POPSEQ

Anchoring and ordering contig assemblies from next generation sequencing data by population sequencing

Dissertationsschrift zur Erlangung des Grades eines Doktors der Naturwissenschaften an der Technischen Fakultät der Universität Bielefeld

von

Martin Mascher

28. Oktober 2013

Gedruckt auf alterungsbeständigem Papier (ISO 9706).

Contents

List of Figures

List of Tables

List of abbreviations

A. thaliana	*Arabidopsis thaliana*
BAC	Bacterial artificial chromosome
BAM	Compressed binary SAM format (see below)
B. distachyon	*Brachypodium distachyon*
BLAST	Basic local alignment search tool
bp	Basepair
BWA	Burrows-Wheeler aligner
cDNA	Complementary DNA of messenger RNA
cM	CentiMorgan
cv.	Cultivar
DH	Doubled haploid
DNA	Deoxyribonucleic acid
E. coli	*Escherichia coli*
EST	Expressed sequence tag
F_1, F_2, ...	First, second, ... filial generation
FP contig	Fingerprint contig
GATK	Genome Analysis Toolkit
Gb	Giga base pair
GBS	Genotyping-by-sequencing
HSP	High scoring pair
H. vulgare	*Hordeum vulgare*
IBSC	International Barley Genome Sequencing Consortium
Indel	Short insertion and deletion polymorphism
kb	Kilo base pair
LD	Linkage disequilibrium
MAD	Median absolute deviation
MTP	Minimum tiling path
MxB	Morex × Barke
Mb	Mega base pair
N50	Weighed average contig size, half of the assembly is contained in contigs larger than the N50
NGS	Next generation sequencing
nt	Nucleotide
OWB	Oregon Wolfe Barleys
PCR	Polymerase chain reaction
RIL	Recombinant inbred line

RNA	Ribonucleic acid
RNA-seq	RNA sequencing using NGS technology
SAM	Sequence Alignment/Map format
VCF	Variant call format
SNP	Single nucleotide polymorphism
Vrs1	Six-rowed spike gene
WGS	Whole genome shotgun

The continuing rapid fall in the cost of computer components is making it possible for most DNA sequencing laboratories to have their own small computer. The fact that DNA sequencing is now a fast procedure, and the availability of computers gives the possibility of more efficient overall strategies for sequence determination.

– Rodger Staden, 1979

1 Introduction

Next generation sequencing (NGS) provides the opportunity to rapidly and at relatively low cost establish gene space assemblies for virtually any species. These assemblies consist of tens to hundreds of thousands of short contiguous pieces of DNA sequence (contigs) and often represent only the low-copy portion of the genome. Despite the limitations of such assemblies, they have been widely proposed as surrogates for draft genome sequences for the purposes of gene isolation, genomics-assisted breeding and the assessment of diversity within and between species (Brenchley et al., 2012; IBSC, 2012; Xu et al., 2012; Guo et al., 2012). However in most cases, particularly those concerning large and complex genomes, they remain disconnected collections of short sequence contigs that are not embedded in a genomic context. Bringing these fragments together into a tentative linear order, or even associating contigs with individual chromosomes or chromosome arms, has been a major and costly undertaking. In a recent example, the International Barley Genome Sequencing Consortium (IBSC, 2012) had reported a gene space assembly of the 5.1 Gb genome of barley. The development and use of a BAC-based physical map, BAC end sequences, flow-sorted and chromosome-arm survey sequences, fully sequenced BAC clones and conserved synteny were all required to fully contextualize only 410 Mb of genomic sequence IBSC (2012). These genomic resources provide an established path towards a reference sequence by sequencing a minimum tiling path of overlapping BAC clones and hierarchically (Feuillet et al., 2012). The development of the necessary resources requires a substantial amount of time, labor and finances which makes this strategy prohibitive for smaller and more poorly resourced research communities, e.g. research in non-model organisms or orphan crops. The establishment of a BAC-based reference sequence of the maize genome took about seven years, required the coordinated effort of several laboratories and cost about US $50 million (Chandler and Brendel, 2002; Martienssen et al., 2004; Schnable et al., 2009). Similarly, the reference sequence of a single 1 Gb chromosome of hexaploid wheat has not been finished five years after the publication of a physical map (Paux et al., 2008).

Emerging technologies such as longer sequence reads (Schadt et al., 2010), optical mapping (Lam et al., 2012a) and novel assembly algorithms (such as ALLPATHS-LG (Gnerre et al., 2011)) may speed up the process of data collection and analysis as well as increase the contiguity and completeness of WGS assemblies, but their applicability to large genomes where abundant sequence repeats (the bane of any assembler), arising from paralogous duplications, repetitive elements, ancestral duplications and polyploidy, remains to be assessed.

It has been common practice to associate mapped genetic markers with sequence resources based on sequence similarity in order to link genetic and physical maps (Chen et al., 2002; Wei et al., 2007). While the order of BAC contigs on a physical is in the order of thousands, NGS technology produces hundres dof thousands of sequence contigs. For example, (IBSC, 2012) reported an assembly that consists of over 350,000 contigs longer than 1 kb. The number of markers afforded by conventional genotyping strategies is simply not commensurate with the large number of short sequence contigs.

In the absence of an appropriate molecular or analytical method to establish short-range connectivity (i. e. to link physically close sequence contigs), we used the power of genetic segregation to directly and linearly arrange sequence contigs into closely associated recombination bins along a target genome. We show that whole genome survey sequencing of a small experimental segregating population and genetic mapping of the millions of observed single nucleotide polymorphisms (SNPs) detected therein can vastly improve the quality and utility of highly fragmented NGS shotgun assemblies. We illustrate the approach using the complex 5.1 Gb genome of cultivated barley (*Hordeum vulgare*) by comparing the output to a gene space assembly that has been partially ordered using extensive physical and genetic mapping resources (IBSC, 2012). Our results are congruent with the current sequence assembly but increase the amount of genetically anchored contig sequences by a factor of three. Most importantly, the whole effort cost US < $100,000 and was completed in a matter of months. This new assembly has greater value for comparative genetic studies, gene isolation and genomics-assisted breeding. In principle the approach, which we term POPSEQ, may be used for any species for which a segregating population can be derived and maintained.

1.1 Overview

Chapter 2 will give an introduction to the two key concepts that underlie POPSEQ: genetic mapping and genome sequencing. After a short description of the fundamentals of genetic linkage analysis, we will focus on the technical aspects of map construction in plant mapping populations. We will introduce next generation technologies as the main drivers of advances in genomic research during the past decade and describe the main strategies for the assembly of entire genomes. This chapter closes with a review of previously established methods for anchoring sequence assemblies or physical maps.

Chapter 3 will describe the POPSEQ method as we developed it for the genetic anchoring of the whole-genome shotgun assembly of barley cv. 'Morex'. We will describe the utilized software tools, the incorporated data sets and the key algorithm of POPSEQ: the integration of SNP markers detected by whole genome sequencing of a segregating population into a framework genetic map of this same population.

In Chapter 4, we will evaluate the outcome of POPSEQ against the published physical and genetic framework of barley as well as check the consistency of POPSEQ when different framework maps are utilized.

Chapter 5 will illustrate potential applications of an ordered gene-space assembly as provided by POPSEQ for the purposes of sequencing-based gene isolation and comparative genetic studies. We will also describe how a POPSEQ assembly may assist the genetic anchoring of physical maps and enables reference-based genetic mapping.

In Chapter 6, we will discuss how the raw data quality affects the outcome of POPSEQ, the general limitations of genetic anchoring, how POPSEQ can be adapted to other plant and animal species and which further improvements to the method are possible.

1.2 Publications

Parts of this thesis have been published or are intended to be published in the following articles:

> Mascher, M, Muehlbauer, GJ, Rokhsar, DS, Chapman, JA, Barry, K, Muñoz-Amatriaín, M, Close, TJ, Wise, RP, Schulman, AH, Himmelbach, A, Mayer, KFX, Scholz, U, Poland JA, Stein, N, Waugh, R (2013) Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ), Plant J., *in press.*

> Mascher, M, Richmond, TA, Gerhardt, DJ, Himmelbach, A, Clissold, L, Sampath, D, Ayling, S, Steuernagel, B, Pfeifer, M, D'Ascenzo, M, Akhunov, ED, Hedley, PE, Gonzales, AM, Morrell, PL, Kilian, B, Blattner, FR, Scholz, U, Mayer, KF, Flavell, AJ, Muehlbauer, GJ, Waugh, R, Jeddeloh, JA, Stein, N (2013). Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. Plant J., **76**:494–505.

> Mascher, M, Wu, S, St. Amand, P, Stein, N, Poland, A (2013). Application of genotyping-by-sequencing on semiconductor sequencing platforms: A comparison of genetic and reference-based marker ordering in barley, PLoS ONE, **8**:e76925.

> Ariyadasa, R, Mascher, M, Nussbaumer, T, Schulte, D, Frenkel, Z, Pousarebani, N, Zhou, R, Steuernagel, B, Gundlach, H, Taudien, S, Felder, M, Platzer, M, Himmelbach, A, Schmutzer, T, Hedley, PE, Muehlbauer, GJ, Scholz, U, Korol, A, Mayer, KFX, Waugh, R, Landgridge, P, Graner, A, Stein, N (2013) A sequence-ready physical map of barley anchored genetically by two million SNPs, *in revision.*

1.3 Acknowledgements

First of all, I thank my supervisor Jens Stoye for his generous interest in my work, which made the completion of my dissertation possible. I am grateful to Bernd Weisshaar for his general support of my PhD project and being a referee of my thesis.

POPSEQ is embedded in the vital and active barley genome sequencing community that thrives by openly sharing ideas and data. Among our numerous collaborators, I want to point out especially Robbie Waugh and Jesse Poland for pushing our projects forward.

My PhD work at IPK Gatersleben has been made a very pleasant and highly productive time by the constant support and fruitful collaboration of many colleagues. Uwe Scholz, who manages the large and diverse bioinformatics research group, succeeds splendidly in fostering a most stimulating research environment. His invaluable counsel in strategic decisions and more down-to-earth matters was crucial for the timely completion of my thesis. Nils Stein introduced me to the barley community, gave me the opportunity to present my work in national and international conferences and was always eager to discuss results and to suggest new directions. I enjoyed many stimulating discussions with Swetlana Friedel who supervised my first small research project at IPK.

A large part of the experimental work to collect the sequence data for this thesis and our other projects has been mostly conducted by members of the research group Genome Diversity. For keeping up a steady flow of high-quality sequence data for me to analyse, I thank Axel Himmelbach, Ruvini Ariyadasa, Matthias Jost and Neele Wendler.

The systems administrators Steffen Flemming, Thomas Münch and Heiko Miehe have been keeping our compute servers up and running.

Among those that should not remain unmentioned are those ladies and gentlemen that I had the pleasure to share the same office with at one – long or short – time during the last three years: Daniel Arend, Sebastian Beier, Christian Colmsee, Jinbo Chen, Thomas Schmutzer, Danuta Schüler, Doreen Stengel, Burkhard Steuernagel and Stephan Weise.

Last but not least, I express my heartfelt gratitude to my family for their unrelenting trust and support.

2 Background

POPSEQ is an amalgamation of genome sequencing and genetic mapping, two disciplines that have for a long time been recognized as “intrinsically commensurate” (Chen et al., 1993). The purpose of this chapter is to furnish the reader with an introduction into both fields and to familiarize them with recent developments of, and state-of-the-art methods for, genetic map construction, genome assembly and the genetic anchoring of sequence contigs or physical maps.

In the first section of this chapter, we will briefly review the key concepts of genetic linkage analysis. Subsequently, we give a short introduction to next generation sequencing technologies and strategies for genome assembly. We present a survey on the paths other genome sequencing projects have taken to order their assemblies and describe the integrated physical and genetic map of barley in more detail as this resource set the bar for how well a whole genome shotgun assembly can be anchored with the currently available technologies and will used be in Chapter 4 to as a benchmark for POPSEQ.

2.1 Genetic mapping

Gregor Mendel did not have any mechanistic insights as to how genetic information is maintained in organisms and transmitted to the next generation. Observing recurring patterns of inheritance of a small number of traits in crosses of pea led him to the formulation of his two laws of segregation and independent assortment (Mendel, 1866). Diploid organisms possess two copies of each of theirs genes, one on each chromosome of a pair. Segregation means that only one randomly selected copy of these genes is passed to the offspring. Independent assortment refers to the fact that genes for different traits are passed on to the next generation independently.

At the time of their initial discovery as “stained bodies” inside cells, chromosomes were not put into relation with inheritance. In the 1880s, chromosomes were first proposed to be carriers of genetic information though this idea was initially much contested. The rediscovery of Mendel’s laws strongly supported the chromosomal theory of inheritance and in 1910, Thomas Hunt Morgan could prove that physical co-localization of two genes on the same chromosome correlates with co-inheritance of these genes. Shortly afterwards, Morgan’s student Sturtevant published the first genetic map which ordered six loci along the sex chromosome of fruit fly (Sturtevant, 1913). The monograph of Morgan et al. (1922) laid down the chromosomal theory of inheritance as the governing principle of classical genetics.

Chromosomes are complexes of DNA and proteins. Initially, it was not clear whether genetic information is encoded in the DNA or in the proteins. The definite proof that DNA is the conveyor of genetic information was given by Avery et al. (1944). Less than ten years after this discovery, Watson and Crick (1953) unraveled the molecular structure of DNA. Another nine years passed and Matthaei et al. (1962) published the first connection between a DNA triplet and an amino acid. The "genetic code" – the table that assigns each DNA triplet to exactly one amino acid – is the reason why it is desirable to obtain the sequence of nucleotides in a DNA molecule. Variation in this sequence translates into variation of proteins and finally into phenotypic variation between organisms.

In the following sections, we will describe how meiotic recombination is related to genetic linkage and elaborate on the practical aspects of linkage analysis. We will describe how linkage is experimentally measured by high-throughput genotyping technologies and how genetic mapping in plants is performed nowadays. We will focus on the two types of mapping populations, recombinant inbred lines and doubled haploid lines, which we applied for POPSEQ in barley. Finally, we outline the computational procedures involved in genetic map construction from genotypic data.

2.1.1 Meiotic recombination and genetic linkage

The concept of genetic linkage was first described by Bateson et al. (1908) in sweet pea (*Lathyrus odoratus*). They observed that the two traits flower color and shape of pollen grains deviated considerably from what would be expected by Mendelian inheritance. They named this phenomenon "coupling", but did not relate their findings to chromosomes.

Thomas Hunt Morgan obtained similar results when performing crossing experiments in the fruit fly *Drosophila* and developed the hypothesis that linked genes are physically located on the same chromosome. He also proposed the concept of meiotic recombination, i. e. the exchange of genetic material between homologous chromosomes during meiosis (Morgan, 1910). At about the same time, Janssen discovered the cytological phenomenon of meiotic chiasmata (Janssens, 1909) which provided a mechanistic explanation for Morgans's hypothesis. Meiosis is a type of cell division that occurs during the formation of the reproductive cells (gametes) of an organism. Janssens (1909) observed by light microscopy that homologous chromosomes overlay each other to form X-shaped structures called chiasmata or crossovers. These crossovers bring about a new configuration of genetic material from both chromatids, resulting in a reshuffling of the allelic content between homologous chromosomes (Figure 2.1).

Genetic linkage describes the phenomenon that alleles of two different genes

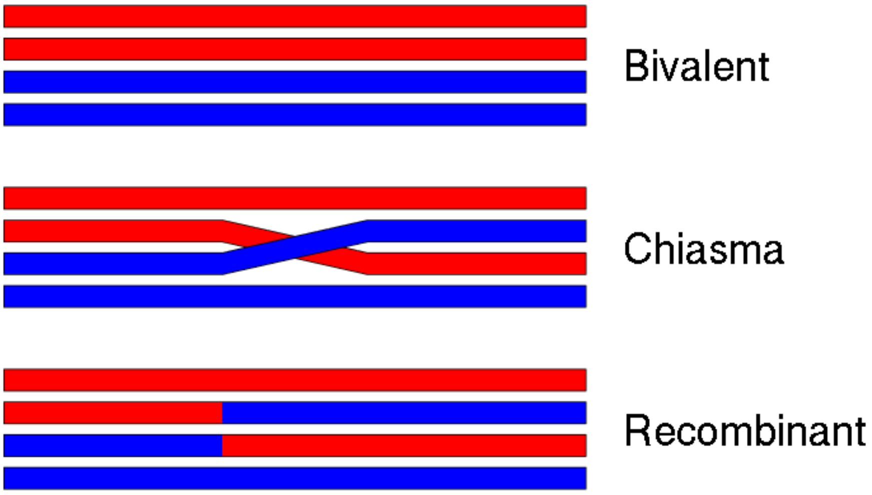


Figure 2.1: Graphical representation of meiotic recombination in a diploid species. Prior to meiosis, chromosomes are duplicated, creating two sister chromatids for both the maternal and the paternal homolog. In the prophase of meiosis, the two pairs of sister chromatids align to form a bivalent. Non-sister chromatids may intertwine leading to breakage and rearrangement of reciprocal strands. This physical strand exchange generates recombinants. For the proper segregation of bivalents, at least one chiasma per chromosome seems to be necessary. This figure was adapted from Figure 1.1 in Schneider (2005).

(or more generally chromosomal loci) that are located close to each other on the same chromosome are inherited together. Mendel originally assumed that two different genes are inherited independently of each other and formulated this idea in the law of independent assortment (Figure 2.2). However, this law is violated when two genes are found on the same chromosome. The independent assortment of genes to gametes is effected by the random distribution of genetic material from paternal and maternal chromosomes by meiotic recombination. If genes are located on the same chromosome, they do not assort independently but have a tendency to be co-inherited.

In a single meiosis, there is only a limited number of chiasmata or crossovers. The location of crossovers is distributed randomly (though not uniformly) along the length of the chromosomes. Thus, the probability of a crossover to occur between a given pair of loci is a function of the physical distance of these loci on the chromosome, i. e. two loci on the same chromosome are more likely to separated by a crossover the farther away they are from each other. In the most extreme case of complete linkage, only parental gametes would be created and the inheritance of two traits would appear as if controlled by a single gene (Figure 2.2).

The number of recombination events between two loci in a certain number of meioses can be determined by tracking the inheritance of genetic markers in a mapping population. The recombination frequencies between loci can then be employed to calculate the genetic distance between them. By computational means, the distances between a large number of markers can be used to construct a genetic linkage map, that is to assign markers to linkage groups corresponding to chromosomes and order them along the chromosomes.

2.1.2 Genetic markers

Genetic markers are manifestations of inheritable variation between the individuals of a species. Genetic markers pertain to a specific region of the genome region, a locus. As observable alterations in these loci, markers can be used to trace the inheritance of DNA sequence in genetic linkage analysis.

Early genetic maps were constructed with morphological markers, such as the shape, size or color of specific organs, and comprised several dozens of loci. Such maps were made for several species, including *Drosophila* (Bridges and Brehme, 1944), maize (Emerson et al., 1935) and barley (Immer and Henderson, 1943). The paucity of scorable phenotypic variation severely limited the resolution of genetic maps. In a seminal paper, Botstein et al. (1980) proposed the use of DNA sequence variation as markers for genetic linkage analysis. DNA sequence variation is present in genome sequences in the form of single nucleotide polymorphism (SNPs), short insertion insertions and deletions (indels) and larger scale structural variation such as presence-absence (PAV) or copy number variation (CNV). DNA sequence variation is abundant in genomes. For example, there is one SNP about every 100 nucleotides between elite inbred lines of maize (Ching et al., 2002). Thus, sequence variation can supply a large number of polymorphic markers.

Before next generation sequencing made it possible to sequence a large number of individuals, the main challenge of genotyping was to discover sequence variation and convert it into assayable DNA markers. With technological advances in molecular biology and in DNA sequencing in particular, marker technologies were improved with the aim to score as many markers as cost efficiently as possible. For an overview of the various techniques for marker discovery and assaying and their historic development, we refer the reader to the reviews of O'Hanlon et al. (2000) and the book of Henry (2012). Here, we will focus on two current state-of-the-art high-throughput marker platforms, genotyping microarrays and resequencing by NGS technology.

One of the first high-throughput genotyping plaforms that allowed the parallel interrogation of a large number of polymorphisms were oligonucleotide SNP arrays (Wang et al., 1998). Fluorescently labeled oligonucleotide probes are synthesized for both alleles of a bi-allelic SNP together with non-polymorphic flanking sequence and arrayed on a microarray. Sample DNA is hybridized to

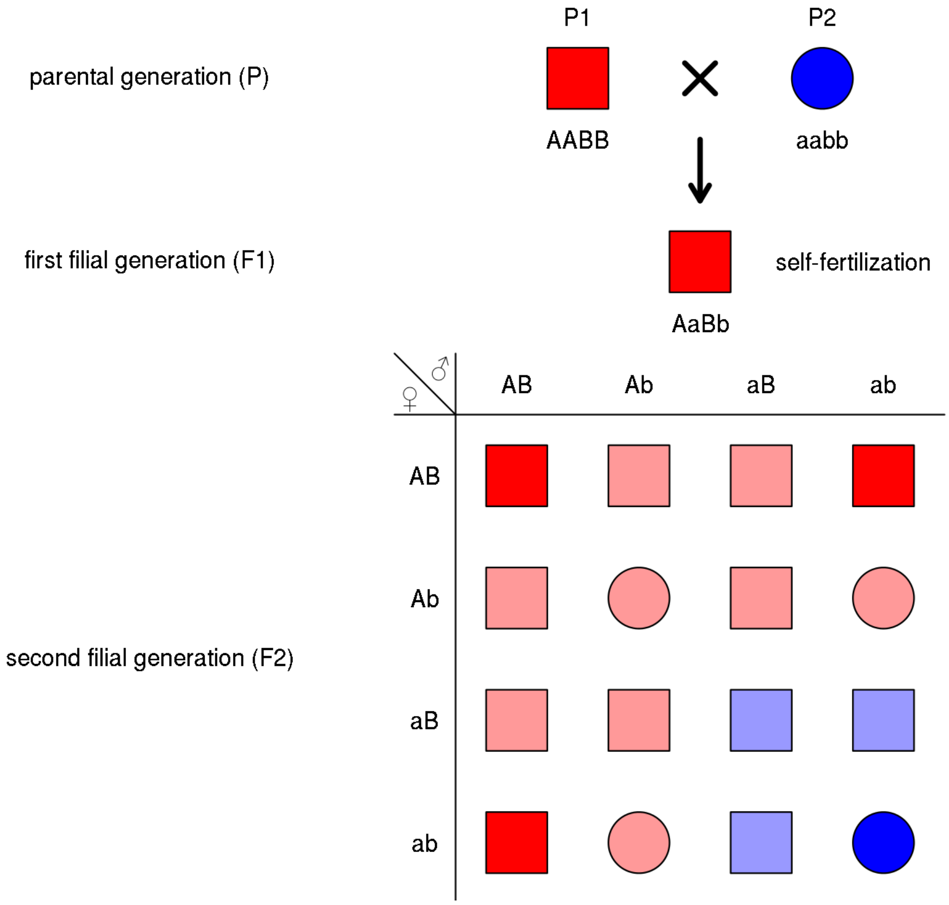


Figure 2.2: Segregation in a dihybrid cross. Two individuals differing in two traits (shape and color) are crossed to each other. The “red” and “square” alleles are dominant, so that the F1 generation has the P1 phenotype. If the shape and color loci are located on different chromosomes, a phenotypic ratio of 9:3:3:1 is observed as predicted by Mendel’s law of independent assortment. In case of complete linkage between shape and color, i. e. if both genes were located in close physical proximity, no recombination between parental genotype would take place and both traits would segregate in a 3:1 ratio as if they were controlled by a single gene. Light colored genotypes would then be absent from the F_2 generation.

this chip and the hybridization signals are recorded by high-resolution imaging and evaluated by computer software. Currently, SNP chips with several million probes are commercially available (The International HapMap Consortium, 2007).

As with previous low-throughput methods, the processes of marker discovery and genotyping are separate in array-based genotyping. One commonly used approach of comprehensive marker discovery is to sequence the transcriptomes or genomes of a limited set of individuals, whose members are selected for high diversity based a small number of markers or by morphological criteria in order to maximize the SNP harvest. The SNPs detected in the reference panel will be converted into an oligonucleotide probe assay to be included on a SNP array, which is then used to genotype a larger number of individuals. This marker discovery process can engender ascertainment bias (Lachance and Tishkoff, 2013). Sequence variation that is absent in members of the discovery panel cannot be assessed, downwardly biasing estimates of genetic diversity when individuals from a diverged population are genotyped.

Genotyping by resequencing with NGS technology allows the detection and typing of markers, most commonly SNPs, to occur simultaneously. These two processes in the context of next generation sequencing are referred to as variant and genotype calling. Variant calling finds sites that are polymorphic in the sequenced individuals by inspecting the location of mapped reads and aggregating nucleotides at each position of a reference sequence. Afterwards, genotype calling determines the genotype (e.g. homozygous for one allele of a bi-allelic SNP or heterozygous in diploids) at each variant site (Nielsen et al., 2011). Alignment of sequence reads and variant/genotype calling is possible even when no reference sequence for read mapping is available. Overlap between reads can be detected, for example, through hash- (Elshire et al., 2011) or graph-based methods (Iqbal et al., 2012).

One of the first application of resequencing for the genome-wide discovery of genetic markers was described by Altshuler et al. (2000). This article introduced the combined use of restriction-enzyme mediated complexity reduction and shotgun sequencing. Because of the technical limitations of capillary sequencing, DNA from several individuals was pooled prior to sequencing and it was not possible to simultaneously genotype these individuals at the detected variant sites.

Barcoded sequencing with NGS technology (see section 2.2.2) has enabled the multiplexed sequencing of DNA from a large number of individuals. Initially, barcoded sequencing was combined with targeted enrichment of small genomic intervals to obtain high coverage across all samples (Craig et al., 2008). Huang et al. (2009) used shallow-coverage whole genome resequencing to genotype a plant mapping population. They sequenced barcoded samples from sixteen recombinant inbred lines (RILs) from a cross between Indica and Japonica rice on the Illumina GenomeAnalyzer to ∼0.02x coverage. They discovered over one million SNPs and adopted a sliding window approach to call genotypes because missing data predominated at each single SNP due to the low read depth. This method was effective because the rice reference sequence could provide an ordered physical framework to place markers and the number of recombination events per RIL was limited. Low-coverage whole genome sequencing reaches its limits when unrelated individuals are genotyped and no inference from parental haplotypes can be made. High coverage whole sequence of a large number of individuals, however, is still so costly as to require the joint effort of international consortia (The 1000 Genomes Project Consortium, 2010; Weigel and Mott, 2009).

Several methods for complexity-reduced resequencing (i. e. sequencing a smaller portion of the genome) have been developed. In hybridization-based targeted enrichment, oligonucleotide probes (also called baits) are designed for genomic regions of interest. Genomic DNA fragments are hybridized to these probes and only bound fragments are sequenced. Most conspicuously, hybridization-based enrichment is implemented in whole exome captures assays, where a comprehensive set of mRNA-coding exons is used as the target space. Whole exome capture assays have been designed for a number of species with large and complex genomes, such as human (Asan et al., 2011), mice (Fairfield et al., 2011), maize (Liu et al., 2012a) and barley (Mascher et al., 2013b). Exome capture enables deep coverage resequencing of the protein-coding portion of the genome. Compared to whole genome sequencing, a much higher number of individuals can be sequenced with the same financial effort. Exome capture has been used for mutation discovery (Bamshad et al., 2011) and studies of genetic diversity (Li et al., 2010c). As an alternative to whole exome capture, smaller hybridization assays targeting selected candidates genes can be designed (Cosart et al., 2011; Salmon et al., 2012). In any case, hybridization-based capture requires a substantial effort for the design of the enrichment assays. Depending on the target size, thousands to millions of oligonucleotides need to be designed. Moreover, the probe design requires a reference sequence.

An alternative to hybridization-based approaches are methods employing restriction-enzyme digestion for complexity reduction. Genomic DNA is digested with one or more restriction endonucleases and only short DNA fragments adjacent to cut sites are sequenced. The various approaches differ in the choice of enzymes, which affects the size of the targeted regions, as well as in the details of the wet-lab protocols for processing digested DNA into libraries suitable for high-throughput sequencing (Davey et al., 2011). Genotyping-by-sequencing (GBS), though somewhat inaccurate, has become a generic term to denote some of these techniques. GBS has been implemented in many crop species (Elshire et al., 2011; Poland et al., 2012b; Truong et al., 2012) and has been recognized as a cost-efficient tool for genome-assisted agronomic research even in species where there is no reference sequence available (Poland and Rife, 2012; Poland et al., 2012a).

2.1.3 Plant mapping populations

In plants, genetic maps are usually constructed by genotyping experimental mapping populations, which are the progeny of crosses between a small number of (usually two) parents. Population development in many plant species takes advantage of self-fertilization. The majority of flowering plant species are hermaphroditic, i. e. male and female reproductive organs are present in the same individual, often even in the same flower. If the predominant mode of reproduction of a species involves the fertilization of egg cells with pollen from the same individual, this species is called inbreeding. Inbreeding is in contrast to outcrossing where zygotes are formed by the fusion of gametes from different individuals. In species with separate sexes, only outcrossing can occur. Outcrossing refers only to the most common mode of reproduction and does not preclude the possibility of (artificial) self-fertilization. For example, the wild progenitors of maize and sugar beet are outcrossing species. However, there are no barriers to creating inbred lines of maize and sugar beet.

When individuals have only one sex, as is the case in mammals or dioecious plants such as wild grapevine or poplar, inbreeding is not possible. Self-fertilization may also by prevented through self-incompatibility mechanisms in hermaphrodites. For example, individuals of the sea squirt *Ciona intestinalis* produce both eggs and sperm, but cannot self-fertilize. In self-incompatible plant species – for example, the close relative of barley *Hordeum bulbosum* or the cereal crop rye (*Secale cereale*) – no seeds develop if flowers have been pollinated by pollen of their own or from other flowers of the same plant. The self-incompatibility response has been shown to be effected by special genes involved in cell-cell interaction and self-recognition based on molecular signals (Haring et al., 1990)

In barley and other self-fertile species such as wheat, rice or maize, mapping populations can be created by crossing two inbred parents, i. e. almost completely homozygous plants whose ancestors have been selfed for several generations. The first filial F_1 generation will be heterozygous at each site that is polymorphic between the parents. The chromosomes of the F_2 generation obtained by selfing F_1 plants are mosaics of the parental haplotypes as a consequence of meiotic recombination (Figure 2.3).

Markers that are polymorphic between the parents can be scored in the F_2 plants and the resulting genotype matrix be used for genetic map construction. However, as chromosomes are heterozygous for half of their genetic length in F_2 plants, several marker techniques cannot correctly score or work unreliably for heterozygous loci. Most markers assaying presence-absence variation by probes located in present/missing sequence cannot distinguish between one or two 'present' alleles. Given sufficient read depth at polymorphic sites, SNPs when genotyped with NGS technology are co-dominant markers, that is heterozygotes and homozygotes can be distinguished. If, however, coverage is low, heterozygotes cannot be scored reliably because only one allele is ran-

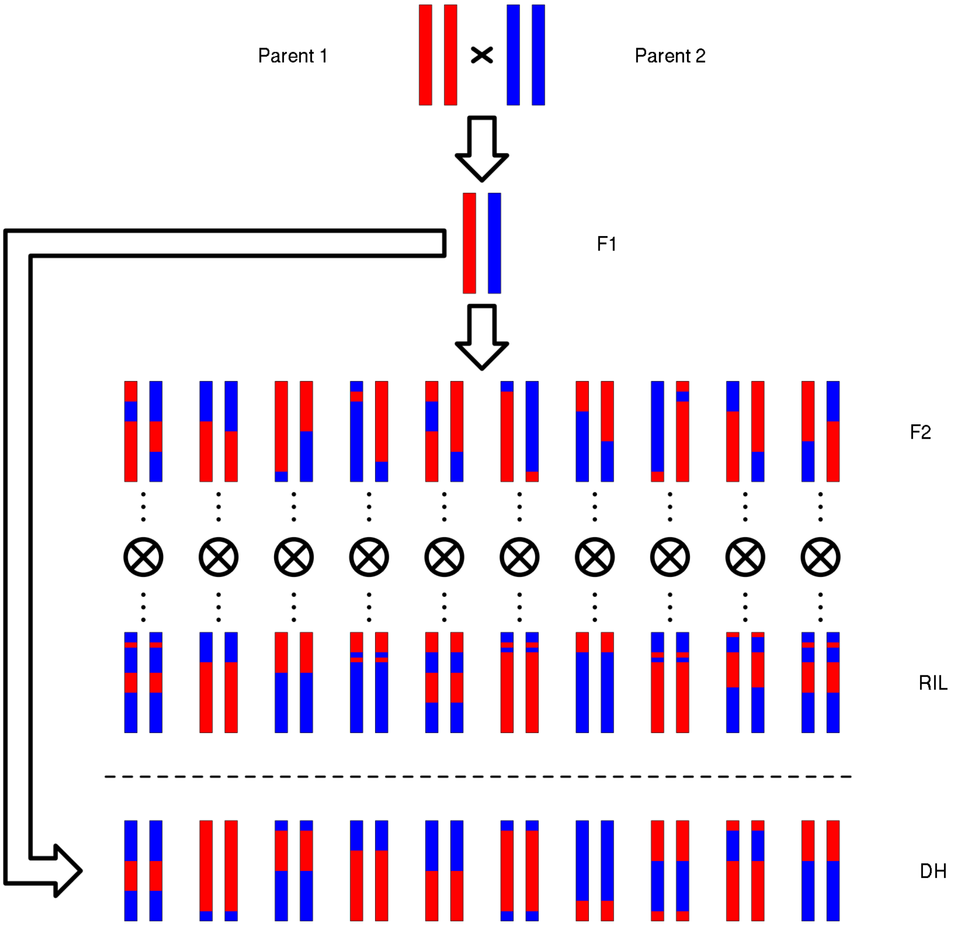


Figure 2.3: Different types of mapping populations in selfing organisms. F_2 populations are created by selfing the offspring of a cross between two parental inbred lines. Through repeated selfing of the F_2 individuals, recombinant inbred lines (RILs) are created. The process of generating completely homozygous lines can be accelerated through the use of doubled haploid (DH) lines obtained from gametes of the F_1 generation.

domly sampled. Another disadvantage of F_2 populations is the small number of recombination events per chromosome. For example, the genetic length of barley chromosomes is between 100 and 200 cM, that means there are one to two crossovers per meiosis. A higher number of recombination events could provide better genetic resolution.

These disadvantages are avoided when recombinant inbred lines (RILs) are genotyped. RILs are developed by single-seed descent from F_2 individuals (Figure 2.3). A plant of the F_2 generation is selfed and a single seed is harvested and grown to a new F_3 plant, which again is selfed and a single seed is selected. This process is usually repeated for four to six rounds. The size of the heterozygous portion of genome is halved in each generation, so that the expected residual heterozygosity of a RIL F_8 after seven rounds of selfing would be $2^{-7} \approx 0.8\%$. As more meioses occur during population development, more recombination breakpoints can be detected, which increases mapping resolution.

RILs have been a powerful tool for genetic mapping in many species such a mouse (Swank and Bailey, 1973), maize (Burr et al., 1988), tomato (Paran et al., 1995) or barley (Comadran et al., 2012). RILs are particular valuable for quantitative genetics. After sufficient rounds selfing, the offspring obtained by selfing a RIL is genetically identical to its parent. Thus, a single genotype can be grown several times in different years and in different environments. This repeatability greatly increases the power of statistical analyses aimed at assessing the interaction of genotype and environment in shaping quantitative traits such as plant height or yield.

The major drawback of RILs is the long time it takes to generate them. The life cycle of a plant from seed to seed has to be completed multiple times. Moreover, in winter annuals, young plants have to vernalized, that is they have be grown in low temperatures for a prolonged time period to simulate winter. Otherwise, they would not flower. A shortcut to produce entirely homozygous plants that obviates the need for several rounds of selfing is the artificial production of doubled haploid plants (Figure 2.3). Haploid gametes of the F_1 generation are induced to develop into diploid plants by a special treatment. In barley, for instance, haploid pollen progenitor cells (microspores) can develop into embryos when subjected to heat or chemical stress (Olsen, 1987). These haploid embryos can be grown into viable barley plants. Treatment with the chemical agent Colchicine is used to double the chromosome number. A different method for haploid induction in barley are wide crosses with *Hordeum bulbosum*. Barley plants are pollinated with pollen from its relative *H. bulbosum* (Kasha and Kao, 1970). During early cell divisions after the formation of the zygote, *H. bulbosum* DNA is eliminated and only the haploid set of barley chromosomes is retained (Sanei et al., 2011). Conceptually similar approaches of doubled haploid induction – wide crosses or using gametophyte progenitors – are available for many other plant species (see Forster et al. (2007) for a recent review).

Irrespective of the type of population, the parents of a mapping population should be genetically diverse in order to show a high degree of polymorphism throughout the genome and to avoid regions of identity-by-descent where both parents have inherited a chromosomal segment from a common ancestor. Yet, excessive genetic distance – created, for example, by crossing crops with wild progenitors – is to be avoided as this may result in sterile progeny or segregation distortion, that is deviations from Mendelian segregation patterns for single markers or groups of markers which may be caused by gametic or post-zygotic selection (Semagn et al., 2006).

2.1.4 Genetic map construction

A genetic linkage map is a linear arrangement of markers reflecting the order of markers on the chromosomes. Distances between markers are given in terms of recombination frequencies. Linkage maps are collinear with physical maps, i. e. genetic markers occur (by definition) in the same order in a genetic map as they do in a physical map. Genetic map distances, however, may differ from physical distances as the recombination rate is not distributed uniformly along the length of the chromosomes.

Given the genotypic data of a mapping population, marker order can be established with the help of computational algorithms. The algorithmic problem of constructing a genetic map from marker data can be divided into three subproblems (Cheema and Dicks, 2009):

(i) Marker grouping. A grouping is a partition of markers into linkage groups. Ideally, there should be a one-to-one correspondence between linkage groups and chromosomes.

(ii) Marker ordering. The most likely order of markers within each linkage group is determined. If a linkage group consists of n markers, there are $\frac{n!}{2}$ possibilities of ordering them. The algorithmic problem of establishing an optimal marker ordering has been shown to be an instance of the symmetric wandering salesman problem, an NP-hard problem (Schiex and Gaspin, 1997).

(iii) Marker spacing. The last step in genetic map construction is to assign a map distance to the interval between each adjacent pair of markers. The unit of map distance is named centiMorgan in honor of Thomas Hunt Morgan. One centiMorgan is equivalent to one recombination event in 100 meioses. There are two commonly used functions to calculate the map distances given the recombination fraction between two loci. The function of Haldane (1919) assumes that all crossover events occur independently of each other, while the function of Kosambi (1943) accounts for positive interference, the phenomenon that the distance of two crossovers on the same chromosome is on average larger than would be expected by chance.

In addition to its algorithmic complexity, genetic map estimation is complicated by errors in raw data collection (Hackett and Broadfoot, 2003). Missing genotype calls may result in an incorrect marker order as recombination events may be missed. Vice versa, genotyping errors can inflate the length of a genetic map because of false-positive crossovers. Segregation distortion may render populations unsuitable for genetic map construction. Since the need to order several hundred DNA markers has arisen in the early 1980s, many software programs for genetic mapping employing diverse algorithmic approaches have been developed. A detailed review of genetic mapping software has been written by Cheema and Dicks (2009).

2.2 Next generation sequencing technologies

After Sanger et al. (1977b) described a method to determine the nucleotide sequence of a DNA strand, the basic principle of DNA sequencing remained unchanged for the next almost thirty years. Sanger sequencing combines terminated reverse strand synthesis with fragment size determination to resolve the order of nucleotides on a DNA template strand. The incorporation of labeled nucleotides, dideoxynucleotides (ddNTPs), terminates the synthesis of a new DNA strand from an existing template at a certain point. As many template strands and a proportionate amount of ddNTPs are used, the stochastic nature of DNA replication will result in DNA molecules of all possible sizes. The fragments are separated by molecular weight and the labels attached at the end of each fragment are read out sequentially.

Initially, ddNTPs were labeled radioactively and size fractionation was performed by gel electrophoresis, making DNA sequencing a laborious and time-consuming procedure. Subsequently, DNA sequencing machines were developed that performed size separation by capillary electrophoresis, detected nucleotides from fluorescent dyes and recorded them automatically. Nevertheless, daily throughput of the latest generation of Sanger machines did not exceed a few Mb per day and cost about US $500 per Mb (Kircher and Kelso, 2010). Consequently, the Human Genome Project used several hundred DNA sequencers distributed over twenty sequencing centers in six countries (International Human Genome Sequencing Consortium, 2001) to sequence the 3 Gb human genome.

At the time of its publication, the complete sequence of the human genome had been likened in importance to the moon landing (Pääbo, 2001). Stretching this simile a little further, the achievements of genomics since the initial release of the human genome sequence are tantamount to having established economy class flights to lunar colonies. The tremendous increase in sequencing throughput – and the commensurate drop in sequencing costs – during the last decade surpassed in magnitude the exponential growth in compute power during the second half of the 20th century as described by Moore's law

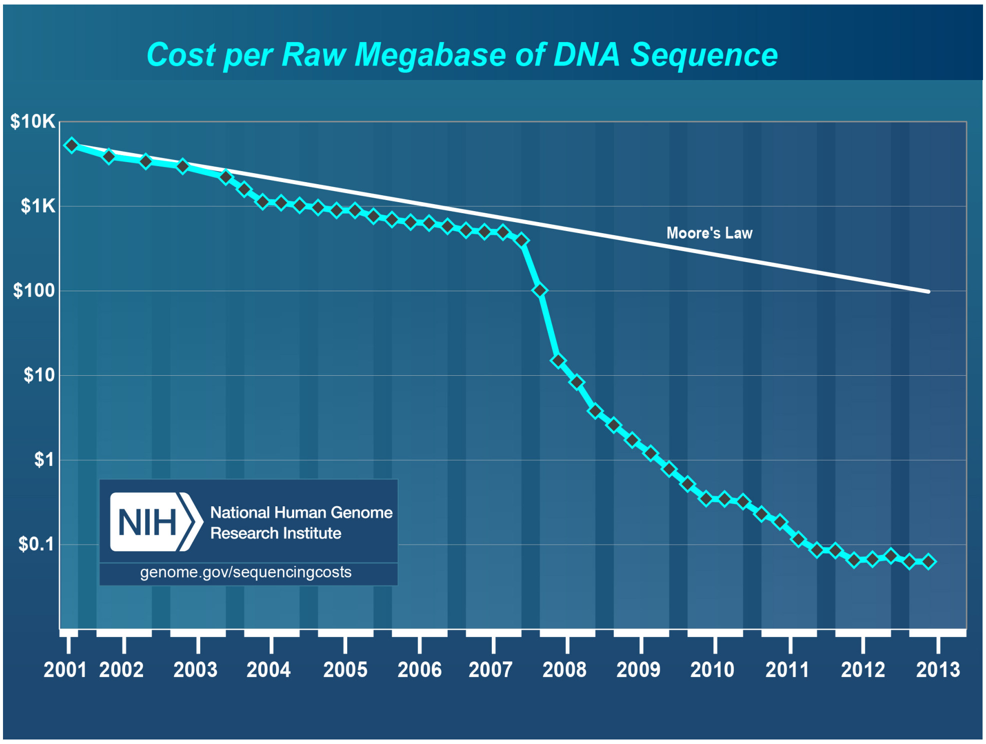


Figure 2.4: Development of sequencing costs since 2001 compared to Moore's law, which states that the number of transistors on an integrated circuit doubles every two years. This figure was taken from `http://genome.gov/sequencingcosts`.

(Figure 2.4).

This development unthought-of at the turn of the century has made genome sequencing the method of choice for many research projects in basic and applied science. Large-scale analyses of genetic diversity within (The 1000 Genomes Project Consortium, 2010) and between (Prado-Martinez et al., 2013) species achieving single nucleotide resolution have become possible. Genomics has given rise to the concept of personalized medicine (Hamburg and Collins, 2010). In plant science, the increased throughput of genome sequencers has enabled the sequence analysis of genomes that are up to seven times larger than the human genome, such as the genomes of barley (IBSC, 2012), wheat (Brenchley et al., 2012) and spruce (Nystedt et al., 2013). Resequencing a large number of cultivated and wild accessions of rice (Huang et al., 2012) and maize (Hufford et al., 2012) has provided insights into the origin of domesticated crops and identified targets of selection by ancient farmers as well as by modern breeders.

"The main principle underlying NGS technologies is sequencing-by-synthesis" (Nielsen et al., 2011). As small single-stranded DNA molecules are enzymatically copied, new DNA molecules complementary to the template strands are built up. This process, performed simultaneously for millions of DNA fragments, is recorded by high-resolution imaging or electronic sensors and the signals are converted in a sequence of A, C, G and T by computer software. Since 2005, several technologies exploiting this principle have hit the market. The promises they initially held (and which they have largely been able to deliver) have caused these devices to be heralded as the next generation of sequencing, the long-awaited successors of Sanger technology. Some of them, such as SOLiD or Helicos, have not met with commercial success. New platforms such as Oxford Nanopore (`https://www.nanoporetech.com`) have already appeared on the horizon.

In the following description of the basic working principles and key performance parameters of NGS sequencers, we will focus on two plaforms that have been in widespread use for *de novo* assembly and resequencing studies – 454 and Illumina sequencing. A brief comparison of their specifications is given in Table 2.1. This overview also includes two other technologies, IonTorrent and PacBio sequencing, which have been launched more recently.

2.2.1 Roche 454

The first commercially available NGS platform was 454 sequencing (Margulies et al., 2005). 454 technology couples bead-based emulsion PCR with highly parallel pyrosequening in microreactors. DNA is fractionated to single-stranded fragments of 500 – 1,000 nucleotides in size. Adapter sequences are ligated to the fragments to mediate the binding of fragments to the surface of micro-beads. Initially, each bead is bound to only a single fragment which is then amplified to several millions of copies through emulsion PCR. In emulsion PCR, a single DNA molecule attached to a bead is isolated in water droplets in an oil phase. A PCR reaction then produces clonal copies that are attached to the bead. The beads coated with the amplified fragments are then loaded onto a plate of microwells. The size of the microwells allows exactly one bead per reaction well.

The sequencing reaction itself is based on pyrosequencing (Ronaghi et al., 1998). Unlike Sanger sequencing which works by chain termination with labeled nucleotides, pyrosequencing detects the release of pyrophosphate when a nucleotide is incorporated into a template strand. Pyrosequencing implements the principle of sequencing-by-synthesis. Taking a single strand of DNA as a template, the complementary strand is synthesized by a DNA polymerase. Solutions of unmodified deoxyribonucletoides (dNTP) are sequentially added to the immobilized template DNA. If the dNTP added in the current cycle is complementary to the first unpaired base on the template strand, this dNTP is incorporated into the template resulting in the release of pyrophosphate. Other enzymes convert this pyrophosphate into a chemiluminescent light signal that is recorded by a CCD camera. This reaction is performed separately

Table 2.1: Overview of sequencing platforms. Specifications according to vendors. The information reflects the state-of-the-art as of October 2013.

	ABI 3730xl	GS FLX Titanium XL+	HiSeq 2000	Ion Proton	PacBio RS II
vendor	Life Technologies[1]	Roche 454[2]	Illumina[3]	Life Technologies[4]	Pacific BioScience[5]
amplification	clonal	emulsion PCR	bridge PCR	emulsion PCR	no amplification
read length (bp)	800 – 1,000	500 – 1,000	2 × 100	up to 200	500 – 20,000
no. of reads per run	96	1×10^6	6×10^9	70×10^6	45×10^3
sequence output per run	900 kb	700 Mb	600 Gb	10 Gb	200 Mb
run time	2 h	23 h	11 d	2 – 4 h	30 – 120 min

[1] `http://www.appliedbiosystems.com`

[2] `http://www.454.com`

[3] `http://www.illumina.com`

[4] `http://de-de.invitrogen.com`

[5] `http://www.pacificbiosciences.com`

and in parellel in each microwell.

The major source of errors in pyrosequencing are runs of homopolymers. If several identical nucleotides occur in a row, more than one dNTP is incorporated in one cycle of the machine. For small homonucleotide tracts (< 6), the signal strength is proportional to the number of incorporated nucleotides. Higher signal intensities, however, cannot be disambiguated due to a saturation effect.

The first sequencing machine marketed by 454 Life Sciences (now a subsidiary of Roche) was the GenomeSequencer 20 (GS20). This device had a throughput of 20 Mb of sequence per run (1 day) and provided read lengths up to 100 bp. This constituted a 100-fold increase of sequence throughput compared to Sanger sequencing technology. The first application was a whole genome shotgun assembly of the bacterium *Mycoplasma genitalium* (Margulies et al., 2005). Shortly afterwards, Green et al. (2006) applied 454 technology to sequence ∼15,000 unique fragments of ancient Neanderthal DNA. Subsequently, both throughput and read length of the 454 platform have been successively increased by improvements in both sequencing chemistry and optics. 454 sequencing has been used for resequencing studies in humans (Wheeler et al., 2008) and to assemble eukaryotic genomes (Diguistini et al., 2009; Velasco et al., 2007), often in conjunction with Sanger or Illumina reads. 454 technology has been extensively applied in RNA sequencing both for transcriptome reconstruction and SNP discovery (see for example Barbazuk et al. (2007) and Emrich et al. (2007)). Currently, Roche offers the GS FLX Titanium XL+ instrument that achieves read lengths of up to 1,000 bp, making it comparable to Sanger sequencing. Limited to an output of only 700 Mb per day, however, the instrument is not on par with other current sequencing platforms.

2.2.2 Illumina

Illumina sequencing combines sequencing-by-synthesis using reversible dye-terminator chemistry with bridge PCR to generate a high-density array of clusters of amplified DNA fragments (Bentley et al., 2008). It was commercially launched in 2006 by the company Solexa, which has subsequently been acquired by Illumina. In Illumina sequencing, DNA is fragmented and special adapter sequences are ligated to both ends of the fragments. These adapters bind to primers fixated on a glass surface (a so-called flow cell). The immobilized DNA molecules are then amplified to form DNA colonies (or clusters) where ∼1,000 clonal copies are localized within one micrometer of each other. Actual DNA sequencing is carried out through sequencing-by-synthesis with reversible dye-terminators. In each cycle of the machine, nucleotides labeled with fluorescent dyes are incorporated into a growing DNA strand. These nucleotides terminates the activity of the polymerase. After each cycle an image of the flow cell is taken. Subsequently, the fluorescent dye is cleaved enzymatically to prepare the newly synthesized DNA strand for the incorporation of another nucleotide in the next cycle. The images taken in each cycle are converted into base calls. Densities of DNA clusters of up to ten million per cm^2 are possible with Illumina sequencing, allowing several hundred million reads to be sequenced in one run of the machine. Initially, read length on the Illumina platform was limited to 35 bp because the reliability of base calls decreases in later cycles. Improvement in sequencing chemistry and base calling software have since enabled an increase to 100 bp in the standard mode of the HiSeq 2000. The rapid run mode of the HiSeq 2000 or the MiSeq achieve read lengths of 150 and 250 bp, respectively, but have a lower throughput.

Paired-end sequencing is possible on the Illumina platform. A single short DNA fragment is sequenced from both ends, resulting in a pair of reads for each fragment. The two reads of a fragment can be aligned to a reference sequence where one uniquely mapping read in a pair may guide the alignment of a second repetitive reads thus improving the mapping rate. While paired-end sequencing originally referred to sequencing fragments of any length (Roach et al., 1995), the term "paired-end reads" has nowadays come to denote reads from fragments of about 200 to 800 bp in size. Due to limitations of bridge PCR, it is not possible to sequence fragments larger than ∼1 kb on the Illumina platform. However, special library preparation protocols enable mate-pair sequencing, i. e. paired-end sequencing with a distance of 1 to 20 kb between the two reads of a pair. Prior to sequencing, long fragments are circularized by ligating both ends of a fragment. The circularization site is labeled. Circular DNA is fragmented to ∼200 – 500 kb and only labeled fragments corresponding to the ends of the original long fragments are sequenced in paired-end mode. Mate-pair sequening has the same applications as Sanger mate-pairs in "double barrel shotgun sequencing" (Roach et al., 1995; Pevzner and Tang, 2001). They can be used to bridge gaps between sequence contigs and provide linking information that can be leveraged to order and orientate contigs.

Multiplexing is crucial to many applications of Illumina sequencing. Even using a single lane of an eight-lane flow cell would result in an excessive coverage of small genomes. For example, the 150 Mb genome of the model plant *Arabidopsis thaliana* can be sequenced to 200-fold coverage with one lane of a HiSeq 2000. For genotyping applications, however, it is desirable not to sequence one sample to high coverage, but to sequence samples from many different individuals to shallow coverage (1 to 10x). Several samples can be sequenced in one Illumina lane by employing barcoding. Oligonucleotides are ligated to fragment ends prior to sequencing. The index oligos are sequenced in a separate index read and used to assign reads back to individual samples by bioinformatical means.

The enormous throughput of Illumina sequencing has enabled whole genome shotgun sequencing and *de novo* assembly of large and complex genomes. Benefits and shortcoming of this application will be discussed in more detail in section 2.3.3. Apart from *de novo* assembly, Illumina sequencing is currently the method of choice for large-scale resequencing projects, such as the human 1000 genomes project (The 1000 Genomes Project Consortium, 2010) or the Arabidopsis 1001 genomes project (Weigel and Mott, 2009). The enormous output of Illumina sequencing machines has spurred the development of new assembly and alignment algorithms (Trapnell and Salzberg, 2009) as well as new file formats (Li et al., 2009a) and hardware infrastructures for data storage (Schatz et al., 2010). Besides whole genome resequencing, Illumina sequencing is used in conjunction with restriction-enzyme mediated or hybridization-based complexity reduction methods, such as genotyping-by-sequencing or exome capture, for population-scale genotyping (see section 2.1.2).

Their two high-throughput devices, the GenomeAnalyzer and later the HiSeq 2000, have secured Illumina a comfortable market share – a position only to be challenged when new technologies will offer a throughput and reliability comparable to the HiSeq 2000. While there has been little technological improvement of the HiSeq 2000 instrument itself over the last few years, the company Moleculo (now acquired by Illumina) has developed a technique to produce synthetic long reads on the HiSeq 2000. Fragments of ∼6 − 8 kb in size are labeled with individual barcodes that allow demultiplexing after pooled sequencing and separate assembly of each fragment. Assembly programs can treat these long fragments as long input reads (Voskoboynik et al., 2013).

2.3 Genome assembly strategies

No sequencing technology is currently able to determine the sequence of a chromosome from one end to the other with a single sequencing read. Instead, all sequence machines yield (probably billions of) DNA pieces ('reads') from 100 to a few thousand base pairs in lengths. The process of genome assembly is putting together overlapping reads into "contigs", contiguous stretches of sequence. This definition of "contig" was given by Staden (1979). A second meaning of the word "contig" was introduced by Coulson et al. (1986) as a group of overlapping clones in a fingerprinting project. Nowadays, "clone" in this context refers to bacterial artificial chromosomes (BACs), genomic fragments of 50 − 300 kb that are maintained and multiplied with the help of *E. coli* bacteria. To distinguish both concepts of 'contig', we will refer to the first kind of contigs as sequence contigs and to the second kind as physical contigs or BAC contigs. Likewise, "assembly" can either refer to the grouping and ordering of sequence reads into a contiguous chain of nucleotides or to the grouping and ordering of overlapping BAC clones into physical contigs.

Shotgun sequencing denotes the random fragmentation (shotgunning) of DNA and subsequent sequencing and assembly of these fragments to recapture the original sequence. It had been originally proposed as a theoretical concept by Staden (1979). One of the first shotgun libraries was prepared and sequenced from a 4,257 bp fragment of bovine mitochondrial DNA (Anderson, 1981). A shotgun library can be made either from genomic DNA – as is the case for whole genome shotgun sequencing – or from only a subset of the complete genome – as is the case for hierarchical shotgun sequencing. Initially, shotgunning meant random subcloning, that is the random fragments were cloned into vectors for *in vivo* amplification prior to Sanger sequencing. As next generation sequencing employs PCR-based amplification methods, no subcloning is necessary anymore. Instead, a large subsample of all shotgun fragment is sequenced and assembled with the help of computers

In the following sections, we will describe computational methods for DNA fragment assembly as well as the two main approaches to assemble the sequence of complete genomes: hierarchical and whole genome shotgun sequencing.

2.3.1 DNA fragment assembly

Sequence assembly is an algorithmic problem that involves the alignment and merging of adjacent reads to form a longer contiguous sequence. Some assemblers faithfully follow this paradigm. Overlap-layout-consensus (OLC) assemblers compare all sequencing reads against each other to find overlaps. Groups of overlapping reads are brought into the presumably correct order and are then merged into a contiguous consensus sequence (Kececioglu and Myers, 1995). The layout step can be thought of as a problem in graph theory. Reads are considered as nodes of a graph that are joined by edges if they overlap. A path through this graph should visit each node exactly once. In other words, OLC assemblers search for a Hamiltonian path. This problem is NP-hard, i. e. there is currently no general solution for it that achieves polynomial time complexity.

Pevzner et al. (2001) phrased DNA fragment assembly in a different way. They split sequence reads into k-mers, that is nucleotide sequences of fixed length k, where k is less or equal to the read length. A so-called de Bruijn graph is constructed where the nodes are all k-mers that are present in the sequence reads. Two k-mers are connected by an edge if they share a common $(k-1)$-mer. An assembly of the reads is then defined as a traversal of the de Bruijn graph that visits each edge only once, i. e. an Eulerian path of the de Bruijn graph. Finding an Eulerian path is a computationally tractable problem. However, (Medvedev et al., 2007) noted that a valid genome assembly should incorporate all reads (i. e. subpaths of the de Bruijn graph) and formulated the genome assembly problem as the Shortest De Bruijn Superwalk problem, which they showed to be NP-hard. Nevertheless, as the number of nodes of k-mer-based de Bruijn graphs is smaller than in read overlap graphs, de Bruijn graph assemblers run faster in practice than OLC assemblers and can cope with much larger amounts of input data. The drawback of de Bruijn graph method is the loss of connectivity information provided by long reads.

In the presence of repetitive DNA, a major challenge in fragment assembly is to correctly resolve repeat structures. Regardless of the algorithmic approach – OLC or de Bruijn – repeats will manifest themselves in complex graph structures that are difficult to disentangle and often result in short contigs that represent multiple collapsed copies of the same repeat element from distinct genomic regions. Further difficulties arise form sequencing errors that may obscure read overlaps in OLC assemblers or give rise to spurious nodes and edges that appear as tips or bubbles in the de Bruijn graph. Each assembly pipeline applies different heuristic methods to deal with repeats and sequencing errors. Repeat structures are commonly resolved by incorporating paired-end and mate-pair data during the assembly process in order to bridge gaps caused by unresolved repetitive sequence and to link physically close contigs. This scaffolding process can be performed directly by the assembler (Gnerre et al., 2011; Li et al., 2010b) or be accomplished by dedicated software (Boetzer et al., 2011; Lindsay et al., 2012). Sequence errors may be detected by inspecting the distribution of k-mers. Single base substitution error or short indels will show up as low frequency k-mers that can be eliminated prior to assembly by discarding or correcting reads that contain putatively erroneous k-mers (Kelley et al., 2010). Alternatively, the number of $(k-1)$-mers that connects pairs of nodes of the de Bruijn graph can be tracked and badly supported edges be pruned from the graph (Li et al., 2010b)

In the absence of a ground truth against which assembly results can be validated, several efforts of comparing assembler performance on real and simulated data have been carried out (Kumar and Blaxter, 2010; Salzberg et al., 2012; Earl et al., 2011), arriving at the equivocal verdict that no single assembler serves best all purposes. The choice of assemblers and their parameters in sequencing project is usually made *ad hoc* and governed by the amount and quality of the input data, the size and structure of the genome of interest and the available compute resources. Because of the quadratic time complexity of the all-against-all comparison in the alignment step, OLC assemblers cannot cope with Illumina sequencing data and therefore de Bruijn graph assemblers are used for WGS assemblies from whole genome shotgun data. Assemblies of single BAC clones or of transcriptome data sets from 454 data can still be performed with OLC assemblers (Kumar and Blaxter, 2010; Steuernagel et al., 2009).

2.3.2 Hierarchical shotgun sequencing

Map-based or hierarchical sequencing approaches reduce the complexity of the assembly process by partitioning the genome into smaller pieces that are sequenced and assembled autonomously (Figure 2.5). The individual assemblies are patched together afterwards to obtain one contiguous pseudomolecule for each chromosome.

The most common method to divide a genome into pieces is the construction of one or more libraries of bacterial artificial chromosomes (BACs) (Shizuya

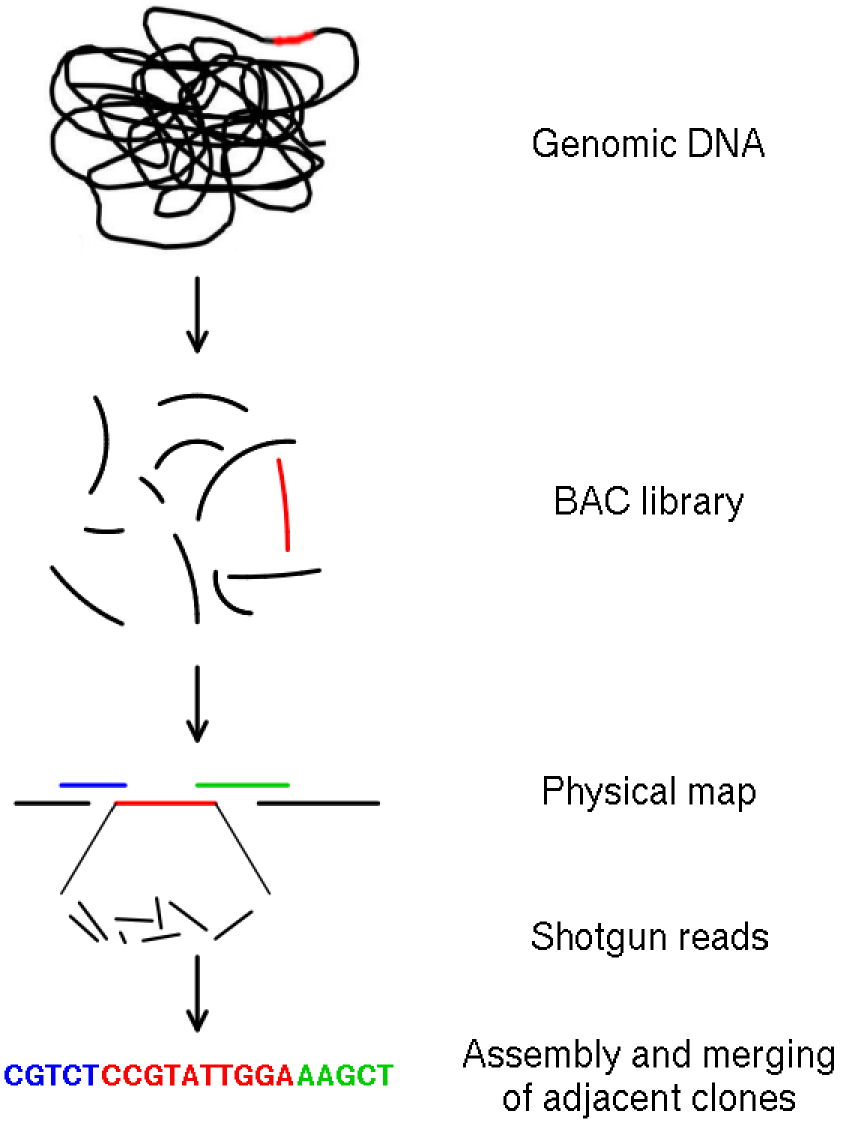


Figure 2.5: Schematic overview of hierarchical shotgun sequencing
A BAC library is constructed from genomic DNA. A genome-wide physical map is constructed and a minimum set of overlapping BACs that completely cover the genome is selected for shotgun sequencing. DNA of each BAC is fragmented and sequenced. Shotgun reads are assembled by computer programs into a contiguous sequence. The assemblies of adjacent clones within a physical contig are merged. This figure was adapted from the International Human Genome Sequencing Consortium (2001).

et al., 1992). Genomic DNA is fragmented to pieces that are between 50 and 300 kb in size by partial digestion with a restriction endonuclease or by mechanical shearing. These fragments are size fractionated by pulsed-field gel electrophoresis and are subsequently cloned into a plasmid of the bacterium *Escherichia coli.* Plasmid vectors are transformed into *E. coli* cells, which are then grown in an antimicrobial medium, selecting for transformed cells that have acquired resistance through their plasmid. Single transformed bacteria grow into spatially isolated colonies that can be picked and arrayed. This results in a library of individually addressable DNA fragments that can be maintained in a convenient 96 or 384 well plate format in laboratory freezers. Many physical mapping projects used several different BAC libraries, where the initial partial digestion was performed with different restriction enzymes in order to avoid the under-representation of genomic regions due to a paucity of a single type of restriction sites (The International Human Genome Mapping Consortium, 2001; Chen et al., 2002; Wei et al., 2009; IBSC, 2012).

In principle, it can be imagined to sequence all BACs of a library. For large genomes, however, this would involve sequencing several hundred thousand clones to obtain sufficient clone coverage in order to avoid gaps in the final assembly (Lander and Waterman, 1988). Adjacency of BAC clones is commonly not identified by sequencing all clones and subsequent *in silico* sequence analysis, but by the experimental construction of a restriction map where overlaps between BACs are uncovered by similarity of restriction patterns. Digestion of BAC DNA with a restriction enzyme results in a reproducible pattern of restriction fragment lengths, so-called fingerprints. Overlapping BAC clones partially share their fingerprints. Thus, similar fingerprints can be used to detect sequence overlap between adjacent BACs and to group them into contigs. A set of BAC contigs is called a physical map. Following the initial description of clone fingerprinting and contiging (Coulson et al., 1986; Olson et al., 1986), improved wet-lab methods for enzymatic fingerprinting (Gregory et al., 1997; Luo et al., 2003), mathematical models of clone overlap (Lander and Waterman, 1988) as well as integrated software toolkits for the fingerprint contig assembly (Soderlund et al., 1997) have been developed. Moreover, novel methods for comparing restriction fragments based on their sequence instead of their size have been established by van Oeveren et al. (2011).

Once a physical map has been set up, a minimum set of BAC clones that cover each contig with as little redundancy as possible is determined. Each clone of this so-called minimum tiling path (MTP) is then shotgun sequenced and assembled. The sequence assemblies of adjacent BAC clones are combined into a non-redundant sequence applying *ad hoc* work flows for overlap detection, elimination of misassemblies and merging of overlapping sequence contigs (see for example Kent and Haussler (2001) and Wei et al. (2009) for the assembly pipelines of the map-based sequences of the human and maize genomes).

Hierarchical shotgun sequencing has been used to assemble the reference genome sequences of several plant and animal genomes (Table 2.2), most notably the human genome (International Human Genome Sequencing Consortium, 2001). Map-based sequencing is nowadays recognized as an established – though tedious – path towards the reference sequence of a large and complex genome where abundant repeat structures would compromise the effectiveness of a whole genome shotgun strategy (Feuillet et al., 2012). The major drawback of the hierarchical strategy is the substantial amount of time, labor and finances required for the development, maintenance and use of the necessary resources.

Map-based assemblies still have been published in recent years and further ones are still in the making. For instance, the high-quality genome sequences of swine (Groenen et al., 2012) or zebra fish (Howe et al., 2013) have been published recently, and hierarchical shotgun sequencing is the approach adopted by the wheat and barley sequencing projects. These genome projects are carried under the auspices of international sequencing consortia similar in structure and goals to the Human Genome Project. Finished map-based sequences are the cumulation of long-time collaborative efforts of laboratories in many different countries. This strategy is only possible for species receiving the attention of large research communities backed by strong scientific or economic interests, as is the case for genetic model organisms or species of agronomic importance. Poorly resourced research communities, however, have no choice but to endorse a WGS strategy.

2.3.3 Whole genome shotgun sequencing

In whole genome shotgun sequencing, the complete genomic DNA is randomly fragmented and sequenced, omitting the intermediary steps of BAC library and physical map construction. As early as 1979, Rodger Staden wrote that "with modern fast sequencing techniques and suitable computer programs it is now possible to sequence whole genomes without the need of a restriction map" (Staden, 1979). This predated the lively debate (Weber and Myers, 1997; Green, 1997) on how to assemble the human genome by a mere eighteen years. A restriction map was still used to obtain the first complete viral DNA genome sequence (Sanger et al., 1977a). Gardner et al. (1981) were the first to use shotgun sequencing to obtain a complete genome, the sequence of the cauliflower mosaic virus. After the first bacterial genomes (Fleischmann et al., 1995; Blattner et al., 1997) had been successfully completed using the whole genome shotgun strategy, it became the method of choice for prokaryotic genome projects.

Weber and Myers (1997) proposed that paired-end sequencing of randomly selected DNA fragments of different insert sizes would enable the completion of the human genome sequence in a more rapid and cost-efficient way than would be possible with the hierarchical method adopted by the international consortium. This opinion was immediately challenged by Green (1997). His

Table 2.2: Progress in genome sequencing

organism	common name	size	year	strategy	remark	reference
Bacteriophage MS2		3,569 nt	1976		first completed RNA genome	Fiers et al. (1976)
Bacteriophage ϕX174		5,375 nt	1977		first completed DNA genome	Sanger et al. (1977a)
Cauliflower mosaic virus CM1841		8,031 nt	1981	WGS	first use of WGS to sequence a complete genome	Gardner et al. (1981)
H. influenzae		1.8 Mb	1995	WGS	first finished genome of a free-living organism	Fleischmann et al. (1995)
E. coli		4 Mb	1997	WGS		Blattner et al. (1997)
S. cerevisiae	baker's yeast	12 Mb	1996	clone-by-clone	first finished eukaryotic genome	Goffeau et al. (1996)
C. elegans	nematode worm	100 Mb	1998	clone-by-clone	first finished genome of a multicellular organism	Caenorhabditis elegans Sequencing Consortium (1998)
D. melanogaster	fruit fly	120 Mb	2000	clone-by-clone and WGS	first application of WGS to a eukaryotic genome	Adams et al. (2000)
A. thaliana	thale cress	150 Mb	2000	clone-by-clone	first sequenced plant genome	Arabidopsis Genome Initiative (2000)
H. sapiens	human	3.2 Gb	2001	clone-by-clone and WGS	initial draft sequence	International Human Genome Sequencing Consortium (2001); Venter et al. (2001)
O. sativa	rice	450 Mb	2002	WGS	initial draft sequence of the rice genome	Goff et al. (2002); Yu et al. (2002)
O. sativa	rice	450 Mb	2005	clone-by-clone	map-based finished sequence of the rice genome	International Rice Genome Sequencing Project (2005)
Z. mays	maize	2.3 Gb	2009	clone-by-clone		Schnable et al. (2009)
A. melanoleura	giant panda	2.4 Gb	2009	WGS	first eukaryotic genome assembled only from NGS reads	Li et al. (2010a)
B. distachyon	purple false brome	270 Mb	2010	WGS	draft assembly of a model grass	The International Brachypodium Initiative (2010)
F. albcollis	flycatcher	1.1 Gb	2012	WGS	draft assembly used for evolutionary studies	Ellegren et al. (2012)
S. scrofa	pig	2.6 Gb	2012	clone-by-clone and WGS	hierarchical shotgun assembly supplemented with WGS data for gap-filling	Groenen et al. (2012)
H. vulgare	barley	5.1 Gb	2012	WGS	gene space assembly representing only 1.9 Gb	IBSC (2012)
T. aestivum	bread wheat	17 Gb	2012	WGS	partial gene space assembly	Brenchley et al. (2012)
P. abies	Norway spruce	19.6 Gb	2013	WGS	gene space assembly representing only 4.3 Gb	Nystedt et al. (2013)

main points of criticism were the difficulties in finishing a WGS assembly, i. e. filling the gaps between contigs would be more difficult as contig ends would have to be amplified from possibly repetitive genomic DNA, whereas in a clone-by-clone approach, gaps could be traced to a small genomic region from which clones could be slated for additional raw data collection. Moreover, in Green's opinion, the computer simulations of Weber and Myers (1997) to prove the feasibility of their approach were oversimplified and did not take into account the complex (and at the time mostly unknown) structure of repeat families in the human genome.

This debate was not resolved graciously, but led to a schism. Craig Venter left the human sequencing consortium and raised private capital to sequence the human genome using the WGS approach. The ensuing competition between the public consortium and the Celera corporation culminated in the simultaneously publication of two human genome sequences (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001). A detailed comparison of the Celera whole genome shotgun assembly and the map-based sequence revealed overall agreement in local sequence content (Istrail et al., 2004) with most discrepancies between both assemblies arising from misplaced scaffolds. The whole genome assembly could fill gaps of the clone-by-clone sequence, which, in turn, was superior in resolving highly similar repeats. However, Waterston et al. (2002) noted that Venter et al. (2001) had had access to the publicly available clone assemblies of the international consortium and incorporated them into their assembly. Thus, the Celera assembly cannot be considered a genuine WGS assembly. Moreover, Green (2002) expressed doubts as to whether the Celera assembly was really faster and cheaper compared to the map-based sequence, given the generous support from a vendor of sequencing machines.

These controversies now seem arcane as the whole genome shotgun assembly of large and complex eukaryotic genomes has gained irresistible momentum with the previously unimagined rapid accumulation of sequence data through next generation sequencing technology. For example, sequencing reads equivalent to 100x coverage of the human genome can now be collected from a single flow cell of an Illumina HiSeq 2000 within ten days' time.

The first genome sequence assembled only with next generation sequencing data was the Giant Panda genome by Li et al. (2010a). Subsequently, the same group reported *de novo* assemblies of two human genomes from Illumina reads (Li et al., 2010b). Comparison of these two assemblies to the human reference sequence indicated a local sequence accuracy greater than 99 %. However, assembly contiguity was inferior with contig sizes of a few kilobases and scaffold sizes of less than of 62 kb and 446 kb (Li et al., 2010b). These initial results highlighted the importance of long-distance mate-pair data to link and order sequence contigs in low-copy regions across assembly gaps corresponding to repetitive elements. Exploiting this principle further, the assembler ALLPATHS-LG (Gnerre et al., 2011) has been developed to utilize a narrowly defined set of short and long insert libraries. It has been shown to deliver high-quality draft assemblies of the human and mouse genomes with megabase-sized scaffolds. This combined recipe of library preparation and assembly algorithm comes with the cost that the input data is limited to reads from specific libraries. In particular, commonly used libraries of 300 – 500 bp insert size cannot be utilized by ALLPATHS-LG.

Although deep read coverage can partially compensate for the shortcomings of NGS technology, such as the short read lengths and higher error rates when compared to capillary sequencing, the weaknesses of WGS assemblies originally pointed out by Green (1997) still persist. Repetitive DNA is not well represented in WGS assemblies and finishing a collection of hundred thousands of pieces separated by gaps of unknown size is an undertaking not contemplated by even the most audacious. Furthermore, concerns have been raised about the completeness and accuracy of genome assemblies constructed solely from next generation sequencing data. Alkan et al. (2011) examined the *de novo* assemblies of two human individuals published by Li et al. (2010b). Among other things, Alkan et al. (2011) found that the assemblies were 16 % shorter than the map-based reference genome. Missing sequence was mostly due to collapsed repetitive regions. Alkan et al. (2011) urged the scientific community to enforce standards for genome quality. Chain et al. (2009) had already proposed a quality scale for completeness and correctness of assemblies and annotation of microbial genomes. Feuillet et al. (2011) adopted this scale for the classification of plant genome assemblies. So far, only the rice and Arabidopsis genomes have achieved the highest grade of “finished genome sequence” (Feuillet et al., 2011).

Many researchers take a pragmatic stance towards the shortcomings of WGS assemblies. An incomplete and imperfect assembly is still better than no assembly at all. So-called gene-space assemblies are created from whole genome shotgun reads. These assemblies are highly fragmented and often consist of more than one million contigs with an N50 of a few kilobases. The cumulative length of all sequence contigs is usually substantially shorter than the actual genome size and only low-copy regions (the “gene space”) are expected to be correctly represented. Often little effort has been spent to advance these assemblies beyond what a short read assembler outputs, and substantial improvements would often not have been possible due the limitation of short read NGS technology in resolving long and often nested repeat structures or recently duplicated regions. Gene space assemblies are considered as interim solutions as long as a full reference sequence based on an improved WGS or clone-by-clone assembly is not available. Such gene space assemblies have recently been reported for barley (IBSC, 2012), Norway spruce (Nystedt et al., 2013) and chickpea (Jain et al., 2013). The assembly of barley was less than half as long as the true 5 Gb genome size. However, 86 % of all barley genes are expected to be represented by this resource based on a comparison to core metabolic genes of *A. thaliana* (IBSC, 2012). While the WGS assembly of barley was partially anchored to genetic maps and a genome-wide physical map, neither genetic nor physical anchoring information was provided for the spruce genome.

2.4 Methods for anchoring sequence assemblies

Anchoring is the process of integrating different type of genomic datasets into a common structure. One data domain is used as a backbone to which the other domains are linked. Most commonly, physical maps or sequence assemblies are assigned to chromosomal locations given in terms of coordinates on a genetic map. In this situation, a physical map – a local grouping and ordering of BAC clones – or a sequence assemblies – a local grouping and ordering of sequence reads – is complemented with global information regarding the position of its constituents, BAC or sequence contigs. Likewise, the genetic backbone is populated with information about gene content in the vicinity of each marker and genetic distance given in terms of recombination frequencies can be related to physical distance given in terms of megabases.

This section describes previous methods that were used to anchor physical maps and sequence assemblies to genetic maps. It also includes a detailed description of the physical and genetic framework of barley, a resource that we will later employ to illustrate the feasibility of POPSEQ in barley.

2.4.1 Integrating physical and genetic maps

The integration of physical and genetic maps is not a standardized approach, but each genome project made good use of the methods available at the time in a rather *ad hoc* manner. One or more genetic maps are associated with physical contigs through the identification of BAC clones harboring genetic markers. Positioning of sequence or physical contigs on the chromosomes is not only achieved through genetic marker information alone, but is supplemented by short-range connectivity afforded by a restriction map.

Physical maps can be linked to genetic markers in the absence of sequence information. To this end, BACs are combined into multidimensional pools to enable the simultaneous assay of BACs corresponding to several genome equivalents. These pools are screened by DNA hybridization, PCR-based methods or with the help of microarrays (reviewed by Ariyadasa and Stein (2012)). For small genomes, these methods can achieve high efficiency. For example, the genome-wide physical map of the ~150 Mb genome of *A. thaliana* (Mozo et al., 1999) was completely anchored and ordered through marker hybridization. This map consisted of only 27 contigs with three to seven contigs per chromosome, providing extensive connectivity within contigs. In larger genomes, however, there may be several hundred BAC contigs per chromosome and a substantial proportion of mostly short contigs remains unanchored (Wei et al., 2007; IBSC, 2012)

In clone-by-clone sequencing projects, genetic mapping, physical mapping and sequencing are usually performed in parallel (Cone et al., 2002). BAC end sequencing is often performed prior to full shotgun sequencing of clones in the minimum tiling path. BAC end sequences are obtained by two Sanger sequencing reactions that are initiated from universal primers adjacent to both insert sites. After quality trimming, BAC end sequences are on average ∼500 bp in size (Kelley et al., 1999). In the absence of more comprehensive sequence data sets, BAC end sequences can provide useful insights into the genome organization of a species, such as the gene content or the amount of repetitive sequences (Messing et al., 2004; Mao et al., 2000). When full or partial sequence information of BAC clones is available, these sequences can be searched for the sequence of known genetic markers *in silico*. BAC end sequences constitute a random subsample of the genome and originate from repetitive regions with a probability that is proportional to the repeat content of the genome. In the case of highly repetitive genomes, care must be taken not to erroneously associate marker sequences with unrelated copies of transposable elements or with paralogous copies of a gene. Yuan et al. (2000) masked repetitive sequence in EST and BAC end sequences of rice using a database of known repetitive elements and subsequently associated the cleaned data set with genetic markers by applying stringent sequence alignment filters. Another approach to masking repetitive DNA is tabulating the occurrence of k-mers and annotating repeat elements as regions of highly abundant k-mers (Kurtz et al., 2008).

Genetic anchoring of physical maps has greatly benefited from recent advances in high-throughput marker technologies. Luo et al. (2013) genotyped five-dimensional BAC pools with a 10,000 feature SNP chip, thus obviating the need for more laborious DNA hybridizations with single probes. Anchoring the physical map of barley has profited from the availability of a large number of GBS markers (IBSC, 2012). Likewise, the relative ease with which whole genome assemblies can be constructed by high-throughput sequencing can augment the map-associated sequence information to be mined for the presence of marker sequences.

2.4.2 The sequence-enriched physical and genetic map of barley

Barley (*Hordeum vulgare*) is an important source of human and animal nutrition, supplying the malting and brewing industries. It is among the earliest domesticated crop plants and is adapted to a wide range of environmental conditions. It is an inbreeding diploid species that serves as a model for genomic research in the Triticeae tribe which also includes the important cereal crops wheat and rye. A wealth of genomic resources such as dense genetic maps, ESTs and cDNA libraries and a gene expression atlas have been developed in the past two decades (reviewed in Schulte et al. (2009)). Moreover, comprehensive germplasm collections from cultivars and wild accessions (van Hintum and Menting, 2003), as well as extensive mutant collections (Druka et al., 2011)

Table 2.3: Features of the barley physical map.

FP contigs	9,265
Clones in contigs	517,202
Singletons	53,805
Contigs containing:	
>100 clones	3,151
50 – 100 clones	1,538
25 – 49 clones	1,478
3 – 24 clones	3,285
2 clones	1,035
Map length	4.9 Gb
Average contig size	538 kb
N50	904 kb

provide a sound foundation from which genetic studies into developmental and morphological processes can take their starting point. The full exploitation of these resources for basic research and crop improvement, however, has so far been hampered by the lack of a reference genome sequence. The extent to which genomic research in a crop species may be spurred by the availability of a reference genome is aptly illustrated by the large number of agronomically important rice genes that have been successfully cloned after the release of the rice genome sequence (reviewed in Huang et al. (2013)). Before the advent of next generation sequencing (NGS) technology, the scale of a barley genome project had seemed daunting, owing to the large size (5 Gb) and the high repeat content of the barley genome. After initial studies involving a small number of BACs (Steuernagel et al., 2009; Taudien et al., 2011) and shallow whole genome sequencing (Wicker et al., 2008) proved the utility of NGS for assembling a large and complex genome, sequencing the barley genome came within reach. The International Barley Genome Sequencing Consortium (IBSC) coordinated research towards a reference sequence of barley (Schulte et al., 2009) through a map-based sequencing strategy encompassing restriction-based fingerprinting that had successful precedents in the mapping and sequencing projects of several plant and animal genomes (Arabidopsis Genome Initiative, 2000; Mouse Genome Sequencing Consortium, 2002; International Rice Genome Sequencing Project, 2005; Schnable et al., 2009).

IBSC (2012) reported a major milestones on the way towards a complete sequence of the barley genome (Figure 2.6). A genome-wide physical map from six independent BAC libraries (Schulte et al., 2011) of the barley cultivar

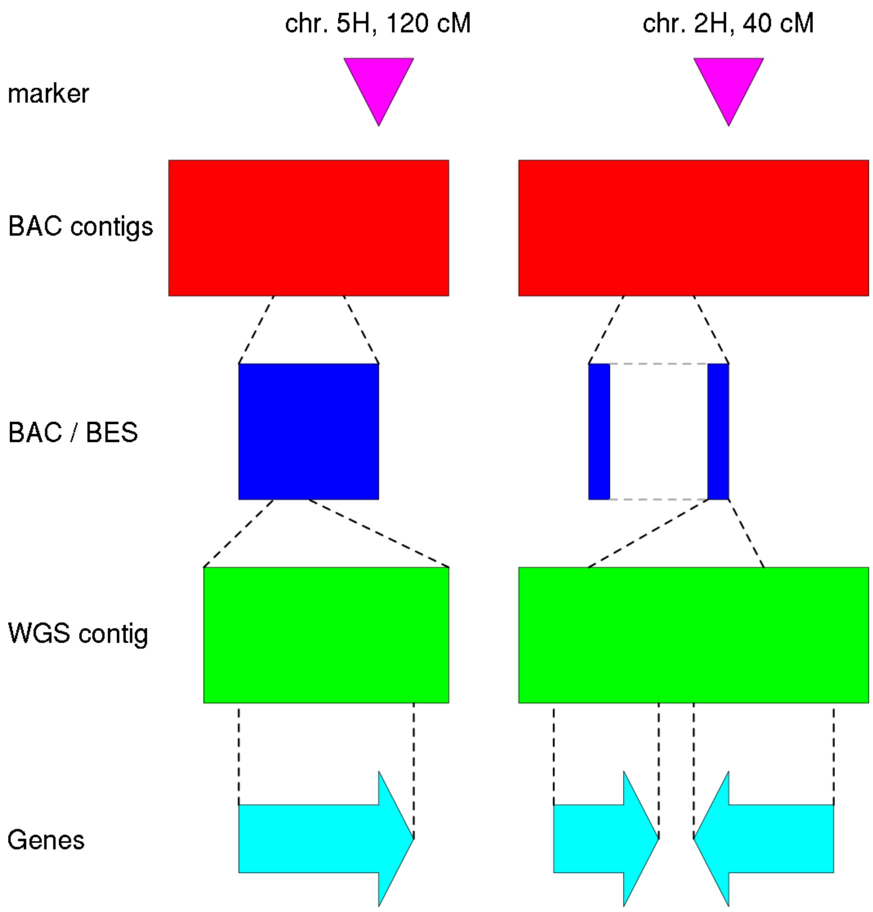


Figure 2.6: Schematic representation of the sequence-enriched physical and genetic framework of barley A genome-wide physical map was constructed by high information content fingerprinting (Luo et al., 2003). Survey sequence information was obtained by fully sequencing several thousand BACs as well as end-sequencing of ∼ 500,000 clones (BES, BAC end sequences). This sequence data was complemented by a whole genome shotgun assembly that provided a comprehensive representation of the gene space of barley. All sequence resources were used to integrate the physical map with various genetic maps. Thus, 80 % of the length of the physical map were assigned to chromosomal locations.

Morex was constructed by high-information content fingerprinting of ∼571,000 BAC clones. Automatic assembly and manual curation resulted in a final set of 9,265 BAC contigs (fingerprinted [FP] contigs) . This assembly spans 4.9 Gb, representing 96 % of the 5.1 Gb barley genome (Table 2.3). Survey sequence data derived from BAC end sequences, sequencing of complete BAC clones, whole genome shotgun contigs and shotgun sequencing of sorted chromosome arms was integrated with the physical map. A total of 5,341 gene-containing and 937 randomly selected BAC clones were sequenced on the Illumina HiSeq 2000 or 454 platforms. BAC end sequencing was performed for ∼300,000 clones. This provided 1.1 Gb of BAC-associated sequence directly integrated with the physical map. In addition to BAC sequence data, WGS assemblies were performed with sequence data from three barley cultivars which were sequenced to 30x – 50x genome coverage. These assemblies were highly fragmented and represented only about 40 % of the barley genome (Table 2.4).

Genetic markers were assigned to physical contigs by PCR-based marker screening, microarray hybridizations with EST-derived probes and Illumina GoldenGate assays for 3,072 SNPs. In addition to these experimental procedures, repeat-masked BAC end sequences, sequenced clones and WGS contigs were searched for marker sequences from ten different genetic maps, including two GBS maps. Overall, 4,556 BAC contigs amounting to 80 % of the physical map length could thus be assigned to genetic positions. Moreover, shotgun sequence data obtained from flow-sorted chromosome arms was used to assign an additional 1,881 contigs to chromosome arms.

The annotation of the transcribed portion of the genome was performed with the help of previously available full-length cDNA sequences of barley (Matsumoto et al., 2011) or by mapping RNA-seq from eight developmental stages against the WGS contigs. In total, a set 26,159 high-confidence gene models with homology to sorghum, rice, Brachypodium or Arabidopsis were defined. Another 53,220 gene models were supported only by RNA-seq reads, but not by gene family clustering. The vast majority (92 %) of high-confidence gene loci could be assigned to genetic positions or at least to chromosome arms.

Table 2.4: Features of the whole genome shotgun assembly of barley cv. 'Morex'

no. of contigs	2.7 million
cumulative length	1.9 Gb
mean contig length	700 bp
no. of contigs > 1 kb	376,261
length of contig > 1 kb	1.1 Gb
N50	1,425 bp

In summary, a multi-layered framework integrating various sources of physical and genetic mapping data with sequence information has been assembled to bring a large part of the barley gene space in a tentative linear order. While this resource does not constitute a draft genome, it has already been used to study the genetic diversity between a small panel of barley cultivars and one wild barley accession (IBSC, 2012), to study genome-wide patterns of copy number variation (Muñoz Amatriaín et al., 2013) and to design a whole exome capture assay (Mascher et al., 2013b). In the near future, it will serve as a valuable reference for further genetic research and breeding applications.

2.4.3 Genome zippers: anchoring by collinearity

Synteny (or more correctly, conserved synteny) is the collinear arrangement of orthologous genes in the genomes of two related species. It was first described for X-linked genes in mammalian species (Ohno, 1973) and has since been observed in other clades (see for example Trachtulec and Forejt (2001) and Tang et al. (2008)).

The genome zipper approach employs genetic maps or chromosomal survey sequence data together with synteny between an unsequenced genome and a sequenced relative to establish a virtual gene order. Exploiting the high degree of syntenic conservation between grasses (Moore et al., 1995), a genome zipper has been first created for barley chromosome 1H (Mayer et al., 2009). Flow cytometry was used to isolate chromosome-specific DNA which was subsequently sequenced on the 454 platform to ∼1x coverage. Evidence form cytology and genetic marker data confirmed that flow sorting achieves up to 95 % purity (i. e. ∼5 % of the sequence reads originate from other chromosomes). About 90 % of EST-based markers previously mapped to 1H could be found in the sequence reads. This dataset allowed Mayer et al. (2009) to estimate the gene content of chromosome 1H through similarity searches against the complete reference genomes of rice and sorghum. Gene models from syntenic regions of these genomes were associated with putatively orthologous EST markers on chromosome 1H. The order of genes in the reference genomes was then lifted to the orthologous genetic markers to create a virtual gene map of chromosome 1H. Overall, ESTs corresponding to almost 2,000 gene models in rice and sorghum could thus be brought into a tentative linear order.

The 1H genome zipper has been since been extended to all seven chromosomes of barley (Mayer et al., 2011). Whole chromosome arms were isolated by flow-sorting and sequenced on the 454 platform. In addition to the 1H genome zipper, genes were assigned to chromosome arm not only by bioinformatical comparison, but also through experimental procedures. Chromosomal DNA was hybridized to a microarray containing oligonucleotide probes for 25,000 to 32,000 genes to assign genes to chromosome arm. Furthermore, ∼23,500 non-redundant full-length cDNA sequence (Sato et al., 2009; Matsumoto et al., 2011) were integrated into this genome zipper. Overall, more than 21,000 barley genes could be assembled into a putative linear order. Prior to the publication of the barley physical and genetic framework, the barley genome zipper had been the most comprehensive genetic resource for any Triticeae species.

Analogous genome zipper have since been generated for several other grass species. Virtual gene maps have been created for the rye genome (Martis et al., 2013) and for several whole chromosomes or chromosome arms of hexaploid bread wheat. Up to now, genome zippers have been published for wheat chromosome arm 1AL (Lucas et al., 2013) as well as the whole chromosomes 3A (Akhunov et al., 2013), 3B (Shatalina et al., 2013), 4A (Hernandez et al., 2012), 5A (Vitulo et al., 2011) and all chromosomes of homeologous group 7 (7A, 7B, 7D) (Berkman et al., 2013). Synteny to rice, Brachypodium and sorghum was also used to genetically anchor the genome-wide physical map of the wheat D genome progenitor *Aegilops tauschii* (Luo et al., 2013) and the gene space assembly of bread wheat (Brenchley et al., 2012). Apart from the Triticeae, genome zippers have been created for perennial rye grass (*Lolium perenne*), an important forage grass, by interpolating gene models from barley into a framework genetic map of ryegrass (Pfeifer et al., 2013).

Genome zippers have proved to be a highly useful resource for genome-wide analyses and gene isolation. Map-based cloning projects can mine genome zippers for new genetic markers or candidate genes in a target region by exploiting the collinear gene order in fully sequenced reference genomes. For example, Mizuno et al. (2012) developed DNA markers for fine-mapping a flowering time gene in einkorn wheat (*Triticum monococcum*) using the barley genome zipper. Furthermore, the syntenic analysis of survey sequence data from supernumerary chromosomes (B chromosomes) of rye traced back the origin of these chromosomes to the standard nuclear genome and to the organellar genomes (Martis et al., 2012).

The genome zipper method of anchoring based on syntenic relationships has several drawbacks compared to ordering sequence contigs based on physical or genetic maps specifically constructed for the species of interest. Genome zippers either rely on flow sorted chromosomes or a genetic map as a skeleton that is populated with gene models whose position and order is inferred by synteny to a related species. However, this inference is confounded by breaks of synteny. Wicker et al. (2011) performed a comparative study of gene content in the group 1 chromosomes of barley and bread wheat (i. e. 1A, 1B, 1D and 1H) and found a frequent accumulation of non-syntenic genes which are in most cases non-functional pseudogenes. This may upwardly bias estimates of gene number inferred from low-pass sequence data and the analysis of conserved synteny.

The sequence information that is incorporated into genome zippers does not provide a suitable reference for mapping NGS reads collected in a resequencing project. Low-coverage survey sequencing (∼1x) with 454 reads does not allow an assembly and individual reads are too short to function as a reference. Neither can EST- or cDNA-based assemblies serve as a good reference for mapping genomic NGS reads, as they do not contain intronic sequence. If the sequence resource subjected to genome zipping is a more comprehensive genomic sequence assembly, the number of anchored sequence contigs is bounded by the number of syntenic genes. For instance, Akhunov et al. (2013) assembled 454 reads obtained from flow-sorted wheat chromosome 3A (corresponding to ∼20x coverage) into ∼240,000 contigs larger than 500 bp. However, syntenic anchoring tagged only ∼3,600 contigs. Even though this assembly represents 40 % of the physical length of the chromosome and a substantially higher proportion of the low-copy regions on 3A, only a small fraction of reads mapped to this assembly can be associated with a chromosomal locations.

Flow-sorting itself is beset with some technical difficulties. The purity of isolated chromosomes or chromosome arm ranges between 80 and 95 %. Contamination from other chromosomes may affect the analysis of syntenic conservation (Akhunov et al., 2013). For sorting chromosomes of wheat and barley, special cytogenetic stocks of wheat are used that carry unusual chromosome configurations where one or several chromosome arms are missing or substituted by chromosomal segments from alien species. In certain cases, chromosome arms may be inaccessible to flow-sorting. In the case of chromosome 1H of barley, it was not possible to obtain purified DNA from the long and short arms separately as 1HL addition lines are sterile (Islam and Shepherd, 2000). Moreover, these cytogenetic stocks and the varieties used in their production have not been characterized comprehensively and may confront their users with unexpected surprises. For example, the ditelosomic line used by Wicker et al. (2011) contained a heavily rearranged chromosome arm 1DS.

2.4.4 Direct anchoring of sequence contigs

Sequence contigs can be anchored to a genetic map without a genome-wide physical map as an intermediary layer. In this case, no short-range connectivity information contained in physical contigs is used and it is not possible to place sequence contigs without marker information to the same genetic bin as it would be possible when using a physical map where marker associations position all sequence contigs belonging to the same physical contig. Consequently, a copious amount of genetic markers is necessary to anchor a WGS assembly that is fragmented into hundred thousands of kilobase-sized contigs. The required marker density can only be supplied by NGS technology.

Several methods for high-throughput genotyping of genetic mapping populations using next generation sequencing technology have been developed. Genotyping by shallow survey sequencing (0.05–0.1x) in the model species rice has been shown to yield genetic maps of unprecedented density (Huang et al., 2009). However, the high resolution of recombination breakpoints (∼40 kb) was provided by inferring marker order from a high-quality reference sequence. This approach cannot be applied to species with genomes of draft or even pre-draft quality as sequence contigs are not organized in pseudomolecules representing the linear chromosomes.

The question of how several millions of markers provided by NGS technology may be used to bring contigs into a linear order has only tentatively been raised. Andolfatto et al. (2011) used restriction-enzyme digestion with a frequent cutter and subsequent multiplexed sequencing of a population of 94 individuals to assign 8 Mb of unassembled contigs to linkage groups. Similarly, a reduced representation genotyping-by-sequencing method has been instrumental for anchoring the barley physical map to a genetic map (Poland et al., 2012b; IBSC, 2012).

Jia et al. (2013) performed reduced representation sequencing of 490 individuals from an F_2 population to anchor a WGS sequence assembly of the wheat D genome progenitor *Aegilops tauschii*. With the resulting ∼150,000 markers, 13,688 sequence scaffolds could be anchored. As this assembly incorporated paired-end and mate-pair data from 45 libraries, very long scaffolds could be generated (N50: 58 kb) and the cumulative length of all anchored contigs was 1.3 Gb (∼30 % of the total assembly length). Additional contigs were anchored with the genome zipper method.

Saintenac et al. (2013) aligned the sequence tags obtained by genotyping-by-sequencing of 178 DH lines of hexaploid wheat to a shotgun sequence assembly of wheat chromosome 3A. Only 19 % of all tags genetically mapped to 3A could be aligned to a sequence contig and only 4 % all contigs received a marker. This leads to the conclusion that even though GBS interrogates an order of magnitude more markers than a medium-sized SNP array, the number of markers is still insufficient to genetically anchor a highly fragmented shotgun assembly.

Genetic anchoring of sequence contigs is not limited to classical genetic mapping with bi-parental populations. Genovese et al. (2013) used an association mapping approach to assign 70 previously unanchored scaffolds of the human genome to their correct chromosomal locations. Association mapping exploits linkage disequilibrium (LD), the non-random association of genetic markers at physically close loci. In bi-parental mapping population, linkage disequilibrium can be entirely explained by a small number of recombination events and its extent is the length of a chromosome. In a large panel of unrelated individuals, linkage disequilibrium decays over much smaller distances. Recombination has reshuffled ancestral chromosomes during hundreds to thousands of generations. However, LD may be created by factors other than meiotic recombination and may extend over much larger ranges. Long-distance LD – even between loci on different chromosomes – may, for instance, be a consequence of population history or natural selection (see Pritchard and Przeworski (2001) for a review). Genovese et al. (2013) exploited LD patterns peculiar to admixed populations. When reproductively isolated populations (such as Africans and Europeans) remix as in African Americans, the genomes of admixed individuals are mosaics of both parental haplotypes with a limited number of recombination breakpoints, somewhat similar to the chromosomes in plant mapping populations. In other words, the genomes of admixed descendants exhibit medium-range linkage disequilibrium which provides better mapping resolution than classical family-based genetic mapping, but avoids the confounding effects of a long and complex population history. Genovese et al. (2013) used whole genome shotgun sequence data from The 1000 Genomes Project Consortium (2012) to find anchored SNPs in high linkage disequilibrium with 139 SNPs on unanchored scaffolds and were thus able to position 4 Mb of euchromatic sequence with a resolution of 10 to 100 kb.

Anchoring by admixture mapping – or more generally association mapping – is unlikely to be successful on a genome-wide scale, i. e. to serve as the primary tool in anchoring a sequence assembly to a framework genetic map. The decay of LD is too fast in natural populations of many species for example barley and maize (Morrell et al., 2005; Remington et al., 2001). Furthermore, it is challenging to quantify the impact of population structure on patterns of LD when the demographic history of a species is unknown. Even in genome-wide association studies, which aim at the identification of genes responsible for a single trait, complex population structure can confound the outcome to an extent that only the availability of new data sets and analytical methods can identify the source of error years after the original study (Larsson et al., 2013).

Thus, it is likely that genetic mapping in bi-parental populations will continue to be the method of choice for anchoring contig assemblies for next years. To the best of our knowledge, genotyping by whole genome shotgun sequencing for genetic mapping in bi-parental populations has not been used as a tool for the *de novo* development of linearly ordered draft genome assemblies.

3 The POPSEQ method

The POPSEQ method combines (i) an experimental layout that is routinely implemented in genetic mapping experiments with bi-parental populations, (ii) standard next generation sequencing protocols, (iii) established computational pipelines for *de novo* assembly, SNP genotyping and genetic map construction, as well as (iv) a novel algorithm for integrating the resulting datasets into a common structure (Figure 3.1).

In this chapter, we will give an in-depth description of POPSEQ using barley as an example. We briefly describe the software that is integrated into a pipeline to perform the tasks of converting NGS raw data into genotypic data and of genetic map construction. We then portray the work flow for anchoring the sequence assembly of barley with the help of whole genome shotgun sequence data from two segregating populations.

We note that POPSEQ requires also a WGS assembly as a basis for read mapping. We have not performed sequence assembly ourselves, but used the WGS of barley cultivar Morex that is available from IBSC (2012). Likewise, to improve the utility of a POPSEQ assembly, gene models should be defined on the assembly. This task has also already been performed by IBSC (2012) using established algorithms (see Section 2.4.2).

3.1 Software used in the POPSEQ pipeline

The POPSEQ pipeline incorporates established tools to perform short read alignment, SNP and genotype calling as well as genetic map construction. The BWA/SAMtools (Li and Durbin, 2009; Li, 2011a) pipeline was used to map Illumina sequence reads to the assembly of barley cultivars 'Morex'. There are a number of alternative SNP calling pipeline, such as BWA/GATK (DePristo et al., 2011) or the SOAP mapper and SNP caller (Li et al., 2009b). We chose BWA/SAMtools as it has been previously employed for SNP calling in barley (IBSC, 2012) and has been an easy-to-use, multi-purpose tool that worked reliably in our hands. We have not been able to perform a comparison of different mappers and SNP callers on the POPSEQ data. As read alignment and SNP calling took two weeks using ~30 CPU cores for one population, a thorough benchmark of different pipelines would have explored the limits of our compute infrastructure. Systematic comparisons of different pipelines by other groups have yielded equivocal results. Lam et al. (2012b) benchmarked GATK and SAMtools for one human sample sequenced to high (48x) coverage.

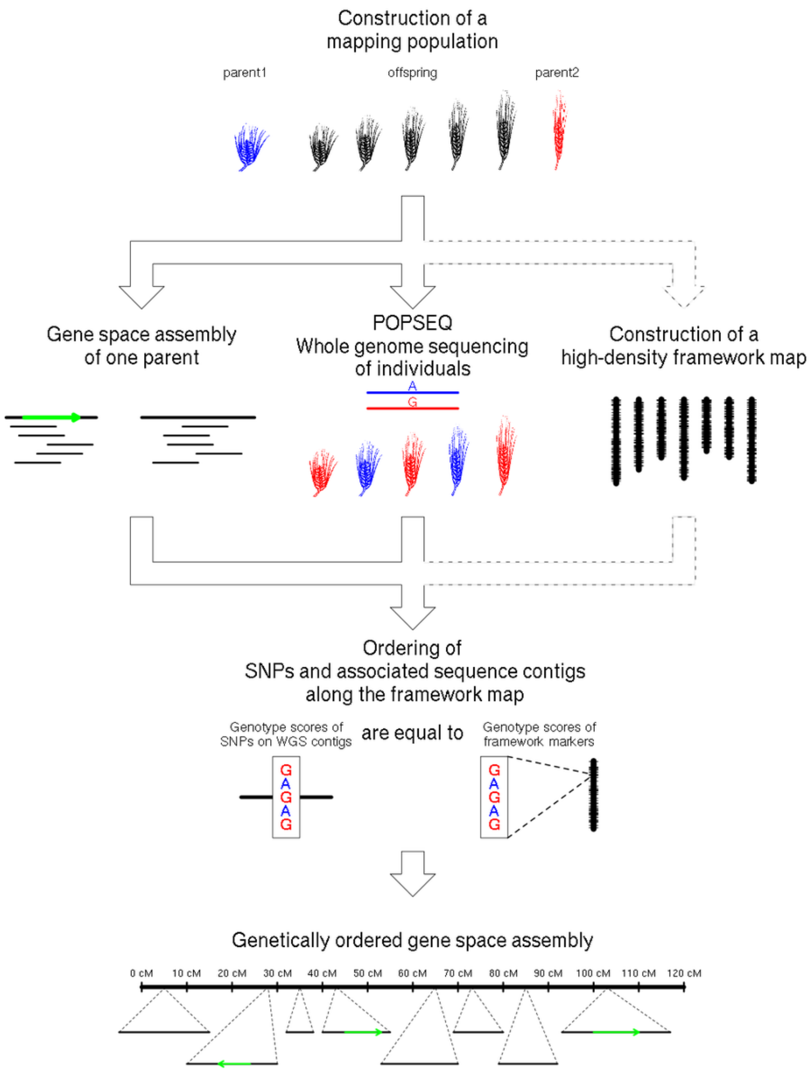

Construction of a mapping population
parent1
offspring
parent2
Gene space assembly of one parent
POPSEQ Whole genome sequencing of individuals
A
G
Construction of a high-density framework map
Ordering of SNPs and associated sequence contigs along the framework map
Genotype scores of SNPs on WGS contigs
are equal to
Genotype scores of framework markers
G A G A G
G A G A G
Genetically ordered gene space assembly
0 cM
10 cM
20 cM
30 cM
40 cM
50 cM
60 cM
70 cM
80 cM
90 cM
100 cM
110 cM
120 cM

Figure 3.1: Overview of POPSEQ. **(a)** A segregating population (80 – 100 individuals) is constructed from a biparental cross. **(b)** A whole genome shotgun assembly is generated of one parent and used to construct a gene-space assembly. On this assembly, gene models (green arrows) are defined using RNA-seq. In parallel, POPSEQ, and if necessary, genotyping-by-sequencing (GBS), is performed on the population and a medium density framework genetic map encompassing thousands of loci is calculated. **(c)** SNPs detected and typed by POPSEQ along with associated WGS contigs are integrated into the framework map through nearest-neighbor search. **(d)** The result of POPSEQ is a sequence assembly in linear order that contains comprehensive information of the gene space. It can be enhanced by conducting POPSEQ on additional populations. This figure was taken from Mascher et al. (2013a).

They found $> 98\%$ concordance between SAMtools and GATK and a sensitivity of $98 - 99\%$ when compared to the results of a high-density SNP array. By contrast, O'Rawe et al. (2013) reported a concordance of only 57 % when five SNP calling pipelines were applied to exome sequence data from 15 individuals. Notably, the human 1000 Genomes Consortium which oversees the most comprehensive set of low-coverage resequencing data collected in any species does not enforce the use of a single variant caller, but each research group may use their preferred variant calling pipeline (The 1000 Genomes Project Consortium, 2012). To the best of our knowledge, there is currently no single best practice for SNP and genotype calling from NGS data. Moreover, the exact choice of tools for read alignment and SNP calling is not of consequence to illustrate the principle of POPSEQ. The integration of SNPs into a framework map requires only SNP genotypes recorded in a generic marker-by-individual matrix. The output of any pipeline can be easily converted to this format.

In contrast to NGS variant calling, genetic map construction is less polyphonic. Several studies have shown the results of different mapping software to be largely equivalent with the remaining inconsistencies being caused by uncertainties present in the raw data (Wu et al., 2008; Close et al., 2009; Bowers et al., 2012; Lorieux, 2012; Truco et al., 2013; Nagy et al., 2012). The various tools differ mostly in their algorithmic approaches and running times. We chose MSTMAP for its excellent performance with a large number (> 1000) of markers and its ease of use in a UNIX environment.

3.1.1 BWA

BWA (Burrow-Wheeler aligner) is arguably the most commonly used tool to map reads from whole genome shotgun sequencing experiments to a reference sequence. BWA creates an index of a reference genome (or a set of WGS contigs) using the block sorting algorithm of Burrows and Wheeler (1994). The memory-efficient construction of the Burrows-Wheeler transform is performed using the algorithm of Hon et al. (2007). It uses the backwards search algorithm of Ferragina and Manzini (2000) to enumerate all alignments of a read to the reference with less than a specified edit distance. BWA allows short indels in reads, a feature that had been absent from initial hash- or BWT-based approaches (Li et al., 2008; Langmead et al., 2009). BWA was specifically designed for mapping paired-end Illumina reads against complex genomes. It is aware of repeats and calculates a mapping score that quantifies the uniqueness of an alignment and can be used by downstream tools to discard alignments to repetitive regions.

BWA takes reads as input (gzipped) FASTQ text files where each read is represented by four lines. Two of them contain the nucleotide sequence and the associated quality values and two are header lines whose content depends on the sequencing instrument. Read alignments are reported in the SAM format (Li et al., 2009a). BWA includes a read trimmer that cuts off the ends of sequence reads when quality drops below a user-defined threshold.

3.1.2 SAMtools

We used SAMtools to process BAM files and to perform SNP and genotype calling. On the one hand, SAMtools is a suite of small, simple utility programs for handling SAM and BAM files (Li et al., 2009a). On the other hand, it provides a comprehensive statistical framework for variant and genotype calling from NGS data (Li, 2011a). Commonly used utility functions of SAMtools are:

view
This commands provides basic viewing and filtering functionality for SAM and BAM files. It also performs the compression of SAM to BAM files and the decompression of BAM files.

sort
This command takes a BAM file as input and sorts the entries according to their mapping positions on a reference. BAM files have to be sorted prior to indexing.

index
Indexing of BAM files allows the fast retrieval of reads from specified genomic regions. SAMtools implements the hierarchical binning scheme of UCSC genome browser (Kent et al., 2002) together with linear indexing. The number of bins per chromosome (or contig) is fixed and is determined beforehand according to the size of the chromosomes / contigs. Overlap searches between features require only the comparison of identical and adjacent bins. A similar indexing method has been developed for generic TAB-separated files (e.g. GFF or VCF files) by Li (2011b). All reads from a specified interval can be retrieved from an indexed BAM file using the **view** command.

rmdup
This commands removes duplicate read pairs with identical mapping locations. These pairs originate most likely not from independent fragments, but are artifacts from PCR amplification.

Other SAMtools commands perform merging, concatenation and header manipulations of BAM files.

Besides being the swiss army knife of NGS data handling, SAMtools includes sophisticated algorithms for variant and genotype calling (Li, 2011a) implemented in the commands **samtools mpileup** and **bcftools**. SAMtools detects and types SNPs and short insertion and deletion polymorphisms (indels). Both variant sites and genotype calls are accompanied by a Phred-scaled likelihood which measures the probability that the variant is a false-positive or the genotype call is incorrect. SAMtools performs multi-sample variant calling, i. e. the read alignments of several samples are inspected simultaneously. **samtools mpileup** aggregates the mapped reads at each – not necessarily polymorphic – site and collects additional information such as the read depth, the distance to the ends of reads and the average base quality score. This output is piped to **bcftools**, which performs the actual variant calling and genotype likelihood estimation.

The SAMtools variant calling pipelines takes sorted and indexed read alignments in BAM format and a reference FASTA file as input. Moreover, variant calling can be performed only on a subset of the genome by specifying an interval file in BED format (Kent et al., 2002). This functionality can be used to parallelize the single-threaded SAMtools variant calling pipeline by partitioning the genome into bins and applying SAMtools to each bin in parallel. The output of the SAMtools variant calling pipeline is a tab-separated text file in the variant call format (VCF) developed by Danecek et al. (2011). VCF is the current standard format for reporting variant calls. Like the SAM/BAM format for read alignment, VCF was developed in frame of the human 1000 genomes project to facilitate the data exchange between different laboratories (Danecek et al., 2011). For fast feature retrieval, VCF files can be indexed with Tabix (Li, 2011b).

3.1.3 MSTMAP

MSTMAP (Wu et al., 2008) is a program for genetic map construction that builds on graph-theoretical methods.

Clustering of markers into linkage groups is performed by recording the distance between any two markers into a complete graph, where nodes are markers and an edge between two markers is weighted by their Hamming distance. Edges between markers from different linkage groups will have a high weight and are pruned from the graph according to a user-defined threshold.

Subsequently, marker order is established in each linkage group by finding a path with minimal weight that visits each node, i. e. they find a minimum weight traveling salesman path. The weight function is semi-linear because when two markers A, B are enclosed in an interval C, D, the probability of a recombination event between A and B is smaller or equal than between C and D, which is a direct consequence of the chromosomal theory of inheritance. As the weight function is semi-linear, a traveling salesman man path can be rapidly calculated as a minimum spanning tree (Wu et al., 2008). This derivation assumes error-free and complete genotypic data. MSTMAP includes several heuristics to deal with noisy or incomplete data. “Bad” markers with an excessive proportion of genotyping errors can be detected and removed and missing data be imputed with an EM algorithm if so desired. MSTMAP expects input, and writes output, in custom textual formats.

MSTMAP is particularly suited to map calculation from dense genotypic datasets (1,000s to 10,000s of markers). It is considerably faster than other algorithms (Wu et al., 2008), but yielded maps similar to the output of other programs when applied to real data (Truco et al., 2013; Nagy et al., 2012).

3.2 Barley populations and sequence data

We generated whole genome shotgun sequence data from members of two experimental populations (Table 3.1). One was a population of recombinant inbred lines (RILs) from a cross between barley cultivars Morex and Barke (MxB) (Comadran et al., 2012). The entire population comprises 2,407 F_8 RILs that were generated by single-seed descent from independent F_2 individuals.

The second population consisted of 82 doubled haploid (DH) lines from the Oregon Wolfe Barleys (Costa et al., 2001). This extremely diverse population had been generated from a cross between two specialized morphological marker stock lines. One parent (OWBDom) is homozygous for several dominant mutant genes, while the other parent (OWBRec) is homozygous for several recessive mutant genes (Wolfe et al., 1990). Doubled haploid plants had been generated using the bulbosum method (Costa et al., 2001). Both populations had been used earlier for genetic map construction (Costa et al., 2001; Comadran et al., 2012) and OWB had been characterized phenotypically (Cistue et al., 2011).

DNA sequencing was performed on the Illumina HiSeq 2000 sequencing platform at the DOE Joint Genome Institute (JGI). DNA from individual plants was fragmented and barcoded according to standard protocols. Eight samples were sequenced per lane in paired-end mode (2 × 100 bp) on an Illumina HiSeq 2000, yielding ca. 1x coverage per line.

In addition to whole genome shotgun sequencing, we sequenced the two parents and 92 F_8 RILs of the Morex × Barke population (including the 90 individuals that were sequenced by WGS for POPSEQ) by using a reduced-representation approach (Poland et al., 2012b). Prior to sequencing, DNA was co-digested with a rare-cutting, methylation-sensitive enzyme (*Pst*I) and a common-cutting enzyme (*Msp*I). Restriction fragments with two different restriction sites were selected by PCR and sequenced on one lane of the Illumina HiSeq 2000 instrument at IPK Gatersleben. The size of the regions on the barley genome targeted by this approach is about 10 Mb (Figure 3.2).

3.3 From FASTQ to marker-by-genotype matrix

Sequencing reads were mapped against the Morex WGS assembly (IBSC, 2012) with BWA version 0.6.2 (Li and Durbin, 2009). The BWA command **aln** was

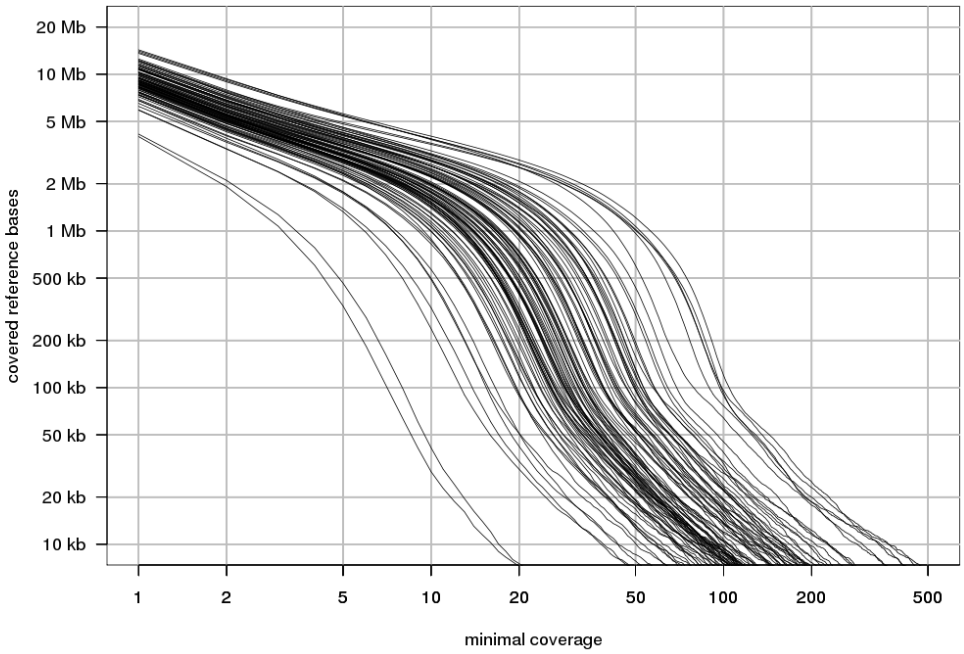


Figure 3.2: The size of the genomic intervals on the Morex WGS assembly with a specific minimal coverage is depicted. Each line corresponds to one of the 94 samples genotyped by GBS (92 RILs and the parents Morex and Barke). In all except two lines, more than 5 Mb of sequence were covered by at least one read. In all lines, less than 200 kb had more than 100-fold coverage.

called with the parameter "`-q 15`" for quality trimming, otherwise default parameters were used. After removing duplicate reads with **samtools rmdup**, variant positions and genotypes of individuals at variant positions were called with the **samtools mpileup** / **bcftools** pipeline version 0.1.18 (Li, 2011a) with default parameters. Additionally, the parameter "`-D`" was used for **samtools mpileup** to record per-sample read depth at variant positions.

The resulting VCF file was filtered with a custom AWK script (Mascher et al., 2013c). This script removed SNPs with a SAMtools quality score below 40 and further filtered the SAMtools genotype calls: a homozygous genotype call was retained if there was at least one read supporting it and its SAMtools genotype quality was at least 3. In the MxB data, a heterozygous call was retained if there were at least three supporting reads and its score was at least 5. In the OWB doubled haploid population, heterozygous calls were always discarded. Genotype calls not matching the specified criteria were set to missing. A variant position was removed if (i) more than 10 % of all samples were called heterozygous, (ii) there were more than 80 % missing data or (iii) the

Table 3.1: Sequence data generated in this study.

	MxB WGS	OWB WGS	MxB GBS
Population	Morex × Barke RIL F8	Oregon Wolfe Barleys DH	Morex × Barke RIL F8
Sequencing technology	WGS; HiSeq 2000	WGS; HiSeq 2000	GBS; HiSeq 2000
No. of sequencing lanes	12	12	1
No. of sequenced individuals	90 (+parents)	82 (+parents)	92 (+parents)
Coverage per sample	~1x	~1x	~1x (10 Mb represented)
No. of detected SNPs	5.1 M	6.5 M	21,397
Average no. of present genotype calls per marker	33	31	58

minor allele frequency (in the non-missing data) was smaller than 5 %. These parameters were chosen specifically for a population of homozygous individuals. For double haploid lines, no genuine heterozygous SNPs are expected. For F_8 RILs, theory predicts a residual heterozygosity of ~1 %. We expect a comparable rate of heterozygous calls erroneously called homozygous when only one allele was sampled and accepted this error rate rather than using a higher stringency, which would have meant that much fewer loci could have been genotyped. For an unselected biparental population, we expect a 50 % minor allele frequency. However, to account for missing data and stochastic sampling, the minimum minor allele frequency was set to 5 % and up to 10 % heterozygous calls were allowed. This mapping and variant calling pipeline is versatile and we applied it in adapted form to the analysis of WGS, exome capture and GBS data (Mascher et al., 2013a,b,c).

3.4 Framework maps

The large amount of missing data (Figure 3.3) in the WGS SNPs precluded the use of these markers as input for *de novo* genetic linkage map construction. In order to assign chromosomal locations to these SNPs, we placed them into three framework genetic maps that had been constructed from high-quality (i. e. near complete) genotypic data. We used as frameworks for POPSEQ two genetic maps of the Morex × Barke population – a published one and one computed by ourselves – as well as a published map of the OWB population.

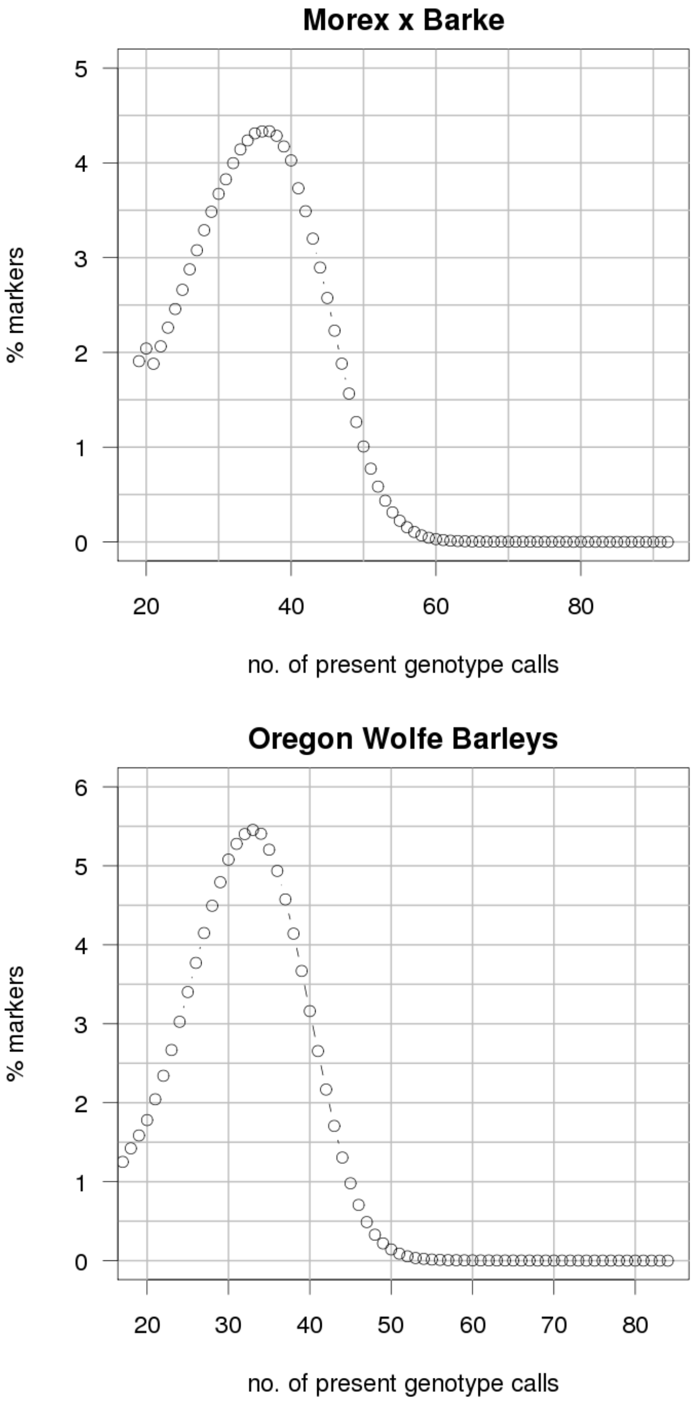


Figure 3.3: Distribution of the number of successful genotype calls at variant positions detected in the WGS data of the Morex × Barke and OWB populations. Variant positions with more than 80 % missing data were not used for downstream analysis.

3.4.1 Morex × Barke iSelect map

A genetic map of the Morex × Barke population had been developed by array-based genotyping using an Illumina 9K iSelect chip (Comadran et al., 2012). SNPs had been discovered by transcriptome sequencing of 10 barley varieties. After several filtering and quality control steps, 5,010 high-quality SNPs were selected for inclusion on the array. The addition of 2,854 SNPs from previous assays (Close et al., 2009) or other resequencing studies gave a total of 7,864 SNP assays on the iSelect chip. Of these SNPs, 3,973 were polymorphic in the Morex × Barke population and were typed on 360 individuals of the population o construct a genetic linkage map (Comadran et al., 2012), which we will call the "iSelect map" for short.

3.4.2 Morex × Barke GBS map

We used GBS data from the Morex × Barke population (Table 3.1) to construct a new linkage map of this population. Compared to array-based genotyping, genotyping-by-sequencing has lower per-sample genotyping costs and does not require any prior knowledge of polymorphisms between the parents of the mapping population. Instead, marker discovery and scoring occur simultaneously, making GBS suitable for species without any, or having only poorly developed, genomic resources.

- Deconvolution and adapter trimming
  Reads were deconvoluted with a custom AWK script (Mascher et al., 2013c) that performs an exact string matching of the read starts to the index sequences. Custom demultiplexing was necessary as the GBS protocol of Poland et al. (2012b) does not use standard Illumina barcodes with a dedicated index read. Instead, index sequences are found at the 5′ end of the first read. Adapter sequences were removed with cutadapt version 1.1.
  (`http://code.google.com/p/cutadapt`). Adapter sequences ligated to the 3′ ends of fragments were present in about 30 % of the reads as the protocol of Poland et al. (2012b) often produces fragments shorter than 100 bp. Trimmed reads shorter than 30 bp were discarded.

- Mapping and SNP calling
  Read mapping, SNP and genotype calling and filtering were performed essentially as described above for the WGS data. Since only single ends were used, BWA (commands **aln** and **samse**) was used for alignment. Additionally, only SNPs meeting the following criteria were considered for genetic map construction: (i) less than 10 % missing data; (ii) no more than 10 % heterozygous genotypes; (iii) $\frac{|A-B|}{A+B} < 0.7$, where A and B denote the counts of the parental alleles. In the absence of heterozygous

calls, criterium (iii) is equivalent to a minimum minor allele frequency of 17.6 %. A total of 4,058 SNPs passed these filters.

- Genetic map construction
  Genetic map construction was performed with MSTMAP using the following parameters:
  ```
  population_type DH,
  distance_function kosambi,
  cut_off_p_value 0.00001,
  no_map_dist 20,
  no_map_size 2,
  missing_threshold 0.8,
  estimation_before_clustering no,
  detect_bad_data yes,
  objective_function COUNT.
  ```

  The Kosambi map distance function (Kosambi, 1943) was used. The population type was set to DH (doubled haploid) as was recommended by the manual for advanced RIL populations. The `missing_threshold` filters all markers with more than 20 % missing or heterozygous calls and is less stringent than the missing data filter we applied prior the map construction. The `cut_off_p_value` specifies a threshold for clustering markers into linkage groups. The `no_map` parameters are used to discard small sets of isolated markers (in this case, a set of up to 2 markers separated by more than 20 cM from the next marker).

The resulting map contained seven linkage groups with more than one marker. Two markers went into a linkage group of their own and were discarded. According to the obtained orders, orientations and distances between markers, the linkage groups corresponded to the seven barley chromosomes and were highly collinear with the IBSC reference map (Figure 3.4). The relationship between genetic positions in the new map and the iSelect map was obtained through LOWESS regression (R function `loess`, smoother span 0.3) similar to the approach of Duffy (2006). Interpolation into the iSelect map of WGS SNP positions integrated to the GBS framework was performed with the `loess` model (R function `predict`).

3.4.3 OWB GBS map

A bin map had been constructed previously from GBS data of 82 OWB doubled haploid (DH) lines by Poland et al. (2012b). SNP calling in this dataset had been performed with the TASSEL pipeline (Elshire et al., 2011; Bradbury et al., 2007). The output of this pipeline are 64 bp long sequence tags which harbor SNPs. Unlike SNPs called by the BWA/SAMtools pipeline, the TASSEL tags were not directly positioned on the Morex WGS assembly. To

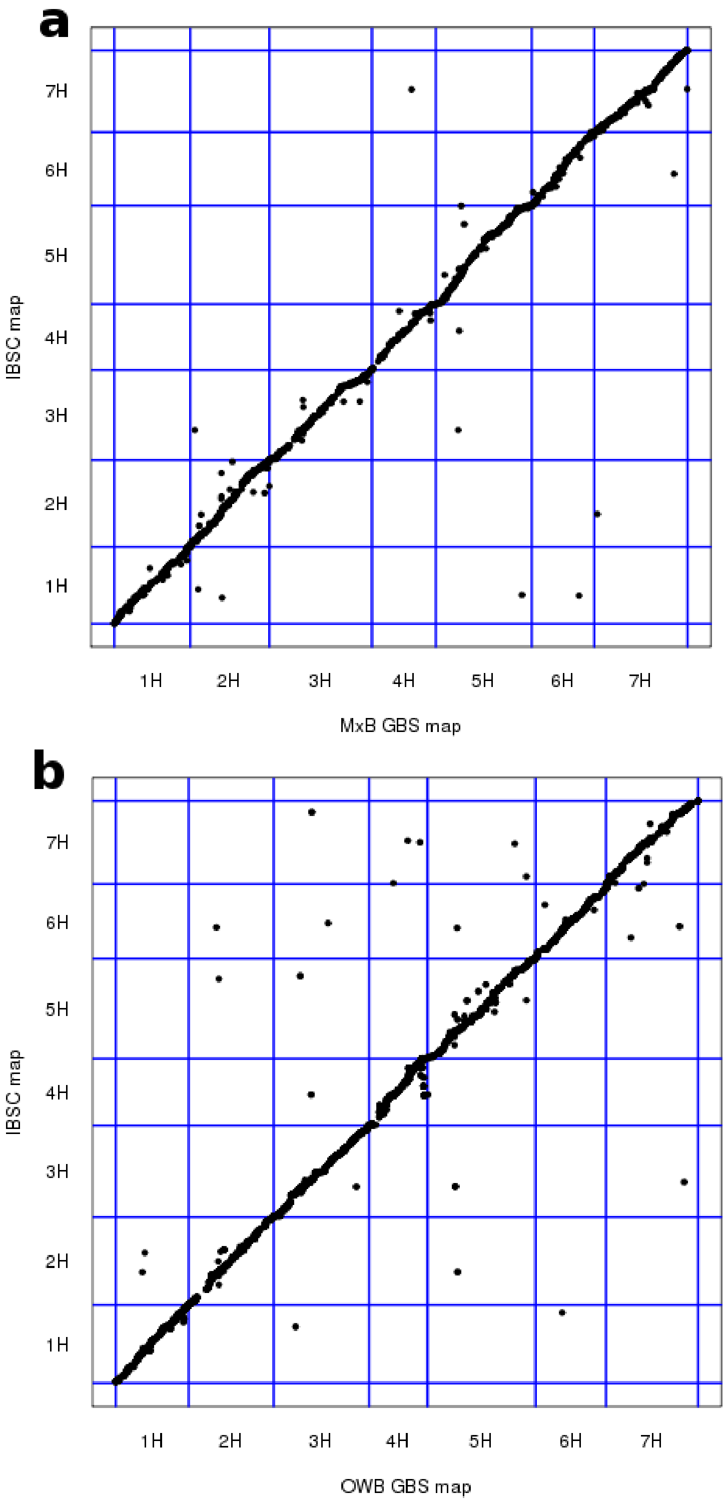


Figure 3.4: Collinearity between the IBSC map and the GBS maps of **(a)** the Morex × Barke and **(b)** OWB population.

position the tag sequences on the Morex reference assembly, we aligned them against the contigs with **bwa aln** and **bwa samse** (Li and Durbin, 2009). Only alignments with the best possible mapping score of 37 were considered in order to avoid erroneous SNP placements caused by paralogous sequences. The genetic position of WGS contigs in the IBSC map and in the OWB GBS map as inferred from the alignments were mostly collinear (Figure 3.4).

SNPs with missing data for the parents or more than 10 % missing data on the DH lines were not considered as framework markers. Interpolation of framework marker positions into the iSelect map was performed as described for the MxB GBS map.

3.5 Mapping SNPs and WGS contigs to the framework map

We assigned genetic positions to WGS SNPs, i. e. SNP markers that were detected in the whole genome shotgun data of the two populations. SNPs were placed according to the positions of their nearest neighbors in the set of framework markers. We used the Hamming distance, i. e. the number of non-identical genotype calls, as a measure of distance. A Hamming distance of zero would indicate that there are no detectable recombination events between a WGS SNP and a framework marker. There may, however, be undetectable crossovers that are hidden because of missing data. Similarly, even in the absence of missing data, crossovers would be possible between WGS SNPs and framework markers as the spacing of markers in the framework map may are not be dense enough to capture every recombination event in the population.

The nearest neighbors were searched for in the set of

(i) 1,723 non-redundant iSelect SNPs for the MxB data,

(ii) 4,056 GBS SNPs used for construction of the MxB GBS map for the MxB data,

(iii) 4,632 non-redundant OWB GBS SNPs for the OWB data.

A SNP was considered redundant if there was another SNP with the same genotype (on the non-missing data) and the same genetic position. SNPs called from WGS data were used to anchor WGS contigs if they were scored unequivocally on at least 20 % of the individuals in the population. For distance calculations, SNPs were numerically represented as vectors with values 0, 1, 2 or 3, where 0 denotes the 'Morex' allele, 2 denotes the 'non-Morex' allele, '1' denotes a heterozygous call and 3 is used to mark missing data. The algorithm to position SNPs relative to a framework map is given in formal notation as Algorithm 1.

We searched for the nearest neighbors of each WGS SNP in the set of framework SNPs (Figure 3.5). For this purpose, we computed the Hamming distance

Algorithm 1 Placing of WGS SNPs relative to a framework map. Parameters that can be varied (maximal Hamming distance for nearest neighbors, tolerance for nearest neighbors from other chromosomes, maximal MAD) or depend on the species (number of chromosomes/linkage groups) are colored blue.

Input: M genotype vectors of WGS SNPs

F genotype vectors of framework SNPs

P framework marker positions (chromosome and genetic position in cM), $\{ (\mathrm{chr}_f, \mathrm{cM}_f) : f \in F \}$

Output: R positions of WGS SNPs, $\{ r_m := (\mathrm{chr}_m, \mathrm{cM}_m) : m \in M \}$

```
for all $m \in M$ do
    $D_m \leftarrow \{ d_{m,f} := \text{hamming\_distance}(m, f) : f \in F \}$
    $F_m \leftarrow \{ f \in F : d_{m,f} = \min(D_m) \text{ and } d_{m,f} \leq 2 \}$
    $T \leftarrow \{ t_i := |\{ f_m \in F_m : \mathrm{chr}_{f_m} = i \}|, i = 1, \dots, 7 \}$
    if there exists $t_c \in T$ such that $t_i \geq 0.8 \cdot \sum_1^7 t_i$ then
        $\mathrm{chr}_m \leftarrow c$
        $C \leftarrow \{ \mathrm{cM}_{f_m} : f_m \in F_m \text{ and } \mathrm{chr}_{f_m} = \mathrm{chr}_m \}$
        $\mathrm{cM}_m \leftarrow \mathrm{median}(C)$
        $\mathrm{MAD}_m \leftarrow \mathrm{MAD}(C)$
        if $\mathrm{MAD}_m < 5$ then
            $r_m \leftarrow (\mathrm{chr}_m, \mathrm{cM}_m)$
        else
            $r_m \leftarrow (\mathrm{NA}, \mathrm{NA})$
        end if
    else
        $r_m \leftarrow (\mathrm{NA}, \mathrm{NA})$
    end if
end for
return $R$
```

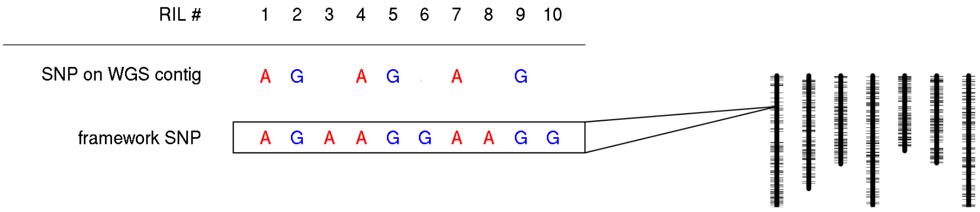


Figure 3.5: Schematic representation of nearest neighbor search. We searched for the nearest neighboring genotype vector in the set of framework markers in order to find the correct place where to put a WGS SNP into the framework map. The chromosomal location of all framework markers is known. If the genotype calls agree in the individuals of the populations, SNPs on WGS contigs and framework SNPs are genetically close to each other, so that we place SNPs and contigs to the same chromosomal position or bin in the framework map. Missing data is represented in the figure by omitted genotypes in the WGS SNP.

(i. e. the number of non-identical, non-missing genotypes) between a WGS SNP and all framework markers. This calculation was performed with a custom C program whose source code can be retrieved from `ftp://ftp.ipk-gatersleben.de/barley-popseq`. All further analysis steps involving the filtering and aggregation of positional information were performed in R (`http://www.R-project.org`). We set the following requirements to ensure the consistency of genetic positions if a WGS SNP had more than one nearest neighbor.

(i) The Hamming distance between a WGS SNP and its nearest neighbor is not larger than 2.

(ii) At least 80 % of all nearest SNPs lie on the same chromosome.

(iii) The median absolute deviation (MAD) of the cM positions of framework markers on the chromosome with most markers is below a given threshold. This threshold was 5 cM for the OWB map and the Morex × Barke iSelect framework. As the genetic length of a barley chromosome is between 100 and 200 cM, this threshold corresponds to a ∼5 % tolerance range, within which the genetic positions of SNPs at identical physical positions may vary as a consequence of sequencing or genotype calling errors. As we used the population type "DH" for the MxB RILs (as required by MSTMAP for advanced RILs) the MxB GBS map overestimated the map length by a factor of ∼3 and we allowed a maximal MAD of 15 (Figure 3.6d)

If these criteria were not met, the genetic position of a WGS SNP was set to missing (NA). Otherwise, the cM coordinate of a SNP was defined as the median cM position of its nearest neighbors. Each of the filters we used removed about 5 – 20 % of the SNPs (Figure 3.6) depending on the framework map. When these filters were applied in combination, 6.5 – 15.5% of the SNPs remained unanchored (Table 3.2). SNP anchoring was most effective for the OWB population with over 91 % anchored SNPs, presumably because the OWB framework map was dense enough to capture a larger fraction of the recombination events in the OWB doubled haploid population in contrast to the maps of the Morex × Barke RILs where there are more recombination events per chromosome.

Sequence contigs of the Morex WGS assembly were genetically positioned with the help of WGS SNPs that are located on them. A WGS contig was assigned to a genetic position if at least 80 % of its SNPs had been mapped to the same chromosome and the median absolute deviation of the cM coordinates of the SNPs was less than 5 (15 for MxB GBS). The cM position of a contig was set to the median cM position of all SNPs located on the contig. If chromosome assignments disagreed or the variation of cM positions was too large, the WGS contigs was considered unanchored. Depending on the framework map, 80.8 % to 99.6 % of the anchored SNPs were used to anchor contigs. The higher the fraction of SNPs on anchored contigs, the fewer discrepancies there were between genetic positions of anchored SNPs on the same WGS contig. The two Morex × Barke maps performed best as almost all anchored SNPs could be used to place contigs. The excellent performance of the iSelect map is according to our expectations as this map was calculated from very high quality data with no missing calls and went through manual curation steps (Comadran et al., 2012). It is encouraging to see that the MxB GBS map that was constructed in a single-step procedure without manual intervention performed equally well. In the OWB GBS map, a higher number of errors may have already been embedded into the framework map. The OWB GBS map was constructed from a smaller number of doubled haploid individuals which limited the number of genetic bins and may have complicated the assignment of WGS SNPs to framework markers in the presence of missing data, thus explaining the smaller proportion of SNPs on anchored WGS contigs.

Table 3.2: Number of anchored WGS SNPs

	MxB iSelect	MxB GBS	OWB
no. of all SNPs	5,123,696	5,123,696	6,543,684
no. of anchored SNPs	4,381,020	4,429,475	6,117,837
percentage of all SNPs	85.5 %	86.4 %	93.5 %
no. of SNPs on anchored contigs	4,361,605	4,400,265	4,941,509
percentage of all anchored SNPs	99.6 %	99.3 %	80.8 %

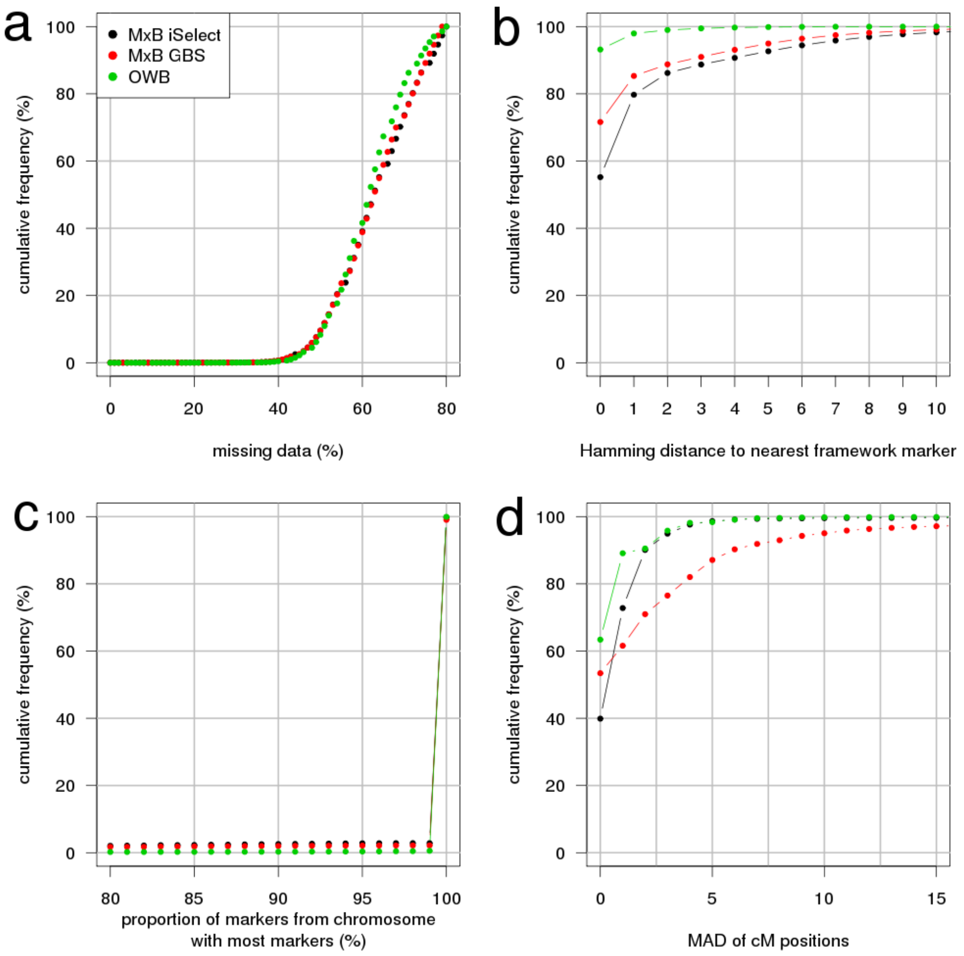


Figure 3.6: Parameters of the POPSEQ anchoring algorithm. Only WGS SNPs with less than 80 % missing data were considered for anchoring **(a)**. We set up filters that checked for the Hamming distance to the nearest framework marker **(b)**, the proportion of framework markers from the chromosome with most markers **(c)** and the median absolute deviation of cM positions of framework markers from the chromosome with most markers **(c)**. Colors refer to the three framework maps we used. The legend for all panels is given in **(a)**.

An overview of the outcome of these computations is given in Table 3.3 and will be discussed in more detail when these anchoring results are compared with each other and to the IBSC map for validation purposes in Chapter 4. In section 4.2, we will also assess the impact of different parameter settings on the outcome of the anchoring process.

Table 3.3: Anchoring statistics.

	MxB WGS	OWB WGS	MxB WGS	MxB GBS	IBSC
Framework map	iSelect[1]	OWB GBS	MxB GBS	iSelect	iSelect
No. of anchored SNPs	4,381,020	6,117,837	4,429,475	17,115	498,165
No. of anchored contigs	498,856	591,779	512,293	10,253	138,443
Size of anchored contigs	927 Mb	978 Mb	934 Mb	49 Mb	410 Mb
	(50%)	(52%)	(50 %)	(3 %)	(21%)
Median length of anchored contigs	1,006 bp	973 bp	977 bp	3,734 bp	1,775 bp
No. of anchored HC genes[2]	16,682	15,743	16,729	3,276	14,923
	(64%)	(60%)	(64%)	(13%)	(57%)
No. of anchored LC genes[3]	28,337	29,033	28,559	3,003	19,415
	(56%)	(55%)	(56%)	(6%)	(38%)

[1] The Morex × Barke framework map described in IBSC (2012) and Comadran et al. (2012).
[2] High confidence genes as described in IBSC (2012).
[3] Low confidence genes as described in IBSC (2012).

4 Proof-of-principle of POPSEQ

The ideal validation of POPSEQ would be to resynthesize a genome with an already finished high-quality sequence assembly, such as the rice or *A. thaliana* genome. However, our major goal when we initially obtained the POPSEQ sequence data was to improve the genomic resources for barley. The development of POPSEQ as a generic method to establish an ordered sequence assembly of any species arose rather by serendipity. Though there is currently no finished reference sequence or even draft genome of barley, comprehensive data sets including BAC sequence data, a physical map and various genetic maps have been collated in recent years. By comparing the results of POPSEQ against these resources, we will show that POPSEQ can not only reproduce the outcome of previous efforts to integrate sequence resources with genetic and physical mapping data, but also constitutes a substantial improvement of these resources.

In this chapter, we check (i) whether POPSEQ is consistent with available short-range linkage information, (ii) whether it is consistent with the published physical and genetic framework of barley (IBSC, 2012), and (iii) whether it is consistent with itself. If different framework maps or different mapping populations are used, POPSEQ should yield the same results, that is two POPSEQ maps should be collinear.

4.1 Comparison to sequenced bacterial artificial chromosomes (BACs)

We ascertained whether the genetic anchoring generated by POPSEQ was consistent with available short-range connectivity information. IBSC (2012) had sequenced 6,278 bacterial artificial chromosomes (BACs). Individual BACs were sequenced to 'Phase 1' quality (i. e. no mate-pair information was used for scaffolding) and consisted of five to ten sequence contigs on average. WGS contigs were compared with megablast version 2.2.26 (Zhang et al., 2000) to 6,278 fully sequenced BACs. We applied very stringent filters (100 % identity over 1,000 bp) to the megablast HSPs in order to avoid spurious hits resulting from paralogous copies of a gene or larger duplicated regions. Using these criteria, we identified 3,902 clones that harbored at least two WGS contigs that were mapped by POPSEQ. The genetic positions of all pairs of contigs on the same BAC were compared. Our hypothesis was that in the majority of cases, pairs of contigs from the same BAC clone (i. e. within a physical distance of less than 200 kb) would exhibit the same genetic location. Indeed, 95 % of the contig pairs were placed within a 3 cM window on the ordered assembly (Table 4.1). Discordant chromosome assignments were found for only 1.7 % of the contig pairs, and a further 3.3 % had a genetic distance larger than 3 cM. We inspected 17 BACs with discordant chromosome assignments and with hits to at least five anchored contigs. For each of these BAC, the chromosome assignments of its contigs were tabulated. If at least 30 % of all contigs on a BAC were anchored to the chromosome with the second highest number of contigs, the BAC was deemed problematic and we checked whether it had been sequenced twice or its length (the cumulative length of its assembled sequence contigs) was unusually large (>180 kb). Nine out of 17 BACs fulfilled these criteria and in these cases, we considered it more likely that the BAC sequence data was incorrect than that POPSEQ was wrong.

4.2 Comparison to the integrated physical and genetic map of barley

We compared the POPSEQ anchoring of WGS contigs relative to the MxB iSelect framework to the released integrated sequence-enriched genetic and physical map of barley (IBSC, 2012), whose backbone for integration of all other genetic markers had also been the MxB iSelect map. Overall, 498,856 contigs with a cumulative length of 927 Mb (49.5 % of the total cv. 'Morex' WGS sequence assembly) could be ordered along the iSelect map (Table 3.3), more than doubling the 410 Mb that was anchored with the help of a genome-wide physical map to the same genetic framework. More than 77,000 WGS contigs (representing 315 Mb of sequence) were assigned by both methods to specific genetic positions. Chromosome assignments disagreed in 2.2 % of the cases and cM coordinates differed by more than 5 cM in 7.0 % of the cases, similar to the 2 – 8% false positive rate observed in PCR-based screening of BAC libraries (IBSC, 2012). In general terms, incongruence appears to occur largely in the highly repetitive and extensive genetic centromeres (Figure 4.1). We believe this to be most likely the product of misplaced repetitive sequence-containing or chimeric BAC contigs in the barley physical map. Thus, employing POPSEQ alongside a fully sequenced minimum tiling path would highlight errors in a physical map and its associated anchoring information, and could thereby be valuable in establishing a robust clone-by-clone assembly of a target genome.

We wished to assess the influence of different parameter settings during the SNP and contig placement steps of POPSEQ (Algorithm 1, Figure 3.6) by performing the anchoring procedures with different stringency settings and comparing the results to the IBSC map. In Algorithm 1, the following parameters can be varied (i) the maximal Hamming distance between WGS SNPs and framework markers, (ii) the tolerated fraction of framework markers from

Table 4.1: Percentage of WGS contig pairs assigned to the same BAC which are positioned farther apart than the specified distance.

Distance	MxB WGS (iSelect)	MxB WGS (GBS map)	OWB
$>$ 0.5 cM	29.28%	29.61%	35.40%
$>$ 1 cM	14.97%	16.19%	21.95%
$>$ 1.5 cM	8.86%	9.32%	20.86%
$>$ 2 cM	5.83%	6.00%	15.99%
$>$ 2.5 cM	4.38%	3.21%	15.58%
$>$ 3 cM	3.25%	2.23%	11.73%
$>$ 3.5 cM	2.48%	1.79%	11.48%
$>$ 4 cM	1.86%	1.45%	8.42%
$>$ 4.5 cM	1.45%	1.25%	8.12%
$>$ 5 cM	0.99%	1.06%	5.81%
$>$ 5.5 cM	0.88%	0.93%	5.68%
$>$ 6 cM	0.85%	0.82%	3.91%
$>$ 6.5 cM	0.77%	0.77%	3.79%
$>$ 7 cM	0.67%	0.71%	2.61%
$>$ 7.5 cM	0.61%	0.69%	2.54%
$>$ 8 cM	0.59%	0.63%	1.89%
$>$ 8.5 cM	0.55%	0.59%	1.88%
$>$ 9 cM	0.52%	0.52%	1.49%
$>$ 9.5 cM	0.45%	0.49%	1.47%
$>$ 10 cM	0.43%	0.47%	1.33%
different chr.	1.66%	1.79%	2.77%

chromosomes other than the chromosome with most framework markers and (iii) the median absolute deviation (MAD) of cM positions of framework markers from the chromosome with most framework markers. In addition to these parameters, the maximal amount of missing data of the WGS SNPs is variable and influences the overall number of SNPs considered for integration into a framework. Analogous to the above criteria (ii) and (iii) for SNP placements, the tolerance for occasional SNPs from other chromosomes and the MAD of all SNPs on a WGS contig can be set differently when aggregating the genetic position of WGS SNPs on the contig level.

In addition to the default parameter set described in section 3.5, we performed the POPSEQ anchoring of WGS contigs against the Morex × Barke

Table 4.2: Evaluation of different parameter sets for POPSEQ

	criteria[1]	no. of anchored contigs	no. of anchored contigs shared with IBSC	disagreement in chr. assignments	disagreement in cM positions
lax	`tol=0.3, MAD=5, miss=0.7, dist=3`	726,483	83,719	5.0 %	7.4 %
default[2]	`tol=0.2, MAD=5, miss=0.8, dist=2`	498,856	77,860	2.2 %	7.0 %
stringent	`tol=0.1, MAD=3, miss=0.8, dist=1`	449,481	73,922	2.0 %	6.7 %
very stringent	`tol=0, MAD=2, miss=0.9, dist=0`	273,745	58,925	1.5 %	6.5 %
default, ≥ 2 SNPs per contig	`tol=0.2, MAD=5, miss=0.8, dist=2`	343,574	63,570	1.8 %	5.8 %

[1] `tol`, tolerance, i. e. the maximal proportion of markers from other chromosomes
`MAD`, median absolute deviation of cM positions of markers from the chromosome with most markers
`missing`, maximal allowed proportion of missing data for WGS SNPs
`dist`, maximal allowed Hamming distance between WGS SNPs and framework markers

[2] set of parameters as described in section 3.5

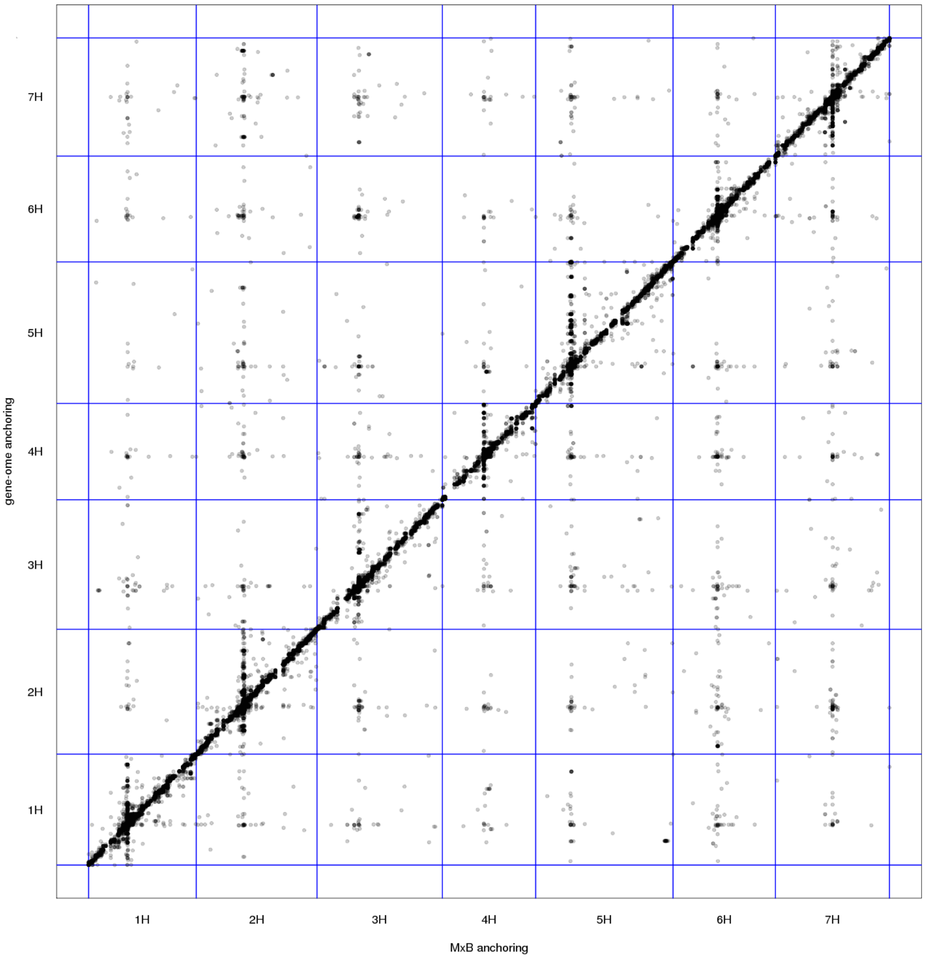


Figure 4.1: Collinearity between the POPSEQ anchoring of the Morex WGS assembly to the MxB iSelect framework (x-axis) and the anchoring of the same assembly reported by IBSC (2012) (y-axis). Each dot is a Morex WGS contigs anchored to both frameworks. 90.8 % of all contigs are within 5 cM of the diagonal. This figure was taken from Mascher et al. (2013a).

iSelect map with one parameter set that was less stringent and two parameter sets that were more stringent (Table 4.2). As expected, more stringent criteria resulted in a smaller number of anchored contigs and higher agreement with the IBSC map. However, even at the highest stringency level, the fractions of markers disagreeing in either chromosome assignment or cM position between the two maps decreased both by less than one percentage point when compared to the default parameter set. The advantage of apparently higher agreement is offset by the overall smaller amount of anchored contigs and consequently a smaller number of contigs that are anchored both by POPSEQ and the IBSC map.

We hypothesized that the disagreement between both maps may be caused by contigs that are anchored by a single erroneously placed SNP. When we required that contig be anchored by at least two SNPs (Table 4.2), the number of anchored contigs decreased by 31 %. The number of discordant chromosome assignments decreased by 0.4 percentage points and the number of discordant cM positions dropped by 1.2 percentage points.

Varying the parameters greatly affected the number of anchored contigs. Using the most stringent parameter set, only half as much contigs could as positioned as when using the most permissive criteria. The disagreement between the POPSEQ maps and the IBSC map fluctuated in a narrow range of 7.8 % and 12.4 %. The default parameters we chose are rather permissive and position a large number of contigs. The threshold for disagreement (5 cM) is also chosen rather arbitrarily. Half of all markers that are placed on the chromosomes by POPSEQ and by the IBSC anchoring, but whose cM positions differ by more than 5 cM, are anchored within 10 cM. In summary, we favored a higher number of (tentatively) anchored contigs over avoiding misplaced contigs at the cost of positioning fewer contigs.

4.3 Using different framework maps for one population

To further investigate the robustness of POPSEQ, we assessed the impact of using a different genotyping platform to construct the framework map. We genotyped the same 90 individuals with a two-enzyme genotyping-by-sequencing (GBS) approach (Table 3.1). We used the *de novo* genetic map comprising 4,056 bi-allelic SNP markers we had constructed with MSTMAP to place WGS contigs into this map using the algorithm described in section 3.5. Altogether, 927 Mb of sequence represented by 512,293 sequence contigs could be ordered (Table 3.3), with 94.3% also linked to the iSelect framework. Importantly, the genetic coordinates of contigs were consistent among the underlying framework maps (Figure 4.2): chromosome assignments were discordant in 0.1 % of the cases, and the map position of only 0.6 % of the contigs differed by more than 5 cM. Though a smaller number of WGS SNPs could be used to place WGS contigs in the MxB GBS map compared to the iSelect map (Table 3.2), overall anchoring results were very similar.

We note that if we only used the SNP markers (∼17,000) provided by GBS, we would be able to anchor only 49 Mb of sequence (Table 3.3), because the number of anchored contigs is bounded by the number of available SNPs. This is similar to the results of Saintenac et al. (2013) who could only associate 4 % of shotgun contigs from wheat chromosome 3A with GBS markers genetically assigned to 3A.

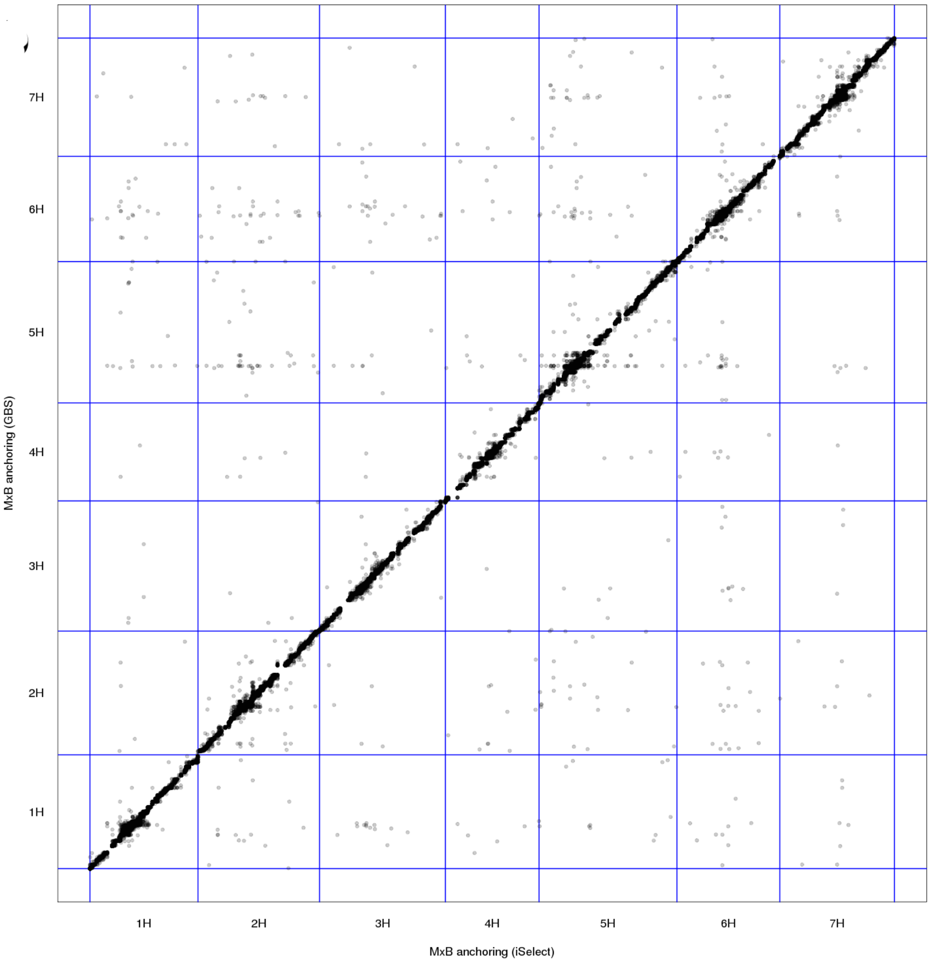


Figure 4.2: Collinearity between two POPSEQ anchorings of the Morex WGS assembly using different framework maps. Each dot is a Morex WGS contig anchored to both frameworks. The position in the anchoring to the MxB iSelect map is given on the x-axis, the anchoring to the MxB GBS map is given on the y-axis. 99.2 % of the contigs are within 5 cM of the diagonal. This figure was taken from Mascher et al. (2013a).

4.4 Using different populations

A last validation step to assess the robustness of the POPSEQ anchoring process, we used a genetic map constructed with the help of a different population to anchor the Morex WGS assembly and compared it to the anchoring obtained using the MxB iSelect framework.

We used the Oregon Wolfe Barley (OWB) population, from which there is genetic map available from GBS on 82 doubled haploid (DH) lines. We survey sequenced these 82 individuals to ca. 1x whole genome coverage each (Table 1) and, by performing the same steps as for MxB, assigned genetic positions to 591,779 WGS contigs corresponding to ∼1,000 Mb of sequence. Of these contigs, 42% (295 Mb) were not anchored to the MxB iSelect framework. In most cases, these contigs either harbored no polymorphisms between Morex and Barke or SNPs were not assayed in a sufficient number of RILs to reach our threshold for inclusion. Contigs anchored to both MxB and OWB maps had highly congruent chromosome assignments (99.6 % agreement, Figure 4.3). Only 6.4 % of all contigs were placed more than 5 cM apart in the two anchored assemblies (falling to 2.1% if we increase the threshold to 7 cM). Given that we were comparing populations constructed with different parents and levels of recombination (ca. half in a DH population compared to RILs), this was not completely unexpected. However, the use of independent populations for anchoring has considerable value: the cumulative length of contigs anchored to either the MxB or OWB map is 1.22 Gb, an increase of one third compared to the use of only a single population. Additional polymorphisms in OWB thus enabled the placement of contigs that were identical between Morex and Barke. More importantly, the POPSEQ ordered assembly positions an additional 5,213 annotated high-confidence genes on the barley genome when compared to the previous release (IBSC, 2012).

This analysis also highlights that it is no prerequisite for POPSEQ that the WGS assembly that is to be anchored is constructed from one parent of the mapping population as Morex is not a parent of the OWB population.

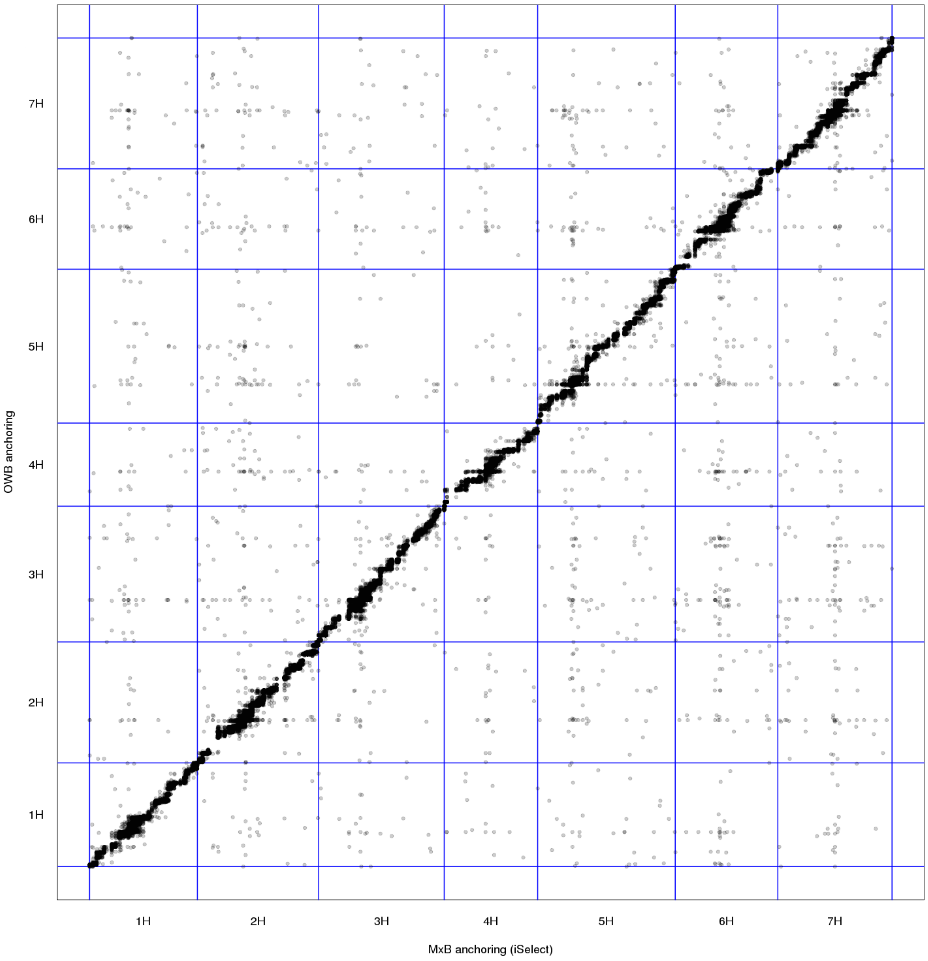


Figure 4.3: Collinearity between the POPSEQ anchoring of WGS contigs against the Morex × Barke framework (x-axis) and against the OWB framework (y-axis). 93.2 % of all contigs are within 5 cM of the diagonal. This figure was taken from Mascher et al. (2013a).

5 Applications of POPSEQ in genome-assisted research

The genome sequence of a species is not an end in itself. But a genome constitutes a "research infrastructure" for biology (Olson, 1993), providing to a wide range of studies in basic and applied research a stepping stone that either makes possible or greatly accelerates the achievement of their aims. Many of these applications do not strictly necessitate a finished reference genome, i. e. near-complete pseudomolecules for each chromosome, but can also be carried out with a partially ordered sequence assembly (possibly supplemented by physical mapping resources) that represents the majority of gene models.

In this chapter, we describe how a POPSEQ assembly may (i) enable reference-based genetic mapping, (ii) function as a hub for gene isolation, (iii) facilitate genetic anchoring of physical maps and (iv) empower comparative genomics. The proper accomplishment of these goals would necessitate the collection of new data. In the absence of these, we tried, as far as possible, to illustrate what may be achieved by simulation with the data we had at hand.

5.1 Reference-based genetic mapping

A reference genomes partially obviates the need for genetic linkage analysis to arrange markers as their order can be obtained by alignment against the reference sequence. This facilitates, for example, the comparison of different genetic maps. In the absence of a reference sequence, comparing marker order between different linkage maps is not straightforward. Different sets of markers are used to construct maps of different populations, and have to be used because markers polymorphic in one population may be monomorphic in another population. Consensus maps are genetic maps that combine linkage maps calculated from genotypic data of several populations. Consensus maps can be constructed directly from the genotypic data of several populations (Beavis and Grant, 1991; Jansen et al., 2001) or by merging the maps calculated for each population without using the original genotypes. Merging can be performed by manual alignment of shared markers in in the case of low-density genetic maps. Denser maps require the use of more elaborate graph-theoretical methods (Yap et al., 2003; Wu et al., 2011).

If an ordered gene space assembly has been established by POPSEQ, marker sequences from different genetic maps can be aligned to the contigs and their

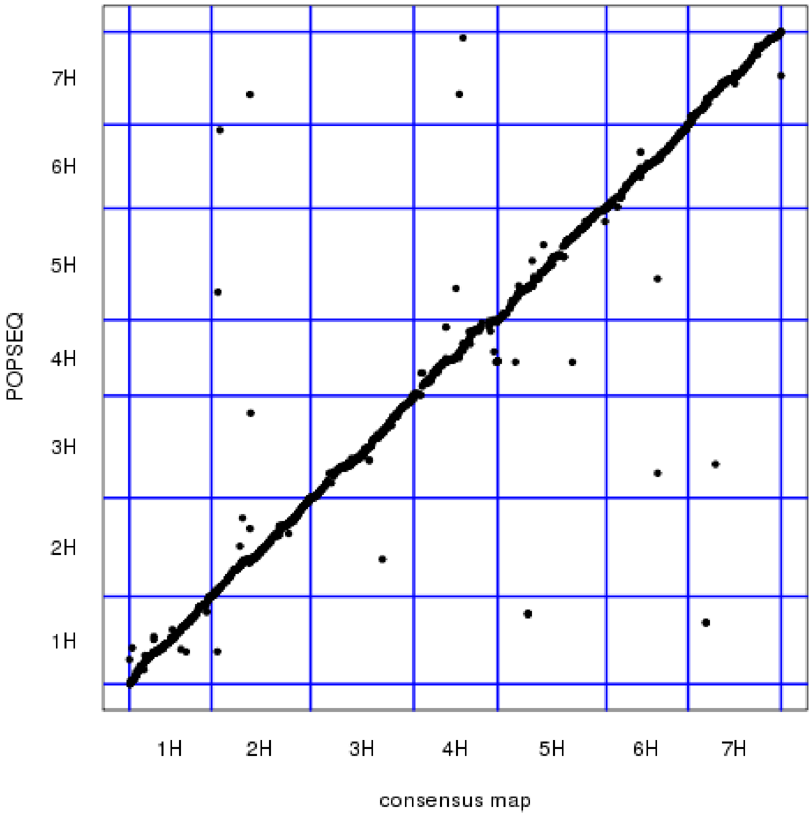


Figure 5.1: Collinearity of the POPSEQ ordering with a recent consensus map. Marker sequences of Close et al. (2009) were mapped to the Morex WGS assembly with BWA-SW (Li and Durbin, 2010). Positions of markers as inferred from the POPSEQ anchoring of their assigned contigs were compared to marker positions in the consensus map of Muñoz-Amatriaín et al. (2011). Out of 2,994 marker placed in the consensus map, 2,436 (81.4 %) could be placed on WGS contigs anchored by POPSEQ. Chromosome assignments agreed between both maps for all except 29 (1.2 %) markers. Marker orders within linkage groups were highly correlated (Spearman's rank correlation = 98.7 %).

chromosomal location be inferred from the genetic positions as given by POPSEQ (Figure 5.1). In addition to simplifying the comparison of existing genetic maps, the analysis of new genotypic data can be performed with the help of a POPSEQ assembly. Without the need for constructing a genetic map *de novo*, genotypic data can be visualized (Figure 5.2) or subjected to further scrutiny such as the selection of suitable recombinants in map-based cloning project. Moreover, a POPSEQ assembly makes it possible to establish marker order in populations where it would be difficult or impossible to construct a robust genome-wide genetic map such as in very small populations (less than 40 individuals), a set of selected recombinant individuals or in advanced backcross populations. In addition, gene models or physical map information associated with a WGS assembly provide valuable information about the gene content in the neighborhood of a given marker.

An important advantage of this reference-based approach to large bi-parental mapping populations is that higher levels of multiplexing can be employed, reducing the genotyping cost per sample. For GBS, higher multiplexing gives fewer data points for each sample and more missing genotype calls. The ordering of markers on a reference framework, however, reduces the need for complete data across the full population (i. e. more missing data is tolerable), while there will still be a surplus of markers for identifying recombination breakpoints for any given line in the mapping population. At current sequencing costs with 192-plex GBS libraries, the per sample genotyping cost for new populations has dropped below US $10 while still producing very high-resolution genetic maps for mapping segregating traits. The strength of GBS for assaying large populations is dependent on obtaining as many independent sequence reads as possible (Poland and Rife, 2012). This determines the amount of missing data as well as the level of multiplexing that can be reasonably utilized. As new sequencing platforms develop, GBS will be preferentially targeted to platforms that produce more reads rather than just longer reads (Poland and Rife, 2012). This is in contrast to most applications of next generation sequencing platforms, particularly those focused on whole genome sequencing and assembly.

5.2 Gene isolation

A genetic map in the literal sense is “map of genes”. Most genes are entities that are only known by their actions, that is by the phenotype caused by a loss-of-function mutation, and not by their sequence. Even though each draft genomes comes equipped with a set of gene models positioned by computational means and associated with a broadly defined function by varying degrees of sequence similarity to known genes of other organisms, these putative transcripts do not constitute a definitive proof of gene function. Associating the function with the sequence of a gene, is commonly referred to as gene cloning. It involves

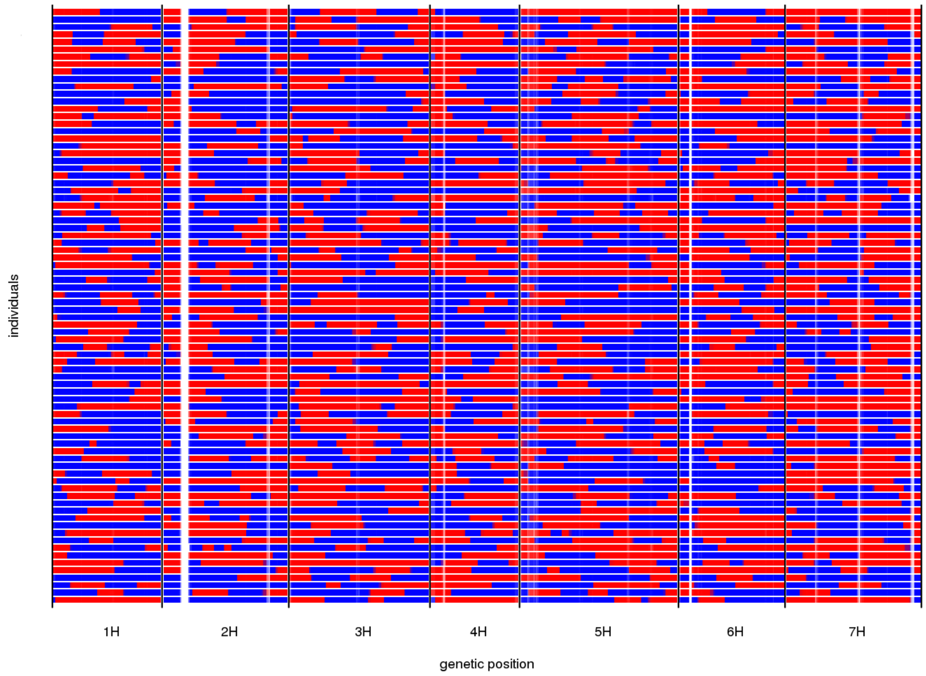


Figure 5.2: Graphical genotypes of 82 individuals from the Oregon Wolfe Barley population. SNP tags from Poland et al. (2012b) were positioned on the Morex WGS contigs and ordered according to the POPSEQ position of their respective contigs. White space on chromosomes 2H, 5H and 7H indicates regions with no polymorphic markers between the parents. In context of a map-based cloning project, individuals with a recombination event close to target interval may be selected for further investigation. This figure is taken from Mascher et al. (2013a).

the processes of locating (mapping) and copying (cloning) a gene out of a genome as an isolated piece of DNA whose exact nucleotide sequence it is then straightforward to determine. After the advent of DNA-based genetic markers, positional cloning has become a cornerstone application of genetic mapping. In the mapping step, an approximate genomic region harboring the gene of interest is delimited through genetic mapping. Subsequently, a local physical map covering the interval between flanking marker is constructed and mined for candidate genes. Though conceptually simple, the practical implementation of map-based cloning has been a formidable exercise, whose success is all but guaranteed and which may take years to accomplish.

Though not strictly necessary for map-based cloning, genomic sequence resources can greatly expedite gene isolation. Pinpointing the causal mutation in an interval delimited by tightly linked markers can become a simple exercise involving the comparison of SNPs found by resequencing against genome annotation databases, if a reference genome is available (Schneeberger and Weigel, 2011).

Mapping-by-sequencing is the application of next generation sequencing to identify genes that underlie phenotypic traits. The first implementation of mapping-by-sequencing was ShoreMap (Schneeberger et al., 2009). The authors successfully identified the mutation responsible for a dwarf phenotype induced by EMS mutagenesis in the model plant *Arabidopsis thaliana.* Mapping-by-sequencing is conceptually similar to bulk-segregant analysis, where DNA from several individuals of a segregating population would be pooled and assayed with conventional molecular markers. The first step of mapping-by-sequencing is the construction of a mapping population. In the original ShoreMap approach, a mutant plant is crossed to a plant from different genetic background. Plants showing the mutant phenotype are selected for sequencing from the F_2 population and pooled DNA is sequenced on a high-throughput sequencing platform. F_2 plants showing that mutant phenotype inherited the chromosomal segment in the vicinity of the causal mutation only from the mutant parent. A chromosomal interval most likely containing the causal mutation is then identified by searching for a group of SNPs with the expected characteristic allele frequency pattern. The region delimited by these SNPs is then queried for candidate genes as well as for sequence polymorphism that are likely to disrupt gene function, such as premature stop codons or frameshifts.

Several enhancements and modifications of the ShoreMap approach have been published recently. Hartwig et al. (2012) demonstrated that the mapping population can also be constructed by crossing the mutant to a wild-type plant with the same genetic background. Only SNPs induced by mutagenesis are then available as genetic markers differentiating the parents of the population. Takagi et al. (2013a) showed that mapping-by-sequencing is not limited to binary traits. They mapped quantitative trait loci by sequencing pools composed of individuals from the phenotypic extremes of a segregating population. Takagi et al. (2013b) combined the inspection of bulked allele frequencies with local *novo* assembly to map genes in regions missing from the genome of the reference genotype.

All these incarnations of mapping-by-sequencing have in common that they require an ordered reference sequence. As the individuals of the mapping population are not genotyped individually but sequenced together in one or several pools, the data available at each marker are not discrete genotype calls for each member of the population, but allele frequencies across the pool. This precludes the possibility of constructing a genetic map from this data alone. Instead, genetic marker positions have to be inferred from an ordered reference sequence. In the absence of such a reference, Galvao et al. (2012) proposed to use syntenic gene order as a surrogate. In a proof-of-principle experiment, they ordered the gene models of *Arabidopsis thaliana* through collinearity to its relative *Brassica rapa* and were able to delineate an approximate interval harboring the causal mutation. Conceptually similar to the genome zipper method (see section 2.4.3), the approach of Galvao et al. (2012) suffers from the same limitations. Synteny-based mapping-by-sequencing is not applicable to organisms that do not have close relatives with fully sequenced genomes to provide the scaffold for establishing a virtual gene order. Even if such proxies are available, a breakdown of collinearity in target regions may preclude the possibility of identifying causal genes. Moreover, the set of syntenic loci that can be incorporated into the analysis is restricted to genic regions.

POPSEQ enables the construction of an ordered reference sequence in non-model species with comparative ease and rapidity. Once a POPSEQ assembly has been established, it can serve as a linear axis on which to hinge mapping-by-sequencing. We illustrate this assertion by using the POPSEQ assembly of barley to fine-map the Vrs1 gene in the OWB barley population with the help of the GBS data, and to fine-map the rough awn trait of barley by using exome capture resequencing data.

5.2.1 Mapping the Vrs1 gene

Spikes of barley are composed of two alternating rows of spikelet triplets. In two-rowed barley, the two lateral spikelets of a triplet are sterile and do not produce grains, whereas they are fertile in six-rowed barley. The two-rowed phenotype confers an adaptive advantage for seed dispersal in the wild (Komatsuda et al., 2007). Consequently, all wild barleys are two-rowed and the six-rowed type is only found in cultivars and weedy barleys that resulted from hybridization between wild and cultivated barley (Komatsuda et al., 2007). The row-type is controlled by a single gene (Vrs1), a homeobox transcription factor, which had been isolated by positional cloning (Komatsuda et al., 2007).

The parents of the OWB population differ in row type. OWBrec has the recessive six-rowed allele and OWBdom has the dominant two-rowed allele. In the doubled haploid progeny, the phenotype is segregating in the expected 1:1 pattern. Vrs1 phenotypes of the OWB population have been recorded for all 82 DH individuals and are available from GrainGenes (Carollo et al., 2005). The SNP genotypes available from the OWB GBS map (Poland et al., 2012b) would have allowed us to find markers tightly linked to Vrs1 by direct inspection of segregation patterns. We chose, however, to create bulked allele frequencies *in silico* in order to simulate mapping-by-sequencing. The OWB DH population was divided into subgroups of individuals showing either the dominant (two-rowed) or the recessive (six-rowed) phenotype. The allele frequencies of the recessive allele in both pools were computed at each individual GBS marker and averaged in 1 cM bins. Then we plotted the allele frequency along the genetic length of the seven barley chromosomes as supplied by POPSEQ (Figure 5.3). The allele frequency ranges between 30 and 70 percent in regions unlinked to the row type. On chromosome 2H, however, there is a frequency pattern that is clearly divergent between both pools. The

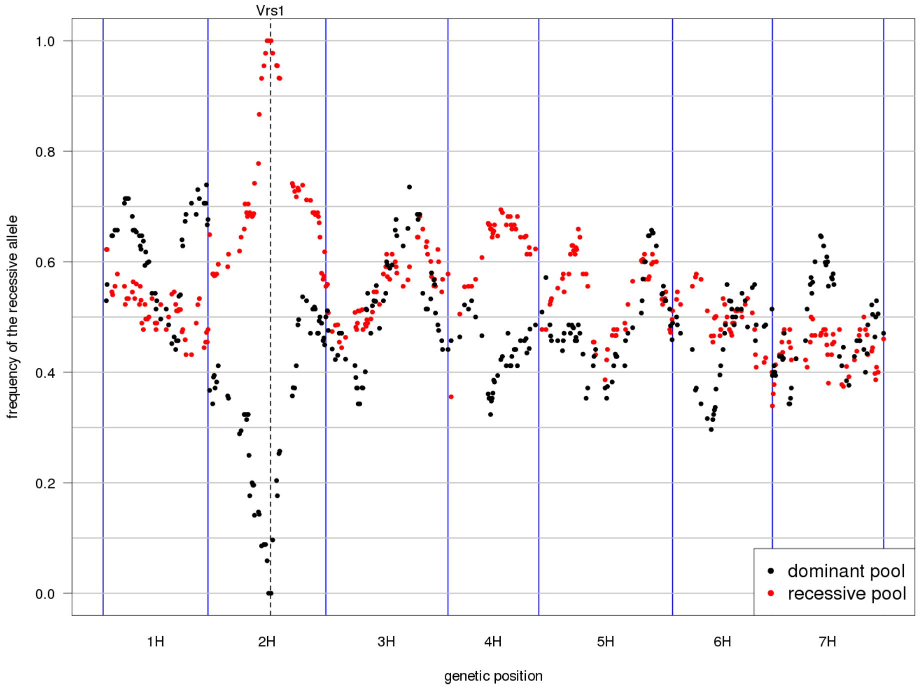


Figure 5.3: Frequency of the recessive allele in phenotypic pools of the OWB population. Plants of OWB population were assigned to the dominant pool if they had the two-rowed phenotype and to the recessive pool if they had the six-rowed phenotype. The allele frequency is plotted along the genetic length of the seven chromosomes of barley (separated by blue lines). The allele frequency was averaged in 1 cM bins. This figure was taken from Mascher et al. (2013a).

peak on 2H coincides with the known position of Vrs1 on the long arm of chromosome 2H (Komatsuda et al., 2007).

In summary, using the GBS data from 82 DH lines and the POPSEQ marker order, we were able to map the Vrs1 gene to a genetic interval of about 2 cM in size. Through further fine-mapping in a larger population, deeper sequencing or the analysis of differential gene expression, cloning of the causal gene might be possible. For a genuine mapping-by-sequencing experiment in barley, sequencing would most likely be performed using the whole exome capture assay developed by Mascher et al. (2013b). This hybridization-based enrichment method targets ~60 Mb of mRNA-coding exons, in contrast to only 10 Mb of the genome that are adjacent to restriction sites targeted through the GBS approach of Poland et al. (2012b). The inclusion of the majority of barley genes may make the single-step identification of causal genes from resequencing data of bulked segregants possible. Thus, exome capture balances better complexity reduction and sequencing load for the purpose of mapping-by-sequencing.

5.2.2 Mapping-by-sequencing with exome capture

The barley whole genome shotgun assembly is gene-focussed and enables a gene-based resequencing strategy that is relevant to both academic and applied interests. Hybridization-based exome capture is an established method that implements this strategy (Bamshad et al., 2011). Briefly, whole genomic DNA is hybridized to pools of oligonucleotide probes that are specific to a set of exons, capturing sequences that are homologous to the targeted regions. The probes are immobilized either by covalent attachment to a glass support (Hodges et al., 2007) or by biotin-streptavidin linkage to an insoluble matrix such as magnetic beads (Bainbridge et al., 2010). The latter approach offers the advantage that both baits and target are in solution during hybridization, decreasing hybridization time. Non-homologous sequences are removed by washing and the hybridized portion eluted and sequenced. As the region targeted for sequencing is greatly reduced, sequencing costs per genome are dramatically lower, allowing high coverage depth of targets and sensitive and accurate variant and genotype calling. Moreover, downstream computational costs per genome for data management, read mapping and variant calling are correspondingly reduced.

Restricting attention to only the mRNA-coding part of the genome can be sufficient to elucidate the molecular basis of natural or induced genetic variation. In biomedical research, exome capture has been successfully applied for the discovery of coding mutations underlying human disease (see the review of Bamshad et al. (2011)) and mutant phenotypes in mice (Fairfield et al., 2011). In maize, a haplotype map (Gore et al., 2009) has been constructed by resequencing only low-copy regions of the genome in different genotypes, and sequence polymorphisms within genic regions have been estimated to contribute a large fraction of the natural variation to quantitatively inherited traits (Liu et al., 2012b).

Exome capture is particular attractive for resequencing studies in large and complex Triticeae genomes where abundant transposable elements and incomplete gene-space assemblies preclude the analysis of a large proportion of NGS reads originating form repetitive DNA. To implement exome capture in barley, we have developed and employed an in-solution hybridization-based sequence capture platform to selectively enrich for a 61.6 megabase coding sequence target that includes predicted genes from the genome assembly of the cultivar Morex as well as publicly available full length cDNAs and *de novo* assembled RNA-seq consensus sequence contigs. The platform provides a highly specific capture with substantial and reproducible enrichment of targeted exons both for cultivated barley and related species (Mascher et al., 2013b). Almost three-quarters (73.7 %) of high-confidence exonic sequence and 40.7 % of low confidence exon sequence annotated on the basis of the barley draft genome assembly (IBSC, 2012) are represented by our target regions. When four captured samples are sequenced on one lane of the Illumina HiSeq 2000, ∼80 – 90

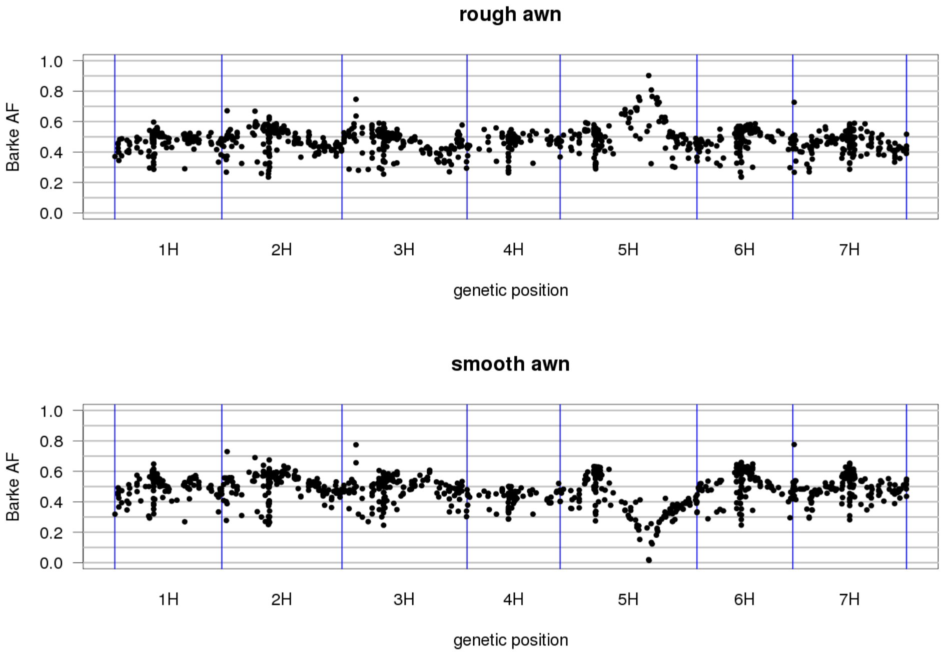
% of all target regions have at least 10x coverage.

We evaluated the applicability of exome capture for mapping-by-sequencing in conjunction with the POPSEQ assembly of barley. We wished to fine-map the rough awn trait in the Morex × Barke population. Wild barley as many other grasses has rough awns with small barbs to facilitate seed dispersal or seed burial (Elbaum et al., 2007). While necessary for survival in the wild, barbs are harmful as they may hurt the throats of animals when barley is used as fodder. A mutation in any of a small number of barley genes (Franckowiak et al., 1997) can cause awns to remain smooth, either because no or only rudimentary awns develop. Smooth-awned varieties of barley are preferable when barley is used for animal nutrition.

As Morex is a smooth-awned cultivar, whereas Barke has rough awns, the awn smoothness is segregating in a 1:1 ratio in the Morex × Barke RIL population and has been mapped as a single gene (Nils Stein et al., unpublished results). For further fine-mapping, we selected from a population of 360 F_8 RIls two bulks consisting either of 180 rough-awned or 180 smooth-awned plants. DNA of plants from each bulk was pooled, subjected to exome capture and subsequent high-throughput sequencing on the Illumina HiSeq 2000. Reads were mapped against whole genome shotgun assembly of barley cultivar Morex using the same procedure as described in Mascher et al. (2013b). SNP calling was performed with the program SNVer (Wei et al., 2011), which can perform allele frequency estimation from pooled sequencing data, whereas other programs such as SAMtools or GATK are only suited for multiple individual samples. Allele frequencies were averaged on an FP contig level and visualized for both pools along the seven chromosomes of barley by using the positional information for the FP contigs provided by POPSEQ (Figure 5.4). Clear peaks of allele frequencies are found in both pools on the long arm of chromosome 5H that coincide with the known map location of the rough awn gene (Franckowiak et al., 1997). In summary, we have shown that a genetically anchored gene assembly serves as an effective backbone to order SNPs along the chromosomes. The visualization of allele frequencies in phenotypically differentiated bulks along a genetically anchored WGS assembly provides an effective means for fine-mapping traits.

5.3 Anchoring physical maps

Gene-space assemblies are not to be considered as finished products. Lacking in completeness – in particular in the repetitive portion of the genome – and contiguity, WGS assemblies of large and complex genomes of flowering plants or mammals have so far not attained the quality of a draft genome. Though they are considered as highly useful resources for the purposes of gene isolation or diversity studies, some global analyses such as studies of the expansion and contraction of gene or repeat families would greatly benefit from a less


rough awn

smooth awn

Figure 5.4: The allele frequencies of the Barke allele is plotted along the length of the seven barley chromosome for pools containing each 180 smooth-awned or rough-awned members of Morex × Barke F_8 RIL population. Allele frequencies were averaged in physical contigs and the genetic positions of contigs were inferred from their POPSEQ anchoring (see Section 5.3).

fragmentary sequence resource.

Hierarchical shotgun sequencing is an established path towards a finished reference genome. Sequence information is organized in several hundreds or thousands of physical contigs. Sequence reads are assembled on a BAC-by-BAC level, avoiding to a large extent the collapse of unrelated repeat elements or paralogous copies of a gene. Similar to sequence contigs, the BAC contigs of a physical map can be ordered along the chromosomes with the help of genetic mapping. As an illustration of how the copious amount of POPSEQ markers can be integrated with survey sequencing data associated with a physical map to anchor this map, we present a new genetic anchoring of the genome-wide physical of barley. Previously, 4,556 physical contigs (3.9 Gb) had been assigned to genetic positions with the help of ∼3,000 EST-based markers and ∼500,000 genotyping-by-sequencing (GBS) markers (IBSC, 2012). Additionally, 1,881 contigs could be assigned to chromosome arms by using 454 sequence data from flow-sorted chromosome arms (IBSC, 2012). Most contigs without genetic or chromosomal positions were either short or lacked sequence or marker information. POPSEQ provided us with an order of magnitude more markers than were available to IBSC (2012). We linked the same BAC sequence information to the Morex WGS assembly and were able to anchor slightly more contigs than was possible with the complex, multi-layered approach of IBSC (2012).

5.3.1 Genetic anchoring of BAC contigs

We used the POPSEQ anchoring of the Morex WGS assembly to anchor BAC contigs of the barley physical map. We first projected the WGS contigs onto the physical map. The whole genome shotgun assembly is a necessary intermediary as the sequence information attached to the physical map is incomplete and consequently cannot serve as an appropriate reference for short-read alignment and SNP calling algorithms. We used the POPSEQ anchorings against the Morex × Barke iSelect framework and the OWB GBS map. The iSelect framework had also been used in the previous effort of anchoring the physical map of barley. By stringent homology search against fully sequenced BACs and BAC end sequences requiring at least 99.5 % sequence identity and a minimum alignment length of 500 bp, we assigned 82,381 WGS contigs to 5,872 BAC contigs (72 % of BAC contigs with associated sequence information).

The genetic position of a physical contig was then set to the median genetic position of all anchored WGS contigs assigned to it. A total of 4,920 and 5,002 BAC contigs could be anchored to the Morex × Barke and OWB maps, respectively. In both cases, three quarters of contig positions were supported by at least two WGS contigs. Out of 4,411 BAC contigs anchored to both maps, 92.8 % (74.3 %) were positioned no farther than 5 cM (2 cM) apart from each other (Figure 5.5a). The proportion of contigs anchored within 1 cM (2 cM) was 56.4 % (74.3 %). A similar degree of agreement between different maps had been found for the anchoring of WGS contigs (see section 4.4). This outcome is the consequence of the different resolutions of underlying genetic maps as well as the procedure of integrating the two maps. By merging anchoring results from both maps, we obtained a set of 5,193 anchored contigs (Table 5.1). The number of anchored contigs varied considerably between distal and peri-centromeric regions (Figure 5.6). In distal regions, the ratio of physical to genetic distance was between 1 and 10 Mb per centiMorgan, while it was 100 – 500 Mb in peri-centromeric regions. Of all contigs anchored to the OWB or Morex × Barke framework, 3,830 (73.8 %) with a cumulative length of 3.5 Gb are also anchored to the published physical and genetic framework (IBSC, 2012). Chromosomal assignments between both maps agreed in 97.6 % of the cases and centiMorgan coordinates were in disagreement in 8.6 % of cases (Figure 5.5b).

Similar to what we had found for the anchoring of WGS contigs (see Section 4.2), discordant contig placements mostly occurred in the genetic centromere. Although the POPSEQ anchoring positions 14 % more contigs than

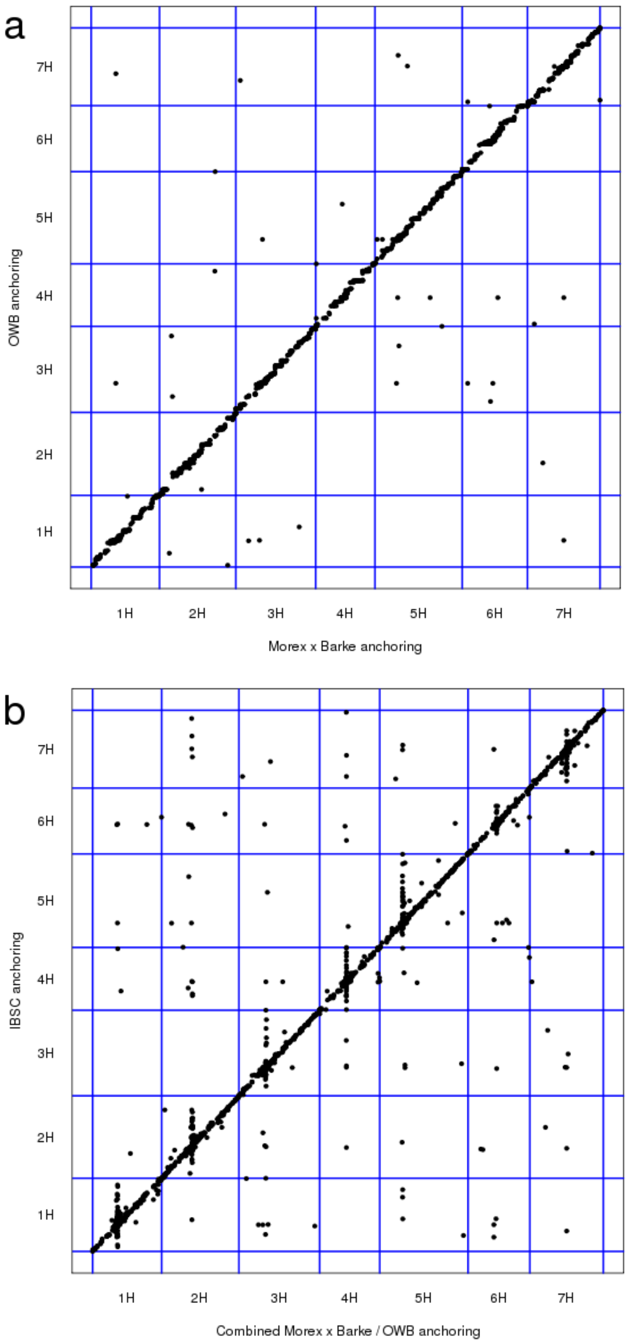


Figure 5.5: Dot plot comparison of different genetic anchorings of the barley physical map. We anchored BAC contigs by POPSEQ to the Morex × Barke and OWB maps. The two panels show **(a)** the collinearity between the two genetic maps and **(b)** the combined POPSEQ anchoring and the previously reported anchoring to various genetic maps (IBSC, 2012).

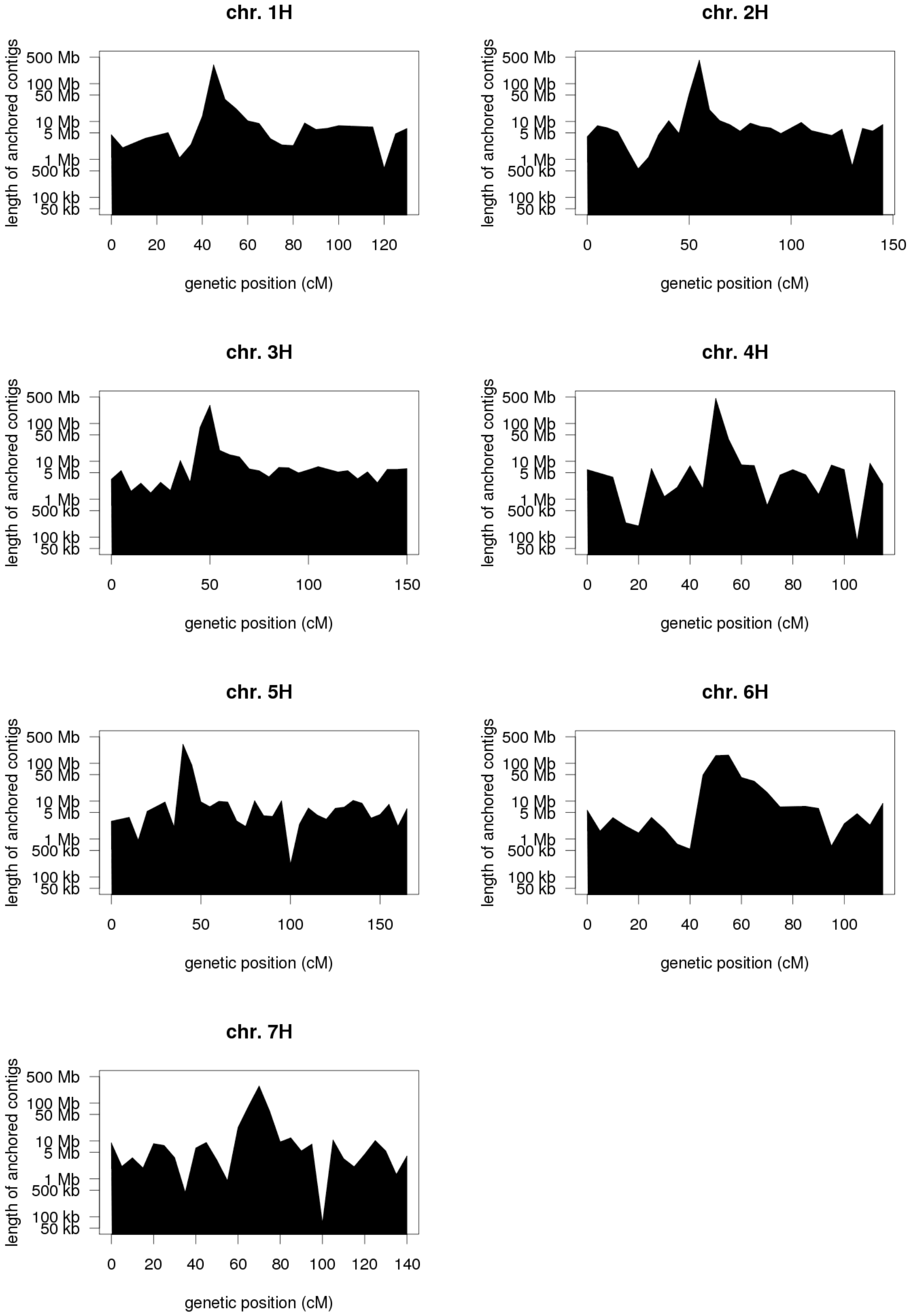


Figure 5.6: Relationship between physical and genetic distance in barley. The length of anchored physical contigs was calculated in 5 cM bins and plotted along the genetic length of each chromosome.

the published physical and genetic framework, the cumulative length of all anchored contigs increases only by 1.3 %. The high number of markers enabled us both to include shorter contigs (mean contig size 761 kb vs. 856 kb) and to exclude some longer contigs with inconsistent marker information. Furthermore, we applied more stringent alignment criteria (500 bp minimum alignment length and $\geq$ 99.5 % identity) compared to the parameters used by IBSC (2012) (200 bp alignment length, 99 % identity) in order to avoid confounding paralogous sequences when assigning WGS contigs to BAC sequences.

5.3.2 Genetic anchoring of single BAC clones

Led by the observation that POPSEQ is able to anchor shorter contigs, we attempted anchoring single, fully-sequenced BAC clones. Instead of aggregating anchoring information per physical contig, we averaged genetic positions on a per-BAC level. A total of 6,243 (99.4 %) of all sequenced BACs harbored WGS contigs and 5,591 (89.1 %) could be anchored to the Morex × Barke or OWB framework (Table 5.1). The genetic positions of BACs and their corresponding FP contigs agreed in 97.6 % of cases. As the number of discordant chromosome assignments was three times higher than the number of discordant cM positions, disagreement between both anchoring methods arises most likely from single wrongly placed clones that are located on different chromosomes from their assigned physical contig. We found pairs of BACs on 71 FP contigs that were anchored to different chromosomes. For this analysis, BACs were required to harbor at least two WGS contigs that were consistently anchored in both the Morex × Barke and OWB frameworks. Among the anchored BACs, there were also 278 singleton clones (i. e. sequenced clones that are not assigned to an FP contig) that could now be assigned to chromosomal locations to guide their assignment to contigs based on sequence similarity.

Presently, POPSEQ anchoring of the barley physical map is limited by the paucity of high quality sequence information integrated with the BAC contigs. Although more than 300,000 BACs have been end-sequenced, these sequences are short (<1,000 bp) and mostly originate from repetitive regions because they are distributed randomly across the genome. As the physical length of all contigs anchored by POPSEQ amounts to 90 % the physical length of all contigs with associated WGS contigs, we anticipate that a substantial increase of anchoring efficiency can only be achieved when more BAC sequence information will be available.

The full power of POPSEQ will only be deployed in the presence of a completely sequenced minimum tiling path. Our preliminary analysis of ∼6,200 fully sequenced clones showed that we can reasonably expect the vast majority of clones to harbor an anchored WGS contig. Alternatively, the sequenced MTP clones may serve as a reference for read mapping after removal of sequence redundancy introduced by overlapping clones. SNP and genotype call-

Table 5.1: POPSEQ anchoring statistics of FP contigs and BAC clones.

	FP contigs			BACs
POPSEQ data	MxB	OWB	MxB + OWB	MxB + OWB
no. of all FP contigs/clones	9,265	9,265	9,265	6,278
no. of FP contigs/clones harboring WGS contigs	5,872	5,872	5,872	6,243
no. of FP contigs/clones harboring anchored WGS contigs	5,295	5,417	5,720	6,189
no. of anchored WGS contigs assigned to FP contigs/clones	53,212	57,942	71,112	34,932
no. of SNPs on anchored WGS contigs assigned to FP contigs/clones	1,040,419	1,013,161	2,053,580	1,231,514
no. of anchored contigs/clones	4,920	5,002	5,193	5,591
length of anchored contigs/clones	3.85 Gb	3.90 Gb	3.95 Gb	703 Mb

ing may be performed on the sequence scaffolds of the physical contigs which would then be directly anchored to a genetic map without the intermediate step of a WGS assembly. The genetic anchoring of individual clones will enable the further validation of FP contig integrity, the identification of erroneously placed clones and the genetic positioning of singleton clones.

5.4 Comparative genomics

Before NGS technology became available, sequencing the complete genome of a species was considered a daunting task. Whole genome sequencing apart from human was confined to a few model species such as *C. elegans*, *D. melanogaster* or *A. thaliana* which had to serve the purposes of many research communities. In recent years, it has become increasingly clear that while genomic resources of model organisms have played crucial roles in unveiling fundamental physiological and molecular mechanisms, they may be ill-suited to advance more narrowly defined research goals in other organisms. Furthermore, model organisms sometimes turned out not to be as good models as was initially thought. For example, the modest extent of syntenic conservation between monocots and dicots limits knowledge transfer from *Arabidopsis* to agronomically important cereals (Brendel et al., 2002). Model organisms do not aid research that explicitly requires whole genome comparison, for example studies into genomic footprints of speciation and adaptation to a specific habitat or environmental stresses. Moreover, the model species of genomics chosen for their past and present importance to biomedical or agronomic research need not be the models of less conspicuous fields such as ecology or comparative biology (Feder and Mitchell-Olds, 2003).

Now that next generation sequencing has enabled the accumulation of genomic resources in non-model species without relying on the reference genomes of more or less closely related species as proxies, Stapley et al. (2010) proposed to “genomicize“ ecological model organisms. Projects have been launched that aim to sequence the genomes of 1,000 plant and animal species (`http://ldl.genomics.org.cn`). Likewise, Varshney et al. (2009) have put forward genomics as a means to boost research in orphan crops, that is plant species that may be major sources of human nutrition in developing countries but have few research tools to assist their improvement.

POPSEQ assemblies can be constructed with only moderate costs and effort and may thus be an attractive means to assist genomic research in non-model organisms. In the following two sections, we will discuss how POPSEQ assemblies may be used to study conserved gene order between species and may also provide the foundation for more sophisticated studies to localize focal points of sequence variation between individuals of the same or different species.

5.4.1 Collinearity

POPSEQ enables genome zipping the other way round. Instead of using the (assumed) synteny between two species to project gene order from one species to the other, the linear arrangement of a large number of genes can be determined for both species independently by genetic linkage mapping. The subsequent comparison of gene order between both species can identify global syntenic blocks, duplicated regions as well as breakdowns of microcollinearity due to rapid genome rearrangements.

Gene models on a WGS assembly can be defined without using genetic anchoring information. Genomic coordinates can be either determined by *de novo* prediction using mathematical models (see, for example, Stanke and Waack (2003)) or mapping transcribed sequence such as ESTs or full-length cDNAs to a genome assembly. Moreover, next generation sequencing of cDNA libraries (RNA-seq) can provide comprehensive annotation of transcribed regions (Roberts et al., 2011). In barley, for instance, a combination of full-length cDNA sequences and RNA-seq reads mapped to the Morex WGS assembly was used to predict ∼26,000 gene models of high confidence. While gene model annotation can be performed on sequence contigs without making reference to chromosomal locations, the linear order of sequence contigs and associated gene models makes it possible to identify blocks of conserved synteny. We illustrate this by identifying the well-known syntenic blocks between barley and the model grass *Brachypodium distachyon* The International Brachypodium Initiative (2010); Mayer et al. (2011) by comparing the positions of orthologous genes in the two genomes (Figure 5.7).

A POPSEQ assembly comprises a large fraction of the transcribed genes of a species. For example, 80 % of barley high confidence genes are located on contigs anchored either to the MxB or OWB framework. Thus, a POPSEQ assembly contains substantially more information than would be available using only comparative linkage mapping. For example, Baxter et al. (2011) have employed a reduced representation sequencing approach to construct a linkage map of the moth *Plutella xylostella* and used it to identify blocks of conserved synteny to the sequenced genome of silkmoth. Moreover, they assembled their sequence tags to contigs of size 100 – 600 bp and annotated these contigs as originating from genes by BLAST searches against the UniRef90 protein database. Less than 300 contigs of genic origin could be defined. In stark contrast to the comprehensive picture of the gene space provided by a POPSEQ assembly, their approach is able to establish syntenic links for only a tiny fraction of genes.

5.4.2 Evolutionary and population genomics

Genomics has been acknowledged as a powerful means to study evolutionary processes across several individuals of a single species (Luikart et al., 2003)

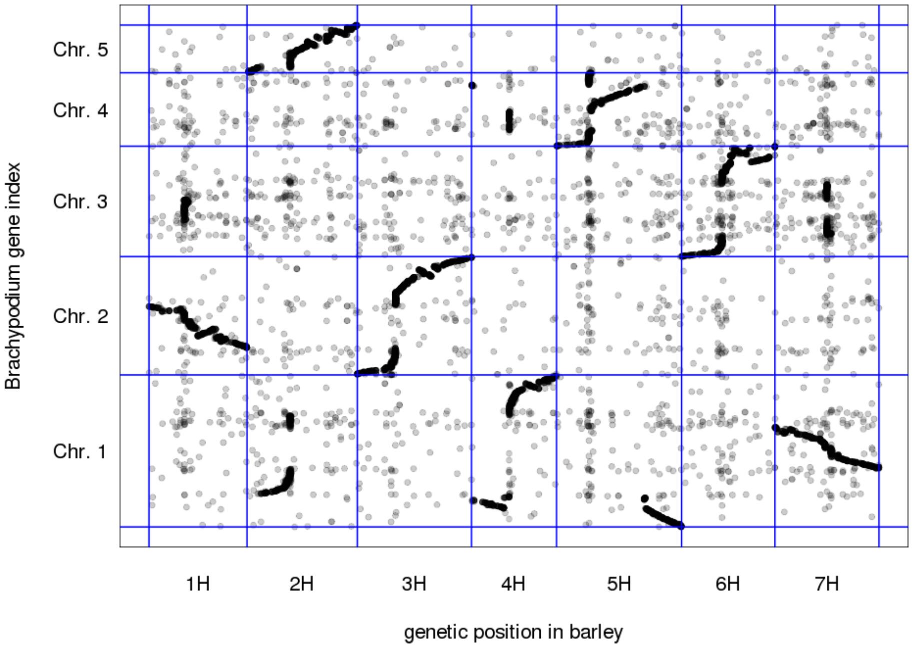


Figure 5.7: Syntenic blocks between *H. vulgare* and *B. distachyon*. The gene order in barley is plotted on the x-axis along genetic distance as given by POPSEQ. The gene order of *B. distachyon* is plotted on the y-axis along a physical scale as given by the reference genome of *B. distachyon* (The International Brachypodium Initiative, 2010). Note the suppressed recombination in the genetic centromere of barley (see Section 6.2). To identify pairs of orthologous genes, protein sequences of barley high confidence genes were aligned by reciprocal BLASTP (Altschul et al., 1990) against the protein sequences of *B. distachyon* gene models (The International Brachypodium Initiative, 2010) (version 1.0). The locations of best bidirectional BLAST hits were visualized with R. This figure was taken from Mascher et al. (2013a).

and also across species boundaries (Sousa and Hey, 2013) to gain insights into how evolutionary forces such as adaptation to environmental conditions, natural selection or random genetic drift shape the genomes of individuals and species. These fields have greatly benefited from the "democratization of sequencing" (Ekblom and Galindo, 2011) engendered by NGS technology. Genomic resources of non-model organisms can now quickly be assembled in order to support specific research aims. The recent study of Ellegren et al. (2012) gives a good example of how a genetically ordered draft genome sequence can be used for evolutionary studies. Ellegren et al. (2012) sequenced the 1.1 Gb genome of collared flycatcher (*Ficedula albicollis*). The assembly of this comparatively small and less repetitive genome resulted in an N50 scaffold size of 7.3 Mb. With the help of this rather extensive short-range connectivity information, a low-density linkage map and conserved synteny with the sequenced zebra finch genome, 89 % of the assembly could be assigned to chromosomal locations. Ellegren et al. (2012) then resequenced individuals of *F. albicollis* and its close, still interfertile relative *F. hypoleuca* to detect genomic intervals that underlie nascent speciation. Without the help of an ordered reference sequence, such a study would not have been possible because signals of selection aggregated on the level of single contigs would have been diluted beyond recognizability.

In an agronomic context, the International Oryza Map Alignment Project (Jacquemin et al., 2013) aims at sequencing eighteen genomes of species from the genus *Oryza*, i. e. relatives of cultivated rice. Starting from the premise that a single reference genome is not sufficient to assess the natural diversity across an entire genus, this project wants to establish a comprehensive genomic infrastructure (i) to empower studies into the dynamics of genome structure on an evolutionary time scale, (ii) to make informed decisions regarding the systematic conservation of genetic diversity endangered by the destruction of natural habitats, and (iii) to identify regions in the genomes of wild *Oryza* species to be introgressed into elite cultivars in order to improve agronomic traits such as yield and stress resistance. For these purposes, various groups have already established WGS assemblies, physical maps and genetic linkage maps for many of the eighteen species (Jacquemin et al., 2013) – resources that could probably benefit from population sequencing data to an extent similar to what we achieved in barley.

Studies similar to the ones of Ellegren et al. (2012) and Jacquemin et al. (2013) can be envisaged in the Triticeae tribe with the help of POPSEQ assemblies for cultivated species such as barley, bread wheat, durum wheat and rye as well as for related wild species. Phylogeny in the Triticeae is all but resolved and complicated by a reticulate structure as a consequence of frequent interspecies hybridizations giving rise to allopolyploids (Escobar et al., 2011) A detailed comparison might be able to highlight a genomic landscape of speciation similar to the one found in flycatchers as well as to identify target genes of agronomic interest.

A resequencing project involving a large number of Triticeae genomes would most probably not employ whole genome shotgun sequencing, but a reduced representation strategy such as whole exome capture. Due to the enormous genome size of Triticeae, sequencing a large panel of accessions to even as little as 1 or 2x coverage would involve a considerable financial investment. Moreover, low-coverage whole genome shotgun resequencing does not allow the accurate genotyping of rare variants and is restricted to the analysis of genetic variation that occur at frequencies above 1 to 5 % (Casals and Bertranpetit, 2012). Tennessen et al. (2012) used exome capture resequencing of the protein-coding part of the 2,440 human genomes to identify more than half a million sequence variants that were previously unknown and have a minor allele frequency below 0.5 %.

The barley exome capture assay can enable resequencing of other *Hordeum* species. We have shown that exome capture hybridization can work efficiently across species boundaries (Mascher et al., 2013b), see Figure 5.8. As the effort of designing another whole exome capture assay specific to other *Hordeum* species is not likely to be taken in the near future, the barley exome capture assay may function as an effective surrogate. While a substantial proportion (50 – 80 %) of the reads could be mapped to the Morex assembly, read mapping is more efficient if an appropriate mapping reference is used, as we had demonstrated by mapping captured *Hordeum pubiflorum* reads against a newly constructed *H. pubiflorum* shotgun assembly (Mascher et al., 2013b) . Using the same approach for tetraploid *H. bulbosum* was less efficient due to issues of assembly quality most likely caused by low read depth and the presence of two homeologous copies of each genomic region (Mascher et al., 2013b). We estimate the necessary minimal genome coverage to produce a *de novo* assembly suitable as a reference for mapping exome capture reads (and POPSEQ data) at 15 – 20x for a diploid species. When an assembly is available, target positions of the barley exome capture can be redefined in terms of the new assembly.

In conclusion, POPSEQ assemblies of selected representative Triticeae may serve as a reference for performing diversity studies based on exome-resequencing in the Triticeae.

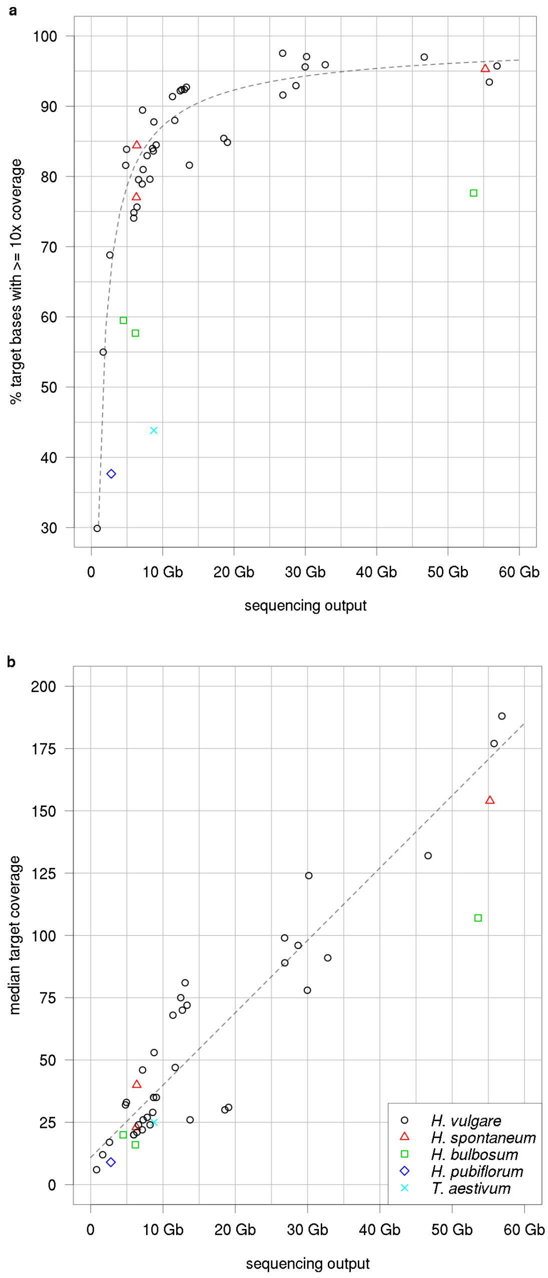


Figure 5.8: The percentage of target regions with at least 10-fold coverage **(a)** and the median coverage **(b)** of target regions are plotted as a function of the raw sequencing output. Different symbols are used for samples from different species. The legend is given in **(b)** for both panels. Regression lines were obtained by fitting the model $\log(1-y) \sim \log(x)$ **(a)** or a linear model **(b)** to the data points of *H. vulgare*. This figure was taken from Mascher et al. (2013b).

6 Discussion and outlook

POPSEQ produces a genetically anchored gene-space assembly by combining experimental procedures with a bioinformatical pipeline for data integration. The final outcome of POPSEQ is determined by a multitude of factors. Among other variables, the genome structure, the assembly strategy, the sequencing technology and the genome coverage influence the quality of the sequence assembly that is used as a reference for short read mapping. Genetic map construction is affected by the mode of reproduction of a species, by the recombination landscape of its genome and by technical aspects of genotyping. The depth of coverage of the population sequencing data determines the amount of missing data and consequently the final resolution with which WGS SNPs can be integrated into a framework map.

We have performed a proof-of-principle POPSEQ experiment in barley, a diploid, inbreeding species. These agreeable characteristics of its genome have made barley a model organism for genomic research in the Triticeae as they greatly facilitate linkage analysis – an apparent advantage that may be offset by its huge and highly repetitive genome, which severely obstruct *de novo* assembly. Though *de novo* assembly is an integral part of POPSEQ, we have not performed *de novo* assembly ourselves, but made use of a recently published gene-space assembly (IBSC, 2012).

In this chapter, we will discuss to what extent POPSEQ is affected by the quality of the underlying genome assembly and sequencing depth of the population sequencing data. We will elaborate on the general limitations of the mapping resolution that can be achieved in Triticeae as a consequence of heavily suppressed recombination in the genetic centromere and discuss what challenges are to be faced when adapting POPSEQ to polyploid and outbred species. Finally, we give an outline how the POPSEQ algorithm could be validated, benchmarked more comprehensively and improved in further studies.

6.1 Impact of assembly quality and sequencing depth

POPSEQ would obviously benefit from an improved sequence assembly with an overall smaller number of contigs, a larger cumulative contig size and consequently a smaller degree of fragmentation. Completeness of the assembly, in particular the correct resolution of repeat structures, would improve read mapping. Likewise, the faithful representation of paralogous copies of genes could prevent collapsed sequences and thus allow a higher number of variant

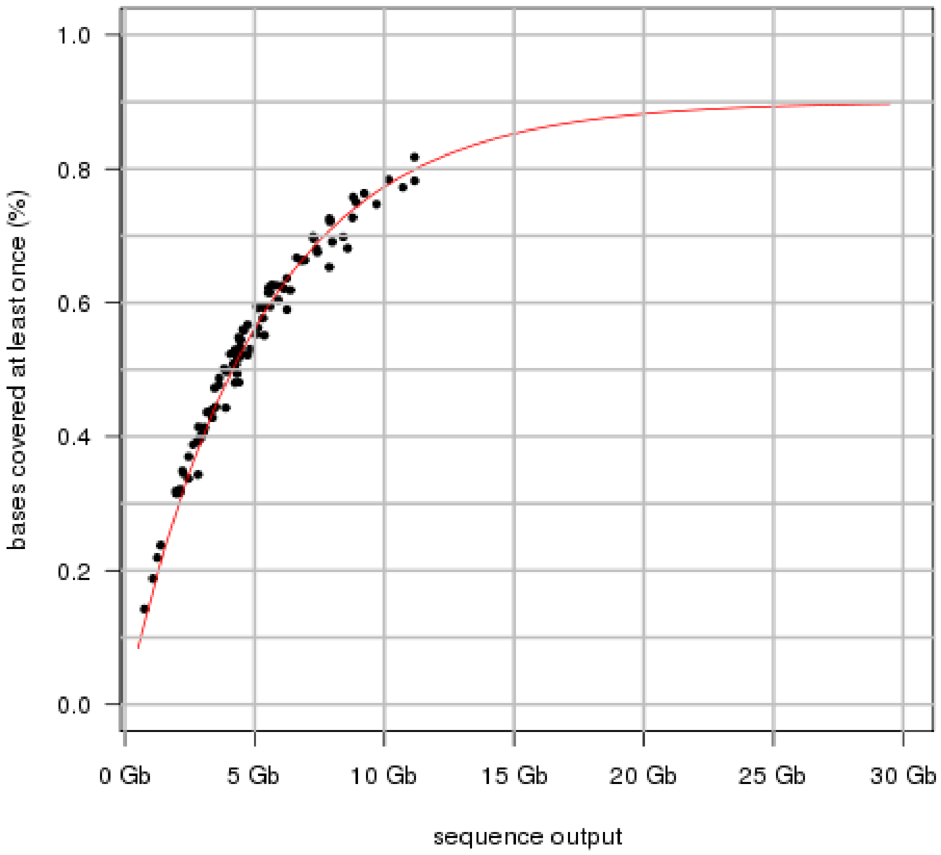


Figure 6.1: Observed and expected sequence coverage according to the model of Lander and Waterman (1988). The number of reference bases that are covered by at least one sequence read. Each dot represents the sequence data from one individual of the Morex × Barke population. The red line is the theoretical genome coverage (given the sequencing output and the expected genome size) according to Lander and Waterman (1988). The expected coverage according to the original formula was multiplied by a factor of 0.9 to fit the observed values more closely. The original formula overestimated the genome coverage, most likely because it did not take unmappable reads into account. This picture is taken from Mascher et al. (2013a).

sites to be correctly genotyped. If other factors remain unaltered, larger contigs and scaffolds that connect multiple contigs harbor more SNPs which can be used to place them.

We also wished to estimate the impact of an assembly even inferior in quality to the one we have used to perform POPSEQ in barley. POPSEQ is attractive especially to research communities in species where no or severely limited resources – in particular funding – are available. It may therefore be desirable to economize on sequence coverage. The set of WGS contigs (the "Morex assembly") used for barley POPSEQ had been computed from sequence data of Illumina libraries with fragment sizes of 350 bp and 2.5 kb (IBSC, 2012). Though large insert mate-pair libraries can be used to establish links between contigs, and may be required input for some assemblers (Gnerre et al., 2011), the construction of such libraries is not straightforward and often yields suboptimal results such as a high fraction of PCR duplicates or short insert read-pairs (Belova et al., 2013). We therefore explored how POPSEQ performed with an assembly made only from short insert paired-end reads. We sequenced the same 350 bp insert library used for the construction of the current barley reference assembly on two lanes of an Illumina HiSeq 2000, yielding ∼15x haploid genome coverage and assembled the reads using the same program as before (CLC assembly cell, `http:/www.clcbio.com`).

As the read coverage was about three times lower than used by IBSC (2012) and did not utilize mate-pair information, we expected the assembly to be of worse quality. The cumulative length of the resulting assembly was shorter (1.6 Gb vs. 1.9 Gb) and the contig N50 was smaller (1,238 bp vs. 1,450 bp). However, contigs of this size are sufficient to function as a reference for read mapping and to enable structural gene annotation via RNA-seq as well as SNP detection. Notably, almost half of the contigs (49.8 %) anchored by POPSEQ to the MxB iSelect framework are shorter than 1,000 bp. In species with smaller and less repetitive genomes, WGS assembly is expected to yield fewer and longer contigs that would potentially harbor a higher number of SNPs per contig (dependent upon the level of polymorphism in the POPSEQ population). Alternatively, larger contigs may compensate for lower levels of polymorphism.

The accuracy of POPSEQ could be improved if the members of one or several mapping populations would be sequenced to higher depth. With the sequencing depth used in this study for the Morex × Barke and OWB populations (1x – 2x), the sequencing reads of each individual cover only ∼50 % of the assembly. Doubling the amount of sequencing data per individual would result in an expected genome coverage of ∼80 % according to the model of Lander and Waterman (Figure 6.1), thus reducing the number of missing genotype calls per individual. An increase in sequencing depth is mandatory for highly heterozygous populations such as F_2s in selfing organisms or F_1s in outcrossing species in order to correctly type heterozygous SNPs. An increase in the number of sequenced individuals (resulting in a proportional increase in the sequencing load) could improve the genetic resolution of the final map. However, regions of severely limited recombination would remain recalcitrant even if larger populations would be utilized.

6.2 Limitations of genetic anchoring in the Triticeae

Recombination frequency is not distributed uniformly along the chromosomes. In humans, for instance, recombination rate varies in a range of about 0.1 to 4 cM per Mb (Kong et al., 2002). A salient feature of Triticeae genomes is the extremely unfavorable ratio of genetic and physical distance in large genomic intervals – so-called genetic centromeres – of each chromosome. The cytogenetically defined centromere is the chromosomal domain where sister chromatids are linked and where spindle fibers attach to separate them during cell division. Discrepancies between physical and genetic distances of single loci to the physical centromere have been observed in barley through cytological methods before the era of molecular marker maps (see for example Künzel (1982)). Through the comparison of comprehensive chromosome-wide linkage maps to cytogenetic maps, large genetic centromeres – comparatively gene-poor regions including the physical centromeres and additionally encompassing up to half of the physical length of a chromosome – emerged as a common feature of grass, and especially Triticeae, chromosomes (Gill et al., 1996; Künzel et al., 2000; Sadder and Weber, 2002)

One shortcoming of the current POPSEQ assembly of barley in contrast to a true draft genome, is the lack of resolution in peri-centromeric regions (Figures 5.6 and 6.2), which is a direct consequence of the severely reduced recombination frequency in the genetic centromere. This deficiency does not only impede genetic anchoring of the WGS assembly and of the physical map, but also hampers map-based cloning. It is not uncommon that even in large mapping populations, closely flanking markers of a target gene situated in the genetic centromere are still remaining on opposite chromosome arms (Okagaki et al., 2012; Shahinnia et al., 2012). Several hundred megabases (encompassing several dozens of BAC contigs) may correspond to a genetic interval of less than 1 cM and the ordering of physical contigs with respect to each other is lacking for these regions.

The limited resolution of our map in centromeric regions may be improved through populations that provide higher mapping resolution, e. g. a large number (>1,000) of recombinant inbred lines. Genome-wide high-density genotyping of several hundred or even thousands of individuals has been made possible by cost-effective genotyping-by-sequencing (Elshire et al., 2011; Poland et al., 2012b). But even with huge mapping populations, genetic linkage analysis is likely to reach its limits in regions of severely repressed recombination.

Alternative methods of physical mapping will have to be explored. In mammalian species, radiation hybrid mapping is a commonly used technique for

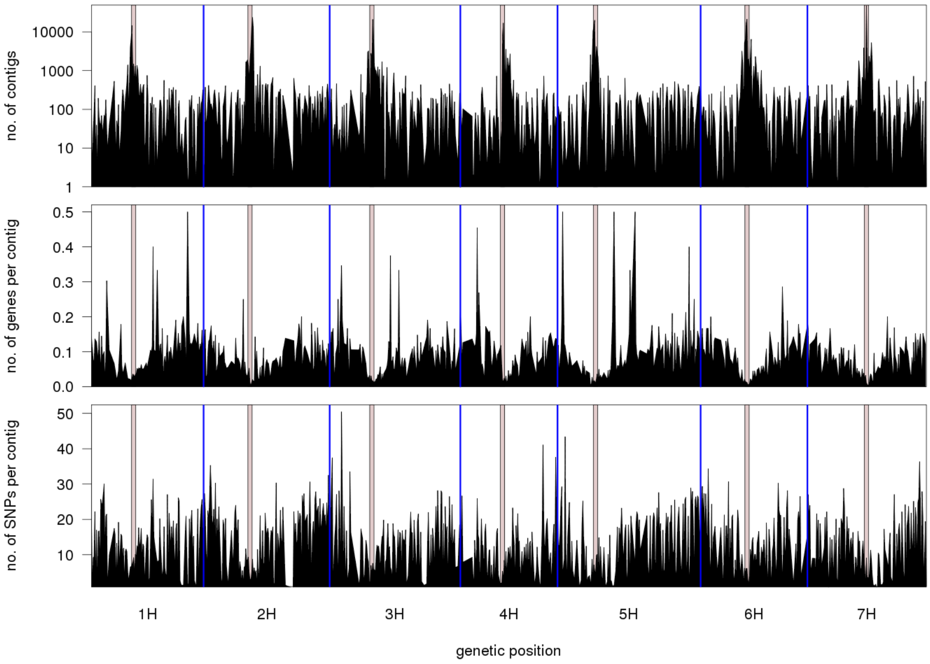


Figure 6.2: Suppressed recombination in the genetic centromere of barley. The number of WGS contigs per genetic bin (upper panel), the number of genes per bin (middle panel) and the number of SNPs between Morex and Barke (lower panel) are plotted on the y-axis. Features were aggregated in 1 cM bins. The x-axis on each panel gives the genetic position of each bin according to POPSEQ. Centromere positions are highlighted in red. This figure was adapted from Mascher et al. (2013a).

physical mapping. Chromosomal breaks leading to the deletion of large chromosomal segments are induced by gamma radiation. In large panels of radiation hybrids, the presence or absence of markers is scored and the order and distance of markers is established by patterns of co-deletion. Radiation hybrid panels have already been implemented in wheat (Kalavacharla et al., 2006). In barley, analogs of radiation hybrid panels have been created by the activity of gametocidal chromosomes (Masoudi-Nejad et al., 2005). Gametocidal chromosomes (GC) are single alien chromosomes added to wheat. If GC additions are combined with barley additions, chromosomal aberrations can be induced in a specific chromosome of barley and the resulting co-deletion patterns be used for map construction as in radiation hybrid panels (Masoudi-Nejad et al., 2005).

Additional physical mapping techniques includes mapping by fluorescent *in situ* hybridizations (FISH) or optical mapping. In FISH mapping, two probes labeled with different fluorophores are hybridized to isolated chromosomes. The distribution of fluorescence signals is used to order the two probes relative to the physical centromere. However, this process is laborious and time-consuming as suitable single-copy probes need to be developed and hybridization patterns have to be manually inspected with a microscope. FISH mapping has, for instance, been applied for mapping a small number of BAC clones in rice (Cheng et al., 2001). Optical mapping is a technique for constructing a high-resolution restriction map where fragment lengths are determined by visualizing restriction sites along a single linearized DNA molecule. Optical mapping is a high-throughput techniques and been used to validate and improve the genome-wide physical maps of rice and maize (Zhou et al., 2007, 2009).

In any mapping approach, the short-range connectivity information afforded by a physical map will be indispensable. Information about clones within fingerprinted contigs will make it possible to extend anchoring information to physically close regions within the same FP contig. Similarly, once an MTP has been sequenced, overlapping adjacent physical contigs can be merged into larger sequence scaffolds. This process would result in an improved linear order of physical contigs in the same genetic bin.

6.3 POPSEQ for polyploid and outbred species

We have performed POPSEQ in barley, a self-fertile, diploid plant species, Most animals, however, are out-bred and polyploidy is common amongst plants. Inbreeding greatly facilitate genetic map construction as populations of entirely homozygous individuals can be developed. In polyploids, high levels of sequence similarity between the homeologous copies of a chromosome complicate genome assembly, or complex patterns of chromosome pairing in meiosis impede genetic map construction.

Nevertheless, genome assembly and linkage analysis is anything but impossible in outbred species and polyploids. In the following, we will discuss how genetic mapping and sequence assembly can be and have been performed in polyploid and outbred species.

6.3.1 Polyploids

The cells of polyploid organisms contain more than two sets of homologous chromosomes. While there are some polyploid animal species (Otto and Whitton, 2000; Wertheim et al., 2013), polyploidy is considered a hallmark of plant genomes. Most flowering plants have been polyploids at one time of their evolutionary history (Adams and Wendel, 2005). Polyploidy has been attributed a key role in domestication, for example by capturing genetic diversity of several progenitors genomes (Dubcovsky and Dvorak, 2007) or by setting reproductive barriers between crops and their wild ancestors (Dempewolf et al., 2012).

Important examples of polyploid crops are tetraploid and hexaploid wheat, oat, potato, rapeseed, sugar cane, cotton and tobacco.

The success of POPSEQ in a polyploid species will large depend on whether and how good a framework genetic map can be constructed. In polyploid species segregation patterns can be classified as multisomic, disomic or intermediate (Stift et al., 2008; Li et al., 2012). Multisomic inheritance occurs when all homologs of a chromosome can randomly pair to produce all possible allelic combinations. In disomic species, the n related copies of a chromosome can be grouped into $n/2$ homeologous groups each comprising a pair of homologous chromosomes. Pairs of homeologous, but not homologous chromosomes are highly similar on the DNA sequence level but do not pair during meiosis. Intermediate stages between these extremes are common and may involve preferential pairing of certain homologs.

If inheritance is disomic, linkage analysis can be carried out as in diploid species because homeologous groups correspond to distinct linkage groups. Theoretical studies have developed analytical tools to model segregation and perform linkage analysis in multisomic polyploids (Luo et al., 2004, 2006; Li et al., 2012). Genetic maps of multisomic or intermediate polyploids have been constructed for alfalfa (Julier et al., 2003), strawberry (Lerceteau-Kohler et al., 2003) or trout (Sakamoto et al., 2000). Alternatively, genetic maps may be constructed in the diploid progenitors (Echt et al., 1994; Choi et al., 2007) or artificially produced diploids (Jacobs et al., 1995; Zhang et al., 2002).

If a framework genetic map of a polyploid has been established, the success of placing marker detected by whole genome resequencing of a mapping population relative to this map depends on how well SNPs can be genotyped in polyploids. Multisomic species are mostly autopolyploids that have resulted from the whole genome duplication of a single progenitor (Stift et al., 2008). Like in diploids, multiple copies of homologous chromosomes will be represented by only a single locus in the haploid sequence assembly. Consequently, NGS reads from all homologous copies will collapse in a single location of the reference sequence and will have to be separated into alleles during genotype calling. Polyploid genotype calling is currently implemented in the UnifiedGenotyper of the Genome Analysis Toolkit (DePristo et al., 2011).

Disomic species are mostly allotetraploids that arose from hybridization of two or more distinct ancestors from different species (Stift et al., 2008). Sequence divergence between the progenitors may be sufficient to enable assembling the subgenomes separately. Alternatively, the genome of still extant diploid progenitors (Ling et al., 2013; Jia et al., 2013) or artificial diploids (Potato Genome Sequencing Consortium, 2011) may be used as proxies for the subgenomes of the polyploid. If homeologous chromosomes can be efficiently disambiguated, NGS reads can be sorted to subgenomes and genotype calling be performed as in diploids.

Among polyploid crop species, hexaploid bread wheat is by far the agronomically most important one in terms of both cultivated area and annual production (http://faostat.fao.org). Having conducted proof of principle in barley, the notion of advancing the closely related bread wheat genome by adopting POPSEQ is of particular interest. The main difference between wheat and barley is the ploidy level, making the wheat genome three times larger than the barley genome. However, disomy greatly simplifies genetic map construction in wheat. Wheat chromosomes can be partitioned into 21 homeologous groups denoted (1A, 1B, 1D, 2A, 2B, 2D and so forth), which correspond to 21 linkage groups. Even though genomic resources are less well developed in wheat than in many other crop species, the construction of sequence-based high-density genetic maps is now routine in hexaploid wheat (Poland et al., 2012b; Saintenac et al., 2013). Wheat is an inbreeding crop further facilitating linkage analysis. Several populations of recombinant inbred lines are already available within the academic and commercial sectors and are ripe for exploitation (Nelson et al., 1995; Manickavelu et al., 2011). The challenge of distinguishing homeologous sequences in the assembly has been largely overcome: sub-genome specific shotgun assemblies have been recently released (Brenchley et al., 2012) and chromosome-specific survey sequences have also been generated (Hernandez et al., 2012). Krasileva et al. (2013) adapted phasing algorithms originally designed for heterozygous diploids to assign SNPs to the subgenomes of tetraploid wheat with a success rate of >98 %.

Wheat will be the last of the world's major crops to be fully sequenced. While the ultimate goal should be a clone-by-clone sequence of wheat with a quality on par with the maize genome, POPSEQ may open the way to obtain with comparative ease an effective surrogate that would be valuable to basic research and breeding applications.

6.3.2 Outbred species

Outbreeding is the only mode of reproduction when an individual has only a single sex or self-fertilization is prevented by mechanisms of self-incompatibility. Hermaphroditism is not as wide spread in animals as it is plants. Most vertebrates – with the notable exception of sequential hermaphrodite fish species – are unisexual (Ghiselin, 1969). Self-incompatible plant species include many fruit and nut trees (Klein et al., 2007), such as apple, avocado, plum, sweet cherry, pear, almond, brazil nut. Self-incompatibility is also common among grasses (Baumann et al., 2000) and is present, for instance, in rye and the important forage grasses *Festuca pratensis* and ryegrass (*Lolium perenne*).

Linkage analysis in families of siblings from a cross between heterozygous parents is more complicated than in the progeny of homozygous lines (Maliepaard et al., 1997). Markers differ in the number of alleles and the number of heterozygous parents, and it can be impossible to determine the linkage phase of a marker, i. e. from which grandparent it was inherited. These problems make it difficult to determine recombination frequencies.

Genetic maps in outbred plant populations can be created by the two-way pseudo-testcross strategy (Grattapaglia and Sederoff, 1994). Two heterozygous parents are crossed. Markers that are heterozygous in one parent and absent from the other parent segregate in a 1:1 ratio in the F_1 generation and their segregation patterns can be used for linkage analysis. This analysis can be performed for markers heterozygous in either parent to construct maternal and paternal linkage maps. Theoretical frameworks for this analysis were provided by Ritter et al. (1990) and Wu et al. (2002) and are implemented in the programs JoinMap (Stam, 1993) and OneMap (Margarido et al., 2007). The two-way pseudo-testcross strategy was first applied in eucalyptus (Grattapaglia and Sederoff, 1994) and has since been used to create linkage maps in, among others, poplar (Cervera et al., 2001), grapevine (Di Gaspero et al., 2007), ryegrass (Studer et al., 2012) and *H. bulbosum* (Salvo-Garrido et al., 2001).

As an alternative to using heterozygous parents as they occur in natural population, inbreeding can be induced artificially. In rye, for example, self-fertile plants can be obtained at a low frequency by selfing of a normally self-incompatible plant (Nilsson and Lundqvist, 1960). These plants and their self-fertile progeny were utilized to generate F_2 (Korzun et al., 2001) or RIL populations (Milczarski et al., 2011) for genetic map construction. Similarly in ryegrass, self-incompatibility can be overcome by heat treatment (Wilkins and Thorogood, 1992) or through introgression of a self-compatible gene from the related species *Lolium temulentum* (Yamada, 2001).

Linkage analysis can be performed even in animal species where it is impossible or difficult to generate large experimental mapping populations. Human genetic maps have been created through the analysis of genotypic data obtained from three-generation families, each consisting of four grandparents, two parents and several children (Donis-Keller et al., 1987; Dib et al., 1996). The map of Dib et al. (1996) incorporated more than 2,000 markers, thus being comparable in density to the framework maps we used for POPSEQ in barley. Similar approaches based on three-generation pedigrees have been employed to obtain genetic maps for other primates, such as rhesus macaque (Rogers et al., 2006), baboon (Rogers et al., 2000) and vervet monkey (Jasinska et al., 2007), as well as for dog (Mellersh et al., 1997) and cat (Menotti-Raymond et al., 1999).

In some animal species, the development of recombinant inbred lines is possible. Inbred strains can be created through repeated mating of siblings over several generations. Like in self-fertile species, RILs can also be created from a cross between inbred individuals. Instead of selfing members of the F_1 generation, male and female F_1 individuals are mated and again siblings of the F_2 generation are intercrossed. Compared to self-mating, the decrease of heterozygosity per generation is smaller and population development takes longer (Broman, 2005). RIL population are available for laboratory animals such as mouse (Williams et al., 2001), rat (Pravenec et al., 1996) and fruit fly (Nuzhdin et al., 1997).

Whenever a linkage map can be constructed, POPSEQ is possible. Apart from more complex linkage analysis, outbred species pose a challenge to POPSEQ by high levels of heterozygosity. Whole genome resequencing of individuals of a mapping population needs be carried out with higher coverage to correctly type heterozygous variants. Likewise, both whole genome shotgun assembly and physical map construction (Moroldo et al., 2008) are complicated by the presence of multiple alleles in one individual. Nevertheless, whole-genome shotgun assemblies have been reported for a wide range of outbred species such as mammals (Table 2.2) or a highly heterozygous grapevine genotype (Velasco et al., 2007), and low coverage resequencing for genotyping is routinely performed in humans (The 1000 Genomes Project Consortium, 2012).

6.4 Validation and improvement of the POPSEQ algorithm

We have performed a proof-of-principle study in barley to illustrate the feasibility of POPSEQ. A more thorough evaluation of the POPSEQ algorithm and its parameters was not possible as there is no gold-standard genome sequence of barley against which to validate our results. Ideally, POPSEQ should be validated in a species with a finished high-quality reference genome. Resequencing data of 132 rice RILs has been published recently (Gao et al., 2013). It was used to construct a high-density recombination bin map using the reference-genome based approach of Huang et al. (2009) and to improve the genome assemblies of the parents of the RIL population, which were also sequenced to high coverage (36- and 64-fold). This dataset constitutes an excellent resource for benchmarking POPSEQ. Rice has one of best genome sequences of any plant species (Feuillet et al., 2011). While the map-based reference sequence of rice was constructed for the cultivar Nipponbare (International Rice Genome Sequencing Project, 2005), draft genome sequences from WGS data of the two parents of the sequenced RIL population are available (Gao et al., 2013) or can be assembled *de novo* from (subsets of) the sequence data. The impact of assembly quality and contiguity could be explored by comparing assembled contigs directly to the finished reference genome. The correctness and exactness of a POPSEQ anchoring might be determined by comparing the assigned genetic order of contigs to their positions in the reference genome. Likewise, the impact of varying parameters of the algorithm might be evaluated. Downsampling of the read data of the RILs from their original fourfold coverage would allow the assessment of how missing data affects the placement of SNPs and WGS contigs and impacts framework map construction. One caveat of a such study would be that the rice genome is small ($\sim$450 Mb), less rich in transposable elements and does not feature such extreme ratios between physical and genetic distance as are found in Triticeae genomes. Thus, conclusions drawn form the rice model may not be fully applicable to more complex genomes.

Several steps of the POPSEQ pipeline may be subject to future change or improvement. POPSEQ currently relies on the integration of markers typed by WGS sequencing into a framework map constructed from a much smaller set of high-quality markers. It can be envisaged that POPSEQ WGS data is directly used for genetic map construction without the use of a framework map as an intermediary. Howe et al. (2013) used microarray technology to genotype an F_2 population of 430 individuals at $\sim$150,000 SNP positions detected by WGS sequencing of the parents of a segregating population. The genotypic data comprising more than 64 million data points was then used as input for MSTMAP. Another way of ordering genetic markers discovered by WGS sequencing would be to construct a genetic map directly from the variant calls made from WGS sequencing data. To incorporate millions of markers into a genetic map, faster algorithms leveraging the power of compute clusters for parallel processing need to be designed and implemented. The current algorithms for genetic map construction are not capable of utilizing millions of markers. Howe et al. (2013) estimated that map construction from their entire 150,000 marker × 430 individuals dataset would take 800 Gb of RAM and three months of compute time with MSTMAP, the fastest algorithm to date. Howe et al. (2013) found that MSTMAP scales linearly with the number of individuals, but exponentially with the number of markers. MSTMAP – in its current implementation – is not parallelized and cannot make use of a compute cluster. Howe et al. (2013) split their data into 15 subsets, computed maps from each set and merged them afterwards. Recently, a new software, Lep-MAP, has been reported to run faster than MSTMAP on large data sets (up to 100,000 markers) and to handle genotyping errors and missing data more efficiently (Rastas et al., 2013). However, the most intriguing feature of Lep-MAP is that it is specifically designed for outbred species, being able to perform haplotype phasing and to utilize data from several families simultaneously.

As an alternative to incorporating all WGS SNPs into a genetic map, *a priori* selection of the most informative markers could greatly reduce the computational burden. The number of genotyped markers in a WGS dataset exceeds by far the number of recombination bins in a mapping population. Gao et al. (2013) found $\sim$ 170,000 high-quality SNPs in a rice RIL population which could be partitioned into only 3,524 recombination blocks. Thus, map construction from the complete data set might not be necessary. Instead, a comparatively small subset of markers that nevertheless captures all cross-overs may be chosen to obtain a framework map with the maximum possible resolution. All other SNP markers would then be integrated into this framework map by the method described in Chapter 3. A major challenge in implementing this strategy would be the selection of informative framework markers without prior knowledge of marker order and in the presence of missing data and experimental noise. Previous efforts of assigning SNPs to recombination bins have taken advantage of the order implied by a map-based reference sequence (Huang et al., 2009; Xie et al., 2010) and averaging in comparatively large physical bins ($\sim$ 100 kb), which is not possible when a highly fragmented WGS assembly is used as a reference for genotyping.

We have performed SNP and genotype calling against the reference sequence of Morex WGS contigs. Limitations of the assembly are also limitations of POPSEQ. For instance, collapsed repetitive regions or recently duplicated paralogs are not accessible for genetic anchoring by POPSEQ. Iqbal et al. (2012) have developed the Cortex assembler as an implementation of colored de Bruijn graph to provide a reference-free framework for variant detection and genotyping. In addition to overlap information, colored de Bruijn graphs encode sample identity by assigning colors to nodes. Iqbal et al. (2012) used the Cortex assembler for variant calling and genotyping of single samples and across multiple samples. Future research might explore how specific properties of plant mapping population such as known linkage phase (when parental genotypes are included) and expected segregation ratios can be used in conjunction with the de Bruijn graph structure to exclude false positive SNPs due to paralogous duplications or even to group and genetically order contigs. Colored de Bruijn might also be superior to conventional assemblers tailored to haploid genomes for the purpose of assembling the genome of highly heterozygous individuals.

6.5 Conclusion

Low coverage (ca. 0.01 – 0.1x) NGS survey sequencing of the small genome (0.4 Gb) of the model crop plant rice, has previously been used as a tool to generate many thousands of genetic markers for both bi-parental linkage studies and genome-wide association studies (Huang et al., 2009). The effectiveness of this 'genotyping by re-sequencing' was afforded by the availability of a high quality reference sequence, a small target genome with comparatively few repeats and innovative statistical approaches to data analysis. Here, we have explored a different application of NGS combined with classical genetic analysis that may find application in many species, particularly those with recalcitrant, large or poorly characterized genomes, among them economically important species such as wheat, sugarcane, pine or *Miscanthus*.

We explored POPSEQ as a method for genetically anchoring and ordering *de novo* NGS assemblies, and have demonstrated its potential by resynthesizing and improving a recently released sequence assembly of the large (5.1Gb) and complex (> 80% repetitive sequence, ancestrally duplicated) barley genome. We used sequence data from two different mapping populations and used the large number of detected SNPs to integrate the sequence assembly with two established framework maps as well as a genetic map computed from GBS data. At its core, POPSEQ exploits the power of genetic segregation combined with shallow (1 – 2x per line) survey sequencing of one or more small experimental populations to genetically anchor NGS sequence assemblies. It is independent of physical mapping and all other genomic resources typically developed in large genome sequencing projects and should be amenable to application in many population types.

We have shown that POPSEQ is both robust and reproducible. Using different genetic maps and mapping population, we obtained comparable results with a concordance of about 95 %. Further validation would require the comparison of a POPSEQ assembly against a finished reference genome. Thus, POPSEQ is neither dependent upon the choice of mapping population nor genotyping platform used for framework map construction. If more extensive short-range connectivity is established by longer sequence contigs or scaffolds, a sliding window approach may be used for genotype calling and framework map construction from POPSEQ data alone, avoiding the need for GBS or SNP mapping platforms. In addition, partitioning of polymorphic sites according to their parental origin may be performed prior to *de novo* assembly, for example by using the colored de Bruijn graph method (Iqbal et al., 2012). The raw sequence reads from POPSEQ (the equivalent of 50x for each parent) should then be sufficient to compute the reference sequence assemblies that will ultimately be ordered along the genetic map.

POPSEQ performs reasonably well with highly fragmented sequence assemblies from short-insert libraries. We were able to construct a *de novo* WGS assembly only from short Illumina reads that showed assembly statistics comparable to an assembly that incorporated mate-pair information. POPSEQ can thus avoid the technical difficulties associated with construction and characterization of large-insert libraries. The simultaneous use of several mapping populations through sequence-based consensus map construction is straightforward, with the same caveats as observed in any genetic map integration. The outcome is not merely an ultra-dense genetic map of anonymous loci: at each genetic position, comprehensive information on the gene space may be obtained through RNA-seq based structural annotation.

The POPSEQ resource we developed here both reproduces and substantially improves the multi-layered gene space assembly that was the result of a large collaborative effort by the IBSC over many years. By comparison, POPSEQ is inexpensive, rapid and conceptually simple, the most time-consuming step being the construction of a mapping population. In relation to the latter, while we used both doubled haploid and recombinant inbred lines, other population types including early generation inbred lines (e.g. F_4s) would also be suitable. Subsequent steps including sequence assembly from short insert libraries, genotyping-by-sequencing (if required) and integrative computational analyses can be conducted quickly. We stress that we do not advocate abandonment of on-going genome projects that are pursuing a clone-by-clone strategy. On the contrary, we believe these may profit from POPSEQ. BAC contigs can be validated though genetic mapping of each single clone and the high number of mapped genetic markers should allow virtually any fully sequenced physical contig to be accurately placed.

For an uncharacterized > 5 Gb diploid genome, with between 14 – 30 HiSeq lanes used for (i) producing a *de novo* sequence assembly for read mapping (2-8 lanes); (ii) genotyping-by-sequencing for map construction (1 lane); (iii) shallow population sequencing (minimum 12 lanes for each population of $\sim$90 lines, though depth can be varied); (iv) deep RNA-seq for structural gene annotation ($>$2 lanes) amounting to US \$50,000 - \$100,000 in sequencing costs, and (v) a medium-sized compute server (32 CPU cores, 512 GB RAM, 3 TB of disk space) it was possible to generate a *de novo* linear gene space assembly.

We propose that POPSEQ may contribute to fundamental research in plant genetics as well as in crop improvement. However, its application is not restricted to plants. The fast and steady advances in sequencing technology will further increase the power of POPSEQ with deeper coverage of larger and outbred populations. As long as the inherent complexity of genomes restricts the assembly of pseudomolecules by shotgun sequencing, POPSEQ provides a rapid, low-cost, and effective method for developing a useful 'interim reference' genome sequence in most species where it is possible to construct a genetic map.

Bibliography

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* **8**: 135–141.

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. 2000. The genome sequence of Drosophila melanogaster. *Science* **287**: 2185–2195.

Akhunov ED, Sehgal S, Liang H, Wang S, Akhunova AR, Kaur G, Li W, Forrest KL, See D, Simkova H, et al. 2013. Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. *Plant Physiol* **161**: 252–265.

Alkan C, Sajjadian S, Eichler EE. 2011. Limitations of next-generation genome sequence assembly. *Nat Methods* **8**: 61–65.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**: 513–516.

Anderson S. 1981. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res* **9**: 3015–3027.

Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama-Morita T, Stern DL. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res* **21**: 610–617.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**: 796–815.

Ariyadasa R, Stein N. 2012. Advances in BAC-based physical mapping and map integration strategies in plants. *J Biomed Biotechnol* **2012**: 184854.

Asan Y, Jiang H, Tyler-Smith C, Xue Y, Jiang T, Wang J, Wu M, Liu X, Tian G, Wang J, et al. 2011. Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol* **12**: R95.

Avery OT, MacLeod CM, McCarty M. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J Exp Med* **79**: 137–158.

Bainbridge MN, Wang M, Burgess DL, Kovar C, Rodesch MJ, D'Ascenzo M, Kitzman J, Wu YQ, Newsham I, Richmond TA, et al. 2010. Whole exome capture in solution with 3 Gbp of data. *Genome Biol* **11**: R62.

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. 2011. Exome sequencing as a tool for mendelian disease gene discovery. *Nat Rev Genet* **12**: 745–755.

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS. 2007. SNP discovery via 454 transcriptome sequencing. *Plant J* **51**: 910–918.

Bateson W, Saunders ER, Punnett RC. 1908. Experimental studies in the physiology of heredity. In *Reports to the evolution committee of the Royal Society*. Harrison and sons, London.

Baumann U, Juttner J, Bian X, Langridge P. 2000. Self-incompatibility in the grasses. *Ann Bot* **85**: 203–209.

Baxter SW, Davey JW, Johnston JS, Shelton AM, Heckel DG, Jiggins CD, Blaxter ML. 2011. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE* **6**: e19315.

Beavis W, Grant D. 1991. A linkage map based on information from four F2 populations of maize (Zea mays L.). *Theor Appl Genet* **82**: 636–644.

Belova T, Zhan B, Wright J, Caccamo M, Asp T, Simkova H, Kent M, Bendixen C, Panitz F, Lien S, et al. 2013. Integration of mate pair sequences to improve shotgun assemblies of flow-sorted chromosome arms of hexaploid wheat. *BMC Genomics* **14**: 222.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.

Berkman PJ, Visendi P, Lee HC, Stiller J, Manoli S, Lorenc MT, Lai K, Batley J, Fleury D, Simkova H, et al. 2013. Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnol J* **11**: 564–571.

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, et al. 1997. The complete genome sequence of escherichia coli k-12. *Science* **277**: 1453–1462.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**: 578–579.

Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**: 314.

Bowers JE, Bachlava E, Brunick RL, Rieseberg LH, Knapp SJ, Burke JM. 2012. Development of a 10,000 locus genetic map of the sunflower genome based on multiple crosses. *G3 (Bethesda)* **2**: 721–729.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**: 2633–2635.

Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, et al. 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**: 705–710.

Brendel V, Kurtz S, Walbot V. 2002. Comparative genomics of Arabidopsis and maize: prospects and limitations. *Genome Biol* **3**: REVIEWS1005.

Bridges CB, Brehme KS. 1944. *The mutants of Drosophila melanogaster*. Publication of the Carnegie Institution, Washington D.C.

Broman KW. 2005. The genomes of recombinant inbred lines. *Genetics* **169**: 1133–1146.

Burr B, Burr FA, Thompson KH, Albertson MC, Stuber CW. 1988. Gene mapping with recombinant inbreds in maize. *Genetics* **118**: 519–526.

Burrows M, Wheeler DJ. 1994. A block-sorting lossless data compression algorithm. In *Technical report 124*. Digital Equipment Corporation, Palo Alto, CA.

Caenorhabditis elegans Sequencing Consortium. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**: 2012–2018.

Carollo V, Matthews DE, Lazo GR, Blake TK, Hummel DD, Lui N, Hane DL, Anderson OD. 2005. GrainGenes 2.0. an improved resource for the small-grains community. *Plant Physiol* **139**: 643–651.

Casals F, Bertranpetit J. 2012. Genetics. Human genetic variation, shared and private. *Science* **337**: 39–40.

Cervera MT, Storme V, Ivens B, Gusmao J, Liu BH, Hostyn V, Van Slycken J, Van Montagu M, Boerjan W. 2001. Dense genetic linkage maps of three Populus species (Populus deltoides, P. nigra and P. trichocarpa) based on AFLP and microsatellite markers. *Genetics* **158**: 787–809.

Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, et al. 2009. Genomics. Genome project standards in a new era of sequencing. *Science* **326**: 236–237.

Chandler VL, Brendel V. 2002. The Maize Genome Sequencing Project. *Plant Physiol* **130**: 1594–1597.

Cheema J, Dicks J. 2009. Computational approaches and software tools for genetic linkage map estimation in plants. *Brief Bioinformatics* **10**: 595–608.

Chen EY, Schlessinger D, Kere J. 1993. Ordered shotgun sequencing, a strategy for integrated mapping and sequencing of YAC clones. *Genomics* **17**: 651–656.

Chen M, Presting G, Barbazuk WB, Goicoechea JL, Blackmon B, Fang G, Kim H, Frisch D, Yu Y, Sun S, et al. 2002. An integrated physical and genetic map of the rice genome. *Plant Cell* **14**: 537–545.

Cheng Z, Presting GG, Buell CR, Wing RA, Jiang J. 2001. High-resolution pachytene chromosome mapping of bacterial artificial chromosomes anchored by genetic markers reveals the centromere location and the distribution of genetic recombination along chromosome 10 of rice. *Genetics* **157**: 1749–1757.

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genet* **3**: 19.

Choi SR, Teakle GR, Plaha P, Kim JH, Allender CJ, Beynon E, Piao ZY, Soengas P, Han TH, King GJ, et al. 2007. The reference genetic linkage map for the multinational brassica rapa genome sequencing project. *Theor Appl Genet* **115**: 777–792.

Cistue L, Cuesta-Marcos A, Chao S, Echavarri B, Chutimanitsakun Y, Corey A, Filichkina T, Garcia-Marino N, Romagosa I, Hayes PM. 2011. Comparative mapping of the Oregon Wolfe Barley using doubled haploid lines derived from female and male gametes. *Theor Appl Genet* **122**: 1399–1410.

Close TJ, Bhat PR, Lonardi S, Wu Y, Rostoks N, Ramsay L, Druka A, Stein N, Svensson JT, Wanamaker S, et al. 2009. Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics* **10**: 582.

Comadran J, Kilian B, Russell J, Ramsay L, Stein N, Ganal M, Shaw P, Bayer M, Thomas W, Marshall D, et al. 2012. Natural variation in a homolog of Antirrhinum CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. *Nat Genet* **44**: 1388–1392.

Cone KC, McMullen MD, Bi IV, Davis GL, Yim YS, Gardiner JM, Polacco ML, Sanchez-Villeda H, Fang Z, Schroeder SG, et al. 2002. Genetic, physical, and informatics resources for maize. On the road to an integrated map. *Plant Physiol* **130**: 1598–1605.

Cosart T, Beja-Pereira A, Chen S, Ng SB, Shendure J, Luikart G. 2011. Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics* **12**: 347.

Costa J, Corey A, Hayes P, Jobet C, Kleinhofs A, Kopisch-Obusch A, Kramer S, Kudrna D, Li M, Riera-Lizarazu O, et al. 2001. Molecular mapping of the Oregon Wolfe Barleys: a phenotypically polymorphic doubled-haploid population. *Theor Appl Genet* **103**: 415–424.

Coulson A, Sulston J, Brenner S, Karn J. 1986. Toward a physical map of the genome of the nematode Caenorhabditis elegans. *Proc Natl Acad Sci USA* **83**: 7821–7825.

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA, et al. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. *Nat Methods* **5**: 887–893.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* **12**: 499–510.

Dempewolf H, Hodgins KA, Rummell SE, Ellstrand NC, Rieseberg LH. 2012. Reproductive isolation during domestication. *Plant Cell* **24**: 2710–2717.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–498.

Di Gaspero G, Cipriani G, Adam-Blondon AF, Testolin R. 2007. Linkage maps of grapevine displaying the chromosomal locations of 420 microsatellite markers and 82 markers for R-gene candidates. *Theor Appl Genet* **114**: 1249–1263.

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

Diguistini S, Liao NY, Platt D, Robertson G, Seidel M, Chan SK, Docking TR, Birol I, Holt RA, Hirst M, et al. 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biol* **10**: R94.

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES. 1987. A genetic linkage map of the human genome. *Cell* **51**: 319–337.

Druka A, Franckowiak J, Lundqvist U, Bonar N, Alexander J, Houston K, Radovic S, Shahinnia F, Vendramin V, Morgante M, et al. 2011. Genetic dissection of barley morphology and development. *Plant Physiol* **155**: 617–627.

Dubcovsky J, Dvorak J. 2007. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* **316**: 1862–1866.

Duffy DL. 2006. An integrated genetic map for linkage analysis. *Behav Genet* **36**: 4–6.

Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO, Buffalo V, Zerbino DR, Diekhans M, et al. 2011. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res* **21**: 2224–2241.

Echt CS, Kidwell KK, Knapp SJ, Osborn TC, McCoy TJ. 1994. Linkage mapping in diploid alfalfa (Medicago sativa). *Genome* **37**: 61–71.

Ekblom R, Galindo J. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**: 1–15.

Elbaum R, Zaltzman L, Burgert I, Fratzl P. 2007. The role of wheat awns in the seed dispersal unit. *Science* **316**: 884–886.

Ellegren H, Smeds L, Burri R, Olason PI, Backstrom N, Kawakami T, Kunstner A, Makinen H, Nadachowska-Brzyska K, Qvarnstrom A, et al. 2012. The genomic landscape of species divergence in Ficedula flycatchers. *Nature* **491**: 756–760.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**: e19379.

Emerson RA, Beadle GW, Fraser AC. 1935. A summary of linkage studies in maize. In *Cornell University Agricultural Experiment Station Memoir 180*. Cornell University, New York.

Emrich SJ, Barbazuk WB, Li L, Schnable PS. 2007. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res* **17**: 69–73.

Escobar JS, Scornavacca C, Cenci A, Guilhaumon C, Santoni S, Douzery EJ, Ranwez V, Glemin S, David J. 2011. Multigenic phylogeny and analysis of tree incongruences in Triticeae (Poaceae). *BMC Evol Biol* **11**: 181.

Fairfield H, Gilbert GJ, Barter M, Corrigan RR, Curtain M, Ding Y, D'Ascenzo M, Gerhardt DJ, He C, Huang W, et al. 2011. Mutation discovery in mice by whole exome sequencing. *Genome Biol* **12**: R86.

Feder ME, Mitchell-Olds T. 2003. Evolutionary and ecological functional genomics. *Nat Rev Genet* **4**: 651–657.

Ferragina P, Manzini G. 2000. Opportunistic data structures with applications. In *Proceedings of the 41st Annual Symposium on the Foundations of Computer Science*, pages 390–398. IEEE.

Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K. 2011. Crop genome sequencing: lessons and rationales. *Trends Plant Sci* **16**: 77–88.

Feuillet C, Stein N, Rossini L, Praud S, Mayer K, Schulman A, Eversole K, Appels R. 2012. Integrating cereal genomics to support innovation in the Triticeae. *Funct Integr Genomics* **12**: 573–583.

Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, Min Jou W, Molemans F, Raeymaekers A, Van den Berghe A, et al. 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**: 500–507.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**: 496–512.

Forster BP, Heberle-Bors E, Kasha KJ, Touraev A. 2007. The resurgence of haploids in higher plants. *Trends Plant Sci* **12**: 368–375.

Franckowiak J, Lundqvist U, Konishi T. 1997. New and revised names for barley genes. *Barley Genet Newslett* **26**: 4–8.

Galvao VC, Nordstrom KJ, Lanz C, Sulz P, Mathieu J, Pose D, Schmid M, Weigel D, Schneeberger K. 2012. Synteny-based mapping-by-sequencing enabled by targeted enrichment. *Plant J* **71**: 517–526.

Gao ZY, Zhao SC, He WM, Guo LB, Peng YL, Wang JJ, Guo XS, Zhang XM, Rao YC, Zhang C, et al. 2013. Dissecting yield-associated loci in super hybrid rice by resequencing recombinant inbred lines and improving parental genome sequences. *Proc Natl Acad Sci USA* **110**: 14492–14497.

Gardner RC, Howarth AJ, Hahn P, Brown-Luedi M, Shepherd RJ, Messing J. 1981. The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by m13mp7 shotgun sequencing. *Nucleic Acids Res* **9**: 2871–2888.

Genovese G, Handsaker RE, Li H, Altemose N, Lindgren AM, Chambert K, Pasaniuc B, Price AL, Reich D, Morton CC, et al. 2013. Using population admixture to help complete maps of the human genome. *Nat Genet* **45**: 406–414.

Ghiselin MT. 1969. The evolution of hermaphroditism among animals. *Q Rev Biol* **44**: 189–208.

Gill KS, Gill BS, Endo TR, Boyko EV. 1996. Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat. *Genetics* **143**: 1001–1012.

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* **108**: 1513–1518.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* **296**: 92–100.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al. 1996. Life with 6000 genes. *Science* **274**: 563–567.

Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, et al. 2009. A first-generation haplotype map of maize. *Science* **326**: 1115–1117.

Grattapaglia D, Sederoff R. 1994. Genetic linkage maps of eucalyptus grandis and eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics* **137**: 1121–1137.

Green P. 1997. Against a whole-genome shotgun. *Genome Res* **7**: 410–417.

Green P. 2002. Whole-genome disassembly. *Proc Natl Acad Sci USA* **99**: 4143–4144.

Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, et al. 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**: 330–336.

Gregory SG, Howell GR, Bentley DR. 1997. Genome mapping by fluorescent fingerprinting. *Genome Res* **7**: 1162–1168.

Groenen MA, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens HJ, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**: 393–398.

Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, Wang Z, et al. 2012. The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions. *Nat Genet* **45**: 51–58.

Hackett CA, Broadfoot LB. 2003. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity* **90**: 33–38.

Haldane JBS. 1919. The combination of linkage values and the calculation of distance between the loci of linked factors. *J Genet* **8**: 299–309.

Hamburg MA, Collins FS. 2010. The path to personalized medicine. *N Engl J Med* **363**: 301–304.

Haring V, Gray JE, McClure BA, Anderson MA, Clarke AE. 1990. Self-incompatibility: a self-recognition system in plants. *Science* **250**: 937–941.

Hartwig B, James GV, Konrad K, Schneeberger K, Turck F. 2012. Fast isogenic mapping-by-sequencing of ethyl methanesulfonate-induced mutant bulks. *Plant Physiol* **160**: 591–600.

Henry RJ, editor. 2012. *Molecular Marker Technology in Plant Science.* John Wiley & Sons.

Hernandez P, Martis M, Dorado G, Pfeifer M, Galvez S, Schaaf S, Jouve N, Simkova H, Valarik M, Dolezel J, et al. 2012. Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant J* **69**: 377–386.

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, et al. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39**: 1522–1527.

Hon WK, Lam TW, Sadakane K, Sung WK, Yiu SM. 2007. A space and time efficient algorithm for constructing compressed suffix arrays. *Algorithmica* **48**: 23–36.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**: 498–503.

Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, Guan J, Fan D, Weng Q, Huang T, et al. 2009. High-throughput genotyping by whole-genome resequencing. *Genome Res* **19**: 1068–1076.

Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, et al. 2012. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**: 497–501.

Huang X, Lu T, Han B. 2013. Resequencing rice genomes: an emerging new era of rice genomics. *Trends Genet* **29**: 225–232.

Hufford MB, Xu X, van Heerwaarden J, Pyhajarvi T, Chia JM, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM, et al. 2012. Comparative population genomics of maize domestication and improvement. *Nat Genet* **44**: 808–811.

IBSC. 2012. A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**: 711–716.

Immer FR, Henderson MT. 1943. Linkage studies in barley. *Genetics* **28**: 419–440.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. *Nature* **436**: 793–800.

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet* **44**: 226–232.

Islam AR, Shepherd KW. 2000. Isolation of a fertile wheat-barley addition line carrying the entire barley chromosome 1H. *Euphytica* **111**: 145–149.

Istrail S, Sutton GG, Florea L, Halpern AL, Mobarry CM, Lippert R, Walenz B, Shatkay H, Dew I, Miller JR, et al. 2004. Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc Natl Acad Sci USA* **101**: 1916–1921.

Jacobs J, Van Eck H, Arens P, Verkerk-Bakker B, te Lintel Hekkert B, Bastiaanssen H, El-Kharbotly A, Pereira A, Jacobsen E, Stiekema W. 1995.

A genetic map of potato (solanum tuberosum) integrating molecular markers, including transposons, and classical markers. *Theor Appl Genet* **91**: 289–300.

Jacquemin J, Bhatia D, Singh K, Wing RA. 2013. The International Oryza Map Alignment Project: development of a genus-wide comparative genomics platform to help solve the 9 billion-people question. *Curr Opin Plant Biol* **16**: 147–156.

Jain M, Misra G, Patel RK, Priya P, Jhanwar S, Khan AW, Shah N, Singh VK, Garg R, Jeena G, et al. 2013. A draft genome sequence of the pulse crop chickpea (Cicer arietinum L.). *Plant J* **74**: 715–729.

Jansen J, De Jong A, Van Ooijen J. 2001. Constructing dense genetic linkage maps. *Theor Appl Genet* **102**: 1113–1122.

Janssens FA. 1909. The chiasmatype theory. A new interpretation of the maturation divisions. *Cellule* **25**: 389–411.

Jasinska AJ, Service S, Levinson M, Slaten E, Lee O, Sobel E, Fairbanks LA, Bailey JN, Jorgensen MJ, Breidenthal SE, et al. 2007. A genetic linkage map of the vervet monkey (Chlorocebus aethiops sabaeus). *Mamm Genome* **18**: 347–360.

Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X, et al. 2013. Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**: 91–95.

Julier B, Flajoulot S, Barre P, Cardinet G, Santoni S, Huguet T, Huyghe C. 2003. Construction of two genetic linkage maps in cultivated tetraploid alfalfa (Medicago sativa) using microsatellite and AFLP markers. *BMC Plant Biol* **3**: 9.

Kalavacharla V, Hossain K, Gu Y, Riera-Lizarazu O, Vales MI, Bhamidimarri S, Gonzalez-Hernandez JL, Maan SS, Kianian SF. 2006. High-resolution radiation hybrid map of wheat chromosome 1D. *Genetics* **173**: 1089–1099.

Kasha KJ, Kao KN. 1970. High frequency haploid production in barley (Hordeum vulgare L.). *Nature* **225**: 874–876.

Kececioglu JD, Myers EW. 1995. Combinatorial algorithms for dna sequence assembly. *Algorithmica* **13**: 7–51.

Kelley DR, Schatz MC, Salzberg SL. 2010. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol* **11**: R116.

Kelley JM, Field CE, Craven MB, Bocskai D, Kim UJ, Rounsley SD, Adams MD. 1999. High throughput direct end sequencing of BAC clones. *Nucleic Acids Res* **27**: 1539–1546.

Kent WJ, Haussler D. 2001. Assembly of the working draft of the human genome with GigAssembler. *Genome Res* **11**: 1541–1548.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996–1006.

Kircher M, Kelso J. 2010. High-throughput DNA sequencing–concepts and limitations. *Bioessays* **32**: 524–536.

Klein AM, Vaissiere BE, Cane JH, Steffan-Dewenter I, Cunningham SA, Kremen C, Tscharntke T. 2007. Importance of pollinators in changing landscapes for world crops. *Proc Biol Sci* **274**: 303–313.

Komatsuda T, Pourkheirandish M, He C, Azhaguvel P, Kanamori H, Perovic D, Stein N, Graner A, Wicker T, Tagiri A, et al. 2007. Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *Proc Natl Acad Sci USA* **104**: 1424–1429.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, et al. 2002. A high-resolution recombination map of the human genome. *Nat Genet* **31**: 241–247.

Korzun V, Malyshev S, Voylokov A, Börner A. 2001. A genetic map of rye (Secale cereale L.) combining RFLP, isozyme, protein, microsatellite and gene loci. *Theor Appl Genet* **102**: 709–717.

Kosambi D. 1943. The estimation of map distances from recombination values. *Ann Eugenics* **12**: 172–175.

Krasileva KV, Buffalo V, Bailey P, Pearce S, Ayling S, Tabbita F, Soria M, Wang S, Consortium I, Akhunov E, et al. 2013. Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biol* **14**: R66.

Kumar S, Blaxter ML. 2010. Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics* **11**: 571.

Künzel G. 1982. Differences between genetic and physical centromere distances in the case of two genes for male sterility in barley. *Theor Appl Genet* **64**: 25–29.

Kurtz S, Narechania A, Stein JC, Ware D. 2008. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* **9**: 517.

Künzel G, Korzun L, Meister A. 2000. Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. *Genetics* **154**: 397–412.

Lachance J, Tishkoff SA. 2013. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *Bioessays* **35**: 780–786.

Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M, et al. 2012a. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat Biotechnol* **30**: 771–776.

Lam HY, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, O'Huallachain M, Gerstein MB, Kidd JM, Bustamante CD, et al. 2012b. Detecting and annotating genetic variations using the HugeSeq pipeline. *Nat Biotechnol* **30**: 226–229.

Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**: 231–239.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Larsson SJ, Lipka AE, Buckler ES. 2013. Lessons from Dwarf8 on the strengths and weaknesses of structured association mapping. *PLoS Genet* **9**: e1003246.

Lerceteau-Kohler E, Guerin G, Laigret F, Denoyes-Rothan B. 2003. Characterization of mixed disomic and polysomic inheritance in the octoploid strawberry (Fragaria x ananassa) using AFLP mapping. *Theor Appl Genet* **107**: 619–628.

Li H. 2011a. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987–2993.

Li H. 2011b. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**: 718–719.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589–595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**: 1851–1858.

Li J, Das K, Liu J, Fu G, Li Y, Tobias C, Wu R. 2012. Statistical models for genetic mapping in polyploids: challenges and opportunities. *Methods Mol Biol* **871**: 245–261.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, et al. 2010a. The sequence and de novo assembly of the giant panda genome. *Nature* **463**: 311–317.

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J. 2009b. SNP detection for massively parallel whole-genome resequencing. *Genome Res* **19**: 1124–1132.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, et al. 2010b. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**: 265–272.

Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H, Albrechtsen A, Andersen G, Cao H, Korneliussen T, et al. 2010c. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* **42**: 969–972.

Lindsay J, Salooti H, Zelikovsky A, Măndoiu I. 2012. Scalable genome scaffolding using integer linear programming. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 377–383. ACM.

Ling HQ, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y, et al. 2013. Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature* **496**: 87–90.

Liu S, Ying K, Yeh CT, Yang J, Swanson-Wagner R, Wu W, Richmond T, Gerhardt DJ, Lai J, Springer N, et al. 2012a. Changes in genome content generated via segregation of non-allelic homologs. *Plant J* **72**: 390–399.

Liu S, Ying K, Yeh CT, Yang J, Swanson-Wagner R, Wu W, Richmond T, Gerhardt DJ, Lai J, Springer N, et al. 2012b. Changes in genome content generated via segregation of non-allelic homologs. *Plant J* **72**: 390–399.

Lorieux M. 2012. Mapdisto: fast and efficient computation of genetic linkage maps. *Mol Breeding* **30**: 1231–1235.

Lucas SJ, Akpinar BA, Kantar M, Weinstein Z, Aydinoglu F, Safar J, Simkova H, Frenkel Z, Korol A, Magni F, et al. 2013. Physical mapping integrated with syntenic analysis to characterize the gene space of the long arm of wheat chromosome 1A. *PLoS ONE* **8**: e59542.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. *Nat Rev Genet* **4**: 981–994.

Luo MC, Gu YQ, You FM, Deal KR, Ma Y, Hu Y, Huo N, Wang Y, Wang J, Chen S, et al. 2013. A 4-gigabase physical map unlocks the structure and evolution of the complex genome of Aegilops tauschii, the wheat D-genome progenitor. *Proc Natl Acad Sci USA* **110**: 7940–7945.

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J. 2003. High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* **82**: 378–389.

Luo ZW, Zhang RM, Kearsey MJ. 2004. Theoretical basis for genetic linkage analysis in autotetraploid species. *Proc Natl Acad Sci USA* **101**: 7040–7045.

Luo ZW, Zhang Z, Leach L, Zhang RM, Bradshaw JE, Kearsey MJ. 2006. Constructing genetic linkage maps under a tetrasomic model. *Genetics* **172**: 2635–2645.

Maliepaard C, Jansen J, Van Ooijen J. 1997. Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. *Genet Res* **70**: 237–250.

Manickavelu A, Kawaura K, Imamura H, Mori M, Ogihara Y. 2011. Molecular mapping of quantitative trait loci for domestication traits and β-glucan content in a wheat recombinant inbred line population. *Euphytica* **177**: 179–190.

Mao L, Wood TC, Yu Y, Budiman MA, Tomkins J, Woo S, Sasinowski M, Presting G, Frisch D, Goff S, et al. 2000. Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. *Genome Res* **10**: 982–990.

Margarido GR, Souza AP, Garcia AA. 2007. OneMap: software for genetic mapping in outcrossing species. *Hereditas* **144**: 78–79.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–380.

Martienssen RA, Rabinowicz PD, O'Shaughnessy A, McCombie WR. 2004. Sequencing the maize genome. *Curr Opin Plant Biol* **7**: 102–107.

Martis MM, Klemme S, Banaei-Moghaddam AM, Blattner FR, Macas J, Schmutzer T, Scholz U, Gundlach H, Wicker T, Simkova H, et al. 2012. Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proc Natl Acad Sci USA* **109**: 13343–13346.

Martis MM, Zhou R, Haseneyer G, Schmutzer T, Vrana J, Kubalakova M, Konig S, Kugler KG, Scholz U, Hackauf B, et al. 2013. Reticulate Evolution of the Rye Genome. *Plant Cell* **in press**.

Mascher M, Muehlbauer GJ, Rokhsar DS, Chapman J, Schmutz J, Barry K, Munoz-Amatriain M, Close TJ, Wise RP, Schulman AH, et al. 2013a. Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J* **in press**.

Mascher M, Richmond TA, Gerhardt DJ, Himmelbach A, Clissold L, Sampath D, Ayling S, Steuernagel B, Pfeifer M, D'Ascenzo M, et al. 2013b. Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond. *Plant J* **76**: 494–505.

Mascher M, Wu S, St Amand P, Stein N, Poland J. 2013c. Application of Genotyping-by-Sequencing on Semiconductor Sequencing Platforms: A Comparison of Genetic and Reference-Based Marker Ordering in Barley. *PLoS ONE* **8**: e76925.

Masoudi-Nejad A, Nasuda S, Bihoreau MT, Waugh R, Endo TR. 2005. An alternative to radiation hybrid mapping for large-scale genome analysis in barley. *Mol Genet Genomics* **274**: 589–594.

Matsumoto T, Tanaka T, Sakai H, Amano N, Kanamori H, Kurita K, Kikuta A, Kamiya K, Yamamoto M, Ikawa H, et al. 2011. Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Physiol* **156**: 20–28.

Matthaei JH, Jones OW, Martin RG, Nirenberg MW. 1962. Characteristics and composition of RNA coding units. *Proc Natl Acad Sci USA* **48**: 666–677.

Mayer KF, Martis M, Hedley PE, Simkova H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H, et al. 2011. Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* **23**: 1249–1263.

Mayer KF, Taudien S, Martis M, Simkova H, Suchankova P, Gundlach H, Wicker T, Petzold A, Felder M, Steuernagel B, et al. 2009. Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol* **151**: 496–505.

Medvedev P, Georgiou K, Myers G, Brudno M. 2007. Computability of models for sequence assembly. In *Algorithms in Bioinformatics*, pages 289–301. Springer.

Mellersh CS, Langston AA, Acland GM, Fleming MA, Ray K, Wiegand NA, Francisco LV, Gibbs M, Aguirre GD, Ostrander EA. 1997. A linkage map of the canine genome. *Genomics* **46**: 326–336.

Mendel G. 1866. Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn* **44**.

Menotti-Raymond M, David VA, Lyons LA, Schaffer AA, Tomlin JF, Hutton MK, O'Brien SJ. 1999. A genetic linkage map of microsatellites in the domestic cat (Felis catus). *Genomics* **57**: 9–23.

Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund CA, Mayer KF, et al. 2004. Sequence composition and genome organization of maize. *Proc Natl Acad Sci USA* **101**: 14349–14354.

Milczarski P, Bolibok-Brągoszewska H, Myśków B, Stojałowski S, Heller-Uszyńska K, Góralska M, Brągoszewski P, Uszyński G, Kilian A, Rakoczy-Trojanowska M. 2011. A high density consensus map of rye (Secale cereale L.) based on DArT markers. *PLoS One* **6**: e28495.

Mizuno N, Nitta M, Sato K, Nasuda S. 2012. A wheat homologue of PHYTOCLOCK 1 is a candidate gene conferring the early heading phenotype to einkorn wheat. *Genes Genet Syst* **87**: 357–367.

Moore G, Devos KM, Wang Z, Gale MD. 1995. Cereal genome evolution. Grasses, line up and form a circle. *Curr Biol* **5**: 737–739.

Morgan TH. 1910. Chromosomes and heredity. *Am Nat* **44**: 449–496.

Morgan TH, Sturtevant AH, Muller HJ, Bridges CB. 1922. *The mechanism of Mendelian heredity.* Heny Holt, New York.

Moroldo M, Paillard S, Marconi R, Fabrice L, Canaguier A, Cruaud C, De Berardinis V, Guichard C, Brunaud V, Le Clainche I, et al. 2008. A physical map of the heterozygous grapevine 'Cabernet Sauvignon' allows mapping candidate genes for disease resistance. *BMC Plant Biol* **8**: 66.

Morrell PL, Toleno DM, Lundy KE, Clegg MT. 2005. Low levels of linkage disequilibrium in wild barley (Hordeum vulgare ssp. spontaneum) despite high rates of self-fertilization. *Proc Natl Acad Sci USA* **102**: 2442–2447.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.

Mozo T, Dewar K, Dunn P, Ecker JR, Fischer S, Kloska S, Lehrach H, Marra M, Martienssen R, Meier-Ewert S, et al. 1999. A complete BAC-based physical map of the Arabidopsis thaliana genome. *Nat Genet* **22**: 271–275.

Muñoz Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, Ariyadasa R, Spannagl M, Nussbaumer T, et al. 2013. Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biol* **14**: R58.

Muñoz-Amatriaín M, Moscou MJ, Bhat PR, Svensson JT, Bartoš J, Suchánková P, Šimková H, Endo TR, Fenton RD, Lonardi S, et al. 2011.

An improved consensus linkage map of barley based on flow-sorted chromosomes and single nucleotide polymorphism markers. *Plant Genome* **4**: 238–249.

Nagy ED, Guo Y, Tang S, Bowers JE, Okashah RA, Taylor CA, Zhang D, Khanal S, Heesacker AF, Khalilian N, et al. 2012. A high-density genetic map of Arachis duranensis, a diploid ancestor of cultivated peanut. *BMC Genomics* **13**: 469.

Nelson JC, Deynze AE, Sorrells ME, Autrique E, Lu YH, Negre S, Bernard M, Leroy P. 1995. Molecular mapping of wheat. Homoeologous group 3. *Genome* **38**: 525–533.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12**: 443–451.

Nilsson NH, Lundqvist A. 1960. The origin of self-compatibility in rye. *Hereditas* **46**: 1–19.

Nuzhdin SV, Pasyukova EG, Dilda CL, Zeng ZB, Mackay TF. 1997. Sex-specific quantitative trait loci affecting longevity in Drosophila melanogaster. *Proc Natl Acad Sci USA* **94**: 9734–9739.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**: 579–584.

O'Hanlon P, Peakall R, Briese D, et al. 2000. A review of new PCR-based genetic markers and their utility to weed ecology. *Weed Res* **40**: 239–254.

Ohno S. 1973. Ancient linkage groups and frozen accidents. *Nature* **244**: 259–262.

Okagaki RJ, Cho S, Kruger WM, Xu WW, Heinen S, Muehlbauer GJ. 2012. The barley UNICULM2 gene resides in a centromeric region and may be associated with signaling and stress responses. *Funct Integr Genomics* **13**: 33–41.

Olsen FL. 1987. Induction of microspore embryogenesis in cultured anthers of hordeum vulgare. The effects of ammonium nitrate, glutamine and asparagine as nitrogen sources. *Carlsberg Res Commun* **52**: 393–404.

Olson MV. 1993. The human genome project. *Proc Natl Acad Sci USA* **90**: 4338–4344.

Olson MV, Dutchik JE, Graham MY, Brodeur GM, Helms C, Frank M, MacCollin M, Scheinman R, Frank T. 1986. Random-clone strategy for genomic restriction mapping in yeast. *Proc Natl Acad Sci USA* **83**: 7826–7830.

O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, et al. 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* **5**: 28.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu Rev Genet* **34**: 401–437.

Pääbo S. 2001. Genomics and society. The human genome and our view of ourselves. *Science* **291**: 1219–1220.

Paran I, Goldman I, Tanksley S, Zamir D. 1995. Recombinant inbred lines for genetic mapping in tomato. *Theor Appl Genet* **90**: 542–548.

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W, et al. 2008. A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* **322**: 101–104.

Pevzner PA, Tang H. 2001. Fragment assembly with double-barreled data. *Bioinformatics* **17 Suppl 1**: S225–233.

Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci USA* **98**: 9748–9753.

Pfeifer M, Martis M, Asp T, Mayer KF, Lubberstedt T, Byrne S, Frei U, Studer B. 2013. The perennial ryegrass GenomeZipper: targeted use of genome resources for comparative grass genomics. *Plant Physiol* **161**: 571–582.

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, et al. 2012a. Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* **5**: 103–113.

Poland JA, Brown PJ, Sorrells ME, Jannink JL. 2012b. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE* **7**: e32253.

Poland JA, Rife TW. 2012. Genotyping-by-sequencing for plant breeding and genetics. *Plant Gen* **5**: 92–102.

Potato Genome Sequencing Consortium. 2011. Genome sequence and analysis of the tuber crop potato. *Nature* **475**: 189–195.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature* **499**: 471–475.

Pravenec M, Gauguier D, Schott JJ, Buard J, Křen V, Bila V, Szpirer C, Szpirer J, Wang JM, Huang H, et al. 1996. A genetic linkage map of the rat derived from recombinant inbred strains. *Mamm Genome* **7**: 117–127.

Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69**: 1–14.

Rastas P, Paulin L, Hanski I, Lehtonen R, Auvinen P. 2013. Lep-MAP: fast and accurate linkage map construction for large SNP datasets. *Bioinformatics* **in press**.

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci USA* **98**: 11479–11484.

Ritter E, Gebhardt C, Salamini F. 1990. Estimation of recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents. *Genetics* **125**: 645–654.

Roach JC, Boysen C, Wang K, Hood L. 1995. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics* **26**: 345–353.

Roberts A, Pimentel H, Trapnell C, Pachter L. 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**: 2325–2329.

Rogers J, Garcia R, Shelledy W, Kaplan J, Arya A, Johnson Z, Bergstrom M, Novakowski L, Nair P, Vinson A, et al. 2006. An initial genetic linkage map of the rhesus macaque (Macaca mulatta) genome using human microsatellite loci. *Genomics* **87**: 30–38.

Rogers J, Mahaney MC, Witte SM, Nair S, Newman D, Wedel S, Rodriguez LA, Rice KS, Slifer SH, Perelygin A, et al. 2000. A genetic linkage map of the baboon (Papio hamadryas) genome based on human microsatellite polymorphisms. *Genomics* **67**: 237–247.

Ronaghi M, Uhlen M, Nyren P. 1998. A sequencing method based on real-time pyrophosphate. *Science* **281**: 363, 365.

Sadder T, Weber G. 2002. Comparison between genetic and physical maps in Zea mays L. of molecular markers linked to resistance against Diatraea spp. *Theor Appl Genet* **104**: 908–915.

Saintenac C, Jiang D, Wang S, Akhunov E. 2013. Sequence-based mapping of the polyploid wheat genome. *G3 (Bethesda)* **3**: 1105–1114.

Sakamoto T, Danzmann RG, Gharbi K, Howard P, Ozaki A, Khoo SK, Woram RA, Okamoto N, Ferguson MM, Holm LE, et al. 2000. A microsatellite linkage map of rainbow trout (Oncorhynchus mykiss) characterized by large sex-specific differences in recombination rates. *Genetics* **155**: 1331–1345.

Salmon A, Udall JA, Jeddeloh JA, Wendel J. 2012. Targeted capture of homoeologous coding and noncoding sequence in polyploid cotton. *G3 (Bethesda)* **2**: 921–930.

Salvo-Garrido H, Laurie D, Jaffe B, Snape J. 2001. An RFLP map of diploid Hordeum bulbosum L. and comparison with maps of barley (H. vulgare L.) and wheat (Triticum aestivum L.). *Theor Appl Genet* **103**: 869–880.

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, et al. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res* **22**: 557–567.

Sanei M, Pickering R, Kumke K, Nasuda S, Houben A. 2011. Loss of centromeric histone H3 (CENH3) from centromeres precedes uniparental chromosome elimination in interspecific barley hybrids. *Proc Natl Acad Sci USA* **108**: 498–505.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M. 1977a. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**: 687–695.

Sanger F, Nicklen S, Coulson AR. 1977b. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74**: 5463–5467.

Sato K, Shin-I T, Seki M, Shinozaki K, Yoshida H, Takeda K, Yamazaki Y, Conte M, Kohara Y. 2009. Development of 5006 full-length CDNAs in barley: a tool for accessing cereal genomics resources. *DNA Res* **16**: 81–89.

Schadt EE, Turner S, Kasarskis A. 2010. A window into third-generation sequencing. *Hum Mol Genet* **19**: R227–240.

Schatz MC, Langmead B, Salzberg SL. 2010. Cloud computing and the DNA data race. *Nat Biotechnol* **28**: 691–693.

Schiex T, Gaspin C. 1997. CARTHAGENE: constructing and joining maximum likelihood genetic maps. *Proc Int Conf Intell Syst Mol Biol* **5**: 258–267.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112–1115.

Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jorgensen JE, Weigel D, Andersen SU. 2009. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat Methods* **6**: 550–551.

Schneeberger K, Weigel D. 2011. Fast-forward genetics enabled by new sequencing technologies. *Trends Plant Sci* **16**: 282–288.

Schneider K. 2005. Mapping populations and principles of genetic mapping. In *The handbook of plant genome mapping*, pages 3–19. Wiley-VCH Verlag.

Schulte D, Ariyadasa R, Shi B, Fleury D, Saski C, Atkins M, deJong P, Wu CC, Graner A, Langridge P, et al. 2011. BAC library resources for map-based cloning and physical map construction in barley (Hordeum vulgare L.). *BMC Genomics* **12**: 247.

Schulte D, Close TJ, Graner A, Langridge P, Matsumoto T, Muehlbauer G, Sato K, Schulman AH, Waugh R, Wise RP, et al. 2009. The International Barley Sequencing Consortium – At the threshold of efficient access to the barley genome. *Plant Physiol* **149**: 142–147.

Semagn K, Bjørnstad Å, Ndjiondjop M. 2006. Principles, requirements and prospects of genetic mapping in plants. *Afr J Biotechnol* **5**: 2569–2587.

Shahinnia F, Druka A, Franckowiak J, Morgante M, Waugh R, Stein N. 2012. High resolution mapping of Dense spike-ar (dsp.ar) to the genetic centromere of barley chromosome 7H. *Theor Appl Genet* **124**: 373–384.

Shatalina M, Wicker T, Buchmann JP, Oberhaensli S, Simkova H, Dolezel J, Keller B. 2013. Genotype-specific SNP map based on whole chromosome 3B sequence information from wheat cultivars Arina and Forno. *Plant Biotechnol J* **11**: 23–32.

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, Simon M. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc Natl Acad Sci USA* **89**: 8794–8797.

Soderlund C, Longden I, Mott R. 1997. FPC: a system for building contigs from restriction fingerprinted clones. *Comput Appl Biosci* **13**: 523–535.

Sousa V, Hey J. 2013. Understanding the origin of species with genome-scale data: modelling gene flow. *Nat Rev Genet* **14**: 404–414.

Staden R. 1979. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* **6**: 2601–2610.

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

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**: i215–225.

Stapley J, Reger J, Feulner PG, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP, Slate J. 2010. Adaptation genomics: the next generation. *Trends Ecol Evol* **25**: 705–712.

Steuernagel B, Taudien S, Gundlach H, Seidel M, Ariyadasa R, Schulte D, Petzold A, Felder M, Graner A, Scholz U, et al. 2009. De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. *BMC Genomics* **10**: 547.

Stift M, Berenos C, Kuperus P, van Tienderen PH. 2008. Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: a general procedure applied to Rorippa (yellow cress) microsatellite data. *Genetics* **179**: 2113–2123.

Studer B, Byrne S, Nielsen RO, Panitz F, Bendixen C, Islam MS, Pfeifer M, Lubberstedt T, Asp T. 2012. A transcriptome map of perennial ryegrass (Lolium perenne L.). *BMC Genomics* **13**: 140.

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

Swank RT, Bailey DW. 1973. Recombinant inbred lines: value in the genetic analysis of biochemical variants. *Science* **181**: 1249–1252.

Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, Uemura A, Utsushi H, Tamiru M, Takuno S, et al. 2013a. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J* **74**: 174–183.

Takagi H, Uemura A, Yaegashi H, Tamiru M, Abe A, Mitsuoka C, Utsushi H, Natsume S, Kanzaki H, Matsumura H, et al. 2013b. MutMap-Gap: whole-genome resequencing of mutant F2 progeny bulk combined with de novo assembly of gap regions identifies the rice blast resistance gene Pii. *New Phytol* **200**: 276–283.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and collinearity in plant genomes. *Science* **320**: 486–488.

Taudien S, Steuernagel B, Ariyadasa R, Schulte D, Schmutzer T, Groth M, Felder M, Petzold A, Scholz U, Mayer KF, et al. 2011. Sequencing of BAC pools by different next generation sequencing platforms and strategies. *BMC Res Notes* **4**: 411.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**: 64–69.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

The International Brachypodium Initiative. 2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463**: 763–768.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–861.

The International Human Genome Mapping Consortium. 2001. A physical map of the human genome. *Nature* **409**: 934–941.

Trachtulec Z, Forejt J. 2001. Synteny of orthologous genes conserved in mammals, snake, fly, nematode, and fission yeast. *Mamm Genome* **12**: 227–231.

Trapnell C, Salzberg SL. 2009. How to map billions of short reads onto genomes. *Nat Biotechnol* **27**: 455–457.

Truco MJ, Ashrafi H, Kozik A, van Leeuwen H, Bowers J, Reyes Chin Wo S, Stoffel K, Xu H, Hill T, Van Deynze A, et al. 2013. An Ultra High-Density, Transcript-Based, Genetic Map of Lettuce. *G3 (Bethesda)* **3**: 617–631.

Truong HT, Ramos AM, Yalcin F, de Ruiter M, van der Poel HJ, Huvenaars KH, Hogers RC, van Enckevort LJ, Janssen A, van Orsouw NJ, et al. 2012. Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLoS ONE* **7**: e37565.

van Hintum T, Menting F. 2003. Diversity in ex situ genebank collections of barley. In R von Bother, T van Hintum, H Knupffer, K Sato, editors, *Diversity in Barley (Hordeum vulgare)*, pages 247–257. Elsevier Science B.V., Amsterdam.

van Oeveren J, de Ruiter M, Jesse T, van der Poel H, Tang J, Yalcin F, Janssen A, Volpin H, Stormo KE, Bogden R, et al. 2011. Sequence-based physical mapping of complex genomes by whole genome profiling. *Genome Res* **21**: 618–625.

Varshney RK, Close TJ, Singh NK, Hoisington DA, Cook DR. 2009. Orphan legume crops enter the genomics era! *Curr Opin Plant Biol* **12**: 202–210.

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J, et al. 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2**: e1326.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. *Science* **291**: 1304–1351.

Vitulo N, Albiero A, Forcato C, Campagna D, Dal Pero F, Bagnaresi P, Colaiacovo M, Faccioli P, Lamontanara A, Simkova H, et al. 2011. First survey of the wheat chromosome 5A composition through a next generation sequencing approach. *PLoS ONE* **6**: e26421.

Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Koh W, Passarelli B, Fan HC, Mantalas GL, Palmeri KJ, et al. 2013. The genome sequence of the colonial chordate, Botryllus schlosseri. *Elife* **2**: e00569.

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, et al. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077–1082.

Waterston RH, Lander ES, Sulston JE. 2002. On the sequencing of the human genome. *Proc Natl Acad Sci USA* **99**: 3712–3716.

Watson JD, Crick FH. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**: 737–738.

Weber JL, Myers EW. 1997. Human whole-genome shotgun sequencing. *Genome Res* **7**: 401–409.

Wei F, Coe E, Nelson W, Bharti AK, Engler F, Butler E, Kim H, Goicoechea JL, Chen M, Lee S, et al. 2007. Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet* **3**: e123.

Wei F, Zhang J, Zhou S, He R, Schaeffer M, Collura K, Kudrna D, Faga BP, Wissotski M, Golser W, et al. 2009. The physical and genetic framework of the maize B73 genome. *PLoS Genet* **5**: e1000715.

Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H. 2011. Snver: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res* **39**: e132–e132.

Weigel D, Mott R. 2009. The 1001 genomes project for Arabidopsis thaliana. *Genome Biol* **10**: 107.

Wertheim B, Beukeboom LW, van de Zande L. 2013. Polyploidy in animals: effects of gene expression on sex determination, evolution and ecology. *Cytogenet Genome Res* **140**: 256–269.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872–876.

Wicker T, Mayer KF, Gundlach H, Martis M, Steuernagel B, Scholz U, Simkova H, Kubalakova M, Choulet F, Taudien S, et al. 2011. Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *Plant Cell* **23**: 1706–1718.

Wicker T, Narechania A, Sabot F, Stein J, Vu GT, Graner A, Ware D, Stein N. 2008. Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *BMC Genomics* **9**: 518.

Wilkins P, Thorogood D. 1992. Breakdown of self-incompatibility in perennial ryegrass at high temperature and its uses in breeding. *Euphytica* **64**: 65–69.

Williams RW, Gu J, Qi S, Lu L. 2001. The genetic structure of recombinant inbred mice: high-resolution consensus maps for complex trait analysis. *Genome Biol* **2**: R46.

Wolfe R, Franckowiak J, et al. 1990. Multiple dominant and recessive genetic marker stocks in spring barley. *Barley Genet Newslett* **20**: 117–121.

Wu R, Ma CX, Painter I, Zeng ZB. 2002. Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. *Theor Popul Biol* **61**: 349–363.

Wu Y, Bhat PR, Close TJ, Lonardi S. 2008. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet* **4**: e1000212.

Wu Y, Close TJ, Lonardi S. 2011. Accurate construction of consensus genetic maps via integer linear programming. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* **8**: 381–394.

Xie W, Feng Q, Yu H, Huang X, Zhao Q, Xing Y, Yu S, Han B, Zhang Q. 2010. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proc Natl Acad Sci USA* **107**: 10578–10583.

Xu Q, Chen LL, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, Jiao WB, Hao BH, Lyon MP, et al. 2012. The draft genome of sweet orange (Citrus sinensis). *Nat Genet* **45**: 59–66.

Yamada T. 2001. Introduction of a self-compatible gene of Lolium temulentum L. to perennial ryegrass (Lolium perenne L.) for the purpose of the production of inbred lines of perennial ryegrass. *Euphytica* **122**: 213–217.

Yap IV, Schneider D, Kleinberg J, Matthews D, Cartinhour S, McCouch SR. 2003. A graph-theoretic approach to comparing and integrating genetic, physical and sequence-based maps. *Genetics* **165**: 2235–2247.

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* **296**: 79–92.

Yuan Q, Liang F, Hsiao J, Zismann V, Benito MI, Quackenbush J, Wing R, Buell R. 2000. Anchoring of rice BAC clones to the rice genetic map in silico. *Nucleic Acids Res* **28**: 3636–3641.

Zhang J, Guo W, Zhang T. 2002. Molecular linkage map of allotetraploid cotton (Gossypium hirsutum L. × Gossypium barbadense L.) with a haploid population. *Theor, Appl Genet* **105**: 1166–1174.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**: 203–214.

Zhou S, Bechner MC, Place M, Churas CP, Pape L, Leong SA, Runnheim R, Forrest DK, Goldstein S, Livny M, et al. 2007. Validation of rice genome sequence by optical mapping. *BMC Genomics* **8**: 278.

Zhou S, Wei F, Nguyen J, Bechner M, Potamousis K, Goldstein S, Pape L, Mehan MR, Churas C, Pasternak S, et al. 2009. A single molecule scaffold for the maize genome. *PLoS Genet* **5**: e1000711.