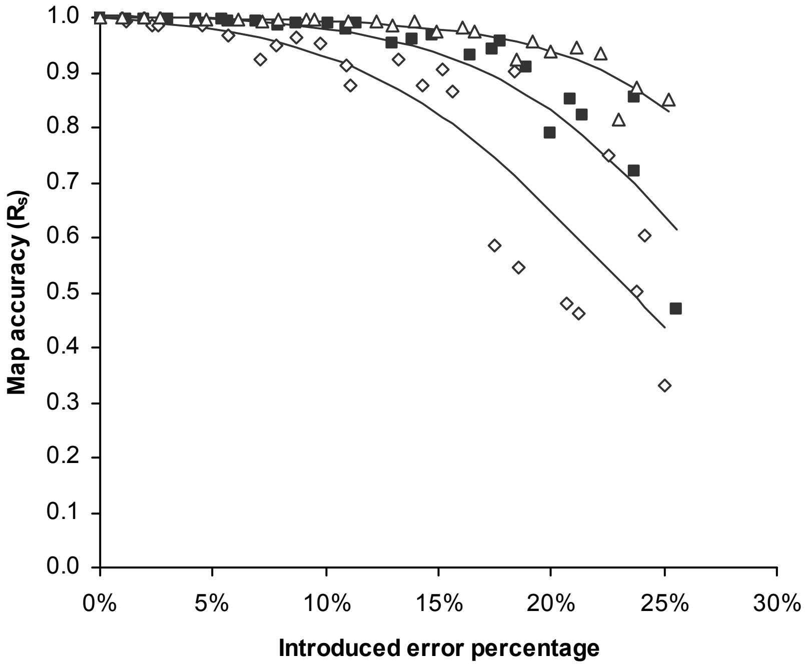


Figure 6. The rank correlation coefficient between the calculated map and the original simulated map after using SMOOTH for different levels of scoring errors based on simulated data sets with 100 loci. The population size is indicated by: '◊' for 50; '■' for 100 and 'Δ' for 150 individuals.

Experimental data: To compare the results from the simulations with real data, the software was tested on a data set from the high-density map of potato (ISIDORE *et al.* 2003). After cleaning the data with SMOOTH, the data were visually inspected for any undetected errors. This revealed a systematic error caused by a group of markers based on AFLP primer combinations from a batch of newly isolated DNA. The confusion of genotypes was solved by removing these individuals from the new set of markers.

When all data ambiguities were removed, a vast amount of redundancy was observed. For instance in chromosome I and IV, a large cluster of cosegregating markers, presumably the centromeric region, contained more than half of the total amount of markers in both the maternal and paternal map. Finally, by deleting the redundant markers from all linkage groups, framework maps were obtained that only consist of unique corrected markers. These markers were converted into bin signatures by restoring all missing values that were not flanked by recombination events. A bin is a unique and most accurate representation of a marker at a certain genetic position. A bin contains at least 1 marker and can not be divided within the given population. Bins are numbered consecutively, based on the recombination events. As a consequence, the bin numbers can be directly translated into map units. Both parental framework maps were free from ambiguities and all the original markers were fit into the most likely bin by maximum likelihood. The map was inspected for possible inconsistencies with the original markers and some minor corrections were made to the bins. Redundant and empty bins were removed; bins that appeared to contain a recombination event were split up and missing values in the bin signature were restored if possible.

To illustrate the difference between a framework map as described above and a conventional map obtained with JoinMap or RECORD, two linkage groups are shown in Figure 7. These linkage groups were derived from the high-density map of potato and represent the paternal map of chromosome III and IX respectively. Linkage group III comprised 124 AFLP markers and linkage group IX comprised 190 AFLP markers. No clustering of markers was observed for linkage group III, but linkage group IX contained a centromeric cluster of 27 cosegregating markers. The marker order from RECORD is basically the same as the order in the framework map. However, four markers with an exceptionally high number of scoring errors are positioned at the end of the linkage group; a commonly observed artifact of mapping software. Major ordering ambiguities can be observed around the centromeric cluster in linkage group IX. JoinMap produces a map which is in length roughly similar to the framework map. However, some map inflation can be observed at both ends of the linkage groups. Ordering ambiguities are more abundant in marker dense areas: markers from the centromeric cluster with elevated levels of singletons are pushed away from the centromeric region and dispersed towards the distal ends of the map.

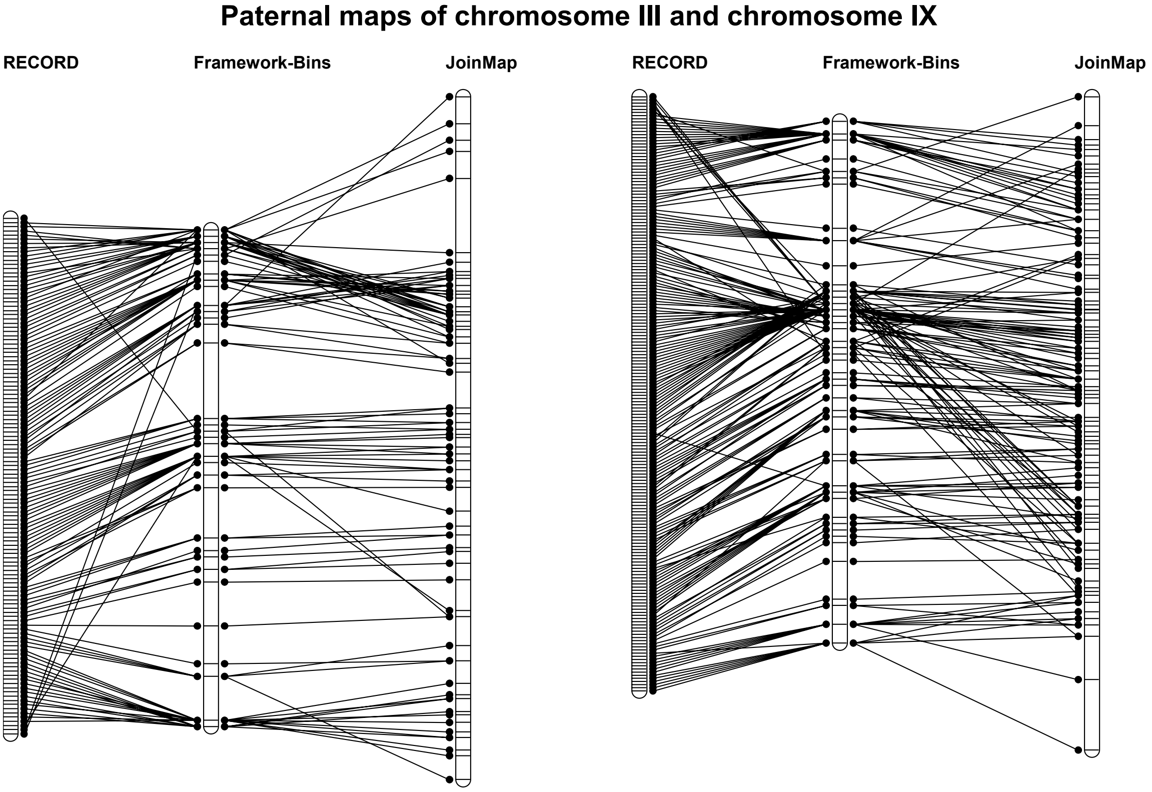


Figure 7. Comparison between three methods of linkage map construction on two different linkage groups. The framework map in the middle is obtained by SMOOTH and the original markers were fit in the bins. Flanking maps have been constructed from the original data set by RECORD and JoinMap. The paternal map of potato linkage group III from the high-density map of potato (ISIDORE *et al.* 2003) is shown on the left, the paternal map of potato linkage group IX is shown on the right. Relative marker positions are displayed by aligning the results of the three methods of linkage map construction. RECORD produces a marker order; distances are proportional to the number of markers. The distances between the bins in the framework map depend on the number of recombination events, which are transformed into centiMorgans. The markers on the map produced by JoinMap are displayed at their corresponding cM position.

In the data set of the high density genetic linkage map of potato, the number of singletons for each marker was calculated by comparing the original data of each marker with the signature of its most likely bin (see Figure 8). The average number of singletons per marker was 3.9 in 130 individuals (3.0%). In contrast with the simulations, the distribution of singletons in the experimental data was not random. In the maternal map, one third of the markers did not contain any singletons, which provides a verification for the framework map. However, some markers contained up to 38 singletons. In fact, 10% of the markers were responsible for more than half of the scoring errors. Despite the fact that in reality singletons are not randomly distributed, SMOOTH was able to detect them to enable the construction of a solid framework map.

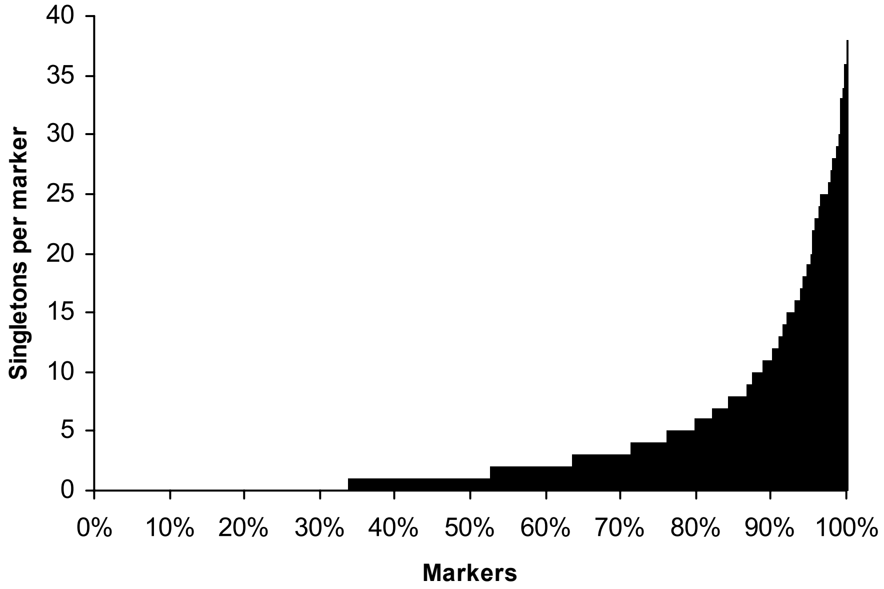


Figure 8. Data quality of the ultra dense map of potato (ISIDORE *et al.* 2003). The number of singletons per marker is calculated after determining the most likely position of the markers on the framework map. The markers are sorted from left to right according to their increasing amount of singletons.

DISCUSSION

Singletons, whether or not caused by biological phenomena or human error, seriously hamper high-density genetic linkage map construction. For calculating a reliable linkage map or marker order, these singletons have to be removed. We devised a statistical method to detect and remove singletons from high-density genetic linkage data. The approach is based on predicting marker scores on the basis of the available neighboring data points, which are more abundant in denser maps. Although in denser maps the rising amount of errors becomes increasingly difficult to handle with current mapping software, this new method takes advantage of the redundancy in high density data sets. The excess of markers within a close genetical range, are the basis of a reliable estimate of the marker score. By removing highly unlikely marker scores from the data, the true recombination events will remain in the data and facilitate marker ordering.

SMOOTH has been extensively tested on simulated data. The results have provided convincing evidence that more than 95% of the singletons can be detected. With a large amount of errors present in the data, a reliable and accurate map can only be constructed when applying SMOOTH. Therefore we conclude that the program has great utility in high density mapping, which is proven by the successful application to the experimental data set of potato for the construction of an accurate framework map.

It is advisable to use SMOOTH for data sets with at least 100 loci per linkage group, because the error detection is dependent on the amount of neighboring markers. Although the program is therefore not intended for globally smoothing low-density maps, it can be useful in cleaning up marker dense clusters in low density maps. These marker clusters are regularly observed in genetic maps (STROMMER *et al*. 2002) and are often situated around the centromere of the chromosome where recombination is suppressed.

The error detection works less on the two distal ends of the chromosome and close to recombination events. Here, the predicted value of the data points is based on two sets of data points with contradicting information. In these situations, there is a risk of removing data points that are correct. However, the consequences of removing too many data points are not severe. In fact, the removal of correct data points in the vicinity of recombinations causes a local decrease of the effective population size and has therefore the same effect as the removal of an individual offspring genotype from the mapping population. The consequences of these unjustified removals can be solved by correcting the framework map using the original data. This verification of the framework map is done by maximum likelihood comparison of the original markers with the framework bins. Moreover, the risk of cleaning data points that were not erroneous is sufficiently reduced by employing the method in an iterative process with the mapping algorithm.

The verification of the framework map by refitting the original data does not provide indisputable evidence for the true marker order. Nevertheless, it provides a detailed overview of the ambiguities in the data. The accuracy of the ultra-dense marker order can only be assessed in simulation studies where the true marker order is known. For potato, the consistency of the genetic map with a physical map is expected to provide the evidence for the current marker order.

The program has been applied for the construction of the ultra-dense genetic linkage map of potato. All linkage groups of this map contain more than 100 markers. Accurate mapping of these large linkage groups was not possible, despite the even small amounts of scoring errors. Most of these errors could be erased by manual re-evaluation of the AFLP gels, but in spite of these time-consuming efforts, accurate marker ordering was still severely complicated. With SMOOTH, the ambiguities of the data were removed to construct a framework map that provided accurate marker placement.

To a certain extent, error detection is available in the current version of MapMaker (LINCOLN and LANDER 1992). Instead of removing possible errors, MapMaker takes the possibility for a data point to be erroneous into account and avoids potential map inflation. The errors remain in the data set and still cause ordering problems, therefore MapMaker is not adequate to calculate high density maps.

Besides backcross populations, the concept of SMOOTH can also be suitable for analyzing other populations like F_2. Dominance will nevertheless decrease the detection power of singletons. In this case, the marker density should be higher than in backcross populations to ensure a reliable singleton detection.

In conclusion, with the advent of ultra-dense genetic linkage maps, a completely new approach of data analysis is required. In combination with RECORD (VAN OS *et al*. submitted), this method provides a fast and accurate way of positioning genetic markers along an unambiguous framework map.

ACKNOWLEDGEMENTS

We are very grateful to Dr. Fred van Eeuwijk for providing valuable comments. This work was carried out under the EU FAIR programme grant: FAIR5-PL97-3565.

LITERATURE CITED

GRATTAPAGLIA, D., and R. SEDEROFF, 1994 Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers, Genetics **137**: 1121-1137.

ISIDORE, E., H. VAN OS, S. ANDRZEJEWSKI, J. BAKKER, I. BARRENA, G. J. BRYAN, B. CAROMEL, H. J. VAN ECK, B. GHAREEB, W. DE JONG, P. VAN KOERT, V. LEFEBVRE, D. MILBOURNE, E. RITTER, J. ROUPPE VAN DER VOORT, F. ROUSSELLE-BOURGEOIS, J. VAN VLIET and R. WAUGH, 2003 Toward a marker-dense meiotic map of the potato genome: Lessons from linkage group I. Genetics **165**: 2107-2116.

KLEIN, P. E., R. R. KLEIN, S. W. CARTINHOUR, P. E. ULANCH, J. DONG, J. A. OBERT, D. T. MORISHIGE, S. D. SCHLUETER, K. L. CHILDS, M. ALE and J. E. MULLET, 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.

LANDER, E. S., P. GREEN, J. ABRAHAMSON, A. BARLOW, M. J. DALY, S. E. LINCOLN and L. NEWBURG, 1987 MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics **1**: 174-181.

LINCOLN, S. E. and E. S. LANDER, 1992 Systematic detection of errors in genetic linkage data. Genomics **14**: 604-610.

NILSSON, N. O., T. SÄLL and B. O. BENGTSSON, 1993 Chiasma and recombination data in plants: are they compatible? Trends Genet **9**: 344-348.

STAM, P., and J. VAN OOIJEN, 1995 JoinMap version 2.0: Software for the calculation of genetic linkage maps. CPRO-DLO, Wageningen.

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

STROMMER, J., J. PETERS, J. ZETHOF, P. DE KEUKELEIRE and T. GERATS, 2002 AFLP maps of Petunia hybrida: building maps when markers cluster. Theor Appl Genet **105**: 1000-1009.

VAN OS, H., P. STAM, R. G. F. VISSER and H. J. VAN ECK, 2005 RECORD: a novel method for ordering loci on a genetic linkage map. Theor Appl Genet (accepted for publication).

VOS, P., R. HOGERS, M. BLEEKER, M. REIJANS, T. VAN DE LEE, M. HORNES, A. FRIJTERS, J. POT, J. PELEMAN, M. KUIPER and M. ZABEAU, 1995 AFLP: a new technique for DNA fingerprinting. Nucl. Acid Res. **23**: 4407-4414.

Chapter 5:

A 10,000 Marker Ultra-Dense Genetic Recombination Map as a New Tool for Anchoring a Physical Map and Fast Gene Cloning in Potato

Hans van Os, Sandra Andrzejewski, Erin Bakker, Imanol Barrena, Glenn J. Bryan, Bernard Caromel, Bilal Ghareeb, Edwige Isidore, Walter de Jong, Paul van Koert, Véronique Lefebvre, Dan Milbourne, Enrique Ritter, Jeroen N. A. M. Rouppe van der Voort, Françoise Rousselle-Bourgeois, Joke van Vliet, Robbie Waugh, Jaap Bakker, Richard G. F. Visser and Herman J. van Eck

Submitted

ABSTRACT

An ultra-dense genetic linkage map with nearly 10,000 AFLP loci was constructed from a heterozygous diploid potato population. It is among the densest meiotic maps ever constructed. A fast marker ordering algorithm was used in combination with genotyping error-detection software to obtain “skeleton bin maps”. A “bin” is a position on the genetic map with a unique segregation pattern and separated from adjacent bins by a single recombination event. Subsequently all marker loci were assigned to the bins on the map by maximum likelihood based on their original segregation pattern. From the markers that were heterozygous in either the maternal or paternal parent, 98% could be fit in the bins. Of the markers that were heterozygous in both parents (bridge markers), only 79% could be fit. In both parental maps the twelve chromosomes could be identified. In addition, the paternal map includes a small unassigned linkage group with a severe segregation distortion. Recombination frequencies or marker positions are non-randomly distributed across the map. Putative centromeric regions showed extensive marker clustering while putative recombination hot spots resulted in large intervals up to 15 cM without markers. The markers derived from EcoRI/MseI and SacI/MseI primer combinations clustered more frequently than those derived from PstI/MseI primer combinations. The distribution of recombination events per chromatid indicated an overall absence of the occurrence of 0 chiasmata and excluded absolute chiasma interference within arms. The ultra-high density map has been used for anchoring of BAC-contigs for a sequence ready potato physical map and gene cloning.

INTRODUCTION

Genetic linkage maps constitute a necessary prerequisite to study the inheritance of both qualitative and quantitative traits, to develop markers for marker-assisted breeding (MAB) and for map-based gene cloning. Multi-locus molecular marker techniques, such as AFLP (VOS *et al.* 1995), can be used to generate large numbers of markers in a relatively short time, facilitating the construction of dense genetic linkage maps. High-density genetic linkage maps have already been constructed in crop plant species such as rice (HARUSHIMA *et al.* 1997: 2,275 markers), maize (VUYLSTEKE *et al.* 1999: 1,539 and 1,355 markers mapped in two populations), wheat (BOYKO *et al.* 2002: 732 markers), potato and tomato (TANKSLEY *et al.* 1992: ca. 1,000 markers; HAANSTRA *et al.* 1999: 1,175 markers), pepper (PARAN *et al.* 2004: 2262 markers mapped in 6 populations), sorghum (BOWERS *et al.*, 2003: 2512 markers), cotton (RONG *et al.*, 2004: 3347 markers) and papaya (MA *et al.*, 2004: 1501 markers).

A genome-wide ultra-dense genetic map results in the global saturation of the genome with marker loci, which if concentrated on a single mapping population, can be useful for all other mapping applications. Usually, map-based cloning of genes responsible for interesting traits requires local marker saturation around the target gene. This targeted marker saturation is generally achieved with Bulked Segregant Analysis (MICHELMORE *et al.* 1991). Ultra-dense genetic maps avoid this time consuming and costly step, which has to be achieved in separate experiments for every trait locus targeted. Moreover, expected average between-marker distances that are smaller than the average insert length of a BAC library generally allow chromosome landing (TANKSLEY *et al.* 1995). In addition, ultra-dense genetic maps also facilitate the genetic anchoring of a physical map. If large-insert genomic clones and contigs can be directly identified with markers from an ultra-dense genetic map, they can be anchored to their corresponding positions in the genome. Besides these applications, the ultra-dense map will become the reference map that facilitates marker exchange and map alignment within the research community working on any given organism, provided that the marker information contained within the map is transferable to other genotypes or populations. High transferability of AFLP markers between populations has been amply demonstrated by using the AFLP catalogue for potato (ROUPPE VAN DER VOORT *et al.* 1997) and barley (WAUGH *et al.* 1997; QI and LINDHOUT 1997). The transferability of other single locus marker types, such as RFLPs, STSs and SSRs, is more obvious and has therefore not been questioned.

The construction of ultra-dense genetic linkage maps has been confronted with two major problems. Currently available computer programs for linkage mapping are incapable of handling data sets of several thousands of markers, and results in prohibitively long calculation times. Moreover, even small frequencies of scoring error result in high rates of ordering ambiguities between markers within short genetic distances. Two recently developed computer programs, referred to as RECORD (VAN OS *et al.* 2005a) and SMOOTH (VAN OS *et al.* 2005b), have tackled these problems. RECORD employs a marker-ordering algorithm based on minimization of the total number of recombination events in any given marker order (VAN OS *et al.* 2005a). SMOOTH is a statistical genotyping error removal utility that calculates the probability of a data point being a ‘singleton’, based on neighboring marker information. A singleton appears to be the result of an apparent double recombination event at either side of a single marker locus. More likely singletons represent artifacts due to scoring errors, technical or biological phenomena such as methylation polymorphisms and gene conversion. The observation of singletons depends on their context of flanking markers. Therefore, singletons are removed in an iterative process, singleton removal, re-ordering of markers, singleton removal, re-ordering etc., thereby gradually relaxing the statistical threshold of singleton identification (VAN OS *et al.* 2005b). The loss of a few percent of the data is obviously less damaging to the map, than having similar levels of genotyping errors. Using these two computer programs results in a framework of ordered ‘bins’ in which all recombination events in the population have been identified. A ‘bin’ is a position on the genetic map with a unique segregation pattern and is separated from adjacent bins by a single recombination event. This ordered set of ‘bins’ is considered to be a ‘skeleton bin map’ to which all original marker data can be fit, using a maximum likelihood method. This approach also provides a quality estimate for each marker that is based on the deviation between the observed marker segregation pattern and the expected segregation pattern as defined by the position of the bin in the skeleton bin map. A bin may contain a number of co-segregating markers and is defined by a segregation pattern. This pattern is called the ‘bin signature’, and it represents an accurate genetic position on the map within a given population. The unit of distance of the skeleton bin map is expressed in recombination events. In saturated linkage maps all recombination events are captured. As a consequence, application of the Kosambi mapping function is not necessary to compensate for unnoticed double recombination events. A more comprehensive description of the method is provided in ISIDORE *et al.* (2003) and VAN OS *et al.* (2005a, 2005b) and is illustrated by Figure 1.

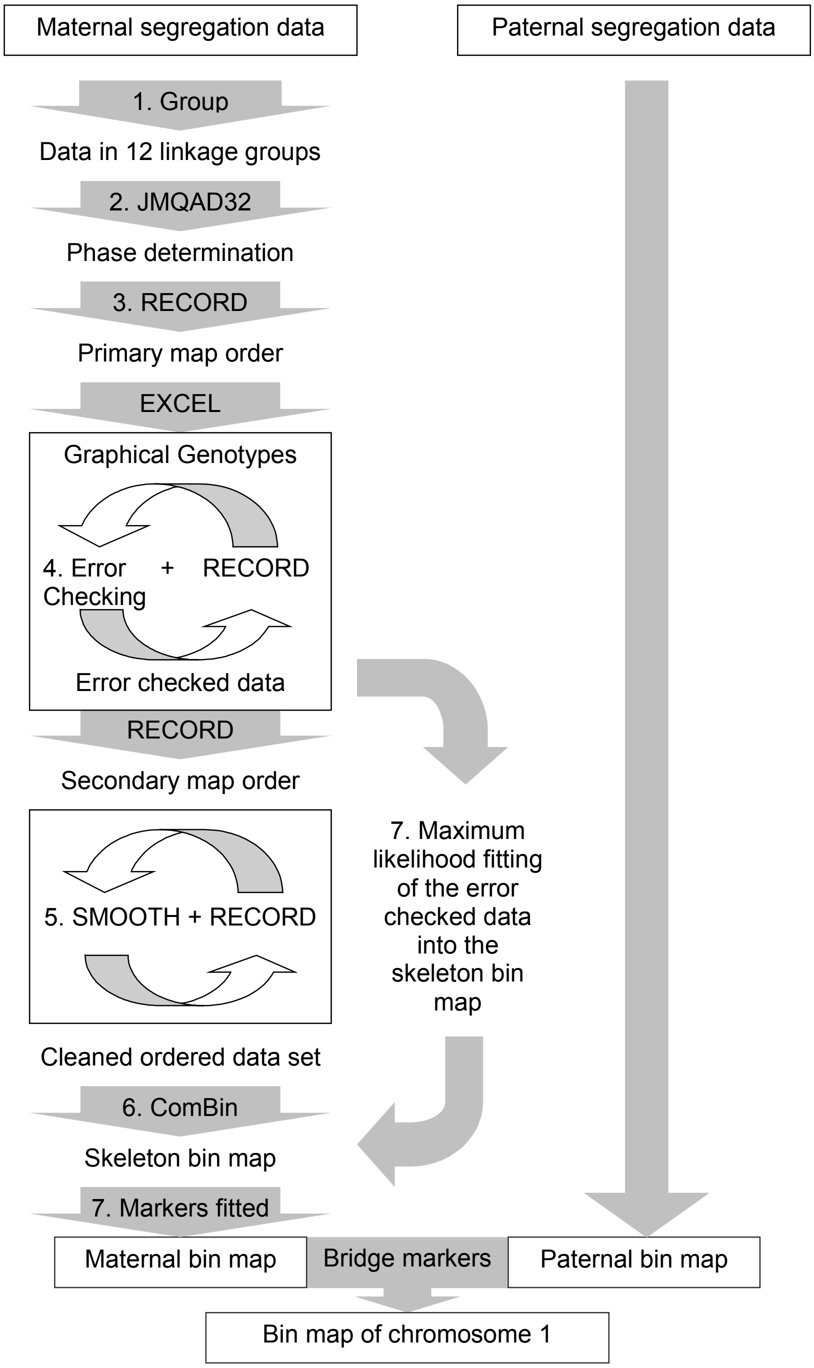


Figure 1: Overview of the method used to construct the ultra dense map of potato as described by ISIDORE *et al.* (2003). The numbered gray arrows represent the computer programs used.

In this paper, we present - to our knowledge - the densest meiotic linkage map yet produced for any species. The ultra-dense map of potato covers all linkage groups and contains nearly 10,000 markers in total. The non-random pattern of marker distribution provides insight into the positions of putative recombination hot spots and centromeric regions. The distribution of recombination events per chromatid provides information on chiasmata. Given an estimated genome size of 840 Mb (BENNETT *et al.* 1997), and assuming random marker distribution, this level of marker saturation will expedite all map based cloning efforts in potato, as well as the anchoring of BAC contigs for the construction of a sequence-ready potato physical map.

MATERIALS AND METHODS

Plant material: A cross between two diploid heterozygous potato clones, SH83-92-488 × RH89-039-16 (hereafter referred to as SH × RH) resulted in an F1 mapping population of 136 individuals. The same mapping population has been used to clone the nematode resistance gene Gpa2 against Globodera pallida (VAN DER VOSSEN *et al.* 1998), and the Phytophthora infestans R-gene R3a (HUANG *et al.* 2004; HUANG *et al.* 2005). Genomic DNA was extracted from frozen leaf tissue according to VAN DER BEEK *et al.* (1992).

Marker analysis: AFLP markers (VOS *et al.*, 1995) were generated with templates of three different restriction enzyme combinations, *Eco*RI/*Mse*I, *Sac*I/*Mse*I and *Pst*I/*Mse*I, and by applying three selective nucleotides to AFLP primers at the *Eco*RI, *Sac*I and *Mse*I side and two selective nucleotides to the primers at the *Pst*I side. A total of 381 primer combinations, listed at http://potatodbase.dpw.wau.nl/UHDdata.html, were used to generate markers. Amplification products were separated by electrophoresis and visualized by autoradiography as described in ISIDORE *et al.* (2003).

The autoradiograms were analyzed manually or with the aid of the computer program Cross-Checker (BUNTJER 2000b), which is available at http://www.dpw.wur.nl/pv/. The names of the markers indicate the enzymes used, the selective nucleotides and the size of the fragment; for instance EAACMCAA_507.0 is an AFLP marker derived from a primer combination with enzymes *Eco*RI and *Mse*I, selective nucleotides AAC and CAA, and a mobility that corresponds to a fragment with an estimated size of 507.0 basepairs. Fragment mobility estimates were inferred relative to a 10-base ladder (Sequamark, Research Genetics) using reference gels provided by Keygene NV, Wageningen, Netherlands. Assigning linkage groups to the 12 potato chromosomes was done with a set of AFLP markers with known position (ROUPPE VAN DER VOORT *et al.* 1997) and other markers, including RFLPs, SSRs, CAPS and SCARs.

Map construction: The marker data were split into three sets based on their segregation type. Markers that were heterozygous in the maternal parent (SH) and absent in the paternal parent (RH) were scored as <ab×aa>; 'paternal' markers heterozygous in RH and absent in SH were scored as <aa×ab>; markers segregating in both parents were denoted <ab×ab>. The maternal and paternal data sets were divided into 12 linkage groups with module GROUP, included in JoinMap 2.0 (STAM and VAN OOIJEN 1995).

Addition of the <ab×ab> markers resulted in the merger of initially separate parental groups due to spurious linkage caused by erroneous markers. Erroneous markers result from non-allelic bands of identical mobility. Such alleles are superimposed on gel, and in this way two markers <ab×aa> and <aa×ab> are perceived as one single <ab×ab> marker, drawing two unrelated paternal groups into one artifactual group. Similarly, artifactual markers can also result from two markers from the same parent. In total 65 of these 'sticky' markers were removed manually to ensure a stable grouping down to a LOD threshold of 6.

A preliminary marker order and the linkage phase was calculated with the 'quick and dirty' mapping module JMQAD32 from JoinMap 2.0 (STAM and VAN OOIJEN 1995). This algorithm calculates the marker order by minimizing the sum of adjacent recombination frequencies (SARF). It is the fastest algorithm available and sufficiently accurate to determine whether markers are linked in coupling phase or in repulsion. Linkage phase between linked markers was determined and indicated in the dataset following the format rules of JoinMap

2.0 for a first backcross population. Information on linkage phase is required to obtain the correct recombination frequency between markers linked in repulsion.

The order of markers in the linkage groups was then re-calculated with RECORD (VAN OS *et al.* 2005a) which requires data in 'BC1' format. After this second ordering of the markers, the data were displayed in map order as a color-coded 'graphical' genotype in Microsoft Excel using a conditional cell formatting formula. Using this display, singletons could be marked easily. They were re-evaluated by visual inspection of the autoradiograms and corrected if necessary.

The corrected data was ordered for a third time with RECORD and remaining singletons were removed with SMOOTH (VAN OS *et al.* 2005b) in iterations with RECORD.

The program ComBin (BUNTJER *et al.* 2000a; available at http://www.dpw.wur.nl/pv/) was used for final inspection. ComBin removes the redundancy due to co-segregating markers and draws connections between non-redundant marker bins without the assumption that a chromosome is a linear structure. Side branches result from singletons, and any alternative connection between pairs of markers (or bins) is allowed as well. When non-linear structures were visualised by ComBin, further data inspection was performed. When ComBin analysis results in a linear figure, it can be concluded that the linkage group is free from data ambiguities. When all ambiguities identified with Combin have been replaced with missing values, the co-segregating markers are used to infer 'bin signatures'. A bin signature comprises the consensus segregation pattern of marker loci, which do not recombine and are thus incorporated in the bin. The resulting bins form a 'skeleton bin map' of the potato linkage groups. Subsequently the bins are filled with marker loci. Please note that marker loci represent the real observed segregation data, including ambiguous data points, whereas the bin signatures represent the least ambiguous consensus segregation obtained so far.

The mapping of the bridge markers, which are heterozygous in both parents <ab×ab>, is based on the information offered by the skeleton bin map. When the telomeric maternal and paternal bin signatures are superimposed (<ab×aa> 1:1 + <aa×ab> 1:1 = <ab×ab> 3:1), a putative bridge bin signature results. This method of postulation of all putative bridge bin signatures follows the method of the two-way pseudo-testcross proposed by GRATTAPAGLIA and SEDEROFF (1994) in reverse direction. Depending on linkage phase in coupling or repulsion of the parental markers ({0-} or {1-}, and {-0} or {-1}), the postulated bridge bin can take four alternative 3:1 segregation patterns as bridge bin signature ({00}, {01}, {10} and {11}). The bridge markers were fit into the putative bridge bins by maximum likelihood. A LOD threshold of 15 ($p < 0.001$) was used to avoid false positive assignment of bridge markers to bridge bins. This threshold was determined by a permutation test. After fitting 10,000 random markers into the bins, less than 0.1% of the markers fit into the framework map with LOD score higher than either 4 or 15, for 1:1 and 3:1 segregating markers respectively. Chromosome orientation follows DONG *et al.* (2000) with the short arm north and the long arm south, except for the linkage groups homologous to chromosome 7, 11 and 12, which are in opposite orientation.

RESULTS

Markers and progeny: The diploid mapping population SH × RH, comprising 136 individuals, was analyzed with a total of 381 AFLP primer combinations derived from three different enzyme combinations. A total number of 10,305 clearly scorable markers were recorded. Additional SSR, CAPS, RFLP, SCAR and phenotypic marker loci were analyzed on the population, which raised the number of markers to 10,365. This implies a data set of 1.4 million data points.

After inspection of the data, one offspring individual (SH×RH#153-D3) was abandoned, because it frequently displayed AFLP fragments that were not present in either of the parents. Another individual (SH×RH#160-D6) contained approx. 50% more bands than the other individuals, probably due to DNA contamination from a sibling and was also excluded from the analysis. Analysis of the similarity between the individuals revealed 4 duplicates (SH×RH#64-B8 with SH×RH#73-B11; SH×RH#166-D9 with SH×RH#167-D10; SH×RH44-J1 with SH×RH45-J2; and SH×RH86-L12 with SH×RH88-M2), which possibly results from errors during the clonal propagation of the population. The data of the duplicates were merged and conflicting scoring data were removed. The genetic resolution of the population and thus the unit for map distance is therefore based on 130 informative individuals.

Mapping: The total data set was split into maternal, paternal and biparental data sets (Table 1). Among the total number of 10,365 markers, 4187 were segregating due to polymorphism in the maternal parent <ab×aa>, 3413 segregated from the paternal parent <aa×ab>, and 2765 markers were heterozygous in both parents. The latter type of markers, being referred to as bridge markers <ab×ab>, were used to align the maternal and paternal maps. Summation of the parental-specific markers and the bridge markers resulted in 6952 maternal loci and 6178 paternal loci.

Table 1: Number of markers per enzyme combination per parent

AFLP enzyme combination	Segregation type (1:1 or 3:1) and parental zygosity			Total number of markers	Number of primer combinations per enzyme combination	Average number of markers per primer combination
	Ab×aa SH-marker	aa×ab RH-marker	ab×ab bridge marker			
*Eco*RI/*Mse*I	2558	2099	1746	6403	208	31
*Sac*I/*Mse*I	754	690	523	1967	79	25
*Pst*I/*Mse*I	842	604	489	1935	94	21
Other markers	33	20	7	60		
Total	4187	3413	2765	10365	381	27

The maternal data set could be split into 12 linkage groups at a LOD threshold of 6. For the paternal data, linkage groups II through XI were obtained at LOD 6, but linkage groups I and XII remained associated up to a LOD threshold of 12. This was due to coincidental correlation between the segregation patterns of loci in these two groups. The 24 linkage groups from SH and RH were aligned with the expected potato chromosome. In addition to the 12 known paternal linkage groups, a small highly skewed unassigned linkage group was obtained which contained only 13 markers with a length of approx. 10 cM. This group was heterozygous in the paternal clone (RH), and unassigned (U) to any particular chromosome and is therefore referred to as RHU. The linkage group RHU was omitted from further analyses.

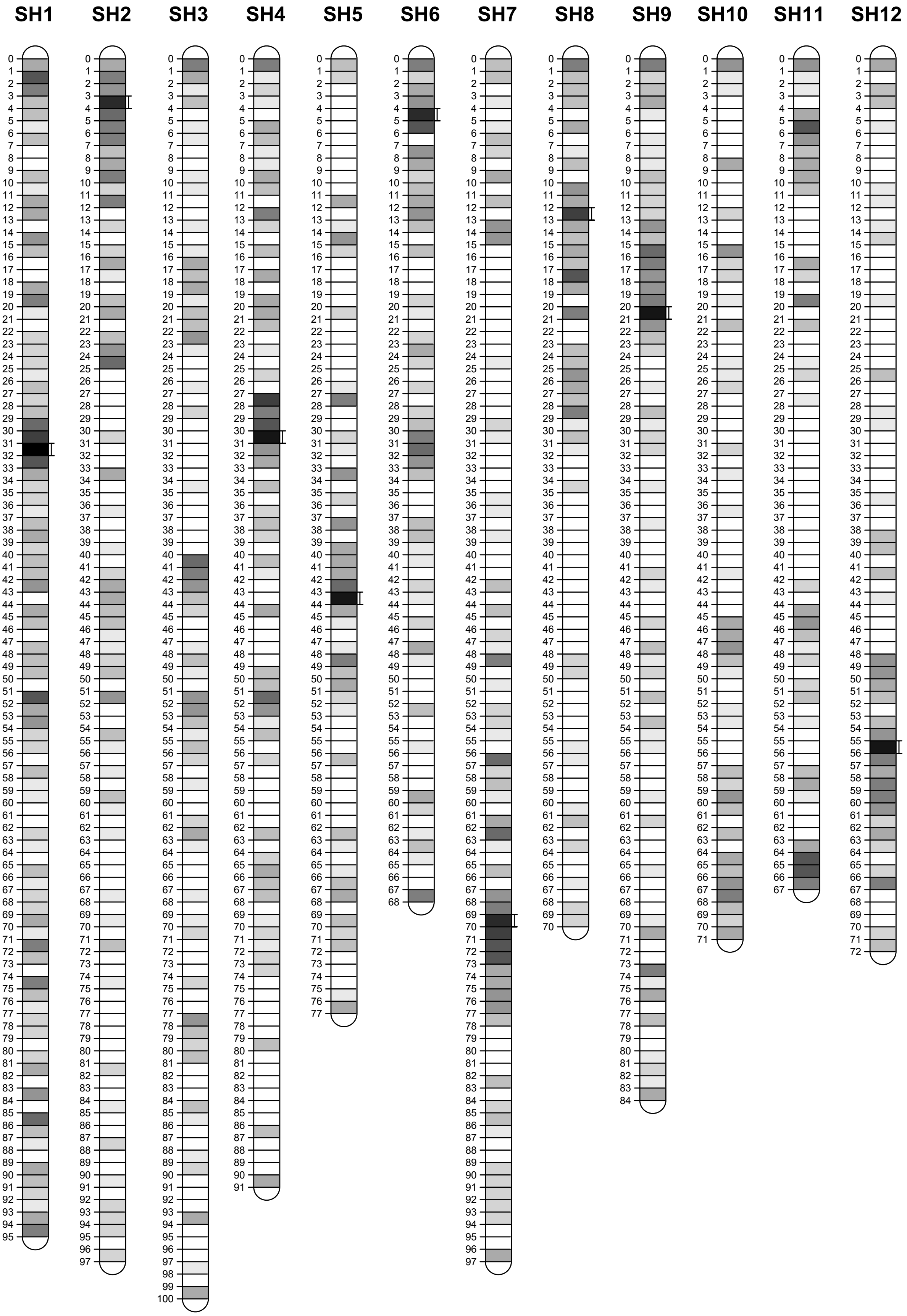
The 24 data sets representing the different linkage groups from both parents were subjected to re-examination for putative scoring errors and to statistical identification of singletons using the computer programs RECORD, and SMOOTH as described in ISIDORE *et al.* (2003). During data inspection, we noticed one individual (SH×RH57-J12) where virtually all markers from the paternal parent in the linkage group corresponding to chromosome VIII were present. This phenomenon is probably due to non-disjunction of this chromosome in the first meiotic division, resulting in a trisomic state. The resulting systematic errors were replaced with missing values. Following the mapping method described above and in ISIDORE *et al.* (2003), marker data resulted into a skeleton bin map.

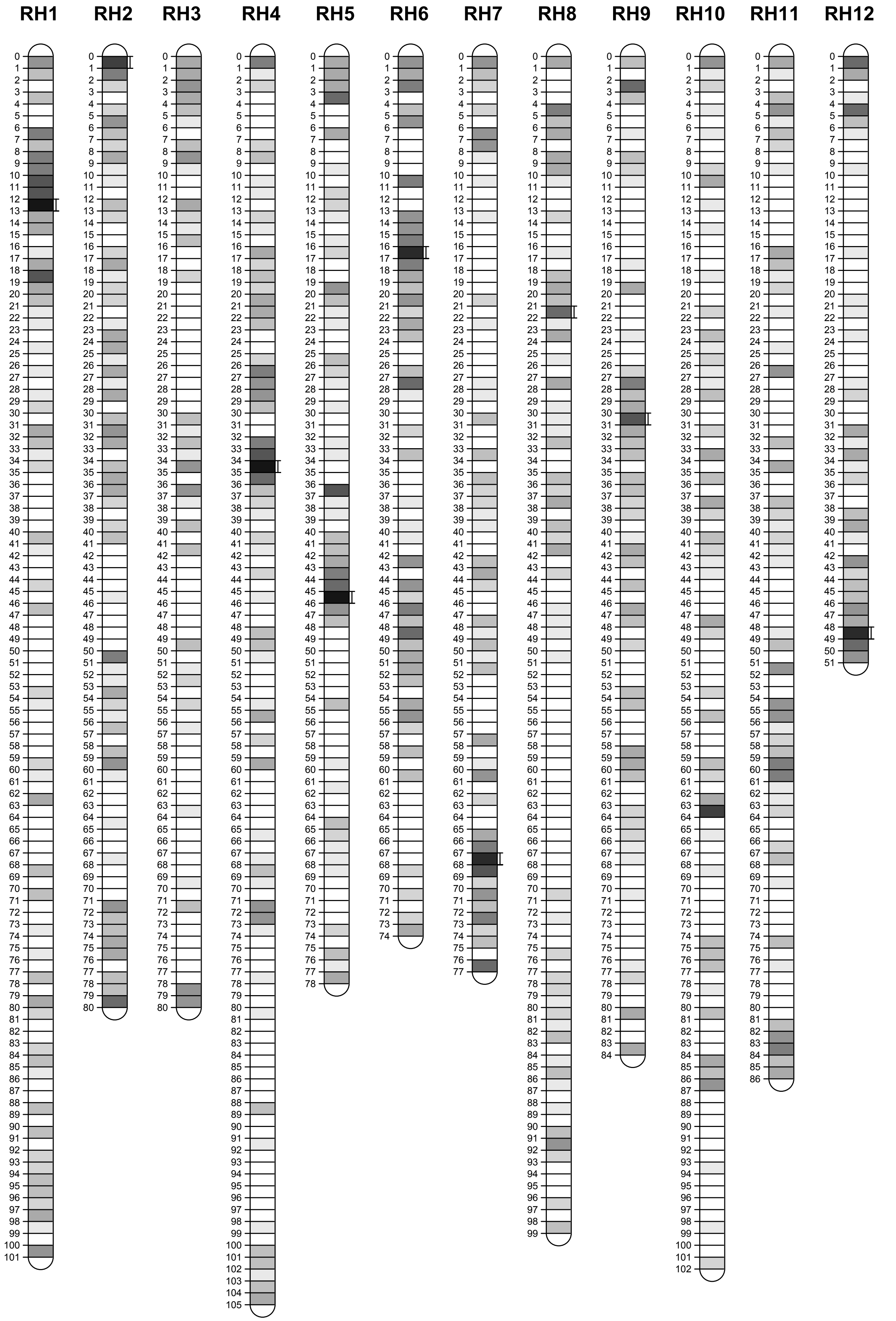
Skeleton bin map: Twelve maternal and twelve paternal skeleton maps deduced from bin signatures provide a representation of the recombination events captured in this mapping population. In total, 569 maternal and 549 paternal bin signatures were obtained. Most adjacent bin signatures differ by only one offspring genotype score, which represents the recombination event between the adjacent bins. In other cases, bin signatures differed for two or more offspring genotype scores, suggesting two or more recombinations between adjacent bins. Inclusion of empty bins, to accommodate for multiple recombination events between marker loci resulted in a skeleton bin map spanning 977 and 1005 recombination events in the maternal and paternal map, respectively (Table 2, Figure 2). All bins, including the empty bins, are numbered consecutively. With 130 offspring in the mapping population one bin represents 1/130 cM. Hence, the genetic length of the parental maps is 767 cM for the maternal map and 773 cM for the paternal map.

Table 2: Overview of the number of markers, bins and recombination events per parent and per linkage group.

Potato Chromosome	SH maternal map			RH paternal map			ab×ab Bridge markers	Total marker number
	Markers	Filled Bins	Rec. events	markers	Filled bins	Rec. events		
I	971	77	94	634	56	100	270	1875
II	311	53	96	262	54	79	145	718
III	193	54	99	124	33	79	145	462
IV	493	55	90	385	57	104	198	1076
V	279	38	76	359	44	77	278	916
VI	265	43	67	366	44	73	180	811
VII	386	54	96	270	43	76	144	800
VIII	209	37	69	155	51	98	106	470
IX	314	53	83	190	44	83	168	672
X	130	36	70	164	47	101	179	473
XI	200	33	66	175	46	85	240	615
XII	367	36	71	237	30	50	172	776
Unassigned markers	69			92			540	701
Total	4187	569	977	3413	549	1005	2765	10365

Fitting of original data into the skeleton bin map: The original scoring data (after the manual verification of singletons) were fit into the bins of the skeleton bin map by maximum likelihood. Subsequently, the marker content of every bin was examined. Application of SMOOTH to remove singletons may have resulted in unjust removal of correct data, and thus causing a reduction in the effective population size. This visual inspection of the original scoring data, specifically near the position of the recombination events, allowed for the correct repositioning of markers into adjacent empty bins. In this way, the unjust removal of putative singletons by SMOOTH is restored. Obviously, after these final improvements to the skeleton bin map, the marker data had to be fit into the bins again.


SH1
SH2
SH3
SH4
SH5
SH6
SH7
SH8
SH9
SH10
SH11
SH12


RH1
RH2
RH3
RH4
RH5
RH6
RH7
RH8
RH9
RH10
RH11
RH12
0
101
80
80
105
78
74
77
99
84
102
86
51

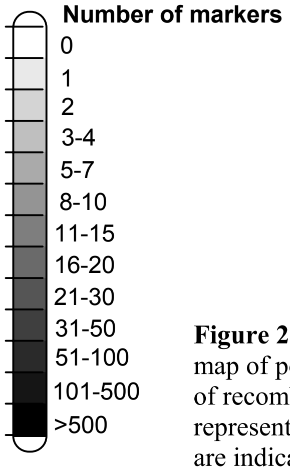


Figure 2. (pages 68-69) The distribution of AFLP markers on the ultra-dense genetic linkage map of potato. The number on the left of the linkage group indicates the cumulative number of recombination events counted from the top. The number of markers in each bin is represented by shades of gray according to the color legend. Putative centromere positions are indicated with "I" alongside the chromosome.

Not all markers, however, were allocated to map positions. From the maternal markers, 46 markers did not reach the threshold of LOD 4, another 22 markers were manually deleted from the most skewed bin SH05B044 and 1 marker fit into two linkage groups with equal likelihood. From the paternal markers, 45 markers did not reach the threshold of LOD 4, 46 markers were removed from bin RH12B049 and also 1 marker could not be fit into one bin unambiguously.

Of the bridge markers, 15 were linked to non-homologous maternal and paternal linkage groups and 525 markers did not reach the stringent threshold of LOD 15. The final numbers of markers within the bins of the skeleton map are listed in Table 2 and can be retrieved via http://potatodbase.dpw.wau.nl/UHDdata.html. Figure 2 shows the skeleton bin map.

Data quality: All mapped markers shown in the online database have been provided with a quality label. This label is based on the deviation between the observed data of this marker and the expected segregation pattern as recorded in the bin signature. Because the dimension of genetic distances due to recombination events is independent of the dimension of distance due to singletons, this deviation can be considered as a distance perpendicular to the map. Hence, listing the number of singletons per marker is useful as a quality measure representing of goodness of fit of the marker in the bin. Singletons did not occur randomly among the markers. Many markers were without singletons and the 10% of the markers with the poorest data quality account for over half the total amount of 33489 singletons in the data set (VAN OS *et al.* 2005b).

Segregation distortion: In the maternal map, segregation distortion was observed for all markers of linkage group V in the maternal map. Moderate segregation distortion (44:86) started at one telomeric end, increased to a highly distorted ratio of 26:104 (χ^2=46.8; p<0.0001) at bin 45 (SH05B045) and declined to 50:80 at the other telomeric end.

In the paternal map, linkage groups I and XII showed segregation distortion. The skewed interval on chromosome I ranged from bin RH01B001 to RH01B042, with bin RH01B021 showing the highest segregation distortion 35:95 (χ^2=27.7; p<0.0001). The first two bins of the short arm of chromosome XII did not show significant skewness (54:76), but skewness increased towards the other end. The telomeric bin RH12B051 showed the strongest segregation distortion: 21:109 (χ^2=59.6; p<0.0001).

Markers in the proximity of the highly skewed bins RH01B021 and RH12B051 showed correlated segregation patterns. This required an elevated LOD threshold to separate markers in the two linkage groups I and XII. Correlated segregation patterns between loci from different linkage groups are a violation of Mendel's law of independent assortment of allele pairs. Possibly interacting allele pairs with strong effects on pollen or embryo viability, germination, or tuber formation are located on I and XII.

Map saturation and marker distribution: Figure 2 provides a clear illustration of the length and saturation of the linkage groups. Shades of grey, rather than listing 10,000 marker names, offer an indication of over- and under saturated regions. The similarity between the maternal and paternal map is striking with respect to map length and the positions of strong clustering of markers. But also the lack of clusters at chromosome III and X is congruent between maternal and paternal maps. The largest cluster is observed on chromosome I, where the bins SH01B32 and RH01B13 contain 539 and 373 marker loci, respectively. Taking the most densely populated bin as the putative position of the centromere, the following putative centromeric bins (and number of markers in brackets): SH01B32 (539), RH01B13 (373), SH02B04 (72), RH02B01 (47), SH04B31 (212), RH04B35 (155), SH05B44 (113), RH05B46 (174), SH06B05 (52), RH06B17 (97), SH07B70 (95), RH07B68 (80), SH08B13 (43), RH08B22 (16), SH09B21 (114), RH09B31 (27), SH12B56 (199) and RH12B49 (100).

Despite the saturation of the map, gaps are observed. The largest gap is on chromosome VIII, spanning 14 recombinations in the maternal parent and 20 recombinations in the paternal parent. These gaps are probably due to recombination hot spots, but could also indicate fixation (homozygosity) of the potato genome in this region.

Distribution of recombination events and chiasmata: The distribution of marker alleles observed in the offspring genotypes allows a reconstruction of the number of recombination events in the chromatids transmitted from the parents. Analysis of the distribution of recombination events per chromatid displayed that the vast majority of the 3119 (=130*24-1) chromatids were either without recombination (44 %), or showed a single recombination event (48 %). The precise numbers of chromatids are 1379 (44 %), 1505 (48%), 228 (7.3%) and 7 (0.22%) chromatids showing 0, 1, 2 and 3 recombination events, respectively. No significant differences in recombination frequencies were observed between the female and male meiosis.

Knowing the genetic position of the 1982 recombination events captured in this mapping population, we can investigate the distribution of chiasmata. However, the distribution of chiasmata can not be directly obtained from the distribution of recombination events. Table 3 shows how the ratio of the number of recombination events depends on the number of chiasmata.

Table 3. Ratio of the number of recombinations per chromatid dependent on the number of chiasmata

Number of	Number of chiasmata						
recombinations	0	1	2	3	4	5	6
0	1	1/2	1/4	1/8	1/16	1/32	1/64
1	0	1/2	2/4	3/8	4/16	5/32	6/64
2	0	0	1/4	3/8	6/16	10/32	15/64
3	0	0	0	1/8	4/16	10/32	20/64
4	0	0	0	0	1/16	5/32	15/64
5	0	0	0	0	0	1/32	6/64
6	0	0	0	0	0	0	1/64

From this table we would expect a higher number of 0 recombinations, and not 50% single recombinants. Apparently there is a mechanism during meiosis that stimulates the formation of at least 1 chiasma.

The 228 chromatids revealing two recombination events are also intriguing, because these may reveal information on chiasma interference relative to the position of the centromere. Therefore we wish to test whether or not chiasma interference is limited to chromosome arms, and if the centromeres play a role in the process of chiasma interference. In other words, would a first chiasma more strongly inhibit the formation of a second chiasma on the same chromosome arm, and hardly interfere with the formation of a chiasma on the other chromosome arm? The 228 chromatids with two recombination events were analyzed by counting the number of recombination events per chromosome arm, taking the most densely populated bin as the putative position of the centromere. Chromosomes III, X and XI, which are without clear centromeres, were omitted from analysis, leaving 174 cases with two recombination events from the remaining chromosomes. The assumption of equal arm lengths would result in an expected ratio of 1:1 between cases with one recombination per arm and cases with both recombinations on one arm.

The 174 double recombination events were distributed over 125 chromatids with one recombination event at either side of the putative centromere, and 49 cases with two recombination events in one arm. These 49 cases however, were mainly observed in the long arms of typically acrocentric or telocentric chromosomes (90 %). The short arms of acrocentric and the metacentric chromosome V contributed only four cases of double recombination events within an arm (10 %). We therefore conclude that the over-representation of cases with one recombination per arm is more likely a reflection of the difference in arm length, rather than providing strong evidence for a maximum of one chiasma per arm. Five of the seven chromatids, which displayed three recombination events, were observed in chromatids belonging to chromosome III and X without a clear centromeric marker cluster.

Distribution of AFLP markers derived from different restriction enzyme combinations: Markers have been generated from AFLP templates based on three different enzyme combinations: *Eco*RI/*Mse*I, *Sac*I/*Mse*I and *Pst*I/*Mse*I. The genomic position of the markers is determined by the position of the six-cutter restriction site, whereas *Mse*I only ‘trims’ fragment length to a size range optimal for polyacrylamide gel electrophoresis. Hence, for each enzyme combination, the marker distribution on the genetic maps reflects the distribution of the six-cutter restriction sites. The effect of the selective nucleotides is considered negligible in view of the many primer combinations tested. The consequences of the AFLP enzyme combination on the position of the markers can be examined by using two different approaches. Firstly, we test for under-representation of methylation sensitive *Pst*I markers in the putative centromeric cluster. Secondly, we compare the average distance between marker loci as a measure for marker clustering per enzyme combination. Thirdly, we examine the effect of the number of C+G residues in the enzyme recognition site.

*Pst*I markers in particular should have a non random distribution, reflecting the methylation status of the genomic DNA. AFLP template from *Pst*I digested DNA should represent only hypo-methylated gene rich regions of the genome. The complexity of *Pst*I/*Mse*I AFLP template is approximately fourfold lower as compared to *Eco*RI/*Mse*I or *Sac*I/*Mse*I template, because equally complex AFLP fingerprints were obtained with only two selective nucleotides added to the core *Pst*I primer (+2/+3 primer combinations). In contrary, *Eco*RI and *Sac*I markers were generated with +3/+3 primer combinations. When comparing the fraction of *Pst*I markers in the putative centromeric clusters relative to the fraction of *Pst*I markers at other regions of the genome, the linkage groups III, X and XI, which lack a clear putative centromeric marker cluster, were excluded. A total 2508 markers, one third of the total number of mapped 1:1 segregating markers, was counted in 18 putative centromeric bins. These 18 bins contained only 209 (8.3 %) *Pst*I markers, whereas among all 7439 mapped 1:1 segregating markers 1446 (19.2 %) are *Pst*I markers. This observation provides clear evidence for an under-representation of *Pst*I markers in the putative centromeric marker clusters.

Finally, the effect of four C+G residues in the recognition site of *Sac*I, and two C+G residues in *Eco*RI is examined. Euchromatic regions differ from centromeric heterochromatic regions.

Plant genomes have a strong under-representation of C+G residues (33-36% in dicots; KARLIN and MRÁZEK 1997). Specifically the repetitive DNA in the centromeric heterochromatin is more A+T rich, and the gene rich euchromatic regions are less biased. This could also affect the distribution of markers of *Eco*RI versus *Sac*I markers. However, the representation of *Sac*I markers (569) in the putative centromeric marker clusters (569/2508 = 22.7 %) is not significantly different from the ratio observed for *Eco*RI markers. Therefore we conclude, *Sac*I and *Eco*RI markers cluster equally in the putative centromeric bins.

An alternative way to study marker distribution is based on the distances between neighboring markers. To compensate for the unequal number of markers per linkage group and per enzyme combination, a random subset of 1024 *Eco*RI, *Pst*I and *Sac*I markers was drawn. As a control, a fourth subset was comprised of 1024 randomly drawn bins, including empty bins. All 1000 intermediate distances within a subset were counted. The results shown in Figure 3 indicate that all markers, including *Pst*I markers are strongly clustered, as compared to the control. The level of marker clustering, however, differs among the AFLP enzyme combinations: *Pst*I markers showing the lowest amount of clustering.

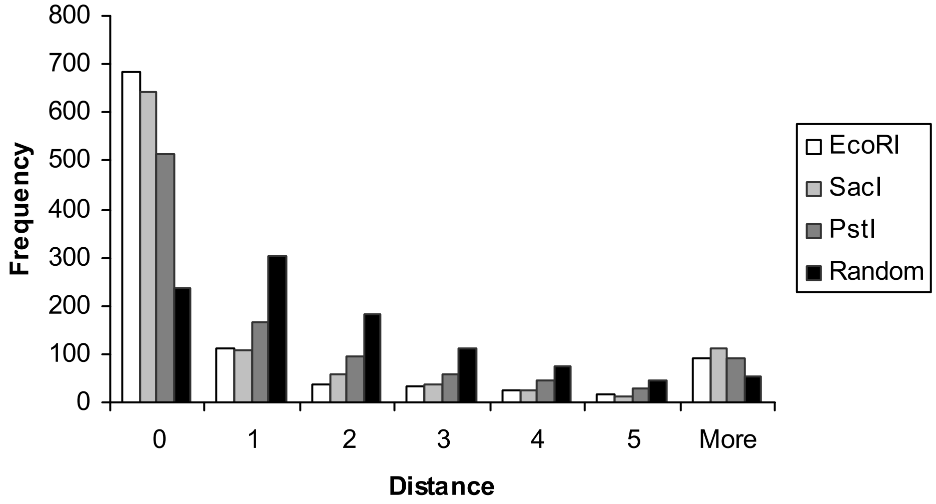


Figure 3. Frequency distribution of distances (recombination events) between neighboring marker loci to represent the marker clustering of *Eco*RI, *Sac*I and *Pst*I AFLP markers as compared to random genetic sites. Equally sized subsets of 1024 markers were randomly chosen from each different enzyme combination. Each distance between two neighboring markers from the subset was calculated, resulting in 1000 distances per enzyme combination. Within each enzyme combination the frequency of each distance was calculated. The degree of clustering is dependent on the amount of distances with value 0. The degree of clustering for random genetic positions was visualized by calculating the frequency of intermediate distances between 1024 randomly chosen bins.

The frequency of distances between neighboring markers larger than 5 recombination events is also higher than expected based on a random distribution. This can be explained by the presence of stretches of empty bins caused by recombination hot spots or local fixation (lack of heterozygosity).

In conclusion, we state that the occurrence of recombination is not random. This is explained by the occurance of local hot spots and cold spots for recombination, assuming a physically random distribution of *Eco*RI, *Sac*I and *Pst*I recognition sites in the nucleotide sequence of the potato genome.

DISCUSSION

The saturation of the potato genome with marker loci: With 381 AFLP primer combinations and a mapping population of 130 individuals, more than 10,000 markers were generated of which 93% could be accurately assigned to a genetic position. Previous maps of potato were already available (BONIERBALE *et al.* 1988, GEBHARDT *et al.* 1989, TANKSLEY *et al.* 1992, JACOBS *et al.* 1995), but varied from 100 to no more than 500 markers. Even when comparing this map with recently published high-density maps of papaya (MA *et al.* 2004: 1501 loci), cotton (RONG *et al.* 2004: 3347 loci) and sorghum (BOWERS *et al.* 2003: 2512 loci), this potato map is the densest meiotic map in any plant species yet obtained. Obviously, the level of DNA polymorphism makes a large difference between such efforts in potato or papaya in terms of data collection. However, with respect to data analysis it was noticed that the available mapping software could not cope with such large data quantities. Linkage groups with over 1000 markers cannot be handled with current software such as JoinMap, and even small amounts of errors caused severe marker ordering problems. Therefore, new approaches were devised, resulting in software (RECORD; VAN OS *et al.* 2005a) that could produce accurate marker orders in a relatively short time. The necessity to remove scoring errors was recognised and performed with SMOOTH (VAN OS *et al.* 2005b). The combination of these two programs made it possible to construct a reliable and robust framework map. The framework map consists of bins, which are positions on the genetic map with a unique segregation pattern and separated by recombination events. Thanks to the high density of the markers, it was possible to determine the position of most of the recombination events on the map. Since a direct translation from bin to centiMorgans can be made, a consecutive numbering of the bins is sufficient for indicating the positions of the genetic markers.

Exploitation of the ultra dense map: The primary goal of this research was to construct an ultra-dense genetic linkage map of potato with the purpose of saturating the genome with markers for gene cloning via BAC landing. In view of the average insert size of a BAC library and the estimated genome size of 840 Mb (BENNETT *et al.* 1997) the current number of marker loci should suffice. Proof of concept was recently obtained by the cloning of the late blight resistance gene *R3a* (HUANG *et al.* 2004; HUANG *et al.* 2005) and the construction of a BAC contig comprising the wart disease resistance gene *Sen1-4* (BRUGMANS *et al.* 2005). Both studies demonstrated that marker spacing was in accordance with the expected physical distance. Nevertheless, the genetic structure of the *R3a* and *Sen1-4* locus also showed remarkable differences. The *R3a* gene was mapped relative to two bins (1.5 cM), collectively containing 27 AFLP markers (marker-dense). A 1748 offspring high resolution map resulted in 35 sub-bins of 0.06 cM. However, the recombination events were unevenly distributed leaving ten AFLP and two CAPS markers co-segregating with resistance and stretches of sub-bins without markers. In contrast, the *Sen1-4* locus was roughly mapped relative to six bins (3.5 cM) with only nine AFLP markers (marker-poor). However, these nine AFLPs landed on overlapping BAC-clones, resulting in a single ~1Mb contig. These two examples suggest that marker-poor and empty bins indicate a favourably low Mb/cM ratio, whereas marker-rich bins indicate a high Mb/cM ratio. Therefore, under-saturated regions on the map do not necessarily present a problem for map-based cloning efforts.

Empty bins and oversaturated bins may indicate alternating recombination hot spots and cold spots on the genome. Consecutive empty bins could also indicate a local absence of marker polymorphism due to fixation of one allele. On both parental maps of chromosome VIII a long stretch of up to 19 empty bins could represent an example of either. At this moment, almost half of the bins in the framework map remain empty (44 %). Eventually the construction of a genetically anchored physical map should provide more insight in the cause of empty bins.

Marker distribution: Three different enzyme combinations have been chosen to generate markers. In a pilot study, a maternal genetic map was produced with 19 EcoRI/MseI primer combinations. In this study, it was recognised that with this single enzyme combination, a considerable portion of the genome remained unpopulated with markers. Therefore, it was decided to generate AFLP markers from DNA template prepared with three different restriction enzymes *viz. EcoR*I, *Sac*I and *Pst*I. The AT rich recognition site directing the distribution of *Eco*RI markers and the CG rich *Sac*I markers nevertheless both resulted in strong clustering, absorbing approximately one third of all the markers. For mapping purposes a more dispersed genetic distribution is preferred, but for applications such as the genetic anchoring of a physical map this is probably not a drawback. For linkage mapping of trait loci *Pst*I markers are recommended, because these are biased to non-methylated regions. There is, however, a drawback with *Pst*I markers: in almost every fingerprint, several bands were observed in the progeny that were absent in both parents. These putative methylation polymorphisms will increase the number of singletons. Furthermore, *Pst*I markers should be used with caution for BAC landing. Extra bands will appear, because of the absence of methylation in bacteria.

The highly similar distribution of *Eco*RI and *Sac*I markers demonstrates that the effect of clustering due to unequal levels of recombination outweighed the effect of differences in A/T composition in euchromatic versus heterochromatic regions. Why are these clusters so sharply confined to a single bin position? This seems to be contradicting multiple publications on AFLP maps, where clustering is obvious but extending over a wider region. In our view, the interaction between (1) the mapping algorithm and (2) the quality of the data set, explains the presence of these sharp marker clusters. First, it was demonstrated that singletons have little effect on the performance of the mapping algorithm of RECORD, but methods that use the distance between marker pairs cannot avoid inflation of map length (VAN OS *et al.* 2005a). Second, the rigorous removal of singletons will reduce the distance between closely linked markers. Usually, distances between markers are the sum of the distances caused by recombination events and the distances caused by singletons. Modest numbers of singletons (1-2 %) overshadow the effect of suppressed recombination, and will flatten the marker cluster.

Centromeric suppression of recombination is the obvious explanation for marker clustering. First, the clear congruence of the maternal and paternal homologous linkage groups excludes other adventitious heterochromatic regions as the cause of marker clustering. Second, the relative position of the clusters coincides with expectations based on cytological observations (TANKSLEY *et al.* 1992; DONG *et al.*, 2000). For example, linkage groups homologous to chromosome 2 are telocentric; the short arm being reduced to the nucleolar organiser. Chromosome VI is known for its very small short arm, and chromosome V is metacentric. In view of the sharp demarcation of the marker dense clusters, we conclude that mapping the centromeric position has an accuracy of the size of one bin: 0.8 cM. The centromeric positions of the paternal linkage groups have been confirmed using half-tetrad analysis in a 4x × 2x mapping population (MENDIBURU and PELOQUIN 1979; PARK *et al.* manuscript in preparation).

Analysis of meiotic recombination and chiasmata: Recently, HILLERS and VILLENEUVE (2003) investigated the control mechanisms of meiotic crossing over in *Caenorhabditis elegans*, which averages only one crossover per chromosome pair per meiosis. A tendency was revealed to restrict the number of crossovers, irrespective of the physical length. Pairs of fusion chromosomes composed of two or even three whole chromosomes enjoyed only a single crossover in the majority of meioses. This observation parallels the work of GERATS *et al.* (1985), who describe a relationship between the length of the deletion in the short arm of Petunia chromosome VI and the recombination frequency between markers in the long arm. The recombination frequency increased in the long arm with an increasing the length of the deletion in the short arm. Both cases in *C. elegans* and *Petunia* demonstrate that the occurrence of a pre-set amount of recombination events is highly regulated and even two recombination events are considered 'a crowd' (VAN VEEN and HAWLEY 2003). In this study, marker saturation allowed the detection of every recombination event. The fraction of chromatid arms with more than one recombination event was only 1.6%. This small fraction still represents a substantial number of 49 cases. Therefore, in our view there is no reason to assume absolute chiasma interference.

In this study singletons have not been interpreted as indicative for double recombination events. Most likely they are caused by inaccurate scoring, but some data points can also be caused by gene conversions, mutations and other biological phenomena (VAN OS *et al.* 2005b). RONG *et al.* (2004) have chosen for an alternative interpretation in a similar situation. They have concluded that negative chiasma interference could explain the unexpectedly abundant double recombinants.

Towards a sequence ready physical map: Currently, a physical map of the potato genome is being constructed from the paternal clone RH using *Eco*RI+0/*Mse*I+0 fingerprints of individual BAC clones (DE BOER *et al.* 2004). The anchoring of several thousand BAC-contigs to this genetic map will be achieved by application of AFLP on 0.4 genome equivalent pools of BACs. AFLP loci that have been mapped are easily recognised in fingerprints of 0.4 genome equivalent BAC pools. Deconvolution of the pooling design allows the identification of the BAC clones and the contig, which carries the mapped AFLP locus. A genetically anchored physical map will culminate in a sequence ready minimal tiling path of BAC contigs of specific chromosomal regions. Within the International Solanaceae Genome Project (SOL) for comparative genome studies (http://sgn.cornell.edu/solanaceae-project/), as well as within the potato genome sequencing consortium, this ultra dense linkage map and the anticipated genetically anchored physical map will have a valuable role.

ACKNOWLEDGEMENTS

This work was carried out under the European Union FAIR (Agriculture and Fisheries) program grant FAIR5-PL97-3565.

LITERATURE CITED

BAKKER, E., U. ACHENBACH, J. BAKKER, J. VAN VLIET, J. PELEMAN, B. SEGERS, S. VAN DER HEIJDEN, P. VAN DER LINDE, R. GRAVELAND, R. HUTTEN, H. J. VAN ECK, E. COPPOOLSE, E. VAN DER VOSSEN, J. BAKKER and A. GOVERSE, 2004 A high-resolution map of the *H1* locus harbouring resistance to the potato cyst nematode *Globodera rostochiensis*. Theor. Appl. Genet. **109**: 146-152.

BASTIAANSSEN, H. J. M., 1997 Marker assisted elucidation of the origin of 2n-gametes in diploid potato. PhD thesis Wageningen University, 150 pp. ISBN 90-5485-759-5.

BENNET, M. D., A. V. COX and I. J. LEITCH, 1997 http://www.rbgkew.org.uk/cval/database1.html.

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.

BOWERS, J. E., C. ABBEY, S. ANDERSON, C. CHANG, X. DRAYE, A. H. HOPPE, R. JESSUP, C. LEMKE, J. LENNINGTON, Z. LI, Y. R. LIN, S. C. LIU, L. LUO, B. S. MARLER, R. MING, S. E. MITCHELL, D. QIANG, K. REISCHMANN, S. R. SCHULZE, D. N. SKINNER, Y. W. WANG, S. KRESOVICH, K. F. SCHERTZ and A. H. PATERSON, 2003 A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. Genetics **165**: 367-386.

BOYKO, E., R. KALENDAR, V. KORZUN, J. FELLERS, A. KOROL, A. H. SCHULMAN and B. S. GILL, 2002 A high-density cytogenetic map of the *Aegilops tauschii* genome incorporating retrotransposons and defense-related genes: insights into cereal chromosome structure and function. Plant Mol. Biol. **48**: 767-790.

BRUGMANS, B., R. G. B. HUTTEN, N. ROOKMAKER, R. G. F. VISSER and H. J. VAN ECK, 2005 Exploitation of a marker dense linkage map of potato for positional cloning of a wart disease resistance gene. Theor Appl Genet, in press.

BUNTJER J. B., H. VAN OS and H. J. VAN ECK, 2000a ComBin: Software for ultra-dense mapping Plant and Animal Genome Conference VIII, San Diego. http://www.intl-pag.org/pag/8/abstracts/pag8038.html

BUNTJER J. B., H. VAN OS and H. J. VAN ECK, 2000b Construction of ultra-dense maps using novel software Plant and Animal Genome Conference VIII, San Diego. http://www.intl-pag.org/pag/8/abstracts/pag8039.html

DE BOER, J. M., T. J. A. BORM, B. BRUGMANS, E. R. BAKKER, J. BAKKER, R. G. F. VISSER and H. J. VAN ECK, 2004 Construction of a genetically anchored physical map of the potato genome. Plant & Animal Genomes XII Conference, San Diego. http://www.intl-pag.org/pag/12/abstracts/W54_PAG12_245.html

DIB, C., S. FAURE, C. FIZAMES, D. SAMSON, N. DROUOT, A. VIGNAL, P. MILLASSEAU, S. MARC, J. HAZAN, E. SEBOUN, M. LATHROP, G. GYAPAY, J. MORISSETTE and J. WEISSENBACH, 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, Z. HUSAIN, R. DREDGE, M. J. DALY, K. A. INGALLS, T. J. O'CONNOR, C. A. EVANS, M. M. DEANGELIS, D. M. LEVINSON, L. KRUGLYAK, N. GOODMAN, N. G. COPELAND, N. A. JENKINS, T. L. HAWKINS, L. STEIN, D. C. PAGE and E. S. LANDER, 1996 A comprehensive genetic map of the mouse genome. Nature **380**: 149-152.

DONG, F., J. SONG, S. K. NAESS, J. P. HELGESON, C. GEBHARDT and J. JIANG, 2000 Development and applications of a set of chromosome-specific cytogenetic DNA markers in potato. Theor Appl Genet **101**: 1001-1007.

GEBHARDT, C., E. RITTER, T. DEBENER, U. SCHACHTSCHABEL, B. WALKEMEIJER and F. SALAMINI, 1989 RFLP analysis and linkage mapping in Solanum toberosum. Theor. Appl. Genet. **78**: 65-75

GERATS, A. G. M., P. DE VLAMING and D. MAIZONNIER, 1985 Recombination behavior and gene transfer in *Petunia hybrida* after pollen irradiation. Mol Gen Genet **198**: 57-61.

GRATTAPAGLIA, D., and R. SEDEROFF, 1994 Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers, Genetics **137**: 1121-1137.

HAANSTRA, J. P. W., C. WYE, H. VERBAKEL, F. MEIJER-DEKENS, P. VAN DEN BERG, P. ODINOT, A. W. VAN HEUSDEN, S. TANKSLEY, P. LINDHOUT and J. PELEMAN, 1999 An integrated high-density RFLP-AFLP map of tomato based on two *Lycopersicon esculentum x L. pennellii* F_2 populations. Theor. Appl. Genet. **99**: 254-271.

HARUSHIMA, Y., M. YANO, A. SHOMURA, M. SATO, T. SHIMANO, Y. KUBOKI, T. YAMAMOTO, S. YANG LIN, B. A. ANTONIO, A. PARCO, H. KAJIYA, N. HUANG, K. YAMAMOTO, Y. NAGAMURA, N. KURATA, G. S. KHUSH and T. SASAKI, 1998 A high-density rice genetic linkage map with 2275 markers using a single F_2 population. Genetics **148**: 479-494.

HILLERS, K. J. and A. M VILLENEUVE, 2003 Chromosome-wide control of meiotic crossing over in *C. elegans*. Curr Biol **13**: 1641-1647.

HUANG, S., V. G. A. A. VLEESHOUWERS, J. S. WERIJ, R. C. B. HUTTEN, H. J. VAN ECK, R. G. F. VISSER and E. JACOBSEN, 2004 The *R3* resistance to *Phytophthora infestans* in potato is conferred by two closely linked *R* genes with distinct specificities. Mol. Plant. Microbe. Interact. **17**: 428-435.

HUANG, S., E. A. G. VAN DER VOSSEN, H. KUANG, V. G. A. A. VLEESHOUWERS, N. ZHANG, T. J. A. BORM, H. J. VAN ECK, B. BAKER, E. JACOBSEN and R. G. F. VISSER, 2005 Comparative genomics enabled the isolation of the *R3a* late blight resistance gene in potato. The Plant J. in press.

ISIDORE, E., H. VAN OS, S. ANDRZEJEWSKI, J. BAKKER, I. BARRENA, G. J. BRYAN, B. CAROMEL, H. J. VAN ECK, B. GHAREEB, W. DE JONG, P. VAN KOERT, V. LEFEBVRE, D. MILBOURNE, E. RITTER, J. ROUPPE VAN DER VOORT, F. ROUSSELLE-BOURGEOIS, J. VAN VLIET and R. WAUGH, 2003 Toward a marker-dense meiotic map of the potato genome: Lessons from linkage group I. Genetics **165**: 2107-2116.

JACOBS, J. M. E., H. J. VAN ECK, P. ARENS, B. VERKERK-BAKKER, B. TE LINTEL HEKKERT, H. J. M. BASTIAANSSEN, A. EL-KHARBOTLY, A. PEREIRA, E. JACOBSEN and W. J. STIEKEMA, 1995 A genetic map of potato (Solanum tuberosum) integrating molecular markers, including transposons, and classical markers. Theoretical and Applied Genetics **91**: 289-300.

KARLIN, S. and J. MRÁZEK, 1997 Compositional differences within and between eukaryotic genomes. Proc Natl Acad Sci USA **94**: 10227-10232.

KING, J., L. A. ROBERTS, M. J. KEARSEY, H. M. THOMAS, R. N. JONES, L. HUANG, I. P. ARMSTEAD, W. G. MORGAN and I. P. KING, 2002 A demonstration of a 1:1 correspondence between chiasma frequency and recombination using a *Lolium perenne/Festuce pratensis* substitution. Genetics **161**: 307-314.

MA, H., P. H. MOORE, Z. LIU, M. S. KIM, Q. YU, M. M. FITCH, T. SEKOKIA, A. H. PATERSON and R. MING, 2004 High-density linkage mapping revealed suppression of recombination at the sex determination locus in papaya. Genetics **166**: 419-436.

MENDIBURU, A. O. and S. J. PELOQUIN, 1979 Gene-centromere mapping by 4x-2x matings in potatoes. Theor Appl Genet **54**: 177-180.

MICHELMORE, R. W., I. PARAN and R. V. KESSELI, 1991 Identification of markers linked to disease-resistance genes by Bulked Segregant Analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc. Natl. Acad. Sci. USA. **88**: 9828-9832.

MURRAY, J. C., K. H. BUETOW, J. L. WEBER, S. LUDWIGSEN, T. SCHERPBIER-HEDDEMA, F. MANION, J. QUILLEN, V. C. SHEFFIELD, S. SUNDEN, G. M. DUYK, J. WEISSENBACH, G. GYAPAY, C. DIB, J. MORISSETTE, G. M. LATHROP, A. VIGNAL, R. WHITE, N. MATSUNAMI, S. GERKEN, R. MELIS, H. ALBERTSEN, R. PLAETKE, S. ODELBERG, D. WARD, J. DAUSSET, D. COHEN and H. CANN, 1994 A comprehensive human linkage map with centimorgan density. Science **265**: 2049-2054.

QI, X. and P. LINDHOUT, 1997 Development of AFLP markers in barley. Mol. Gen. Genet. **254**: 330-336.

PARAN, I., J. ROUPPE VAN DER VOORT, V. LEFEBVRE, M. JAHN, L. LANDRY, M. VAN SCHRIEK, B. TANYOLAC, C. CARANTA, A. BEN CHAIM, K. LIVINGSTONE, A. PALLOIX and J. PELEMAN, 2004 An integrated genetic linkage map of pepper (*Capsicum spp.*) Molecular Breeding **13**: 251-261.

RONG, J., C. ABBEY, J. E. BOWERS, C. L. BRUBAKER, C. CHANG, P. W. CHEE, T. A. DELMONTE, X. DING, J. J. GARZA, B. S. MARLER, C. H. PARK, G. J. PIERCE, K. M. RAINEY, V. K. RASTOGI, S. R. SCHULZE, N. L. TROLINDER, J. F. WENDEL, T. A. WILKINS, T. D. WILLIAMS-COPLIN, R. A. WING, R. J. WRIGHT, X. ZHAO, L. ZHU and A. H. PATERSON, 2004 A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (*Gossypium*). Genetics **166**: 389-417.

ROUPPE VAN DER VOORT, J. N. A. M., H. J. VAN ECK, J. DRAAISTRA, P. M. VAN ZANDVOORT, E. JACOBSEN and J. BAKKER, 1997 An online catalogue of AFLP markers covering the potato genome. Mol. Breed. **4**: 73-77.

STAM, P., and J. VAN OOIJEN, 1995 JoinMap version 2.0: Software for the calculation of genetic linkage maps. CPRO-DLO, Wageningen.

TANKSLEY, S., M. GANAL, J. PRINCE, M. DE VICENTE, M. BONIERBALE, P. BROWN, T. FULTON, J. GIOVANNONI, S. GRANDILLO, G. MARTIN, R. MESSEGEUR, J. MILLER, L. MILLER, A. PATERSON, O. PINEDA, M. RÖEDER, R. WING, W. WU and N. YOUNG, 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 OOIJEN, J. W. and R. E. VOORRIPS, 2001 JoinMap® Version 3.0, Software for the calculation of genetic linkage maps. Plant Research International, Wageningen, The Netherlands.

VAN OS, H., P. STAM, R. G. F. VISSER and H. J. VAN ECK, 2005a RECORD: a novel method for ordering loci on a genetic linkage map. Theor. Appl. Genet. In press.

VAN OS, H., R. G. F. VISSER and H. J. VAN ECK, 2005b SMOOTH: a statistical method for successful removal of genotyping errors from high-density genetic linkage data. Theor. Appl. Genet. In press.

VAN VEEN, J. E. and R. S. HAWLEY, 2003 Meiosis: when even two is a crowd. Curr Biol. **13**: R831-833.

VOS, P., R. HOGERS, M. BLEEKER, M. REIJANS, T. VAN DE LEE, M. HORNES, A. FRIJTERS, J. POT, J. PELEMAN, M. KUIPER and M. ZABEAU, 1995 AFLP: a new technique for DNA fingerprinting. Nucl. Acid Res. **23**: 4407-4414.

VUYLSTEKE, M., R. MANK, R. ANTONISE, E. BASTIAANS, M. L. SENIOR, C. W. STUBER, A. E. MELCHINGER, T. LUBBERSTEDT, X. C. XIA, P. STAM, M. ZABEAU and M. KUIPER, 1999 Two high-density AFLP linkage maps of *Zea mays* L.: analysis of distribution of AFLP markers. Theor. Appl. Genet. **99**: 921-935.

WAUGH, R., N. BONAR, E. BAIRD, B. THOMAS, A. GRANER, P. HAYES and W. POWELL, 1997 Homology of AFLP products in three mapping populations of barley. Mol. Gen. Genet. **255**: 311-321.

Chapter 6:
General Discussion

An ultra dense genetic linkage map of potato based on AFLP markers is a useful tool for a wide range of breeding applications as well as for basic science. The aim of this research project was to develop such an ultra dense map for potato with a main purpose for BAC landing and research in structural genomics. As the genome of potato is not yet sequenced, the ultra dense map serves as a compromise in sequence information by AFLP, which can be exploited in subsequent genetical studies.

This thesis describes the development and testing of methods for data analysis in ultra dense genetic mapping projects. Problems and pitfalls concerning data analysis are discussed and practical application of the methods is demonstrated.

LINKAGE MAPPING IN POTATO

The first genetic linkage map of potato (BONIERBALE *et al.* 1988) made use of RFLP markers, which originated from the molecular genetic map of tomato. It was found that there was a high colinearity between the potato and tomato genome. This colinearity and the absence of an adequate numbering of the potato chromosomes, led to the numbering of the potato linkage groups in accordance with the numbering for the tomato chromosomes. This numbering, based on homoeology with tomato, is now commonly used in potato genetics (e.g. TANKSLEY *et al.* 1992).

Whereas the first genetic map of potato was based on one parent only, the second genetic map of potato (GEBHARDT *et al.* 1989) was constructed on the basis of three linkage groups per chromosome. The first linkage group consisted of alleles from the female parent, segregating in a 1:1 ratio. The second linkage group comprised 1:1 segregating alleles from the male parent. Since the mapping population was a backcross, both parents had at least one allele in common. When this common allele is polymorphic, the third linkage group can be based on both parents and includes alleles segregating 3:1. These common alleles are also called allelic bridges (RITTER *et al.* 1990). Finally, these three linkage groups have been merged according to the positions of these allelic bridges. This approach has led to a common mapping procedure in outbreeding species, called: the two-way pseudo-testcross (GRATTAPAGLIA and SEDEROFF 1994). This procedure is based on the fact that the segregating F1 generation of two heterozygous individuals can be seen as the superposition of two BC1 generations. A major difference with the BC1 generation is the phase difference. In outbreeding populations, markers can segregate in coupling or in repulsion phase. Phase determination is essential for an accurate marker placement. Other genetic linkage maps in potato were constructed by BONIERBALE *et al.* 1988; GEBHARDT *et al.* 1989, 2001; JACOBS *et al.* 1995; VAN ECK *et al.* 1995; MILBOURNE *et al.* 1998.

Cultivated potato in the western world is an autopolyploid species with 4 chromosome sets that are all capable of pairing and recombining with each other. If a genetic linkage map was constructed from crosses between tetraploids, 6 allelic combinations are possible per locus in the gametes. Therefore, it is obvious that the determination of crossing-over frequency in such crosses is a complicated task. To circumvent this, the ultra dense map has been constructed from a cross between two dihaploid clones. Computer software is available that allows mapping in tetraploid species if no dihaploid clones are available (HACKETT *et al.* 2003).

The optimal cross for map construction should be a balance between maximum polymorphism, and a high level of homology for chromosome pairing and recombination to take place. In general, a high polymorphism can be obtained by using distantly related parents for BC or F_2 populations. However, the genetic distance between the parents used, can affect the recombination frequencies in the cross. The more closely related the parents in the cross, both for parents within the same species as well as for parents from different species, the greater the homology. This results in increased chromosome pairing, and higher recombination frequencies based on crossing-over. Therefore, longer genetic maps are expected. However, a high level of homology between the parents is countered by a decrease in polymorphism, and the recombination events might go unnoticed. An advantage of using a non-inbred species for map construction is the high level of polymorphism, even within species. This often allows the use of an intraspecific F1 population for map construction, thereby avoiding the chance of reduced recombinations, due to a lack of homology, encountered in interspecific crosses. A clear disadvantage of working with a cross-fertilizing, highly heterozygous crop like potato is the general phenomenon of inbreeding depression. Consequently, many uncharacterized (sub)lethal loci are expected to be present in the potato genome. These become visible as distorted segregation of alleles in one of the parents, due to gametic selection, or as (partial) absence of genotypic classes in the progeny due to zygotic selection. As a consequence, clusters of markers with distorted segregation are expected to arise. They may vary for each cross and will mainly depend on the parents used. Distorted segregation was found on chromosome V in the female parent and chromosomes I and XII in the male parent in the ultra dense mapping population.

DEVELOPMENT OF GENERIC TOOLS FOR ULTRA DENSE MAPPING

The construction of ultra dense maps requires adequate tools and techniques. One requirement is a molecular marker technique that can produce many reliable markers in short time with not too much labor. AFLP (VOS *et al.* 1995) is such a molecular marker technique and makes the production of 10,000 unique molecular markers feasible.

The first problem to encounter when analyzing ultra dense mapping data sets is the quantity of the data in terms of the calculation time. A data set of 10,000 markers cannot be analyzed with the average available mapping software. Therefore, the development of new programs was indispensable. New programs need to be extensively tested first with simulated data, to make sure that the software produces the desired results.

A linkage-mapping program consists of two elements, the target function or minimization criterion and the (heuristic) search algorithm. To increase the calculation speed, one has two options. The first option is to develop a minimization criterion that is less calculation intensive. JoinMap minimizes the least squared differences between the observed recombination frequencies and the calculated recombination frequencies in the map. This requires the calculation of all pairwise recombination frequencies between all markers. As the map grows, the calculation of the minimization criterion becomes more complex and thus slower. The evaluation of markers on different map positions, which is a prerequisite for accurate mapping, becomes therefore too time consuming. The program RECORD (VAN OS *et al.* 2005a) uses the total amount of recombination events in the map. As compared to the least squared differences, this evaluation criterion is much less time consuming. In fact, it is a good compromise between statistical rigor and common sense. The second option for decreasing the calculation time is choosing a smart search algorithm. The most accurate search is an exhaustive search in the complete solution space. For a map of 100 markers, this requires an evaluation of $\frac{100!}{2} = 4.7 \times 10^{157}$ possible marker orders. When the evaluation of one order takes one second, an exhaustive search would take 1.5 x 10^{150} years. This illustrates the need of heuristic search algorithms. RECORD uses a search algorithm that adds one random marker from the data set to the existing marker order. This marker is placed on its most likely position, i.e. the position on which it has the lowest increase on the target function. As this approach is input order dependent, the final result may lead to a local minimum in the solution space. Therefore the whole procedure is repeated for ten times and the best possible order is chosen from these ten solutions.

The second problem that arises when analyzing ultra dense mapping data sets is the amount of singletons. A singleton is a data point that has recombined with both its directly neighboring data points. Singletons do not have a large impact on the order of markers when the marker density is low. However, in ultra dense mapping data sets, even relatively small amounts of singletons (1%-5%) cause substantial ordering ambiguities and map inflation (LINCOLN and LANDER 1992). Therefore we state that, whether or not singletons are due to biological phenomena, they have to be removed. SMOOTH (VAN OS *et al.* 2005b) is a statistical singleton removal program. It calculates the probability of each data point for being a singleton by using 15 neighboring data points on either side within a given marker order. The closest markers in the order are given the highest weight in calculating the expected score. The difference between the expected score and the observed score is a measure for the probability of the data point for being a singleton. This probability is more accurate, when the density of markers in the map is higher. SMOOTH has been applied to the data set in an iterative process with RECORD. Each time RECORD has calculated the marker order, SMOOTH removes all singletons higher than the threshold for the probability. After every step, the threshold for removal is slightly released. This process is continued for 15 cycles down to a probability threshold of $p = 0.7$. At this stage almost all singletons have been removed from the data. Although it is advisable to use SMOOTH only in high-density data sets, it might also be applicable to locally dense marker clusters in maps of otherwise low or normal density.

During an early stage of the project, the program ComBin (BUNTJER *et al.* 2000) has been developed. The ComBin mapping procedure consists of three stages: the binning step, in which cosegregating markers are gathered in bins; the threading step, in which bins are ordered and connected with each other like threading beads on a string; and the grouping step, in which strings that belong to the same linkage group are connected. The binning step is used to remove redundant data from the data set. In fact, the bin signature should be an exact duplicate of the marker segregation pattern. However, two otherwise identical markers from which one contains a missing data point for a single offspring individual also lack any observable recombination, and will thus be classified in a single bin. The bin signature can be derived from the ’best’ marker, *i.e.* the marker with the least amount of missing values or from the complementation of the data from all the markers in the bin. Whereas data complementation may lead to a maximum of information, non-existent segregation patterns can be composed from markers with a lot of missing data.

The bins are connected with each other in the threading step. Although the conventional use of recombination distances falsely suggest a continuous scale, a linkage map is basically built upon discrete events. The true discrete nature of a linkage map becomes more apparent in saturated data sets. The threading process comes down to an iterated scanning of direct neighbors of a bin. Although a bin theoretically has two neighbors, the data may provide multiple neighbors for a bin. These neighboring bins are connected with the first bin up to programmable limitations. The scanning of neighbors is continued in every bin that is connected to the string until no further neighbors can be found at a distance of one recombination event. From that point the scanning is continued up to a given amount of recombination events between the bins, by postulating empty bins between bins that differ for more than one recombination event. The theoretical empty bin has a signature that is derived from the two existing bins, except for the allele data of the individuals that have recombined. These will be replaced by missing data. Finally, the strings can be connected by the grouping step at a distance larger than allowed during the threading process. The grouping method will determine the distance between strings, which is defined as the lowest number of recombinations between a pair of bins from both strings.

While analyzing the ultra dense data set with ComBin, it became clear that the amount of singletons was too large to cope with. In fact, the strings that were produced more resembled 'Christmas trees' rather than linear chromosomes. Although the bin concept behind ultra dense mapping was a reasonable approach, from that moment we were forced to adopt a different strategy by the present data quality. It was after the development of SMOOTH that we decided to use the cleaned data set as input for ComBin. The results that were produced by ComBin with the cleaned data set were completely different from what we had seen before. The number of unique bins was much lower, because the removal of singletons caused many markers to become redundant. Even the strings that were produced were almost linear. Some side branches still remained, due to the presence of some systematic errors like unintentional mix up of individuals before isolating a fresh batch of DNA. With ComBin, these remaining errors could be detected and successfully removed. At this point, the framework map was available consisting of unique and corrected bins. We derived the following definition for a bin. A bin is a unique and most accurate representation of a marker at a certain genetic position. A bin contains at least 1 marker and cannot be divided within the given population. Bins are numbered consecutively, based on the recombination events. As a consequence, the bin numbers can be directly translated into map units. Bins are not a statistical average of recombination frequencies as is the case with the positions of markers in conventional genetic maps based on centiMorgans. Based on this definition, it is allowed to update the bin signatures by filling in all missing values that are not flanked by recombination events.

In the final stage of the ultra dense mapping process, the bin signatures were verified by fitting the original marker data into the bins. This verification of the framework map is done by maximum likelihood comparison of the original markers with the framework bins. Every marker is placed in its most likely bin and the number of singletons is a measure for the accuracy of each marker. The average amount of singletons in the markers segregating in one parent only was 4.1%.

The dominant allelic bridges in the data, derived from the AFLP markers segregating in both parents, can be mapped with a much lower accuracy than the markers segregating in one parent only (see also: MALIEPAARD *et al.* 1997). The allelic bridges provide a rough estimate for the linkage between the two parental maps. Thanks to the large number of available allelic bridges, we were able to observe a reasonable alignment of the two separate maps.

GENETICAL STUDIES WITH THE ULTRA DENSE MAP

Besides AFLP markers, other types of markers were applied on the ultra dense mapping population, like RFLPs, SSRs, CAPs, SCARs, etc. These markers, of which the chromosomal location was known, provided the chromosome numbers to the linkage groups in the ultra dense linkage map.

Significant segregation distortion was observed on chromosome V in the maternal map and chromosome I and XII in the paternal map. Segregation distortion is due to selection on the gametes and/or progeny between the moment of crossing the parents and the moment of harvesting leaf material from the progeny for DNA extraction. Selection might have occurred on traits like pollen viability, germination, stress tolerance, earliness and disease resistance, or (sub)lethal loci. Also the unassigned linkage group RHU in the paternal map containing 13 markers, showed severe segregation distortion. The true identity of this linkage group might be revealed by the physical map.

An inspection of the amount of recombinations per chromatid showed that the occurrence of 0 chiasmata is underrepresented. A number of 49 chromatid arms in the population contained a double recombination. Obviously the amount of chiasmata in one chromatid arm is not limited to only one. Previous studies had not provided clear evidence for double recombinations in potato, but after the accurate removal of artifacts in the map with SMOOTH, the occurrence of more than one chiasma per chromosome arm is evident.

The distribution of recombinations and AFLP markers is not similar in potato as demonstrated by the presence of both gaps and clusters in the map. The biggest clusters of markers are observed around the centromeres, where recombination is suppressed. The few gaps present in the map do not show a regular pattern as the clusters. These gaps are probably due to recombination hot spots, but could also indicate fixation of the potato genome in this region.

Among the AFLP markers, differences in distribution can be found between restriction enzyme combinations. The main observation is the lower amount of Pst markers in the centromeric clusters due to methylation, which gives Pst primer combinations an advantage in obtaining a more even distribution of markers. However, due to methylation polymorphisms, the amount of singletons in Pst primer combinations is higher.

After fitting the entire original scoring data into the bins, an accurate estimate of the amount of singletons can be made. Considering the 7,600 markers segregating in one parent only, the amount of missing values is 12.8% and the amount of singletons is 4.1%. This percentage is compared with the amount of data points that are different between the four duplicate individuals, which equals 3.9%. However, if the singletons have occurred at random, 2.0% singletons should have caused 3.9% different data points between duplicates. Clearly this percentage is too low in comparison with the amount of singletons found in the map. Obviously, the singletons do not occur at random, but more than halve of them are present in both duplicate clones. Thus, singletons are not only caused by scoring errors, but also by laboratory errors and/or biological phenomena.

APPLICABILITY OF ULTRA DENSE MAPS

The map has been successfully used as a resource for markers linked to resistance genes. The AFLP markers are transferable to other populations (ROUPPE VAN DER VOORT *et al.* 1997) and recombinants can be produced to fine map the specified region (BAKKER *et al.* 2004). With these markers, the genes can be cloned, but the applicability of the ultra-dense map ranges further than specific target regions. It is an important database for the genome wide anchoring of BACs into a physical map of potato. Especially in regions on the genome rich in recombination events, molecular markers provide the blueprint for contig construction. However, ordering BACs within the centromeric clusters is still difficult.

To enable research on potato, a detailed database is being constructed that can be accessed through Internet. It will be completed with a reference catalogue of AFLP annotated gels, to transfer the AFLP markers from the ultra-dense map to other populations.

The methods described in this thesis have been applied in a progeny from two highly heterozygous clones of potato. But in principal, without any modification, these methods work in backcross populations, as well as in haploid or doubled haploid populations. Dense and accurate mapping is virtually impossible in recombinant inbred line populations due to the high number of double recombination events, but with some modifications in SMOOTH, F_2 populations can be analyzed as well. It may be advisable in F_2 populations to split the two parental maps, especially in the case of dominant scoring of the markers.

Finally, this robust, easy and fast approach is a model for other ultra-dense maps. Being the densest map ever obtained through meiotic recombination in any species, it will prove the utility in basic genetical studies as well as applied potato-breeding research.

LITERATURE CITED

BAKKER, E., U. ACHENBACH, J. BAKKER, J. VAN VLIET, J. PELEMAN, B. SEGERS, S. VAN DER HEIJDEN, P. VAN DER LINDE, R. GRAVELAND, R. HUTTEN, H. J. VAN ECK, E. COPPOOLSE, E. VAN DER VOSSEN, J. BAKKER and A. GOVERSE, 2004 A high-resolution map of the *H1* locus harbouring resistance to the potato cyst nematode *Globodera rostochiensis*. Theor. Appl. Genet. **109**: 146-152.

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.B., H. VAN OS and H. J. VAN ECK, 2000 ComBin: Software for ultra-dense mapping. Plant and Animal Genome Conference VIII, San Diego. http://www.intl-pag.org/pag/8/abstracts/pag8038.html

GEBHARDT, C., E. RITTER, T. DEBENER, U. SCHACHTSCHABEL, B. WALKEMEIJER and F. SALAMINI, 1989 RFLP analysis and linkage mapping in Solanum toberosum. Theor. Appl. Genet. **78**: 65-75

GEBHARDT, C., E. RITTER and F. SALAMINI, 2001 RFLP map of the potato, pp. 319–336 in DNA-Based Markers in Plants, Advances in Cellular and Molecular Biology of Plants, Vol. 6, Ed. 2, edited by R. L. PHILLIPS and I. K. VASIL. Kluwer Academic Publishers, Dordrecht, The Netherlands.

GRATTAPAGLIA, D., and R. SEDEROFF, 1994 Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers, Genetics **137**: 1121-1137.

HACKETT, C. A., B. PANDE and G. J. BRYAN, 2003 Constructing linkage maps in autotetraploid species using simulated annealing. Theor. Appl. Genet. **106**: 1107-115.

JACOBS, J. M. E., H. J. VAN ECK, P. ARENS, B. VERKERK-BAKKER, B. TE LINTEL HEKKERT, H. J. M. BASTIAANSSEN, A. EL-KHARBOTLY, A. PEREIRA, E. JACOBSEN and W. J. STIEKEMA, 1995 A genetic map of potato (Solanum tuberosum) integrating molecular markers, including transposons, and classical markers. Theoretical and Applied Genetics **91**: 289-300.

LINCOLN, S. E. and E. S. LANDER, 1992 Systematic detection of errors in genetic linkage data. Genomics **14**: 604-610.

MALIEPAARD, C., J. JANSEN, and J. W. VAN OOIJEN, 1997 Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. Genet. Res. **70**: 237-250.

MILBOURNE, D., R. MEYER, A. COLLINS, L. RAMSAY, C. GEBHARDT and R. WAUGH, 1998. Isolation, characterisation and mapping of simple sequence repeat loci in potato. Mol. Gen. Genet. **259**: 233-245.

RITTER, E., C. GEBHARDT, and F. SALAMINI, 1990 Estimation of recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents. Genetics **125**: 645-654.

ROUPPE VAN DER VOORT, J. N. A. M., H. J. VAN ECK, J. DRAAISTRA, P. M. VAN ZANDVOORT, E. JACOBSEN and J. BAKKER, 1997 An online catalogue of AFLP markers covering the potato genome. Mol. Breed. **4**: 73-77.

TANKSLEY, S., M. GANAL, J. PRINCE, M. DE VICENTE, M. BONIERBALE, P. BROWN, T. FULTON, J. GIOVANNONI, S. GRANDILLO, G. MARTIN, R. MESSEGEUR, J. MILLER, L. MILLER, A. PATERSON, O. PINEDA, M. RÖEDER, R. WING, W. WU and N. YOUNG, 1992 High density molecular linkage maps of the tomato and potato genomes. Genetics **132**: 1141-1160.

VAN ECK, H., J. ROUPPE VAN DER VOORT, J. DRAAISTRA, P. VAN ZANDVOORT, E. VAN ENCKEVORT, B. SEGERS, J. PELEMAN, E. JACOBSEN, J. HELDER and J. BAKKER, 1995 The inheritance and chromosomal localization of AFLP markers in a non-inbred potato offspring. Mol. Breed. **1**: 397-410.

VAN OS, H., P. STAM, R. G. F. VISSER and H. J. VAN ECK, 2005a RECORD: a novel method for ordering loci on a genetic linkage map. Theor. Appl. Genet. In press.

VAN OS, H., R. G. F. VISSER and H. J. VAN ECK, 2005b SMOOTH: a statistical method for successful removal of genotyping errors from high-density genetic linkage data. Theor. Appl. Genet. In press.

VOS, P., R. HOGERS, M. BLEEKER, M. REIJANS, T. VAN DE LEE, M. HORNES, A. FRIJTERS, J. POT, J. PELEMAN, M. KUIPER and M. ZABEAU, 1995 AFLP: a new technique for DNA fingerprinting. Nucl. Acid Res. **23**: 4407-4414.

Summary

The research in this thesis deals with the construction and specifics connected to the construction of an ultra-dense genetic linkage map of potato. An F1 population of 130 individuals from a cross between two heterozygous diploid potato clones was analyzed with 381 AFLP primer combinations which yielded nearly 10,000 markers. During an early stage of data analysis, it was noticed that the available mapping software could not cope with these data quantities. Linkage groups with over 1000 markers cannot be handled with current software such as JoinMap, and even small amounts of errors caused severe marker ordering problems. Therefore, new approaches were developed, resulting in the program RECORD (Chapter 3) that could produce accurate marker orders in a relatively short time. The necessity to remove scoring errors was recognised and performed with the computer program SMOOTH (Chapter 4). The combination of these two programs made it possible to construct a reliable and robust framework map. The framework map consists of bins, which are positions on the genetic map harbouring different amounts of AFLP markers with a unique segregation pattern and separated by recombination events. Thanks to the high level of saturation offered by 10,000 markers, it was possible to determine most of the recombination events in the population. Finally, all marker data were fit into the bins of the skeleton bin map by maximum likelihood.

Chapter 2 describes the pilot study of the new mapping procedure with linkage group I, the largest linkage group in terms of the number of markers. LG I consists of 95 maternal bins and 101 paternal bins. The 1260 AFLP markers are not evenly distributed along the genetic map. Clustering was observed in one bin in each of the parental maps and despite the marker saturation, gaps of up to seven bins were found. Markers derived from EcoRI/MseI, SacI/MseI and PstI/MseI enzyme combinations showed different genetic clustering. Approximately three-quarters of the markers placed into a bin were considered to fit well, based on an estimated residual ‘error-rate’ of 0-3%. However, twice as many PstI-based markers fit badly, suggesting that parental PstI-site methylation patterns had changed in the population.

The new software RECORD (REcombination Counting and ORDering) is presented in Chapter 3 and can be used for the ordering of loci on linkage groups. The cost function of this method is based on the minimization of the total number of recombination events per linkage group. The search algorithm is a heuristic procedure, combining elements of branch-and-bound with local reshuffling. Since the criterion proposed does not require intensive calculations, the algorithm rapidly produces an optimal ordering. A simulation study was performed to compare the performance of RECORD and JoinMap. RECORD is much faster and less sensitive to missing observations and scoring errors, since the optimization criterion is less sensitive to the effect of the scoring errors. In particular, RECORD performs better in regions of the map with high marker density.

The statistical method that was developed to remove genotyping errors from genetic linkage data during the mapping process is described in Chapter 4. The program SMOOTH calculates the difference between the observed and predicted values of data points based on data points of neighboring loci in a given marker order. Highly improbable data points are removed by the program in an iterative process with a mapping algorithm that recalculates the map after cleaning. SMOOTH has been tested with simulated data and experimental mapping data from the ultra-dense map of potato. Simulations demonstrate that this method is able to detect a high amount of scoring errors and enables mapping software to successfully construct a very accurate high-density map.

Finally, the complete ultra-dense map of potato is presented in Chapter 5. A skeleton bin map was derived, spanning 977 and 1005 recombination events in the maternal and paternal map, respectively. From the markers that were heterozygous in only one of the parents, 98% could be fit in the bins with a LOD threshold of 4. Of the markers that were heterozygous in both parents (bridge markers), only 79% could be fit with a LOD threshold of 15. These thresholds were determined by a permutation test. After fitting 10,000 random markers into the bins, less than 0.1% of the markers fit into the framework map with LOD scores higher than either 4 or 15, for 1:1 and 3:1 segregating markers respectively. In both parental maps the twelve chromosomes could be identified. In addition, the paternal map includes a small unassigned linkage group with a severe segregation distortion. Singletons did not occur randomly among the markers. Many markers were without singletons and the 10% of the markers with the poorest data quality accounted for more than half of the total amount of 33489 singletons in the data set. Segregation distortion was observed in the maternal map on linkage group V and in the paternal map on linkage groups I and XII. Markers are non-randomly distributed across the map. Putative centromeric positions showed extensive marker clustering while putative recombination hot spots resulted in large intervals up to 15 cM without markers. The markers derived from EcoRI/MseI and SacI/MseI enzyme combinations clustered more frequently than those derived from PstI/MseI enzyme combinations. The distribution of recombination events per chromatid suggested an absence of the occurrence of 0 chiasmata and a presence of more than one chiasma per chromosome arm.

The marker saturation of the ultra-dense map can be used for gene cloning and BAC landing. Proof of concept was recently obtained by the cloning of the late blight resistance gene R3a and the construction of a BAC contig comprising the wart disease resistance gene Sen1-4 (see Chapter 5 for references). The ultra-dense map will thus not only be very helpful with the anchoring of BAC-contigs for a sequence ready potato physical map, but will also prove its value for further gene cloning projects.

Samenvatting

Het onderzoek dat wordt beschreven in dit proefschrift, behandelt de problemen die optreden en de oplossingen die verkregen zijn bij het maken van een ultradichte genetische koppelingskaart van aardappel. Een F1 populatie van 130 individuen van een kruising tussen twee heterozygote diploïde aardappelklonen werd geanalyseerd met behulp van 381 AFLP primercombinaties en dit resulteerde in ongeveer 10.000 merkers. Tijdens de eerste stadia van het analyseproces werd al snel duidelijk dat de beschikbare software niet kon omgaan met zulke grote hoeveelheden aan gegevens. Koppelingsgroepen met meer dan 1000 merkers kunnen niet door JoinMap worden behandeld, en zelfs een klein percentage aan singletons (i.e. dubbele recombinaties of foutscores) leidt tot grote onnauwkeurigheid in het vaststellen van de merkervolgorde. Als gevolg hiervan werden nieuwe concepten ontwikkeld waaronder het software programma RECORD (Hoofdstuk 3), dat in een relatief korte tijd nauwkeurige merkervolgordes kan produceren. De noodzaak om singletons te verwijderen werd beschreven in Hoofdstuk 4 waar ook een statistisch computerprogramma (SMOOTH) gepresenteerd wordt dat de singletons verwijderd. De combinatie van deze twee programma's maakte het mogelijk om via een dataset waarin tegenstrijdigheden verwijderd waren, een betrouwbaar en robuust geraamte van de kaart te maken. Het geraamte van de kaart bestaat uit compartimenten, die elk een positie op de genetische kaart voorstellen, begrensd door recombinatiegebeurtenissen. Elk compartiment ontleent zijn identiteit en positie aan een uniek uitsplitsingspatroon en bevat variabele aantallen AFLP merkers. Dankzij het hoge verzadigingsniveau als gevolg van de 10.000 merkers, was het mogelijk om de meeste recombinatiegebeurtenissen in de populatie vast te stellen. Tenslotte werden alle ongekuiste merkergegevens aan de compartimenten van het geraamte van de kaart toegekend op basis van de methode van de grootste aannemelijkheid.

Hoofdstuk 2 beschrijft het eerste resultaat van de nieuwe karteringsprocedure met koppelingsgroep I, de grootste groep qua merkeraantallen. Koppelingsgroep I bestaat uit 95 maternale compartimenten en 101 paternale compartimenten. De 1260 AFLP merkers zijn niet gelijkmatig verdeeld over de genetische kaart. In de kaarten van beide ouders kwamen de meeste merkers voor als cluster in één compartiment en werden gaten van tot wel zeven lege compartimenten gevonden ondanks de verzadiging van het chromosoom met merkers. Merkers afkomstig van de verschillende enzymcombinaties EcoRI/MseI, SacI/MseI en PstI/MseI lieten een verschillende mate van clustering zien. Ongeveer driekwart van de merkers paste zeer goed in de compartimenten, uitgaande van een geschat 'singletongehalte' van 0-3%. De PstI merkers daarentegen bereikten bijna twee keer zo vaak een singletongehalte van meer dan 3%, hetgeen suggereert dat de ouderlijke methylatiepatronen waren veranderd in de nakomelingen.

De nieuwe software RECORD (REСombinaties tellen en ORDenen) wordt beschreven in Hoofdstuk 3 en kan worden gebruikt voor het ordenen van loci in koppelingsgroepen. Het beoordelingscriterium van deze methode is gebaseerd op het minimaliseren van het totale aantal recombinatiegebeurtenissen per koppelingsgroep. Het zoekalgoritme is een heuristische procedure, die elementen van 'branch-and-bound' met 'local reshuffling' combineert. Aangezien het voorgestelde criterium geen intensieve berekeningen nodig heeft, produceert het algoritme snel een optimale volgorde. Een simulatiestudie werd uitgevoerd om de prestaties van RECORD en bestaande software (JoinMap) te vergelijken. RECORD is veel sneller en minder gevoelig voor singletons, aangezien het optimalisatiecriterium minder gevoelig is voor het effect van singletons. In het bijzonder excelleert RECORD in regionen van de kaart met hogere merkerdichtheid.

De statistische methode die ontwikkeld is voor het identificeren en verwijderen van singletons uit de genetische merkergegevens tijdens het karteringsproces, is beschreven in Hoofdstuk 4. Het programma SMOOTH berekent het verschil tussen de waargenomen en voorspelde waarden van een specifiek datapunt, gebaseerd op datapunten van flankerende loci in een gegeven merkervolgorde. Hoogst onwaarschijnlijke datapunten worden verwijderd door het programma in een iteratief (i.e. zichzelf herhalend) proces met een karteringsalgoritme, dat de kaart opnieuw berekent na het verwijderen. SMOOTH is getest met behulp van simulaties en experimentele karteringsgegevens van de ultradichte kaart van aardappel. De simulaties tonen aan dat deze methode goed blijft presteren zelfs als het percentage singletons onrealistisch hoog is (tot 20%). Het stelt karteringssoftware in staat tot het succesvol vervaardigen van een zeer nauwkeurige hogedichtheidskaart.

Tenslotte is de complete ultradichte kaart van aardappel gepresenteerd in Hoofdstuk 5. Een geraamte van de kaart was verkregen met een lengte van respectievelijk 977 en 1005 recombinaties in de maternale en paternale kaart. Van alle merkers die heterozygoot waren in één van beide ouders, kon 98% in dit geraamte ingepast worden met een LOD van tenminste 4. Van de merkers die heterozygoot waren in beide ouders konden hoogstens 79% worden ingepast met een LOD van tenminste 15. Deze drempelwaarden zijn vastgesteld door middel van een permutatietest. Na het passen van 10.000 willekeurige merkers in de compartimenten, pasten slechts 0,1% van de merkers in het geraamte van de kaart met LOD-waarden hoger dan 4 of 15 respectievelijk voor de 1:1 en 3:1 uitsplitsende merkers. De twaalf chromosomen konden in de kaarten van beide ouders worden geïdentificeerd. Daarnaast bevatte de paternale kaart een kleine onbekende koppelingsgroep met een ernstig afwijkende uitsplitsingsverhouding. Singletons waren niet gelijkmatig verdeeld over de merkers. Veel merkers bevatten helemaal geen singletons en de 10% merkers met de laagste kwaliteit waren goed voor meer dan de helft van de in totaal 33489 aanwezige singletons in de gegevens. Scheve uitsplitsingsverhoudingen werden gevonden in koppelingsgroep V van de maternale kaart en koppelingsgroepen I en XII van de paternale kaart. Merkers zijn niet gelijkmatig verdeeld over de kaart. Mogelijke centromeerposities lieten enorme merkerclustering zien, terwijl mogelijke recombinatie hotspots resulteerden in grote gaten tot wel 15 cM. Merkers afkomstig van EcoRI/MseI en SacI/MseI enzymcombinaties clusterden meer dan die afkomstig van PstI/MseI enzymcombinaties. De verdeling van recombinatiegebeurtenissen per chromatide suggereerde een afwezigheid van 0 chiasmata en de aanwezigheid van meer dan één chiasma per chromosoomarm.

De verzadiging van merkers van de ultradichte kaart kan worden gebruikt voor het cloneren van genen en BAC-landing. Bewijzen voor de waarde van de ultradichte kaart hiervoor zijn recent verkregen door middel van het cloneren van het aardappelziekte resistentiegen R3a en de vervaardiging van een BAC-contig dat het wrattenziekte resistentiegen Sen1-4 omvat (zie Hoofdstuk 5 voor literatuurverwijzingen). Daarnaast zal de ultradichte kaart erg nuttig zijn bij het genetisch verankeren van BAC-contigs van een fysische kaart van aardappel, zodat chromosoomspecifieke en minimaal overlappende BACs geleverd kunnen worden voor DNA-sequentieanalyse. Ook voor toekomstige en nog lopende projecten voor gen-clonering zal de ultradichte kaart waarschijnlijk zijn blijvende waarde bewijzen..

Nawoord

Het is zover: het proefschrift is nu bijna klaar om gedrukt te worden. Graag wil ik op deze plaats iedereen van harte bedanken die direct of indirect een bijdrage aan dit proefschrift heeft geleverd.

Beste Herman, als mijn directe begeleider en copromotor heb jij verreweg de belangrijkste rol vervuld in mijn promotiewerk. Ik was nog bezig met een afstudeervak, toen je mij al aannam om computer-AIO te worden bij de aardappelgroep. We hebben vele interessante en soms ook lange gesprekken gevoerd over wetenschap, maar ook over maatschappelijke en persoonlijke zaken. Ik had vaak wel moeite met de hoge eisen die je aan mij en mijn werk stelde, maar ik denk dat dat achteraf ook wel nodig is geweest.

Richard, jij bent in de loop van mijn AIO-tijd mijn belangrijkste promotor geworden door het stokje van Evert Jacobsen over te nemen. In die tijd heb je je langzamerhand ook steeds meer bemoeid met mijn werk. Dankjewel dat je uiteindelijk de vaart erin gezet hebt om ervoor te zorgen dat ik dit werk kan verdedigen.

Mijn twee paranimfen, Jaap en Manga, ik ben blij dat jullie me willen bijstaan tijdens de laatste loodjes van mijn promotie. Jaap, hoe vaak hebben we niet samen gepingpongd, choco de luxe of cola gedronken en ons hart uitgestort over hardlopen en fietsen (jouw interesse) en muziek en theater (mijn interesse)? Wat mij betreft zeker niet te vaak! Manga, ik ben er erg trots op dat ik de ceremoniemeester van jouw bruiloft met Saskia mocht zijn. De band die we met elkaar hebben komt ongetwijfeld ook doordat ik jouw geboorteland Kameroen heb bezocht. Ik hoop dat we alledrie snel weer volledig aan de slag kunnen en ik wens jullie veel succes met jullie eigen laatste loodjes van de promotie.

Verder wil ik hier mijn kamergenoten bedanken. Marieke, ik kijk terug op een gezellige tijd die we voornamelijk achter onze computer tegenover elkaar hebben doorgebracht. Nelleke, we zijn al een tijdje geen collega's meer, maar ik stel het contact dat we hebben enorm op prijs. Bedankt voor je steun, je gezelligheid en alle keren dat ik bij je mag blijven eten. Ook de rest van mijn kamergenoten: Guusje, Yuling, Wole, Asun, Niek, Luisa, dankjewel.

Beste Piet, als ik er even niet meer helemaal uitkwam, was jij er om de orde in chaos te scheppen. Informeel ben je voor mij ook altijd een beetje een promotor geweest.

Annie, het hart van plantenveredeling, dankjewel voor de zorg over de financiële en secretariële zaken. Ik weet nog goed hoe ik als jong studentje schoorvoetend over de drempel van het secretariaat kwam om mijn cijfers op te halen en Annie me aankeek en zei: "Zo meneer Van Os, u heeft het gehaald hoor."

Hoewel we niet zo vaak een bijeenkomst hebben gehad die pas na 10 uur 's ochtends werd gehouden, wil ik ook de leden van de aardappelgroep bedanken voor het meedenken en bediscussiëren van mijn werk waaronder Carolina Celis, Ronald Hutten en Jaap Buntjer. Hierbij wil ik ook graag de partners van het UHD-project bij nematologie en in het buitenland meenemen: Sandra Andrzejewski, Erin Bakker, Imanol Barrena, Glenn Bryan, Bernard Caromel, Bilal Ghareeb, Edwige Isidore, Walter de Jong, Paul van Koert, Véronique Lefebvre, Dan Milbourne, Enrique Ritter, Jeroen Rouppe van der Voort, Françoise Rousselle-Bourgeois, Joke van Vliet, Robbie Waugh en Jaap Bakker.

Verder heel de vakgroep (ik bedoel natuurlijk laboratorium voor plantenveredeling) dankjulliewel voor de informele zaken rond en tijdens het werk en dan bedoel ik niet alleen het klaverjassen. Van al mijn collega's wil ik Arnold toch nog even speciaal bedanken voor sessies die we samen gehad hebben om mijn gedachten te kunnen ordenen en weer structureel aan de slag te gaan met schrijven. Arnold, jij was mijn eerste afstudeervakbegeleider, ik ben blij dat je me een week eerder voorgaat met je promotie.

Ongeveer gelijktijdig met mijn werk als AIO, ben ik in de musicalwereld gestapt. Eerst bij de musicalvereniging Sempre Sereno, maar al gauw waren de nevenactiviteiten van de vereniging onder de bezielende leiding van Willem van Roekel uitgegroeid tot een serieuze aparte stichting. Roekeloos was al die jaren een geweldige uitlaatklep, maar ook een goede leerschool. Willem en alle andere 'Roekelozen', dankjewel voor de jarenlange intensieve samenwerking. Ik wens jou en je stichting veel succes in alle nog komende producties.

Het meest dierbaar is me mijn familie. Lieve Mirjam, Irene, Paul, Eline, Mi Sun en Se Woong, bedankt dat jullie er zijn. Paul, we hebben veel opgetrokken in de tijd dat we samen op het huis pasten. Het was een fijne tijd! Lieve Sam, Anna en Floris, wat is het toch geweldig om jullie oom te kunnen zijn.

Tot slot wil ik mijn ouders bedanken. Lieve papa en mama, ik ben jullie enorm dankbaar voor de geweldige steun en het vertrouwen in mijn keuzes. Ik hoop dat jullie nog jaren kunnen genieten van het pensioen en de kleinkinderen en van elkaar.

Hans

Curriculum Vitae

Hans van Os werd op 15 augustus 1976 geboren in Langbroek. In juni 1994 behaalde hij aan het Revius Lyceum te Doorn het VWO diploma. In datzelfde jaar begon hij zijn studie plantenveredeling en gewasbescherming aan de Wageningen Universiteit. Bij het laboratorium voor plantenveredeling heeft hij als student onderzoek verricht naar de rol van artificiële intelligentie bij de voorspelling van F1 hybriden en merkergestuurde selectie vergeleken met fenotypische selectie bij Arabidopsis. Zijn stage heeft hij doorgebracht in Nieuw Zeeland waar hij werkte aan een detectiemethode van aspergevirus II en het mappen van apomixie in Hieracium. In juni 1999 behaalde hij zijn doctoraal diploma. In april van datzelfde jaar trad hij als AIO in dienst bij het laboratorium voor plantenveredeling aan de Wageningen Universiteit in het project over de constructie en toepassing van een multifunctionele ultra-dichte genetische kaart van aardappel. De resultaten van dit door de EU gefinancierde project staan beschreven in dit proefschrift.

The research presented in this thesis was performed at the Laboratory of Plant Breeding of Wageningen University and were part of the programme of the graduate school 'Experimental Plant Sciences'. Financial support was kindly provided by the European Union grant FAIR5-PL97-3565.

Printed by Ponsen & Looijen BV, Wageningen