T.J.A.Borm

Construction and use of a physical map of potato

Promotor:	Prof. dr. R. G. F. Visser Hoogleraar in de plantenveredeling.
Co-promotor:	Dr. ir. H. J. van Eck Universitair docent, leerstoelgroep plantenveredeling

Samenstelling promotiecommissie:
Prof. dr. ir. T. Gerats (Radboud Universiteit Nijmegen)
Prof. dr. M. Groenen (Wageningen Universiteit)
Dr. ir. H. Janssen (Plant Research International)
Prof. dr. ir. J. Bakker (Wageningen Universiteit)

Dit onderzoek is uitgevoerd binnen de onderzoekschool Experimental Plant Sciences

T.J.A.Borm

Construction and use of a physical map of potato

Proefschrift
ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van wageningen Universiteit
Prof. dr. M. J. Kropff
In het openbaar te verdedigen
op dinsdag 11 november 2008
des namiddags te half twee in de Aula

Construction and use of a physical map of potato
T.J.A. Borm

PhD thesis, Wageningen University, the Netherlands
With references – with abstracts in English and Dutch

ISBN 978-90-8585-237-7

Contents

Chapter 1

General introduction

T.J.A.Borm, H.J. van Eck and R.G.F. Visser

Originating from Southern America, cultivated potatoes were introduced to other parts of the world from the early 16th century onward (Glendinning 1983). As shown in Table 1, today cultivated potato is one of the major human food crops (Table 1). Potatoes are rich in energy, proteins and vitamin C (Scott et al. 2000); protein and energy yield per acre is higher than for cereals (Horton 1988), and only soybean surpasses potato in protein yield per unit area (Tarn 1992). Its ability to adapt to various climates ranging from tropical to temperate and altitudes (Doehlonan and Sleper 1995) also makes potatoes accessible to most of the world population. The importance of potato for the human food supply has been recognized by the United Nations, declaring 2008 the UN International Year of the Potato (http://www.potato2008.org).

With nearly 80% of seed tuber exports in the hands of Dutch companies, and occupying in excess of 20% of the arable land in the Netherlands, potato is also commercially a very important crop in the Netherlands. Besides export of seed potato and being an important staple food crop, potato starch and starch derivatives play an important role in a range of industrial application like paper-making, textiles and food processing.

Potato is a crop that is constantly under an array of biological and environmental threats (Bradshaw and Ramsay 2005), and therefore, and to improve crop quality and yield, potato breeders need to utilize the pool of genetic variation provided by both wild and cultivated potato. Thorough, systematic characterization of the potato genome is expected to contribute to reaching this goal (Stuber et al. 1999), and thereby help safeguard an abundant supply of healthy food to an increasing world population.

Table 1: Rankings of the seven (by quantity) most important human food crops.

rank	**crop**	**Quantity (*10^9 kg)**	**Area (*$10^9 m^2$)**	**Yield (kg/m^2)**
#1	Maize	660.45	1421.30	0.47
#2	Wheat	597.70	2146.30	0.28
#3	Rice, paddy	595.87	1501.00	0.40
#4	Potatoes	320.11	191.93	1.67
#5	Vegetables, nec	240.28	169.85	1.42
#6	Cassava	196.92	177.19	1.11
#7	Soybeans	193.82	846.67	0.23

Table produced using a five year average (2001-2005) of FAO data for the whole world (FAOSTAT 2007).

There are several ploidy levels observed in potato, ranging from monoploid (2n=x=12) to hexaploid (2n=6x=72), with cultivated potato (*Solanum tuberosum* Group Tuberosum)

commonly being tetraploid. Issues fundamental to linkage analysis in polyploid species (Luo et al. 2006) make genetic mapping using diploid clones a far more favorable approach, and fortunately dihaploids (2n=2x=24) can be obtained though anther culture or parthenogenesis (Hermsen and Verdenius 1973; Ortiz and Peloquin 1994). Several genetic linkage maps have been produced of diploid potato (e.g. Bonierbale et al. 1988; Gebhardt et al. 1989, 2001; Jacobs et al. 1995; van Eck et al. 1995; Milbourne et al. 1998), and recently an ultra dense linkage map of potato consisting of more than 10,000 markers has been published (van Os et al. 2006).

Bacterial Artificial Chromosome (BAC, Shizuya et al. 1992) libraries are a commonly used tool to study a genome in more detail than a genetic linkage map affords. With DNA fragments cloned from the donor organism routinely between 80 and 300 kb in length, the BAC cloning system allows stable cloning of a complete gene, including some of its interesting context in the genome (e.g. promoter regions and paralogues from a gene cluster), which can then be used to perform a variety of different types of research. Until now, potato BAC libraries have been used mainly for map-based cloning of disease resistance genes and the construction of local physical maps (van der Vossen et al. 2000, 2003; Ballvora et al. 2002; Paal et al. 2004; Kuang et al. 2005; Chen et al. 2004), for hybridisation screening (Fu et al. 2001), or for the purpose of cytogenetic chromosome identification (Song et al. 2000). Every time a BAC library is constructed and used for a specific purpose, considerable effort is put in assigning some of the clones to linkage groups in a genetic map, and the implied repetition of labor led to the realization that a systematic, genome-wide, alignment of BACs in a BAC library to a genetic map would facilitate and accelerate potato research. In crops like sorghum (Klein et al. 2000), soybean (Wu et al. 2004) and rice (Chen et al. 2002), genome-wide, genetically anchored physical maps have been constructed from BAC libraries. First step in construction of such an integrated physical and genetic map is ordering of BAC clones into physically overlapping groups of BAC clones, called contigs, most often performed on the basis of individually generated fingerprints (Sulston et al. 1988; Soderlund et al. 1997, 2000), followed by anchoring these contigs to a genetic map by screening them for genetically mapped markers.

The male parent of the ultra dense genetic map, RH89-038-16 (abbreviated "RH") (Rouppe van der Voort et al. 1997), shows excellent vigor and male fertility, and has been used as parental clone in two other diploid mapping populations (Rouppe van der Voort et al. 1998, 2000), making it, from a molecular genetic viewpoint, a thoroughly characterized potato genotype, and hence an ideal target for the construction of a genome wide genetically anchored physical map of potato. Genome wide physical maps can also serve as a template or scaffold to sequence a genome (e.g. Matsumoto et al. 2005), and

recently, members of the Potato Genome Sequencing Consortium (PGSC, http://www.potatogenome.net) have commenced sequencing the whole 850Mb potato genome (Arumrhanathan and Earle 1991), with the first sequenced chromosome scheduled to be delivered by 2009. This sequencing effort is based on the BAC libraries and the integrated genetic and physical map that are the subject of this thesis.

Objective of this thesis

The objective is to construct a genome wide, genetically anchored, BAC-based physical map to facilitate potato research in general, and sequencing the potato genome specifically. Construction, fingerprinting and characterization of the BAC library is discussed in chapter 2 of this thesis. Chapters 3 and 4 discuss genetic mapping methods and software leading to a refined version of the ultra dense genetic map to which BAC contigs will be anchored using the method presented in chapter 6. During fingerprinting of the BAC library unexpected fragments sizing deviations were observed in the fingerprints leading to inferior quality contigs if left uncorrected, and chapter 5 discusses the measures that have been taken to repair this. Physical map construction and anchoring of the contigs to the ultra dense genetic map using a novel, highly efficient, anchoring method are the subject of chapter 6.

Chapter 2

Construction and characterization of a fingerprinted BAC library of potato as the resource for sequencing the potato genome

T.J.A.Borm, T. Jesse, J. de Boer, B. Brugmans, J.S. Werij, H.J. van Eck and R.G.F. Visser

Abstract

We have constructed and characterized a BAC library and derived resources (DNA isolates, DNA pools) of diploid potato (*Solanum tuberosum* ssp. Tuberosum) genotype RH89-038-16, the male parent of the ultra dense, 10,000 marker map of potato recently published. The BAC library provides approximately 11.7 times coverage of the haploid genome, and is stored in 204 384-well micro-titre plates. Average insert size is 131 kb, with approximately 3% empty vector clones. This BAC library was fingerprinted using an AFLP™ based fingerprinting protocol, resulting in between 9.6 and 10.2 times coverage of the potato genome in usable fingerprints. BACs containing chloroplast DNA derived inserts (3.8% of the clones) were identified using PCR and a preliminary physical map built from the fingerprints using the FingerPrinted Contigs program (FPC), and later confirmed using BAC end sequences obtained from genbank. We demonstrate that, though it is not possible to precisely identify the causes of failures in the fingerprinting process, it is possible to exercise quality control on the fingerprints, and reject unwanted fingerprints based on objective criteria. This rejection results in great improvements in time required to compute physical maps, and, as many of the rejected fingerprints are artifacts caused by various types of failures in the fingerprinting and band-calling process, we assume that removal will prevent accidental inclusion of these fingerprints in contigs, thereby improving the quality of a genetically anchored physical map that is being constructed. Selected clones from this BAC library are currently being sequenced by members of the Potato Genome Sequencing Consortium (PGSC) (http://www.pototogenome.net), and this effort is expected to result in the complete elucidation of the potato genome sequence.

Introduction

BAC libraries are generally considered valuable resources for genomic research. Previously constructed potato BAC libraries have been instrumental for map-based cloning of disease resistance genes and the construction of local physical maps (Van der Vossen et al. 2000, 2003; Ballvora et al. 2002; Paal et al. 2004; Kuang et al. 2005; Chen et al. 2004), for hybridisation screening (Fu et al. 2001), and for cytogenetic chromosome identification (Song et al. 2000). In crops like sorghum (Klein et al. 2000), soybean (Wu et al. 2004) and rice (Chen et al. 2002) genome-wide, genetically anchored, physical maps have been constructed from fingerprinted BAC libraries, accelerating research into these crops and providing a basis for sequencing (e.g. Matsumoto et al. 2005).

Recently an ultra dense genetic map of potato, consisting of more than 10,000 markers has been published (van Os et al. 2006), and the male parent of this genetic map, RH89-038-16 (abbreviated "RH") (Rouppe van der Voort et al. 1997) has also been used in two other mapping populations (Rouppe van der Voort et al. 1998, 2000), making it, from a molecular genetic viewpoint, a thoroughly characterized potato genotype, and hence an ideal target for the construction of a genome wide genetically anchored physical map of potato.

Construction of a genome-wide, fingerprint-based, physical map requires a specified coverage of the genome in fingerprints of sufficient quality. Fingerprinting failures, extrachromosomal DNA-inserts and empty vector clones all result in fewer acceptable fingerprints and hence in reduced genome coverage for a BAC library of a given size. Plant BAC libraries intended for fingerprint based physical map construction are often characterized for empty vector clones and contamination with organellar DNA as well as being fingerprinted (e.g. Tomkins et al. 2002; Coe et al. 2002 and Nelson et al. 2005 for maize and Wu et al. 2004 for soybean). However, this is not always the case (e.g. Chen et al. 2002 for rice), and in the latter case generally little information on library quality and contaminants is available. Prior characterization may in some cases have led to rejection of BAC libraries in their entirety, but to our knowledge never to re-arraying or selective fingerprinting as such actions would represent an expensive disruption of a high throughput fingerprinting process. Given that production of BAC libraries nowadays is a well established technique, using optimized protocols routinely yielding BAC libraries of high quality, it can be questioned if the additional effort required for characterization prior to or in parallel with fingerprinting is useful, or if perhaps similar information can be deduced from the fingerprints.

In the present paper we will describe the construction and fingerprinting of an approximately 10 genome-equivalent, high insert length, potato BAC library and derived

resources, and present a method for BAC library and fingerprint characterization based on analysis of raw fingerprint data and preliminary Finger-Printed Contigs built using FPC (Soderlund et al. 2000). Central premise in these analyses is that unacceptable fingerprints can be detected, either because the fingerprints themselves deviate from the expected, or because the fingerprints form groups that deviate from the expected.
In combination with the ultra-high density linkage map of potato (van Os et al. 2006), this BAC library is being used for the ongoing construction of a genome-wide genetically anchored physical map, which serves as the basis for the sequencing of the 850 Mbp (Arumuganathan and Earle 1991) potato genome by the members of the Potato Genome Sequencing Consortium (PGSC, http://www.potatogenome.net), scheduled to deliver the completely sequenced potato genome by the end of 2010.

Materials and methods I (molecular biology)

BAC vector isolation

The BAC vector pIndigoBAC536 DNA was isolated using the Qiagen Gigaprep isolation kit. The vector DNA was completely linearized with HindIII or EcoRI (New England Biolabs (NEB)), dephosphorylated with CIP (NEB) and the product was run on a 1% agarose gel without ethidium bromide. The DNA fragments of 7.5 kb were excised from the gel and electro-eluted. The quality of each vector batch was assessed with a self ligation with T4 ligase (Promega) and a ligation with lambda fragments cut with HindIII or EcoRI and subsequent transformation of the ligated product into electro-competent *E. coli* DH10B cells (Research Genetics).

Preparation of partially digested DNA of Potato

High-molecular weight DNA was prepared from young leaf nuclei of *Solanum tuberosum* genotype RH89-039-16. The leaf tissue was stored at -80°C directly after harvest. Nuclei were extracted by grinding 40 grams tissue in liquid nitrogen to a fine powder and further processing was performed as described by Budiman et al. (2000) and De Scenzo and Wise (1996). The samples were tested for digestion with 500 Units HindIII or 500 Units EcoRI, and electrophoresis was performed under pulsed field conditions (CHEF DR-III, BioRad). Samples that showed complete digestion and no breakdown of DNA in the negative controls with 0 Units restriction enzyme were used for further partial digests. A titration was performed with 0, 30, 60, 90, 120 and 150 Units HindIII or with 0, 60, 70, 80, 90 and 100 Units EcoRI for 20 minutes at 37 °C. The incubations were stopped by adding 0.1 volume 0.5 M EDTA and the partially digested fragments were separated on a 1% agarose gel by pulsed field gel electrophoresis (1-45 sec, 14 hrs, 14°C, 120° angle, 5.8 V/cm, 0.5

TBE buffer). Two fractions (100-150 and 150-250 kb) were excised and used in a second sizing PFGE step (5 sec, 14 hrs, 14°C, 120° angle, 5.8 V/cm, 0.5 TBE buffer). For each fraction the fragments larger than 100 kb were electro-eluted, quantified on 1% agarose gel and used for ligation with the pIndigoBAC536 vector.

BAC library construction

The ligations were performed in 100 µl volume, with approximately 100 ng eluted High Molecular Weight (HMW) DNA and 20 ng linearized pIndigoBAC536 vector. Before adding the T4 ligase the ligation mixture was incubated at 55°C for 10 min and then cooled down to room temperature. The ligation reaction was performed at 16°C overnight. Further processing of the ligation mix was performed according to Budiman (2000). The BAC library was stored in 384 well plates at -80°C.

High throughput BAC DNA isolations

BAC DNA was isolated using a high throughput protocol, adapted from Sambrook et al. (1989). BAC clones were replicated directly from 384-well storage plates using a 96 pin replicator tool into 96-well deep well plates containing 1.5 ml of Terrific Broth (Tartof and Hobbs 1987) supplemented with 12.5mg/L chloramphenicol per well. These culture plates were sealed by airpore tape sheets (Qiagen) and incubated for 24 hours at 37°C at 300 rpm. Cultures were pelleted for 15 minutes at 3000 rpm (a Heraeus Multifuge 3 S-R was used for all centrifugal steps). Plates were decanted and carefully tapped onto paper towels to remove as much culture supernatant as possible. Bacterial pellet size uniformity was observed, and any deviations noted. Pellets were resuspended in 150 µl of 50 mM Tris, 10 mM EDTA (pH 8), supplemented with 100 µg/ml RNAse A. Cells were lysed by adding 300 µl of 0.2 M NaOH /1% SDS) per well. The plates were vortexed at low speed, and left to sit at room temperature for 5 minutes. After adding 250 µl cold KAc buffer (5 M acetic acid adjusted to pH 4.8 with KOH) per well, the plates were sealed using a layer of parafilm and a polypropylene sealing mat and carefully inverted 10 times to mix. After overnight incubation in a refrigerator to allow complete formation of a cell debris precipitate, the plates were centrifuged for 20 minutes at 4600 rpm , and 310 µl of cleared lysate was carefully transferred to a new deep well plate. To precipitate the DNA, 220 µl of isopropanol was added and plates were sealed using a layer of parafilm and a polypropylene sealing mat, inverted a few times to mix, and left to sit on the laboratory bench for a couple of minutes or, alternatively, overnight in the refrigerator. After centrifugation (25 min at 4600 rpm) supernatant was drained and plates carefully tapped on a paper towel to remove remaining droplets. DNA pellets were washed using 0.7 ml of 70% ethanol, plates drained, and left upside-down on a paper towel for 30 minutes to dry.

50 µl of TE0.1 buffer (10 mM Tris-HCl, 0.1 mM EDTA pH 8) was added to the DNA pellet and plates were sealed using polypropylene sealing mats, lightly vortexed and left overnight in the refrigerator to dissolve the DNA before freezing (-20°C) for long term storage.

BAC insert size analysis

BAC DNA of a sample of BACs, obtained from either the high throughput DNA isolation method described above or alternatively obtained from an essentially unmodified miniprep as described by Sambrook et al. (1989) with 3ml culture volumes, was digested with NotI (NEB) to completion and separated by field inversion gel electrophoresis (BioRad FIGE MAPPER™) on a 1% agarose gel in 0.5x TBE, with a linear run time, forward (3-30s) reverse (1-10s), 14 hrs and 160 Volts, along with a mid-range PFGE marker I (NEB). The BAC insert sizes were determined conservatively with an estimated error of ~ 5 kb for each insert.

BAC DNA pool construction

High throughput DNA isolation was performed in batches of 96 BACs (a quarter library plate) at a time. After transferring 310 µl of lysate to a new deep well plate for individual BAC clone DNA isolations, approximately 200-250 µl of lysate is left in each well of the deep well plate. To construct Quarter Plate Pools (QPPs) this remaining lysate is collected (for each 96-BAC batch separately) into a small container and filtered over synthetic aquarium wool packed in a 5ml pipette tip to remove remaining debris. For each deep well plate approximately 25ml of lysate was collected into a 50ml tube. A 5ml sample of lysate of the four QPPs belonging to the same 384 well storage plate was combined in a new 50ml tube to construct a Full Plate Pool (FPP). DNA was precipitated by adding 0.7 volumes isopropanol (or about 15 ml) to each pool (QPP or FPP), inverting a few times to mix, then left to sit on the laboratory bench for a couple of minutes or, alternatively, overnight in the refrigerator. After centrifugation (45 min, at 3600 rpm), supernatant was discarded and tubes carefully tapped on a paper towel to remove remaining droplets. DNA pellets were washed using 7.5ml of 70% ethanol, tubes drained and left upside-down on a paper towel for 1 hour to dry. 600 µl of TE0.1 buffer (10 mM Tris, 0.1 mM EDTA, pH 8) was added, tubes were lightly vortexed and left overnight in the refrigerator to dissolve the DNA. Dissolved DNA was transferred to Eppendorf tubes and frozen (-20°C) for long term storage.

DNA yield testing using Not-I digests

Yield of the DNA isolations was routinely tested for the majority of FPPs and QPPs and

for a sample of the individual clones by inspecting (ethidium bromide stained 1% agarose gel, run for 5 hours at 100V) a Not-I digest (2 µl pool DNA or 1 µl individual clone DNA, 1 unit Not-I in its applicable buffer, in a total volume of 25 µl) for the presence of both vector and insert bands.

BAC Fingerprinting using AFLP™

BACs were fingerprinted using AFLP using EcoRI/MseI restriction enzymes, essentially as described by Vos et al. (1995). During restriction and adapter ligation 2 µl BAC-DNA, isolated using the high throughput method, was substituted for the 0.5 µg genomic DNA. A single stage AFLP PCR reaction using primers without selective extension was performed using unlabeled MseI-adapter primers and differently labeled EcoRI-adapter primers (either FAM, NED or JOE), to allow multiplexing of three PCR products into a single capillary of a MegaBACE (GE Healthcare) capillary sequencer. Prior to electrophoresis PCR products were combined with 20 µl 5 pM et-ROX-labeled et-ROX 900 size ladder (GE Healthcare), mixed, and subsequently cleaned of excess salts and unincorporated primers using the AutoSeq 96 G50 cleanup kit (GE Healthcare) as per manufacturers' instructions. Capillary electrophoresis was performed at KeyGene, and electropherograms were exported from the MegaBACE machine in a raw data format ("RSD files") and processed (sizing and band-calling) using KeyGENE proprietary tools (Xpose and BAC-Xtractor), as described by Srinivasan et al. (2003). Band calling resulted in one file per BAC clone containing a list of AFLP fragment sizes, multiplied by 10 to facilitate processing using FPC (which requires integer band sizes). These files are hereafter referred to as "bands files".

PCR with organellar DNA specific primers

A selection of BACs was screened for chloroplast DNA derived inserts using PCR with three different chloroplast DNA specific primer pairs (Hamilton 1999, Aoki and Ito 2000, Taberlet et al. 1991). To detect mitochondrial DNA derived inserts, two primer pairs were designed and used on 96 of the 191 Full Plate Pools (FPPs). Primers sequences are shown in Table 1. In all cases PCR was performed on 0.5 µl DNA samples, using 2µl 10x PCR buffer, 4 mM $MgCl_2$, 0.1 mM dNTPs, 0.2 µM of each primer, 0.3 units SuperTaq, and H_2O to a volume of 20µl, using the following cycle conditions: 3'@94°C, 30 cycles (30"@94°C, 30"@50°C, 2'@72°C), 5'@72°C, hold@4C. Product was visualized on a 1% agarose gel stained with ethidium bromide.

Table 1: primer sequences used to detect organellar DNA contamination

Organelle	Target	Forward primer sequences	Reverse primer sequences	Product
Chloroplast	TrnLF	CGAAATCGGTAGACGCTACG	ATTTGAACTGGTGACACGAG	1015bp
Chloroplast	MatK	TAGATATACTAATACCCTACCCTGT	TGCCCGGGATTCGAACCCGGAACTA	1344bp
Chloroplast	PsbA	CGAAGCTCCATCTACAAATGG	ACTGCCTTGATCCACTTGGC	495bp
Mitochondrion	atp6	GGGAGGAGGAAACTCAGTA	GAATGCTCCACGACTAAGTAT	686bp
Mitochondrion	cob	AACCCCGAGCAATCTTAGTTA	GCGGCCAGATGAAGAAGAC	537bp

Materials and methods II (data analysis)

Introduction

During BAC library construction and AFLP-based fingerprinting there are many possible causes of failure, and it is likely that observable symptoms will vary accordingly. Some of the failures we expected to encounter and their possible effects are shown in Table 2, with arrows used to indicate chains of failures. As an example, reduced bacterial growth may lead to a low DNA yield, which may lead via a low template concentration et cetera. There are, however, multiple other routes leading to a low template DNA concentration, such as problems with DNA isolation, partial restriction, and partial ligation.

Different modes of failure may result in the same symptoms and multiple failures may coincide. Therefore it may be impossible to determine the cause of a failure using only the fingerprinting data. Detecting which fingerprints are unacceptable may, however, still be possible:

1. Given that bands detected in a fingerprint by the band calling software may have been produced by fundamentally different processes (e.g. PCR, crosstalk or bands called in noise), we expect that fragment size distributions may also be different. After establishing statistical distribution of AFLP fragment sizes, we can test if the bands in a fingerprint represent a sample from this distribution or not.
2. Given that a certain genome-coverage of a chromosomal region by (putatively randomly cloned) BAC clones should be reflected in a similar coverage (within a contig) by their fingerprints, we expect that contamination of the BAC library with organellar DNA, even at a sub-1%-level, will result in a contig of clearly excessive coverage. The fingerprints of such a contig may represent BACs containing organellar DNA.
3. Similarly, failures during DNA isolation followed by AFLP on template consisting predominantly of *E.coli* genomic DNA may lead to recognizable fingerprint patterns, and such patterns should be assembled into a large contig of excessive coverage, even at a sub 1% failure rates.

Both methods, analysis of fragment distributions and detection of excessive coverage in a contig, are used, in combination with other experiments, to classify fingerprints into different categories.

Table 2: Types of failure expected during BAC library construction and fingerprinting

Failure point	**Type of failure**	**Possible effect(s)**
Library construction	Chloroplast DNA	**Chloroplast fingerprint**
	No or short insert	Low AFLP template complexity ->
	Chimeric or long insert	High AFLP template complexity ->
Replication and growth	No colonies	No DNA ->
	Reduced growth	Low DNA yield ->
DNA isolation	-> No DNA	No AFLP template ->
	-> Low BAC DNA yield	Low AFLP template concentration ->
	Excess *E.coli* genomic DNA	***E.coli* fingerprint**
Restriction and ligation	No restriction or ligation	No AFLP template ->
	Partial ligation	Low AFLP template concentration ->
	Partial restriction	Low AFLP template concentration -> High AFLP template complexity ->
PCR	No primers	Empty fingerprint ->
	-> No AFLP template	**PCR artifacts (spurious bands)**
	-> Low template concentration	Low signal strength -> **PCR artifacts (spurious bands)**
	-> Low AFLP template complexity	**PCR artifacts (spurious bands)**
	-> High AFLP template complexity	Low signal strength -> **PCR artifacts (missing bands)**
	PCR failure	Empty fingerprint ->
Electrophoresis	Capillary failure	**No fingerprints in the affected capillary**
	Electro-kinetic injection failure	Empty fingerprint ->
	No sample	Empty fingerprint ->
	Partial injection failure	Low signal strength ->
Crosstalk correction	Signal out of range	**No fingerprints in the affected capillary**
	Bad correction parameters	**Spurious bands in other detector channels of the same capillary**
Band calling	-> Low signal strength	Low detection threshold ->
	-> Empty fingerprint	Low detection threshold ->
	-> Low detection threshold	**Detection of bands in noise or (residual) crosstalk from other channel(s)**

Arrows indicate chains of failure: an effect that is the same as a type of failure occurring lower in the table. Eventually chains of failures lead to a symptom observable in the fingerprint (shown in bold type). Effects leading to AFLP-like fingerprints are shown highlighted (gray); others (non-highlighted) may lead to non-AFLP-like fingerprints.

Statistical analysis of band distributions and band counts

It is assumed that some failed fingerprints will have a different distribution of fragment sizes than normal AFLP fingerprints because they arise through fundamentally different processes. All statistical analyses were performed using the R software package (Ihaka and Gentleman 1996). Where necessary simple PERL scripts were written to pre-process, filter and convert data to a suitable format. The Kolmogorov-Smirnov test (KS-test) (Chakravarti et al. 1967) was used throughout to obtain p-values indicating the probability that either the pair of fingerprints being tested are two samplings from the same distribution or indicating the probability that a single fingerprint is a sampling of the specified distribution.

Histograms of fragment lengths were obtained by counting the number of bands in one base-pair intervals in the bands files. These histograms were normalized by dividing the counts per interval by a factor so that the area under the histogram became equal to one. Either these normalized histograms were integrated numerically to obtain a cumulative distribution (in case of known empty fingerprint histograms) or a truncated geometric distribution was fitted to the normalized histogram, and subsequently this truncated geometric distribution was integrated numerically to obtain a cumulative distribution.

The equation describing the truncated geometric distribution governing AFLP fragments sizes was obtained from literature (Koopman and Gort 2004), and modified slightly to accommodate the length of the primer sequences flanking the AFLP restriction fragment: With p the probability of observing an A *or* T nucleotide in the sequence, we can trivially obtain the probabilities of observing an A, T, C or G nucleotide respectively ($p_A=p_T=p/2$ and $p_C=p_G=(1-p)/2$). Assuming EcoRI and MseI restriction enzymes and given:

l_{extra} :The primer contribution to the total length of the fragment,

l_{min} :The minimum fragment size that can be detected,

l_{max} :The maximum fragment size that can be detected and

the truncated geometric distribution becomes:

$$G(l)=\frac{(1-A)\times A^{l-l_{min}-l_{extra}}}{1-A^{l_{max}-l_{min}+1}}$$

With $A=(1-p_G*p_A*p_A*p_T*p_T*p_C)*(1-p_T*p_T*p_A*p_A)$. Fitting the observed AFLP fragment frequencies to this distribution yields an estimate of the AT fraction of the potato genome.

The KS-test is used in three different ways:

1. To compare fingerprints that are known not to result from AFLP (“Known Empty” or KE fingerprints) with the cumulative distribution of these KE fingerprints, to investigate the possibility of using this cumulative distribution to detect other “empty” fingerprints.

2. To compare (putatively normal) AFLP fingerprints to individual KE fingerprints, interpreting p-values above a threshold as an indication that the BAC fingerprint is an empty fingerprint.
3. To compare all fingerprints with a range of truncated geometric distributions (equivalent to a range in AT:GC nucleotide composition), with consistent low p-values indicating deviance.

Based on tests 2 and 3, fingerprints were divided into four classes: deviant but not empty ("FD"), empty and deviant ("ED"), neither deviant nor empty ("FN", or "good") and empty but not deviant ("EN") fingerprints. Based on the distribution of the number of bands encountered in fingerprints previously classified as "good", fingerprints were additionally classified as having a normal or an abnormal number of bands.

Preliminary physical map construction

Prior to assembly of the fingerprints into contigs using the FPC program, a simple program was used to remove all fragments smaller than 100 or larger than 650 base pairs from the bands files. Additionally, all BACs with either no bands or more than 250 bands after this fragment size filtering were removed from the analysis to accommodate limitations of the FPC program. Initial contigs were constructed using FPC, using "equation 2", a fixed sizing tolerance of 5 and a probability cut-off of 10^{-12}, with all other parameters remaining at their default settings. The contigs thus obtained were automatically refined by using "the deQer" to re-analyze contigs which contain in excess of 5 fingerprint patterns which seem to contradict (as determined by FPC) the contigs' consensus bands map. No further optimization of the contigs was attempted.

Analysis of fragment count versus BAC length

To qualitatively asses the possibility of using the number of bands in a fragment as a predictor of the BAC length, the FIGE size was plotted against the number of bands for individual BAC clones for which FIGE sizing data was available.

Identification of BACs with chloroplast DNA derived inserts

A sample of BACs from a single large contig in the preliminary physical map, together with some BACs from other contigs, were subjected to PCR with chloroplast DNA specific primers to establish if this particular contig contains BACs with chloroplast DNA derived inserts. To confirm the chloroplast identity of this contig, BAC end sequences for the BAC library were obtained from genbank, and a BLAST based search was conducted (all default settings, however without filtering of regions of reduced complexity) using the known chloroplast sequence of *Solanum tuberosum* (genbank accession ID nc008096,

pseudo-circularized by adding the first 1200 base-pairs of the sequence to the end), as a query. BLAST hits were subsequently filtered using a PERL script, retaining hits with e-values < 10^{-06}, >=95% identity and covering >= 95% of the BAC ends' lengths.

Identification of fingerprints with a putative *E.coli* genomic DNA fingerprint pattern

It was occasionally observed that a single culture plate yielded many similar, low signal strength, fingerprints. Similar fingerprints were also occasionally observed (personal communication Taco Jesse, Jan de Boer) at KeyGene in AFLP fingerprints of BACs derived from other organisms. It was surmised that these fingerprints were caused by some handling error, leading to loss of BAC DNA and (AFLP) amplification of residual *E.coli* genomic DNA. It was expected that FPC would group these in a single contig, and that we would be able to detect such "*E.coli* fingerprints" occasionally occurring in other plates by virtue of their inclusion in this "*E.coli* contig".

Results

High throughput BAC DNA isolation and fingerprinting

The BAC libraries consist of 204 numbered 384-well plates. Table 3 summarizes BAC DNA isolation and fingerprinting results. In total 423 (0.57%) cultures exhibited reduced or no growth. Of the 1344 BACs grown in duplicate, 14 exhibited reduced or no growth and 11 exhibited inconsistent growth. 764 Quarter Plate Pool (QPP) and 191 Full Plate Pool (FPP) DNA pool samples were constructed, and all these exhibited reproducible BAC DNA yield and clear separation of vector and inserts in the Not-I digests.

Table 3: DNA isolation and fingerprinting results

	#plates / #quarter plates	#BACs / %BACs fingerprinted	#Genome equivalents
Total number of BACs in libraries	204 / 816	78336 / 107%	11.5
Fingerprinting attempted	191 / 764	73344 / 100%	10.8
Re-fingerprinted because of bad quality	2.25 / 9	864 / 1.18%	0.13
Re-fingerprinted randomly	1.25 / 5	480 / 0.65%	0.07
Duplicate electrophoresis and band-calling	9.29 / 37.2 (*)	3568 / 4.86%	0.52
Reduced bacterial pellet size after growth	0.30 / 1.22 (*)	117 / 0.16%	0.02
No bacterial pellet after growth	0.76 / 3.04 (*)	292 / 0.40%	0.04
Permanent fingerprinting failures	0.24 / 0.97 (*)	93 / 0.13%	0.01

(*) indicates equivalent numbers as these BACs are scattered over multiple library plates. The number of genome equivalents was based on an average insert size of 127 k basepairs.

Insert sizes

Insert sizes obtained for a sample of clones of each of the 6 ligations are shown in Table 4.

Table 4: Insert size distributions

Ligation:	H1	H2	H3	H4	H5	H1-H5	E1	Total
Plates:	001-037	038-071	072-097	098-105	106-119	001-119	120-204	001-204
# Sized:	80	95	92	55	2	324	266	590
No insert:	1 (1%)	2 (2%)	2 (2%)	2 (4%)	0 (0%)	7 (2%)	11 (4%)	18 (3%)
<100kb:	0 (0%)	5 (5%)	4 (4%)	2 (4%)	0 (0%)	11 (3%)	22 (8%)	33 (6%)
100-120kb:	21 (26%)	16 (17%)	3 (3%)	9 (16%)	0 (0%)	49 (15%)	83 (31%)	132 (22%)
120-140kb:	44 (55%)	30 (32%)	38 (41%)	18 (33%)	0 (0%)	130 (40%)	88 (33%)	218 (37%)
140-160kb:	11 (14%)	22 (23%)	30 (33%)	7 (13%)	0 (0%)	70 (22%)	41 (15%)	111 (19%)
160-180kb:	3 (4%)	8 (8%)	12 (13%)	7 (13%)	1 (50%)	31 (10%)	14 (5%)	45 (8%)
>180kb:	0 (0%)	12 (13%)	3 (3%)	10 (6%)	1 (50%)	26 (8%)	7 (3%)	33 (6%)
Avg. (A):	127kbp	135kbp	137kbp	138kbp	180kbp *	134kbp	119kbp	127kbp
St.dev. (A):	21kbp	38kbp	33kbp	42kbp	10kbp *	34kbp	38kbp	37kbp
Avg. (B):	129kbp	138kbp	140kbp	143kbp	180kbp *	137kbp	124kbp	131kbp
St.dev. (B):	15kbp	32kbp	26kbp	33kbp	10kbp *	28kbp	30kbp	30kbp

Averages of the 5 HinDIII ligations ("H1-H5") and overall averages are shown against a gray background. Average (Avg.) size in kilo base pairs and standard deviations (St.dev.) have been calculated including (A) and excluding (B) empty vector clones. Figures marked "*" are based on only two clones, so should be treated with caution.

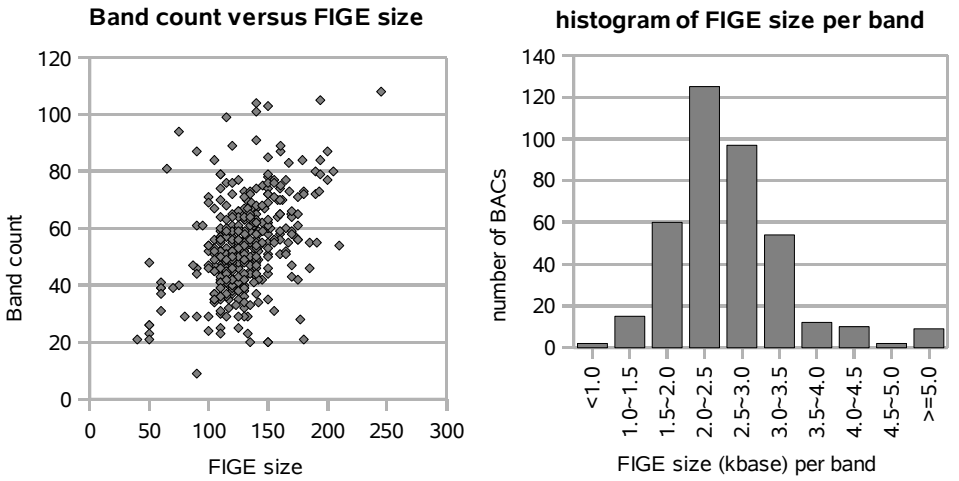


Figure 1: Fingerprint band count versus FIGE size (in kilo base pairs) for 386 identifiable, individually sized BACs (left), and a histogram of the amount of BAC DNA (according to the FIGE size) represented by each fingerprint band (right). On average there is one fingerprint band per 2.6 kbp of (FIGE sized) DNA, with a standard deviation of 0.95 kbp.

BAC insert size and fingerprint fragment counts are shown in Figure1 (left for individual BACs). Dividing the size by the number of bands in a BAC, the average amount of DNA represented by each detected fingerprint band is obtained, a histogram of which is also shown in Figure 1 (right). This figure clearly demonstrates that there is a poor correlation between the number of bands per fingerprint and the (FIGE) size of BACs.

Statistical analysis of band distributions

In each MegaBACE capillary there are nominally three BAC AFLP samples plus the et-ROX labeled size standard. In a fraction of the available sample slots, BAC AFLP samples were intentionally omitted. The size standard, however, was always present. Fingerprints resulting from such sample slots do not represent valid AFLP fingerprints, and we shall refer to these as "Known Empty" (KE) fingerprints. Figure 2 shows normalized histograms of (light gray) fragment sizes obtained all other (putatively normal) fingerprints and (dark gray) fragment sizes obtained from KE fingerprints, and a truncated geometric distribution (black) that was fitted to the green histogram. The observable difference in distributions between putatively normal (truncated geometric) and KE fingerprints (flat with peaks at positions suggesting et-ROX crostalk) immediately demonstrates that empty wells do not result in AFLP-like fingerprints. The fitted truncated geometric distribution corresponds to an estimated AT-content of the potato genome of 56.6%.

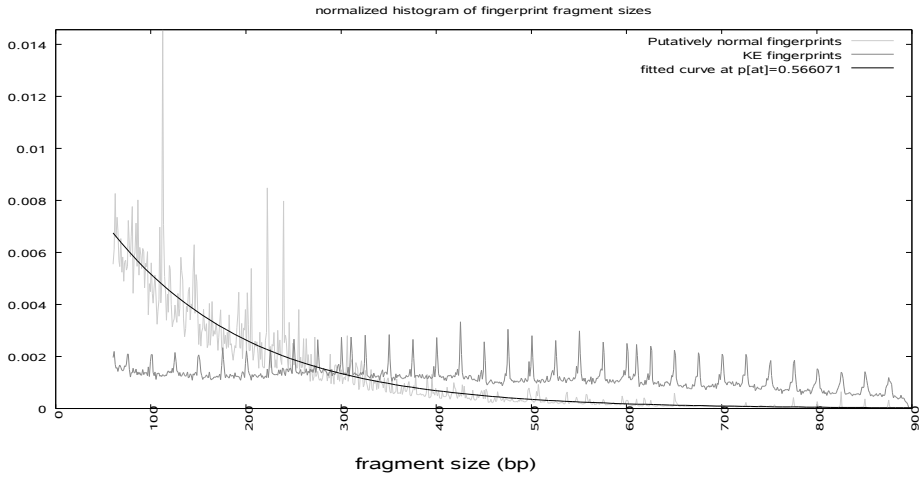


Figure 2: Normalized histograms (area under curve equals 1) of fragment sizes detected in putatively normal (light gray) and "known empty" (dark gray) fingerprints. Normalization allows a direct comparison of the histogram to a probability density function. The black curve is a truncated geometric distribution fitted to the putatively normal fingerprints (resulting in an estimate of the AT-content of the potato genome of 56.6%).

Using the Kolmogorov-Smirnov test (KS-test) to test each individual KE fingerprint against the cumulative distribution obtained from the histogram of all KE fingerprints (shown in Figure 2) results in 287 of the 384 individual KE fingerprints being discarded as *not* belonging to this distribution with a p-value of <=0.001. Moreover, only one KE fingerprint seems to belong to the same distribution as the histogram at a p-value>=0.99. This effectively rules out the use of the KE histogram based cumulative distribution in a KS-test to positively identify KE fingerprints. Visual inspection of the KE fingerprints (see Figure 3 for a sample) confirmed this; there seems to a continuum of different band distributions. The KE fingerprints evidently not being samples drawn from a single distribution, subsequently each KE fingerprint was assumed to represent a separate distribution. All 78,176 putatively normal BAC fingerprints were tested against each KE fingerprint, retaining the highest scoring match, classifying fingerprints based on their respective p-values (KS test #1). Next, all putatively normal fingerprints were tested against a range of truncated geometric distributions, corresponding to a range in AT nucleotide content from 45% to 70% in 5% increments, and the highest p-value was noted for each fingerprint (KS test #2). Combined results for KS test #1 and KS test #2 are summarized in table 5. We classified fingerprints as either Empty or Full (E or F) using KS test #1 and a p-value threshold of 0.10, and as either Deviant or Normal (D or N) using KS test #2 and a p-value threshold of 0.90. Each fingerprint is placed in a combined class: “ED”, “FD”, “EN” or “FN”, with 188, 2498, 71 and 75,419 fingerprints respectively. Figure 4 shows histograms for each of these groups. Figure 7 shows a representative sample of fingerprints from each of these 4 classes, in addition to some “*E.coli*” and “chloroplast” fingerprints.

Table 5: Classification of fingerprints based on the maximum p value obtained in two series of KS tests.

		p-value class in KS test #2 (range of geometric distributions)					
		p<=0.01	p<=0.1	0.1<p<0.9	p>=0.9	p>=0.99	total
p-value class in KS test #1 (KE fingerprints)	p<=0.01	208	2137	38036	6014	779	46187
	p<=0.1	218	**2306**	**58560**	**9985**	1222	70851
	0.1<p<0.9	38	**192**	**5809**	**1065**	128	7066
	p>=0.9	177	**188**	**64**	**7**	1	259
	p>=0.99	137	141	7	1	0	149
	total	433	2686	64433	11057	1351	78176
KE fingerprints		378	382	2	0	0	384

Figures with a white background are the number of fingerprints in a combined class; a light gray background highlights totals for each test. The bottom row shows results for KE fingerprints, which were only tested against the range of geometric distributions. The figures in bold type represent mutually exclusive events, and at a chi-square test statistic of 3775.6 and 4 degrees of freedom, it is unlikely that the KS tests yield independent results.

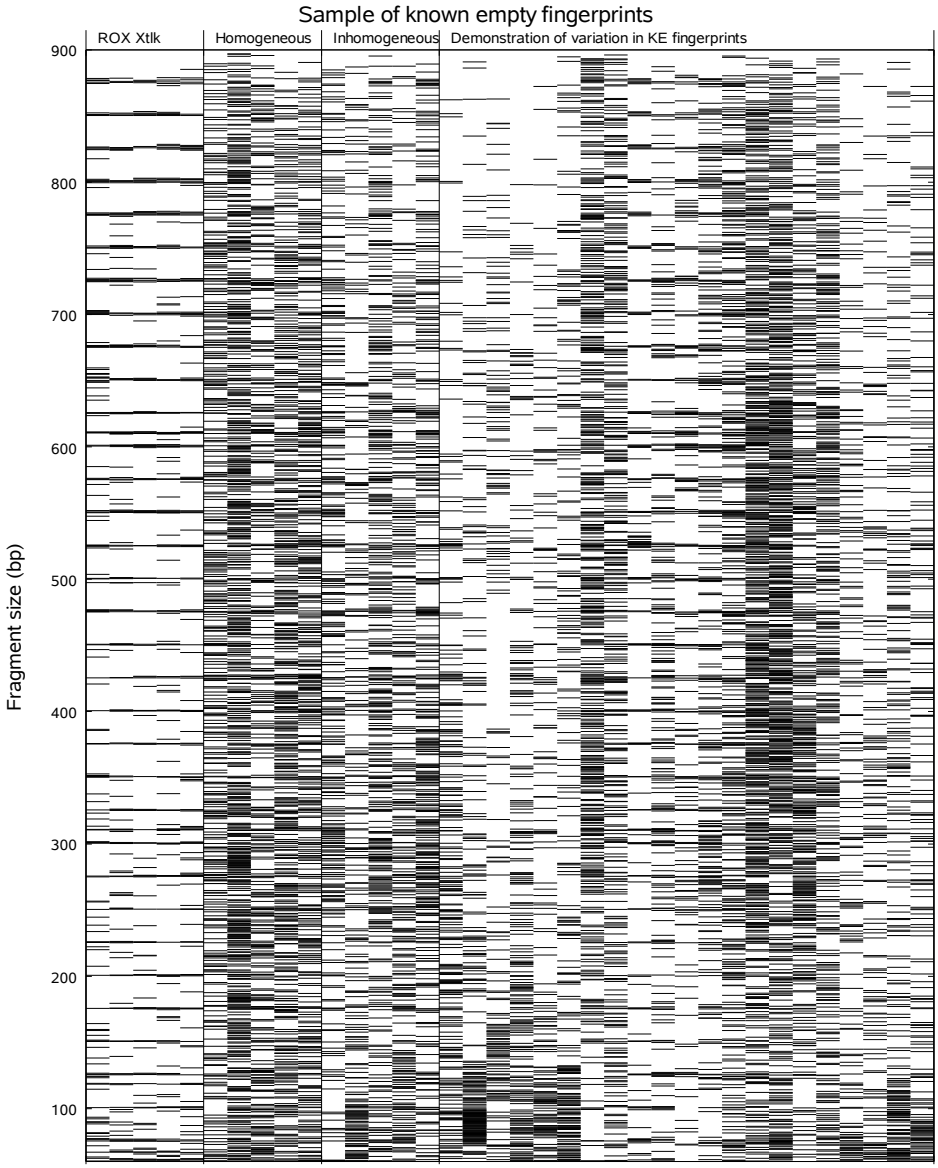


Figure 3: Pseudo–gel image of some "known empty" (KE) fingerprints. At least three visually distinguishable classes exist: Excessive cross-talk ("ROX Xtalk"), homogeneous distribution with many bands ("homogeneous") and inhomogeneous non-AFLP-like distribution ("inhomogeneous"). The right part ("demonstration of variation in KE fingerprints") is a sample of KE fingerprints showing the variation in band distribution encountered. After visually clustering the KE fingerprints (data not shown) many of the fingerprints in this sample are significantly different from any of these clusters in a KS test.

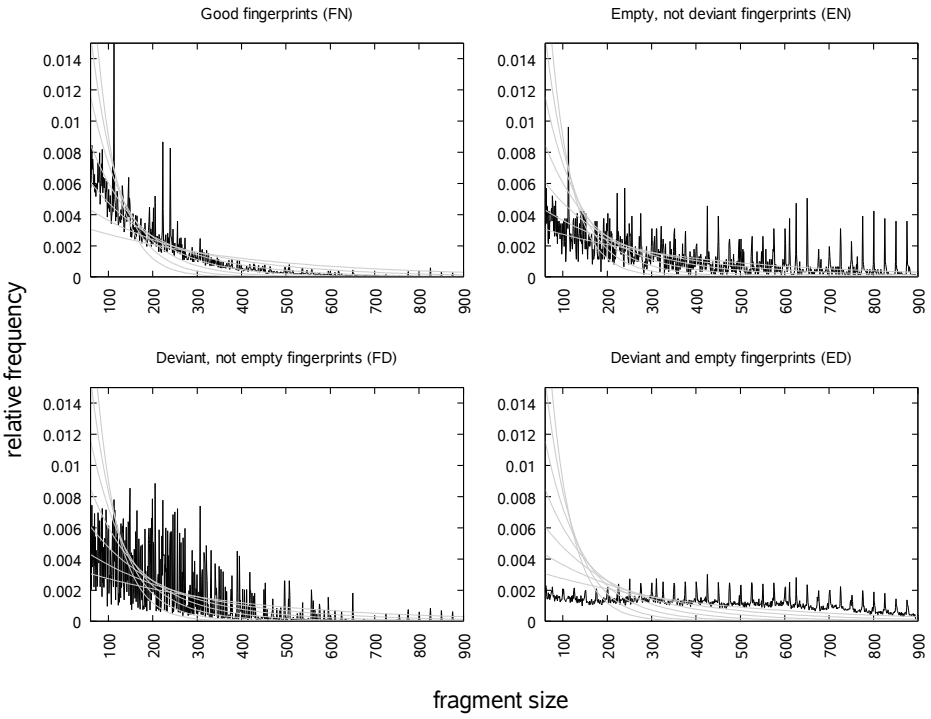


Figure 4: Normalized histograms (area under curve equals 1) of fragment sizes after classification of fingerprints as either Normal or Deviant (N or D) using KS test #2 and as either Empty or Full (E or F) using KS test #1, both as described in the text. The superimposed gray curves are the truncated geometric distributions corresponding to an AT nucleotide content ranging from 45% (flattest) to 75% (steepest) in 5% increments

Analysis of band counts

A wide variation in the number of bands per fingerprint was observed, and Figure 5 shows two normalized histograms of band counts: One for fingerprints classified previously as "FN" ("good", black), and one for all the "other" fingerprints (previously classified as "EN", "FD" or "ED", gray). On average "good" fingerprints have 56.5 bands each, with a standard deviation of 26.1. As shown in Figure 5, the other fingerprints on average have many more bands, in a much wider distribution. Visual inspection of putative "*E.coli* fingerprints" also reveals that many of these have a relatively large number of bands. Assuming a normal distribution of the number of bands per BAC, rejection of BACs with more than 137 bands would theoretically lead to a false rejection rate of 0.1% (75.4 BACs from the combined library), however, at this threshold 1638 (2.2% of 75,419) BACs are rejected from the "good" group versus 1386 (50% of 2757) from the "other" group.

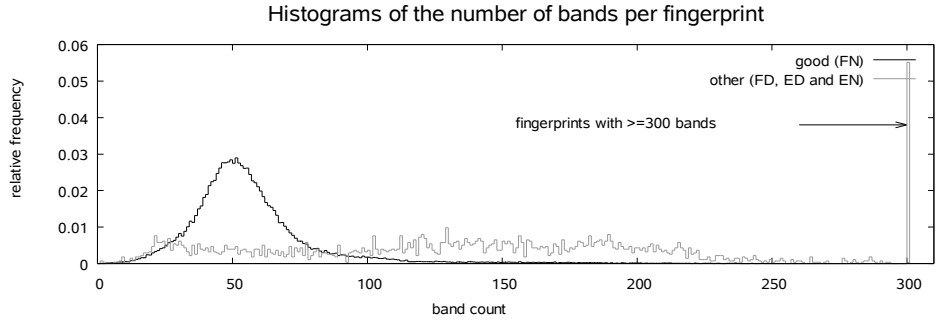


Figure 5: Normalized histograms (area under curve equals 1) of the number of bands per "good" (black) or "other" (gray) fingerprint after classification of fingerprints using KS test #1 and #2 as described in the text.

Fingerprinting reproducibility

Two sets of clones were processed entirely in duplicate, starting from culture initiation through growth, DNA extraction, AFLP, electrophoresis and band calling. The first set of 480 BACs from 5 quarter library plates was chosen randomly (denoted "random"), and the second set were 864 BACs from 9 quarter library plates whose initial fingerprints were visually determined to be of insufficient quality (having few and/or low intensity peaks) (denoted "bad Q"). Figure 6 shows the reproducibility of fingerprinting for both sets of BACs, based on the number of bands detected in each fingerprint, before and after "cleaning" (removal of the fingerprints classified as empty, deviant and/or having an abnormal number (>137) of bands in the previous paragraphs).

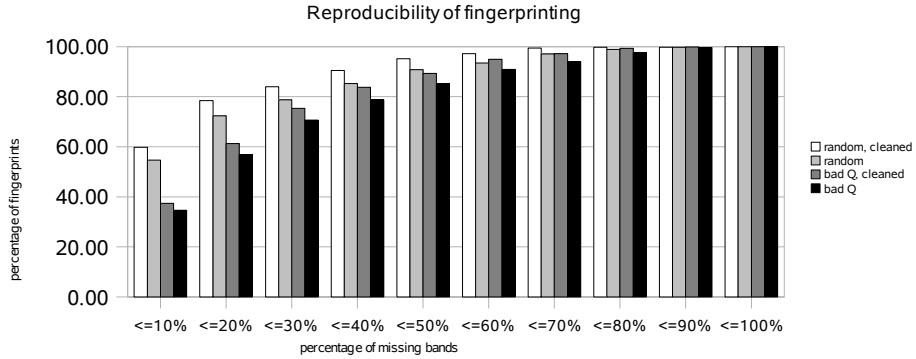


Figure 6: Reproducibility of fingerprinting as deduced from band counts in duplicate fingerprints. For two sets of 480 "random" and 864 "bad quality" BACs respectively, fingerprinting was performed in duplicate. BACs were classified (X-axis) according to the percentage difference in band count. Results are shown before and after ("cleaned") removal of non-AFLP-like fingerprints and fingerprints with an abnormal number of bands. As an example, we read from this figure that in the "random,cleaned" set of fingerprints, 79% of the clones displayed 20% or less difference in band count.

Preliminary physical map construction using FPC

Two preliminary physical maps were constructed using the FPC program, version 6.4 (Soderlund et al. 2000): Before removal of any fingerprints ("uncleaned") and after removal of empty or deviant fingerprints and removal of fingerprints with an abnormal (>137) number of bands ("cleaned"). FPC refers to contigs by a contig-number, and this number and the exact composition of the contigs produced, may vary between runs, making identification of homologous contigs between runs difficult. Where necessary, we have identified homologous contigs in different map versions on the basis of such contigs having a large number of fingerprints in common. We will adhere to the FPC convention of referring to contigs by their contig-ID. Based on their composition, contigs #8773 and #8788 in the "uncleaned" dataset are homologous to contigs #8671 and #8727 respectively in the "cleaned" dataset.

Table 6 summarizes the number of contigs and their size distribution, both before and after execution of the "deQer" (reanalysis of contigs with more than 5 "questionable clones"). The vast majority (70.5%) of the fingerprints removed by cleaning is assembled into a single contig (ctg8788) in the "uncleaned" dataset. Of the removed fingerprints 14.7% would be considered singletons and only 11.3% (0.6% of the total number of fingerprints) would be placed in any other contig. Table 7 summarizes the effect of cleaning on the composition of the two largest contigs, the singletons pool and the remaining contigs. 149 Fingerprints (0.2%) had in excess of 250 or zero bands after band filtering, and these had to be dropped from analysis to accommodate limitations of FPC.

Table 6: Size distribution of contigs obtained from FPC

	Before executing the "deQer"						After executing the "deQer"					
	"uncleaned"			"cleaned"			"uncleaned"			"cleaned"		
	#ctgs	#FPs	#Qs	#ctgs	#FPs	#Qs	#ctgs	#FPs	#Qs	#ctgs	#FPs	#Qs
singletons	-	13510	-	-	12957	-	-	13510	-	-	12957	-
2-3	3820	8915	0	3823	8937	0	3964	9244	0	3935	9190	0
4-10	3429	21195	10	3439	21266	10	3531	21845	10	3521	21784	10
11-30	1087	16958	205	1083	16891	208	1157	18145	171	1141	17902	174
31-100	146	6541	353	145	6500	351	154	6585	128	152	6530	134
101-300	1	118	16	0	0	0	1	110	0	0	0	0
>300	1	10790	4201	1	7190	3132	2	8588	426	2	5378	245
total	8484	78027	4785	8491	73741	3701	8809	78027	734	8751	73741	563

Size distribution of contigs were obtained from analysis of FPC contigs, both before and after execution of the "deQer" (reanalysis of contigs with more than 5 "questionable clones".), and before and after cleaning of the fingerprints as discussed. Prior to contig construction any bands below 100bp or above 650 bp were removed and, after this, clones having either zero or in excess of 250 remaining bands were removed

Table 7: Shift in some contig sizes introduced by cleaning fingerprints.

		"uncleaned"					
		dropped:	singletons:	ctg8773:	ctg8788:	other ctg:	total:
"cleaned"	removed by cleaning:	148	654	28	3126	478	4434
	dropped:	1	0	0	0	0	1
	singletons:	0	12856	0	24	77	12957
	ctg8671:	0	0	2719	2	0	2721
	ctg8727:	0	0	1	2645	11	2657
	other ctg:	0	0	0	43	55363	55406
	total:	149	13510	2748	5840	55929	78176

The time required to compute these maps (on the same computer: 2.0 GHz AMD Athlon, 2GB RAM, NetBSD 3.1) varied considerably: 10.4 and 26.7 hours for the "cleaned" and "uncleaned" dataset respectively. With the settings that were used in this paper, FPC did not appear to be limited by input/output delays or memory shortage. During runs, CPU usage was consistently above 99% and memory requirement consistently below 25%. It was, however, also observed (data not shown) that less stringent parameter settings could lead to an increase of analysis time, and all else being equal, to non-completion because of program crashes caused by memory shortage for the "uncleaned" dataset.

Organellar DNA contamination levels

One pair of (homologous) contigs (ctg8671 and ctg8773 in the "cleaned" and "uncleaned" versions of the FPC map), containing 2721 and 2748 fingerprints of 2600 and 2625 clones respectively, was selected because they exhibited excessive coverage. This excessive coverage suggested an organellar DNA origin of the fingerprints. These contigs appeared relatively well-built. From this pair of contigs, clones from one particular deep well culture plate were selected along with some negative controls and subjected to PCR with three chloroplast specific primer combinations. None of the negative controls was positive with any of the primers, while all of the BACs from the target contigs were positive with at least one primer pair, establishing that this pair of contigs (henceforth denoted "chloroplast contig") contains BAC clones with chloroplast derived inserts. As an illustration of the high fingerprint homology seen within the chloroplast contig, Figure 7 shows, amongst others, a sample of some of the chloroplast fingerprints.

Comparison of BLAST results (chloroplast sequence compared against the available BAC end sequences) with the fingerprint based chloroplast contig assignments resulted in Table 8. A chi-square test shows that classification based on BAC end sequence is highly correlated to classification based on contig-ID (p-value $< 10^{-16}$), and we conclude that

identification of chloroplast DNA derived BAC clones through either method is essentially interchangeable. Based on both fingerprinting results and BAC end sequences the chloroplast contamination level is approximately 3.8%. In 0.09% of the cases BAC end sequences of BACs assigned to the chloroplast contig find no significant match with the chloroplast sequence, and in 0.15% of the cases BACs with chloroplast-like BAC end sequences are not located in the chloroplast contig. It is also interesting to note that many (29%) of the clones whose fingerprints were removed through cleaning also had no BAC end sequence. This set of clones also constitutes 28% of the number of clones without BAC end sequence, strongly suggesting that there is an intrinsic reason why fingerprinting and BAC end sequencing failed for this group.

Table 8: Classification of clones according to BAC end sequence (vertical) and fingerprint contig (horizontal).

	"uncleaned"		"cleaned"				
	CP+	CP-	CP+	CP-	removed	no fingerprint	total
BAC end CP+	2423	192	2404	176	35	212	2827
BAC end CP-	116	66839	111	64141	2703	4642	71597
No BAC end seq.	86	3594	85	2488	1107	232	3912
total	2625	70625	2600	66805	3845	5086	78336

Note that the figures shown are the number of clones, not fingerprints, and that the "total" column shows totals for either the "uncleaned" *or* the "cleaned" columns plus the "no fingerprint" column

The fraction of clones with mitochondrial DNA in our BAC library was estimated from the fraction of Full Plate Pools (FPPs) containing mitochondrial DNA derived inserts using $p=1-\sqrt[N]{1-f}$, where p is the fraction of clones containing a particular insert, f is the observed fraction of pools containing that insert, and N is the number of clones in each pool (384 for FPPs). For both mitochondrial DNA specific primer pairs used, Table 9 shows the number and percentage of positive FPPs (of 96 tested) as well as the percentage of clones we expect to contain mitochondrial DNA derived inserts based on these figures.

Table 9: Mitochondrial DNA screening results.

Primer pair:	atp6	cob	atp6 **or** cob
Positive FPPs (96 tested):	6 (6.3%)	4 (4.2%)	7 (7.3%)
% of clones positive (expected):	0.017%	0.011%	0.020%

96 of the 191 Full Plate Pools (FPPs) were screened using two mitochondrial DNA specific primers. The last column shows the number of pools where either one or both of the primer pairs produced a positive result. We expect that less than 0.1% of the BACs contain mitochondrial DNA.

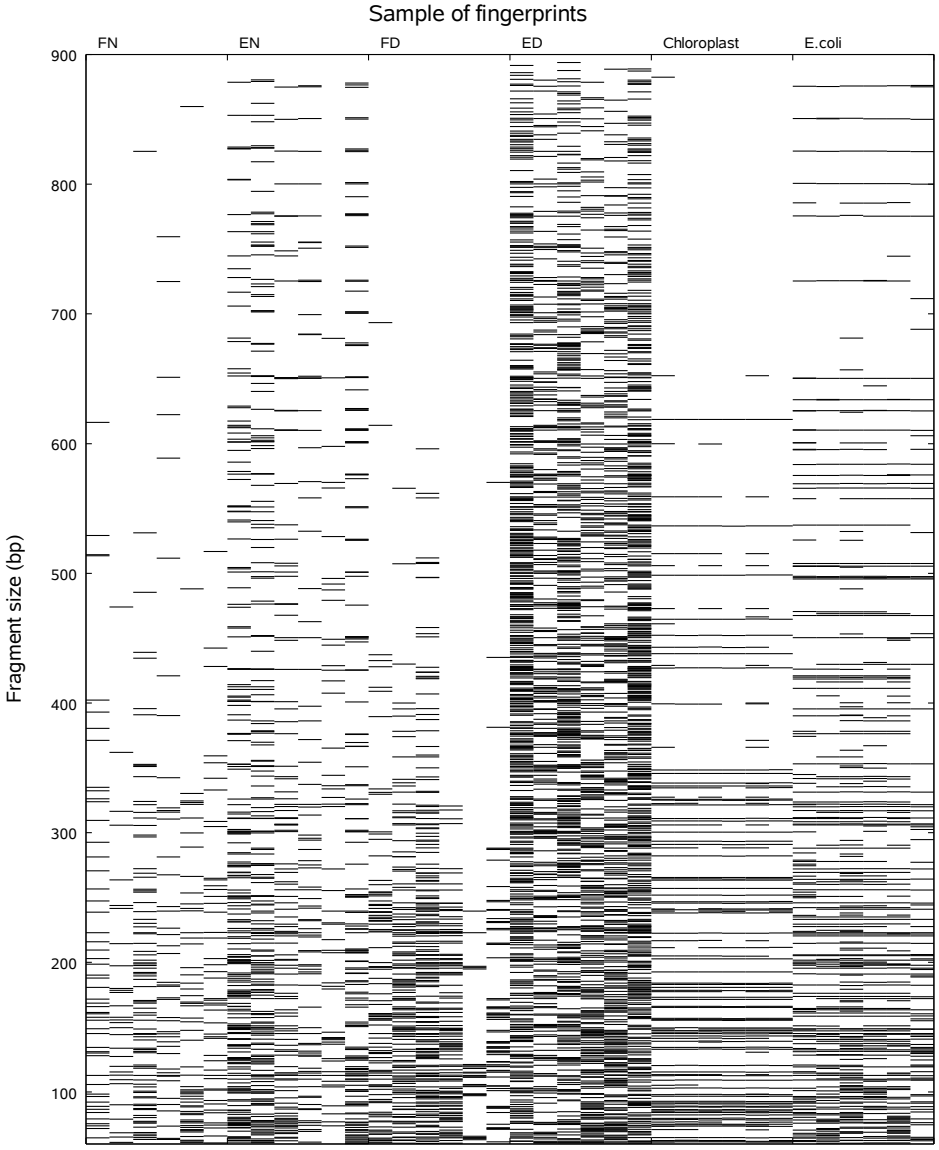


Figure 7: Sample of classified fingerprints. From left to right, 6 fingerprints each: "FN" - Fingerprints that were neither classified as empty nor deviant, "EN" - Fingerprints classified as empty but not deviant, "FD" - Fingerprints classified as deviant but not empty, "ED" - Fingerprints classified as empty and deviant, "Chloroplast" - Fingerprints from the chloroplast contig and "*E.coli*" - Fingerprints exhibiting putative contamination with *E.coli* genomic DNA.

Identification of putative *E.coli* fingerprints

One pair of (homologous) contigs (ctg8727 and ctg8788 in the "cleaned" and "uncleaned" versions of the FPC map), containing 2657 and 5840 fingerprints of 2580 and 5615 clones respectively was selected because they exhibited excessive coverage (compatible with a small percentage of putative *E.coli* fingerprints), and because visual inspection of the fingerprints revealed that many were similar to putative *E.coli* fingerprints encountered in other AFLP based BAC fingerprinting projects at KeyGene, involving different DNA donor organisms (personal communication Taco Jesse, Jan de Boer). This inspection, however, also revealed considerable variation between what were considered *E.coli* fingerprints, and a considerable number of fingerprints (visually estimated between 10%-15% in the "cleaned" version) not appearing to be *E.coli* fingerprints at all. As shown in Table 7, a considerable number of the fingerprints in ctg 8788 is removed through cleaning.

Discussion

BAC library construction

A small difference in average insert size was observed between HindIII (137kbp) and EcoRI (124kbp) BAC clones. We also observed (data not shown) that the cloning efficiencies of the HindIII fractions were always higher than those of the EcoRI fractions. The EcoRI fractions seem to result in significantly fewer BAC-clones per ligation when starting with the same quantity of HMW DNA and pIndigoBAC-536 vector (in estimated ng from gel) and when starting with the same excised PFGE-gel fractions. Reasons for this finding could be a combination of differences in EcoRI and HindIII ligation dynamics and differences in the distribution of EcoRI and HindIII restriction sites over the genome. The differences in ligation efficiency between HindIII and EcoRI were also observed during construction of other plant BAC libraries at KeyGene in the 1999 to 2006 time frame (personal communication Taco Jesse).

High throughput BAC DNA isolation

Approximately 0.56% of the BAC cultures produced a small bacterial pellet or no pellet at all. Of the 1344 clones grown in duplicate, 11 of the 14 clones exhibiting reduced or no growth did so inconsistently, indicating that the majority of reduced (or absent) bacterial pellet sizes are caused by sporadic (~0.44%) inoculation failures. The remainder places an upper limit on the number of empty wells in this BAC library (~0.12%). Long growth (24 hours) in a rich medium (Terrific Broth) does not seem to affect DNA quality and quantity greatly. We have been able to successfully perform PCR, restrictions and ligations on

DNA of individual BACs, on QPPs and FPPs. Long growth (putatively to saturation) may have been the key to the constant pellet size and BAC DNA yield we observed. In all, the BAC DNA isolation protocol used here has proven to be reliable and reproducible.

AFLP Fingerprinting and band calling

Raw electropherograms were processed using KeyGene proprietary software tools (Xpose and BACXtractor). Xpose handles dye signal cross-talk removal and size standard calling, BACXtractor is responsible for band calling and data collection of the actual BAC fingerprints. Dye signal cross-talk removal is necessary because the emission spectra of the different dyes and sensor spectral sensitivity ranges overlap (Hadd et al. 2000; Yin et al. 1996). For sequencing applications, cross-talk removal can be implemented as described by Yin et al. (1996), and this requires that the device is operated within a linear detection range, and that at every base-pair (peak) position there is essentially one and only one labeled DNA fragment present. In genotyping applications the first requirement is met by adjustment of reactions and volumes so that the resulting signal is guaranteed to be well within the linear range of the detector. The second requirement, however, cannot be met; the differently labeled fragments are essentially unrelated, and the corresponding peaks may unpredictably overlap in the electropherograms. Therefore, in genotyping applications, dye signal cross-talk must essentially be treated as a (possibly drifting) machine constant, which must be corrected by application of a pre-determined cross-talk correction matrix. Although we were not provided with details on the precise procedure followed to obtain these matrices, we have inferred from the raw electropherograms produced by the MegaBACE machine (".rsd files") and cross-talk correction matrices supplied to us that they are (re-)determined at (irregular) intervals. Despite the evident effort we find that a large number of fingerprints bear evidence of residual cross-talk at the positions of the larger (>650 bp) et-ROX size standard fragments, both in the band files (as clearly shown in the histograms in Figures 2 and 4) and in cross-talk corrected electropherograms (data not shown), suggesting that the procedure followed to obtain these cross-talk correction matrices was sub-optimal. We do not know if crosstalk from smaller et-ROX size standard fragments is truly less frequent or only not visible in these histograms because smaller true AFLP fragments simply occur more frequently. Residual cross-talk from other dyes can not easily be detected in these histograms because such cross-talk will not alter the normal AFLP fragment distribution.

Problems with residual cross-talk may also in part explain the relatively poor reproducibility in band numbers: 19% of the randomly duplicated fingerprints has a >20% difference in band count (Figure 6). Nelson et al. (2005), observed an effect of similar magnitude (only 75% of the bands were shared between repeated fingerprints of the same

BAC clones) in High Information Content Fingerprints (HICF) obtained from a maize BAC library, and attributed this to band-calling problems. It has been observed that an overall lower signal strength in the raw electropherograms resulted in lower detection thresholds being used in BACXtractor, which in turn led to detection of (minor) residual cross-talk peaks or even to detection of peaks in random detector noise. It has also been observed that seemingly comparable electropherograms resulted in different threshold settings. The main controls offered by the BACXtractor software for band detection threshold setting are called “signal rise” and “scale level” (Srinivasan et al. 2003, both set at 20%), both of which are a percentage of some measure of the overall signal intensity in a particular trace. Using relative thresholds for each separate trace allows one to effectively compensate for considerable variation in signal intensity caused by variations in the underlying processes. Relative thresholds are, however, not without danger. Besides an apparent absence of a lower, absolute, bound on the threshold (leading to detection of bands in what is essentially detector noise), the measure of overall signal intensity used may in itself be flawed, leading to occasionally inappropriate threshold settings. In the BACXtractor software the measure appears to be based on an average of some percentile of higher signal values, which may or may not include (part of) the signal peak caused by residual unincorporated primers, and which may be influenced by the total number of DNA fragments in a trace. The lack of insight into and lack of control over this important aspect of band calling may eventually prompt us to re-process the raw data using different sizing and band calling software. In view of the occasionally high levels of residual cross-talk observed in the cross-talk corrected electropherograms (putatively caused by incorrect cross-talk correction matrices), such an exercise may be of limited practical value.

Statistical analysis of fingerprints

The statistical analysis of individual fingerprint patterns was designed to detect and remove deviant (in particular non-AFLP-like) fingerprints from the dataset. This filtering is performed based on similarity to fingerprints that were known a priori to be non-AFLP-like, based on dissimilarity to theoretical AFLP fragment length distributions and based on an unexpectedly large number of bands being present in the fingerprint. Different types of failure in the fingerprinting process can result in similar effects observable in fingerprints (Table 2). Multiple types of failure can occur simultaneously in the same fingerprint and a failure may not be complete. For example a BAC fingerprint may have a few et-ROX size standard cross-talk bands in addition to it's own AFLP fingerprint bands. Because we lack models describing the fragment size distributions expected for the different types of failures (For “Known Empty” fingerprints the available data suggests that there is a mixture of multiple underlying distributions), and because partial fingerprinting failures

and/or simultaneously occurring failure modes leads us to expect a continuous range of mixtures of distributions, we have not attempted to classify fingerprinting failures in any detail, and we only remove the most affected fingerprints from the FPC analysis. Originally we tried to use a fixed value for the AT nucleotide fraction in the theoretical model of AFLP fragment length distribution. This, however, required us to use a low p-value threshold for fingerprint rejection to prevent false rejection of fingerprints apparently derived from genomic regions with locally different nucleotide composition, and this in turn led to false acceptance of some other fingerprints. The use of a range of values for the AT nucleotide fraction renders the filtering process less sensitive to naturally occurring local variations in nucleotide composition, in essence rejecting only those fingerprints whose band distribution does not fit any reasonable truncated geometric distribution.

Preliminary physical map construction

If not removed, the majority (88.7%) of rejected fingerprints would end up in either the singletons pool or in one particularly large contig. Therefore it can be argued that prior removal is unnecessary. However, on close inspection, the large contig appears to consists of some normal fingerprints entangled in a multitude of bad quality fingerprints, and it is expected that at least some of the remaining (11.3%) rejected fingerprints will also give rise to entanglement and overall quality deterioration in other contigs.

Contig construction by FPC can conceptually be divided in two phases: First FPC compares every fingerprint to every other fingerprint and divides the fingerprints into groups according to their pairwise scores, then FPC tries to determine the most likely order of fingerprints within a group (contig). Assuming that pairwise fingerprint comparison is the most computationally expensive phase during map construction, one would expect the time it takes to construct a map to scale with $\mathbf{O}(N^2)$ for maps containing N fingerprints. Comparison of the time needed to compute two versions of the map, one "uncleaned" containing 78,027 fingerprints and one "cleaned" containing 73,741 fingerprints reveals a 2.5 fold difference in time – much more than the ~1.1 fold difference expected based on the number of fingerprints alone. As discussed by Flibotte et al. (2004) the problem of ordering of clones within a contig resembles the well-studied Traveling Salesman Problem (TSP), which is NP-hard and without solution in polynomial time ($\mathbf{O}(N^K)$). Therefore an approximation algorithm must be used, some versions of which have been shown to be NP-complete (Golumbic et al. 1994). The clone ordering algorithms are applied to individual contigs, and we assume that it is the different number of clones contained in contigs ctg8727 (of the "cleaned" dataset, containing 2656 fingerprints) and ctg8788 ("uncleaned", 5840 fingerprints) that is responsible for most of

the required extra computational expense. Given that this contig (ctg8727/ctg8788) does not appear to be a valid assembly of fingerprints derived from potato genomic DNA, but rather a collection of fingerprints bearing evidence of putative *E.coli* contamination and non-related fingerprints sharing a putatively significant number of bands either due to artifacts (e.g. Cross-talk) or entirely coincidentally, it may even be sensible to remove it altogether.

The statistical models used by FPC assume that all fragment sizes in a fingerprint occur equally frequently. In theory and in practice, however, the AFLP process produces a highly skewed distribution. The fact that fingerprint fragment distributions are generally not constant was recognized by Sulston et al. (1988), and it was suggested that in extreme cases the score looses its intrinsic meaning (as a probability). More recently Wendl (2005) demonstrated that, even if all fragments are equally probable, the conventional scores dramatically overestimate the probability of coincidental fingerprint overlap. Unfortunately Wendl (2005) also demonstrates that proper calculation of probabilities becomes computationally prohibitively expensive for fingerprints with more than a few bands. To our knowledge no model incorporating fragment size dependent a-priori fragment probabilities has yet been published, but a preliminary investigation on our part indicates that such a model would also be computationally prohibitively expensive. There is, in our opinion, with our dataset, no sensible way to interpret the scores obtained by FPC as probabilities, or even to sensibly relate them to each other, other than by observing that they meet some threshold criterion. As the gel length appears as a constant in the equations, we left FPC's setting for the gel length at it's default (3300 bp) value, and have only adjusted the value for cut-off to meet our requirements.

Besides smaller fragments, fragments larger than 650 bp were removed from analysis because, based on visual inspection of the fingerprints and evident peaks in the histogram, we believe that a large proportion of these larger fragments are caused by cross-talk from the et-ROX size standard. It should also be noted that physical maps were computed only for characterization of the BAC library, in particular the identification of clones containing chloroplast derived inserts and putative *E.coli* genomic DNA contaminated fingerprints. We assume that the relatively large percentage (20%) of singletons, and the large number of disjoint contigs (>8500) we observed is at least partially caused by the use of stringent parameter settings.

In the AFLP fingerprints, approximately one quarter of the AFLP fragments is between 60bp and 100bp in length, and inclusion of these bands in the computations (using the same cut-off) leads to excessive memory and CPU use and eventually crashes FPC. We assume that this is caused by excessively large groups of fingerprints being created based on essentially non-related fingerprints sharing a seemingly significant number of bands

because of a flawed statistical model. Given the skewness of the AFLP fragment size distribution and the problems this causes, we expect that we will need to carefully optimize FPC tolerance and cut-off settings in combination with our threshold for small fragment removal to arrive at a physical map more suitable for other purposes.

Organellar DNA contamination levels

Chloroplast-DNA derived clones were identified based on their inclusion in single a large contig (ctg8671 and ctg8773 respectively in the "cleaned" and "uncleaned" versions of the preliminary FPC map), and a sample of these clones was verified to be chloroplast-derived using PCR. Later, when the BAC end sequences became available, the identity of the chloroplast contig was verified. The match between the sets of clones identified as chloroplast derived by either inclusion in the chloroplast contig or by BAC end sequences is near perfect, and therefore further screening, for instance using filter hybridization is considered redundant. The percentage of BAC clones (3.8%) containing chloroplast DNA derived inserts is within the range reported by other authors (e.g. Lin et al. 2006, McGrath et al. 2004, Yim et al. 2002, Budiman et al. 2000). Pools constructed from the fingerprinted part BAC library were screened using two sets of primers targeting mitochondrial DNA based on the sequence of two genes: Mitochondrial ATPase subunit 6 (atp6) and Mitochondrial apo-cytochrome b (cob) (Sugiyama et al. 2005). A BLAST search was performed in genbank to predict the mitochondrial DNA specificity of the primer-pairs in silico, and only one significant match to one of the primers outside mitochondrial DNA was found, making it likely that these primers will only amplify truly mitochondrion derived DNA. As incorporation of mitochondrial DNA into the genome has been reported for other organisms (e.g. Stupar et al. 2001), this assay only gives us an estimated upper bound of the number of mitochondrial DNA derived inserts in our BAC libraries: 0.02%. This figure is comparable to the level of mitochondrial DNA contamination reported for other BAC libraries (e.g. McGrath et al. 2004, d'Alençon et al. 2004, Guimarães 2008, Ratnayaka 2005). The effect on genome coverage of any mitochondrial DNA derived BAC clones is, however, negligible.

Insert sizes, empty vector clones and genome coverage

Of 590 clones for which insert size was determined, 18 (3%) were found to be empty vector clones, and the remaining 572 had an average insert size of 131 kbp. Extrapolating this to the entire library, we expect approximately 2400 empty vector clones and 11.7 times coverage of the potato genome, which is a slightly higher estimate than would be obtained if the empty vector clones are accounted for in the average insert size. After fingerprinting 191 of the 204 384-well microtitre plates (94%), 5.7% of the fingerprints

were rejected, 3.8% of the clones contained chloroplast DNA derived inserts and 3.4% of the fingerprints were located in a contig (ctg8727/ctg8788) containing predominantly bad quality and putative *E.coli* fingerprints. As some of the fingerprints entangled in this contig may be recoverable, we estimate the total coverage of the haploid potato genome by usable fingerprints to be between 9.6 and 10.2 times. As the diploid potato clone used as a source organism for this BAC library is highly heterozygous genome, the probability of encountering any particular heterozygous DNA fragment may be considerably lower, and in this context it may be more appropriate to estimate coverage as between 4.8 and 5.2 times the diploid potato genome.

Putative *E.coli* fingerprints

Use of AFLP for BAC fingerprinting involves an intrinsic risk not present in other BAC fingerprinting methods: The AFLP reaction will amplify any template present in the reaction, and in absence of a sufficient quantity of BAC DNA to use as AFLP template (for instance caused by DNA isolation failures), the template may predominantly be *E.coli* genomic DNA that was accidentally co-precipitated. Despite the fact that the *E.coli* genome is much longer than BAC inserts, the dynamics of the AFLP reaction, as discussed by Han et al. (1999) will result in a pattern of bands rather than a continuous smear. In such cases, because the banding pattern may essentially be caused by subtle differences in template amplification efficiencies, the exact fingerprint pattern may be much more variable than normal fingerprints due to variations in chemical composition (e.g. salts) and PCR temperature profiles. Though we have no conclusive evidence, we assume that fingerprints in one particular contig (ctg8727 and ctg8788 in the “cleaned” and “uncleaned” versions respectively of the preliminary physical map) are predominantly *E.coli* genomic DNA derived. This assumption is based on a combination of different observations: Firstly that many of these fingerprints exhibit relatively low signal intensities (putatively caused by low template concentration). Secondly that the fingerprints are more complex (more bands) and appear more variable within a single (high coverage) stack than other fingerprints (e.g. fingerprints in the chloroplast contig). Thirdly, and most significantly, that highly similar fingerprints occurred in other, totally unrelated (non-potato) AFLP fingerprinted BAC libraries (personal communication Taco Jesse). Possibly due to their variability, but possibly also because some fingerprints are the result of a mixture of BAC and *E.coli* DNA, the contig also seems to capture, or entangle, a fair number of other fingerprints (visually estimated 10-15%). Without application of the “deQ-er” the contig even entangles all clones from the chloroplast contig.

Conclusion

We have constructed, using two restriction enzymes, a large insert (~131 kb) BAC library of potato. Less than 4% of the clones contain chloroplast DNA derived insert, approximately 3% are expected to be empty vector clones while the contamination with mitochondrial DNA was found to be negligible. Individual BACs from 191 of the 204 384-well micro-titre plates were fingerprinted using an AFLP fingerprinting method, resulting in an approximately 10-fold coverage by fingerprints of the haploid potato genome. We demonstrate that fingerprint data, even without spending much effort on optimizing physical maps, can be used to identify clones containing chloroplast DNA derived inserts and putative contamination with *E.coli* genomic DNA, and can be used to exercise quality control on the fingerprints.

Chapter 3

A Universal Maximum Likelihood Pairwise Linkage Estimator

T.J.A.Borm, J. de Boer, H.J. van Eck and R.G.F. Visser

Abstract

Linkage analysis in full-sib families descending from non-inbred parents is generally considered more complicated than linkage analysis in mapping progenies derived from inbred or homozygous parents. Markers can segregate from one or both parents, up to four alleles may segregate and markers can be linked in cis- or in trans- phase on both parental homologous chromosomes independently. Multiple alleles partaking in a single individuals' marker score may also lead to another complication. Sometimes different alleles may not be distinguishable in every individual in a progeny (e.g. short repeat SSR markers with overlapping peak position distributions or AFLP markers co-dominantly scored based on band intensities with overlapping band intensity distributions). This leads to marker score ambiguities and a mixture of Mendelian segregation types within a marker score. To our knowledge no generalized framework for dealing with marker scoring ambiguities or mixed segregation types exists. In this paper we introduce a suitable marker scoring system, and derive, from first principles, a Universal Maximum Likelihood Pairwise Linkage Estimator (UMLPLE). This UMLPLE can simultaneously deal with any combination of (mixed) marker segregation types and all types of marker scoring ambiguity. We mathematically prove that relative linkage phase of a pair of markers can be obtained immediately from the UMLPLE, and provide a reference implementation written in the PERL programming language.

Introduction

Linkage analysis in progenies derived from pure, inbred, lines is generally considered easier than linkage analysis in full sib families derived from heterozygous parents, because, in the former case, only two alleles per locus need to be considered and determining linkage phase is trivial. Some agronomically important crop species, however, either have long generation times precluding the development of homozygous lines (e.g. trees), or suffer from severe inbreeding depression (Uitewaal et al. 1987), and for such species F1 mapping populations from non-inbred parents (also referred to as CP-populations) have been used to construct linkage maps. Linkage analysis in such CP populations needs to consider that for each locus multiple alleles from one or both parents may segregate in an offspring, and that these alleles may not be distinguishable between parents. On the one hand most of these complexities can be circumvented by separate analysis of parental gametes as proposed by Grattapaglia and Sederoff (1994), who consider the CP mapping population as a pseudo double testcross, and on the other hand to handle CP populations linkage analysis software has been developed using the concept of segregation types (Stam 1995): each marker has an attribute that specifies from which parent and how it will segregate in an offspring.

Maximum likelihood linkage estimators were initially developed by the founders of classical genetics (e.g. Mather 1938, Allard 1956). For use in full sib families additional maximum likelihood estimators were developed for specific combinations of segregation types (Ritter et al. 1990), with all possible combinations covered by Maliepaard et al. (1997). Wu et al. finally (2002a) extended these maximum likelihood estimators to deal with sex specific recombination frequencies.

Although between them these ML linkage estimators cover all combinations of marker segregation types, they cannot deal with the effects of ambiguous marker scores. For instance, for some SSR markers, process variation and resulting fragment sizing variation (Vemireddy et al. 2007; Amos et al. 2007) may result in overlapping allele fragment size distributions. Use of SSRs with small (e.g. single nucleotide) difference in allele sizes aggravates this problem (van de Wiel, personal communication), making proper binning of the SSR alleles occasionally impossible. Similarly, problems with secondary SSR peaks (Kirov et al. 2000) and SSR allele-dosage determination (Chatet al. 2003, Reid and Kerr 2006, Landergott et al. 2006) may also result in mis-classifications because of overlapping distributions. Likewise, when the band intensity of AFLP markers is used to co-dominantly score these markers, the intensity distributions for null, single and double dose phenotypes may overlap (Jansen et al. 2001), again giving rise to ambiguity. Genotyping errors, which would be introduced by accepting mis-classifications, could largely be

prevented by allowing ambiguous scores in regions where distributions overlap. Although an individual scored with an ambiguous marker score is less informative, it does not represent mis-information, and is more informative than a missing score.

Besides these (and other) examples where ambiguities are intrinsic to the marker technology, experimental failures may also lead to ambiguity. Consider for example a PCR marker that segregates from both parents in an offspring (n.b. <a0 x a0> segregation type), with a restriction enzyme (CAPS) recognition site in one of the parental alleles (effectively resulting in a <a0 x b0> segregation type). Failure to add the restriction enzyme in part of the progeny would still result in a marker segregating from both parents, but to properly map it, mapping software must either deal with ambiguous marker scores (equivalent to a mixture of segregation types), or part of the scoring data must be discarded.

It is evident that different types of markers and different types of failures may require dealing, simultaneously, with several different types of ambiguity. This may explain why the capability of dealing with ambiguous marker scores, to our knowledge, is largely missing from mapping software. Joinmap (Stam 1993, 1995), for instance, only allows "h-" and "k-" (equivalent to allele dominance) scores to be used for <hk x hk> (e.g. co-dominantly scored AFLP markers), and no other types of ambiguity whatsoever for any other segregation type.

Another issue associated with linkage analysis in CP populations is the problem of determining linkage phase. Two genetically closely linked markers may either be located on the same (in cis-phase) or on different homologous chromosomes (in trans-phase) in each parent independently, and the segregation patterns of the markers will vary accordingly. To accurately determine linkage, the linkage phase of the markers must either be known a priori, or it must be possible to determine linkage phase from the marker data itself. For inbred line based mapping populations it has long been known that if the ML linkage estimate r, assuming linkage in cis-phase, exceeds 0.5, then it is more likely that that the markers are linked in trans-phase at $r^*=(1-r)$. In full-sib mapping populations there are 4 different combinations of relative linkage phase: markers may be linked in cis-phase (1) or trans-phase (2) in both parents and markers may be linked in cis-phase in one parent and in trans-phase in the other (3) and vice versa (4). Between pairs of these relative linkage phases a similar mathematical relation exists: (1) and (2) are related as are (3) and (4). For phase unknown markers various statistical methods have been employed to determine linkage phase (Ritter and Salamini 1996, Wu et al. 2002b, Axenovich 1996, Cartwright 2007a), and all these methods represent a significant computational overhead.

In this paper a single LOD-score equation and maximum likelihood linkage estimator will be presented that can simultaneously deal with all possible combinations of marker

segregation types and all types of ambiguous marker scores. Because of it's generality we call this the Universal Maximum Likelihood Pairwise Linkage Estimator (UMLPLE). To effectively work with the UMLPLE requires a marker scoring system that is equally universal: It must be able to represent, simultaneously, arbitrary segregation types and arbitrary ambiguities. As, to our knowledge, no suitable scoring system was previously available, we propose a new, comprehensive, comprehensible and concise marker scoring system. Furthermore, we will prove that our UMLPLE has another useful property: it makes linkage phase detection automatic. We provide a reference implementation written in PERL (http://theo.borm.org/pbsw/), and show some results of a simulation study obtained using this program.

Methods

General outline

We will start by establishing our assumptions and basic terminology and then proceed to define our universal marker scoring system. After this we will obtain our Universal Maximum Likelihood Pairwise Linkage Estimator (UMLPLE) in several steps:

1. We establish that in order to calculate the odds or LOD (Log of Odds) score for a pairwise marker observation in a mapping population as a function of parental recombination frequencies it suffices to calculate the odds for each individual in the population separately and combine results afterwards (eq. 1).
2. We obtain tables, assuming a pair of markers allowing full classification (equivalent to a <ab x cd> segregation type) observed in a single individual, giving the component probabilities of the odds equation for every possible marker score combination (eq. 2).
3. We observe that probabilities for ambiguous scores can be obtained by summing the probabilities of the corresponding (underlying) fully classified probabilities, and note that such summation can be implemented conveniently using vector algebra (eq. 3).
4. Noting that several pairwise combinations of marker scores (both ambiguous and non-ambiguous) result in identical individual odds equations, we arrive at a table associating any pairwise marker score with one of 25 individual odds equations (Table 2). By combining this information with equation 1, we arrive at a universal LOD equation (eq. 4).
5. Taking partial derivatives of the LOD equation, which should equal zero at a maximum, we obtain a system of two equations in two unknows (eq. 5 and 6). Solving this system analytically in a few cases (eq. 7 through 17), and numerically

otherwise (eq. 18) we arrive at maximum likelihood linkage estimates: the UMLPLE.

We conclude the mathematical section by proving (mathematically) that linkage phase determination using the UMLPLE is essentially automatic.

Assumptions and basic terminology

We assume that:

1. Meiotic recombination events in parents of a mapping population are statistically independent events.
2. The offspring in a full-sib mapping population are statistically independent samples originating from random gametes
3. There is no gametic nor zygotic selection

We use the following terminology:

1. Parents of the mapping population are called P and Q.
2. Recombination frequencies are r_p and r_q for parent P and Q respectively.
3. Non-recombination frequencies are $n_p=(1-r_p)$ and $nr=(1-r_q)$ respectively.
4. Two markers (in a linkage group) are called M_1 and M_2.
5. The parental alleles of P and Q at a particular locus are denoted 1 and 2 (for P) and 3 and 4 (for Q), regardless if there is an identifiable difference at that locus or not.
6. The four possible Parental Allele Combinations (PACs) at a particular locus in the offspring are denoted [13], [14], [23] and [24], which can be used as a superscript to the marker names (e.g. $M_1^{[13]}$ specifies PAC [13] for marker M_1).

Marker scoring system

Any observed marker phenotype is the result of a particular marker technology and a particular PAC. If a marker allows full genetic classification of the offspring (with segregation type <ab x cd>) there is a one-to-one relationship between phenotypes (<ac>, <bc>, <ad> and <bd>) and the underlying PACs ([13], [14], [23] and [24]). For other marker segregation types and for ambiguous marker scores the same marker phenotype may be caused by more than one particular PAC. As there are 4 different PACs, we expect that there are 24 = 16 different possible combinations of between zero and four different PACs necessary to explain every possible marker phenotype. One of these combinations, however, is invalid. An observed marker phenotype can simply not be explained by the absence of all PACs as at any locus; at least one PAC is always present. The opposite, where any of the four PACs can explain the observed marker phenotype, is equivalent to a missing observation. Such a “marker phenotype” can not distinguish between any of the underlying PACs. Therefore we arrive at the conclusion that only 15 different scores are

sufficient to specify any possible marker score, including ambiguous scores. Hence we arrive at the PAC-centric marker scoring system shown in Table 1: either as a 4 element binary vector, with each element corresponding to a particular PAC ("M vector"), or as list of combined parental alleles of a hypothetical <AB x CD> marker. It is evident that this scoring method makes explicit specification of of marker segregation type unnecessary.

Table 1: Overview of the syntax/data format of the PAC based scoring system.

segregation type	**phenotype**	**explaining PACs**	**M vector**	**marker score**	**shorthand**
None		None	[0,0,0,0]	Invalid	Invalid
None; missing value		[13] or [23] or [14] or [24]	[1,1,1,1]	AC|BC|AD|BD	--
<ab x aa>	aa	[13] or [14]	[1,0,1,0]	AC|AD	A-
	ab	[23] or [24]	[0,1,0,1]	BC|BD	B-
<aa x ab>	aa	[13] or [23]	[1,1,0,0]	AC|BC	C-
	ab	[14] or [24]	[0,0,1,1]	AD|BD	D-
<ab x ab>	aa	[13]	[1,0,0,0]	AC	AC
	ab	[23] or [14]	[0,1,1,0]	BC|AD	BC|AD
	bb	[24]	[0,0,0,1]	BD	BD
<ab x cd>	ac	[13]	[1,0,0,0]	AC	AC
	bc	[23]	[0,1,0,0]	BC	BC
	ad	[14]	[0,0,1,0]	AD	AD
	bd	[24]	[0,0,0,1]	BD	BD
<a0 x a0>	00	[13]	[1,0,0,0]	AC	AC
	a-	[23] or [14] or [24]	[0,1,1,1]	BC|AD|BD	!AC
<ab x a0>	a-	[13] or [14]	[1,0,1,0]	AC|AD	A-
	b0	[24]	[0,0,0,1]	BD	BD
	ab	[23]	[0,1,0,0]	BC	BC
<a0 x ab>	a-	[13] or [23]	[1,1,0,0]	AC|BC	C-
	b0	[24]	[0,0,0,1]	BD	BD
	ab	[14]	[0,0,1,0]	AD	AD
See note		[13] or [24]	[1,0,0,1]	AC|BD	AC|BD
		[13] or [14] or [24]	[1,0,1,1]	AC|AD|BD	!BC
		[13] or [23] or [24]	[1,1,0,1]	AC|BC|BD	!AD
		[13] or [23] or [14]	[1,1,1,0]	AC|BC|AD	!BD

In this table, all 15 combinations of PACs explaining any possible marker phenotype, the resulting PAC-based marker scores and their relationship with the core 7 segregation types (Maliepaard et al. 1997). are shown. For marker scores three alternative notations are given in the last three columns: as a binary M vector, with each binary digit representing a particular PAC, as a list of explanatory PACs separated by a vertical bar or in shorthand notation. The last 4 PAC-based scores correspond to variants of other scores expressed in a different linkage phase; they represent no score that should ever be needed for markers adhering to a definite segregation type, however they can occur as ambiguous scores or in mixed segregation type markers (e.g. in synthetic bridge bin signatures as used in chapter 4 of this thesis). As discussed in the text the combination of no PACs as expressed by the M vector [0,0,0,0] is invalid.

Table 1 also specifies the relationship between our PAC based scores and the traditional segregation type based scoring system as used by Maliepaard et al. (1997). While often called a "null allele", in the traditional scoring system the "0" allele is actually used to signify recessive inheritance and dominance of the "a" allele. The <a0xa0> segregation type markers will be scored as a series of "00" and "a-" (to indicate the unknown single or double dosage of allele "a") scores and not as a series of "00", "0a" and "aa" scores. To score markers where allele dosage can be determined the <abxab> segregation type is actually abused.

Derivation of a universal LOD equation

Using the separate parental recombination frequencies (r_p and r_q) between a pair of markers, we can write the equation for the odds (ratio of probabilities under two assumptions: Linkage and random assortment) of a set of observations on N offspring as follows:

$$odds_{total}(r_p, r_q) = \frac{p_{linked}}{p_{unlinked}} = \frac{\prod_{1}^{N} p_l^i}{\prod_{1}^{N} p_u^i} = \prod_{1}^{N} \frac{p_l^i}{p_u^i} = \prod_{1}^{N} odds_c^i(r_p, r_q), \qquad \text{eq. 1}$$

with $odds_{total}(r_p,r_q)$ being the total odds of the observation, p_l^i and p_u^i represent the likelihoods of observing a given marker phenotype in the i^{th} offspring individual assuming linkage and non-linkage respectively, and $odds_c^i(r_p,r_q)$ the contribution to $odds_{total}(r_p,r_q)$ of the i^{th} offspring individual. Therefore, we only need to derive equations for "component" $odds_c^i(r_p,r_q)$ for a single offspring individual, and can accommodate ambiguous marker scores by using different $odds_c$ equations for different offspring individuals as applicable.

From first principles we can trivially derive tables giving the probabilities of observing (in a single offspring) any of the 16 (4*4) possible PACs for a pair of markers assuming linkage or random assortment, and write these probabilities as a matrices H (linkage) and J (random assortment) respectively:

$$H = \begin{array}{c} \\ M_2^{[13]} \\ M_2^{[23]} \\ M_2^{[14]} \\ M_2^{[24]} \end{array} \begin{array}{c} \begin{array}{cccc} M_1^{[13]} & M_1^{[23]} & M_1^{[14]} & M_1^{[24]} \end{array} \\ \begin{bmatrix} n_p n_q/4 & r_p n_q/4 & n_p r_q/4 & r_p r_q/4 \\ r_p n_q/4 & n_p n_q/4 & r_p r_q/4 & n_p r_q/4 \\ n_p r_q/4 & r_p r_q/4 & n_p n_q/4 & r_p n_q/4 \\ r_p r_q/4 & n_p r_q/4 & r_p n_q/4 & n_p n_q/4 \end{bmatrix} \end{array}, \quad J = \begin{array}{c} \\ M_2^{[13]} \\ M_2^{[23]} \\ M_2^{[14]} \\ M_2^{[24]} \end{array} \begin{array}{c} \begin{array}{cccc} M_1^{[13]} & M_1^{[23]} & M_1^{[14]} & M_1^{[24]} \end{array} \\ \begin{bmatrix} 1/16 & 1/16 & 1/16 & 1/16 \\ 1/16 & 1/16 & 1/16 & 1/16 \\ 1/16 & 1/16 & 1/16 & 1/16 \\ 1/16 & 1/16 & 1/16 & 1/16 \end{bmatrix} \end{array} \qquad \text{eq. 2}$$

Probabilities in matrices J and H in equation 2 represent mutually exclusive compound statistical events (encountering a particular combination of PACs at a pair of marker loci), therefore we can calculate the probability of encountering a member of a subset of these compound statistical events by summation of the probabilities over the components of the

subset (combinations of PACs). This summation can be accomplished conveniently using vector algebra (with T signifying a matrix or vector transpose):

$$odds_c(p_r, q_r) = \frac{(M_1 \times H) \times M_2^T}{(M_1 \times J) \times M_2^T}, \qquad \text{eq. 3}$$

with M_1=**[**$M_1^{[13]}$,$M_1^{[23]}$,$M_1^{[14]}$,$M_1^{[24]}$**]** and M_2=**[**$M_2^{[13]}$,$M_2^{[23]}$,$M_2^{[14]}$,$M_2^{[24]}$**]** marker scores expressed as binary M-vectors as shown in Table 1. As there are 15 possible marker scores at a single locus (Table 1), there are 15 x 15 = 225 possible combinations to describe the joint observation of a pair of marker scores in one offspring individual. These 225 combinations, however, result in only 25 unique $odds_c$ equations, indexed by the letters A-Y in Table 2.

Table 2: Overview of all possible observations at a pair of markers in one offspring individual.

			marker 1 score														
		M_1	[1,0,0,0]	[0,0,1,0]	[0,1,0,0]	[0,0,0,1]	[1,0,1,0]	[0,1,0,1]	[1,1,0,0]	[0,0,1,1]	[1,0,0,1]	[0,1,1,0]	[0,1,1,1]	[1,0,1,1]	[1,1,0,1]	[1,1,1,0]	[1,1,1,1]
		M_2	AC	AD	BC	BD	A-	B-	C-	D-	AC|BD	AD|BC	!AC	!AD	!BC	!BD	--
marker 2 score	[1,0,0,0]	AC	A	B	C	D	E	J	F	I	G	H	N	L	M	K	O
	[0,0,1,0]	AD	B	A	D	C	E	J	I	F	H	G	L	N	K	M	O
	[0,1,0,0]	BC	C	D	A	B	J	E	F	I	H	G	M	K	N	L	O
	[0,0,0,1]	BD	D	C	B	A	J	E	I	F	G	H	K	M	L	N	O
	[1,0,1,0]	A-	E	E	J	J	E	J	O	O	O	O	U	U	P	P	O
	[0,1,0,1]	B-	J	J	E	E	J	E	O	O	O	O	P	P	U	U	O
	[1,1,0,0]	C-	F	I	F	I	O	O	F	I	O	O	T	Q	T	Q	O
	[0,0,1,1]	D-	I	F	I	F	O	O	I	F	O	O	Q	T	Q	T	O
	[1,0,0,1]	AC|BD	G	H	H	G	O	O	O	O	G	H	R	S	S	R	O
	[0,1,1,0]	AD|BC	H	G	G	H	O	O	O	O	H	G	S	R	R	S	O
	[0,1,1,1]	!AC	N	L	M	K	U	P	T	Q	R	S	V	X	W	Y	O
	[1,0,1,1]	!AD	L	N	K	M	U	P	Q	T	S	R	X	V	Y	W	O
	[1,1,0,1]	!BC	M	K	N	L	P	U	T	Q	S	R	W	Y	V	X	O
	[1,1,1,0]	!BD	K	M	L	N	P	U	Q	T	R	S	Y	W	X	V	O
	[1,1,1,1]	--	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O

These 225 combinations result in 25 unique equations, indexed by the letters A-Y, describing the LOD score contribution of this offspring individual. Cells containing "O" are non-informative combinations with a zero contribution to the total LOD score. Gray cells correspond to combinations of the 7 basic segregation types. Cells with white text on a black background show new variants used to accommodate some ambiguously scored markers.

To obtain a universal LOD score equation, it suffices to count the number of progeny falling into each of these 25 classes (with counts denoted CA - CY), and use these as multiplicands, arriving at equation 4.

$$\begin{aligned}LOD(r_p,r_q)=&C_E\log(2n_p)+C_J\log(2r_p)+C_F\log(2n_q)+C_I\log(2r_q)+\\&C_A\log(4n_pn_q)+C_B\log(4n_pr_q)+C_C\log(4r_pn_q)+C_D\log(4r_pr_q)+\\&C_G\log(2(n_pn_q+r_pr_q))+C_H\log(2(r_pn_q+n_pr_q))+C_0\log(1)+\\&C_K\log(4/3(1-r_pr_q))+C_L\log(4/3(1-n_pr_q))+C_P\log(2/3(1+n_p))+\\&C_M\log(4/3(1-r_pn_q))+C_N\log(4/3(1-n_pn_q))+C_Q\log(2/3(1+n_q))+\\&C_R\log(2/3(1+n_pr_q+r_pn_q))+C_S\log(2/3(1+r_pr_q+n_pn_q))+\\&C_T\log(2/3(1+r_q))+C_W\log(4/9(2+r_pn_q))+C_V\log(4/9(2+n_pn_q))+\\&C_U\log(2/3(1+r_p))+C_X\log(4/9(2+n_pr_q))+C_Y\log(4/9(2+r_pr_q))\end{aligned}$$ eq. 4

Maximum likelihood estimates of recombination frequencies

A maximum likelihood estimate can be obtained by finding the location (r_p,r_q) of the maximum of equation 4. It can be proven that equation 4 is a concave function, therefore we may find a solution if we equate derivatives along two different directions in the (r_p,r_q)-plane, and solve the resulting system of two equations with two unknowns. This results in a pair of simultaneous equations (eq. 5 and 6).

$$\begin{aligned}&\frac{C_C+C_D+C_J}{r_p}+\frac{C_A+C_B+C_E}{r_p-1}+\frac{C_G(2r_q-1)}{1-r_p-r_q+2r_pr_q}+\frac{C_P}{r_p-2}+\\&\frac{C_H(1-2r_q)}{r_p+r_q-2r_pr_q}+\frac{C_Kr_q}{r_pr_q-1}+\frac{C_Lr_q}{1-r_q+r_pr_q}+\frac{C_R(1-2r_q)}{1+r_p+r_q-2r_pr_q}+\\&\frac{C_M(r_q-1)}{1-r_p+r_pr_q}+\frac{C_N(1-r_q)}{r_p+r_q-r_pr_q}+\frac{C_S(2r_q-1)}{2-r_p-r_q+2r_pr_q}+\frac{C_U}{r_p+1}+\\&\frac{C_V(r_q-1)}{3-r_p-r_q+r_pr_q}+\frac{C_W(1-r_q)}{2+r_p-r_pr_q}+\frac{C_X(-r_q)}{2+r_q-r_pr_q}+\frac{C_Yr_q}{2+r_pr_q}=0\end{aligned}$$ eq. 5

$$\begin{aligned}&\frac{C_B+C_D+C_I}{r_q}+\frac{C_A+C_C+C_F}{r_q-1}+\frac{C_G(2r_p-1)}{1-r_p-r_q+2r_pr_q}+\frac{C_Q}{r_q-2}+\\&\frac{C_H(1-2r_p)}{r_p+r_q-2r_pr_q}+\frac{C_Kr_p}{r_pr_q-1}+\frac{C_L(r_p-1)}{1-r_q+r_pr_q}+\frac{C_R(1-2r_p)}{1+r_p+r_q-2r_pr_q}+\\&\frac{C_Mr_p}{1-r_p+r_pr_q}+\frac{C_N(1-r_p)}{r_p+r_q-r_pr_q}+\frac{C_S(2r_p-1)}{2-r_p-r_q+2r_pr_q}+\frac{C_T}{r_q+1}+\\&\frac{C_V(r_p-1)}{3-r_p-r_q+r_pr_q}+\frac{C_W(-r_p)}{2+r_p-r_pr_q}+\frac{C_X(1-r_p)}{2+r_q-r_pr_q}+\frac{C_Yr_p}{2+r_pr_q}=0\end{aligned}$$ eq. 6

Unfortunately this system of equations has no known general analytical solution. There are, however, some interesting special cases that can be solved analytically:

1. When $C_G=C_H=C_K=C_L=C_M=C_N=C_P=C_Q=C_R=C_S=C_T=C_U=C_V=C_W=C_X=C_Y=0$ then:

$$r_p=\frac{C_C+C_D+C_J}{C_A+C_B+C_C+C_D+C_E+C_J}$$
$$r_q=\frac{C_B+C_D+C_I}{C_A+C_B+C_C+C_D+C_F+C_I} \quad \text{eq. 7}$$

This corresponds to a combination of marker exhibiting parental and/or fully classified segregation types.

2. When the marker pair contains information on one parent only:
   a) If $C_A=C_B=C_C=C_D=C_F=C_G=C_H=C_I=C_K=C_L=C_M=C_N=C_Q=C_R=C_S=C_T=C_V=C_W=C_X=$ $=C_Y=0$ then r_q is indeterminate, and r_p can be solved directly from:

$$(C_E+C_J+C_P+C_U)r_p^3-(C_E+2C_J+3C_U)r_p^2- (2C_E+C_J+C_P-2C_U)r_p+2C_J=0 \quad \text{eq. 8}$$

   b) If $C_A=C_B=C_C=C_D=C_E=C_G=C_H=C_J=C_K=C_L=C_M=C_N=C_P=C_R=C_S=C_U=C_V=C_W=C_X=$ $=C_Y=0$ then r_p is indeterminate, and r_q can be solved directly from:

$$(C_F+C_I+C_Q+C_T)r_q^3-(C_F+2C_I+3C_T)r_q^2- (2C_F+C_I+C_Q-2C_T)r_q+2C_I=0 \quad \text{eq. 9}$$

3. If the supremum of LOD(r_p,r_q) is located in one of the corners of the (r_p,r_q) plane:
   a) then $r_p=0$ and $r_q=0$ when the following holds true:

$$-C_A-C_E-C_G-C_M-C_P/2+C_R-C_S/2+C_U-C_V/3+C_W/2<0 \text{ and}$$
$$-C_A-C_F-C_G-C_L-C_Q/2+C_R-C_S/2+C_T-C_V/3+C_X/2<0 \text{ and} \quad \text{eq. 10}$$
$$C_B=C_C=C_D=C_H=C_I=C_J=C_N=0.$$

   b) then $r_p=1$ and $r_q=0$ when the following holds true:

$$+C_C+C_H+C_J+C_N-C_P+C_R/2-C_S+C_U/2-C_V/2+C_W/3>0 \text{ and}$$
$$-C_C-C_F-C_H-C_K-C_Q/2-C_R/2+C_S+C_T-C_W/3+C_Y/2<0 \text{ and} \quad \text{eq. 11}$$
$$C_A=C_B=C_D=C_E=C_G=C_I=C_M=0.$$

   c) then $r_p=0$ and $r_q=1$ when the following holds true:

$$-C_B-C_E-C_H-C_K-C_P/2-C_R/2+C_S+C_U-C_X/3+C_Y/2<0 \text{ and}$$
$$+C_B+C_I+C_H+C_N-C_Q+C_R/2-C_S+C_T/2-V_V/2+C_X/3>0 \text{ and} \quad \text{eq.12}$$
$$C_A=C_C=C_D=C_F=C_G=C_J=C_L=0.$$

   d) then $r_p=1$ and $r_q=1$ when the following holds true:

$$+C_D+C_J+C_G+C_L-C_P-C_R+C_S/2+C_U/2-C_X/2+C_Y/3>0 \text{ and}$$
$$+C_D+C_I+C_G+C_M-C_Q-C_R+C_S/2+C_T/2-C_W/2+C_Y/3>0 \text{ and} \quad \text{eq.13}$$
$$C_A=C_B=C_C=C_E=C_F=C_H=C_K=0.$$

4. $C_R=C_S=C_V=C_W=C_X=C_Y=0$ and the supremum of LOD(r_p,r_q) lies on a boundary of the (r_p,r_q) plane:
   a) If we assume $r_p=0$ ($C_C=C_D=C_J=0$), then we can solve r_q from:

$$(a+b+C_Q+C_T)r_q^3-(2a+b+C_T)r_q^2+(2C_T-C_Q-a-2b)r_q+2=0,$$
$$\text{with } a=C_B+C_D+C_H+C_I+C_N \text{ and } b=C_A+C_C+C_F+C_G+C_L$$
$$\text{Then if: } \frac{\partial LOD(0,r_q)}{\partial r_p}\le 0 \text{ we have found the maximum at } (0,r_q) \qquad \text{eq. 14}$$

   b) If we assume $r_p=1$ ($C_A=C_B=C_E=0$), then we can solve r_q from:

$$(a+b+C_Q+C_T)r_q^3-(2a+b+C_T)r_q^2+(2C_T-C_Q-a-2b)r_q+2=0,$$
$$\text{with } a=C_B+C_D+C_I+C_G+C_M \text{ and } b=C_A+C_C+C_F+C_H+C_K$$
$$\text{Then if: } \frac{\partial LOD(1,r_q)}{\partial r_p}\ge 0 \text{ we have found the maximum at } (1,r_q) \qquad \text{eq. 15}$$

   c) If we assume $r_q=0$ ($C_B=C_D=C_I=0$), then we can solve r_p from:

$$(a+b+C_P+C_U)r_p^3-(2a+b+C_U)r_p^2+(2C_U-C_P-a-2b)r_p+2=0,$$
$$\text{with } a=C_C+C_D+C_H+C_J+C_N \text{ and } b=C_A+C_B+C_E+C_G+C_M$$
$$\text{Then if: } \frac{\partial LOD(r_p,0)}{\partial r_q}\le 0 \text{ we have found the maximum at } (r_p,0) \qquad \text{eq. 16}$$

   d) If we assume $r_q=1$ ($C_A=C_C=C_F=0$), then we can solve r_p from:

$$(a+b+C_P+C_U)r_p^3-(2a+b+C_U)r_p^2+(2C_U-C_P-a-2b)r_p+2=0,$$
$$\text{with } a=C_C+C_D+C_G+C_J+C_L \text{ and } b=C_A+C_B+C_E+C_H+C_K$$
$$\text{Then if: } \frac{\partial LOD(r_p,1)}{\partial r_q}\ge 0 \text{ we have found the maximum at } (r_p,1) \qquad \text{eq. 17}$$

In other cases we use Newton's approach to find a maximum, iterating:

$$\vec{x_{n+1}}=\vec{x_n}-\lambda[\boldsymbol{K}\,LOD(\vec{x_n})]^{-1}\nabla LOD(\vec{x_n}) \quad \text{until} \quad |\vec{x_{n+1}}-\vec{x_n}|<1/(2\mathrm{N}) \quad , \qquad \text{eq. 18}$$

with $\vec{x_n}=(r_{p,n},r_{q,n})$, N the number of offspring and $\boldsymbol{K}\,LOD(\vec{x_n})$ the Hessian matrix:

$$\boldsymbol{K}\,LOD(\vec{x_n})=\begin{bmatrix}\frac{\partial^2 LOD(r_p,r_q)}{\partial r_p^2} & \frac{\partial^2 LOD(r_p,r_q)}{\partial r_p \partial r_q}\\ \frac{\partial^2 LOD(r_p,r_q)}{\partial r_q \partial r_p} & \frac{\partial^2 LOD(r_p,r_q)}{\partial r_q^2}\end{bmatrix} \qquad \text{eq. 19}$$

Automatic phase discrimination

Any pair of markers can be linked in cis or in trans with respect to each other in both parents independently, resulting in four possible relative configurations. These linkage configurations are denoted CC, TC, CT and TT with C standing for “cis” and T for

"trans", the first letter applying to parent P, and the second to parent Q. We want to prove that a LOD score under any assumption of relative phase can be calculated immediately from the LOD score under any other assumption of relative phase (with the subscript to the LOD denoting the assumed phase relation):

$$LOD_{CC}(c_p, c_q) = LOD_{TC}(t_p, c_q) = LOD_{CT}(c_p, t_q) = LOD_{TT}(t_p, t_q) \quad \text{eq. 20}$$

with c_p, c_q, t_p=1-c_p and t_q=1-c_q the maternal and paternal recombination frequencies in cis and trans respectively. Substituting c_p, c_q, t_p and t_q for the r_p and r_q parameters in the H matrix (equation 3), we note that the resulting matrices (denoted H_{CC}, H_{TC}, H_{CT} and H_{TT}) will be permutations of each other. The same effect can be achieved using the following permutation matrices:

$$X_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, X_p = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, X_q = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, X_{pq} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}. \quad \text{eq. 21}$$

Obtaining:

$$H_{CC} = X_i \times H_{CC}, \quad H_{TC} = X_p \times H_{CC}, \quad H_{CT} = X_q \times H_{CC} \text{ and } H_{TT} = X_{pq} \times H_{CC} \quad \text{eq. 22}$$

with:

$$H_{CC} = \begin{bmatrix} c_p c_q/4 & t_p c_q/4 & c_p t_q/4 & t_p t_q/4 \\ t_p c_q/4 & c_p c_q/4 & t_p t_q/4 & c_p t_q/4 \\ c_p t_q/4 & s_p s_q/4 & c_p c_q/4 & t_p c_q/4 \\ t_p t_q/4 & c_p t_q/4 & t_p c_q/4 & c_p c_q/4 \end{bmatrix}. \quad \text{eq. 23}$$

For the purpose of this proof we postulate an "absolute" phase relationship between the parental chromosomes and a marker, which we will denote {αα}, {βα}, {αβ} or {ββ}, with the first letter applying to parent P, and the second to parent Q. Markers are in cis-phase on a particular parent if the corresponding letters are the same. Marker scores can be converted from one absolute phase to any other absolute phase using the same permutation matrices as in equation 21:

When "converting" to the same phase: $M' = X_i \times M$,

When converting between {αα} ↔ {βα} or {αβ} ↔ {ββ}: $M' = X_p \times M$,

When converting between {αα} ↔ {αβ} or {βα} ↔ {ββ}: $M' = X_p \times M$, eq. 24

When converting between {αα} ↔ {ββ} or {βα} ↔ {αβ}: $M' = X_{pq} \times M$

In these equations M' is the phase converted marker. Thus we can modify the numerator (the denominator is phase and recombination frequency independent) of $odds_c(r_p, r_q)$ (equation 3):

$$((X_1 \times M_1) \times (X_H \times H_{CC})) \times (X_2 \times M_2)^T = M_1 \times H_{CC} \times (X_H \times X_1 \times X_2) \times M_2^T, \quad \text{eq. 25}$$

with X_1 and X_2 the permutation matrices applied to M_1 and M_2 respectively, and X_H the corresponding permutation matrix applied to H_{CC}. Because the term $X_H \times X_1 \times X_2$ always equals Xi (the identity matrix), we have proven the equivalent of equation 20 for the oddsc(r_p,r_q) of individual progeny, and because the phase relationship affects all progeny similarly, by extension, equation 20.

Therefore it suffices to score and process individual markers as if they are all in any arbitrary phase, calculate (pairwise) maximum likelihood estimates r_p and r_q over the full range (0 to 1) of r_p and r_q, then decide the true relative phase relationship based on the magnitudes of both r_p and r_q, and then obtain phase-corrected recombination frequencies (r_p*,r_q*) as follows:

$$\begin{aligned}
&r_p < 0.5 \text{ and } r_q < 0.5 \Rightarrow phase = CC,\ r_p^* = r_p,\ r_q^* = r_q \\
&r_p > 0.5 \text{ and } r_q < 0.5 \Rightarrow phase = TC,\ r_p^* = n_p,\ r_q^* = r_q \\
&r_p < 0.5 \text{ and } r_q > 0.5 \Rightarrow phase = CT,\ r_p^* = r_p,\ r_q^* = n_q \\
&r_p > 0.5 \text{ and } r_q > 0.5 \Rightarrow phase = TT,\ r_p^* = n_p,\ r_q^* = n_q
\end{aligned} \qquad \text{eq. 26}$$

It is evident that for r_p=0.5 (exactly) and r_q=0.5 (exactly) the relative phase is indeterminate. This is, however, of little consequence as these values signify independent assortment of the markers.

Implementation and simulated examples

The method outlined above has been implemented in a freely available PERL script (http://theo.borm.org/pbsw/). This PERL script takes the segregation patterns of a pair of markers as input, and computes pairwise linkage estimates and LOD score. In addition, the PERL script can generate R-language (Ihaka and Gentleman 1996) output that can be used to visualize the LOD(r_p,r_q) graph and the position of it's apex (the maximum likelihood estimate).

Two simulated fully classified marker segregation paterns were obtained for two simulated linked loci by first generating a sequence of 100 random PACs for the first locus, then swapping the parental alleles of parents P and Q alleles in random subsets of 10 and 20 “offspring” respectively. In this way we obtain two simulated marker scores linked at r_p=0.1 and r_q=0.2. These simulated marker scores were used in two in silico experiments first to demonstrate the effect of relative linkage phase on r_p, r_q and LOD(r_p,r_q), second to demonstrate the effect of introducing some types and degrees of ambiguity. For the first experiment we kept the scores for the first marker constant while applying one of the four permutation matrices (X_i, X_p, X_q or X_{pq}) to the scores of the second marker. For the second experiment we introduced different types and amounts of ambiguity into these markers:

1. Random ambiguity in 50 individuals: individual scores “gained” a random extra PAC
2. Random ambiguity in all individuals: individual scores “gained” a random extra PAC
3. Simulated dominance in 50 individuals: “BC”, “AD” and “BD” changed to “!AC”
4. Simulated dominance in all individuals: “BC”, “AD” and “BD” changed to “!AC”

Results and discussion

Linkage analysis in full sib families descending from non-inbred parents is generally considered more difficult than linkage analysis in a progeny derived from homozygous inbred lines because in CP families the segregation type of markers may vary between loci. In addition the linkage phase of markers is not known a priori (Maliepaard et al. 1997). To this we can add the complexity of dealing with some different types of ambiguous marker scores. These ambiguous marker scores can arise as a consequence of failure to observe a clear difference between marker alleles in part of the progeny and as a result of a material or handling error affecting part of the progeny.

Wu et al. (2002a) demonstrated a sex specific pairwise maximum likelihood linkage estimator for markers adhering to the seven core segregation types defined by Maliepaard et al. (1997), and we extended and rephrased this model to cope with arbitrary ambiguity in marker scores. A new PAC-based marker scoring system was introduced that is ideally matched to our objectives. Other marker scoring systems (e.g. the system used by Joinmap (Stam 1993, 1995) do allow for some ambiguity in marker scores for full-sib families derived from non-inbred parents (“CP” population type in Joinmap), but to our knowledge no other scoring system is currently completely generic in this respect. The result of our effort is a concise derivation from first principles of a universal LOD score equation and corresponding maximum likelihood linkage estimator.

It is interesting to note that, though most components of equation 4 arise as a result of the seven core segregation types, there are four components (corresponding to R,W,X and Y in Table 2) that are novel. These four components are variants of other components expressed in a different linkage phase, and thus do not represent a different segregation type per se. They are, however, necessary components to deal with arbitrarily ambiguous marker scores, as can be illustrated by considering the case of two ambiguously, co-dominantly scored AFLP markers.

Although we have not conducted an in-depth analysis of any issues, if, for a specific purpose, an integrated linkage map is desirable, we propose that if a single, sex-averaged, recombination frequency estimate can be obtained by averaging the phase-corrected sex specific recombination frequencies. This averaging could for instance be done

geometrically, using $r'=(r_p+r_q)/2$ (amounting to a projection onto the line $r_p=r_q$), or conserving the magnitude of the recombination frequency using $r''=sqrt(r_p \times r_p+r_q \times r_q)$.
Without a known analytical solution to the system of equations (eq. 5 and 6), we have to use numerical methods. The special cases where we solve the system analytically are of relevance for three reasons. Firstly, obtaining an analytical solution is generally faster. Secondly, if a solution lies on the boundaries and corners of the (r_p,r_q)-plane, then solving these analytically ensures that the numerical procedure does not need to reach these areas of the solution space, thereby simply avoiding possible divisions by zero in equations 5 and 6. Thirdly, if the special cases do not apply, then it can be proven (e.g. by computing the eigenvalues of the Hessian matrix in terms of r_p and r_q) that the Hessian matrix is a Hermitian definite negative matrix, and that therefore $LOD(r_p,r_q)$ is strictly concave, with therefore a single (global) maximum, which is a convenient property for numerical procedures that might otherwise fail to converge to the global maximum. We have employed Newton's method because of its simplicity and fast convergence, but other methods such as the Expectation Maximization (EM) algorithm (Dempster et al. 1977), with applications in linkage analysis illustrated by Maliepaard et al. (1997) and Wu et al. (2002a) may also be used.
Though some of the possible types of marker scoring ambiguity may currently appear of academic interest only, including them does not complicate our model. Furthermore, we expect that in the future, for marker technologies as it is for sequencing technologies today, the cost of processed data will become a more important concern than the quality of individual raw data points: In linkage maps, like in DNA sequencing, quality can be obtained through redundancy (For example the bin signatures in the high density genetic map of potato (van Os et al. 2006) represent information extracted from a redundant dataset). Being able to effectively deal with scoring ambiguity in other ways than by discarding data may allow a better trade-off to be made.

Illustratation of the effect of different linkage phases

note: The specific marker scores used in this section are given in the appendix.

Figure 1 shows graphs of the $LOD(r_p,r_q)$ score plane for two synthetic loci converted to markers in the 4 possible different relative linkage phase configurations: cis-phase in both parents (top left), trans-phase in parent P and cis-phase in Q (top right), cis-phase in P and trans-phase in Q (bottom left) and trans-phase in both parents (bottom right). As expected, all four graphs reach a maximum LOD score of 24.4 at a phase-corrected recombination frequency of 0.10 in parent P and 0.20 in parent Q. The issue of determining the relative linkage configuration of markers is, in our opinion, an artifact caused by the under-parameterization of the model used for non-sex-specific linkage estimators. We have

mathematically proven (and illustrated in Figure 1) that linkage configuration can be determined directly from the maximum likelihood recombination frequency estimates obtained (n.b. a sex specific recombination frequency of less than 50% signifying linkage in cis, while a recombination frequency of more than 50% signifies linkage in trans). This renders determination of relative linkage configuration of markers in a two point analysis essentially a non-issue.

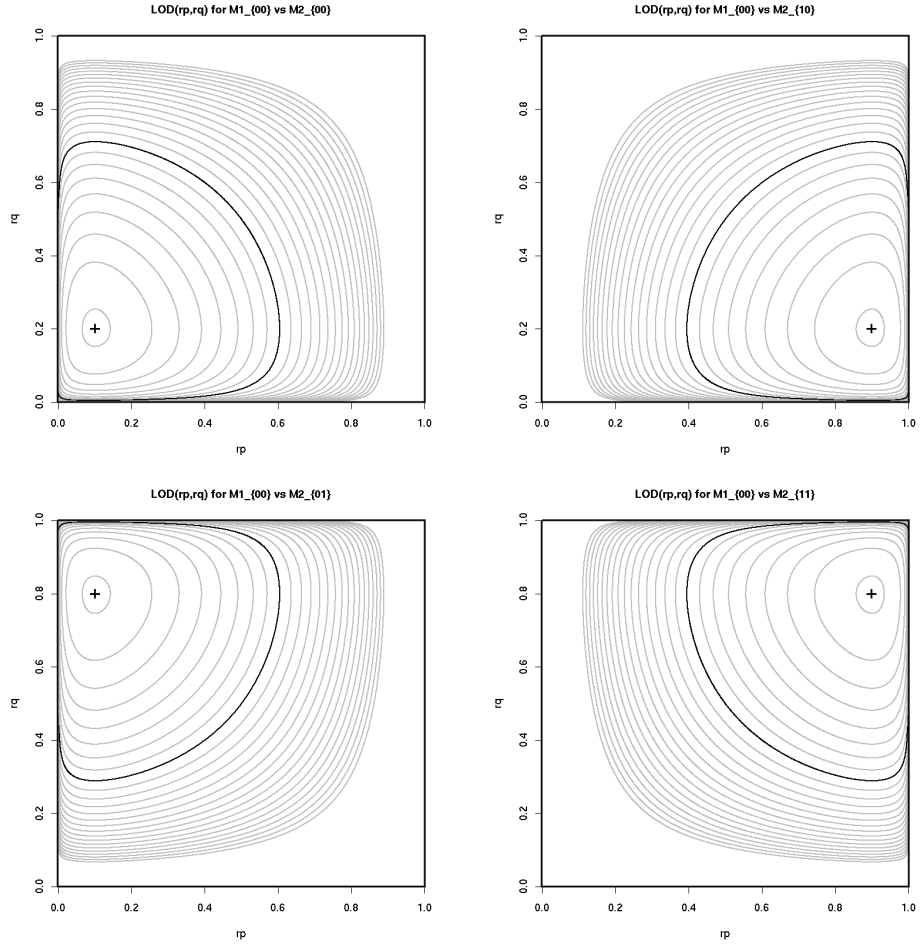


Figure 1: Illustration of the mirror symmetry of LOD(r_p,r_q) graphs for the same data expressed in different relative phase combinations. The 3-D shape of the LOD(r_p,r_q) graph is visualized through concentric equimetric LOD contours at three LOD unit intervals (n.b. a 1000-fold difference in likelihood). The widest contours indicate where LOD(r_p,r_q) equals minus 48, while black contours indicate where LOD(r_p,r_q) equals zero. The apexes of the graphs are indicated by a "+" symbol.

Illustration of the effect of ambiguities in marker scores

note: The specific marker scores used in this section are given in the appendix.

Figure 2 illustrates the effect on the graphs of LOD(r_p,r_q) caused by the introduction of some different types and levels of ambiguity in the marker scores (of the same synthetic loci as shown in Figure 1).

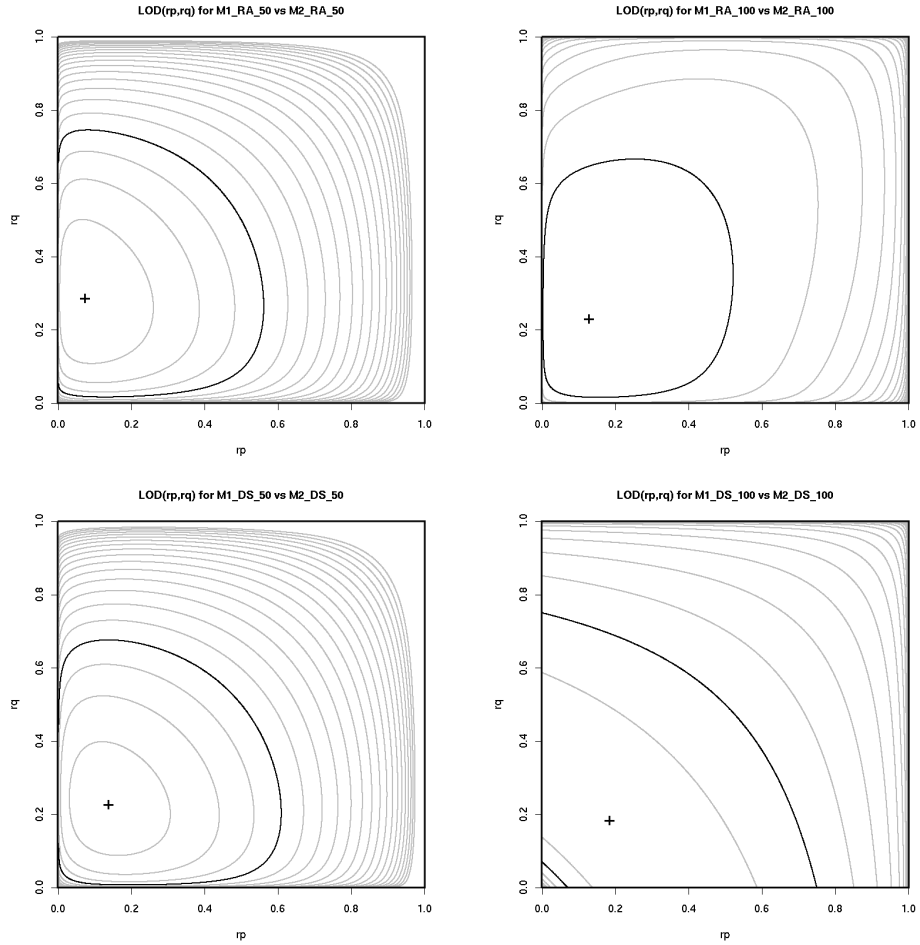


Figure 2: Illustration of the effect of introduction of ambiguities on the LOD(r_p,r_q) graphs for the same loci as shown in figure 1. The 3-D shape of the LOD(r_p,r_q) graph is visualized through concentric equimetric LOD contours at three LOD unit intervals (n.b. a 1000-fold difference in likelihood). The widest contours indicate where LOD(r_p,r_q) equals minus 48, while black contours indicate where LOD(r_p,r_q) equals zero. The apexes of the graphs are indicated by a "+" symbol. The top row shows the effect of introducing random ambiguity in either half (left) or all (right) the offspring. The bottom row shows the effect of converting half (left) or all (right) of the marker scores from fully classified to dominantly scored.

In the top row random ambiguity is introduced in half (left) and all (right) of the individuals of the offspring by adding one extra PAC into to each marker score at random (for instance converting a “AC” score into a “AC|xx” score, with “xx” chosen at random from “AD”, “BC” or “BD”). These graphs reach maximum LOD scores of 11.5 and 2.6 at $(r_p,r_q) = (0.07,0.29)$ and (0.13,0.23) respectively. In the bottom row of figure 2, half (left) or all (right) of the offspring have been converted from a fully classified <abxcd> segregation type to a dominantly scored <a0xa0> segregation type. These graphs reach maximum LOD scores of 10.7 and 4.9 at $(r_p,r_q) = (0.14,0.23)$ and (0.18,0.18) respectively. As illustrated, ambiguities will have an effect on attainable LOD score as well as on the recombination frequency estimates obtained. The latter can at least partially be attributed to the finite sample size (analogous to the discussion of the effect of finite sample size by Maliepaard et al. (1997)). However, as particularly well-illustrated by the bottom right graph in Figure 2, the fact that some of the marker scores (n.b. in case of dominance) do not allow a distinction to be made between parental alleles may also result in a shift of the recombination frequency estimates. As the bottom right graph of Figure 2 is based on completely dominantly scored data the maximum likelihood linkage estimates are equal in both parents and the deviating shape of this graph (in particular LOD scores in excess of zero on the axes where $r_p=0$ and $r_q=0$) can also be attributed to the effect of this dominant scoring: On the $r_p=0$ axis the observed marker scores may be explained by a large recombination frequency in parent Q and vice versa.

Chapter 4

Binmap+ and Homap+: Retrofitting normal and homoplasic markers to framework maps using the Universal Maximum Likelihood Pairwise Linkage Estimator

T.J.A.Borm, J. de Boer, H.J. van Eck and R.G.F. Visser

Abstract

Over the last decades, large scale genetic linkage mapping projects have produced several, increasingly marker dense, high quality, framework linkage maps. Recently a high density, 10,000 marker, bin-based genetic map of potato was published. To exploit such maps effectively, we need to be able to retrofit new markers of arbitrary types and possibly lesser quality without disrupting the underlying framework. Using the recently described Universal Maximum Likelihood Pairwise Linkage Estimator (UMLPLE), Binmap+ achieves this objective in maps of full-sib mapping populations descending from heterozygous, outbreeding parents. Previously, the separate maternal and paternal linkage groups in the high density map of potato were left essentially unaligned. Binmap+ was first used to select a suitable set of high quality markers in order to set up a system of constraints, aligning the maternal bins with the paternal bins to the extent suported by the marker data. After this, Binmap+ was used to re-map all markers. Binmap+ is offered as a web/e-mail based service to the scientific community in combination with the underlying bin-based map of potato, or as source code for use with other maps. In addition we present Homap+, a variant of the Binmap+ program, which is, to our knowledge, the only software capable of mapping homoplasic markers.

Introduction

94 Years after it's inception (Sturtevant 1913), linkage analysis is still an indispensable tool to study genome structure. Even with sequencing costs continually dropping, standing at the brink of the 1000 dollar genome (Mardis 2006) it is good to realize that, to our knowledge, no higher organism has been sequenced nor any gene cloned from it without reference in some way or another to a linkage map. This is not to say nothing has changed in 94 years: Different and refined statistical models, a multitude of marker techniques and various types of segregating populations can now be used to construct maps, and even the natural mechanism (meiosis) causing partial linkage of markers can be replaced by an artificial mechanism (e.g Radiation Hybrid mapping, Cox et al. 1990). Also the number of markers in maps has dramatically changed: Where Sturtevant's 1913 map contained six phenotypic markers, several linkage maps (e.g. Sun et al. 2007, van Os et al. 2006, Kwitek et al. 2004) have now been constructed in collaborative scientific efforts containing more than 10,000 markers. These maps, however dense, are still not complete representations of the underlying genomes, and new markers, mapping various traits, may need to be added at a later stage.

In his procedure for the selection of an optimal set of genotypes from an initially larger segragating population, Vision et al. (2000) makes use of the concept of bins: “An interval along a linkage group within which no breakpoint occur among any members of a given set of individuals but which is bounded by such breakpoints in at least one individual (or by the end of a linkage group)”. Software is available (Brown et al. 2000) to place markers onto such bins. Sun et al. (2007), and van Os et al. (2006) constructed bin-based linkage maps de novo, albeit using a modified definition of what constitutes a bin. Their definition of a bin is “A genetic interval comprising co-segregating markers which can not be ordered further given the recombination events in the mapping population”. Van Os et al. (2006) used software to deal with genotyping errors and missing values in the marker scores (SMOOTH, van Os et al. 2005a), producing consensus bin signatures. Bin signatures represent the segregation pattern of the chromatids as inferred from marker scores. These bin signatures were subsequently ordered (RECORD, van Os et al. 2005b) according to the principle of maximum parsimony. The resultant map is no longer a marker map but a recombination map (Sun et al. 2007), consisting of high quality consensus bin signatures, to which all the (inferior) marker scores (including those which originally defined the bin signatures) were retrofitted. Originally this retrofitting of markers was performed using software restricted to dominantly scored AFLP data (BINMAP and BRIDGEMAP, van Os, unpublished).

Vision et al. (2000) proposed to treat mapping as a two-phase process; the first phase

being the construction of a well-measured framework map, the second stage being the mapping of markers using a reduced set of offspring. This reduction in size of the mapping population, however carefully selected, fundamentally results in a reduced mapping resolution and hence in less accurately positioned markers. We propose to indefinitely extend the second phase to include all new markers added at a later date. We also propose that all markers should be treated equal, and only be retrofitted to the framework map.

Retrofitting of markers to a framework map is markedly different approach from conventional linkage analysis. The distance conventionally estimated between a pair of markers is a distance due to cross-over events in a fraction of the offspring plants. This distance might be inflated by error, and double crossovers go unnoticed. Retrofitting on a saturated scaffold bin map implies a survey to which known position a marker fits best. A distance of zero implies a perfect fit of a marker in a bin, whereas distances larger than zero can be explained only as the result of singletons (scoring errors, gene conversion etc.; van Os et al., 2005a). This separation of distance due to recombination (irrelevant for retrofitting) and distance due to scoring errors offers a measure for data quality, and is referred to in this paper as an apparent error rate. As mapping in full sib progenies of outbreeding species is generally considered more complicated than mapping using F2 or BC1 populations (Maliepaard et al. 1997), it is reasonable to expect that this is also the case for retrofitting markers to such maps. One of the complicating factor is that, if markers are used that segregate from only one parent, this results in separate parental maps (for these markers). These separate parental maps can only be aligned through bridge markers (Ritter et al. 1990) segregating from both parents simultaneously. Depending on the particular type of markers used, these bridge markers may be more (e.g. SSR markers with multiple distinguishable alleles segregating from both parents) or less (e.g. dominantly scored AFLP™ markers) informative, and this, in combination with missing values and marker scoring errors, may lead to ambiguity in the alignment of parental maps (van Os et al. 2006). When retrofitting new bridge markers to ambiguously aligned parental maps, this ambiguity must be accommodated, and we propose to use a system of constraints derived from the raw data for this, allowing for marker placement ambiguity when more precise localization is not supported by the data.

Using the recently developed Universal Maximum Likelihood Pairwise Linkage Estimator (UMLPLE) (Chapter 3 of this thesis) as well as using the system of constraints, the Binmap+ program was implemented for retrofitting markers to framework maps. In its simplest form the constraints define which parental linkage groups represent homologous chromosomes as no proper bridge marker should segregate from non-homologous chromosomes. AFLP, however, is known to occasionally produce homoplasic fragments

(i.e. Non-allelic AFLP fragments of equal mobility on gel) (Meudt and Clarke 2007). Such fragments, when segregating from both parents, can be mapped using our software by removing the constraints. In addition, it was surmised that one can map mixtures of two homoplasic AFLP fragments segregating from only one of the parents by removal of the constraints and simple substitution of one of the underlying parental bin-maps. The result is HOMAP+, a version of the BINMAP+ program specifically tailored to mapping homoplasic markers. Both programs, in combination with the underlying bin-map of potato (van Os et al. 2006), are being offered as a web/email based service (https://secure.potatogenome.net/binmap) to the scientific community. In addition the programs can be downloaded (http://theo.borm.org/pbsw/) as a set of (operating system independent) C and Perl language source code files for use with other framework (bin-) maps. Both programs have been applied to the AFLP markers used to construct the 10,000 marker map of potato (van Os et al. 2006) to obtain first a (conservative) set of constraints and then to re-map all markers, and some results will be discussed.

Materials and methods

Although van Os et al. (2006) did identify the homologous linkage groups in the maternal and paternal maps, no attempt was made to properly integrate these maps. Bridge markers were placed on both maps by locating the pair(s) of a maternal and a paternal bin signature best explaining the observed marker score. If a particular linkage group has M and P bins in the maternal and paternal map respectively, this means that M*P possible combinations of a maternal and a paternal bin were checked, despite the fact that many of these combinations are clearly impossible (and even contradicted by other data). Therefore we must first use bridge marker data to obtain a (conservative) set of constraints limiting which combinations of paternal and maternal bins a marker can be assigned to, and only then we can proceed to re-map all bridge markers using this set of constraints. Finally we attempt to re-map some markers under alternative hypotheses of their mode of inheritance (both normal and homoplasic). As we obtain the set of constraints using the same methods that will eventually use these constraints, the components will be treated in non-chronological order.

Nomenclature and data sources

Naming of bin signatures largely follows van Os et al. (2006): Each bin is denoted by a name starting with a two letter code, SH or RH, identifying the maternal and paternal potato genotypes (SH83-92-488 and RH89-039-16, Rouppe van der Voort et al. 1997), a two digit chromosome number, the letter “B” and a three digit bin number. To this we append either “_C” (for a complete original signature with no missing scores), “_I” (for an

incomplete original signature) or "_P" (for postulated bin signatures). Parental bin signatures will be given as parental type marker scores, using the scoring system defined for use with the UMLPLE (chapter 2 of this thesis): "A-" and "B-" for maternal bin signatures,"C-" and "D-" for paternal bin signatures and "--" for missing values. Bin signatures were retrieved from the previously used software (BINMAP/BRIDGEMAP, van Os unpublished), and marker scoring data was extracted from the database underlying the online presentation of the genetic map (http://www.plantbreeding.wur.nl/potatomap/).

Basic assumptions

1. Although some offspring genotypes may be missing, the existing bin signatures are essentially free from genotyping errors.
2. The order of the existing bin signatures is correct.
3. Although bin signatures with missing values lead to some ambiguity concerning the exact location of recombination events, no recombinations have remained undetected.
4. The positions of recombinations in different offspring do not exactly coincide.
5. Parental chromosomes are co-linear and do not contain any translocated or inverted regions.

Postulating empty-bin parental bin signatures for empty bins

If adjacent complete (no missing values) parental bin signatures differ in L>1 positions, then L>1 offspring must exhibit a recombination in the corresponding interval, and as a consequence of the fourth assumption there should be L-1, as yet unknown, bin signatures between them. Such unknown bin signatures represent empty bins for which no markers were found. If one or more of the bin signatures has missing values (is incomplete), then this results in ambiguity in the number of intermediate bin signatures. Given the third assumption, we can postulate part of the signatures of the empty bins: If existing adjacent bin signatures have identical scores in a particular genotype, then the corresponding intermediate score must be identical, otherwise this score is unknown. Evidently, even if immediately adjacent bin signatures are complete, these postulated empty-bin signatures will inevitably have two or more missing values resulting from the ambiguous expectations on the order of the recombination events that separate those adjacent bins. Despite this, because a missing observation is less detrimental to the LOD score a marker can achieve than a conflicting observation (resulting from using an adjacent existing bin signature), these postulated empty-bin signatures can be successfully used as mapping targets in the bin mapping process to be discussed. Figure 1 illustrates the process of postulation for a subset of progeny in the bin-map of potato.

	#11	#57	#91	#144	#157	#164	#179	#190	#203	4	11	20	38	46	55	59	74	76	82	83	91	97
RH08B049_C	C-	D-	C-	C-	C-	C-	D-	C-	C-	D-	C-	D-	C-	C-	C-	C-	C-	C-	D-	C-	D-	D-
postulated sig #1	C-	--	--	C-	--	C-	D-	--	C-	D-	C-	D-	C-	C-	C-	C-	--	C-	D-	--	D-	--
RH08B051_I	C-	--	D-	C-	D-	C-	D-	--	C-	D-	C-	D-	C-	C-	C-	C-	--	C-	D-	--	D-	--
postulated sig #2	--	--	D-	--	D-	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
RH08B071_C	D-	C-	D-	D-	D-	D-	C-	D-	D-	C-	D-	C-	D-	D-	D-	D-	D-	D-	C-	D-	C-	C-

Figure 1: Postulated empty-bin signatures (gray rows) are derived from flanking existing bin signatures based on the assumption that none of the offspring (columns) exhibits a double recombination in the interval between two existing bin signatures: For offspring where the adjacent existing bin signatures have the same score, this same score is interpolated into the postulated signature. This figure shows a subset of the offspring for an existing range of bins in the genetic map, with all the individuals without recombination within this range removed.

Assigning markers to bins using Binmap+

The method used by Binmap+ to assign markers to bins is a variation of the method used previously (BINMAP/BRIDGEMAP, van Os, personal communication): Markers are assigned to bins by scanning bin signatures for the best match (highest LOD score) with a given marker segregation pattern. If multiple bin positions result in an identical LOD score, then the marker is assigned to multiple bins. For markers segregating from only one of the parents, only the bin signatures for that parent are scanned. For markers segregating in both parents, synthetic bridge-bin signatures are scanned. Synthetic bridge bin signatures are synthesized on the fly for valid (within the given constraints) combinations of a maternal and a paternal bin signature according to Table 1. This generates bridge bin signatures where full classification of all four parental alleles is offered, with ambiguous scores used wherever one of the parental signatures has a missing value. Bridge bin signatures are synthesized in a single linkage phase combination only, and actual marker linkage phase is detected using the UMLPLE as described (chapter 3 of this thesis). For markers assigned to multiple bridge bins, a list of parental bin combinations is retained.

Table 1: Combining the alleles in parental signatures to construct synthetic bridge bin signatures.

maternal allele:	A-	A-	A-	B-	B-	B-	--	--	--
paternal allele:	C-	D-	--	C-	D-	--	C-	D-	--
synthetic bridge score:	AC	AD	A-	BC	BD	B-	C-	D-	--

In a single descendant offspring genotype of the mapping population, a maternal (“A-”, “B-” or “--”) allele is combined with a paternal (“C-”, “D-” or “--”) allele. Result is a synthetic bridge bin signature score which may be fully classified (“AC”, “AD”, “BC” or “BD”), ambiguous (“A-”, “B-”, “C-” or “D-”), or be missing (“--”). Hence synthetic bridge bin signatures may be of a mixed segregation type (e.g. a mixture of scores ordinarily belonging to <ab x cd>, <a0 x 00> and <00 x a0> segregation types).

Assigning markers to bins using Homap+

Homoplasic markers are assigned to bins in much the same way as bridge markers:

1. Homoplasic markers segregating from both parents (i.e. dominantly scored AFLP

markers appearing to be normal bridge markers based on the parental genotypes and exhibiting a nominally 3:1 segregation in the offspring) are mapped exactly like normal bridge markers except for the removal of all normal constraints and the imposition of the constraint that maternal and paternal linkage groups must be different.

2. Homoplasic markers segregating only from the female parent (i.e. dominantly scored AFLP markers appearing to be a maternal marker based on the parental genotypes but with a skewed (nominally 3:1) segregation in the offspring) are mapped exactly like homoplasic markers segregating from both parents except for the substitution of transformed maternal bin signatures for the paternal bin signatures in the analysis. The transformation of maternal bin signatures consists of replacement of the maternal "A" and "B" alleles in the maternal bin signature with "C" and "D" respectively.
3. Homoplasic markers segregating only from the male parent (i.e. dominantly scored AFLP markers appearing to be a paternal markers based on the parental genotypes but with a skewed (nominally 3:1) segregation in the offspring) are mapped exactly like homoplasic markers segregating from both parents except for the substitution of transformed paternal bin signatures for the maternal bin signatures in the analysis.The transformation of paternal bin signatures consists of replacement of the paternal "C" and "D" alleles in the paternal bin signature with "A" and "B" respectively.

Consequence of the way in which homoplasic markers segregating from a single parent are handled is that such markers must not be scored as parental segregation type markers but instead like co-dominantly scored AFLP bridge markers (n.b. using "AC", "BC|AD", "BD" and "!AC" scores). No attempt is made to a priori infer homoplasy from putatively skewed observed segregation ratios.

Apparent error rates

The parental recombination rates obtained through the UMLPLE serve as (parent-specific) estimators of the apparent error rate for each marker: As bin signatures are taken to be complete (covering the entire genome) and correct (no errors, missing values at most), any apparent recombination between a bin signature and a marker must be caused by a scoring error in the marker. By this approach we can effectively separate distance due to error from distance along the linkage map due to recombination events (van Os et al.2005a).

Constraints

As parental chromosomes are assumed to be co-linear, it must be possible (within a

certain linkage group) to align the maternal bins to the paternal bins in such a way that:

1. If maternal bin number m is linked to paternal bin p_1, then maternal bin number m+1 must be linked to some paternal bin number p_2, with $p_1<=p_2$, except if the paternal chromosome is missing one or more bin signatures near its end, in which case maternal bin number m+1 may remain unlinked.
2. If paternal bin number p is linked to maternal bin m_1, then paternal bin number p+1 must be linked to some maternal bin number m_2, with $m_1<=m_2$, except if the maternal chromosome is missing one or more bin signatures near its end, in which case paternal bin number p+1 may remain unlinked.

In short this means that if one would draw a graph of the alignment of parental bins, with the maternal bins on the X-axis and the paternal bins on the Y-axis, then such an "alignment graph" must be continuously increasing – without any sections with a negative first derivative. In practice, because of uncertainty caused by incomplete marker data and not completely informative marker scores, the true "alignment graph" may be located within some bandwidth, producing an "alignment zone", with the bandwidth dictated by the data. If some good bridge marker (albeit possibly with incomplete marker data and not completely informative segregation type) is mapped to a to a range of bin combinations in both parents, with the maternal range running from bin number m_1 to m_2, and the paternal range running from p_1 to p_2, then:

1. No other bridge marker can link a maternal bin $m'<m_1$ to any paternal bin $p'>p_2$
2. No other bridge marker can link a maternal bin $m'>m_2$ to any paternal bin $p'<p_1$

The first constraint will be called a "north-west constraint", and the second constraint will be called a "south-east constraint". The Binmap+ program expects the constraints to be specified by the user as a list of north-west and south-east constraints, with each individual constraint appearing as a combination (m,p) of a maternal and a paternal bin number.

Actual constraints were obtained using only the highest quality bridge markers. Bridge markers were initially mapped using Binmap+ without any constraints, and only those assigned to bin combination(s) reaching a LOD score of at least 20 and with an apparent error rate of zero were retained. The bin ranges to which these markers were mapped were inspected, and markers causing conflicting alignments were removed. For each of the remaining markers the minimum and maximum maternal (m_{min} and m_{max}) and paternal (p_{min} and p_{max}) bin numbers of the range of bin combinations to which the marker was mapped were obtained. Subsequently (m_{min}-1,p_{max}+1) was used as an northwest constraint and (m_{max}+1,p_{min}-1) was used as a southeast constraint. Note that by adding or subtracting one bin as indicated, a slightly wider bandwidth is obtained, and also note that for some markers m_{min} may equal m_{max} and/or p_{min} may equal p_{max}.

Mapping of markers

After obtaining a set of constraints all markers were re-mapped using Binmap+ utilizing these constraints. If we assume that our bin-map (including the postulated empty-bin signatures) is correct and complete (except for missing genotype values) then an ideal marker (with the correct segregation type and no scoring errors) should be assigned to the bins of the map with zero apparent error rate (no recombinations). If, however, we observe a non-zero apparent error rate, this should therefore be the result of a non-ideal marker with erroneous scores in some of the genotypes. High apparent error rates can indicate (amongst others) that markers are composed of multiple homoplasic fragments. Therefore all markers with an apparent error rate >= 0.05 in one or both of the parents or with a LOD score <=6 were rejected and re-mapped using all other possible modes of inheritance, including homoplasy using Homap+. Results were compared with previously obtained map positions.

Results

Constraints

Of the 2208 bridge markers available, 1721 were rejected because their LOD score was too low (<=20), or because their apparent error rate was non-zero. Map positions of the remaining 487 markers were inspected and only one marker (#0014512) in linkage group 10 was found to be problematic by lying outside the alignment zone indicated by the other markers. It was noted that some constraints generated by the remaining 486 markers were superfluous because they were located within areas already covered by other constraints, and these were removed. Table 2 shows the distribution of the number of acceptable bridge markers and resulting non-redundant constraints among the linkage groups.

Table 2: Number of accepted bridge markers per linkage group, and the resulting number of constraints.

Linkage group	1	2	3	4	5	6	7	8	9	10	11	12	total
Accepted markers	43	30	25	48	52	43	30	22	39	54	64	36	486
# northwest constraints	12	10	6	12	3	7	10	8	9	9	11	5	102
# southeast constraints	12	9	7	14	4	9	8	10	12	9	8	4	106
# unconstrained bins	9595	7760	8000	9555	6006	5032	7469	6930	7056	7242	5762	3672	84079
Search space remaining	16%	20%	20%	19%	25%	23%	31%	26%	18%	15%	18%	31%	21%

If a linkage group contains M maternal and P paternal bins, then "# unconstrained bins" = M*P is the unconstrained number of bin combinations. "Search space remaining" is the percentage of valid bin-combinations remaining after application of the constraints.

Figure 2 provides a graphical illustration of the accepted markers and the resulting constraints for linkage group 1. The gray areas in this figure represent invalid

combinations (for bridge markers) of a maternal (X-axis) and a paternal (Y-axis) bin while the white area is the "alignment zone" to which mapping of bridge markers on linkage group one will be constrained. The markers used to obtain the constraints are indicated in the figure by black dots, lines and boxes: Dots for markers linking a single maternal bin to a single paternal bin, vertical lines for markers linking one maternal bin to a range of paternal bins, horizontal lines for markers linking a range of maternal bins to a single paternal bin and boxes for markers linking a range of maternal bins to a range of paternal bins.

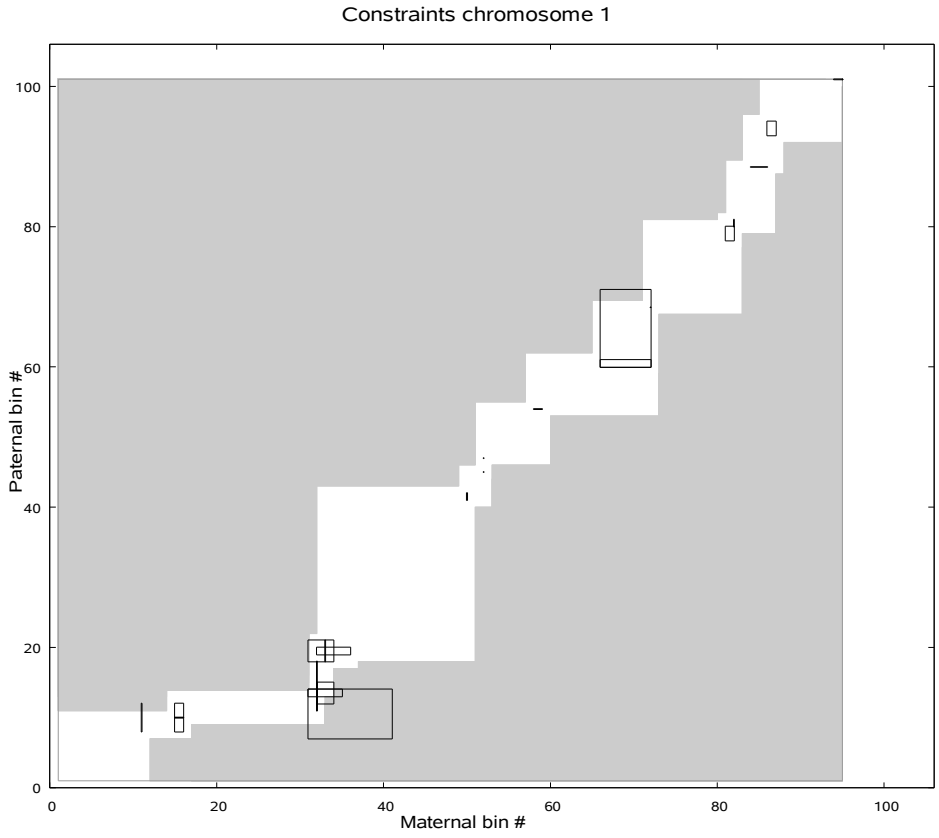


Figure 2: Markers and resulting constraints for linkage group 1. X and Y axes represent the maternal and paternal bins respectively, and the rectangular area from coordinate (1,1) to (95,101) represents the number of possible combinations of a maternal and a paternal bin within this linkage group. Black dots, lines and boxes represent the positions of good quality bridge markers. From these markers the north-west and south-east constraints are obtained. Application of these constraints results in two "forbidden zones" (gray) where no valid bridge marker can be placed, and an "alignment zone" (white area) where all valid bridge markers should be located. Some of the good markers (boxes and lines) overlap the gray areas because there are other markers in that area leading to tighter constraints.

Conflicting north-west and south-east constraints would result in overlapping gray areas and thereby an interrupted “alignment zone”, but this is not the case. It is evident that map integration is still fundamentally incomplete. If full integration were achieved, then a narrow alignment zone should result in all places. Similar graphs can be drawn for other linkage groups as well (data not shown).

Using Binmap+ to map markers

Binmap+ was used to retrofit all markers of the high density genetic map of potato (van Os et al. 2006). Of a total of 9724 markers, 7146 markers passed both rejection criteria (LOD score >6 and apparent error rate <5%). Of the 2578 failing markers only 5 were rejected based on LOD score alone, suggesting a large proportion of missing values for these markers. This was confirmed by inspection of the marker scores. Of the 1690 rejected parental markers, 744 (44.0%) exhibited an apparent error rate of 12.5% or more. Of the 888 rejected bridge markers, 340 (38.3%) exhibited an apparent error rate of 12.5% or more in one or both of the parents. As will be discussed, a high apparent error rate can be indicative of either severe marking scoring errors, incorrect segregation type assignment or homoplasy. Figures 3 and 4 show some LOD(rp,rq) graphs (chapter 2 of this thesis) for examples of good and rejected markers

Of the bridge markers previously assigned to linkage group 5, 42 (Table 3) are now assigned to a different linkage group. The original Bridgemap program (van Os, unpublished) synthesized bridge bin signatures in four different phase combinations as dominantly scored AFLP markers, obtaining marker present and marker absent scores. All of these markers were previously assigned to a particular area in linkage group 5, where the combinations of a maternal bin signature with a paternal bin signature were such that a continuous string of marker present scores would be recorded. In addition, the original Bridgemap program used the same linkage estimator for these dominantly scored AFLP bridge markers as for parental segregation type markers, thereby effectively grossly overestimating the LOD contribution of progeny where a marker presence score in the AFLP bridge markers was matched by a marker presence score in the synthetic bridge bin signatures, while simultaneously underestimating the negative LOD contribution where a marker scores conflict with scores in the synthetic bridge bin signatures. As a consequence, this bin-combination has attracted many markers scored, with some errors, from other loci, in particular near the end of RH chromosome 12, where segregation is very skewed.

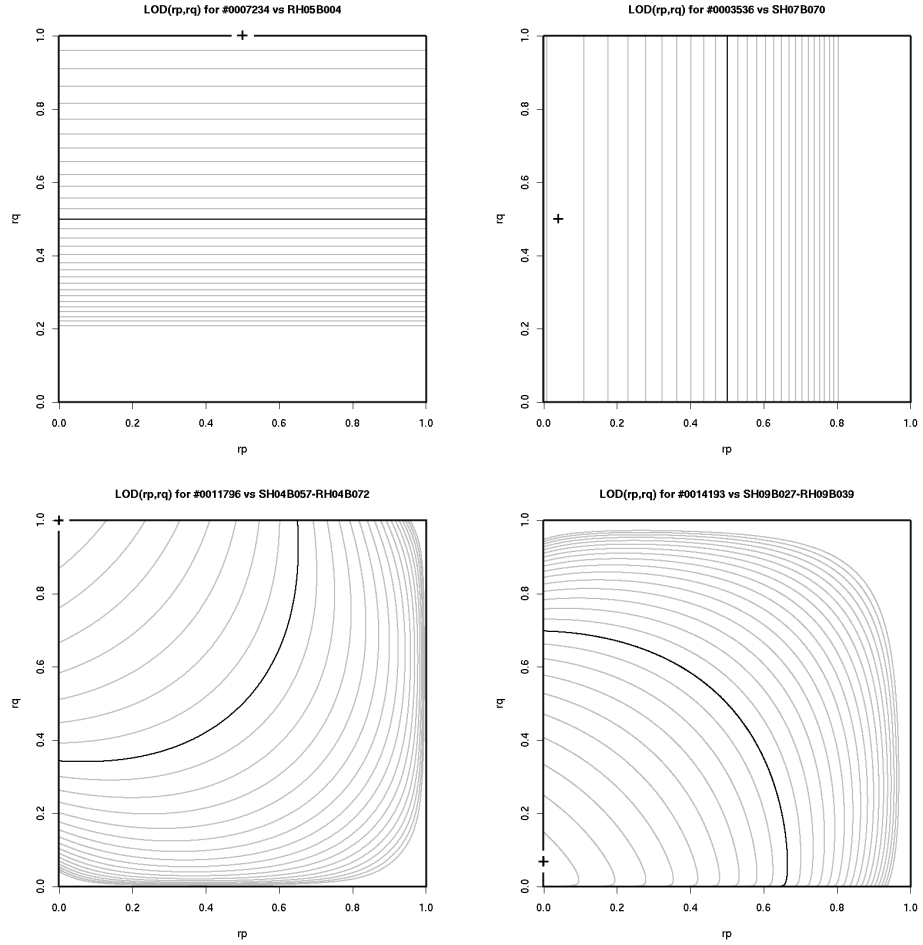


Figure 3: LOD(r_p,r_q) graphs of four accepted markers (passing the LOD score and apparent error rate criteria). The axes in these graphs represent sex-specific recombination frequencies (in this case recombination frequency should be read as apparent error rates between the marker and the bin). The 3-D shape of the LOD(r_p,r_q) graph is visualized through concentric equimetric LOD contours at three LOD unit intervals (n.b. a 1000-fold difference in likelihood). The widest contours indicate where LOD(r_p,r_q) equals minus 48, while black contours indicate where LOD(r_p,r_q) equals zero. The apexes of the graphs are indicated by a "+" symbol. The top two graphs represent markers segregating from a single parent, and in such cases LOD(r_p,r_q) becomes a function of r_q (top left, paternal marker) or r_p (top right, maternal marker) only, and although the apexes in these graphs are still represented by a "+" symbol at r_p=0.5 (left) or r_q=0.5 (right), in reality r_p (left) and r_q (right) are indeterminate. The maximum likelihood linkage estimates (indicated by the "+" at the apexes of the graphs) are: LOD(indeterminate,1)=38.23 for marker #0007234, LOD(0.040,indeterminate)=28.79 for marker #0003536, LOD(0,1)=24.16 for marker #0011796 and LOD(0,0.031)=29.17 for marker #0014193.

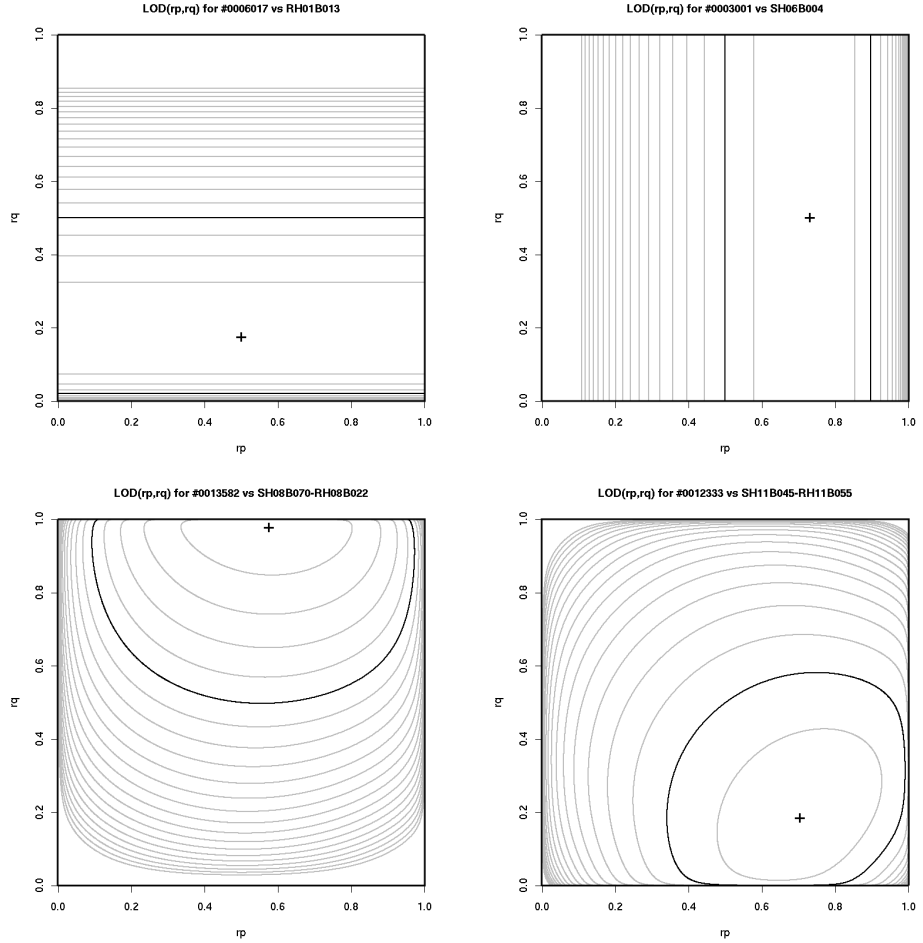


Figure 4: LOD(rp,rq) graphs of four rejected markers. The axes in these graphs represent sex-specific recombination frequencies (in this case recombination frequency should be read as apparent error rates between the marker and the bin). The 3-D shape of the LOD(r_p,r_q) graph is visualized through concentric equimetric LOD contours at three LOD unit intervals (n.b. a 1000-fold difference in likelihood). The widest contours indicate where LOD(r_p,r_q) equals minus 48, while black contours indicate where LOD(r_p,r_q) equals zero. The apexes of the graphs are indicated by a "+" symbol. The top two graphs represent markers segregating from a single parent, and in such cases LOD(r_p,r_q) becomes a function of r_q (top left, paternal marker) or r_p (top right, maternal marker) only, and although the apexes in these graphs are still represented by a "+" symbol at r_p=0.5 (left) or r_q=0.5 (right), in reality r_p (left) and r_q (right) are indeterminate. The maximum likelihood linkage estimates (indicated by the "+" at the apexes of the graphs) are: LOD(indeterminate,0.18)=11.96 for marker #0006017, LOD(0.73,indeterminate)=5.51 for marker #0003001, LOD(0.57,0.98)=14.54 for marker #0013582 and LOD(0.70,0.18)=4.81 for marker #0012333. The markers on the left-hand side fail the selection based on apparent error rate, but pass the LOD selection. The markers on the right-hand side fail both tests.

Table 3: Bridge markers relocated from linkage group 5 to other linkage groups.

		Chromosome 5		New		Alternative	
Marker ID	Marker name	old LOD	new LOD	LG	LOD	LG	LOD
#0011981	CAAGMCAC_146.3H	18.93	3.53	SH07-RH07	9.41	SH04-RH07	13.51
#0011987	PCT/MGA_158.7H	17.73	2.31	SH09-RH09	5.71	SH02-SH09	7.68
#0011991	PTG/MAAT_219.6H	15.48	2.61	SH10-RH10	6.76	SH10-RH01	8.25
#0012001	EAAGMCAT_422.0H	19.6	2.7	SH07-RH07	4.5	SH10-RH01	8.38
#0012004	EAAAMACG_144.2H	15.49	2.36	SH04-RH04	4.29	RH12	5.6
#0012045	EAAGMCTC_198.6H	27.17	10.08	SH01-RH01	10.77	SH05-RH01	12.15
#0012056	EATGMCAG_38H	24.17	5.77	SH08-RH08	8.8	none	
#0012077	EAGGMAGT_16H	18.7	5.05	SH08-RH08	8.76	SH09-RH08	23.75
#0012079	CACAMCTT_497.9H	18.67	3.62	SH01-RH01	5.49	RH06-RH12	7.16
#0012096	PTG/MATG_221.8H	16.19	2.11	SH01-RH01	5.2	RH12	7.58
#0012198	EAGTMCAT_2H	31.48	10.34	SH12-RH12	12.31	RH12	16.04
#0012215	EAACMCAA_6H	27.73	7.2	SH01-RH01	10.2	RH01-RH12	12.56
#0012220	EACAMCGA_88.6H	27.66	8.03	SH01-RH01	9.21	RH12	11.72
#0012225	EATGMCTC_299	27.17	6.13	SH12-RH12	13.05	RH12	19.51
#0012240	EACGMATC_19H	25.41	7.27	SH12-RH12	9.99	RH12	15.52
#0012245	EAGTMCAC_160	25.22	5.09	SH01-RH01	5.09	RH12	9.64
#0012250	EACAMCAG_244.0H	24.72	3.08	SH07-RH07	4.65	RH12	10.46
#0012255	PAG/MAAG_172.3H	23.58	4.61	SH03-RH03	9.9	RH12	13.76
#0012265	EACTMCAT_233.4H	23.09	5.17	SH12-RH12	7.35	RH12	13.13
#0012275	EACAMCTG_585.0H	22.51	3.03	SH07-RH07	7.37	RH12	9.65
#0012280	EAGTMACC_21H	22.5	4.75	SH01-RH01	6.06	RH12	9.52
#0012285	EAGAMAGG_32H	22.28	4.26	SH08-RH08	4.79	RH12	7.26
#0012290	EAGAMCAT_7H	21.18	4.04	SH10-RH10	8.26	SH05-RH10	8.76
#0012295	EAAGMCGT_3H	20.94	4.59	SH12-RH12	5.61	SH05-SH12	9.06
#0012300	EACAMAGG_17H	20.64	4.33	SH12-RH12	4.41	RH12	9.63
#0012305	EATCMCTA_164.2H	20.21	3.83	SH09-RH09	7.93	RH02-RH06	10.35
#0012310	EACTMCTA_213.9H	20.13	5.02	SH01-RH01	6.28	RH12	8.66
#0012316	EAAGMCTC_5H	19.76	4.03	SH01-RH01	7.07	SH02-RH01	11.01
#0012321	EACAMCTG_77.5H	19.44	2.89	SH12-RH12	7.68	SH12-RH04	10.1
#0012326	EATGMCTA_182	19.44	4.38	SH01-RH01	5.33	SH05	7.49
#0012327	EAAGMCGA_171.7H	19.38	2.62	SH11-RH11	3.8	SH05	7.26
#0012329	EAACMCAC_139.8H	18.7	1.38	SH10-RH10	2.83	RH12	8.98
#0012331	EAGAMACC_34H	17.96	3.43	SH11-RH11	7.25	RH08-RH11	13.54
#0012333	EAACMCCA_562.2H	17.24	2.09	SH11-RH11	4.81	RH12	8.63
#0012335	EAAGMCGA_139.2H	16.61	2.74	SH11-RH11	5.02	SH05	6.77
#0012337	EACAMCTT_219.9H	16.49	2.5	SH04-RH04	6.75	RH04-RH08	9.62
#0012342	EATGMCTA_155	16.25	3.26	SH04-RH04	4.49	SH03-SH04	7.2
#0012347	EATGMCAC_26H	16	3.14	SH10-RH10	4.95	RH12	8.5
#0012352	EAGAMCAC_4H	15.42	1.11	SH01-RH01	7.25	SH01-RH04	10.65
#0012529	EAGGMACA_22H	22.77	4.74	SH12-RH12	7.07	SH10-SH12	13.21
#0012532	EACTMCAT_103.9H	22.28	5.96	SH07-RH07	6.07	RH07-RH09	9.11

Bridge markers originally assigned to Linkage Group (LG) 5 (light gray column) now assigned to a different LG ("New" column, shown against a white background). "Alternative" (dark gray) shows the highest scoring alternative mode of inheritance. The "new LOD" score in the "Chromosome 5" column represents the highest scoring position on LG 5 (computed using our universal linkage estimator), not necessarily at the same position to which the marker was originally assigned. "Old LOD" and "new LOD" differ because a different genetic model was used by the original BRIDGEMAP program. Names of the markers where the highest scoring alternative mode of inheritance is compatible with the parental scores of an AFLP bridge marker are highlighted.

Using Homap+ to map rejected markers

The 2578 previously rejected markers were re-mapped using Homap+. Because it was unknown if the parental genotypes were correct (and hence the segregation type of the marker), this was done irrespective of the parental genotype, effectively testing a multitude of possible modes of inheritance. Table 4 summarizes the difference in LOD score between original assignment and highest scoring possible, categorized by type and recombination frequency for all rejected markers. From this table we read that there is a continuous distribution of LOD score differences between original and alternative modes of inheritance, and that for some 374 of the 2578 tested markers (14.5%) the difference is larger than 6, indicating that the alternative mode of inheritance is at least 1,000,000 times more likely. Noteworthy detail is that of these 374 markers, 305 (81.5%) exhibited, under the original mode of inheritance, a recombination frequency (or apparent error rate) >=0.125 in one or both parents.

Table 4: Rejected markers classified by type (maternal, paternal and bridge), recombination frequency and the difference (d_L) between highest scoring mode of inheritance and original bin assignments. If d_L=0, then the original assignment was the best possible.

	LOD difference (d_L) between highest alternative and original score								
Rejected marker type	d_L=0	0<d_L, d_L<=1	1<d_L, d_L<=2	2<d_L, d_L<=3	3<d_L, d_L<=4	4<d_L, d_L<=5	5<d_L, d_L<=6	6<d_L	Total
Maternal, rp<0.125	207	64	73	63	50	31	24	20	532
	(39%)	(12%)	(14%)	(12%)	(9.4%)	(5.8%)	(4.5%)	(3.8%)	(100%)
Maternal, rp>=0.125	32	41	37	29	35	33	29	139	375
	(8.5%)	(11%)	(9.9%)	(7.7%)	(9.3%)	(8.8%)	(7.7%)	(37%)	(100%)
Paternal, rq<0.125	173	55	42	37	26	18	19	44	414
	(42%)	(13%)	(10%)	(8.9%)	(6.3%)	(4.3%)	(4.6%)	(11%)	(100%)
Paternal, rq>=0.125	65	22	26	38	27	33	28	130	369
	(18%)	(6.0%)	(7.0%)	(10%)	(7.3%)	(8.9%)	(7.6%)	(35%)	(100%)
Bridge, r_p and r_q <0.125	530	1	5	3	3	1	0	5	548
	(97%)	(0.2%)	(0.9%)	(0.6%)	(0.6%)	(0.2%)	(0%)	(0.9%)	(100%)
Bridge, r_p or r_q >=0.125	200	16	18	26	25	9	10	36	340
	(59%)	(4.7%)	(5.3%)	(7.6%)	(7.4%)	(2.6%)	(2.9%)	(11%)	(100%)
Total:	1207	199	201	196	166	125	110	374	2578

Figures 5a and 5b show sample results of calculating maximum attainable LOD scores for alternative modes of inheritance. In these figures the maximum LOD scores for assignment of a marker to each chromosome or chromosome combination is shown. Marker #0006017 remains assigned to paternal (RH) chromosome 1, with the best alternative (a bin-combination on paternal chromosome 1 and maternal chromosome 11

denoted SH11-RH01) scoring 11.96-9.49=2.47 LOD units lower (295 times less likely). Marker #0003001 was assigned to maternal chromosome 6 (SH06), but the highest alternative (SH06-RH07) scores 2.19 LOD units higher (155 times more likely). The alternative (SH01-RH08) for marker #0013582 scores 14.2 LOD units higher than the original assignment.

Marker #0006017

	SH01	SH02	SH03	SH04	SH05	SH06	SH07	SH08	SH09	SH10	SH11	SH12	
SH01		1.10	1.50	1.27	1.95	1.44	1.45	0.90	1.20	1.10	3.60	1.27	
SH02			1.02	0.81	1.55	0.76	0.68	0.59	0.82	0.67	3.21	1.00	
SH03				1.37	2.35	1.61	1.30	1.05	1.40	1.43	4.38	1.58	
SH04					1.93	1.19	0.94	0.83	1.01	0.77	2.96	1.01	**Homoplasic**
SH05						1.71	1.64	1.44	1.49	1.51	3.92	1.56	**fragments**
SH06							0.89	0.78	1.19	1.05	3.05	0.98	**segregating from**
SH07								0.42	0.92	1.08	3.22	0.72	**two different**
SH08									0.77	0.54	3.39	0.69	**maternal (SH)**
SH09										0.87	3.71	0.97	**chromosomes**
SH10											2.97	0.91	
SH11												3.62	
SH12													

	RH01	RH02	RH03	RH04	RH05	RH06	RH07	RH08	RH09	RH10	RH11	RH12	
RH01		8.96	7.82	7.83	7.50	7.37	8.22	7.71	7.42	7.62	9.77	7.44	
RH02			1.25	1.92	1.27	1.93	1.05	1.74	0.79	1.82	2.58	4.01	
RH03				2.03	1.31	2.20	1.31	2.44	1.06	1.78	3.05	4.10	
RH04					1.98	2.35	1.55	2.30	1.44	2.01	3.58	4.98	**Homoplasic**
RH05						2.13	0.94	1.72	0.74	1.50	2.25	4.65	**fragments**
RH06							1.45	2.47	1.41	1.94	3.63	4.19	**segregating from**
RH07								1.61	0.57	1.18	2.15	4.10	**two different**
RH08									1.48	2.12	3.00	4.02	**paternal (RH)**
RH09										1.10	2.23	3.76	**chromosomes**
RH10											2.49	4.62	
RH11												5.11	
RH12													

	RH01	RH02	RH03	RH04	RH05	RH06	RH07	RH08	RH09	RH10	RH11	RH12	
SH01		1.85	2.09	2.38	1.58	1.87	1.27	2.20	0.89	1.68	2.91	4.22	
SH02	7.09		1.37	1.69	0.91	1.60	0.81	1.59	0.43	1.08	2.24	3.78	
SH03	7.75	1.42		2.21	1.43	2.09	1.29	2.44	1.13	1.92	2.69	4.17	
SH04	7.45	1.01	2.07		1.06	1.55	0.96	1.89	0.66	1.49	2.55	4.28	**Homoplasic**
SH05	7.52	2.18	2.00	2.55		2.15	1.87	2.25	1.59	1.98	3.73	4.70	**fragments**
SH06	8.07	1.08	1.36	2.04	1.14		1.00	2.14	0.64	1.40	2.08	4.19	**segregating from**
SH07	7.70	0.72	1.10	1.87	0.94	1.43		1.87	0.50	1.18	2.44	4.35	**a maternal and a**
SH08	7.08	0.61	1.07	1.46	0.74	1.35	0.69		0.29	0.99	2.27	3.76	**paternal**
SH09	7.30	1.05	1.37	1.55	1.50	1.70	0.83	2.56		1.26	2.34	3.82	**chromosome**
SH10	7.40	1.07	1.47	1.63	1.05	1.49	0.81	2.06	0.52		2.38	4.20	
SH11	9.49	3.15	3.51	4.45	3.49	4.03	3.21	3.74	2.88	3.75		5.54	
SH12	7.72	0.94	1.10	1.61	0.91	1.63	0.87	1.87	0.61	1.14	2.76		

LG01	LG02	LG03	LG04	LG05	LG06	LG07	LG08	LG09	LG10	LG11	LG12	
1.05	0.36	1.05	0.82	4.29	0.36	0.36	0.74	0.59	1.23	2.66	0.73	**maternal marker**
11.96	1.43	0.26	1.65	0.36	1.23	0.74	2.54	0.26	0.66	1.05	6.35	**paternal marker**
7.34	0.95	2.26	2.23	1.67	1.77	0.84	1.46	0.72	1.40	4.38	5.42	**bridge marker**

Figures 5a and 5b (next page): LOD scores for alternative modes of inheritance of three sample markers. For each marker LOD scores shown in the top three blocks were computed under the assumption that the observed marker segregation patterns are caused by homoplasic AFLP fragments. The bottom three rows show LOD scores under assumption that the marker is non-homoplasic. For each chromosome or chromosome combination, the maximum LOD score attained by the marker is shown. LOD scores in bold type indicate the original map location of the marker, while LOD scores shown against a gray background indicate the highest scoring alternative mode of inheritance.

Figure 5b: for description see figure 5a.

Discussion

Implementation

Depending on settings and the number of markers submitted, Binmap+ and Homap+ can take a long time to execute: From seconds for a single marker using Binmap+ to days for 1000's of markers using Homap+. As keeping a browser window open for days on end is impractical, a web-based front-end with an e-mail based back-end is used. After a user submits data on a web-page, a PERL script run from the web-server checks the submitted data and, if correct, enqueues it for subsequent processing, or discards it otherwise. A second PERL script, working as a wrapper around a C-language computational core, regularly checks the queue for new entries, reads and transforms (stripping of whitespace, comments and names and with marker scores normalized) the data to a compact format that can be easily read by the compiled C-language program responsible for actual processing. After completion, output of the C-language program is retrieved again by the second PERL script, combined with the original input and dispatched to the user as an attachment to an e-mail. As most of the methodology used to map bridge markers is identical to the methodology used to map homoplasic fragments, the only actual difference between the Binmap+ and the Homap+ program lies in the web-based front-end, setting optional switches causing the subsequent programs to behave differently. Splitting of the programs into a multiple parts has several advantages: Separating the web-based front-end from actual processing allows processing to be easily scaled to use multiprocessing if necessary. Use of program in a language more suitable to text interpretation (PERL) as a wrapper around a program in a language more suitable to data processing and fast computations greatly simplifies implementation. Without the C-language computational core (using only PERL), processing takes approximately 20 times longer.

Use of the UMPLE

Linkage analysis in full sib families descending from non-inbred parents is generally considered more difficult than linkage analysis in an F2, BC1 or RIL progeny derived from inbred lines (Maliepaard et al. 1997). One complicating factor, ambiguous marker scores, can arise as a consequence of failure to observe a clear difference between marker alleles in part of the progeny (e.g. co-dominantly scored AFLP markers (Jansen et al. 2001)), as a result of a material or handling error affecting part of the progeny (e.g. some fragments running off an electrophoresis gel) and now also as a result of bridge bin signature synthesis as performed by the Binmap+ and Homap+ programs. To our knowledge, the Universal Maximum Likelihood Pairwise Linkage Estimator (UMLPLE)

(Chapter 3 of this thesis) is the only estimator of linkage intensities between marker pairs that can deal with these (arbitrary) types of ambiguity, hence its application here.
The UMLPLE calculates sex specific recombination rates, which in the context of a framework (bin) map can be reinterpreted as an apparent error rate. Rosa et al. (2002) and Cartwright et al. (2007b) also offer models to estimate error rates from data, but these models are more complicated and do not take differences in marker quality into account. We, however, assume that markers (bin signatures) in the framework (bin) map are essentially error-free. Therefore all errors occur in the retrofitted markers, manifesting themselves as a non-zero recombination rates. Although our assumption may be overly optimistic, the resultant apparent error rates are generally indicative of issues with individual markers; some markers were re-scored (data not shown), producing better LOD scores and lower apparent error rates. Use of SMOOTH (van Os et al. 2005a) during original map construction ensured that every bin signature is supported by multiple marker data points, either directly or through neighboring bin signatures, and so far we have not detected any cases where a systematic and/or substantial number of marker scores contradict a specific score in the bin signatures. We will, however, continue to monitor this issue, and take corrective action if and when necessary

Improvement of the 10000 marker map of potato

Underlying the high density genetic map of potato (van Os et al. 2006) is a pair of non-integrated scaffold maps called bin maps. The scaffold bin maps were created using a multi-stage procedure, with the actual markers only placed onto the bin map in the final phase, using bin signatures as virtual markers. Although the method produces a solid scaffold, there are some issues with the resulting map:

1. Bin-signatures are based on parental segregation type AFLP markers (i.e. <a0x00> or <00xa0> segregation type) only, ignoring bridge markers (<a0xa0> segregation type), primarily because this type of marker is less informative for bin map construction.
2. Bin-signatures were only created for bins actually containing markers, and all markers were assigned to the best fitting existing bin signature (or existing bin combination), never to a better fitting position next to an existing bin signature.
3. Bridge markers were placed onto bin-combinations by synthesizing dominantly scored AFLP-like bridge bin signatures, leading to a loss of information with respect to the full classification of alleles that the bin signatures generally allow.
4. The same linkage estimator was used for bridge and parental markers, ignoring the fact that for the dominant bridge markers the different combinations of marker phenotypes (in this case AFLP fragment absence or presence) contribute differently

to the linkage estimate and LOD score.

5. For any linkage group all combinations of a maternal and a paternal bin were tested, even if such combinations are clearly in conflict with reasonable expectations on the alignment of these co-linear maps.

The first issue is not addressed here, and, given the reduced information content of dominantly scored AFLP bridge markers, it seems unlikely that inclusion would lead to a dramatic improvement in the bin-map. The second issue was resolved by postulation of empty-bin signatures. Both issues three and four were resolved by using the UMLPLE. Comparing the bin-ranges to which bridge markers were assigned previously with their current bin-ranges indicates (data not shown) that resolution of issue three result in a dramatic shrinkage of bin-ranges for some markers. The explanation we can offer for the occasionally large displacements of markers is that this is caused by the previous use of an inappropriate distance estimator in combination with a non-perfect match of the marker to the bin to which it was assigned. The set of markers from linkage group five (Table 3) now finding a position with a higher LOD score on a different linkage group are an extreme example: Originally they were assigned in a single linkage phase to an area on chromosome 5 where the combination of parental bin signatures results in a dominantly scored synthetic AFLP bridge signature entirely consisting of “b-” (band present) scores, and by using the inappropriate maximum likelihood estimator the significance of coinciding “b-” scores in marker and synthetic bridge bin signatures was grossly exaggerated at the cost of more informative coinciding “aa” scores. For markers that do find a perfectly matching (set of) bins this is not an issue; though the old LOD score in itself is meaningless, it can find no better position than a perfectly matching bridge bin signature, and this explains the observation that only a fraction of the markers is affected, and that there is no systematic trend in the positional changes.

Use of constraints, though providing far from perfect map integration, resolves issue 5 at least partially, and results in dramatic increase in computational speed. Moreover, the constraints can easily be updated if more or better data becomes available.

The speed advantage is, however, absent when trying to map homoplasic fragments; not only are all bin combinations between a pair of linkage groups tested, so are all linkage group combinations (except homologous linkage groups), leading to approximately 500-fold increase in computational overhead.

So far we have applied the method to the 9724 publicly available AFLP markers, and, as this dataset was entirely dominantly scored, the only ambiguities that our software has dealt with so far were ambiguities in synthetic bridge bin signatures. During scoring of our AFLP data, band intensity variations were frequently observed in the bridge markers, though often with variability that would make accurate scoring difficult (personal

communication Herman van Eck). Re-scoring of at least some of these markers, may improve our map integration, and we currently also have in excess of 1000 new markers of varying type (AFLP, SSR, STS and domain directed profiling) awaiting detailed analysis, so the map presented here should not be considered a definitive version.

Homoplasic AFLP fragments

Assuming absence of significant skewness in segregation, if a dominantly scored AFLP bridge marker is erroneously interpreted as a parental marker, yet assigned to the correct bin in that parent, then the apparent error rate will be approximately 0.25, the same figure as when a parental marker is erroneously interpreted as a bridge marker. We observed many markers with a relatively high apparent error rate (>=0.125) in one or both of the parents, suggesting that some of these might be caused by genotyping errors in the parents or co-incidental co-migration of AFLP markers. We tested if an alternative hypothesis for the mode of fragment inheritance would result in a higher LOD score. Table 4 indicates that a high apparent error rate under the hypothesis of ordinary segregation is a predictor for the existence of a higher scoring alternative mode of inheritance. However, as illustrated in figure 5 for two of the three markers, the contrast (difference) between the LOD scores achieved under different hypotheses may be low, making outright rejection or acceptance of any particular hypothesis dangerous in many cases: To use the LOD score (in casu the LOD contrast between different hypotheses) as a test value we need to know some measure of the number of different hypotheses we are testing, and establishing such a measure is difficult: closely separated bin signatures and bin signature-combinations will be highly correlated, and therefore do not represent completely independent hypotheses, whereas more distant bin signatures may do so, and the parental bin signatures underlying synthetic bridge bin signatures will also be (partially) correlated to the synthetic bridge bin signatures. Ignoring the latter fact and ignoring the fact that a single chromosome may represent multiple partially dependent hypotheses, we arrive at an estimate equal to the number of chromosome-combinations plus individual chromosomes we are testing: (23*24)/2 + 24 = 300 for potato, safely qualifying the difference in LOD scores between two of the three markers in Figures 5a and 5b as insignificant. To reduce the number of alternative hypotheses we are testing we can re-confirm the parental phenotypes, and restrict ourselves to alternative modes of inheritance that are compatible with the observed parental genotypes. The third marker (in Figure 5b) shows a LOD contrast of more than 14 with respect to the original segregation type, and the parental phenotypes are compatible, making the alternative extremely likely. Interpreting a LOD contrast of >=3 with respect with the original segregation type as indicative and a LOD contrast of >=6 as proof we conclude that between 4% and 8% of the markers are affected. In the majority of

the affected markers we have detected, coincidental co-migration is the most likely explanation, and the figure we have obtained is compatible with previously reported figures in other species (Hansen et al. 1998, O'Hanlon and Peakall 2000) for homoplasy (1%-5%). Though it seems unlikely that coincidental comigration has seriously affected our bin-map, it may require some consideration when aligning genetic maps through allele specific co-migrating AFLP fragments (Rouppe van der Voort et al. 1997).

Application to other framework maps

Assuming that a framework map of sufficient quality, with near perfect marker scores, exists, we see no objection in converting such a map to a bin map. For relatively low density maps this could be straightforwardly done by re-scoring the frame-work markers so that all scores are in the appropriate linkage phase, and replacing the names of markers with appropriate bin-names, followed by postulation of empty-bin signatures. For denser maps, containing co-segregating markers, such markers must be combined into consensus bin signatures first.

Though the programs may appear intimately linked to mapping in full-sib mapping populations, the Homap+ program may also find application in other types of mapping populations: by replacement of the separate maternal and paternal bin-maps with a single map of (e.g.) an F2 or BC1 progeny, it may be possible to map homoplasic markers in such maps too.

Chapter 5

Correction for systematic fragment sizing differences observed between different MegaBACE machines, capillaries and fluorescent labels

T.J.A.Borm, J. de Boer, H.J. van Eck and R.G.F. Visser

Abstract

Reproducible sizing of DNA fragments is crucial for many applications ranging from physical mapping to forensic studies. For instance, to identify SSR alleles differing by 1 base-pair with 99.7% confidence, the standard deviation in sizing may not exceed 0.16 bp. During high throughput fingerprinting of approximately 73,000 potato Bacterial Artificial Chromosome (BAC) clones we discovered systematic sizing variations between MegaBACE 1000 machines, capillaries and fluorescent dye labels, and we devised a way to correct for these systematic errors using the band-called fingerprints of 2404 BAC clones containing chloroplast DNA derived inserts as a source of calibration data. After correction the standard deviation in sizing is approximately halved. Though we observed extreme deviations between machines (of approximately 0.4 bp for fragments between 200 and 450 bp in length) putatively caused by differences in electrophoresis temperature, our data suggests that sizing accuracy (standard deviation) can be improved to less than 0.12 bp for fragments up to 430 bp in length, which would imply a confidence better than 99.997% for scoring SSR alleles differing by one base-pair.

Introduction

Fingerprinting the thousands of Bacterial Artificial Chromosome ("BAC", Shizuya et al. 1992) clones necessary to construct a genome-wide, Finger-Printed Contigs (FPC, Soderlund et al. 2000) based physical map of any eukaryotic species requires a high throughput fingerprinting method. Capillary Electrophoresis (CE) offers several advantages over slab gel electrophoresis (SGE) to separate fingerprint fragments: It does not require (manual) casting and loading of slab gels and is considered more precise and reproducible (Wenz et al. 1998, Lazaruk et al. 1998, Rosenblum et al. 1997, Nelson et al. 2005, Koumi et al. 2004). Furthermore, capillary sequencers like the ABI3730 or MegaBACE 1000 are highly automated. Capillary electrophoresis does, however, require a significant investment in capital and training of personnel. For one-off projects or when normal work flow is optimized for other electrophoresis systems, as well as for projects requiring the quality that only labs employing highly trained personnel using standardized, optimized, protocols can offer, outsourcing offers a viable alternative. We have recently reported (Chapter 2 of this thesis) the construction and characterization of an AFLP™ fingerprinted BAC library of potato. Individual BAC growth, DNA extraction, and AFLP were all performed in-house, while electrophoresis was outsourced to Keygene NV (Wageningen, the Netherlands), using a MegaBACE 1000 capillary sequencer in combination with proprietary BAC-Xtractor software (Srinivasan et al. 2003) to produce band-called fingerprints precise to 0.1 basepairs (personal communication Taco Jesse, Keygene NV). During the course of the project a sample of BACs was fingerprinted in duplo, and it was discovered that there was a systematic difference in sizing between some runs. Close inspection, involving the disassembly of the raw electropherograms ("RSD files") produced by the MegaBACE machines, revealed that two different MegaBACE machines had been used. Because sizing difference appeared systematic and the fingerprinting of the 73,344 BAC clones was approximately 70% complete, and because we surmised that we had sufficient calibration data in the form of fingerprints of BACs containing chloroplast-DNA derived inserts, we decided that we would attempt to apply an error correction in software rather than to repeat fingerprinting. Here we describe the analysis of machine, capillary position and label dependent sizing discrepancies using chloroplast derived BAC fingerprints, and the effect that correction has on resulting physical maps computed using identical settings.

Our objectives are: A) To identify some types of systematic band sizing errors, and B) To apply an error correction to the fingerprints based on these estimates. We make use of the band-called fingerprints of a large set of clones previously identified as containing chloroplast DNA derived inserts (Chapter 2 of this thesis). These fingerprints will be used

as a calibration standard, albeit often with missing bands because the BAC cloning process does not always result in the complete chloroplast genome being inserted into the BAC vector.

Methods

Data collection and identification of usable fingerprint fragments

For every fingerprint the information on the MegaBACE machine identity (#220 and #300) was obtained from the data file (RSD files) describing the raw electropherogram. These RSD files have been processed (sizing and band-calling) using KeyGENE proprietary tools (Xpose and BAC-Xtractor), as described by Srinivasan, 2003. This resulted in one "bands file" per fingerprint containing a list of AFLP fragment lengths. From these "bands files" the DNA fragment lengths were collected for further analysis of size variation. Capillary position data are retrieved from the BAC name, which reflects well position in a 384-well micro-titre plate.

As we already visually observed a major MegaBACE machine dependency in the fragment sizing, bands files of fingerprints grouped into the chloroplast contig (Chapter 2 of this thesis) were gathered and sorted according to MegaBACE machine-ID. For each of these two subsets of chloroplast fingerprints a histogram of the number of DNA fragments per 0.1bp interval was obtained, and within these histograms, several peaks, each representing a distinctive AFLP fragment contained in the chloroplast genome, could be observed. Based on visual inspection, band size-intervals corresponding to well-defined peaks in the histograms were defined. Homologous intervals in the machine-ID dependent subsets were visually aligned. Fingerprint fragments were assigned to these intervals (binning), and average fragment sizes were obtained per interval for separate (machine-ID dependent) subsets. Intervals were subsequently filtered to remove:

1. Adjacent intervals with apparently overlapping fingerprint size distributions.
2. Any intervals starting > 670 bp because of expected et-ROX size standard crosstalk.
3. Intervals where the difference in average fragment size between different machine-IDs had a different direction compared to both directly adjacent intervals (e.g. a fragment appearing longer in one machine than the other while adjacent fragments suggest that it should be shorter)

Identification of other sources of residual errors

In lieu of a "golden standard" we used the average size, taken across all fingerprints, of the fragments within each accepted interval to create a "consensus chloroplast fingerprint". Fingerprints were sorted according to machine-ID and fluorescent label, and

average fragment sizes were obtained per interval. From these averages the corresponding consensus chloroplast fingerprint fragment sizes were subtracted to obtain a fragment size dependent fragment size deviation graph for each combination of machine-ID and fluorescent label. These curves were subsequently subtracted from respective individual chloroplast fingerprints to obtain provisionally corrected fingerprints. Next, these fingerprints were sorted according to capillary scan order and average residual fragment size deviations were obtained by subtraction of the respective (machine and label dependent) consensus chloroplast fingerprint fragment sizes and averaging per capillary.

Curve fitting, fragment size correction and evaluation

For each combination of machine-ID and fluorescent label, a fragment size (s) and capillary position (p) dependent function f(p,s) was fitted to the difference between individual fingerprint fragment size and corresponding consensus chloroplast fragment size using the "lm" linear model fitting module of the R statistics package (Ihaka and Gentleman 1996). The function can conveniently be written using vector algebra:

$$f(p,s)=\begin{bmatrix}1 & p & p^2\end{bmatrix}\cdot\begin{bmatrix}C_{00} & C_{01} & C_{02} & C_{03}\\ C_{10} & C_{11} & C_{12} & C_{13}\\ C_{20} & C_{21} & C_{22} & C_{23}\end{bmatrix}\cdot\begin{bmatrix}1\\ s\\ s^2\\ s^3\end{bmatrix}, \qquad \text{eq. 1}$$

with C00..C23 constants estimated from data. A corrected dataset was generated by subtracting the value of f(p,s) from the observed fragment sizes. A reverse-corrected dataset was generated by adding the value of f(p,s) to the observed fragment sizes. An over-corrected dataset was generated by subtracting twice the value of f(p,s) from the observed fragment sizes. Effectiveness of correction was evaluated by comparing standard deviations computed for each consensus chloroplast fingerprint fragment, by visual comparison of some chloroplast fingerprints and by visual comparison of a set of randomly duplicated fingerprints. In addition, all datasets were fully automatically assembled into contigs using the FPC program (Soderlund et al. 2000) with the following settings: Equation 2, tolerance=5 and cut-off=10^{-12}, all other parameters at their default values and after filtering (Chapter 2 of this thesis). Resulting differences in the distribution of contig sizes and the number of clones remaining in the singletons pool were noted.

Results

In total 122 intervals were identified in the histograms of chloroplast fragment lengths. Of these, 17 exhibited unclear separation between adjacent peaks, suggesting overlapping

fragment size distributions. Five exhibited inconsistent deviation between machines (as judged from adjacent peaks on both sides), and three were larger than 665 bp. The remaining 97 intervals were used in the analysis.

Figure 1 shows the average deviation from the consensus chloroplast fingerprint of each of the six combinations of machine-ID (denoted "220" and "300") and fluorescent label (denoted "F", "J" and "N" for FAM, JOE and NED respectively. All fragments larger than 110 bp that were sized on MegaBACE machine 220 are consistently larger than the average, while the same fragments sized on MegaBACE machine 300 were consistently smaller. Below 110 bp the trend is still apparent. Within each machine, although the effect is much smaller and less consistent across all fragment sizes, the trend can be observed that NED-labeled fragments are sized larger than the same FAM-labeled fragment, which are sized larger than the same JOE-labeled fragment.

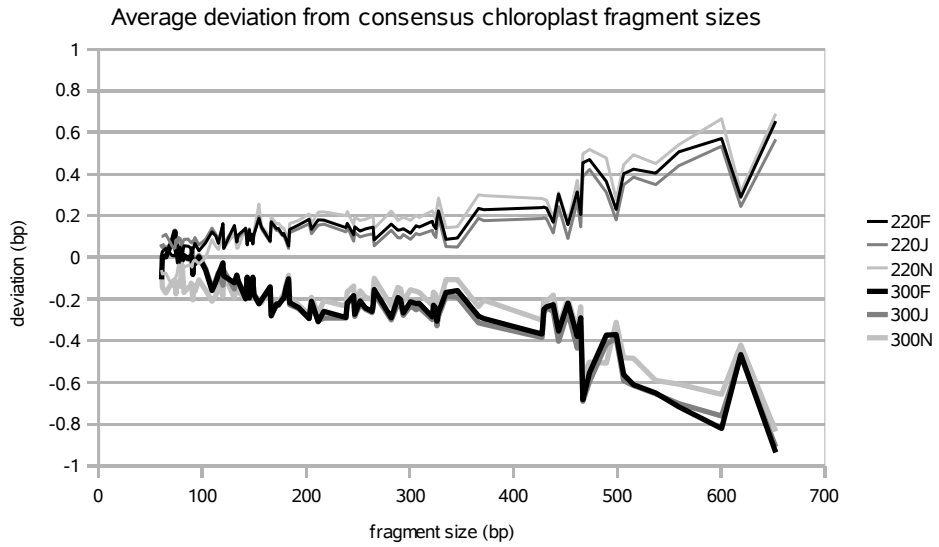


Figure 1: Average deviation per combination of machine-ID and fluorescent label from the consensus chloroplast fragment sizes. Consensus chloroplast fragment sizes were computed as averages (per interval) of fragment lengths across machines and fluorescent labels, and these curves (linearly interpolated between chloroplast fragment sizes) indicate a major (around 0.5 bp) dependency on machine-ID in chloroplast fingerprint fragment sizes and a smaller (around 0.1 bp) dependency on fluorescent label. The imbalance in deviation between both machines is an artifact caused by the fact that fewer fingerprints were produced on machine 300.

Figure 2 shows the average (over all intervals) of the residual deviation per capillary position after provisional correction (subtraction of the machine-ID and fluorescent label

dependent part). The outside lanes in a slab gel often appear to migrate more slowly than the central lanes, leading to a distinctive smile-like pattern, and here a similar pattern can be observed: The distal capillaries in the capillary array appear to result in longer fragment sizes than the central capillaries. It should, however, be noted that the effect observed in capillaries is a small residual effect remaining after sizing with respect to an internal size standard while in slab gel systems the effect is absolute. We have not inspected raw, unsized, electropherograms for individual capillaries, so we do not know if there is any absolute difference in fragment migration velocity between distal and central capillaries in the array.

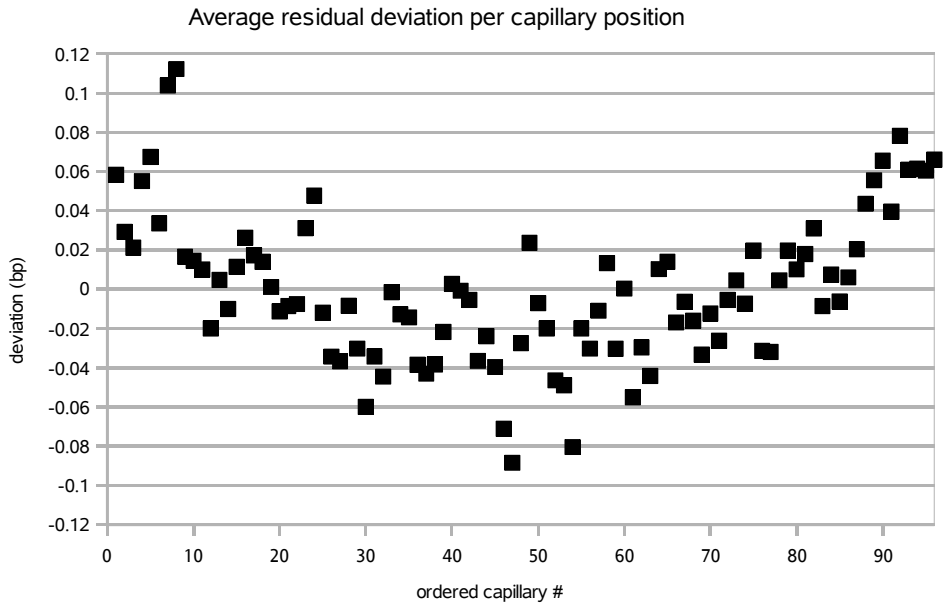


Figure 2: Average residual deviation per capillary position. The residual deviations plotted here are the average, across fragment sizes and per capillary position, of the remainder after subtracting the machine ID and fluorescent label dependent average fragment sizes from the individual fragment sizes, and therefore an indication of the capillary position dependency of fragment sizes. Capillaries are scanned by the MegaBACE hardware in a particular order: A01, B01, C01 ... H01, B02, C02 ... H02, A03 ... F12, G12, H12, and numbered consecutively in this graph: #1 = A01, #8 = H01, #9 = B01 etcetera.

Figure 3 demonstrates the effect on the standard deviation for each consensus chloroplast fingerprint fragment of application of the fitted error correction functions. Overall, correction appears to approximately halve the standard deviation. There is, however, a large machine dependency and therefore the fragment sizes may be better described by separate normal distributions. Within a single machine ID, correction reduces the standard deviation by approximately 20% (detailed data not shown).

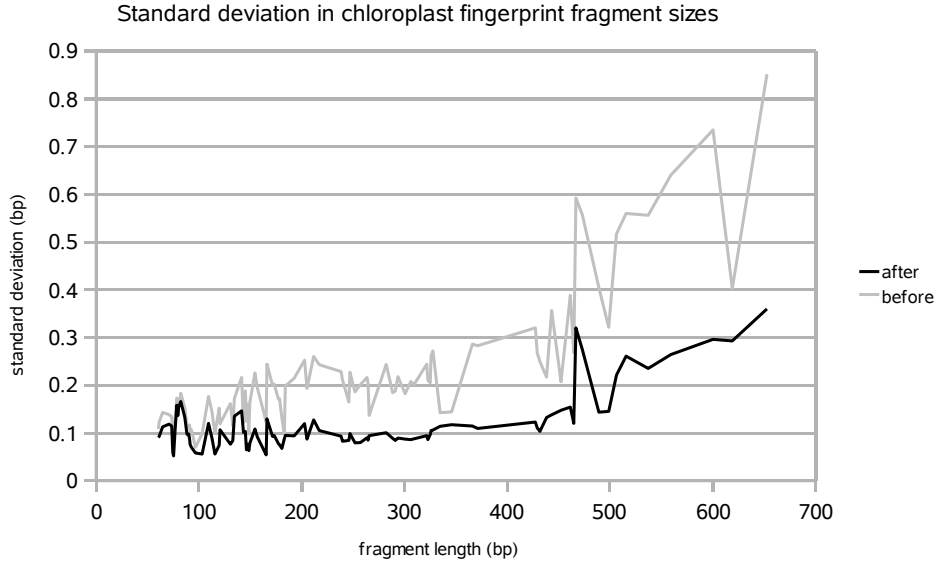


Figure 3: Standard deviation in chloroplast fingerprint fragment sizes across machines, fluorescent labels and capillaries, both before and after correction.

Figure 4 illustrates the effect of reverse, absent and proper correction on a sample of individual fingerprints. The left half of the figure shows the effect on three fingerprints duplicated on both MegaBACE machines. Within each block of six fingerprints, the three pairs of duplicated fingerprints show that:

1. Occasionally bands are missing in duplicate fingerprints.
2. Fragments sized on machine 220 are generally larger than their counterpart on machine 300.
3. Reverse correction increases the difference, making fingerprints appear less similar.
4. Proper correction decreases the difference, making the (pairs) appear more similar.

The right-hand side shows the same for a set of six (non-duplicated) chloroplast fingerprints. Within each block, the difference between fingerprints produced on machine #220 (left triplet) and machine #300 (right triplet) is evident in the uncorrected and reverse corrected fingerprints, and virtually absent in the properly corrected fingerprints. While systematic differences between machines can immediately be observed in Figure 4, label dependent sizing differences that might exist between differently labeled chloroplast fingerprints are smaller and partially obscured by capillary position dependent sizing differences.

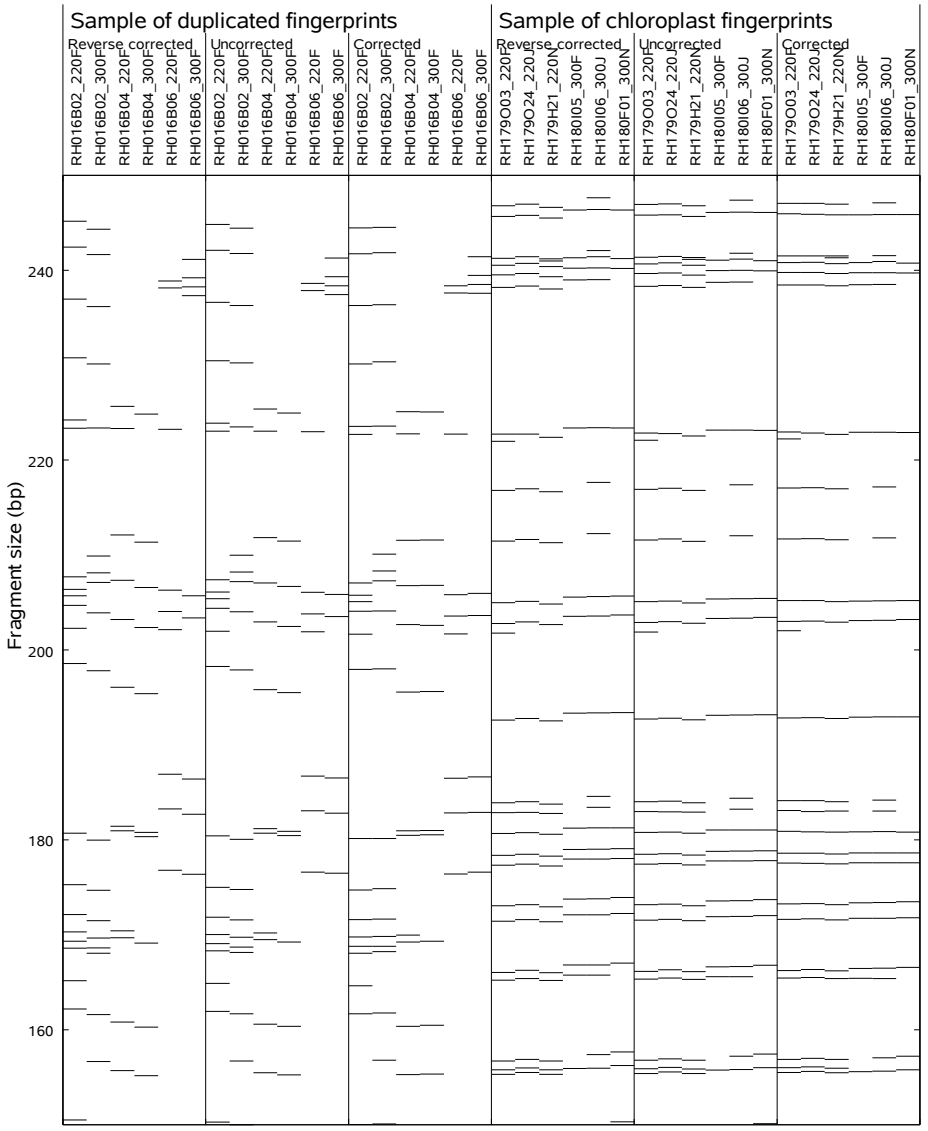


Figure 4: Illustration of the effect of proper and reverse correction of fingerprint fragment sizes on a sample of fingerprints. This pseudo-gel image was synthesized from band-called fingerprints. Note that fingerprint fragments may range from 60 to 900 bp but that only a small region (150 to 250 bp) is shown in this figure to increase clarity. The left half of the figure represents a set of fingerprints for which fingerprinting was duplicated (Chapter 1 of this thesis) using the same label on two different machines, and it should be note that fingerprinting was not entirely reproducible. The right half of this figure shows a set of six chloroplast fingerprints, one for each combination of MegaBACE machine ID and fluorescent label.

After auto-assembly of fingerprints into contigs using FPC, the size distribution of contigs and the size of the singletons pool was noted (Table 1) after application of the "deQer" (re-analysis of contigs with > 5 questionable fingerprints). Proper correction leads to 16.5% fewer singletons, approximately 17% fewer small (2-3 fingerprints) contigs and ~17% more fingerprints in larger (11-100 fingerprints) contigs (excluding the extremely large chloroplast and putative *E.coli* contamination (Chapter 2 of this thesis) contigs). Reverse correction, however, results in 36% more singletons, 72-75% more small contigs and a 47-48% reduction in the number of fingerprints in larger (11-100 fingerprints) contigs. As expected, over-correction, applying the same correction twice, leads to worse results, clearly illustrating that direction and approximate magnitude of the applied corrections are essentially correct. Although the "deQer" has an effect on the contigs themselves, leading to a subtle shift towards smaller contigs, it hardly affects the observed relative differences (data not shown).

Table 1: Size distribution of contigs obtained from FPC using different datasets.

	Reverse-corrected		**Uncorrected**		**Corrected**		**Over-corrected**	
	#ctgs	#FPs	#ctgs	#FPs	#ctgs	#FPs	#ctgs	#FPs
singletons	-	17674	-	12957	-	10817	-	14161
2-3	6781	15824	3935	9190	3280	7709	4273	9983
4-10	3919	22196	3521	21784	3379	21154	3535	21595
11-30	682	10755	1141	17902	1297	20442	1083	16835
31-100	49	1936	152	6530	187	8278	141	5859
101-300	0	0	0	0	0	0	0	0
>300	2	5356	2	5378	2	5341	2	5307
total	11433	73741	8751	73741	8145	73741	9034	73741

In the "reverse-corrected" dataset the correction coefficients are added rather than subtracted from the fragment sizes, in the "over-corrected" datasets the correction coefficient is subtracted twice. Results are shown after removal of any bands below 100bp or above 650 bp from the fingerprints and after execution of the "deQer" (reanalysis of contigs with more than 5 "questionable clones").

Discussion

Although the band size correction reduces the variation in selected chloroplast fingerprint band sizes significantly, this is no guarantee that this will be the case for all fingerprints and all fragments. Indeed, we have discarded five chloroplast fingerprints fragments from analysis because their relative size change between machines was opposite or much smaller than expected based on their immediate context. Application of the size correction

to these fragments exhibiting “abnormal deviations” actually increases the difference. We therefore assume that this will also be the case for some other (non-chloroplast) fingerprint fragments, and that a fraction of the fragments sizing differences will be exaggerated by correction. In the over-corrected dataset the same correction is applied twice, and a deviation of similar magnitude but of different sign (direction) was expected. The physical map computed from the over-corrected dataset is of lesser quality, in particular with more singletons, than the physical map computed from the uncorrected dataset, and this difference may in part be explained by fragments exhibiting “abnormal deviations”. Where correction would already enlarge the size difference between these fragments, over-correction would additionally exaggerate this difference, putatively leading to FPC detecting fewer overlaps between fingerprints and hence more singletons, as has been observed. The fact that FPC auto-assembly of corrected fingerprints leads to fewer singletons and larger contigs, whilst reverse correction or over-correction leads to a deterioration of the resulting physical map is a good indication that on average the fingerprint data are improved by application of the correction, and that only a minority of fragments exhibit abnormal deviations. There is, however, no doubt that the quality of BAC fingerprints would have been better if correction could have been avoided altogether.

We have executed the size correction without reference to an explanatory hypothesis. Other authors have observed sizing differences dependent on fluorescent label, capillary position, denaturing conditions and sizing standard migration abnormalities (e.g. Rosenblum et al. 1997, Poltl et al. 1997, Hahn et al. 2001 and Koumi et al. 2004). We also observed fluorescent label dependent sizing differences. Sizing standard migration abnormalities are unlikely to affect us. These deviations are reproducible and therefore the inferred fragment sizes of our anonymous AFLP fingerprint fragments (though they may be off by several base-pairs) should also be reproducible. Assuming that there were no systematic, machine dependent, errors made in handling, and given that the same chemically denaturing conditions prevailed during electrophoresis (dictated by the commercial ready-made Amersham™ long read matrix used), differences in temperature may offer the most reasonable explanation.

Close inspection of the MegaBACE run parameters as recorded in the raw electropherograms (“RSD files”) revealed that electrophoresis was performed at 44°C. Temperature dependency of DNA migration in CE is a well established fact, both for single stranded (e.g. Kleparnik et al. 1996, Rosenblum et al. 1997, Wenz et al. 1998) and double stranded (e.g. Guttman and Cooke 1991; Guttman 1996) DNA molecules, and several graphs have been produced indicating that for denaturing gels at 44°C there is generally a 0.2 to 0.4 bp per degree Celcius ascent in measured fragment sizes.

Rosenblum et al. (1997) and Wenz et al. (1998) also show that some fragments exhibit a negative temperature coefficient, confirming our observation that some fragments show abnormal migration deviations. The sizing deviations we observe appear more erratic for smaller fragments, which is compatible with the observation that Single Strand Conformational Polymorphisms (SSCPs), are more pronounced in smaller fragments (Kozlowski and Krzyzosiak 2005). The fact that a (moderate) conformational changes or nucleotide composition dependent sizing differences can occur in denaturing gels has also been noted (e.g. Bowling et al. 1991, Glavac and Dean 1993, Konrad and Pentoley 1993, Cordier et al. 1994, Noll et al. 2007). As discussed by Wenz et al. (1998), the dye used to label the size standard (et-ROX) differs in chemical structure from the other dyes (FAM, JOE and NED) and the consequential, putatively temperature dependent, difference in dye – electrophoresis matrix interactions may be responsible for a large part of the observed systematic differences. The evidence that some specific fragments exhibit an abnormal mobility shift can, however, not be ignored, and points to an internal sequence dependency.

It is interesting to note that, in a study using the ABI 3100 capillary sequencer (Sgueglia et al. 2003), even though electrophoresis was performed at an (putatively equilibrated) oven temperature of 60ºC, variation of ambient temperature between 23ºC and 32ºC resulted in sizing differences of more than 1 bp, indicating, in combination with the strong U shape observed in the capillary position dependent sizing (Kuomi et al. 2004) uneven oven temperatures and poor thermostatic control for this type of machine. Though a similar but smaller U shape in the capillary position dependent sizing appears to be present in Koumi's data for the MegaBACE 1000, this is not noticed, possibly because though the data for the ABI 3100 was gathered over 60 runs, only 6 runs were performed on the MegaBACE 1000. If we average MegaBACE data per block of 16 capillaries (data not shown) the U shape is even more evident, and we would like to suggest this is caused by oven temperature variation. If we assume that, on average, for an oven temperature around 44ºC, fragments are sized larger by an estimated 0.3 bp/ºC, our data (Figure 2) suggests that the outside capillaries in the array are ~ 0.35 ºC warmer than the interior capillaries. Unfortunately, we can not verify that temperature variation between machines nor within a machine are the root cause for the machine and capillary position dependent variations as KeyGene is no longer offering the service and has since decommissioned one of the two machines.

The immediate impact of sizing correction on the (preliminary) physical maps we compute should come as no surprise. Using a tolerance of 0.5 bp for detection of overlapping fragments in FPC means that before correction (standard deviation in sizing ~ 0.2 bp) approximately 20% of the fragment overlaps are missed. This situation is further

confounded by the fact that sizing errors are not normally distributed; there is a systematic sizing error of (on average) about 0.4 bp between machines, and this may cause additional fragmentation of contigs.

Conclusion

For precise genotyping and fingerprinting applications, where one relies on a reproducible sizing with respect to a size ladder, a higher (e.g. 60ºC as suggested by Wenz et al. 1998) run-temperature and location of the MegaBACE machine is a temperature controlled environment may be appropriate, whereas for sequencing applications, which are essentially self-sizing, other motivations may play a more important role. We achieve a standard deviation in sizing (albeit only of selected chloroplast fingerprint fragments) of approximately 0.1bp for fragments up to 400bp in length, increasing to 0.35bp for fragments up to 650bp in length, indicating that the advertised precision can be achieved using the MegaBACE 1000 instrument in combination with BAC-Xtractor software. In the higher molecular weight range (>300 bp) this precision compares favorably with previously reported values for the MegaBACE and other capillary instruments (Sgueglia et al. 2003, Koumi et al. 2004). However, even if machine dependency can be discounted (either by using a single machine or putatively using a higher run temperature), label and capillary position dependent sizing correction remains crucial to achieve this.

Chapter 6

Towards a genetically anchored physical map of potato using AFLP Contig Matching

T.J.A.Borm, J. de Boer, T. Jesse, B. Brugmans, J.S. Werij, R.C.B. Hutten, H.J. van Eck and R.G.F. Visser

Abstract

We report on the construction of a first version of a genome wide, genetically anchored, BAC-based physical map of potato. To anchor contigs in this physical map to the recently published ultra dense genetic map of potato, we employed a novel, ultra-efficient, combinatorial anchoring method termed "AFLP Contig matching" that makes use of the fact that the BAC clones in our library were fingerprinted using the same AFLP enzyme combination (EcoRI and MseI) that was used to generate the majority of markers in the ultra-dense genetic map. Screening (with 57 different primer combinations) a set of 90 pools for anchoring we anchor more than 800 contigs. Including the additional effort that was required to convert AFLP fragment mobilities between different electrophoresis systems, we arrive at an average efficiency of requiring less than 8 PCR reactions per anchored contig. In this first version we were able to discern two large contigs (containing 2512 and 2603 fingerprinted BACs) representing a "chloroplast contig" with fingerprints of BACs containing chloroplast DNA derived inserts and an "*E.coli* contig" putatively containing fingerprints of *E.coli* genomic DNA. Around 53,000 BACs were placed in contigs varying in size from 2 to 100 BACs. This corresponds to about 8.2 Genome Equivalents (GE). 1.7 GE of BACs (11,063 BAC fingerprints) found no significant overlap with any other BAC and were therefore classified as singletons.

Introduction

Because of their large cloning capacity, relative ease of construction, maintenance and stability of their insert, Bacterial Artificial Chromosome (BAC, Shizuya et al. 1992) libraries have become a standard tool for various types of genomics research. Applications include chromosome identification and physical mapping of chromosomes through Fluorescent In Situ Hybridization (FISH), purification of centromeric DNA, repeat classification, map based cloning of genes, sequencing and sequence assembly (both locally and genome-wide), structural and comparative genomics and physical mapping (e.g. Mahairas et al. 1999, Siegel et al. 1999, Cardle et al. 2000, Islam-Faridi et al. 2002, Bakker et al. 2003, Luo et al. 2004, Meyers et al. 2004, Osoegawake et al. 2004, Huang et al. 2005, Matsumoto et al. 2005, Kim et al. 2005, Lai et al. 2006, Wicker et al. 2006, Hein et al. 2007, Wei et al. 2007, Zhu et al. 2008). For many of these applications it is crucial to know the genetic position of the BAC clones that are being used, and a variety of screening methods have been used that allow researchers to determine which BAC clones contain a particular genetically mapped DNA fragment. As long as research into a species is restricted to a couple of loci, this "anchoring" of BAC clones to a genetic map can effectively be handled by an ad hoc approach. However, at a certain stage, the question will arise if a scientific community, or even single research group, would be better served if the anchoring would be done, for a single or a few BAC libraries, in a systematic, high throughput manner on a genome-wide scale.

In many BAC based physical maps, collections of BAC clones with DNA inserts derived from overlapping regions of the donor organism's genome are organized into ordered groups called contigs. Although such BAC physical maps have been constructed using other techniques (e.g. Hoheisel et al. 1993, Mozo et al. 1998, 1999, Han et al. 2000), for genome wide physical maps it is currently considered best practice to fingerprint BAC clones individually (Coulson et al. 1986, Ding et al. 1999, 2001, Gregory et al. 1997, Hong et al. 1997, Marra et al. 1997, Zhang et al. 2001, Srinivasan et al. 2003, Luo et al. 2003, Xu et al. 2004, Nelson et al. 2007) and use a fingerprint based clone ordering algorithm (Lander et al. 1988, Sulston et al. 1988, Flibotte et al. 2004) to combine these fingerprints into contigs. The most commonly used computer program for ordering fingerprints into contigs is the Finger Printed Contigs (FPC) program (Soderlund et al. 1997, 2000).

Although some BAC libraries have been screened for genetically mapped markers in a more or less systematic manner without reference to a genome-wide BAC-based physical (contig) map (e.g. Cai et al. 1998, 2001, de Donato et al. 1999), and although genome-wide BAC-based physical maps have been published that have not (yet) been anchored in

a genome-wide fashion (e.g. Katagiri et al. 2005, Han et al 2007, Quiniou et al. 2007), often such efforts are combined. The benefit of anchoring clones from a BAC library whose clones are organized into contigs of a physical map is immediately evident. Instead of just anchoring clones containing a particular genetically mapped DNA fragment, all the clones in a contig are immediately genetically anchored, thereby potentially gaining a foothold into previously genetically uncharted territory of the genome. The anchoring of BAC clones can, however, also be of benefit to physical map construction by suggesting, supporting or advise against mergers between contigs that remained undetected using the fingerprints. In all, there is considerable synergy.

While individual fingerprinting of BAC clones now appears to be the de-facto standard method used to construct genome-wide physical maps, employing highly optimized, labor saving protocols and equipment (e.g. Xu et al. 2004, Nelson et al. 2005), a plethora of different anchoring methods is currently being practiced. If we restrict our examples to recently published genetically anchored physical maps of plants, then Brasica rapa (Mun et al. 2008, STS markers), Sorghum (Klein et al. 2000, AFLP markers), rice (Tao et al. 2001, Chen et al. 2002, using hybridization probes) and Populus (Kelleher et al. 2007, SSRs) can serve as examples. With some marker systems the thousands of BAC clones in a library can be individually screened cost-effectively (e.g hybridization based schemes such as described by Ross et al. 1999, Romanov et al. 2003, Gardiner et al. 2004 and Ren et al. 2003 - although the marker probes may have been pooled to increase efficiency, the probe(-sets) hybridize against individual BAC clones). For PCR-type markers, screening all individual BAC clones individually is generally considered too expensive, and in such cases the BACs themselves are customarily pooled (e.g. Klein et al. 2000).

Recently an ultra dense, AFLP based genetic map of potato has been published (van Os 2006 et al.). Chapter 2 of this thesis describes development of an AFLP fingerprinted BAC library of the paternal parent used in this genetic map and Chapter 5 discusses corrections applied to the fingerprints to improve their reproducibility and ultimately the quality of a resulting physical map. Here we describe the construction of a first version of a physical map from these fingerprints, and the development of an ultra efficient, AFLP based, anchoring method that we have used with markers from 57 primer combinations to anchor BAC contigs to the ultra dense genetic map.

Materials and methods

Overview

Objective is to construct a fingerprint based physical map of potato, and to anchor the contigs in this physical map to the ultra dense genetic map of potato (van Os et al. 2006)

with the express intent to use these contigs to direct the sequencing effort currently being undertaken by the members of the Potato Genome Sequencing Consortium (PGSC, http://www.potatogenome.net).

The ultra dense genetic map was created by screening a segregating full-sib offspring of two diploid potato clones, SH83-92-488 and RH89-039-16 (Rouppe van der Voort et al. 1997), hereafter denoted "SH" and "RH", using AFLP™ (Vos et al. 1995). AFLP markers in the genetic map were obtained from three PCR-templates generated using three enzyme combinations: SacI/MseI, PstI/MseI and EcoRI/MseI, and selective amplification was obtained with AFLP primer pairs (Vos et al. 1995). For the EcoRI/MseI combination three selective nucleotides (Vos et al. 1995) were used on both the EcoRI and the MseI primers, which we will denote as "E+3/M+3" primer combinations. We used primers without selective nucleotides (denoted as "E+0/M+0" primers) to fingerprint the individual BAC clones (Chapter 2 of this thesis). During DNA isolation for individual BAC fingerprinting we constructed Quarter Plate Pools (QPPs) and Full Plate Pools (FPPs), containing DNA from a quarter library plate (96 BAC clones) and a full library plate (384 BAC clones) each (Chapter 2 of this thesis).

As the objective is to anchor physical map contigs to the ultra dense genetic map, we need some criterion or a combination of criteria to determine which BAC clone(s) contain which marker(s) from the ultra dense map. The anchoring method makes use of four important observations:

1. Only those E+3/M+3 AFLP fragments that were heterozygous present in one or both of the parents can segregate and be an AFLP E+3/M+3 marker the ultra dense genetic map, and only E+3/M+3 AFLP fragments that are heterozygous in the paternal (RH) genotype can be mapped in that parent, either as paternal (<00x0a>) or as bridge (<0ax0a>) segregation type markers.
2. E+3/M+3 AFLP fragments are a subset of all possible E+0/M+0 fragments (e.g. Han et al. 1999, ignoring problems putatively caused by varying template complexity). In other words, for each E+3/M+3 marker fragment observed in the genomic DNA of the RH genotype, there is a corresponding E+0/M+0 fragment of the same size that may be visible in the E+0/M+0 fingerprints of the individual BAC clones (that were derived from this genotype).
3. If we screen QPPs of the BAC library using E+3/M+3 AFLP markers we expect a few QPPs to be marker-positive (depending on local genome coverage and distribution of BACs among QPPs), thereby providing another subselection method: only marker-positive QPPs may contain BACs with a particular AFLP E+3/M+3 marker fragment.
4. If BACs are ordered into contigs based on their E+0/M+0 fingerprints, we may

expect some (probably overlapping) BACs present in a single contig to contain the same E+3/M+3 AFLP marker fragment.

Combined, observations 1) and 2) can be used as a selection criterion which we refer to as "fragment matching": If we know the size of a bridge or paternal segregation type AFLP E+3/M+3 marker fragment, then we can make a sub-selection of BAC clones from the fingerprinted part of the library that may potentially contain that marker by looking for an E+0/M+0 fragment of exactly the same size in the BAC fingerprints. Combined, observations 3) and 4) lead to a selection criterion which we refer to as "contig matching": while random contigs may (nominally) contain one or a few unrelated BACs from marker-positive pools (e.g. QPP or FPP), only one contig should contain BACs from many marker-positive pools because this contig contains many marker positive BACs. In combination with our physical map and the heterozygous AFLP markers, neither the "contig matching" nor the "fragment matching" selection criterion is sufficiently powerful to unequivocally assign markers to BAC clones, however, when combined, they are.

In order to turn this combination of selection criteria into a highly efficient map integration method we optimize screening of the QPPs for E+3/M+3 genetic map markers by applying a pooling design. We pool the QPPs into what we call Complex Pooled Pools or CPPs using a pseudo-random pooling design, that allows us to screen relatively few CPPs while still allowing us to resolve the scores to the (many more) underlying QPPs.

One complicating factor that we need to address is the fact that AFLP E+3/M+3 genetic markers in the ultra dense genetic map were developed using radioactively labeled fragments (^{33}P) separated on a Biorad radioactive slab gel system, whereas electrophoresis during AFLP E+0/M+0 BAC fingerprinting was performed on a MegaBACE 1000 capillary sequencer. Because these electrophoresis systems have different characteristics, some effort is needed to comparatively identify and size E+3/M+3 marker fragments in MegaBACE data.

Complex pooling design

As we want to screen our CPPs using a high throughput protocol, employing a MegaBACE 1000, a 96-capillary sequencer, the number of CPPs plus reference samples to be screened should be a convenient subset or multiple of 96. We chose to use 90 CPPs plus 6 reference samples. We distributed four copies of each QPP pseudo-randomly among the CPPs in such a way that each copy of a particular QPP sample occurs in a different CPP and each CPP contains either 33 or 34 QPPs (n.b. 764 QPPs times four copies is not divisible by 90 without remainder). Our choice to use four copies of each QPP was primarily motivated by concerns that small CPPs might constitute insufficiently complex DNA templates to reliably perform AFLP using E+3/M+3 primers (Han et al.

1999). Assuming 10 fold coverage of the haploid potato genome by the fraction of our libraries for which DNA was isolated and QPPs constructed (73,344 clones), then, with four copies of each of the 764 QPPs distributed among the 90 CPPs, each CPP nominally contains DNA of either 3168 or 3264 BAC clones, or approximately 0.44 genome equivalents, which was expected to constitute sufficiently complex AFLP template.

We deconvolute the CPP marker screening results to individual "resolved positive QPPs" using the following algorithm (for each marker individually):

1. We adjust, in silico, in order to take missing observations (e.g. caused by a capillary drop-out) into account, the pooling design by removing those CPPs for which no data is available, establishing what we call an "adjusted pooling design".
2. We count how often each QPP occurs in the adjusted pooling design, obtaining what we call "Target QPP counts". (n.b. in absence of missing observations and pipetting errors this should equal four).
3. For each QPP we count how many of the (adjusted) CPPs containing that QPP were marker-positive, obtaining what we call "observed QPP counts".
4. We compare, for each QPP, the observed count with the target count, and set the corresponding deconvoluted QPP score to positive if and only if the target QPP count equals the observed QPP count.

In this manner, deconvoluted QPP scores can be obtained for any subset of CPPs, regardless of pipetting errors or drop-outs. Note that using this algorithm, a QPP will be resolved false positive if by coincidence all the CPPs containing it are positive.

We evaluated the pooling design in silico by generating 120,000 pseudo-random patterns of between 1 and 12 positive QPPs (10,000 each), computing the resulting pattern of positive and negative CPPs that would result. This pattern of positive and negative CPPs was then used as input to the deconvolution algorithm described above, producing a set of resolved positive QPPs. This set of resolved positive QPPs was compared to the set of originally positive QPPs, and the number of false positives generated noted.

To implement the pooling design, QPP samples obtained previously (Chapter 2 of this thesis) were diluted 1:1 using water to decrease viscosity and increase pipetting volume, and using clean pipette tips for each sample, 20 µl samples of each QPP were combined into CPPs contained in a 96-well deep-well (2.0 ml) micro-titre plate. As the pooling design was executed manually, extreme care was taken to avoid or at least record human pipetting errors. After reading the QPP source and CPP destination well-addresses from paper, the pipette tip was inserted into the relevant QPP well and 20 µl aspirated and, before removing the pipette tip from the well, the QPP address was verified. After transfer, while the pipette tip was still within the relevant well, the CPP address was verified.

High throughput marker screening

AFLP on the CPPs (inluding SH and RH genotypes as reference samples) was performed essentially as described previously (Vos et al 1995). For screening on the MegaBACE platform, fluorescently labeled primers (FAM, JOE and NED) were substituted for the radioactively labeled primers used on the Biorad slab gel system. Electrophoresis on the MegaBACE 1000 machine and band calling (Using BAC-Xtractor, Srinivasan 2003) were performed as described previously for individual BAC DNA fingerprinting (Chapter 2 of this thesis). Besides the band-called ".bands" files, tiff-file representations of the individual electropherograms were exported. The tiff files, sized and stretched to a resolution of 10 pixels per basepair, were used as input to Gelsynth (Borm, unpublished, https://secure.potatogenome.net/gelsynth/), a web-based, pseudo-slab gel image generator, for visualization. For the purpose of marker identification and size conversion (described in the next paragraph), a subset of 19 of the 90 CPP samples plus both parents were screened in also using radioactively labeled E+3 primers and unlabeled M+3 primers in combination with the Biorad slab gel system.

Marker identification and size conversion

We identified markers in the original (Biorad slab-gel based) autoradiographs used for genetic mapping on the basis of their mobility and the marker segregation pattern in the offspring, and used the (reproducible) parental banding pattern to identify the relevant marker fragments in the duplicated Biorad slab gel based 19 CPP samples. Marker fragments were then identified in the MegaBACE dataset both on the basis of approximate mobility, as well as on the basis of the pattern of marker presence and absence observed in the 19 CPPs. The quality of marker identification and size conversion was manually judged and classified as:

1. "Good" - When there is no doubt about identification and correspondence of fragments observed in both electrophoresis systems.
2. "Likely" - When there is little doubt about identification, for instance because a band was not called in one or a few of the MegaBACE capillaries where a band was expected, or because a band was called where none was expected.
3. "Possible" - When there is serious doubt about identification, for instance because overlapping size distributions of adjacent fragments in the MegaBACE capillaries or because of differences in fragment separation between electrophoresis systems leading to ambiguous identification.
4. "Failed" - When the marker fragment could not be identified in either of the gel systems, for instance because of an administrative error (wrong size or wrong primer combination) or because the fragment does not appear to be present in the CPPs.

Within the band-called MegaBACE data, some variation in individual band sizes was observed. Therefore minimum and maximum fragment sizes observed in the MegaBACE data were used to delimit scoring intervals: any band within a particular interval was taken to represent a marker-positive CPP.

Physical mapping and physical map data extraction

Before physical mapping, the filtered and size corrected fingerprints (Chapter 2 of this thesis) were manually curated to remove duplicate fingerprints, so that for every BAC clone only a single representative fingerprint remained. FPC (Soderlund et al. 2000) was used with "equation 2", tolerance=5 and cut-off=10^{-12}, while all other parameters remained at their default values. Besides running the "D-Qer" to reanalyze contigs with in excess of five questionable clones, no further optimization was performed.

Deconvolution of marker screening results and in silico anchoring

Positive CPPs were de-convoluted to positive QPPs, and in silico anchoring was performed for each marker individually. A series of six scripts, each fulfilling a specific task in the anchoring procedure, was used:

1. To extract (from the physical map ".fpc" file), for each contig, a list of BACs contained in that contig and process this list, by replacing each occurrence of a BAC by the QPP containing it, into a list of the QPPs that would be marker positive if all BACs within the contig would be marker-positive: The list of "predicted positive QPPs".
2. To extract CPP marker scores from the band-called CPP files (using the scoring intervals identified during marker identification and size conversion)
3. To deconvolute the CPP screening results as described.
4. To count, for each contig in the physical map, the number of "predicted positive QPPs" matched by a deconvoluted positive QPP, obtaining a set of what we call a "contig match scores" (one for each contig)
5. To count, for each contig, the number of BACs contained within the deconvoluted positive QPPs that contain an AFLP E+0/M+0 fingerprint fragment of the same size as the AFLP E+3/M+3 marker fragment, producing a set of "contig fragment scores" (one for each contig).
6. To combine and sort these "contig match-" and "contig fragment-" scores into an ordered list of scores, producing what we call a list of "candidate anchor points", with each candidate anchor point being a contig identifier with its contig match and contig fragment scores.

Candidate anchor points were manually inspected and where possible, a single contig was

selected as the true anchor point. Anchoring confidence was manually classified as either:

1. "Ok" - If the pooling design could be resolved with few problems and in silico anchoring identifies a single contig with a high "contig match" and a high "contig fragment" score.
2. "Candidate" - If there were either problems resolving the pooling design, *or* if there are multiple contigs with similar "contig match" and "contig fragment" scores *or* if there is a single contig with a high "contig match" and zero "contig fragment" score.
3. "Failed" - If there were severe problems resolving the pooling design, or if more than a few (2-3) contigs were found with similar "contig match" and "contig fragment" scores.

Anchor validation

The actual presence, as predicted by the in silico anchoring, of a particular AFLP marker fragment in anchored BAC clones was verified using AFLP with E+3/M+3 primers on a sample of individual BAC DNA isolates (Chapter 2 of this thesis). AFLP and electrophoresis was performed essentially as described previously (Brugmans et al. 2006), with the notable exception that samples were diluted tenfold prior to electrophoresis.

Results

Physical map construction and genome coverage of the physical map

From the 73,741 fingerprints in the cleaned (Chapter 2 of this thesis) and size corrected (Chapter 5 of this thesis) dataset, 4484 duplicate fingerprints were removed, resulting in a dataset containing 69,257 fingerprints. The two largest contigs, containing 2512 and 2603 fingerprinted BACs represent a "chloroplast contig" with fingerprints of BACs containing chloroplast DNA derived inserts and an "*E.coli* contig ", putatively containing fingerprints predominantly derived from *E.coli* genomic DNA (Chapter 2 of this thesis). 11,063 BAC fingerprints found no significant overlap with any other BAC and thus remained singleton. Assuming an average BAC insert length of 131 kb (Chapter 2 of this thesis) and a haploid genome size of 850 Mbp (Arumagnatan and Earle 1991), and discounting the fingerprints in the chloroplast and *E.coli* contigs, the BACs in the physical map cover the (haploid) potato genome approximately 9.9 times, with approximately 8.2 Genome Equivalents (GE, with respect to the haploid genome size) in contigs and 1.7 GE of BACs remaining singleton. The size distribution of the resulting contigs is shown in Table 1. These estimates represent a redundant coverage of the potato genome. An estimate of the non-redundant genome coverage is not easily obtained, but conservatively assuming that each contig of the 7680 contigs is comprising at least 130 kb of potato DNA, the resulting

estimated 998 Mb exceeds the haploid size of the potato genome of 850 Mb. This suggests that merging of contigs will result in further identification of overlapping DNA.

Table 1: Distribution of contig sizes and resulting (haploid) genome coverage in the 1st build of the physical map.

	Singletons	**Distribution of BACs in contigs**				***E.coli***	**Chloroplast**	**Total**
BACs/contig		**2-3**	**4-10**	**11-30**	**31-100**			
#Contigs		3010	3360	1148	162	1	1	7682
#BACs	11063	7190	20898	18000	6991	2512	2603	69257
Coverage (GE)	1.71	1.11	3.22	2.77	1.08	-	-	9.9 (*)

(*) Please note that while BACs in the *E.coli* and Chloroplast contigs do contribute to the number of BACs in the library, they do not contribute to the estimated total genome coverage of the potato nuclear genome (9.9 times).

Complex pooling design

On a total of 6112 (we actually implemented two separate CPP pooling designs, but used only one for screening) pipetting transactions, 143 (2.3%) mistakes were recorded. Of these, 41 (0.7%) involved inserting the pipette tip into the wrong QPP source well and in 102 (1.7%) cases the pipette tip was inserted into the wrong destination well. The former type of mistake has no consequence as the QPP sample was not transferred to the CPPs, while the latter was accommodated by adjusting the pooling designs accordingly. After adjusting the pooling design to incorporate the pipetting errors that were detected we evaluated its performance in silico. For each marker the true marker-positive QPPs are expected to be concentrated in the contig containing the marker, while false positive QPPs are expected to be distributed randomly among other QPPs. Therefore we did not require deconvolution without false positives in our simulation to record a success, and used the situation where the number of false positives was smaller or equal to the number of truly positives as a criterion. Table 2 shows the average number of false positives and the "success rate" as a function of the number of simulated positive QPPs.

Table 2: Average number of false positives and success rate for various numbers of simulated positive QPPs

#positive QPPs	1	2	3	4	5	6	7	8	9	10	11	12
#false positives	0	0.02	0.11	0.38	0.93	1.86	3.26	5.28	7.97	11.4	15.6	20.5
Success rate	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.88	0.70	0.45	0.24	0.10

The average number of false positives and success rate were determined over 10,000 simulation runs for each number of positive QPPs. A succes was defined as the situation where the number of false positive QPPs is smaller or equal to the number of true (simulated) positive QPPs.

Marker identification and size conversion

For 57 EcoRI/MseI primer combinations the comparative analysis of AFLP fragment mobilities between the radioactive and MegaBACE system has been completed.

Combined, in the ultra dense genetic map, these primer combinations contain 2269 markers, of which 831 were mapped in the maternal (SH) parent only, and can therefore not be used to anchor physical map contigs built from BACs derived from DNA of the paternal parent (RH). The remaining 1338 markers (548 bridge and 790 with paternal segregation type) were available for anchoring purposes. Table 4 summarizes (in the columns labeled "Size Conversion") the results of marker identification and conversion of radioactive AFLP fragment mobility to MegaBACE AFLP fragment mobility for markers within each primer combination.

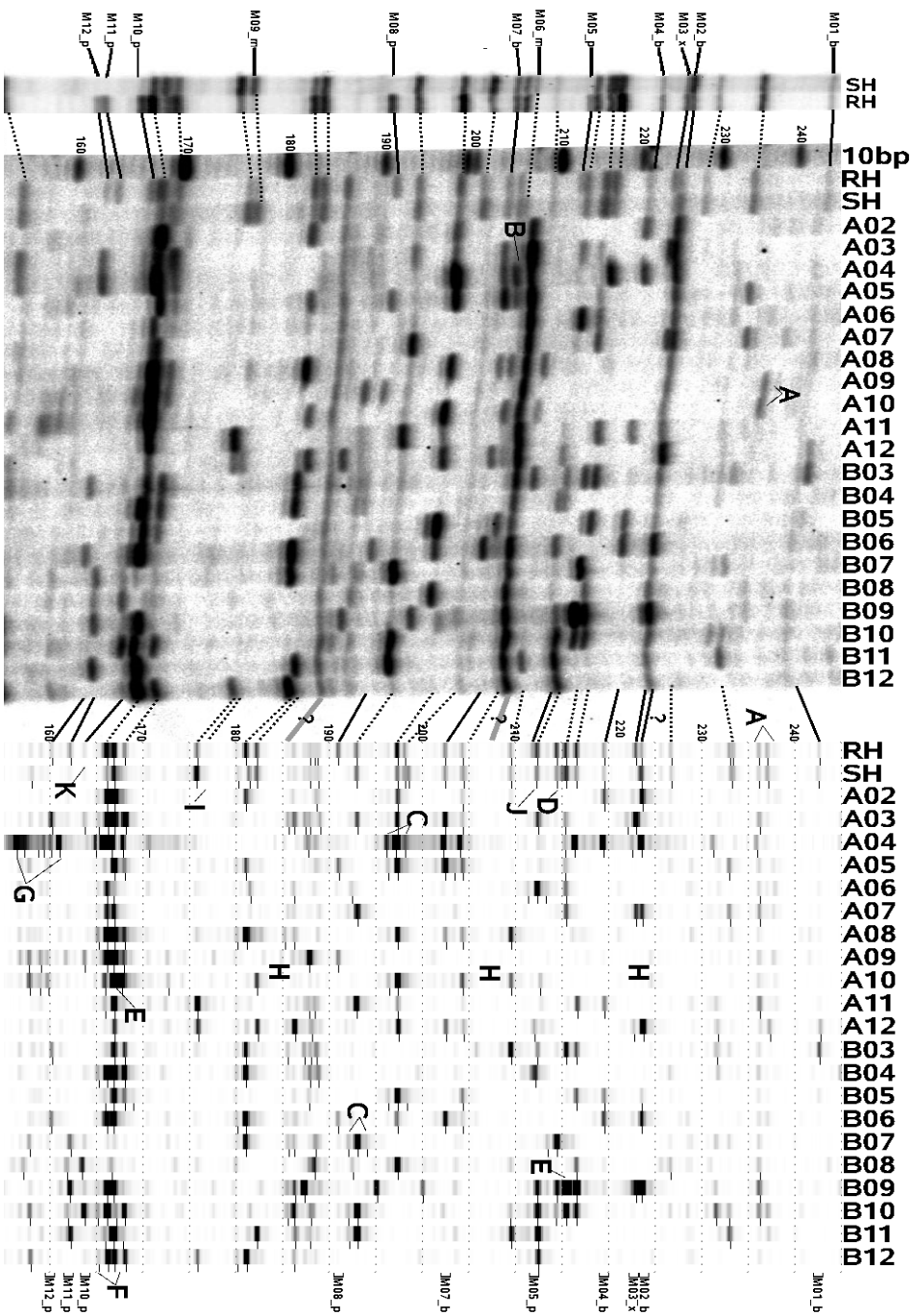
Figure 1 provides an illustration of some of the problems encountered during size conversion. This image shows partial data (only fragments from approximately 150bp to 245 bp) for three different electrophoresis runs. To the left the parental fingerprints in the autoradiographs used to construct the high density genetic map, in the middle both parents and 19 CPPs that were run in duplicate on the Biorad slab gel system, and to the right the same 21 samples run on the MegaBACE capillary sequencer. Note that the order of the fingeprints of SH and RH in the left section are swapped when compared to the parental fingerprints in the middle and right sections. While the left and middle sections correspond to real slab gel images, the section on the right is a pseudo-gel image synthesized from sized and stretched MegaBACE electropherograms, with to the right of each separate electropherogram the positions of the corresponding bands called by BAC-Xtractor (Srinivasan et al. 2003). We see: A) spurious peaks present seemingly randomly in only one of the gel-systems. B) Occasional (putative) "stutter peaks" present below a fragment in the Biorad slab gel images. C) Occasional (putative) "stutter peaks" present on either side of a fragment in the MegaBACE pseudo gel images. D) A band doublet in slab gel images becoming a triplet in MegaBACE data. E) Peaks (present in both electrophoresis systems), where BAC-Xtractor failed to detect a band in the MegaBACE data. F) Peaks separated in capillary electrophoresis were co-migrating as a single peak in a slab gel. G) a smear leading to spurious band calls. H) Constant bands present in all the CPPs in one electrophoresis system, absent in the other. I) A well-separated marker peak apparently missing in the MegaBACE pseudo gel images. J) A marker fragment barely separable from its neighbor in a slab gel, comigrating in the MegaBACE system. K) A failure to detect a (marker) peak in the parental fingerprints in the MegaBACE electropherograms. In addition, for all fragments we observed both variable and systematic differences in fragment mobility between bands sized on different electrophoresis systems and experience difficulties to match fragments based on parental fingerprint patterns alone. Nevertheless, in spite of the many discrepancies, 78% of the 1338 genetic markers could be retrieved from MegaBACE fingerprints, which offers a vast resource of markers for the anchoring of BAC contigs to the potato map. This value

of 78% can also be viewed as an overly conservative estimate of the genome coverage offered by the BAC library because the majority of the failures to detect a genetic marker in the BAC pools are due to electrophoretic artifacts such as overlapping fragment migration and not to fragment absence.

Figure 1 (next page): Comparison of an autoradiogram with the pseudo-gel image produced from a MegaBACE capillary sequencer electropherograms. The figure illustrates the necessity to re-identify genetic markers of the potato map (generated with autoradiograms) in the BAC pools of the physical map (generated with the MegaBACE). As fragment identity is based on fragment mobility a platform dependent fragment specific size conversion is required. Shown is part of three separate electrophoresis runs covering fragment mobilities approximately from 150 to 245 basepairs. To the left are the fingerprints of the parental potato clones for primer combination E+AAG/M+AGC as present in the original autoradiogram used for construction of the high density genetic map. In the middle an autoradiogram is shown with both parents next to 19 complex pooled pools (CPP) samples. To the right are the same 21 samples, only now as pseudo-gel images produced from MegaBACE capillary sequencer electropherograms. To the right of each capillary the bands detected by BAC-Xtractor software are indicated. Lines connecting the images show the re-identification of markers and their mobility shifts. Markers were identified and size-converted using both the approximate fragment locations and their banding pattern in the CPPs. Letters indicate various types of problems discussed in the text.

In silico anchoring of BAC contigs to the potato genetic map

To illustrate the genetic anchoring of BAC contigs, the subsequent steps of this process are shown in Figure 2 and described in the text below. The AFLP E+0/M+0 fingerprints of the BAC clones making up two specific contigs in the physical map (contig 1268 and contig 2558) are depicted as pseudo gel images generated from band-called data. Each of these BAC clones is present in exactly one QPP, identified by the 384 well library plate number (e.g. RH101) followed by Q1, Q2, Q3 or Q4 (identifying the particular quarter library plate); these QPP identifiers are shown as gray text. For each contig this list of QPP identifiers equals the list of “predicted positive QPPs”. The complex pooled pools (CPPs) were screened for three specific markers of the ultra dense genetic map (EACAMCAG_565.5, EAGAMCTG_292.1 and EACCMCAA_243), producing “CPP marker scores” for these markers. Deconvolvution of these “CPP marker scores” produced “deconvoluted QPP scores”. After scanning through all contigs in the physical map, the two contigs depicted in Figure 2 were identified as matches because: A) Their “predicted positive QPPs” predict many of the “resolved positive QPPs” (marked in Figure 2 by a “+” in the applicable fingerprint for each marker) and B) BAC clones present in the “resolved positive QPPs” contain an E+0/M+0 AFLP fragment of the same (MegaBACE) mobility as the E+3/M+3 marker fragment. The first of these markers only identifies contig 1268, and was classified as “Ok”, while the other markers identified both contigs, and were classified as “Candidate”. All three markers are genetically mapped to the same location: RH chromosome 5, bin46. As suggested by the genetic co-localization of the E +3/M+3 markers, and also based on similarity of the E+0/M+0 BAC fingerprints, these contigs can be merged in FPC at a lower threshold setting (data not shown).


M01_b
M02_b
M03_x
M04_b
M05_p
M06_m
M07_b
M08_p
M09_m
M10_p
M11_p
M12_p
SH
RH
10bp
RH
SH
A02
A03
A04
A05
A06
A07
A08
A09
A10
A11
A12
B03
B04
B05
B06
B07
B08
B09
B10
B11
B12
A
B
C
D
E
F
G
H
I
J
K
?

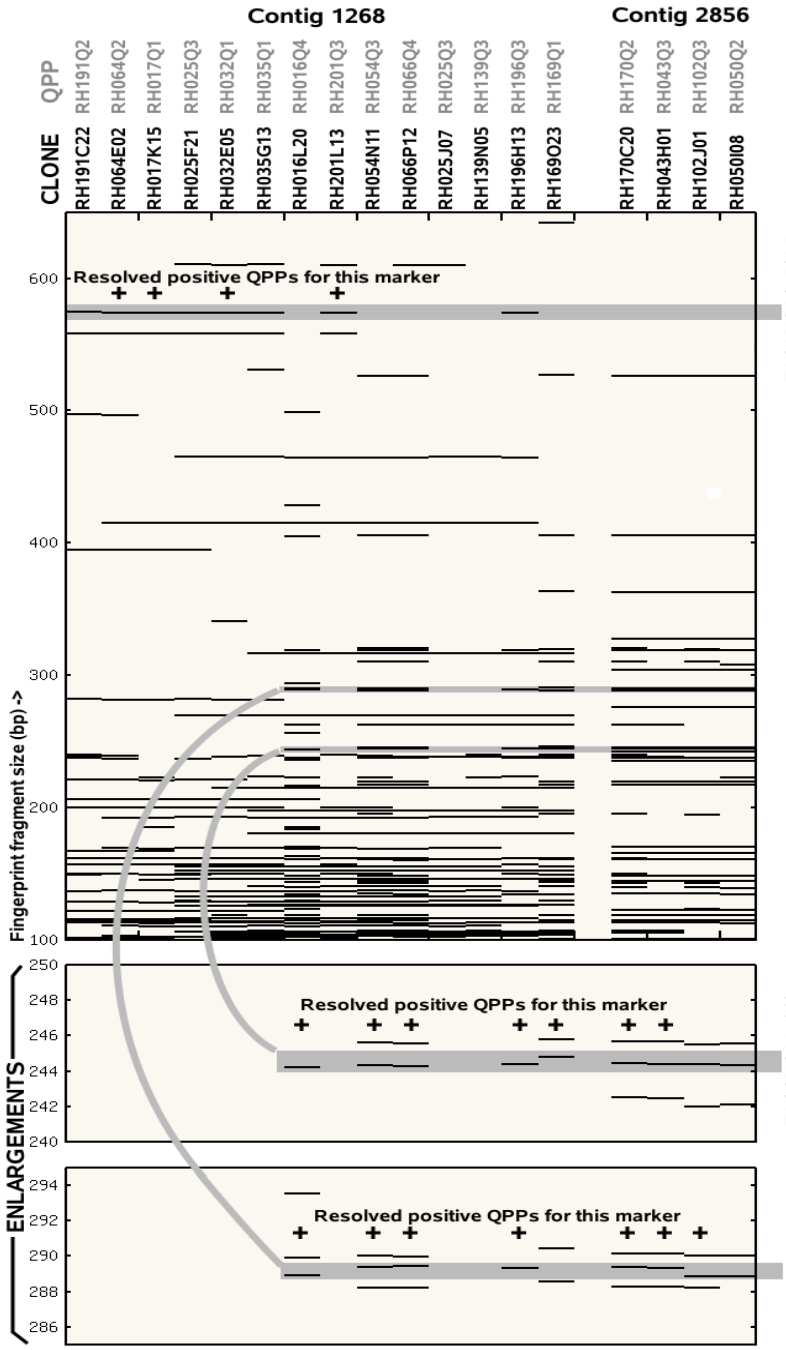


Figure 2: Example of two contigs genetically anchored through AFLP markers to potato chromosome 5. BAC clones within contigs where the resolved positive QPP scores matched the predicted QPP score are indicated by a "+" in this figure. In this particular case one contig is anchored with "Ok" quality using marker EACAMCAG_565.5, while both contigs are anchored as "candidate" through the other two markers. All three markers are located in bin 46 of paternal linkage group 5 of the high density genetic map. Both contigs can be merged based on their fingerprints using FPC at lower stringency settings, and AFLP with E+3/M+3 selective primers on three of the BAC clones validated this anchoring.

Per primer combination results of in silico anchoring are shown in the columns labeled "Anchoring" in Table 4. On average, 73% of the markers that could be identified and size-converted, could be used for anchoring. The column "Overall" shows the compound success rate – of being able to use a genetic map marker to anchor one or more contigs. Because a single marker may anchor multiple contigs, the columns labeled "Contigs" show the number of contigs actually anchored by each primer combination. As these counts are per primer combination, and multiple markers in different primer combinations may anchor the same contigs, the total number of contigs anchored is somewhat lower. It is clear from the results in Table 3, which shows the number of anchored contigs and clones within those contigs per linkage group, that linkage group 1 has most contigs anchored while all other linkage groups have between 46 to 84 contigs with the exception of linkage groups 3 and 8 which have a lower number of anchored contigs, which is largely a reflection of the number of markers per linkage group (van Os et al. 2006).
The anchoring procedure has been validated by performing AFLP on an independent system (using the LiCor slab gel electrophoresis system, Brugmans et al. 2006) on a selection of 118 BAC clones which belonged to 44 anchored contigs and which should contain a specific diagnostic AFLP E+3/M+3 fragment from the genetic map of potato. The 118 BACs were fingerprinted with the appropriate AFLP E+3/M+3 selective primer combinations (38 different combinations in total) (data not shown). In view of the selectivity of E+3/M+3 AFLP primer pairs genomic template (850 Mb for potato) should result in 70-120 fragments per lane. An individual BAC clone represents on average only 130 kb, which is a more than 6000-fold reduction in template complexity. A random E+3/M+3 primer combination on a random BAC usually results only in a vague pattern of mismatch amplification products. This experiment however resulted in all cases in an AFLP fragment of the correct size, confirming correct anchoring of the BAC (and thus the contig). When loading the undiluted AFLP sample onto the gel, a large blob could be observed at the approximate AFLP target position together with some moderately strong other bands, indicating that an excessive amount of DNA was loaded. Even at tenfold dilution the signal for the target AFLP fragment is very strong, while all the other bands are much weaker. The example shown in Figure 2 has also been validated by performing AFLP with E+3/M+3 selective primers on three of the individual BAC clones, which resulted in a single predominant E+3/M+3 amplification product of the expected mobility.

Table 3: Anchored contigs and clones within those contigs per linkage group

Linkage group	1	2	3	4	5	6	7	8	9	10	11	12
#contigs anchored	146	47	32	84	82	74	60	32	52	46	65	78
#clones anchored	1294	388	316	742	731	526	447	943	384	558	572	771

Table 4: Summary of anchoring results per primer combination.

		Size conversion				Anchoring					Contigs		
Primer combination	Available	Good	Likely	Possible	Success rate	Confirmed	OK	Candidate	Success rate	**Overall**	Confirmed	OK	Candidate
EAAAMACG	19	11	1	1	68%	0	4	3	54%	37%	0	4	3
EAACMACG	22	16	0	2	82%	3	12	3	100%	82%	3	12	3
EAACMCAA	32	24	1	4	91%	7	14	6	93%	84%	7	14	10
EAACMCAC	42	17	3	8	67%	5	3	11	68%	45%	5	3	12
EAACMCAG	34	19	3	3	74%	1	14	2	68%	50%	1	14	3
EAACMCAT	27	20	1	1	81%	1	13	4	82%	67%	1	13	4
EAACMCCA	23	11	2	3	70%	1	4	6	69%	48%	1	4	7
EAACMCCT	22	13	1	2	73%	0	4	4	50%	36%	0	4	6
EAACMCGA	15	6	2	1	60%	0	3	3	67%	40%	0	3	4
EAACMCTC	29	20	4	2	90%	1	17	2	77%	69%	1	17	3
EAACMCTG	20	13	2	1	80%	2	9	2	81%	65%	2	9	3
EAACMCTT	34	16	5	6	79%	0	9	4	48%	38%	0	9	5
EAAGMACC	44	19	9	9	84%	6	13	10	78%	66%	6	13	13
EAAGMAGC	31	20	4	1	81%	3	12	6	84%	68%	3	12	7
EAAGMCAC	31	18	1	7	84%	5	10	7	85%	71%	5	10	7
EAAGMCAT	21	12	4	2	86%	0	9	3	67%	57%	0	9	4
EAAGMCTC	36	9	7	5	58%	0	12	5	81%	47%	0	12	7
EACAMACC	22	16	2	3	95%	3	8	6	81%	77%	3	8	8
EACAMAGG	18	14	1	1	89%	0	10	2	75%	67%	0	10	2
EACAMCAC	25	22	0	1	92%	0	16	4	87%	80%	0	16	5
EACAMCAG	17	12	0	3	88%	4	4	4	80%	71%	4	4	5
EACAMCAT	30	19	1	3	77%	0	13	3	70%	53%	0	13	4
EACAMCCA	15	10	2	0	80%	0	9	1	83%	67%	0	9	2
EACAMCCT	19	8	3	2	68%	0	4	7	85%	58%	0	4	11
EACAMCGA	19	13	0	4	89%	2	7	5	82%	74%	2	7	7
EACAMCTG	24	15	3	3	88%	0	10	3	62%	54%	0	10	5
EACCMACA	28	7	5	6	64%	0	5	2	39%	25%	0	5	3
EACCMACT	21	17	0	2	90%	0	11	5	84%	76%	0	11	9
EACCMATC	23	12	1	5	78%	0	7	4	61%	48%	0	7	4
EACCMCAA	22	13	2	2	77%	0	7	5	71%	55%	0	7	7
EACCMCAT	26	16	3	2	81%	0	10	5	71%	58%	0	10	7
EACCMCTA	18	14	0	0	78%	0	8	2	71%	56%	0	8	2
EACCMCTT	31	21	3	0	77%	0	13	8	88%	68%	0	13	11
EACTMCAG	24	16	0	1	71%	0	7	5	71%	50%	0	7	6
EAGAMCAG	14	12	1	0	93%	0	8	4	92%	86%	0	8	5
EAGAMCAT	28	9	3	10	79%	0	4	3	32%	25%	0	4	4
EAGAMCCT	17	13	1	2	94%	0	6	8	88%	82%	0	6	11
EAGAMCTG	17	13	1	1	88%	0	8	4	80%	71%	0	8	5
EAGCMAAG	22	16	2	1	86%	0	13	3	84%	73%	0	13	4
EAGCMACA	24	20	2	2	100%	0	15	3	75%	75%	0	15	6
EAGCMAGT	29	11	1	4	55%	0	8	0	50%	28%	0	8	0
EAGCMATC	19	10	2	2	74%	0	7	2	64%	47%	0	7	4
EAGGMAAC	24	12	1	4	71%	0	5	6	65%	46%	0	5	8
EAGGMAAG	21	12	0	2	67%	0	8	2	71%	48%	0	8	2
EAGGMACA	19	7	5	1	68%	0	6	4	77%	53%	0	6	5
EAGGMACT	20	14	1	2	85%	0	9	3	71%	60%	0	9	4
EAGGMAGA	20	8	2	1	55%	0	6	1	64%	35%	0	6	1
EAGGMAGT	20	8	3	3	70%	0	7	2	64%	45%	0	7	2
EAGGMATG	17	7	2	2	65%	0	3	4	64%	41%	0	3	5
EAGGMCTA	17	13	1	2	94%	0	9	5	88%	82%	0	9	6
EAGTMAGC	17	8	4	2	82%	0	6	5	79%	65%	0	6	6
EAGTMCAA	23	10	1	1	52%	0	9	1	83%	43%	0	9	1
EAGTMCAC	24	18	1	0	79%	0	12	2	74%	58%	0	12	3
EAGTMCCA	23	14	3	4	91%	0	8	2	48%	43%	0	8	3
EAGTMCCT	18	12	2	1	83%	0	10	3	87%	72%	0	10	5
EAGTMCGA	19	13	2	2	89%	0	9	3	71%	63%	0	9	4
EAGTMCTA	22	15	1	0	73%	0	11	3	88%	64%	0	11	5
Total	1338	784	118	145	78%	44	498	225	73%	57%	44	498	298

Success rates are shown separately for size conversion and anchoring (which requires successful size conversion) and overall. The “Contigs” column shows the number of contigs anchored per marker. Because some contigs have been anchored through multiple markers, the actual number of different contigs anchored is somewhat lower.

Discussion

Pooling design

There are three particular risks associated with pooling designs and the way we have implemented it:

1. There can be false positive outcomes. If a series of CPPs that are marker-positive because of a particular combination of QPPs coincidentally contain all copies present in the pooling design of some unrelated QPP, then this QPP will also be assumed positive. For this to happen in our pooling design, there must a specific set of (nominally) four CPPs positive, and if only a modest number of QPPs is positive this does not happen often (Table 2). For larger numbers of positive QPPs we have observed this, and we accommodate this by not requiring that a contigs' set of "predicted QPP scores" explain all resolved QPPs. Our definition of a "succes" as used in Table 2 reflects the fact that if there are many (>5) resolved positive QPPs, then there will be just one or a small number of contigs explaining many of these resolved positive QPPs, while there will be many contigs explaining just one or a few of these resolved positive QPPs. We may assume that the latter matches are caused by false positive QPPs. For smaller number of resolved positive QPPs, there will automatically be fewer false positives. Assuming random cloning of BAC clones and 10-fold coverage of the haploid genome by the BAC library, we expect heterozygous AFLP E+3/M+3 marker fragment to be present on average five times in our BAC library. This marker copy-number is expected to follow a binomial distribution, with approximately 90% of the markers being present in up to eight copies, therefore false positives are expected fairly frequently, but given our definition of a "success", even for a marker leading to eight positive QPPs, we expect an average of 5.28 false positives and a deconvolution success rate of 88%. As we have not screened any of the QPPs using E+3/M+3 AFLP primers we have no definitive data on the occurrence of false positive QPPs.
2. A false-negative score in one of the CPPs results in false negative scores for all of the QPPs it contains. In many respects this problem is much graver than the problem of false positives: a single scoring error will immediately affect many QPPs. It can, however, to a certain extent be detected. If we reverse the pooling design deconvolution process by simulating the combining of QPP scores into CPP scores as it is specified by the pooling design in silico, we can compute, given (deconvolved) QPP scores what the CPP scores should have been. If a large difference is observed between these reconstituted CPP scores and the observed CPP scores, then problems are indicated. After detection of problems, the markers can be

(manually) re-scored to improve the situation. Alternatively, particular CPPs can be declared as having "failed". In particular the last option has proven extremely useful: by visually inspecting the band-called files as pseudo-gel images, CPPs fingerprints containing an excessive number of bands, like previously seen in band-called files of "known-empty" fingerprints (Chapter 2 of this thesis), or CPPs fingerprints containing relatively few bands can be detected. These (putatively) failed fingerprints can be removed from analysis. Although such removal may eventually affect the deconvolution of the pooling design, it illustrates the necessity of robustness in the pooling design: Despite runs of some primer combinations having up to 8 CPPs marked as "bad" and subsequently removed, all primer combinations succeeded in anchoring at least some contigs.

3. The risk we can do least about are undetected pipetting errors during pool construction. This type of error will simultaneously lead to false positives and false negatives that can not trivially be dealt with. If a particular misplaced QPP is negative it may coincidentally be found positive if it is registered to be present in a set of CPPs found positive because of other, unrelated QPPs, while if a misplaced QPP is positive it may not be resolved as such from the pooling design. Unless coincidentally positive because of other unrelated QPPs, the CPP registered to contain the QPP will be negative, resulting in a negative QPP being resolved. This last fact may also indicate a way to identify which QPPs are affected by pipetting errors: such QPPs will, using a large set of essentially random markers, be resolved as positive relatively few times, and then only coincidentally. The number of markers screened so far, and the expected statistical distribution in QPPs containing these markers, however, does not warrant any conclusion regarding the ability to recognize and accommodate those undetected pipetting errors at the moment.

Grosso modo our pooling design behaves as expected, and though false positives and false negatives are assumed to have lowered our anchoring yield (73%), this is more than offset by the fact that the CPPs represent an approximately eight-fold reduction in the number of samples that needs to be screened.

Physical map construction and genome coverage

Coverage of the haploid potato genome by our BAC fingerprint-based physical map may seem sufficient at 9.9 Genome Equivalents (GE) at it is statistically expected to contain 99.99% of the potato genome (Clarke and Carbon 1976). We should, however, keep in mind that potato is a highly heterozygous species, which may mean that it is more appropriate to discuss coverage in terms of the diploid genome size (5 GE, providing 99.3% coverage). BACs from different potato homologous chromosomes may lack co-

linearity, similar as observed in the maize genome, where transposon insertions and deletions by illegitimate recombination resulted in marked differences between the DNA sequence of two maize inbred lines (Fu and Dooner 2002), and the non-random distribution of restriction sites may result in preferential cloning. Being able to construct contigs from fingerprints requires sufficient overlap between fingerprints containing a sufficient number of bands. Given that, in order to be able to construct contigs using FPC we need to restrict ourselves to fingerprint fragments between 100 and 650 bp, we effectively discard approximately 25% of the bands (Chapter 2 of this thesis), and consequently more overlap in fingerprints between BACs is required. The number and size distribution of contigs we observe seems to be in line with automatically assembled physical maps reported previously. The parameters we have selected for automatic map construction are what we believe fairly stringent, though, as discussed previously (Chapter 2 of this thesis), this stringency can not meaningfully be statistically quantified using FPC's cut-off parameter. This results in contigs with few questionable clones. Initial studies (no data shown) indicate that many of these contigs can be merged based on fingerprints alone, though not automatically, and not by relaxing auto-assembly stringency. Some of these non-curated mergers have demonstrated the necessity of manual curation, as some of the merged contigs were found to be anchored (through multiple markers) to different linkage groups or widely separated loci on a single linkage group of the ultra dense genetic map of potato. Assuming correct genetic mapping and correct anchoring, such conflicts can arise either through erroneous fingerprint-based contig merges and through rarely observed chimaeric BAC clones. Therefore it seems prudent to err on the safe side and not merge contigs too liberally.

It can also be argued that, with the sequencing of the whole potato genome currently being undertaken on the basis of the same BAC library (http://www.potatogenome.net/), with the full complement of BAC-end sequences of the library already being available (Zhu et al. 2008), and new sequence data becoming available to Potato Genome Sequencing Consortium (PGSC) partners, that contig merges should in principle only be executed when supported by both fingerprint data and sequence matches between BAC end sequences and sequenced BACs. A major obstacle to this approach of simultaneous sequencing and physical mapping is the presence of many small, unanchored, contigs. This makes a priori selection of sufficient BACs in a minimum tiling path for sequencing almost impossible. Unless a sufficient number of independent (simultaneous) starting points for sequencing can be selected, sequentially sequencing BAC clones (chromosome walking) may take too long.

The contigs in the current version of the physical map are on average slightly smaller and the number of singletons is slightly larger than in a previous version that we computed

(Chapter 2 of this thesis). We believe that this is entirely an artifact caused by the removal of duplicate fingerprints. Many BACs currently in the singletons pool with duplicate fingerprints were previously considered a two-fingerprint contig, and larger contigs containing duplicate fingerprints shrunk accordingly.

Marker re-identification and size conversion

It is a well known fact that the type of electrophoresis system used for AFLP analysis can affect the sizes and the ability to separate similarly sized DNA fragments. As we have seen in Biorad slab gel and MegaBACE capillary sequencer data, this can sometimes influence banding patterns observed in parental fingerprints to such an extent that it is occasionally impossible to identify homologous fragments. Some fragments separable on one electrophoresis system co-migrate on the other electrophoresis system. The use of a sample of 19 CPP lanes duplicated on both Biorad and MegaBACE system allowed us to identify relevant (paternal or bridge) AFLP marker fragments in the MegaBACE data with relatively high confidence and ease. Various artifacts, as illustrated in Figure 1 (with causes often unknown to us), complicate the marker identification and conversion to MegaBACE sizes. Additionally, not all (paternal and bridge) markers present in the genetic map are actually present in the BAC library (or at least the subset represented by the 19 CPPs used for size conversion). Fortunately, marker identification and size conversion has a reasonable yield. In our case, measured over 57 AFLP primer combinations, approximately 78% of the markers could be retrieved. We expect a similar yield for an additional 73 primer combinations awaiting analysis.

Anchoring of the physical map to the genetic map of potato

In essence our AFLP contig matching anchoring algorithm can be divided into two parts: “Contig matching” and “fragment matching”. For “contig matching” to work efficiently, sufficiently high “contig match” scores must be obtained. A single positive QPP selects 96 BACs from the BAC library, which will (nominally) be present in 96 different contigs in the physical map, all scoring an equally good “contig match score” of one. Two QPPs will select 192 BACs from the library, which are present in a multitude of physical map contigs, most of which achieve a “contig match score” of one and only a few contigs achieving the maximum score of two. Etcetera. In general, having more positive QPPs allows a less ambiguous match between a marker with a single contig to be made.

The “fragment matching” part works more efficiently if the marker fragment is sufficiently rare in the individual BAC fingerprints. Therefore well separated marker fragments, with a small MegaBACE scoring interval (little sizing variation), and in particular the larger AFLP fragments (because of the skewed distribution of AFLP

fragment sizes as discussed by Koopman and Gort 2004) work best. Separately neither "contig matching" nor "fragment matching" would (on average) be very efficient at selecting the correct contig in our physical map, but the combination of both is. Despite complicating factors such as false positive and false negative QPPs (resulting from first having to resolve the markers scored on CPPs), the approximate yield of the anchoring method is 73%, measured over 57 primer combinations. Combined with a 78% yield of marker identification and size conversion this results in a gross yield of 57%. Again we expect a similar yield for an additional 73 primer combinations awaiting analysis.

An approach similar to the "contig matching" part of our method has been described recently by Paux et al. (2008), termed "ELEPHANT", using direct screening of full plate pools of an eight times coverage, chromosome 3B specific, BAC-based physical map of hexaploid wheat (Paux, unpublished). In a simulation study reported in this paper, 93% of the simulated markers with 25% of the data missing can be assigned to the correct contig. These markers, however, are homozygous markers that will on average naturally occur eight times in an eight-fold coverage BAC library. The practical trial with 158 SSR markers resulted in a much lower success rate; overall it was 32% and it was 10% for markers present in five to ten pools. Their conclusion therefore is that the system performs best for markers encountered in 10-18 pools (more than the average genome coverage of the BAC library), and can not select the correct contig if fewer than five pools are positive. Translated to our situation, where 90% of the (heterozygous) markers is expected to occur eight times or less in the BAC library, this would mean screening 191 full plate pools (the fingerprinted part of our BAC library), while still being able to assign less than 10% of the markers to the correct contig. This failure to assign markers to the correct contig due to a lack of positive pools could of course be amended by increasing the coverage of our physical map until our heterozygous markers occur sufficiently often on average. This would, however, require excessive coverage, excessive fingerprinting and screening an extremely large set of pools. Application of the "ELEPHANT" methodology alone does therefore not entice us; even taking the difficulties experienced in converting between electrophoresis systems into account, we achieve much higher screening efficiencies with our system. Though it is possible to set the different values for the scores and penalties used by "ELEPHANT" to determine quality of the anchoring, we do not see how this scoring system would easily accommodate heterozygous markers in contigs (putatively) constructed from BACs derived from both homologous chromosomes. Another issue that sets our method apart from the "ELEPHANT" method is that, because it allows lover coverage, it may also identify multiple contigs with a single marker (putatively located near the end), offering valuable clues for contig merges.

Others (e.g. Klein 2000) have used pooling designs where BAC clones are (virtually)

located in a 3-D stack with approximately equal X Y and Z dimensions, pooling clones that are present in planar slices made through this stack in different directions. If the 3-D stack is sliced through in N different directions, such a pooling design is referred to as an N-dimensional pooling design. Applying this methodology to the fingerprinted portion of our BAC library, taking the dimensions (16 rows and 24 columns) of 384-well plates into account, our 3-D stack would have 32 layers consisting of 48 by 48 clones (2 by 3 384-well plates). Slicing through this stack in six directions (perpendicular to the three axes, and in three diagonal directions) would result in three pool-sets representing slices perpendicular to the axes containing 48+48+32=128 pools plus three diagonally sliced pool-sets containing 48+48+48=144 pools. The resulting total of 272 pools is approximately 3 times as many as we are currently using (90 CPPs), and would identify individual BACs that are positive for a marker, whereas our method resolves to the QPP level. Although such a pooling design would be fraught with the familiar problems of false positives and false negatives, we would expect such a pooling design to be able to use a larger percentage of markers to successfully anchor contigs, and consequently be more efficient than the "ELEPHANT" methodology, approaching our AFLP contig matching method. Of course, combining individual BAC DNA isolates into a pooling design represents a monumental task that is best automated using robotic equipment that may not be available even in otherwise well-equipped laboratories. The largest benefit for using a pooling design resolving to individual BAC addresses would come from being able to screen BAC libraries for which no physical map is available.

So far our anchoring method has proven 100% accurate; all of the BAC clones tested using specific E+3/M+3 AFLP primers did indeed contain the expected fragment, thereby also confirming some of the contig merges suggested by the marker screening data and fingerprints. Results obtained so far lead us to conclude that systematic verification of anchoring results, for instance through AFLP with E+3/M+3 primers on individual BAC clones or by FISH (Koo et al. 2008) is unnecessary.

Although other efficient BAC screening methods exist and are regularly used, in particular filter hybridizations using various types of probes, these have not been considered because of assumed incompatibility with the AFLP markers in our genetic map; converting these markers, without any knowledge about their internal sequence, to hybridization probes was considered too risky.

We have observed, amongst other unexplained artifacts, the presence of some extra (with respect to the paternal parent of the ultra dense genetic map and DNA donor of the BAC library, RH) AFLP fragments in the CPPs, we do not know if this is caused by artifacts due to an insufficiently complex AFLP template (Han et al. 1999), or by some other cause. The multiple cloning site used in the pIndigoBAC535 vector that was used to construct

our library contains an EcoRI restriction site. For the BAC clones constructed using the HindIII enzyme, this site remains unused, and is therefore accessible to the AFLP E+3/M +3 protocol in combination with an MseI site located within the BAC insert, essentially resulting in a variable vector-insert fragment. The relative positions of the EcoRI and HindIII cloning sites in the BAC vector are such that only AFLP E+3/M+3 primer combinations with "GAG" as selective nucleotides on the EcoRI side (E+GAG/M+3) can give rise to a variable vector-insert AFLP E+GAG/M+3 fragment. As no such primer combinations have been used, this variable vector-insert fragments offer no explanation for the observed extra fragments.

Conclusion

Here we have presented a first version of a genetically anchored physical map of potato, covering the potato genome approximately 9.9 times. A highly efficient physical-genetic map integration method is used, delivering approximately 800 anchored contigs distributed over the potato genome using just 57 AFLP primer combinations. Using only 90 Complex Pooled Pool samples (CPPs), two parental samples, and 21 duplicate samples for marker identification and size conversion, and assuming that AFLP template and pre-amplification product needs to be prepared only once for the entire project, we achieve an as yet unmatched overall anchoring efficiency, requiring less than 8 PCR reactions per anchored contig. With another 73 primer combinations currently being processed, we expect to be able to eventually anchor approximately 1400 physical map contigs, and in addition provide valuable clues to possible contig merges. Further refinement of the physical map is also expected to result from integrating sequence data delivered by the Potato Genome Sequencing Consortium, and given the ample supply of anchored contigs as starting points for sequencing the potato genome the target of delivering the first completely sequenced potato chromosomes by 2009 seems to be within reach.

Chapter 7

Summary and concluding remarks

It is possible to construct a local, BAC-based, physical map in order to answer a single, specific, biological question. The economy of doing this repetitively, however, in order to answer a multitude of biological questions, is such that the question will arise if this could be done more efficiently on a genome wide scale. As already argued by van Os et al. (2006), the ultra dense genetic map of potato delivers marker saturation on a genome wide scale, negating the need to do this locally, for instance using Bulked Segregant Analysis (BSA). The goal of the physical map construction project was similar: to saturate the potato genome with genetically anchored BAC contigs, negating the need to construct BAC contigs locally.

A central theme in this thesis, in absence of groundbreaking biological discoveries and without a biologically relevant hypothesis to test, is the creation, extraction, conservation, interpretation and integration of information present in or used for construction of the integrated physical and genetic map:

1. In Chapter 2 we try to extract information to characterize our BAC library and fingerprinting process from the fingerprints. By comparing the composition of the chloroplast contig (containing fingerprints derived from BACs containing chloroplast DNA) with BLAST results obtained using BAC-end sequences, the usefulness of contig construction for BAC library characterization is demonstrated.
2. The Universal Maximum Likelihood Pairwise Linkage Estimator (UMLPLE) presented in Chapter 3 attempts to conserve all the information about (sex specific) linkage between markers as present in ambiguous marker scores. While not immediately evident by itself, this capacity to deal with ambiguous marker scores is needed to place (bridge) markers, without loss of information, onto the sometimes incomplete bin signatures (van Os et al. 2005a, van Os et al. 2005b, van Os et al. 2006) of the ultra dense genetic map.
3. Binmap+ and Homap+, both presented in Chapter 4, allow more accurate information on the genetic positions of markers to be obtained by using the UMLPLE, by postulating empty-bin signatures and by using a system of data driven constraints to limit which combinations of a maternal and a paternal bin a bridge marker can be assigned to, effectively integrating the previously separate maternal and paternal maps of potato to the extent supported by data. Homap+ is unique in that it is the only program that we know of that is capable of mapping homoplasic AFLP fragments.

4. Chapter 5 uses fingerprints of BAC clones containing chloroplast DNA derived inserts to obtain information on systematic fragment sizing differences between different MegaBACE capillary sequencer machines, different fluorescent labels and different capillary positions. This leads to a fragment size correction which increases overall fingerprint data quality.
5. In Chapter 6 an efficient method to anchor BAC contigs to the ultra dense genetic map is described which combines information from three sources: Pools screened for AFLP E+3/M+3 markers, AFLP E+0/M+0 fingerprints individual BAC clones and the composition of the contigs in the physical map. By themselves, each of these three sources does not offer enough information to assign markers to BAC contigs and it is only the combination that turns out to be a more efficient anchoring method than the (also AFLP based) method practiced by Klein et al. (2000)

Besides these five chapters, and outside the scope of this thesis, several tools to present, manage, interpret and integrate information from the integrated physical and genetic map and the currently ongoing potato sequencing project (http://www.potatogenome.net/) have been developed. Amongst others:

1. A fully interactive presentation of the integrated physical and genetic map, taylor-made to accommodate and faithfully represent specific aspects of our bin-based ultra dense genetic map, such as ambiguous placement of markers in bins and the different linkage phases of markers.
2. The Submap package was developed to maintain the information in the integrated map in a version control system. By subdividing the information along functional boundaries users can work independently with subsets of the data (e.g. contigs) and merge and revert changes at will.
3. Gelsynth allows users to dynamically generate pseudo-gel images by combiningb trace obtained from individual capillaries on the basis of a set of queries. Queries can include individual traces, MegaBACE run numbers, (partial) BAC or BAC-pool identities and (partial) contigs.
4. The BAC-end-tool allows users to BLAST sequence data against the BAC end sequences (Zhu et al. 2008). It filters results to produce a list containing only the most relevant hits, and displays these in the context of the physical map. Results are gathered and may eventually be used for contig merges.

The integrated map, Gelsynth, the BAC-end-tool and other tools are available through a password-protected website (https://secure.potatogenome.net)

In Chapters 2 and 5 the BAC library and their fingerprints were characterized, filtered and processed. Although it is possible to construct a physical map of unfiltered, unprocessed fingerprints, computation will take much longer and yield contigs of lesser quality. Although we have no conclusive evidence that fingerprints in the "*E.coli* contig" are really derived from *E.coli* genomic DNA, the observation of highly similar patterns in AFLP fingerprints of another physical mapping project (personal communication Jan de Boer) is corroborating our hypothesis. Similarly, we have no conclusive proof that fingerprints that were removed because they were unlike normal AFLP fingerprints are diagnostic of some problem during fingerprinting. We are, however, not aware of any biologically relevant alternative mechanism that might produce such deviant fingerprints. They do, however, cause considerable problems when attempting to construct contigs, while not significantly contributing to genome coverage. It is interesting to note (Table 8 chapter 2) that for 1107 of the 3845 BAC clones that were removed because their fingerprints were unlike normal AFLP fingerprints, BAC end sequencing produced no data, indicating that there is some intrinsic reason why fingerprinting appeared to have failed.

Chapters 3 and 4 have resulted in an improved ultra-dense genetic map, in particular with more precise locations for bridge markers. While the flexibility of the UMLPLE in general and Binmap+ specifically to deal with ambiguous marker scores does not appear to be required to place the dominantly scored AFLP markers on the ultra-dense map, this is, however, not the case. The use of incomplete synthetic parental bin signatures while synthesizing bridge bin signatures, automatically produces such ambiguous marker scores. It would have been possible to take these ambiguous scores only into account for the synthetic bridge bin signatures and not for the marker scores themselves. This would, however, not have simplified the statistical model or indeed implementation significantly. The facility to place homoplasic AFLP fragments onto bins of the ultra dense genetic map has produced some interesting results. Occasionally markers that could previously not be mapped satisfactorily, now find a place on two different chromosomes. Some of these "homoplasic markers" have been used to anchor contigs and have been confirmed by other markers.

As discussed in chapter 6 of this thesis, running 57 AFLP M+3/E+3 Primer combinations on a set of 90 Complex Pooled Pools resulted in more than 800 BAC contigs being anchored. Currently, with analysis of the data of another 70 AFLP M+3/E+3 primer combinations nearly complete, more than 1300 contigs have been anchored to the ultra dense genetic map. Although we could argue that, with these 1300 contigs, the ultra dense genetic map has been covered by BAC contigs, we can not claim that our physical map

has been saturated with markers. In our current version of the physical map, approximately 6,300 contigs remain unanchored. We expect that we can reduce this number of unanchored contigs significantly by merging. Such merging is customarily performed using FPCs (Soderlund et al.1997, 2000) end-merge facility. Using this facility, fingerprint bands placed in the consensus band map (which shows the putative relative position within a BAC clone of each of its fingerprint fragments) at the end of a contig are used, at a lowered stringency to find matches with fingerprint bands of other contigs, without affecting the internal ordering of fingerprints within a contig. During our initial attempts to do so, we have observed that such mergers are not always possible and that some of the mergers suggested are contradicted by the contigs being anchored to different chromosomes. This can of course have several different reasons:

1. Contigs can be anchored to the wrong chromosome. While we have detected a few instances where candidate anchor points proved wrong, overall our anchoring procedure remains highly accurate.
2. BACs or fingerprints can be chimearic. We have detected a few cases were we suspect that either a BAC clone really is chimaeric or the fingerprint is a result of a mixture of BAC clones.
3. The consensus bands ordering algorithm tends to place problematic fingerprint bands, not finding a match with other fingerprint bands in overlapping clones, near the end of the contig. Such problematic bands can be caused by fingerprinting reproducibility issues (Chapter 2) or simpy by the fact that potato is a heterozygous organism, and fingerprints of BAC clones derived from the same locus on the different homologous chromosomes may therefore differ.
4. The distribution of fragment sizes in AFLP fingerprints is extremely skewed towards smaller fragments, and placement in the consensus band map of a number of small AFLP fragments near the end of a contig may result in false detection of overlap between the contigs.
5. There may be uncloned regions (gaps) in the BAC library or undetectably small overlaps between the fingerprints.

This makes us reluctant to execute contig mergers without corroborative evidence. Luckily such additional evidence is occasionally available, but most importantly, more is becoming available through the efforts of the Potato Genome Sequencing Consortium (PGSC, http://www.potatogenome.net):

1. Markers can anchor two contigs, suggesting their merger, which much be confirmed by their fingerprints.
2. Contigs or a chain of three or more contigs may be merged if contigs on either end contain markers from the same linkage group.

3. A BLAST hit of the sequence of a completely sequenced BAC in one contig with the BAC-end sequences in another contigs can suggest a merge which must then be confirmed by their fingerprints.

Both the anchoring of more contigs and the mapping of bridge markers through Binmap+ (chapter 4) will help us merge contigs. More anchored contigs means that more contigs may become be part of a chain containing multiple anchor points on the same chromosome, increasing confidence. More accurate mapping (of bridge markers in particular) will allow a better selection to be made from contigs that are candidates for mergers. For example if three contigs are located on a linkage group in bins 10, 11 and 12 respectively, then a successful direct merger between the contigs present in bin 10 and 12 is more unlikely than a merger incorporating the contig anchored to bin 11 in between. If, before application of the UMLPLE (Chapter 3) and the constraints placed on bridge markers (Chapter 4), the markers used to anchor these contigs were placed on a wider ranges of bins, for example bin 6 to 12, bin 11 and bin 8 to 12 respectively, then relative order of these contigs might have remained unresolved.

As discussed in Chapter 6, there are reasons to combine construction of a genome wide physical map with the genome-wide anchoring to a genetic map of its contigs. By extension, there is a point in case for the argument that de novo construction of a high density, genome wide genetic map should be combined with the de novo construction of a genome wide physical map because it allows selection of appropriate marker and BAC fingerprinting technologies. As we have demonstrated here, matching the marker and BAC fingerprinting technology allowed us to efficiently anchor contigs. Similarly, combination of a physical mapping effort with a sequencing project is an established sequencing strategy (Venter et al. 1996), delivering benefits to both. It is interesting to note that in silico AFLP with E+3/M+3 primers (Rombauts et al. 2003) on completed sequences of anchored BACs results in retrieval of the correct fragment (although rarely of the exact size predicted by observed mobility), allowing quick verification that the correct clone has been sequenced, even if BAC-end sequences are unavailable. For us, AFLP has been an enabling technology.

References

d'Alençon E., Piffanellim P., Volkoff A., Sabau X., Gimenez S., Rocher J., Cérutti P. and Fournier P. (2004): "A genomic BAC library and a new BAC-GFP vector to study the holocentric pest Spodoptera frugiperda", Insect Biochemistry and Molecular Biology, Volume 34, Issue 4, pp. 331-341.

Allard R. (1956): "Formulas and tables to facilitate the calculation of recombination values in heredity", Hilgardia, Volume 24, pp. 235-278.

Amos W., Hoffman J.I., Frodsham A., Zhang L., Best S. and Hill A.V.S. (2007): "Automated binning of microsatellite alleles: problems and solutions", Molecular Ecology Notes, Volume 7, Issue 1, pp. 10-14.

Aoki S. and Ito M. (2000): "Molecular Phylogeny of Nicotiana (Solanaceae) Based on the Nucleotide Sequence of the matK Gene.", Plant Biology, Volume 2, Issue 3, pp. 316-324.

Arumuganathan K. and Earle E.D. (1991): "Estimation of nuclear DNA content of plants by flow cytometry", Plant Molecular Biology Reporter, Volume 9, pp. 229-233.

Axenovich T.I. (1996): "Prediction of linkage phase by parental phenotypes.", Genetic Epidemiology, Volume 13, nr 3, pp. 271-283.

Bakker E., Butterbach P., Rouppe van der Voort J., van der Vossen E., van Vliet J., Bakker J. and Goverse A. (2003): "Genetic and physical mapping of homologues of the virus resistance gene Rx1 and the cyst nematode resistance gene Gpa2 in potato", Theoretical and Applied Genetics, Volume 106, Issue 8, pp. 1524-1531.

Ballvora A., Ercolano M.R., Weiß J., Meksem K., Bormann C., Oberhagemann P., Salamini F. and Gebhardt C. (2002): "The R1 gene for potato resistance to late blight (Phytophthora infestans) belongs to the leucine zipper/NBS/LRR class of plant resistance genes.", Plant Journal, Volume 30, pp. 361-371.

Bonierbale M., Plaisted R.L. and Tanksley S.D. (1988): "RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato", Genetics, Volume 120, pp. 1095-1103.

Bowling J.M., Bruner K.L., Cmarik J.L. and Tibbetts C. (1991): "Neighboring nucleotide interactions during DNA sequencing gel electrophoresis", Nucleic Acids Research, Volume 19, pp. 3089-3097.

Bradshaw J.E. and Ramsay G. (2005): "Utilisation of the Commonwealth Potato Collection in potato breeding", Euphytica, Volume 146, nr 1-2, pp. 9-19.

Brown D. and Vision T. (2000): "MapPop 1.0: Software for selective mapping and bin mapping." Computer program available from http://www.bio.unc.edu/faculty/vision/lab/mappop/. [Accessed 29 June 2008]

Brugmans B., Hutten R.G.B., Rookmaker A.N.O., Visser R.G.F. and van Eck H.J. (2006): "Exploitation of a marker dense linkage map of potato for positional cloning of a wart disease resistance gene." Theoretical and Applied Genetics, Volume 112, nr 2, pp. 269-277.

Budiman M.A., Mao L., Wood T.C. and Wing R.A. (2000): "A Deep-Coverage Tomato BAC Library and Prospects Toward Development of an STC Framework for Genome Sequencing", Genome Research, Volume 10, pp. 129-136.

Cai W.W., Reneker J., Chow C W., Vaishnav M. and Bradley A. (1998): "An anchored framework BAC map of mouse chromosome 11 assembled using multiplex oligonucleotide hybridization", Genomics, Volume 54, Issue 3, pp. 387-397.

Cai W.W., Chow C.W., Damani S., Gregory S.G., Marra M. and Bradley A. (2001): "An SSLP marker-anchored BAC framework map of the mouse genome", Nature Genetics, Volume 29, Issue 2, pp. 133-134.

Cardle L., Ramsay L., Milbourne D., Macaulay M., Marshall D. and Waugh R. (2000): "Computational and experimental characterization of physically clustered simple sequence repeats in plants", Genetics, Volume 156, Issue 2, pp. 847-854.

Cartwright D.A. (2007a): "Determination of marker phases in crosses with many offspring.", Genetics, Volume 176, nr 4, pp. 2637-2650.

Cartwright D.A., Troggio M., Velasco R. and Gutin A. (2007b): "Genetic mapping in the presence of genotyping errors.", Genetics, Volume 176, nr 4, pp. 2521-2527.

Chakravarti, Laha and Roy (1967): Handbook of Methods of Applied Statistics, Volume I, John Wiley and Sons , pp. 392-394.

Chat J., Decroocq S. and Petit R.J. (2003): "A One-Step Organelle Capture: Gynogenetic Kiwifruits with Paternal Chloroplasts.", Proceedings of the Royal Society of London: Biological Sciences, Volume 270, nr 1517, pp. 783-789.

Chen M., Presting G., Barbazuk W.B., Goicoechea J.L., Blackmon B., Fang G., Kim H., Frisch D., Yu Y., Sun S., Higingbottom S., Phimphilai J., Phimphilai D., Thurmond S., Gaudette B., Li P., Liu J., Hatfield J., Main D., Farrar K., Henderson C., Barnett L., Costa R., Williams B., Walser S., Atkins M., Hall C., Budiman M.A., Tomkins J.P., Luo M., Bancroft I., Salse J., Regad F., Mohapatra T., Singh N.K., Tyagi A.K., Soderlund C., Dean R.A. and Wing R.A. (2002): "An integrated physical and genetic map of the rice genome.", Plant Cell, Volume 14, nr 3, pp. 537-545.

Chen Q., Sun S., Ye Q., McCuine S., Huff E. and Zhang H.B. (2004): "Construction of two BAC libraries from the wild Mexican diploid potato, Solanum pinnatisectum, and the identification of clones near the late blight and Colorado potato beetle resistance loci.", Theoretical and Applied Genetics, Volume 108, nr 6, pp. 1002-1009.

Clarke L. and Carbon J. (1976): "A colony bank containing synthetic ColE1 hybrid plasmids representative of the entire *E. coli* genome", Cell, Volume 9, pp. 91-101.

Coe E., Cone K., McMullen M., Chen S.S., Davis G., Gardiner J., Liscum E., Polacco M., Paterson A., Sanchez-Villeda H., Soderlund C. and Wing R. (2002): "Access to the maize genome: an integrated physical and genetic map.", Plant Physiology, Volume 128, nr 1, pp. 9-12.

Cordier Y., Roch O., Cordier P. and Bischoff R. (1994): "Capillary gel electrophoresis of oligonucleotides: prediction of migration times using base- specific migration coefficients.", Journal of Chromatography A, Volume 680, Issue 2, pp. 479-489.

Coulson A., Sulston J., Brenner S. and Karn J. (1986): "Toward a physical map of the genome of the nematode Caenorhabditis elegans.", Proceedings of the National Academy of Sciences of the U S A , Volume 83, Issue 20, pp. 7821-7825.

Cox D.R., Burmeister M., Price E.R., Kim S. and Myers R.M. (1990): "Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes.", Science, Volume 250, Issue 4978, pp. 245-50.

Dempster A.P., Laird N.M. and Rubin D.B. (1977):"Maximum likelihood from incomplete data via EM algorithm", Journal of the Royal Statistical Society Series B, Volume 39, pp. 1-38.

Ding Y., Johnson M.D., Colayco R., Chen Y.J., Melnyk J., Schmitt H. and Shizuya H. (1999): "Contig assembly of bacterial artificial chromosome clones through multiplexed fluorescence-labeled fingerprinting", Genomics, Volume 56, Issue 3, pp. 237-246.

Ding Y., Johnson M.D., Chen W.Q., Wong D., Chen Y.J., Benson S.C., Lam J.Y., Kim Y.M. and Shizuya H. (2001): "Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases", Genomics, Volume 74, Issue 2, pp. 142-154.

Doehlonan J.M. and Sleper D.A. (1995): "Breeding field crops", Iowa state university press, Ames, pp. 419-433.

de Donato M., Gallagher D.S., Davis S.K., Ji Y., Burzlaff J.D., Stelly D.M., Womack J.E. and Taylor J.F. (1999): "Physical assignment of microsatellite-containing BACs to bovine chromosomes", Cytogenetics and Cell Genetics, Volume 87, pp. 59-61.

van Eck H.J., Rouppe van der Voort J., Draaistra J., van Zandvoort P. and van Enckevort E. (1995): "The inheritance and chromosomal localization of AFLP markers in a non-inbred potato offspring", Molecular Breeding, Volume 1, pp. 397-410.

Flibotte S., Chiu R., Fjell C., Krzywinski M., Schein J.E., Shin H. and Marra M.A. (2004): "Automated ordering of fingerprinted clones", Bioinformatics, Volume 20, Issue 8, pp. 1264-1271.

Fu H., Du J., Song J., Jiang J., and Park W.D. (2001): "Potato and tomato Forever Young genes contain class-I patatin promoter-like sequences.", Botanical Bulletin of Academia Sinica, Volume 42, pp. 231-241.

Fu H.H. and Dooner H.K. (2002): "Intraspecific violation of genetic colinearity and its implications in maize", Proceedings of the National Academy of Sciences of the U S A , Volume 99, Issue 14, pp. 9573-9578.

Gardiner J., Schroeder S., Polacco M.L., Villeda S., Fang Z.W., Morgante M., Landewe T., Fengler K., Useche F., Hanafey M., Tingey S., Chou H., Wing R., Soderlund C. and Coe E.H. (2004): "Anchoring 9,371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization", Plant Physiology, Volume 134, Issue 4, pp. 1317-1326.

Gebhardt C., Ritter E., Debener T., Schachtachabel U. and Walkemeijer B. (1989): "RFLP analysis and linkage mapping in Solanum tuberosum", Theoretical and Applied Genetics, Volume 78, pp. 65-75

Gebhardt C., Ritter E. and Salamini F. (2001): "RFLP map of the potato", in "DNA-Based Markers in Plants, Advances in Cellular and Molecular Biology of Plants", ed. Phillips R.L. and Vasil I.K., Vol. 6, Ed. 2, pp. 319-336. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Glavac D. and Dean M. (1993): "Optimization of the single-strand conformation polymorphism (SSCP) technique for detection of point mutations",Human Mutation, Volume 2, Issue 5, pp. 404-414.

Glendinning D.R. (1983): "Potato Introductions and Breeding up to the early 20th century", New Phytologist 1983, Volume 94, pp. 479-505.

Golumbic M.C., Kaplan H. and Shamir R. (1994): "On the complexity of DNA physical mapping", Advances in Applied Mathematics, Volume 15, Issue 3, pp. 251-261.

Grattapaglia D. and Sederoff R. (1994): "Genetic Linkage Maps of Eucalyptus grandis and Eucalyptus urophylla Using a Pseudo-Testcross: Mapping Strategy and RAPD Markers", Genetics , Volume 137, Issue 4, pp. 1121-1137.

Gregory S.G., Howell G.R. and Bentley D.R. (1997): "Genome mapping by fluorescent fingerprinting", Genome Research, Volume 7, Issue 12, pp. 1162-1168.

Guimarães P.M., Garsmeur O., Proite K., Leal-Bertioli S.C.M., Seijo G., Chaine C., Bertioli D.J. and D'Hont A. (2008): "BAC libraries construction from the ancestral diploid genomes of the allotetraploid cultivated peanut", BMC Plant Biology, January 29, 8:14.

Guttman A. and Cooke N. (1991): "Effect of temperature on the separation of DNA restriction fragments in capillary gel electrophoresis", Journal of Chromatography A , Volume 559, Issues 1-2, pp. 285-294.

Guttman A. (1996): "Effect of temperature on separation efficiency in capillary gel electrophoresis", Trends in Analytical Chemistry 1996, Volume 15, Issue 5, pp. 194-198.

Hadd A.G., Goard M.P., Rank D.R. and Jovanovich S.B. (2000): "Sub-microliter DNA sequencing for capillary array electrophoresis.", Journal of chromatography, Volume 894, nr 1-2, pp. 191-201.

Hahn M., Wilhelm J. and Pingoud A. (2001): "Influence of fluorophor dye labels on the migration behavior of polymerase chain reaction - amplified short tandem repeats during denaturing capillary electrophoresis", Electrophoresis, Volume 22, Issue 13, pp. 2691-2700.

Hamilton M.B. (1999): "Four primer pairs for the amplification of chloroplast intergenic region with intraspecific variation", Molecular Ecology, Volume 8, pp. 513-525.

Han T.H., van Eck H.J., De Jeu M.J. and Jacobsen E. (1999): "Optimization of AFLP fingerprinting of organisms with a large-sized genome: a study on Alstroemeria spp.", Theoretical and Applied Genetics, Volume 98, nr 3-4, pp. 465-471.

Han C.S., Sutherland R.D., Jewett P.B., Campbell M.L., Meincke L.J., Tesmer J.G., Mundt M.O., Fawcett J.J., Kim U.J., Deaven L.L. and Doggett N.A. (2000): "Construction of a BAC contig map of chromosome 16q by two-dimensional overgo hybridization", Genome Research, Volume 10, Issue 5, pp. 714-721.

Han Y., Gasic K., Marron B., Beever J.E. and Korban S.S. (2007): "A BAC-based physical map of the apple genome", Genomics, Volume 89, Issue 5, pp. 630-637.

Hansen M., Kraft T., Christiansson M. and Nilsson N.O. (1998): "Evaluation of AFLP in Beta", Theoretical and Applied Genetics, Volume 98, nr 6-7, pp. 845-852.

Hein I., McLean K., Chalhoub B. and Bryan G.J. (2007): "Generation and Screening of a BAC Library from a Diploid Potato Clone to Unravel Durable Late Blight Resistance on Linkage Group IV", International Journal of Plant Genomics, Article-id 51421, 5 pages.

Hermsen J.G.T. and Verdenius J. (1973): "Selection from Solanum tuberosum group phureja of genotypes combining high-frequency haploid induction with homozygosity for embryo-spot", Euphytica, Volume 22, pp. 244-259.

Hoheisel J., Maier E., Mott R., McCarthy L., Grigoriev A., Schalkwyk L., Nizetic D., Francis F. and Lehrach H. (1993): "High resolution cosmid and P1 maps spanning the 14 Mb genome of the fission yeast S. pombe.", Cell, Volume 73, Issue 1, pp. 109-120.

Hong G.F. (1997): "A rapid and accurate strategy for rice contig map construction by combination of fingerprinting and hybridization", Plant Molecular Biology , Volume 35, pp. 129-133.

Horton D.E. (1988): "Potato: truly a world crop". SPAN 1988, Volume 30, Issue 3, pp. 116-118.

Huang S.W., van der Vossen E.A.G., Kuang H.H., Vleeshouwers V.G.A.A., Zhang N.W., Borm T.J.A., van Eck H.J., Baker B., Jacobsen E. and Visser R.G.F. (2005): "Comparative genomics enabled the isolation of the R3a late blight resistance gene in potato", Plant Journal, Volume 42, Issue 2, pp. 251-261.

Ihaka R. and Gentleman R. (1996): "R: A Language for Data Analysis and Graphics", Journal of Computational and Graphical Statistics, Volume 5, nr 3, pp. 299-314.

Islam-Faridi M.N., Childs K.L., Klein P.E., Hodnett G., Menz M.A., Klein R.R., Rooney W.L., Mullet J.E., Stelly D.M. and Price H.J. (2002): "A molecular cytogenetic map of sorghum chromosome 1: Fluorescence in situ hybridization analysis with mapped bacterial artificial chromosomes", Genetics, Volume 161, Issue 1, pp. 345-353.

Jacobs J.M.E., van Eck H.J., Arens P., Verkerk-Bakker B., te Lintel Hekkert B., Bastiaanssen H.J.M., El Kharbotly A., Pereira A., Jacobsen E. and Stiekema W.J. (1995): "A genetic map of potato Solanum tuberosum integrating molecular markers, including transposons, and classical markers", Theoretical and Applied Genetics, Volume 91, pp. 289-300.

Jansen R.C., Geerlings H., van Oeveren A.J. and van Schaik R.C. (2001): "A Comment on Codominant Scoring of AFLP Markers", Genetics, Volume 158, pp. 925-926.

Katagiri T., Kidd C., Tomasino E., Davis J.T., Wishon C., Stern J.E., Carleton K.L., Howe A.E. and Kocher T.D. (2005): "A BAC-based physical map of the Nile tilapia genome", BMC Genomics, Volume 6, 89.

Kelleher C.T., Chiu R., Shin H., Bosdet I.E., Krzywinski M.I., Fjell C.D., Wilkin J., Yin T.M., DiFazio S.P., Ali J., Asano J.K., Chan S., Cloutier A., Girn N., Leach S., Lee D., Mathewson C.A., Olson T., O'Connor K., Prabhu A.L., Smailus D.E., Stott J.M., Tsai M., Wye N.H., Yang G.S., Zhuang J., Holt R.A., Putnam N.H., Vrebalov J., Giovannoni J.J., Grimwood J., Schmutz J., Rokhsar D., Jones S.J.M., Marra M.A., Tuskan G.A., Bohlmann J., Ellis B.E., Ritland K., Douglas C.J. and Schein J.E. (2007): "A physical map of the highly heterozygous Populus genome: integration with the genome sequence and genetic map and analysis of haplotype variation", Plant Journal, Volume 50, 6, pp. 1063-1078.

Kim J.S., Klein P.E., Klein R.R., Price H.J., Mullet J.E. and Stelly D.M. (2005): "Chromosome identification and nomenclature of Sorghum bicolor", Genetics, Volume 169, Issue 2, pp. 1169-1173.

Kirov G., Williams N., Sham P., Craddock N. and Owen: M.J. (2000) "Pooled Genotyping of Microsatellite Markers in Parent-Offspring Trios", Genome Research, Volume 10, Issue 1, pp. 105-115.

Klein P.E., Klein R.R., Cartinhour S.W., Ulanch P.E., Dong J.M., Obert J.A., Morishige D.T., Schlueter S.D., Childs K.L., Ale M. and Mullet J.E. (2000): "A high-throughput AFLP-based method for constructing integrated genetic and physical maps: Progress toward a sorghum genome map", Genome Research, Volume 10, Issue 6, pp. 789-807.

Kleparnik K., Foret F., Berka J., Goetzinger W., Miller A.W. and Karger B.L. (1996): "The use of elevated column temperature to extend DNA sequencing read lengths in capillary electrophoresis with replaceable polymer matrices", Electrophoresis, Volume 17, pp. 1860-1866.

Konrad K.D. and Pentoley S.L. (1993): "Contribution of secondary structure to DNA mobility in capillary gels", Electrophoresis, Volume 14, Issue 5-6, pp. 502-508.

Koo D.H., Jo S.H., Bang J., Park H., Lee S. and Choi D. (2008): "Integration of Cytogenetic and Genetic Linkage Maps Unveils the Physical Architecture of Tomato Chromosome 2.", Genetics, Volume 179, Issue 3, pp. 1211-1220.

Koopman W.J. and Gort G. (2004): "Significance tests and weighted values for AFLP similarities, based on Arabidopsis in silico AFLP fragment length distributions.", Genetics, Volume 167, nr 4, pp. 1915-1928.

Koumi P., Green H.E., Hartley S., Jordan D., Lahec S., Livett R.J., Tsang K.W. and Ward D.M. (2004): "Evaluation and validation of the ABI 3700, ABI 3100 and the MegaBACE 1000 capillary array electrophoresis instruments for use with short tandem repeat microsatellite typing in a forensic environment", Electrophoresis, Volume 25, Issue 14, pp. 2227-2241.

Kozlowski P. and Krzyzosiak W.J. (2005): "Structural factors determining DNA length limitations in conformation-sensitive mutation detection methods", Electrophoresis, Volume 26, Issue 1, pp. 71-81.

Kuang H., Wei F., Marano M.R., Wirtz U., Wang X., Liu J., Shum W.P., Zaborsky J., Tallon L.J., Rensink W., Lobst S., Zhang P., Tornqvist C.E., Tek A., Bamberg J., Helgeson J., Fry W., You F., Luo M.C., Jiang J., Buell R.C. and Baker B. (2005): "The R1 resistance cluster contains three groups of independently evolving, Type I R1 homologues and shows substantial structural variation among haplotypes of Solanum demissum.", Plant Journal, Volume 44, pp. 37-51.

Kwitek 2004: Kwitek A.E., Gullings-Handley J., Yu J., Carlos D.C., Orlebeke K., Nie J., Eckert J., Lemke A., Wendt-Andrae J., Bromberg S., Pasko D., Chen D., Scheetz T.E., Casavant T.L., Bento-Soares M., Sheffield V.C., Tonellato P.J. and Jacob H.J. (2004): "High-Density Rat Radiation Hybrid Maps Containing Over 24,000 SSLPs, Genes, and ESTs Provide a Direct Link to the Rat Genome Sequence", Genome Research, Volume 14, pp. 750-757.

Lai 2006: Lai C.W.J., Yu Q.Y., Hou S.B., Skelton R.L., Jones M.R., Lewis K.L.T., Murray J., Eustice M., Guan P.Z., Agbayani R., Moore P.H., Ming R. and Presting G.G. (2006): "Analysis of papaya BAC end sequences reveals first insights into the organization of a fruit tree genome", Molecular Genetics and Genomics, Volume 276, Issue 1, pp. 1-12

Lander E.S. and Waterman M.S. (1988): "Genomic mapping by fingerprinting random clones: a mathematical analysis.", Genomics, Volume 2, Issue 3, pp. 231-239

Landergott 2006: Landergott U., Naciri Y., Schneller J.J. and Holderegger R. (2006): "Allelic configuration and polysomic inheritance of highly variable microsatellites in tetraploid gynodioecious Thymus praecox agg.", Theoretical and Applied Genetics, Volume 113, Number 3, pp. 453-465.

Lazaruk 1998: Lazaruk K., Walsh P.S., Oaks F., Gilbert D., Rosenblum B.B., Menchen S., Scheibler D., Wenz H.M., Holt C. and Wallin J. (1998):"Genotyping of forensic short tandem repeat (STR) systems based on sizing precision in a capillary electrophoresis instrument", Electrophoresis, Volume 19, Issue 1, pp. 86-93.

Lin Y.R., Chow T.Y., Luo M., Kudrna D., Lin C.C., Wing R.A. and Hsing Y.I.C. (2006): "Two highly representative rice BAC libraries of japonica cv Tainung 67 suitable for rice structural and functional genomic research", Plant Science, Volume 170, Issue 4, pp. 889-896.

Luo M.C., Thomas C., You F.M., Siao J., Shu O.Y., Buell C.R., Malandro M., McGuire P.E., Anderson O.D. and Dvorak J. (2003): "High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis", Genomics, Volume 82, Issue 3, pp.378-389

Luo S., Hall A.E., Hall S.E. and Preuss D. (2004): "Whole-genome fractionation rapidly purifies DNA from centromeric regions", Nature Methods, Volume 1, Issue 1, pp. 67-71

Luo Z.W., Zhang Z., Leach L., Zhang R.M., Bradshaw J.E. and Kearsey M.J. (2006): "Constructing Genetic Linkage Maps Under a Tetrasomic Model.", Genetics, Volume 172, pp. 2635-2645.

Mahairas G.G., Wallace J.C., Smith K., Swartzell S., Holzman T., Keller A., Shaker R., Furlong J., Young J., Zhao S.Y., Adams M.D. and Hood L. (1999): "Sequence-tagged connectors: A sequence approach to mapping and scanning the human genome", Proceedings of the National Academy of Sciences of the U S A , Volume 96, Issue 17, pp. 9739-9744

Maliepaard C., Jansen J. and van Ooijen J.W. (1997): "Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications", Genetics Research Cambridge, Volume 70, pp. 237-250.

Mardis E.R. (2006): "Anticipating the $1,000 genome", Genome Biology, Volume 7, Issue 7, pp. 112.

Marra M.A., Kucaba T.A., Dietrich N.L., Green E.D., Brownstein B., Wilson R.K., McDonald K.M., Hillier L.W., McPherson J.D. and Waterston R.H. (1997): "High throughput fingerprint analysis of large-insert clones", Genome Research, Volume 7, Issue 11, pp. 1072-1084

Mather K. (1938): "The Measurement of Linkage in Heredity". Methuen & Co., London.

Matsumoto T., Wu J.Z., Kanamori H., Katayose Y., Fujisawa M., Namiki N., Mizuno H., Yamamoto K., Antonio B.A., Baba T., Sakata K., Nagamura Y., Aoki H., Arikawa K., Arita K., Bito T., Chiden Y., Fujitsuka N., Fukunaka R., Hamada M., Harada C., Hayashi A., Hijishita S., Honda M., Hosokawa S., Ichikawa Y., Idonuma A., Iijima M., Ikeda M., Ikeno M., Ito K., Ito S., Ito T., Ito Y., Ito Y., Iwabuchi A., Kamiya K., Karasawa W., Kurita K., Katagiri S., Kikuta A., Kobayashi H., Kobayashi N., Machita K., Maehara T., Masukawa M., Mizubayashi T., Mukai Y., Nagasaki H., Nagata Y., Naito S., Nakashima M., Nakama Y., Nakamichi Y., Nakamura M., Meguro A., Negishi M., Ohta I., Ohta T., Okamoto M., Ono N., Saji S., Sakaguchi M., Sakai K., Shibata M., Shimokawa T., Song J.Y., Takazaki Y., Terasawa K., Tsugane M., Tsuji K., Ueda S., Waki K., Yamagata H., Yamamoto M., Yamamoto S., Yamane H., Yoshiki S., Yoshihara R., Yukawa K., Zhong H.S., Yano M., Sasaki T., Yuan Q.P., Shu O.T., Liu J., Jones K.M., Gansberger K., Moffat K., Hill J., Bera J., Fadrosh D., Jin S.H., Johri S., Kim M., Overton L., Reardon M., Tsitrin T., Vuong H., Weaver B., Ciecko A., Tallon L., Jackson J., Pai G., Van Aken S., Utterback T., Reidmuller S., Feldblyum T., Hsiao J., Zismann V., Iobst S., de Vazeille A.R., Buell C.R., Ying K., Li Y., Lu T.T., Huang Y.C., Zhao Q., Feng Q., Zhang L., Zhu J.J., Weng Q.J., Mu J., Lu Y.Q., Fan D.L., Liu Y.L., Guan J.P., Zhang Y.J., Yu S.L., Liu X.H., Zhang Y., Hong G.F., Han B., Choisne N., Demange N., Orjeda G., Samain S., Cattolico L., Pelletier E., Couloux A., Segurens B., Wincker P., D'Hont A., Scarpelli C., Weissenbach J., Salanoubat M., Quetier F., Yu Y., Kim H.R., Rambo T., Currie J., Collura K., Luo M.Z., Yang T.J., J. Ammiraju SS., Engler F., Soderlund C., Wing R.A., Palmer L.E., de la Bastide M., Spiegel L., Nascimento L., Zutavern T., O'Shaughnessy A., Dike S., Dedhia N., Preston R., Balija V., McCombie W.R., Chow T.Y., Chen H.H., Chung M.C., Chen C.S., Shaw J.F., Wu H.P., Hsiao K.J., Chao Y.T., Chu M.K., Cheng C.H., Hour A.L., Lee P.F., Lin S.J., Lin Y.C., Liou J.Y., Liu S.M., Hsing Y.I., Raghuvanshi S., Mohanty A., Bharti A.K., Gaur A., Gupta V., Kumar D., Ravi V., Vij S., Kapur A., Khurana P., Khurana P., Khurana J.P., Tyagi A.K., Gaikwad K., Singh A., Dalal V., Srivastava S., Dixit A., Pal A.K., Ghazi I.A., Yadav M., Pandit A., Bhargava A., Sureshbabu K., Batra K., Sharma T.R., Mohapatra T., Singh N.K., Messing J., Nelson A.B., Fuks G., Kavchok S., Keizer G., E. Llaca LV., Song R.T., Tanyolac B., Young S., Il K.H., Hahn J.H., Sangsakoo G., Vanavichit A., de L. Mattos AT., Zimmer P.D., Malone G., Dellagostin O., de Oliveira A.C., Bevan M., Bancroft I., Minx P., Cordum H., Wilson R., Cheng Z.K., Jin W.W.,

Jiang J.M., Leong S.A., Iwama H., Gojobori T., Itoh T., Niimura Y., Fujii Y., Habara T., Sakai H., Sato Y., Wilson G., Kumar K., McCouch S., Juretic N., Hoen D., Wright S., Bruskiewich R., Bureau T., Miyao A., Hirochika H., Nishikawa T., Kadowaki K. and Sugiura M. (2005): "The map-based sequence of the rice genome", Nature, Volume 436, Issue 7052, pp. 793-800.

McGrath J.M., Shaw R.S., de los Reyes B.G. and Weiland J.J. (2004): "Construction of a Sugar Beet BAC Library From a Hybrid With Diverse Traits", Plant Molecular Biology Reporter, Volume 22, nr 1, pp. 23-28.

Meudt H.M. and Clarke A.C. (2007): "Almost forgotten or latest practice? AFLP applications, analyses and advances.", Trends in Plant Science, Volume 12, Issue 3, pp. 106-117.

Meyers B.C., Scalabrin S. and Morgante M. (2004): "Mapping and sequencing complex genomes: Let's get physical!", Nature Reviews Genetics, Volume 5, Issue 8, pp. 578-U1.

Milbourne D., Meyer R.C., Collins A.J., Ramsay L.D. and Gebhardt C. (1998): "Isolation, characterisation and mapping of simple sequence repeat loci in potato", Molecular and general genetics, Volume 259, pp. 233-245.

Mozo T., Fischer S., Ewert M., Lehrach H. and Altmann T. (1998): Use of the IGF BAC library for physical mapping of the Arabidopsis thaliana genome", Plant Journal, Volume 16, Issue 3, pp. 377-384

Mozo T., Dewar K., Dunn P., Ecker J.R., Fischer S., Kloska S., Lehrach H., Marra M., Martienssen R., Ewert M. and Altmann T. (1999): "A complete BAC-based physical map of the Arabidopsis thaliana genome", Nature Genetics, Volume 22, Issue 3, pp. 271-275.

Mun J.H., Kwon S.J., Yang T.J., Kim H.S., Choi B.S., Baek S., Kim J.S., Jin M., Kim J.A., Lim M.H., Lee S.I., Kim H.I., Kim H., Lim Y.P. and Park B.S. (2008): "The first generation of a BAC-based physical map of Brassica rapa", BMC Genomics, Volume 9, 280.

Nelson W.M., Bharti A.K., Butler E., Wei F.S., Fuks G., Kim H., Wing R.A., Messing J. and Soderlund C. (2005): "Whole-genome validation of high-information-content fingerprinting", Plant Physiology, Volume 139, Issue 1, pp. 27-38.

Nelson W.M., Dvorak J., Luo M.C., Messing J., Wing R.A. and Soderlund C. (2007): "Efficacy of clone fingerprinting methodologies", Genomics, Volume 89, Issue 1, pp. 160-165.

Noll B.O., Debelak H. and Uhlmann E. (2007): "Identification and quantification of GC-rich oligodeoxinucletides in tissue extracts by capillary gel electrophoresis", Journal of Chromatography B, Volume 847, pp. 153-161.

O'Hanlon P.C. and Peakall R. (2000): "A simple method for the detection of size homoplasy among amplified fragment length polymorphism fragments.", Molecular Ecology, Volume 9, Issue 6, pp. 815-816.

Ortiz R. and Peloquin S.J. (1994): "Use of 24-Chromosome Potatoes (Diploids and Dihaploids) for Genetical Anlysis", In "Potato Genetics", Bradshaw J.E., Mackay G.R. (ed.) CAB International, Wallingford, UK, pp 133-172.

van Os H., Stam P., Visser R.G.F. and van Eck H.J. (2005a): "SMOOTH: a statistical method for successful removal of genotyping errors from high-density genetic linkage data.", Theoretical and Applied Genetics, Volume 112, Issue 1, pp. 187-194.

van Os H., Stam P., Visser R.G.F. and van Eck H.J. (2005b): "RECORD: a novel method for ordering loci on a genetic linkage map.",Theoretical and Applied Genetics, Volume 112, Issue 1, pp. 30-40.

van Os H., Andrzejewski S., Bakker E., Barrena I., Bryan G.J., Caromel B., Ghareeb B., Isidore E., de Jong W., van Koert P., Lefebvre V., Milbourne D., Ritter E., Rouppe van der Voort J.N.A.M., Rousselle-Bourgeois F., van Vliet J., Waugh R., Visser R.G.F., Bakker J. and van Eck H.J. (2006): "Construction of a 10,000-Marker Ultra-dense Genetic Recombination Map of Potato: Providing a Framework for Accelerated Gene Isolation and a Genome-wide Physical Map", Genetics, Volume 173, pp. 1075-1087.

Osoegawa K., Zhu B.L., Shu C.L., Ren T., Cao Q., Vessere G.M., Lutz M.M., Seaman J., Zhao S.Y. and de Jong P.J. (2004): "BAC resources for the rat genome project", Genome Research, 14, Volume 4, pp. 780-785.

Paal J., Henselewski H., Muth J., Meksem K., Menéndez C.M., Salamini F., Ballvora A. and Gebhardt C. (2004): "Molecular cloning of the potato Gro1-4 gene conferring resistance to pathotype Ro1 of the root cyst nematode Globodera rostochiensis, based on a candidate gene approach", Plant Journal, Volume 38, nr 2, pp. 285-297.

Paux E., Legeai F., Guilhot N., Adam-Blondon A.F., Alaux M., Salse J., Sourdille P., Leroy P. and Feuillet C. (2008): "Physical mapping in large genomes: accelerating anchoring of BAC contigs to genetic maps through in silico analysis", Functional & Integrative Genomics, Volume 8, pp. 29-32.

Pöltl R., Luckenbach C., Fimmers R. and Ritter H. (1997): "Typing of the short tandem repeat D8S347 locus with different fluorescence markers", Electrophoresis, Volume18, Issue15, pp. 2871-2873.

Quiniou S.M.A., Waldbieser G.C. and Duke M.V. (2007): "A first generation BAC-based physical map of the channel catfish genome", BMC Genomics, Volume 8, 40.

Ratnayaka I., Båga M., Fowler D.B. and Chibbar R.N. (2005): "Construction and Characterization of a BAC Library of a Cold-Tolerant Hexaploid Wheat Cultivar", Crop Science, Volume 45, pp. 1571-1577.

Reid A. and Kerr E.M. (2007): "A rapid simple sequence repeat (SSR)-based identification method for potato cultivars.", Plant Genetic Resources, Volume 5, nr 1, pp. 7-13.

Ren C.W., Lee M.K., Yan B., Ding K.J., Cox B., Romanov M.N., Price J.A., Dodgson J.B. and Zhang H.B. (2003): "A BAC-based physical map of the chicken genome", Genome Research, Volume 13, Issue 12, pp. 2754-2758.

Ritter E., Gebhardt C. and Salamini F. (1990): "Estimation of Recombination Frequencies and Construction of RFLP Linkage Maps in Plants From Crosses Between Heterozygous Parents", Genetics, Vol 125, nr 3, pp. 645-654.

Ritter E. and Salamini F. (1996): "The calculation of recombination frequencies in crosses of allogamous plant species with applications to linkage mapping.", Genetical Research, Volume 67, pp. 55-65.

Romanov M.N., Price J.A. and Dodgson J.B. (2003): "Integration of animal linkage and BAC contig maps using overgo hybridization", Cytogenetic and Genome Research, Volume 102, Issue 1-4, pp. 277-281.

Rombauts S., van de Peer Y. and Rouzé P. (2003): "AFLPinSilico, simulating AFLP fingerprints", Bioinformatics, Volume 19, no. 6, pp. 776-777.

Rosa G.J., Yandell B.S. and Gianola D. (2002): "A Bayesian approach for constructing genetic maps when markers are miscoded.", Genetics, selection, evolution, Volume 34, pp. 353-369.

Rosenblum B.B., Oaks F., Menchen S. and Johnson B. (1997): "Improved single-strand DNA sizing accuracy in capillary electrophoresis", Nucleic Acids Research, Volume 25, nr. 19, pp. 3925-3929.

Ross M.T., LaBrie S., McPherson J. and Stanton V.P. (1999): "Screening large-insert libraries by hybridization.", In Current protocols in human genetics John Wiley and Sons, New York.(eds. Dracopoli N.C., Haines J.L., Korf B.R., Moir D.T., Morton C.C., Seidman C.E., Seidman J.G. and Smith D.R.), pp. 5.6.1-5.6.52.

Rouppe van der Voort J.N.A.M., Wolters P., Folkertsma R., Hutten R., van Zandvoort P., Vinke H., Kanyuka K., Bendahmane A., Jacobsen E., Jansen R. and Bakker J. (1997): "Mapping of the cyst nematode resistance locus Gpa2 in potato using a strategy based on comigrating AFLP markers.", Theoretical and Applied Genetics, Volume 95, pp. 874-880.

Rouppe van der Voort J., Lindeman W., Folkertsma R., Hutten R., Overmars H., van der Vossen E., Jacobsen E. and Bakker J. (1998): "A QTL for broad spectrum resistance to cyst nematode species (Globodera spp.) map to resistance gene clusters in potato.", Theoretical and Applied Genetics, Volume 96, nr 5, pp. 654-661.

Rouppe van der Voort J., van der Vossen E., Bakker E., Overmars H., van Zandvoort P., Hutten R., Klein Lankhorst R. and Bakker J. (2000): "Two additive QTLs conferring broad-spectrum resistance in potato to Globodera pallida are localized on resistance gene clusters", Theoretical and Applied Genetics, Volume 101, nr 7, pp. 1122-1130.

Sambrook J., Fritsch E.F. and Maniatis T. (1989): "Molecular Cloning - A Laboratory Manual, 2nd Edition.", Cold Spring Habour Laboratory Press, New York.

de Scenzo R.A. and Wise R.P. (1996): "Variation in the ratio of physical to genetic distance in intervals adjacent to the Mla locus on barley chromosome 1H.", Molecular & general genetics, Volume 251, pp. 472-482.

Scott G.J., Bers R., Rosegrat M. and Bokanga M. (2000): "Root and tubers in the global food system: A Vision statement to the year 2020". Co-publication of the Centro Internacional de la Papa (CIP), Centro Internacional de Agricultura Tropical (CIAT), International Food Policy Research Institute (IFPRI), International Institute of Tropical Agriculture (IITA) and International Plant Genetic Resources Institute (IPGRI), 27 pp.

Sgueglia J.B., Geiger S. and Davis J. (2003): "Precision studies using the ABI Prism 3100 Genetic Analyzer for forensic DNA analysis", Analytical and Bioanalytical chemistry, Volume 376, Issue 8, pp. 1247-1254.

Shizuya H., Birren B., Kim U.J., Mancino V., Slepak T., Tachiiri Y. and Simon M. (1992): "Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector", Proceedings of the National Academy of Sciences of the U S A, Volume 89, nr 18, pp. 8794-8797.

Siegel A.F., Trask B., Roach J.C., Mahairas G.G., Hood L. and van den Engh G. (1999): "Analysis of sequence-tagged-connector strategies for DNA sequencing", Genome Research, Volume 9, Issue 3, pp. 297-307.

Soderlund C., Longden I. and Mott R. (1997): "FPC: a system for building contigs from restriction fingerprinted clones", Computer Applications in the Biosciences, Volume 13, Issue 5, pp. 523-535.

Soderlund C., Humphray S., Dunham A. and French L. (2000): "Contigs built with fingerprints, markers, and FPCV4.7", Genome Research, Volume 10, Issue 11, pp. 1772-1787.

Song J., Dong F. and Jiang J. (2000): "Construction of a bacterial artificial chromosome (BAC) library for potato molecular cytogenetics research", Genome, Volume 43, nr 1, pp. 199-204.

Srinivasan J., Sinz W., Jesse T., Wiggers-Perebolte L., Jansen K., Buntjer J., van der Meulen M. and Sommer R.J. (2003): "An integrated physical and genetic map of the nematode Pristionchus pacificus", Molecular Genetics and Genomics , Volume 269, nr 5, pp. 715-722.

Stam P. (1993): "Construction of integrated genetic linkage maps by means of a new computer package: JoinMap." The Plant Journal, Volume 5, nr 3, pp. 739-744.

Stam P. (1995): "JoinMap 2.0 deals with all types of plant mapping populations." ,Plant Genome III Abstracts, World Wide Web site: www.intl-pag.org.

Stuber C.W., Polacco M. and Lynn M. (1999): "Synergy of Empirical Breeding, Marker-Assisted Selection and Genomics to Incrase Crop Yield Potantial", Crop Science, Volume 39, pp. 1571-1583.

Stupar R.M., Lilly J.W., Town C.D., Cheng Z., Kaul S., Buell C.R. and Jiang J. (2001): "Complex mtDNA constitutes an approximate 620-kb insertion on Arabidopsis thaliana chromosome 2: Implication of potential sequencing errors caused by large-unit repeats", Proceedings of the National Academy of Sciences of the U S A , Volume 98, nr 9, pp. 5099-5103.

Sturtevant A.H. (1913): "The Linear Arrangement of Six Sex-linked Factors in Drosophila, as Shown by their Mode of Association", Journal of experimental Zoology, Volume 14, pp. 43-59.

Sugiyama Y., Watase Y., Nagase M., Makita N., Yagura S., Hirai A. and Sugiura M. (2005): "The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants.", Molecular Genetics and Genomics, Volume 272, nr 6, pp. 603-615.

Sulston J., Mallett F., Staden R., Durbin R., Horsnell T. and Coulson A. (1988): "Software for genome mapping by fingerprinting techniques", Computer Applications in the Biosciences, Volume 4, Issue 1, pp. 125-132.

Sun Z., Wang Z., Tu J., Zhang J., Yu F., McVett P.B. and Li G. (2007): "An ultradense genetic recombination map for Brassica napus, consisting of 13551 SRAP markers.", Theoretical and Applied Genetics, Volume 114, Issue 8, pp. 1305-1317.

Taberlet P., Gielly L., Pautou G. and Bouvet J. (1991): "Universal primers for amplification of three non-coding regions of chloroplast DNA", Plant Molecular Biology, Volume 17, nr 5, pp. 1105-1109.

Tao Q.Z., Chang Y.L., Wang J.Z., Chen H.M., Islam-Faridi M.N., Scheuring C., Wang B., Stelly D.M. and Zhang H.B. (2001): "Bacterial artificial chromosome-based physical map of the rice genome constructed by restriction fingerprint analysis", Genetics, Volume 158, Issue 4, pp. 1711-1724

Tartof K.D. and Hobbs C.A. (1987): "Improved media for growing plasmid and cosmid clones.", Focus, Volume 9, p. 12.

Tomkins J.P., Davis G., Main D., Yim Y., Duru N., Musket T., Goicoechea J.L., Frisch D.A., Coe E.H. and Wing R.A. (2002): "Construction and Characterization of a Deep-Coverage Bacterial Artificial Chromosome Library for Maize", Crop Science, Volume 42, pp. 928-933.

Uijtewaal B.A., Jacobsen E. and Hermsen T.J.G. (1987): "Morphology and vigour of monohaploid potato clones, their corresponding homozygous diploids and tetraploids and their heterozygous diploid parent", Euphytica, Volume 36, nr 3, pp. 745-753.

Vemireddy L.R., Archak S. and Nagaraju J. (2007): "Capillary Electrophoresis Is Essential for Microsatellite Marker Based Detection and Quantification of Adulteration of Basmati Rice (Oryza sativa)", Journal of Agricultural Food Chemistry, Volume 55, nr 20, pp. 8112-8117.

Venter J.C., Smith H.O. and Hood L. (1996): "A new strategy for genome sequencing", Nature, Volume 381, pp. 364-365.

Vision T.J., Brown D.G., Shmoys D.B., Durrett R.T. and Tanksley S.D. (2000): "Selective mapping: a strategy for optimizing the construction of high-density linkage maps.", Genetics, Volume 155, Issue 1, pp. 407-420.

Vos P., Hogers R., Bleeker M., Reijans M., van de Lee T., Hornes M., Frijters A., Pot J., Peleman J. and Kuiper M. (1995): "AFLP: a new technique for DNA fingerprinting.", Nucleic Acids Research, Volume 23, nr 21, pp. 4407-4414.

van der Vossen E.A.G., Rouppe van der Voort J.N.A.M., Kanyuka K., Bendahmane A., Sandbrink H., Baulcombe D.C., Bakker J., Stiekema W.J. and Klein-Lankhorst R.M. (2000): "Homologues of a single resistance-gene cluster in potato confer resistance to distinct pathogens: a virus and a nematode", The Plant Journal, Volume 23, nr 5, pp. 567-576.

v**an der Vossen E., Sikkema A., te Lintel Hekkert B., Gros J., Stevens P., Muskens M., Wouters D., Pereira A., Stiekema W. and Allefs S.** (2003): "An ancient R gene from the wild potato species Solanum bulbocastanum confers broad-spectrum resistance to Phytophthora infestans in cultivated potato and tomato", The Plant Journal, Volume 36, nr 6, pp. 867-882.

Wei F., Coe E., Nelson W., Bharti A.K., Engler F., Butler E., Kim H., Goicoechea J.L., Chen M., Lee S., Fuks G., Sanchez-Villeda H., Schroeder S., Fang Z., McMullen M., Davis G., Bowers J.E., Paterson A.H., Schaeffer M., Gardiner J., Cone K., Messing J., Soderlund C. and Wing R.A. (2007): "Physical and genetic structure of the maize genome reflects its complex evolutionary history", PLOS Genetics, Volume 3, Issue 7, pp. 1254-1263.

Wendl, M.C. (2005): "Probabilistic Assessment of Clone Overlaps in DNA Fingerprint Mapping via a priori Models'", Journal of Computational Biology, Volume 12, nr 3, pp. 283-297.

Wenz H.M., Robertson J.M., Menchen S., Oaks F., Demorest D.M., Scheibler D., Rosenblum B.B., Wike C., Gilbert D.A. and Efcavitch J.W. (1998): "High-precision genotyping by denaturing capillary electrophoresis", Genome Research, Volume 8, Issue 1, pp. 69-80.

Wicker T., Schlagenhauf E., Graner A., Close T.J., Keller B. and Stein N. (2006): "454 sequencing put to the test using the complex genome of barley", BMC Genomics, Volume 7, 257.

Wu R., Ma C.X., Wu S.S. and Zheng Z.B. (2002a): "Linkage mapping of sex-specific differences.", Genetics Research Cambridge 2002, Volume 79, pp. 85-96.

Wu R., Ma C.X., Painter I. and Zeng Z.B. (2002b): "Simultaneous Maximum Likelihood Estimation of Linkage and Linkage Phases in Outcrossing Species", Theoretical Population Biology, Volume 61, Issue 3, pp. 349-363.

Wu C.C., Nimmakayala P., Santos F.A., Springman R., Scheuring C., Meksem K., Lightfoot D.A. and Zhang H.B. (2004): "Construction and characterization of a soybean bacterial artificial chromosome library and use of multiple complementary libraries for genome physical mapping.", Theoretical and Applied Genetics, Volume 109, nr 5, pp. 1041-1050.

Xu Z.Y., Sun S.K., Covaleda L., Ding K., Zhang A.M., Wu C.C., Scheuring C. and Zhang H.B. (2004): "Genome physical mapping with large-insert bacterial clones by fingerprint analysis: methodologies, source clone genome coverage, and contig map quality", Genomics, Volume 84, Issue 6, pp. 941-951.

Yim Y.S., Davis G.L., Duru N.A., Musket T.A., Linton E.W., Messing J.W., McMullen M.D., Soderlund C.A., Polacco M.L., Gardiner J.M. and Coe E.H. (2002): "Characterization of Three Maize Bacterial Artificial Chromosome Libraries toward Anchoring of the Physical Map to the Genetic Map Using High-Density Bacterial Artificial Chromosome Filter Hybridization", Plant Physiology, Volume 130, pp. 1686-1696.

Yin Z., Severin J., Giddings M.C., Huang W.A., Westphall M.S. and Smith L.M. (1996): "Automatic matrix determination in four dye fluorescence-based DNA sequencing.", Electrophoresis, Volume 17, nr 6, pp. 1143-1150.

Zhang H.B. and Wu C.C. (2001): "BAC as tools for genome sequencing", Plant Physiology and Biochemistry, Volume 39, pp. 195-209.

Zhu W., Ouyang S., Iovene M., O'Brien K., Vuong H., Jiang J. and Buell C.R. (2008): "Analysis of 90 Mb of the potato genome reveals conservation of gene structures and order with tomato but divergence in repetitive sequence composition", BMC Genomics, Volume 9, 286.

Abstract

Feeding the growing world population is one of the biggest challenges for the 21st century. Potato, being the fourth crop in the human diet, after maize, wheat and rice, plays an important role in this respect. Like other crops, potato is exposed to a range of potentially yield-reducing factors: Pathogens, a (possibly changing) bad climate and averse soil conditions. Research into the response of potato to these influences, often determined by hereditary factors, is necessary to meet a growing demand for potatoes. A map of genetically determined properties is crucial for this research. Several techniques are available to produce maps – each with it's own merits and demerits, resulting in maps of different qualities and with different resolutions. Two often used mapping techniques are genetic mapping, where the inheritance of multiple traits ("markers") is studied in offspring using statistical analysis and the markers ordered accordingly, and physical mapping on the basis of "Bacterial Artificial Chromosome" (BAC) libraries. BAC libraries consist of a large number of individual bacterial strains (BAC clones), each containing a randomly sampled section of DNA of the organism being studied. By comparing individual BAC clones with each other, finding out where the donor organism's (the organism being studied) DNA sections overlap, the BAC clones can be ordered into groups or "contigs". Comparison is often done on the basis of so called fingerprints – a pattern consisting of DNA fragments of different lengths, resembling a bar-code pattern. A similarity in fingerprint patterns between two BAC clones indicates that the BAC clones contain similar (overlapping) sections of the donor organism's DNA. Recently an ultra dense genetic map has been published, containing more than 10,000 markers produced using "Amplified Fragment Length Polymorphism" (AFLP™) marker technology. The integrated physical and genetic map that is the subject of this thesis extends this genetic map, and is in itself the starting point for determining the detailed DNA sequence of potato, as is currently being undertaken by an international scientific collaboration within the Potato Genome Sequencing Consortium (PGSC, http://www.potatogenome.net).

First step in creating this integrated physical and genetic map was creation, fingerprinting and characterization of a BAC library, as described in chapter two. BACs were individually fingerprinted using an AFLP based protocol, and (amongst others) these AFLP BAC-fingerprints were compared to a theoretical model of the distribution of fragment lengths in AFLP fingerprints to determine if fingerprinting was successful.

Correction and refinement of some of the mapping algorithms that were used to create the genetic map are discussed in chapters three and four, resulting in refined genetic map locations for the AFLP markers and the capability to process marker scores containing arbitrary types of scoring ambiguities while conserving all available information. An extension to the basic principle offers the possibility to also map AFLP markers derived from different chromosomes that are indistinguishable on the basis of their AFLP fragment length alone.

In chapter five, systematic differences in AFLP BAC fingerprints are discussed that are caused by the use of different machines for capillary electrophoresis, by the use of different fluorescent DNA labels and by different capillary position. These systematic differences are (partially) corrected by using the (abundant) AFLP fingerprints of BAC clones containing (part of) the potato chloroplast genome as a reference sample.
By ordering the AFLP fingerprints of individual BAC clones on the basis of fingerprint similarity, a physical map is produced that is integrated with the genetic map using a novel, ultra efficient, procedure described in chapter six. This procedure, "AFLP contig matching" uses intricate experimental design and combinatorial analysis to obtain an integrated physical and genetic map with the least amount of effort.

Samenvatting

Het voeden van de groeiende wereldbevolking is een van de grootste uitdagingen van de 21ste eeuw. Aardappel speelt hierbij, als vierde voedselgewas na mais, tarwe en rijst een belangrijke rol. Net als andere gewassen staat aardappel bloot aan een reeks potentieel opbrengst-reducerende factoren: pathogenen, een (mogelijk veranderend) slecht klimaat en slechte bodemcondities. Onderzoek naar de vaak door erfelijke factoren bepaalde respons van aardappel op deze invloeden is nodig om ook in de toekomst aan de toenemende vraag te kunnen blijven voldoen. Een kaart van de genetische eigenschappen is hierbij onmisbaar. Verschillende technieken bestaan om zulke kaarten te maken, en elk van deze technieken heeft specifieke voor- en nadelen en levert een kaart van verschillende kwaliteit en met verschillende resolutie. op. Twee veelgebruikte technieken zijn genetische kartering waarbij de overerving van verschillende eigenschappen ("merkers") in nakomelingen met statistische technieken wordt bestudeerd en de merkers op grond daarvan geordend worden, en fysische kartering op basis van "Bacterial Artificial Chromosome" (BAC) banken. BAC-banken bestaan uit een groot aantal bacterie-stammen (BAC-klonen) die elk een willekeurig bemonsterd deel van het te bestuderen genoom bevatten. Door individuele BAC-klonen met elkaar te vergelijken en te bepalen of en hoe de uit het donor-organisme afkomstige stukken DNA overlappen, kunnen de BAC-klonen worden geordend in groepen ("contigs"). Het vergelijken van BAC-klonen vindt veelal plaats op basis van zogenaamde "vingerafdrukken" - een patroon bestaande uit DNA fragmenten van verschillende lengte, lijkend op een bar-code. Hierbij is gelijkenis van vingerafdruken van twee BAC klonen een indicatie dat de BAC klonen gelijkende (overlappende) secties van het DNA van het donor-organisme bevatten. Recent is een genetische kaart gepubliceerd die meer dan 10000 merkers bevat die zijn gebaseerd op "Amplified Fragment Length Polymorphism" (AFLP™) merker-technologie. De geintegreerde fysische en genetische kaart die het onderwerp is van dit proefschrift borduurt hierop voort, en is zelf weer de opmaat tot het bepalen van de volledige DNA base-volgorde van het aardappelgenoom zoals nu in internationaal wetenschappelijk samenwerkingsverband plaatsvindt door het Potato Genome Sequencing Consortium (PGSC, http://www.potatogenome.net).

Eerste stap in het maken van de geintegreerde fysische en genetische kaart was het maken en karakteriseren van een Bacterial Artificial Chromosome (BAC)-bank en het maken van op een AFLP protocol gebaseerde vingerafdrukken van individuele BAC klonen, zoals beschreven in hoofdstuk twee. BAC-vingerafdrukken zijn onder andere vergeleken met een theoretisch model dat de distributie van fragment-lengten in een AFLP vingerafdruk beschrijft om te bepalen of het maken van de vingerafdrukken geslaagd was.

Verbeteringen en verfijningen van sommige van de karteringsalgorithmen die gebruikt zijn om de genetische kaart te maken worden besproken in hoofdstukken drie en vier, en leiden tot verfijnde (genetische) karteringen van AFLP-merkers, en tot de mogelijkheid

om merkers die willekeurige (score-) onzekerheden bevatten met behoud van alle beschikbare informatie op de kaart te plaatsen. Een uitbreiding op het basisprincipe biedt de mogelijkheid om ook op basis van enkel de AFLP-fragment-lengte ononderscheidbare merker-paren afkomstig van verschillende chromosomen te karteren.
In hoofdstuk vijf worden systematische verschillen in AFLP BAC-vingerafdrukken die het gevolg zijn van het gebruik van verschillende machines voor capillaire electroforese, het gebruik van verschillende fluorescente DNA-labels en verschillende capillair-posities besproken. Deze systematische verschillen worden gedeeltelijk gecorrigeerd door de AFLP-vingerafdruken van (veelvuldig voorkomende) BAC-klonen die (delen van) het chloroplast-genoom van aardappel bevatten als referentie-monster te gebruiken.
Door de individuele AFLP vingerafdrukken van BAC-klonen op basis van gelijkenissen in de vingerafdrukken te groeperen en te ordenen is een fysische kaart geproduceerd die in hoofdstuk zes met behulp van een nieuw, ultra-efficient, procede gekoppeld wordt aan de genetische kaart. Dit procede, “AFLP contig matching”, maakt gebruik van een uitgekiend experimenteel ontwerp en combinatoriek om met zo min mogelijk middelen tot een geintegreerde fysische en genetische kaart te komen.

Appendix

The synthetic marker scores used to generate Figures 1 and 2 in Chapter 3 are given below. M1_{00} and M2_{00} represent the basic synthetic marker scores from which other scores are derived M2_{01}, M2_{10} and M2_{11} represent three different linkage phase variants of marker 2; the phase of marker 1 is fixed. M1_ra_50 and M2_ra_50 are variants of M1_{00} and M2_{00} with ambiguity introduced in 50 randomly selected individuals, whereas in M1_ra_100 and M2_ra_100 ambiguity is introduced in all 100 offspring. M1_ds_50 and M2_ds_50 are variants of M1_{00} and M2_{00} with marker scores reduced to "dominant" (AC and !AC) scores in 50 randomly selected individuals, whereas in M1_ds_100 and M2_ds_100 all marker scores are reduced to "dominant" scores.

M1_{00}	BD	AC	AD	BD	AD	BD	BD	BC	BD	AC
	AD	AD	AC	AD	AC	BD	BC	BC	AC	AD
	BC	BD	BC	AC	BD	BD	BD	BD	AD	BD
	BD	BC	AC	AD	BD	AD	BD	BD	BC	BC
	AC	BC	AD	AD	AD	BC	BD	BC	BC	AD
	AC	AC	BC	BC	AD	AC	AD	BC	BD	AD
	BD	AD	AD	BD	AD	AD	BC	BD	AC	BC
	AD	AC	BD	AC	BC	BC	BD	AD	AD	BD
	BC	BD	BC	BC	BD	BC	AD	AC	BD	BC
	AD	AC	BC	BD	BD	BC	AC	BD	AC	AD
M2_{00}	AD	AC	AD	BD	AD	BC	BD	BC	BD	AD
	AD	BD	AD	AD	AC	AD	BD	BC	AD	AD
	BC	BD	BD	AC	BD	BD	BD	BD	AD	AD
	AD	BC	AC	AC	BD	AD	BC	BC	BC	BD
	BC	BC	AD	AC	AD	BC	BD	BC	BC	AD
	AC	AD	AC	BD	AD	AD	AD	BC	BD	AD
	BD	AD	AD	BD	AD	AD	BD	AD	AD	BC
	AD	AD	BD	AC	BC	BC	AC	AD	AD	BD
	BC	BD	BC	BC	AD	BC	AC	AC	BD	BC
	AC	AC	BC	BD	BD	BC	AC	BD	AC	AD
M2_{10}	BD	BC	BD	AD	BD	AC	AD	AC	AD	BD
	BD	AD	BD	BD	BC	BD	AD	AC	BD	BD
	AC	AD	AD	BC	AD	AD	AD	AD	BD	BD
	BD	AC	BC	BC	AD	BD	AC	AC	AC	AD
	AC	AC	BD	BC	BD	AC	AD	AC	AC	BD
	BC	BD	BC	AD	BD	BD	BD	AC	AD	BD
	AD	BD	BD	AD	BD	BD	AD	BD	BD	AC
	BD	BD	AD	BC	AC	AC	BC	BD	BD	AD
	AC	AD	AC	AC	BD	AC	BC	BC	AD	AC
	BC	BC	AC	AD	AD	AC	BC	AD	BC	BD
M2_{01}	AC	AD	AC	BC	AC	BD	BC	BD	BC	AC
	AC	BC	AC	AC	AD	AC	BC	BD	AC	AC
	BD	BC	BC	AD	BC	BC	BC	BC	AC	AC
	AC	BD	AD	AD	BC	AC	BD	BD	BD	BC
	BD	BD	AC	AD	AC	BD	BC	BD	BD	AC
	AD	AC	AD	BC	AC	AC	AC	BD	BC	AC
	BC	AC	AC	BC	AC	AC	BC	AC	AC	BD
	AC	AC	BC	AD	BD	BD	AD	AC	AC	BC
	BD	BC	BD	BD	AC	BD	AD	AD	BC	BD
	AD	AD	BD	BC	BC	BD	AD	BC	AD	AC

M2_{11}	BC	BD	BC	AC	BC	AD	AC	AD	AC	BC
	BC	AC	BC	BC	BD	BC	AC	AD	BC	BC
	AD	AC	AC	BD	AC	AC	AC	AC	BC	BC
	BC	AD	BD	BD	AC	BC	AD	AD	AD	AC
	AD	AD	BC	BD	BC	AD	AC	AD	AD	BC
	BD	BC	BD	AC	BC	BC	BC	AD	AC	BC
	AC	BC	BC	AC	BC	BC	AC	BC	BC	AD
	BC	BC	AC	BD	AD	AD	BD	BC	BC	AC
	AD	AC	AD	AD	BC	AD	BD	BD	AC	AD
	BD	BD	AD	AC	AC	AD	BD	AC	BD	BC
M1_ra_50	BD|AC	AC	AD|AC	BD	AD	BD	BD	BC|AC	BD|BC	AC|AD
	AD|AC	AD	AC	AD|BC	AC|BD	BD|BC	BC	BC	AC|BC	AD|AC
	BC|AC	BD|AC	BC	AC|AD	BD	BD|AC	BD	BD|BC	AD	BD|BC
	BD|BC	BC|AC	AC	AD	BD	AD|BD	BD	BD	BC	BC|AD
	AC|BC	BC	AD	AD	AD	BC	BD|AD	BC	BC	AD|BC
	AC	AC|BD	BC|AC	BC	AD|BC	AC|AD	AD|BD	BC|AD	BD	AD|BD
	BD	AD	AD|AC	BD|AD	AD	AD|AC	BC	BD|BC	AC	BC|AD
	AD|AC	AC|AD	BD	AC|BC	BC	BC|BD	BD|AD	AD	AD|AC	BD|AD
	BC	BD	BC	BC|AC	BD	BC|BD	AD	AC	BD	BC|AC
	AD	AC	BC	BD	BD	BC|AC	AC	BD|BC	AC|AD	AD|BD
M2_ra_50	AD	AC|BC	AD	BD|AD	AD|BC	BC	BD	BC	BD	AD
	AD	BD|BC	AD	AD	AC	AD|BD	BD	BC	AD	AD|BC
	BC	BD|AD	BD|AD	AC|AD	BD|BC	BD|BC	BD|AD	BD	AD	AD
	AD	BC	AC	AC	BD|AD	AD	BC|BD	BC|AD	BC|AC	BD
	BC	BC|BD	AD|AC	AC|BC	AD	BC|AC	BD	BC|BD	BC|BD	AD
	AC	AD|BD	AC|AD	BD	AD	AD|AC	AD|BD	BC|AC	BD|AD	AD|AC
	BD|AD	AD|BC	AD|BD	BD|BC	AD	AD	BD	AD|BC	AD|BC	BC|BD
	AD	AD	BD	AC|BC	BC|BD	BC	AC	AD	AD|AC	BD
	BC	BD|AD	BC|AC	BC	AD|BC	BC	AC	AC|BD	BD|BC	BC
	AC|BC	AC	BC	BD|AC	BD	BC|BD	AC|AD	BD|AC	AC	AD|BC
M1_ra_100	BD|BC	AC|AD	AD|BC	BD|AD	AD|AC	BD|BC	BD|BC	BC|BD	BD|BC	AC|BD
	AD|BC	AD|AC	AC|BC	AD|AC	AC|AD	BD|AD	BC|AC	BC|AD	AC|BC	AD|BD
	BC|AD	BD|BC	BC|BD	AC|AD	BD|BC	BD|BC	BD|AC	BD|AC	AD|AC	BD|BC
	BD|AC	BC|AD	AC|BD	AD|AC	BD|BC	AD|BD	BD|BC	BD|BC	BC|AD	BC|AC
	AC|BD	BC|AD	AD|AC	AD|BD	AD|BC	BC|BD	BD|BC	BC|AD	BC|AD	AD|BD
	AC|AD	AC|AD	BC|AC	BC|AD	AD|AC	AC|AD	AD|BD	BC|AD	BD|AC	AD|BD
	BD|BC	AD|BC	AD|BC	BD|AD	AD|BD	AD|BD	BC|AC	BD|AC	AC|AD	BC|BD
	AD|BD	AC|BC	BD|AD	AC|AD	BC|AC	BC|AC	BD|AD	AD|BD	AD|BC	BD|AC
	BC|AC	BD|AC	BC|AC	BC|AD	BD|AC	BC|BD	AD|AC	AC|AD	BD|AC	BC|AD
	AD|AC	AC|BD	BC|BD	BD|BC	BD|AD	BC|AD	AC|BD	BD|AD	AC|BC	AD|BC
M2_ra_100	AD|BC	AC|BD	AD|BC	BD|AC	AD|AC	BC|AC	BD|BC	BC|AD	BD|AC	AD|AC
	AD|AC	BD|BC	AD|BD	AD|BD	AC|AD	AD|BD	BD|AC	BC|AC	AD|AC	AD|AC
	BC|AC	BD|AD	BD|AD	AC|AD	BD|AC	BD|AD	BD|AD	BD|AD	AD|BD	AD|AC
	AD|BD	BC|AD	AC|BD	AC|AD	BD|AC	AD|BD	BC|AD	BC|AC	BC|AD	BD|BC
	BC|AC	BC|AD	AD|BD	AC|BC	AD|BD	BC|AC	BD|AD	BC|AD	BC|BD	AD|BD
	AC|AD	AD|BC	AC|AD	BD|AC	AD|BD	AD|AC	AD|AC	BC|AD	BD|BC	AD|BC
	BD|AD	AD|BD	AD|BD	BD|BC	AD|AC	AD|BD	BD|AC	AD|AC	AD|BD	BC|BD
	AD|AC	AD|BD	BD|AC	AC|BD	BC|BD	BC|AC	AC|BD	AD|AC	AD|BD	BD|BC
	BC|AD	BD|AC	BC|AD	BC|AC	AD|BC	BC|AD	AC|AD	AC|BC	BD|AC	BC|BD
	AC|BD	AC|BC	BC|BD	BD|BC	BD|AD	BC|AC	AC|BD	BD|BC	AC|AD	AD|AC

M1_ds_50	BD	AC	!AC	BD	AD	BD	BD	BC	!AC	AC
	!AC	!AC	AC	AD	AC	BD	!AC	!AC	AC	AD
	BC	BD	BC	AC	!AC	BD	BD	BD	AD	BD
	BD	BC	AC	AD	!AC	!AC	BD	BD	BC	BC
	AC	!AC	AD	AD	AD	!AC	BD	!AC	BC	!AC
	AC	AC	!AC	!AC	AD	AC	AD	!AC	BD	AD
	BD	AD	AD	!AC	!AC	!AC	BC	BD	AC	!AC
	!AC	AC	!AC	AC	!AC	!AC	!AC	AD	!AC	!AC
	BC	!AC	BC	BC	BD	!AC	AD	AC	!AC	!AC
	AD	AC	!AC	BD	BD	BC	AC	BD	AC	!AC
M2_ds_50	AD	AC	AD	BD	AD	!AC	BD	!AC	BD	!AC
	AD	BD	AD	!AC	AC	AD	BD	BC	!AC	!AC
	BC	BD	!AC	AC	BD	!AC	!AC	BD	!AC	AD
	!AC	BC	AC	AC	BD	AD	BC	BC	BC	BD
	BC	!AC	AD	AC	AD	BC	BD	BC	BC	AD
	AC	!AC	AC	BD	!AC	!AC	AD	!AC	BD	AD
	BD	AD	!AC	!AC	!AC	AD	BD	!AC	!AC	BC
	AD	!AC	!AC	AC	BC	BC	AC	!AC	!AC	BD
	BC	!AC	BC	!AC	AD	!AC	AC	AC	BD	!AC
	AC	AC	BC	!AC	!AC	!AC	AC	BD	AC	AD
M1_ds_100	!AC	AC	!AC	!AC	!AC	!AC	!AC	!AC	!AC	AC
	!AC	!AC	AC	!AC	AC	!AC	!AC	!AC	AC	!AC
	!AC	!AC	!AC	AC	!AC	!AC	!AC	!AC	!AC	!AC
	!AC	!AC	AC	!AC	!AC	!AC	!AC	!AC	!AC	!AC
	AC	!AC	!AC	!AC	!AC	!AC	!AC	!AC	!AC	!AC
	AC	AC	!AC	!AC	!AC	AC	!AC	!AC	!AC	!AC
	!AC	!AC	!AC	!AC	!AC	!AC	!AC	!AC	AC	!AC
	!AC	AC	!AC	AC	!AC	!AC	!AC	!AC	!AC	!AC
	!AC	!AC	!AC	!AC	!AC	!AC	!AC	AC	!AC	!AC
	!AC	AC	!AC	!AC	!AC	!AC	AC	!AC	AC	!AC
M2_ds_100	!AC	AC	!AC	!AC	!AC	!AC	!AC	!AC	!AC	!AC
	!AC	!AC	!AC	!AC	AC	!AC	!AC	!AC	!AC	!AC
	!AC	!AC	!AC	AC	!AC	!AC	!AC	!AC	!AC	!AC
	!AC	!AC	AC	AC	!AC	!AC	!AC	!AC	!AC	!AC
	!AC	!AC	!AC	AC	!AC	!AC	!AC	!AC	!AC	!AC
	AC	!AC	AC	!AC	!AC	!AC	!AC	!AC	!AC	!AC
	!AC	!AC	!AC	!AC	!AC	!AC	!AC	!AC	!AC	!AC
	!AC	!AC	!AC	AC	!AC	!AC	AC	!AC	!AC	!AC
	!AC	!AC	!AC	!AC	!AC	!AC	AC	AC	!AC	!AC
	AC	AC	!AC	!AC	!AC	!AC	AC	!AC	AC	!AC

Curriculum vitae

Ik ben op 30 september 1968 geboren te Eemnes. Na het behalen van het Atheneum diploma (1986, Laar en Berg, Laren N.H.) ben ik Electrotechniek gaan studeren in Delft. Na enige jaren bij onder andere het ATO-DLO (ontwikkeling van beeldverwerkingssoftware en elektronica) en ASML (ontwikkeling besturingssoftware voor machines die gebruikt worden voor de productie van geintegreerde schakelingen) gewerkt te hebben, ben ik in februari 2002 begonnen met promotieonderzoek bij de vakgroep plantenveredeling. In februari 2002 ben ik eveneens begonnen aan de MSc-cursus “Biotechnology” (Wageningen Universiteit) en in augustus 2003 ben ik hiervoor met lof geslaagd. Het promotieonderzoek, waar dit proefschrift het resultaat van is, betrof de constructie van een genetish verankerde fysische kaart van aardappel. Dit onderzoek heeft een vervolg gekregen in het Potato Genome Sequence Consortium (PGSC, htp://www.potatogenome.net), en sinds februari 2007 werk ik bij de vakgroep plantenveredeling aan de sequentiering van aardappel.

Education Statement of the Graduate School
Experimental Plant Sciences

Issued to: **Theo Borm**
Group: **Plantenveredeling**

1) Start-up phase	**date**
> First presentation of your project	
BAC pool mapping - Distance estimation between loci based on correlations between markers in pools of BAC-clones	August 25 2003
> MSc courses	
Complete (18 months) MSc course "Biotechnology"	
> Laboratory use of isotopes	
Delft Technical University, level 3	

2) Scientific Exposure	**date**
> EPS PhD student days	
EPS PhD student day, Vrije Universiteit Amsterdam	June 3 2004
EPS PhD student day, Radboud University Nijmegen	June 2 2005
EPS PhD student day, Wageningen University	September 19 2006
> EPS theme symposia	
Theme 4 symposium Wageningen	December 9 2004
Theme 4 symposium Wageningen	December 12 2005
Theme 4 symposium Nijmegen	December 8 2006
Theme 4 symposium Leiden	December 7 2007
> NWO Lunteren days and other National Platforms	
NWO-ALW Lunteren meeting	April 7-8 2003
NWO-ALW Lunteren meeting	April 5-6 2004
NWO-ALW Lunteren meeting	April 4-5 2005
NWO-ALW Lunteren meeting	April 3-4 2006
> Seminars (series), workshops and symposia	
Dr. Valerie Williamson: Mechanism of Mi-mediated resistance to nematodes in tomato	October 30 2003
Dr. Christiane Gebhardt: Potato function map for pathogen resistance and its uses	October 30 2003
Dr. Jens Kossmann: Plant research at Risoe National Laboratory: Prospects to interface biology with physics	March 21 2003
Dr. Rejeev Varshney: Functional molecular markers in barley: development and applications	December 2 2004
Prof. Nicola Pecchioni: QTLs and candidate genes for acclimatisation-related traits in the Nure x Tremois barley cross	December 7 2004
Dr. Levente Kiss: New approaches to understand the evolution of powdery mildew fungi	September 28 2004
Karin van Haren: BioASP - Bioinformatics Application Service Provider	April 15 2004
Dr. Zhang Xueyoung: glimpse of wheat centromere: sequence analysis and cytogenetic characterization of a 89-Kb BAC clone of Triticum boeoticum	October 4 2005
Prof.dr. Sophien Kamoun: Reprogramming the host: The effector secretome of Phytophthora infestans	October 5 2005
Dr. M. Babadoost: Management of Phytophthora Blight (Phytophthora capsici) of Cucurbits and Peppers: Progress and Frustration.	July 18 2005
Prof. N.J. Talbot: Investigating the molecular genetics of plant infection by the rice blast fungus Magnaporthe grisea	May 3 2006
Prof. Andreas Graner: Comparative Genomics of Barley and Rice: Promises Kept and Pending	May 10 2007
> International symposia and congresses	
1st International congress on Plant Metabolomics, Wageningen (NL)	April 7-11 2002
Plant and Animal Genome XI, San Diego (USA)	January 11-15 2003
EU meeting Apophys, Koeln (DE)	February 13-14 2004
EU meeting Apophys, Dundee (GB)	February 4-5 2005
EU meeting Apophys, Avignon (FR)	February 28 - March 1
Plant Gems, Venice (IT))	October 11-14 2006
Benelux Bioinformatics Conference, Wageningen (NL)	October 17-18 2006
EU meeting Apophys, Vitoria (ES)	January 25-26 2007
First International Potato Genome Sequencing Consortium (PGSC) workshop, Wageningen (NL)	April 18-20 2007
Second International Potato Genome Sequencing Consortium (PGSC) workshop, Wageningen (NL)	April 21-23 2008
> Presentations	
Plant and Animal Genome XI, San Diego (USA)	January 11-15 2003
STW meeting, Wageningen (NL)	June 18 2003
STW meeting, Wageningen (NL)	November 19 2003
EU meeting Apophys, Koeln (DE)	February 13-14 2004
STW meeting, Wageningen (NL)	June 15 2004
Theme 4 symposium Wageningen	December 9 2004
STW meeting, Wageningen (NL)	December 16 2004
EU meeting Apophys, Dundee (GB)	February 4-5 2005
STW meeting, Wageningen (NL)	June 23 2005
STW meeting, Wageningen (NL)	November 29 2005
Theme 4 symposium Wageningen	December 12 2005
EU meeting Apophys, Avignon (FR)	Feb 28 - Mar 1 2006
STW meeting, Wageningen (NL)	June 2 2006
Plant Gems, Venice (IT))	October 11-14 2006
Benelux Bioinformatics Conference, Wageningen (NL) (same poster as Plant Gems)	October 17-18 2006
STW meeting, Wageningen (NL)	November 30 2006
EU meeting Apophys, Vitoria (ES)	January 25-26 2007
First International Potato Genome Sequencing Consortium (PGSC) workshop, Wageningen (NL)	April 18-20 2007
Theme 4 symposium Leiden	December 7 2007
Second International Potato Genome Sequencing Consortium (PGSC) workshop, Wageningen (NL)	April 21-23 2008
> IAB interview	September 18 2006

3) In-Depth Studies	**date**
> EPS courses or other PhD courses	
The analysis of natural variation within crop and model plants	April 22-25 2003

4) Personal development	**date**
> Skill training courses	
Afstudeervakken begeleiden (OWU)	February 5-6 2003
> Organisation of PhD students day, course or conference	
First International Potato Genome Sequencing Consortium (PGSC) workshop, Wageningen (NL)	April 18-20 2007
Second International Potato Genome Sequencing Consortium (PGSC) workshop, Wageningen (NL)	April 21-23 2008
Workshop Physical map/sequence annotation for plant breeders	June 12 2008

The research described in this thesis was performed at the Laboratory of Plant Breeding of Wageningen University and Research Center and was financially supported by the Dutch Technology Foundation STW, the applied science division of NWO and the Technology Program of the Ministry of Economic Affairs (Project no. WPB.5283)