Genotyping by sequencing from sparse sequenced genomes representations from bi- and multi- parental mapping population using a HMM approach

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Vipul Kumar Patel

aus Dortmund

Köln

2016

Die vorliegende Arbeit wurde am Max-Planck-Institute für Züchtungsforschung in Köln in der Abteilung für Entwicklungsbiologie der Pflanzen (Direktor Prof. Dr. George Coupland) angefertigt.

Berichterstatter:	Prof. Dr. George Coupland
(Gutachter)	Prof. Dr. Achim Tresch
Prüfungsvorsitzender:	Prof. Dr. Marcel Bucher

Tag der mündlichen Prüfung: 18.April.2016

Abstract

Genotyping is one key element for successfully carrying out molecular breeding, gene network discovery or assessment of genetic diversity. The onset of next generation sequencing has enabled high-resolution genotyping of thousands or millions of markers per individual in one analysis. Such dense information can be used to identify genetic loci associated with a trait of interest. Development of multiplexing allows sequencing of whole populations in a single run, vastly reducing inputs of time and money per sample. This high throughput genotyping is known as genotyping-by-sequencing (GBS). However, there is a trade-off for using GBS, as the total number of reads per run must be distributed across all samples, leading to a reduction of coverage per sample. The distribution of the total reads is currently not uniform, which leads to samples with only partial sequence coverage.

This thesis presents a solution for handling such data by imputing missing markers based on a Hidden Markov Model approaches for bi- and multi- parental mapping populations. The developed methods were not only validated by simulation studies but also applied to several real mapping population datasets. For the bi-parental mapping population, data were derived from three different taxa (*Arabidopsis thaliana*, *Sorghum bicolor* and *Fragaria vesca*) and for the multi-parental mapping population the *Arabidopsis* multi-parental RIL (AMPRIL) population was genotyped. The successful high resolution genotyping of such mapping populations with sparse sequencing data demonstrates the advantages of the developed method and the positive effects for downstream analysis e.g. for quantitative trait analysis or genome-wide-association studies.

This thesis additionally provides a theoretical approach and implementation for a hybrid correction approach of sequencing errors in third generation sequencing data from Pacific Biosciences.

Zusammenfassung:

Die Genotypisierung ist ein wichtiges Verfahren, insbesondere für eine erfolgreiche molekulare Züchtung, bei der Aufdeckung von Gennetzwerken oder der Ermittlung der genetischen Vielfalt einer Population. Besonders durch die Einführung von „Next-generation-sequenzierung“, gelang es Millionen von neuen und unbekannten Markern pro Individuum zu genotypisieren. Die so gewonnene Informationsdichte erlaubt es, eine effektive Analyse der Beziehung zwischen Genen und deren Eigenschaften aufzudecken. Für solche komplizierten Analysen müssen mehrere hundert Individuen sequenziert werden, was einem hohen Investitionsaufwand entspricht. Mit der Einführung von „multiplexing“ wurde es möglich, Individuen gleichzeitig parallel zu sequenzieren und zu genotypisieren. Diese Methode wird als „Genotyping by sequencing“ (GBS) bezeichnet. Sie hat aber den Nachteil, dass nicht alle Individuen gleichmäßig sequenziert werden. Es gibt somit Individuen, deren Genome nur teilweise sequenziert werden. Dies reduziert die Anzahl der Marker, welche genotypisiert werden können.

In dieser Arbeit stellen wir eine Lösung vor welche mit Hilfe eines statistischen Modells, dem „Hidden Markov Model“ fehlende Informationen vorhersagen kann. Es wurden zwei Modele entwickelt für Populationen von zwei oder mehr Eltern. Die entwickelten Methoden wurden mit simulierten Daten getestet und auf tatsächlich vorhandenen Population angewendet: für Populationen generiert aus zwei Eltern (*Arabidopsis thaliana*, *Sorghum bicolor* and *Fragaria vesca*) und für mehrere Eltern, die *Arabidopsis* multi-parental RIL Population. Die Anwendung unserer Methoden auf diese Populationen half, neue Erkenntnisse und Kandidatengene zu finden.

Zusätzlich zum Thema „Genotyping by sequencing“ wird ein Algorithmus behandelt, welcher die Korrektur von langen Sequenzeninformation geeignet ist, die von der Technologie Pacific Bioscience generiert wurden.

Introduction

Individuals share broadly the same DNA sequences, however, many of the homologous loci show different types of DNA sequences which are referred to as alleles. Genotyping is the process of characterizing the specific alleles within an organism, which can be natural differences or artificially induced mutations. Differentiating genotypes allows understanding sources of phenotypic variations. The individual genetic loci with different alleles underlying the variation in a phenotype are defined as quantitative trait loci (QTL). Hence, correlation between phenotypic and genetic variation can be used to identify regions coding for a particular phenotype or trait of interested. Such knowledge than can be used for an effective breeding process or to unravel further downstream genetic networks.
However, complex trait analyses requires populations of more than hundreds or even thousand of individuals. For this amount of individuals applying standard genotyping methods are not practicable and require huge investment. Next-generation-sequencing (NGS) allows to genotype more than thousand markers in a single run allowing to not only by-pass time-intense genotyping efforts but would also allow for the reconstruction of recombination breaks at great resolution. In combination with multiplexing multiple individuals can be sequenced in parallel. This reduces the sequencing cost per individual, but comes at the price of having only partial genomes sequenced.

The aim of this thesis is to offer possible solution for genotyping such individuals where only sparse sequencing data is available. Therefore the introduction is structured in such way that first an introduction to NGS as technology for genotyping of thousand of markers is given. Further an introduction to the concept of Hidden-Markov-Model (HMM) is done, as this machine learning method is used as a possible solution to predict the genotypes of missing markers. Afterwards the following two chapters will cover cases where such model was applied for genotyping bi- and multi-parental individuals.

High-throughout genotyping

Different technological developments introduced different types of molecular markers used for genotyping, starting with restriction fragment length polymorphism (RFLP) (Botstein, White, Skolnick, & Davis, 1980) and followed by other types of PCR-based makers, for example random amplification of polymorphic DNA (RAPD), cleaved amplified polymorphic sequences (CAPS), simple sequence repeats (SSR), and amplified fragment length polymorphisms (AFLPs). Later if became possible to screen whole genomes for SNPs as well as small insertions and deletions. It has been shown that among the other molecular markers SNPs are highly abundant in different genomes as well in crops (Rafalski, 2002; Sonah et al.) and useful for genome wide screens. Nevertheless achieving high-throughput including thousands to millions of SNPs has been only

possible since the introduction of DNA hybridization microarrays and NGS. NGS is based on sequencing millions of reads in a massively parallel high throughput assay. Different NGS

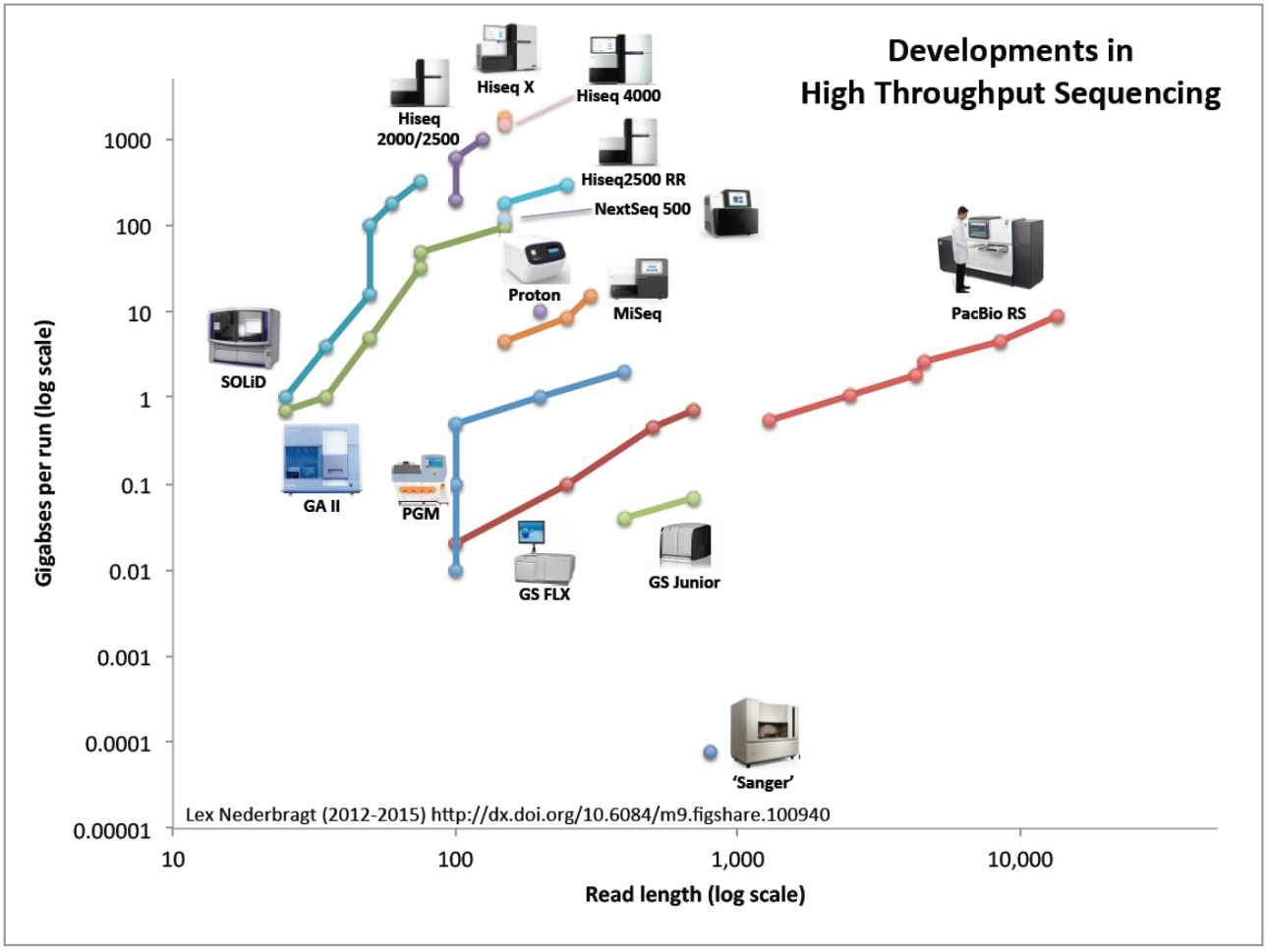


Figure 1. Development of read length and read numbers of different NGS technologies.
The x-axis shows the read length and the y a-axis the number of reads produced. The different colours show the different NGS technologies. The figure was generated by Lex Nederbragt (Nederbragt, 2014)

technologies differ in how the reads are captured, amplified and sequenced (Berglund, Kiialainen, & Syvänen, 2011; Quail et al., 2012). Since the introduction read length and read number are increasing for different technologies (Figure 1).

Figure 1 also shows the appearance of the latest sequencing technology, third generation sequencing, which is based on single molecule sequencing introduced by Pacific Biosciences for generating reads with multiple kb in size. As the majority of the reads analyzed for this thesis were produced using Illumina technology a short description of this particular technique will be given in the next paragraph. The single-molecule sequencing from Pacific Bioscience will be introduced in chapter 3.

Basic concept of Illumina sequencing technology

DNA is extracted, sheared and ligated to adaptor oligos at both ends of each DNA fragment. The adapters contain linker sequences to enable binding to the surface of a so-called flowcell. The flowcell itself is a glass plate containing complementary oligo adaptor sequences fixated on its surface. Ligated molecules bind randomly to the flowcell. After binding of the fragments, a local amplification ("bride amplification") is initiated by adding non-labelled nucleotides producing high-density clusters of the fragment sequences. After several cycles of amplification, sequencing begins by annealing the sequencing primers. Fluorescence-labelled terminator nucleotides are added and incorporated by DNA polymerase, which interrupt polymerase activity after incorporation of individual nucleotides. A laser scanner is used to excite the flourophores of the nucleotides, which then emit a light pulse, which is recorded as an image of the flowcell. The terminator is then enzymatically cleaved out and the next cycle can begin. The image files are converted into sequence data by "basecalling" software (Metzker, 2010).

The Illumina platform also offers paired-end sequencing, where reads are generated from both ends of the fragments. The number of reads increased over time and now allows sequencing of multiple samples on the same flowcell. To reconstruct from which samples reads were derived from a barcoding system was introduced. Those barcodes are short unique nucleotide sequences (commonly around six nucleotides) added to the adapter of each sample, ensuring that each read can be assigned to its source sample (Bystrykh, 2012; Mir, Neuhaus, Bossert, & Schober, 2013; Van der Auwera et al., 2002).

Genotyping using NGS genotyping by sequencing (GBS)

NGS allows screening for thousand of markers in one analysis, which allows identifying allelic variation at high resolutions. As described in the previous section, millions of reads are generated containing short genomic information of the sample. To obtain the allelic differences, the reads have to be transformed into a useful representation. Therefore short reads are aligned to a known reference sequence to order the short reads into a physical representation. The reference sequence is a snapshot of single genome. This approach of aligning read towards a reference genome is known as resequencing. After alignment, the consensus sequence is generated from the aligned short reads for each position of the reference sequence. The difference between the consensus sequence and the reference sequence represent the observed genetic variation from the used genotype compared to the reference sequence.

In general, it is possible to assembly the short reads obtain from the sample into a reference genome and compare the assembled genomes against each other, but this requires a dense sequencing depth and huge computational time where resequencing is cost-effective and a fast method. But it has to be noted as well that resequencing will miss segmental duplications and repeat structures such as active transposons that could differ between genotypes, which could be

available by comparing de-novo assemblies of the genotypes. Genotyping by sequencing (GBS) used the concept of resequencing in a high throughput manner by genotyping for hundred of genomes the genotypes at allelic marker position for each genotype in parallel.

There exists currently two main methods for GBS, the first method based on the whole genome sequencing as described in the previous paragraph and genome complex reduction approaches, i.e. RAD-seq (Baird et al., 2008; Davey et al., 2011; Poland & Rife, 2012). Instead of sequencing fragments produced by random fragmentation of the sample DNA, RAD-seq sequences fragments based on restriction enzyme digestion. Fragments selected for sequencing will have well defined sequence pattern reflecting the restriction recognition site of the restriction enzyme. RAD-seq has two advantages as it reduces the complexity of the genome and allows for high coverage rate at the recognition sites, which allows accurate SNP genotyping. The drawback is the reduction of the resolution compared to whole-genome sequencing as only a limited number of markers can be obtained. To increase the resolution imputation can be applied.
Both methods (whole genome sequencing and complex reductions) are using multiplexing which is highly cost efficient, multiplexing a large number of samples results in low amounts of sequencing data for each of the samples. This can complicate the use of these data for genotyping.

Though GBS is generally a simple concept specific details complicate this simplicity including how to handle heterozygous position, what reads are reliably aligned or how to handle genomes which are not diploid? But the benefit of using NGS for genotyping is the flexibility to screen for known and unknown variations. This allows genotyping by assigning the correct allele for each sample without the requirement of any prior knowledge on SNPs and their alleles.

Genotyping based on sparse sequencing data.
For genotyping of recombinants from controlled crosses, first the variants have to be identified describing the differences between the parental lines. For our purpose we used SNPs as variants (markers). To genotype an individual recombinant line high coverage sequencing can be applied, i.e. for *Arabidopsis thaliana* an average coverage of 10x is sufficiently enough (Figure 2A).
As such sequencing depth can be quite costly (in particular for species with large genomes), low-fold (sparse) sequencing is a reasonable compromise.
Besides sequencing costs which typically limit the amount of data per individual, sequencing coverage is not uniformly distributed across multiplexed samples. Low sequencing depth introduces lack of reads at many of the SNP markers and not every SNP will be aligned by the same number of reads in all samples. In general sparse sequencing will produce low-density genotypes (Figure 2B). Generally SNPs with only one or two reads aligned can lead to a false prediction of the genotype through sequencing errors and wrong alignments. To correct and

impute missing genotypes different approaches have been developed. Those approaches make use of high levels of linkage in recombinant genomes allowing imputing missing informations.

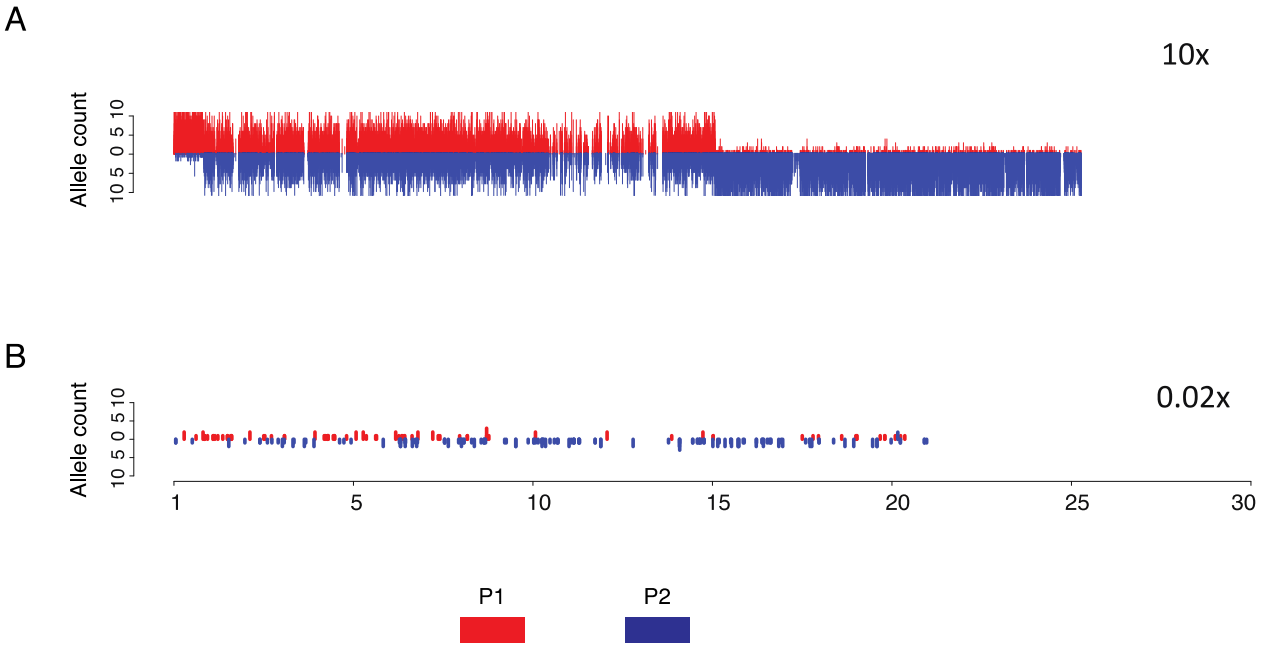


Figure 2. Observed allele support (y axis) for the two parental alleles P1 (red) and P2 (blue) on a genomic region (x axis, in Mb). A) Deeply sequenced sample allows for screening for CO and to identify the correct genotype at each marker position. B) In contrast, sparsely sequenced individuals make it difficult to identify COs and the correct genotypes for each marker position. To genotype such samples, we have to apply more sophisticated methods.

Imputation methods can be divided into two types of approaches, studies of direct related or unrelated individuals. Imputing missing genotypes in recombinant individuals derived from controlled crossed (typically even with known parental genomes) relies on long haplotypes. Identifying such haplotype blocks can be used to impute missing genotypes as each marker in one haplotype block is linked to the same parental genotypes. Imputing natural accessions, e.g. selected in different countries, relies on the ancestral haplotypes segregating in such populations. Such methods have been well studied in the field of human resequencing and genome-wide association studies (Marchini & Howie, 2010).

Here we present two approaches for imputing genotypes from recombinant genomes: sliding window and Hidden-Markov Model (HMM).

Imputation using a simple sliding window approach

Huang et al., 2009 published a sliding window approach for genotyping 150 RILs from a bi-parental population derived from a cross of *indica* and *japonica* rice lines. The average coverage per sample was 0.02x. They applied a sliding window of 15 SNPs per window over 1,226,791 SNPs (3.2 SNPs/kb), counting only informative SNPs labeled by their support for *indica* or *japonica* accessions. The ratio between SNPs supporting *indica* or *japonica* is used to determine the underlying genotype. Given the window size, the expected probability for a certain genotype can

be calculated given the frequency of observed genotypes in that window. The thresholds for assigning genotypes are dependent on the window size. If the window size has to be modified, e.g. the window size has to be increased, as the marker density is too low, the expected probabilities have to be changed. Additionally the threshold for identifying heterozygous regions is problematic to estimate, as it has been defined as a high up and down in a short range. In general, the benefits of a sliding window approach includes that it is quite simple to apply and fast, however sliding window approaches have a low resolution as not each marker is individually imputed.

Hidden-Markov Model (HMM)

To increase the genotype resolution (and accuracy), imputation based on HMM was introduced (Andolfatto et al., 2011; Xie et al., 2010).

A HMM is an extended version of a Markov chain. A Markov chain is a statistical model predicting a future event given the knowledge of previous experiences. The Markov chain can be described as a chain of states $S = \{s_1, s_2, s_3, s_4, \dots, s_n\}$. Each state is representing an observable event. The probability $p_{ij}, i, j = \{1 \dots n\}$ describes the probability of observing an event s_j after observing s_i $i, j = \{1 \dots n\}$. The set of all possible p_{ij} is known as transition matrix.

A Markov chain of order 1 is a model consisting of a finite set of states $S = \{s_1, s_2, s_3, s_4, \dots, s_n\}$ and a transition matrix $T = \{p_{ij}\}, i, j = \{1 \dots n\}, where\ \sum_{i,j}^{n} p_{i,j} = 1$. And for all $s \in S$ the probability of the transition $s_i \rightarrow s_j$ is determined by $P(s_{x+1} = j | s_x = i) = \{p_{ij}\}$, where x represents time.

A Hidden-Markov model (HMM) extends the Markov chain, where observation events are not representing the states. The states are hidden and can only be predicted based on the observation. A HMM is defined by an alphabet S, a set of states Q, a matrix $A = \{p_{ij}\} for\ i, j\ \in Q$, emission probability $e_k(b)$ for every $k \in Q$ and $b\ \in S$ and an initial starting probability for observing a certain state at the beginning.

By combining the transition and emission probabilities a solution space is defined, where all possible combination can be described and each chain of events can be evaluated by their probability. In other words: Let $\pi = (\pi_1, \pi_2, \dots, \pi_L)$ be a possible path generated by our HMM for a given sequence x = ($x_1, x_2, \dots, x_L$). The probability for $P(x, \pi) = p_{o\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i)\, p_{\pi_i \pi_{i+1}}$ with π_{L+1}. We need to calculate each possible path and take the path with the highest probability (π^*): $\pi^* = max_\pi P(x, \pi)$. To reduce the computational time the Viterbi algorithm is used (Rabiner, 1989).

The theory about HMM can be used to solve the interpretation of sparse sequencing data for genotyping and as well can be used for imputing genotypes at maker positions without having any information. Genotyping with a HMM can result in a higher resolution compared to a sliding window approach. The drawback for using a HMM is the estimation of the transition and emission probabilities.

In the following chapter we will use HMMs to correct and impute genotypes using sparse sequencing for bi- and multi-parental mapping populations. For the bi-parental section we developed a similar model as proposed by Andolfatto et al. 2011 and Xie et al., 2010. The differences between these appraoches is in the esitmation of transition and emission probabilities and which information is used to construct the sequence of observations for the HMM. Xie et al. designed a HMM for the genotyping of a RIL mapping population of a cross of rice varieties allowing them to apply expected probabilities for transition and emission probabilities. As for a RIL population only homozygous genotypes are expected. Andolfatto et al. introduced a more general form of the HMM approach primarily designed for a RIL population from flies. They applied a Bayesian approach to calculate the probabilities under the constraint that only one crossing over per chromosome is expected and that the error rate is equal for each individual. We will present an approach predicting the genotypes using a sample-wise error rate and no constraints regarding crossing over rates.

For the multi-parental section we will use a two stage HMM for genotyping homozygous regions first and then assigning the two parental lines to the heterozygous regions in a second step. We will start by introducing a method to genotype bi-parental mapping population as imlemented in the newly developed pipeline TIGER (Trained Individual GenomE Reconstruction).

1. Genotyping by sequencing for bi-parental crosses

This chapter covers genotyping by sequencing for bi-parental crosses. We will start by introducing the method used for genotyping based on the previous introduction chapter. The result section is followed by a description of different projects that used the method. And finally the chapter will be closed by a discussion. The presented methods have been used to investigate whether the absent of *RECQ4a* could increase the recombination rate in *Arabidopsis thaliana* (Rowan, Patel, Weigel, & Schneeberger, 2015). Furthermore two more projects on genotyping a F_2 mapping population of Strawberry plants and for genotyping a *Sorghum bicolor* RIL population.

The methods part will explain how we approach the problem of imputing missing and removing false genotyped information for each given marker position based on NGS data. Imputation and correction is necessary as we are handling sparse sequencing data for each individual, which is challenging. We used a machine learning algorithm based on a HMM to solve this task.

1.1 Method

In this section we present the algorithmic approach for genotyping by sequencing of sparse resequenced data for bi-parental crosses. We called our approach **Trained Individual GenomE Reconstruction** (TIGER).

1.1.1 Premises for using the TIGER pipeline

Before we can apply TIGER to sequencing data, the data itself has to fulfill certain criteria:

1) The sample data comes from a resequencing project; reads can be aligned towards an existing reference sequence from the same species as the recombinants. It not recommended applying a different reference species as chromosomal rearrangements could generate patterns similar patterns to those introduced by recombination. This would generate false training information for the HMM.
2) The mapping population is generated from two parental lines.
3) A genome-wide set of SNPs differentiating both parents needs to be available. The density of the SNP set defines the resolution of identifying genotype blocks.
4) The crossing scheme for the recombinant population needs to known, as it is used for genotype predictions.
5) The pipeline presented here is trained and validated for diploid species.

1.1.2 Marker generation

A common way to generate SNP markers is to resequence the parental lines. For this we followed the standard resequencing workflows, including alignments of whole-genome shotgun reads against a reference sequence. After the alignment process against the reference sequence we

obtained a list of raw SNPs using SHORE for SNP calling (Ossowski et al., 2008). However, raw SNPs calls typically contain false positive SNPs.

Therefore we applied strict SNP filtering. First SNPs located in mitochondria- and chloroplasts DNA are removed, followed by removing SNPs reporting insertion or deletion (InDels) events, as TIGER does not consider insertion and deletions (InDels) for genotyping (see Discussion). Further SNPs are only considered if they are of high quality (minimum SHORE quality score of 30) and found in uniquely aligned reads. Additionally we remove SNPs which are located in a region where the surrounding genomic sequence is not supported by the reference sequence with high read mapping quality, because this indicates possible rearrangement in the close vicinity, which can complicate SNP calls in particular in low fold sequenced genomes. The remaining SNPs are further filtered for overlap with transposable elements to reduce the impact of possible rearrangements in the genotyping call (Wijnker et al., 2013) and a global read coverage filtering is applied to remove too low or too high coverage region. This setup allows only SNPs with read coverage within two standard deviations of the average genome-wide coverage. Finally, we used the segregation patterns of the SNPs in the F_2 population to remove SNPs, which did not show a Mendelian pattern of inheritance to obtain a final set of markers. These filtering steps reduce the genotyping errors that might arise from poor quality markers.

1.1.3 Pre-assignment of genotypes at individual marker positions

Typically recombinant population are large, and sequencing of those requires following multiplex based sequencing protocols. In general the sample DNA is first fragmentised and fragments are selected based on their length and sample-specific adaptors are ligated. This is done for each sample independently; afterwards samples are pooled for sequencing run on a NGS machine (Baird et al., 2008; Mir et al., 2013; Wong, Jin, & Moqtaderi, 2013).

After the reads are de-multiplexed, based on their barcode signature (Craig et al., 2008), the reads of each individual sample are aligned against the reference sequence. Then we record the allele frequency for each parental allele at each SNP marker by counting the number of aligned reads supporting the parental allele. The allele ratio can already be used for genotyping, however, the low amount of reads per marker make these call error prone.

In order to prepare for genotyping with TIGER, the read counts are transformed into three possible genotype states e.g. homozygous state for parent A or B and the heterozygous state AB (parent A and B are synonyms for the parental genomes). To model the low sequencing marker situations additional three states are included AU, BU and UU, where AU or BU represent the uncertainty about the second chromosome and is applied if less than five reads for that position are recorded (U for unknown). UU is being reported if no information could be obtained for that SNP position, i.e.

no reads aligned to this marker. The translation of the allele ratio counts into the alphabet of six states is performed as follows:

1) An allele count threshold is used for labeling homozygous states. The homozygous states (AA or BB) will be assigned if at least five reads are supporting only one of the parents at that SNP position.

2) For all other ratios, we calculated the probability of observing the allele counts from a homozygous or heterozygous background using a multinomial distribution. First we calculate the probability of drawing allele counts for either parent A or parent B in a homozygous background (x_1,x_2) (Equation 1). We assume 1% sequencing error. For a heterozygous background (f(a,b)) we consider that the probabilities for each allele would be equal to p=0.5. To determine the genotype we first compare x_1 and x_2 and take their maximum which is then compared with f(a,b). The last comparison determines if we have a homozygous are heterozygous genotype. If the maximum is greater than f(a,b) then homozygous state (AU or BU) is reported based on which variable was greater (x_1 or x_2).

The transformation of the allele frequency counts at each position into our six genotypes simplifies the construction of the HMM as we do not have to model distributions being emitted from our hidden states.

Equation 1. Probability for drawing one of the parental genotypes (x_1 or x_2) in the homozygous background and the probabilities to draw them together coming from a heterozygous distribution (f(a,b)).

$$x_1 = \frac{(a+b)!}{a!*b!} * 0.99^a * 0.01^b$$

$$x_2 = \frac{(a+b)!}{a!*b!} * 0.99^b * 0.01^a$$

$$f(a,b) = \frac{(a+b)!}{a!*b!} * p^a * p^b, p = 0.5$$

1.1.4 State model of the HMM implemented in TIGER

The pre-assigned genotypes could already represent the final genotype output, but it contains markers with no genotyping information at all which lowers genotyping resolution. Additional to missing information, wrong genotype calls are still possible even given our strict filtering steps. To predict and correct genotypes we use a HMM approach. The connection between the hidden states reflects their relationships towards each other and the probabilities of a transition. A connection from parent A to parent B is interpreted as a recombination event, changing the genotype from homozygous A to homozygous B. The weight is the probability for such an event occurring between two markers. For our purpose the HMM contains three hidden nodes (AA, AB and BB), reflecting the possible genotypes at bi-allelic sites for segregating populations derived from outcrossed parents. In our model, all hidden nodes were connected with themselves and with

each other, and each hidden node had six emission states, reflecting the alphabet that was assigned as genotypes to each marker (Figure 3).

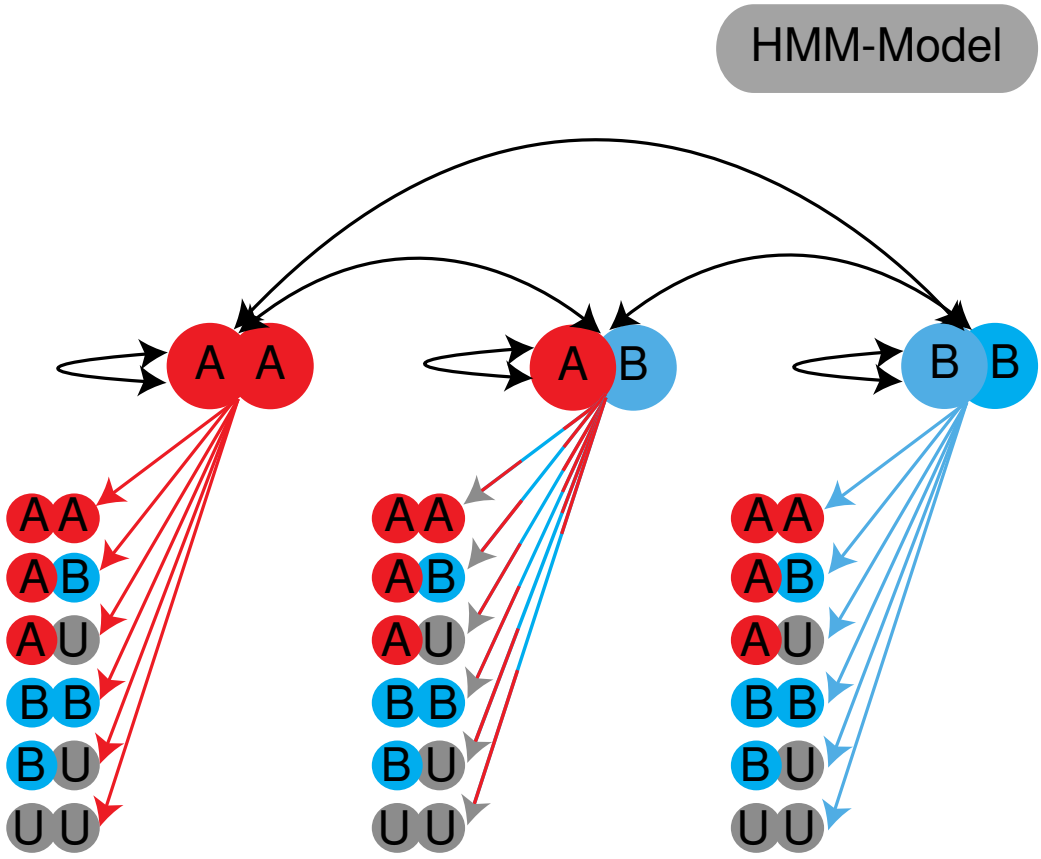


Figure 3. Schematic of the state model used in TIGER. Red indicates parent A, blue indicates parent B and grey absence of information.

1.1.5 Parameter estimations for the Hidden-Markov Model (HMM)

To finalize our model we need to add the missing transition and emission probabilities. There are different ways to estimate those either by training the model using genotyping data (supervised) or by training on the provided data set (unsupervised) (Rabiner, 1989). TIGER estimates the probabilities for each sample separately without the need for any additional information on error rate, allele bias or similar, besides the inbreeding depth of the sample. In order to get sample-specific HMMs and their probabilities we broadly estimate the genotypes using a simple sliding window. For this we first have to estimate the local allele frequencies from all chromosomes by applying a simple sliding window e.g. of 1,000 adjacent markers. In each of the window the sum of allele ratio per marker is calculated, similar to the already presented sliding window approaches (Xuehui Huang et al., 2009). The size of the sliding window should be chosen based on the filtered marker density and the expected noise level. Therefore a graphical output can be produced to determine the optimal window size on selected sample representing the average coverage rate.

The goal of the sliding window is to reduce the noise ratio to recover the distribution of the allele frequency ratios. This step already assigns genotypes to all marker positions (Xuehui Huang et al., 2009) but this does not allow for a highly accurate resolution of the CO breakpoints. In the ideal case the result of the sliding window reflects the allele frequencies 0, 0.5 or 1 based on the ratio of the parental alleles. However, due to random sampling, sequencing errors, parental allele biases mis-alignment and windows that include regions with different allelic states the distribution is distorted and does not allow for unique assignment of an uniform genotype to each of the windows. The frequencies of the resulting allele ratios from the sliding windows can be plotted as a histogram, which represent the observed allele ratio distribution for that genome. To this observed distribution we fit three beta distributions, representing the expected three different allele frequency distributions representative for the three possible genotypes.

To fit the beta distribution a beta-mixture model with an expectation-maximization (EM) algorithm is applied, which is adapted from (Ji, Wu, Liu, Wang, & Coombes, 2005). After fitting the three beta curves, we label each of the underlying allele frequency under each curve accordingly: homozygous for parent A, heterozygous or homozygous for parent B. The area under the curves is limited by 0, 1. We then can apply a supervised learning strategy to obtain the transition and emission probabilities by combining the allele frequencies and the previously genotyped labels based on the beta-mixture model thresholds (Figure 4).

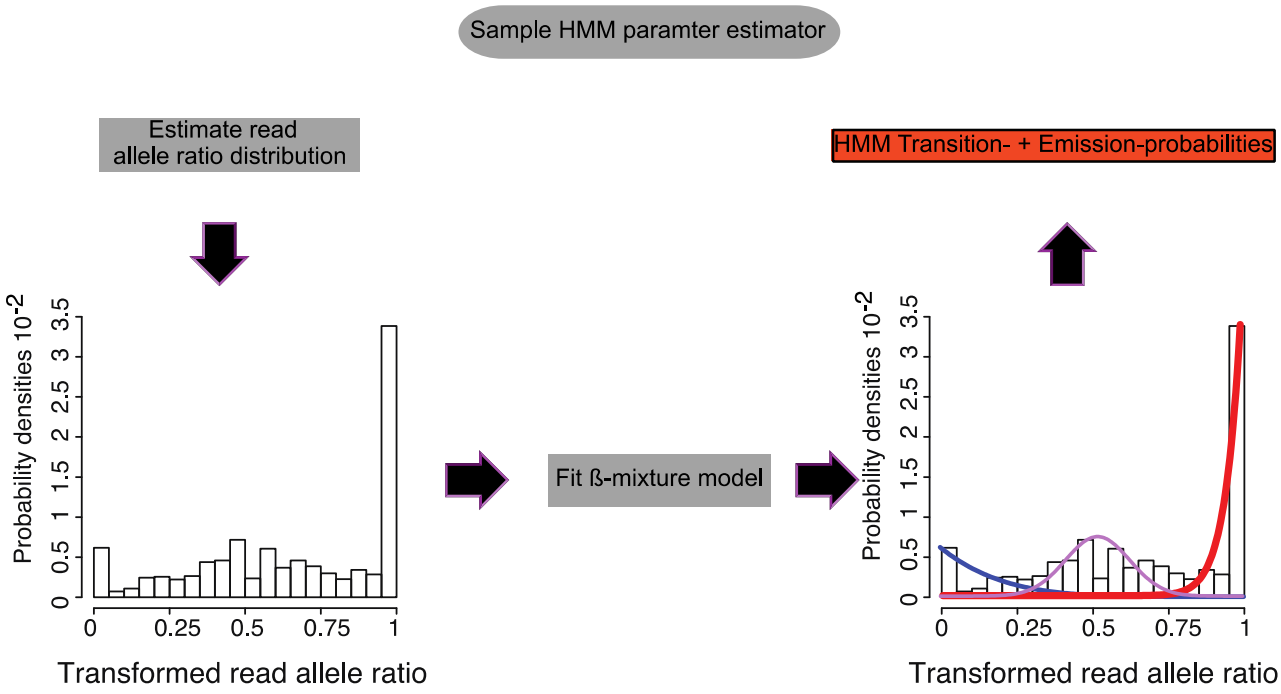


Figure 4. Schematic workflow of the parameter estimator for the HMM

For each sample the read allele ratio is estimated through a sliding window approach. The resulting frequencies are then transformed to values between 0 and 1 as the beta function is only defined at that region. Afterwards the beta-mixture model fit is applied. From the intersection of the three fitted curves and the output from the sliding window we estimate the probabilities for our sample-specific HMM.

1.1.5 Increasing the CO resolution by incorporating removed low quality markers

Tiger is genotyping a set of high quality filtered SNP markers. To increase the resolution for each sample we integrate previously removed markers into regions near the predicted CO breakpoints by a simple gap filling approach. We include from both sides of the breakpoint markers supporting the predicted genotype until the recombination breakpoint. The filling allows one marker not supporting the predicted genotype if their neighboring marker supports again the predicted genotype (Figure 5).

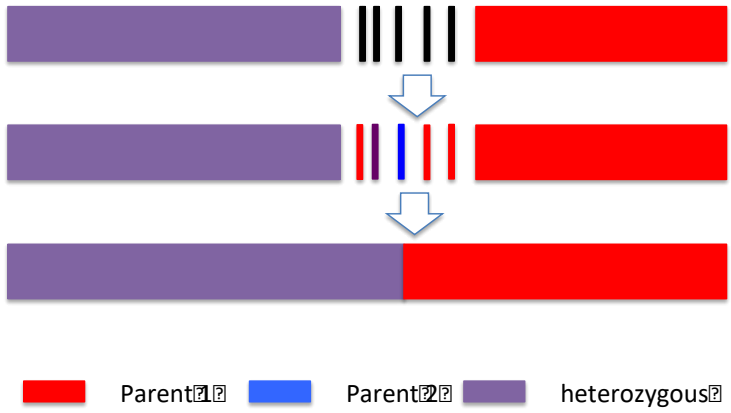


Figure 5. Closing of the predicted CO position by incorporating removed markers (black). The genotype is used to estimate a possible CO by using the last correct genotyped marker.

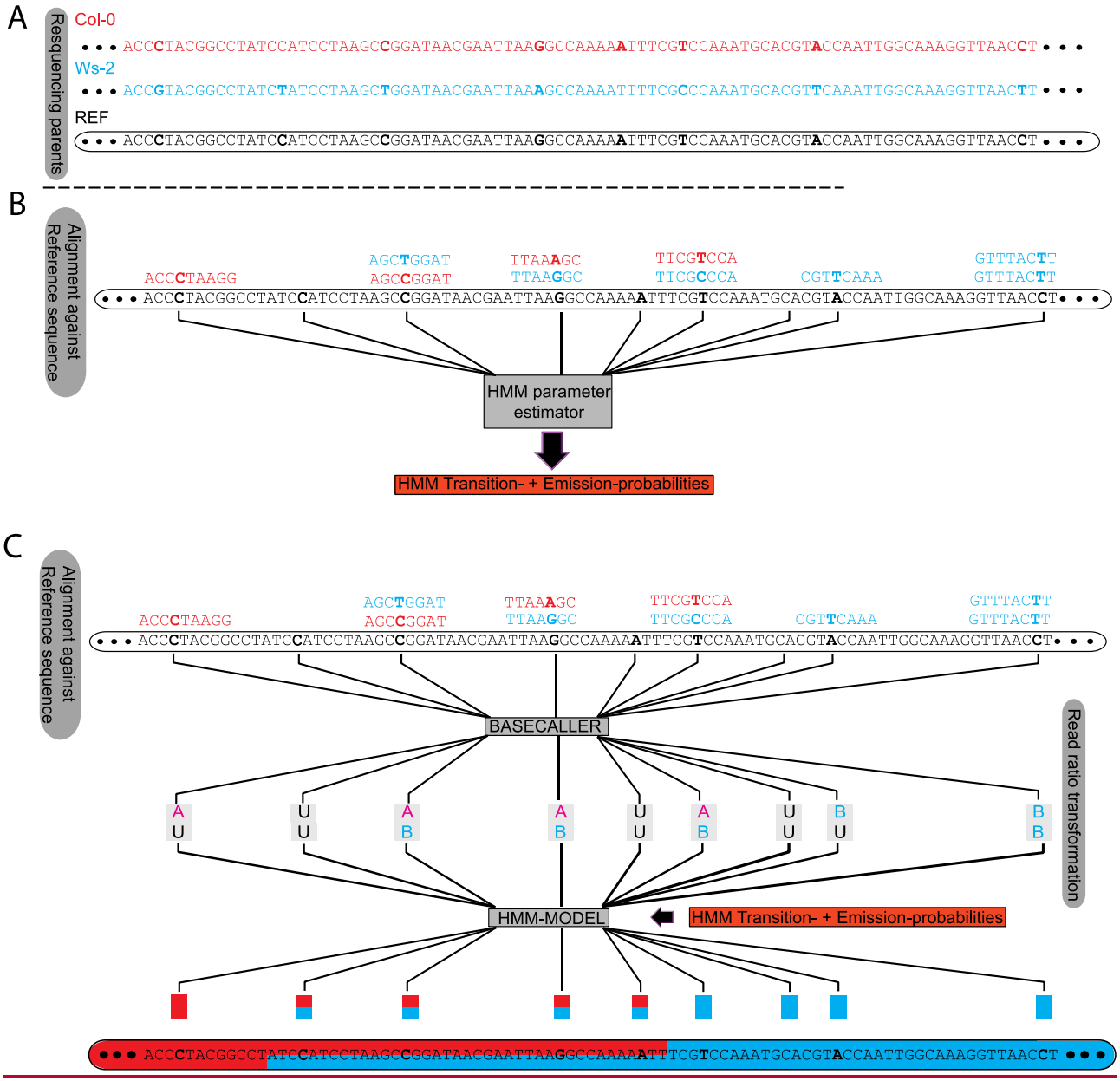


Figure 6. Workflow of TIGER

A) SNP markers for both parental genotypes have to be identified. Using high-density coverage data reveals SNPs. For example, in red *A. thaliana* Col-0 and in blue Ws-2, SNPs in bold. B) Realignment of short read data of a sparse sequenced individual. Short strings represent reads and SNPs in bold, colouring indicated which genotype the reads supports. At each SNP position the number of read counts for each genotype is counted per position and applied to the HMM parameter estimator for estimation of the transition and emission probabilities for the HMM. C) The read counts of each SNP position for each genotype is called and transformed into one of six labels representing different possible genotyping outcomes. The labelled string presentation is now interpreted by the trained HMM to correct and impute missing genotypes.

1.1.6 In-silico validation of TIGER

To validate our pipeline we simulated three different F_2 *A. thaliana* mapping populations, each containing 1,000 samples, with three different read coverage rates (0.1x, 1x and 10x) using the Pop-seq tool (James et al., 2013) with the default recombination landscape from Salome et al 2011. We used a simulated error rate of 1-3% and genotypes were generated at 261,795 high-quality SNP markers. The 1,000 samples were randomly distributed into 10 separate batches and for each batch we determined the number of predicted recombination events, the breakpoint resolution, and the types and genomic positions of errors produced by applying TIGER (Figure 7). We combined the results from 10 bins for each of the simulated coverage rates independently. We compared the difference between the predicted COs and the numbers of expected COs based on the simulated data. The difference between expectation and prediction was always positive, irrespective of coverage rate. This indicates a tendency of TIGER to underestimate the number of COs. As expected by increasing the coverage rate the percentage of COs that were not predicted decreased from 2.5% for the lowest coverage (0.1x) to 0.7% for the highest coverage (10x) (Figure 7).

To estimate the resolution of the predicted CO positions, we used the physical distance as well as the number of markers between the predicted and expected CO position. We found that the resolution improves with increased coverage. More than 90% of the COs were predicted on average within a distance of 2 kb from the expected CO position. The average number of markers between predicted and simulated CO for 0.1x coverage was 7, for 1x 1 and 10x 0. The average resolution at 0.1x was 1,986 bp (Table 1).

Table 1 Simulation results

The difference between predicted and simulated CO positions were used for estimating the quality of the prediction based on simulated data

Average coverage	CO identified				Median resolution (bp)
	≤ 90%	≤ 98%	≤ 90%	≤ 98%	
	Marker numbers		Physical Distance (kb)		
0.1x	38	79	27	222	1985.5
1x	4	10	4	94	937.75
10x	2	3	1	30	0

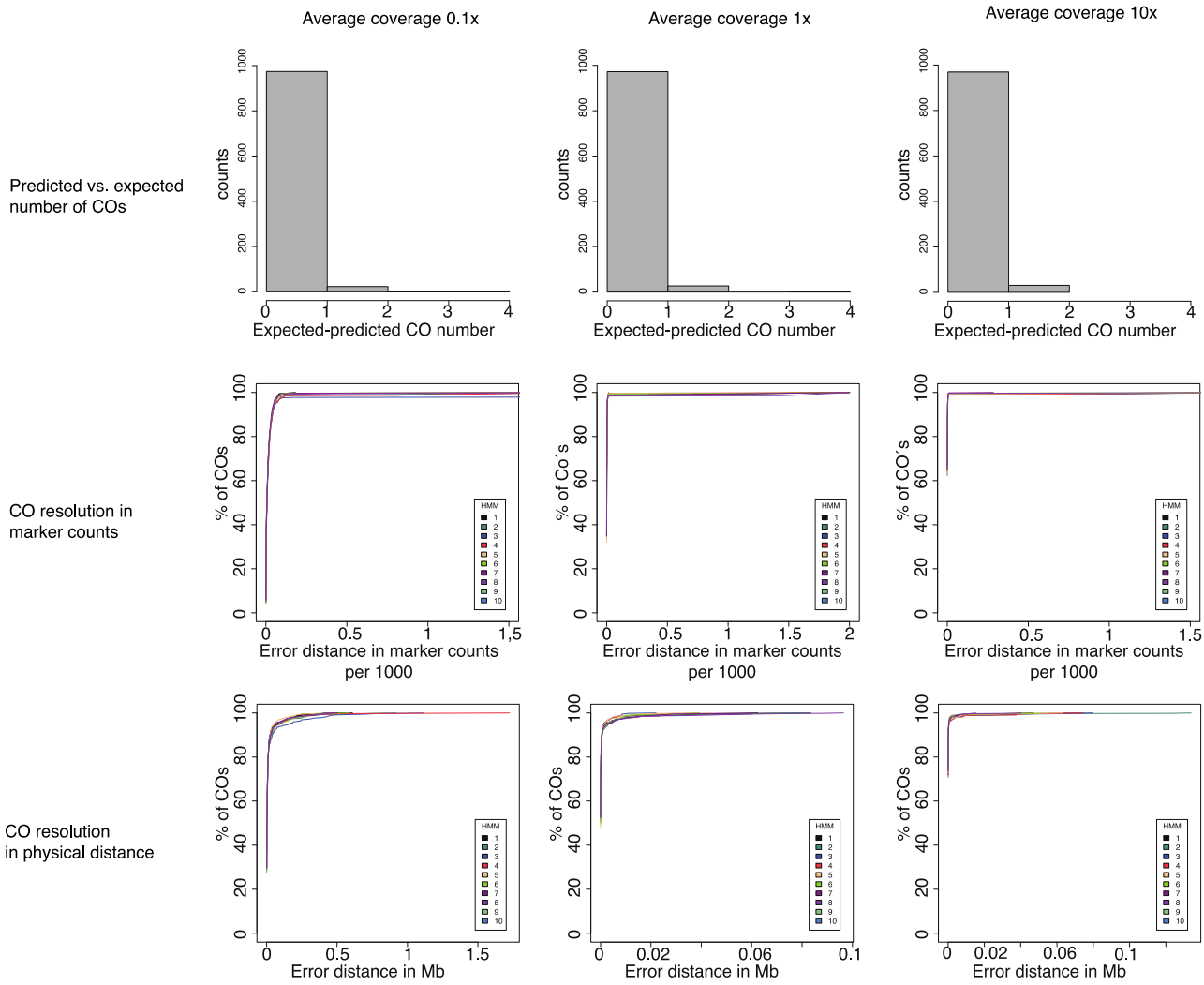


Figure 7. Validation by simulation

We simulated three different coverage rates 0.1x, 1x and 10x (columns). The first row describes the difference between predicted and simulated COs regardless of their location. The middle row shows the difference between predicted and simulated COs in marker counts (x-axis per 1,000 markers and y-axis the percentage of COs located in that interval). The last row shows the same information as the middle row but measured in physical distance (x-axis in Mb).

1.1.7 Errors detected during in-silico validation

Analyzing how many errors and what kind of errors are produced while reconstructing the simulated genome data allowed us to analyze these types of errors.

The most dominant error (89% of all errors) was mis-predicting heterozygous genotypes as homozygous regions. When only one allele was present in the short read data and there was predominant marker support for only one of the parental genotypes, heterozygous regions were falsely predicted as homozygous, especially along the chromosome arms. Most false genotype prediction errors were in the regions located within or next to the centromeres and telomeres. In these regions the median false homozygote error rate was 2.4% for 0.1x coverage and around 0.9% for higher coverage levels. The background error rate for all other types of errors (i.e.

homozygous but simulated heterozygous or the predicted homozygous genotype) was 0.1% regardless of the coverage (Figure 8).

Regions adjacent to the centromere exhibited a high marker diversity, which dropped off at the border of the centromere including repeats, wrong and missing reference sequence information that is the most likely reason for this type of error (Figure 8). To exclude a bias for predicting only for a certain parental homozygous genotype wrong, we analyzed the false homozygous error rate per parental allele. False homozygote regions represented either 42% or 47% for either one of the parental homozygous, indicating that this type of error was not biased towards one of the two parental genotypes.

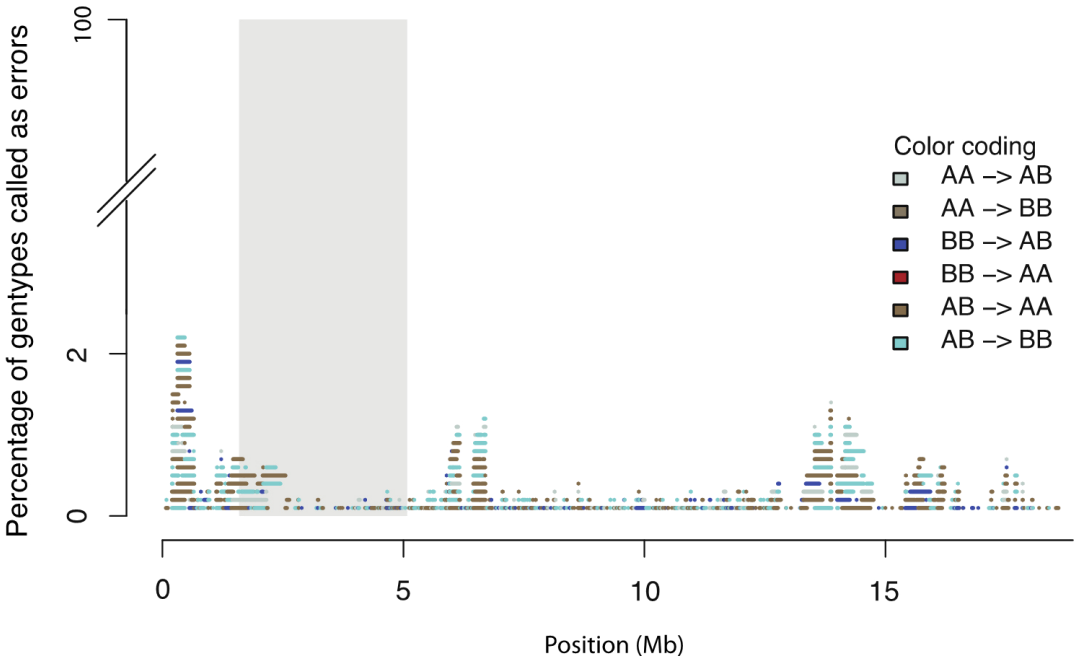


Figure 8. The frequency of different types of genotyping errors produced by TIGER using simulated data. An example of an error profile for Chromosome 4 is shown (results were similar for the other four chromosomes). The grey box indicates the location of the centromere. The error frequencies were obtained from genotypes predicted from read data from 1,000 simulated recombinant individuals. X-axis genomic scale in Mb and y-axis the percentage of wrong predicted genotypes.

1.2 RESULTS

In this section we will describe the application of TIGER applied on real data. We present three different projects where TIGER was used for genotyping individuals from mapping population with sparse sequencing data. Each project covers a different taxa and a different motivation for genotyping individuals from a mapping population.

1.2.1 Applying TIGER on individuals of a mapping population of *A. thaliana*

1.2.1.1 Introduction

In yeast and in humans exists a homologous protein of *A. thaliana* RECQ4a, SGS1 (yeast) and BLM (human). These proteins are involved in resolving CO intermediates (Knoll & Puchta, 2011). It has been shown that RECQ4a can partially restore the meiotic defects in yeast *sgs1* mutants (Bagherieh-Najjar, de Vries, Hille, & Dijkwel, 2005). A defect in RECQ4a in somatic cells of *A. thaliana* increased the CO rate (two to seven-fold) (Hartung, Suer, & Puchta, 2007). Higgins et al. 2011 found out that the RECQ4a is localized at the telomere regions and along the chromosome axes, partially interacting with CO proteins but also found evidence that RECQ4a is actually resolving telomeric bridges (Higgins, Ferdous, Osman, & Franklin, 2011). Therefore we have here conflicting observations regarding the involvement of RECQ4a during meiosis in resolving CO intermediates. To investigate this we (Rowan et al., 2015) developed two mapping population, a wild type and a *recq4a* (mutant) population based on the background of the parental genotypes Col-0 and Ws-2 and mutants with the same background, respectively. Each population consists of 196 individuals. Both populations were sparsely sequenced and genotyped using TIGER.

First we have to generate markers. SNPs were generated by analysing Ws-2 (25x) against the Col-0 reference. From the high-coverage data we found 840,611 SNPs between Ws-2 and Col-0 (TAIR10). After removing the SNPs located in the mitochondria and chloroplasts genomes and those close to indel polymorphisms, 745,273 SNPs remained. An additional 238,111 SNPs were removed after considering only high quality SNPs and those supported only by uniquely aligned reads. We further decreased the marker number to 302,082 after applying filtering for homozygous regions, transposons and coverage filtering. Finally we removed 40,287 SNPs that did not show a Mendelian pattern of inheritance in our F_2 population to obtain a final set of 261,795 markers. The F_2 individuals were sequenced on an Illumina GAIIx analyzer in one flow cell lane using 2 x 150-bp length paired-end reads. Raw reads, which were de-multiplexed and aligned to the TAIR10 reference genome for detection of sequence polymorphisms using the SHORE and GenomeMapper software (Ossowski et al., 2008; Schneeberger et al., 2009). On average 88,856,650 reads were produced, which is an average of 1.03x per sample (Figure 9). The percentage of uniquely aligned reads was 46%, indicating that a problem with the library or possible contamination exists. We did not further investigate this and went on with our pipeline.

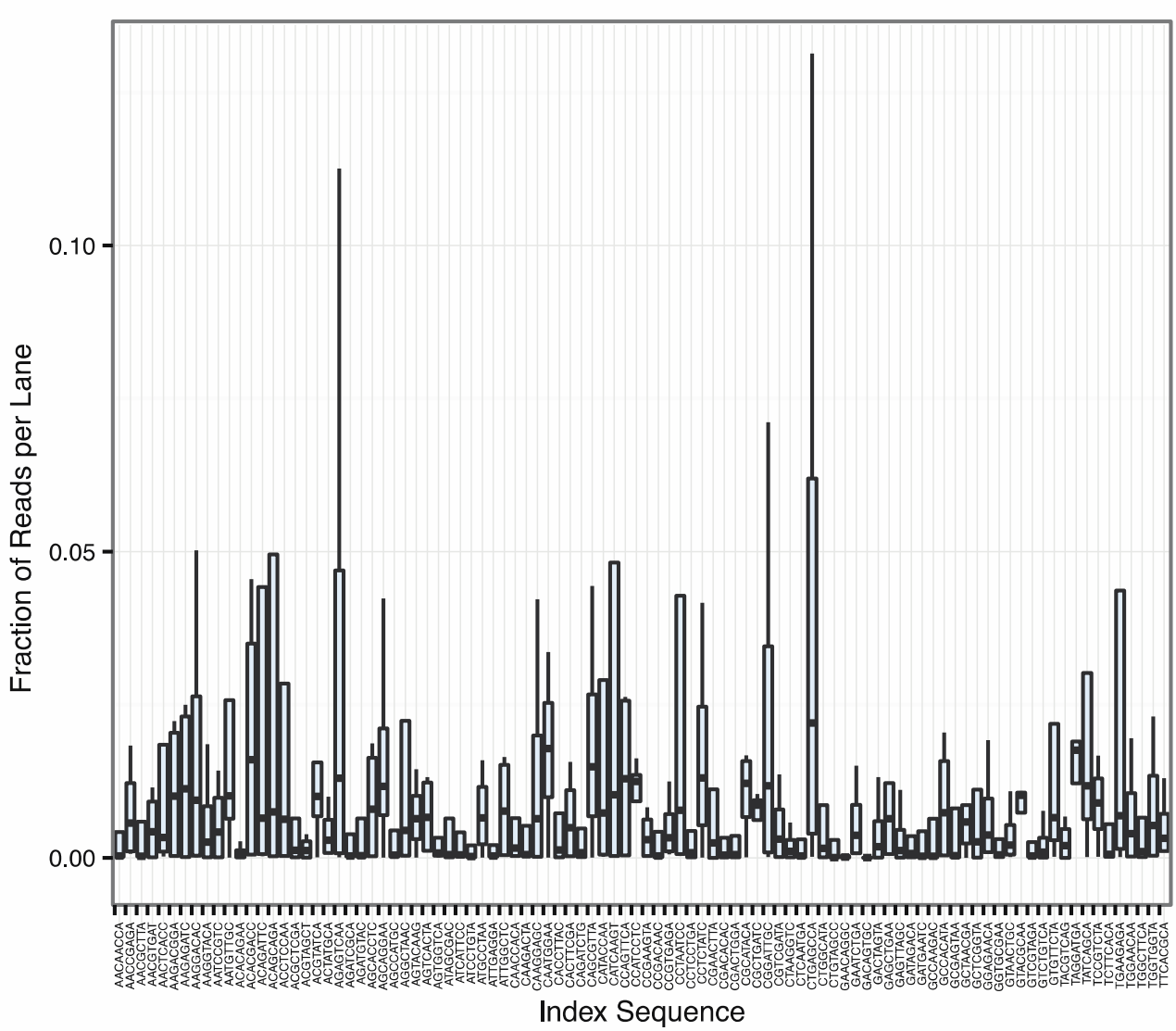


Figure 9. Sequencing statistics for the barcodes used to produce the sequencing data.
The x-axis represents the barcode sequence and the y-axis shows the fraction of reads per lane.

1.2.1.2 Reconstructions of wild type and recq4a F2 sample genomes

An average genome-wide coverage threshold of 0.025x was used to remove samples with too little sequencing data as the accuracy of correct CO breakpoint prediction was strongly reduced at such coverage rates. After filtering, 216 individuals could be reconstructed (110 from the wt population and 106 from the *recq4a* population), overall representing an average coverage of 0.63x and a median coverage of 0.37x. From our simulation studies we already observed several types of errors (Figure 8) but we encountered a new additional type of a possible errors, where small genotype blocks were embedded within larger blocks of a different genotype. We termed these regions “islands”, which could either be false positive or real double recombination events (Figure

10).

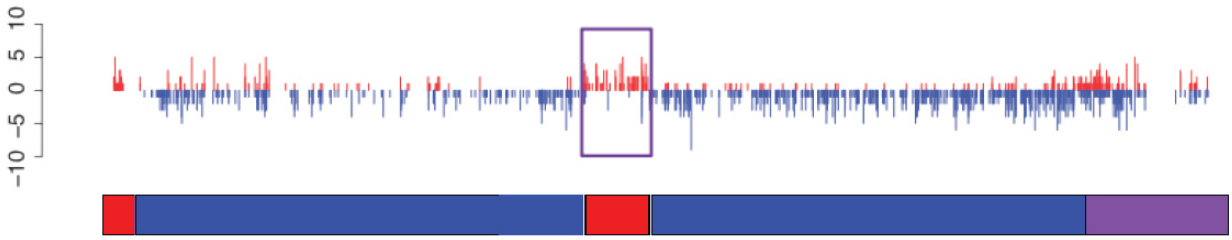


Figure 10. Illustration of an island structure

Illustration of read support (y-axis) for either one of the parental allele (red or blue) for F2 sample for one of the chromosomes (x-axis). The purple box shows a genotyping island showing a homozygous genotype for one of the parental allele

We used the island length distribution for filtering for wrong signals. From 67 islands, 7 were less than 400-kb long and were removed (Figure 11). The remaining 60 islands were categorized as double COs. After the error correction step, the final genome reconstructions can be used for further analyses, i.e. whether the observed CO rate is dependent on the coverage rate. We observed that the CO predictions were hardly affected (Figure 12). After introducing removed markers next to the predicted breakpoint sites, we could resolve the majority of COs to an interval of two kb or less (Figure 13). For validation eleven CO breakpoints were randomly selected for PCR and Sanger sequencing. Eight of the eleven breakpoints were in the predicted two-kb intervals (Table 2).

The observed breakpoint resolution of our experimental F_2 populations closely matched that of the simulated populations at 0.1x coverage. Given the median coverage rate of 0.3x this indicates that the simulation was quite fitting and therefore the expected error rate based on the simulation data should be in the same region for our experimental set. Finally, we verified that the overall frequency of Col-0, Ws-2 and heterozygous genotypes in both populations was consistent with the expected pattern of inheritance (

Table 3) before comparing the CO patterns between both populations.

Table 2 CO An evaluation of CO break point predictions using TIGER and PCR comparison

[a] Genotypes predicted up or downstream of the CO point

[b] Size of the PCR fragment amplified

[c] number of markers covered by Sanger sequencing reads

[d] position of the first and last markers covered by the Sanger sequencing reads

Table taken from (Rowan et al., 2015)

ID	Pop.	Plant ID	Chr.	Pos.	Up[a]	Down[a]	Frag. size[b]	N[c]	First marker[d]	Last marker[d]	Contains Breakpoint
1	wt	125	4	134048	Col-0	Het	966	2	133527	134493	yes
2	wt	145	4	16276940	Het	Col-0	1567	3	16275698	16277265	yes
3	wt	147	1	29632761	Col-0	Het	1448	3	29532395	29533843	yes
4	wt	125	4	10419728	Het	Col-0	675	4	10419347	10420022	yes
5	wt	139	5	24954623	Het	Ws-2	784	5	24954069	24954853	no (Ws-2 only)
6	*recq 4a*	231	3	4957859	Het	Ws-2	1118	6	4957322	4958440	yes
7	*recq 4a*	253	5	26618925	Het	Col-0	1351	2	26568249	26569600	no (Het only)
8	*recq 4a*	261	2	12206209	Ws-2	Het	779	3	12206209	12206988	no (Het only)
9	*recq 4a*	278	1	1850178	Het	Ws-2	1210	6	1849692	1850902	yes
10	wt	308	3	172226	Col-0	Het	1242	7	171475	172717	yes
11	wt	387	3	10014511	Col-0	Het	1012	2	10013641	10014653	yes

Table 3 Mendelian distribution

Population	Genotypes		
	Col-0	WS-2	Col-0/Ws-2
wt	28.00%	21.80%	50.20%
recq4a	23.80%	25.60%	50.70%

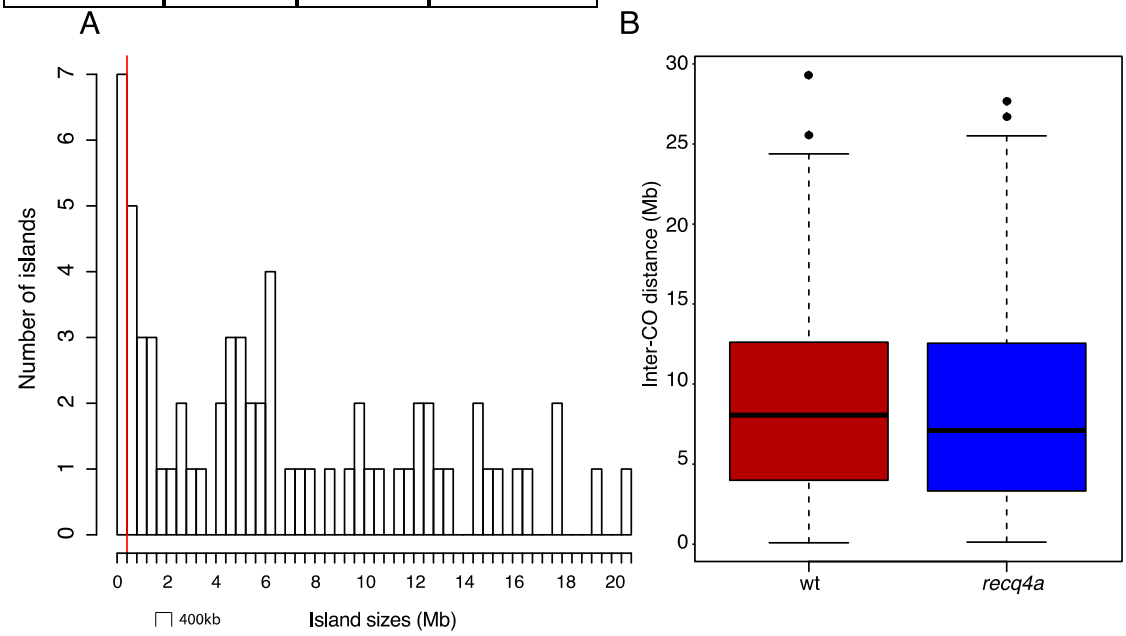


Figure 11. "Island" errors and double COs.

TIGER-generated reconstructions of experimental recombinant individuals produced a type of error where small genotype blocks were embedded in a larger block of a different genotype. A) Histogram depicting the lengths of these small genotype "islands". Some of these islands are errors, others might represent real closely-spaced double COs. The red line indicates the chosen threshold for distinguishing between island errors and true double COs. B) Box plots showing the inter-CO distances for all double COs in the wt compared to the mutant population.

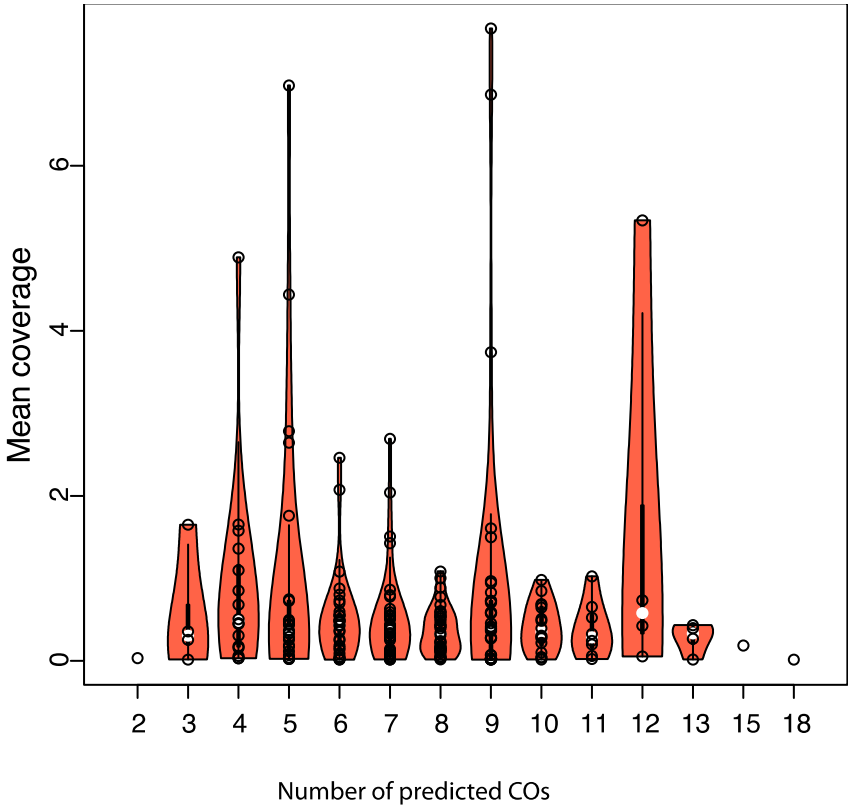


Figure 12. The effect of coverage on CO prediction using TIGER

The density curve for probabilities (orange) is indicated for each number of predicted COs compared to the coverage rate.

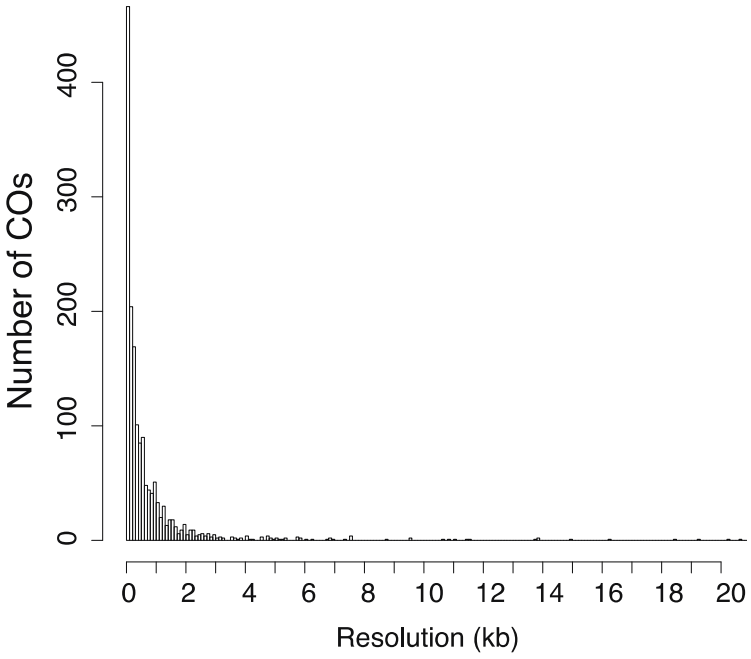


Figure 13. CO Inter-marker distance between predicted CO positions

The x-axis is in kb and the y-axis shows the number of COs in counts

1.2.1.3 RECQ4a does not affect the frequency or distribution of CO events in Col-0 X Ws-2 F2 populations

A comparison of the CO distribution and frequency in the wt and *recq4a* populations by counting the number of CO events in a 100-kb sliding window across the chromosomes was done. The CO frequency was highest on the chromosome arms in the regions adjacent to the centromeres and lowest within the centromere, as it was already described previously for many F_2 populations derived from different pairs of parents (Salomé et al., 2012) (Figure 14A). There are several windows where there appears to be a difference in CO frequency between wt and *recq4a* mutants. To test if these differences were statistically significant we used a χ^2 test with correction for multiple testing, resulting in no significant difference for all windows. Overall the CO frequency and distribution of wt and *recq4a* mutants were highly correlated. Although the *recq4a* population had a slightly higher number of COs per chromosome (1.52) compared with the wt population (1.46) (Figure 14B), this difference was also not statistically significant as determined by a Wilcoxon test (p-value 0.32).

To determine if *recq4a* influences CO interference (where the presence of one CO on a pair of homologues suppresses the formation of a CO nearby), we measured the inter-CO distance in both populations. We only compare double CO on the same chromosome. The mean inter-CO distance was slightly higher in the wt population (15,855,278 bp) than in the mutant population (15,243,188 bp). We again used a Wilcoxon test to test for statically significance, which was not given (p-value 0.15) (Figure 14C). The mean distance between double COs that occurred on the same chromosome (inter-CO double distance) was also slightly higher in the wt population (8,700,616 bp) compared with the mutant population (8,424,029 bp), but again the difference was not significant (Wilcoxon test p-value 0.33). We conclude that the loss of RECQ4a either has no effect on the frequency or distribution of COs or that its effect is so minor that the number of individuals we examined was too few to detect it.

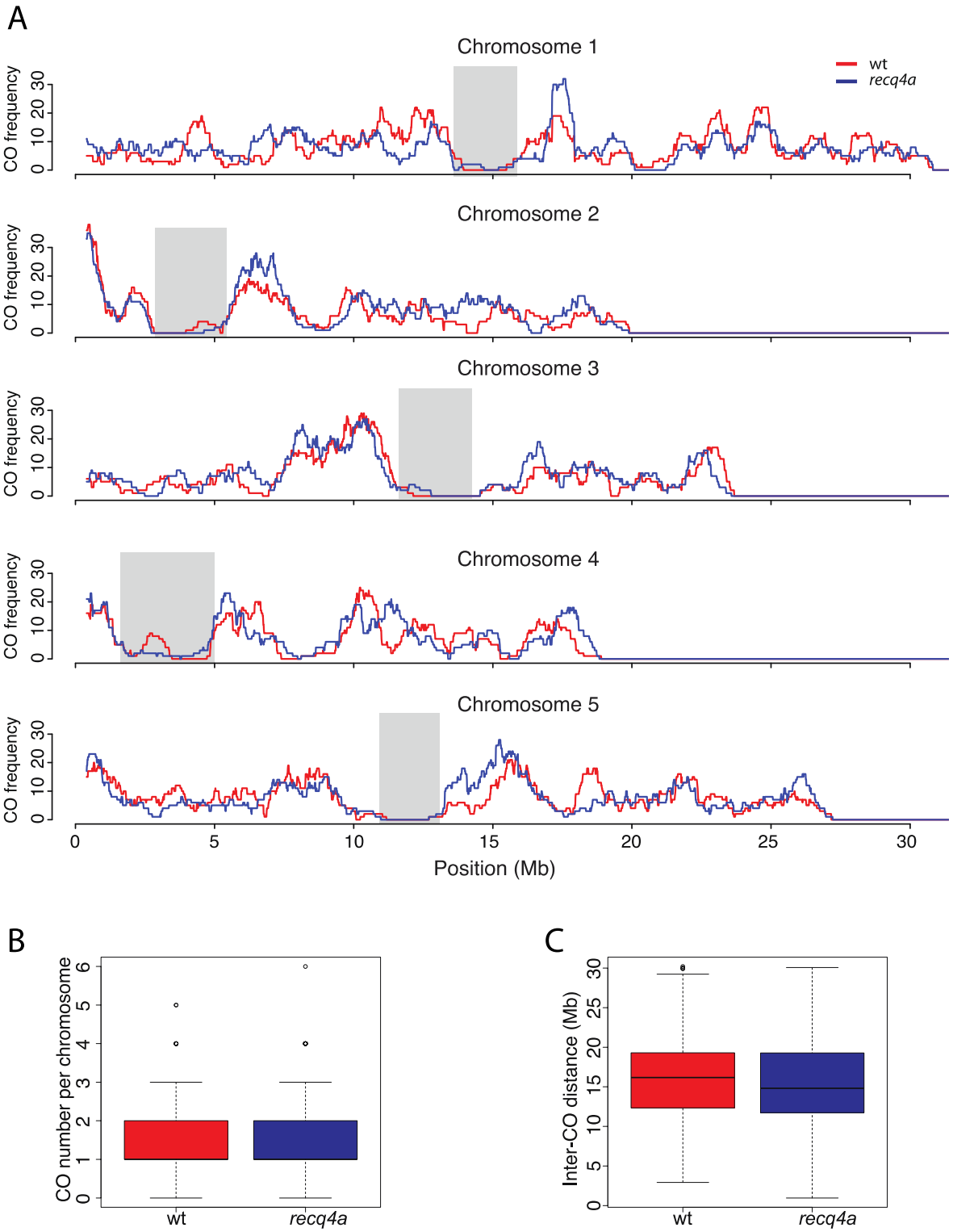


Figure 14. Comparison between *recq4a* and the wild type mapping populations

A) Recombination landscapes are compared between wt (blue) and *recq4a* (red) in a sliding window of 100kb size for each of the five chromosomes. The x-axis is in Mb and the y-axis in CO frequency counts. B) Boxplot of the CO number per chromosomes and C) boxplots of the comparison of the inter CO distances (in Mb) for each mapping population. Figure taken from Rowan et al, 2015.

1.2.1.4 A suppression of COs reveals a 1.8 Mb inversion

By comparing the CO landscape of both populations Beth et al. 2015 found the expected CO suppression regions close to the centromeres and to the chromosome arms and additionally an unexpected suppression from 7 to 9 Mb on the long arm of chromosome four. Since regions of suppressed recombination are thought to contain inversions (Coyne, Aulard, & Berry, 1991), a structural variant analysis using Pindel (Ye, Schulz, Long, Apweiler, & Ning, 2009) were done using the high-coverage Ws-2 short read data. Pindel predicted inversion breakpoints at positions 7,139,542 and 8,914,936 bp. A confirmation was done by PCR and Sanger sequencing, which revealed that the downstream break was coupled with an additional insertion of 389 bp, of which 337 bp had 83% similarity to the CACTA-like transposable element Ptta/En/Spm. A PCR-based screening in Ws-0 revealed that the inversion was not present there. From combined CO data we could pinpoint at 6,989,963 and 8,960,496, which is 150kb and 45kb away from the actual inversion, indicating that only using CO data was quite accurate.

1.2.2 Applying TIGER on a *Fragaria vesca* mapping population

1.2.2.1 Strawberry genome

In 2010 the draft reference genome of woodland strawberry has been published (Shulaev et al., 2011), which is based on the assembly of *F. vesca ssp*. *vesca* accession (H4x4). *F. vesca* is a diploid genome, containing 7 chromosomes and it is the most plausible progenitor of *F. ananassa* (which is the agricultural crop taxa). The genome size was expected to be 240 Mb long, which is nearly double the size of *A. thaliana*. For the reference sequence itself 198 Mb genomic sequence could be anchored to seven pseudo chromosomes, covering 82.9% of the genome (Shulaev et al., 2011). Studying the genome of strawberry offers broad benefits, as it has been already cultivated for centuries, showing high diversity as it grows in different climate ranges, it is self-compatibility, has a short generation time and can be used as a model species for the *Rosaceae* branch which contains crops like apple or peach.

This analysis has been conducted as part of a genotyping project led by Timo Hytönen (University of Stockholm). The focus was on the genetic interaction of the gene *TERMINAL* FLOWER1 (TFL1) in the background of the wild Strawberry *F. vesca* genome. It has already been shown that TFL1 is suppressing flowering as it binds to FD and represses the expression of LEAFY, AP1 and FUL (Hanano & Goto, 2011; Ratcliffe, Bradley, & Coen, 1999). Koskela et al., 2012 showed that TLF1 is active in photoperiods pathway, which have been observed in *Arabidopsis*, where it is linked to the development stage (Conti & Bradley, 2007). The genes of flowering time are nearly fully conserved between annual (i.e. *Arabidopsis*) and perennial (*F. vesca*) plants but not their regulation. Hence the task was to find out which other genes were interacting with TLF1. Therefore a cross between two accessions Hawaii-4 (H4×4) and the mutant *TLF1*, where the H4×4 genome is identical towards the published reference genome (Shulaev et al. 2010), was done. The *TLF1* mutant has a 2 bp deletion in the first exon (Hytönen et al. 2012) leading to non-stop flowering phenotype. A F_2 population was generated and samples were selected showing extreme flowering time phenotype. 40 samples were selected, 20 early and 20 late flowering plants. The samples should not have a continuously flowering phenotype. By collecting extreme flowering time samples, the aim was to reveal a genetic basis of alleles describing the observed phenotype. Therefore, a QTL-analysis was done using the genotypes from the 40 individuals.

1.2.2.2 SNP markers filtering

Resequencing of the *tlf1* mutant was done (5x) and the resulting reads from *tlf1* were mapped against the reference sequence to find SNPs markers for genotyping. As described in the methods part we applied a strict marker filtering protocol. We found 491,527 raw SNPs between both accessions. These SNPs were already filtered for InDels, mitochondrial and chloroplast DNA. After removing SNPs with low quality we ended up with 134,491 SNPs. Further 111,682 SNPs were filtered out through the vicinity step and additionally 2,288 SNPs were removed as they were

located in repetitive regions identified by RepeatMasker (Smit, AFA, Hubley, R & Green, 2010). By removing non-segregating SNPs in our mapping population we reduced the number of SNPs to the final set of 15,225 SNPs.

The distribution of these SNPs showed no evidence of having a bias towards a particular region or chromosome (Table 4).

Table 4 Number of SNP markers per chromsome

Chromosome	Number of SNP markers
1	1,503
2	1,936
3	1,456
4	2,217
5	2,140
6	3,517
7	2,456

1.2.2.3 Sequencing results for the 40 selected samples

Paired end read sequencing data were produced by the ABI Solid technology (Mardis, 2008). Read length was for the first read 48 bp and for its partner 32 bp. The 40 samples were multiplexed using barcoded sequencing 1,348,342,937 short reads were generated. We de-mulitplexed and aligned the sample reads using BWA (Heng Li & Durbin, 2009) with default parameters but using single end mode. On average 15.35% of 32 bp and 25.47% of 48 bp reads were aligned uniquely against the reference sequence, resulting an average coverage of 1.78x per sample. We aligned single end reads instead of paired to receive a higher amount of aligned reads as in pair end mode only 1% of the read could be aligned, indicating a problem with the quality of the read pairs regarding their insert size. The read distribution per sample was variable, the lowest coverage per sample was 0.003x and the maximum 3.03x (Figure 15).

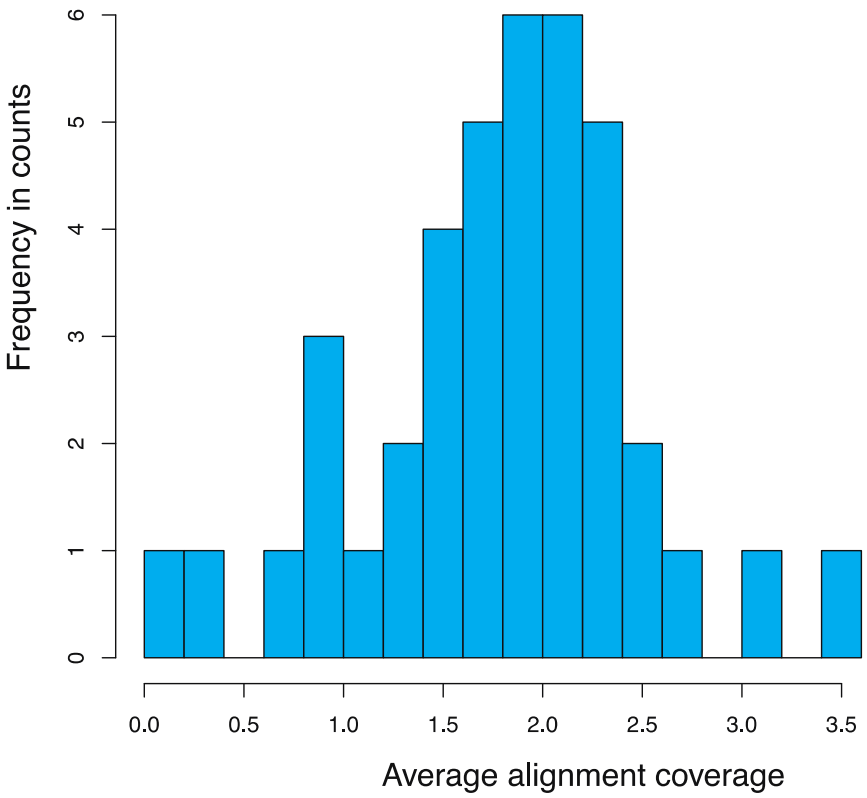


Figure 15. Histogram of the average alignment coverage rate for the 40 samples (x-axis) and their frequency in counts (y-axis).

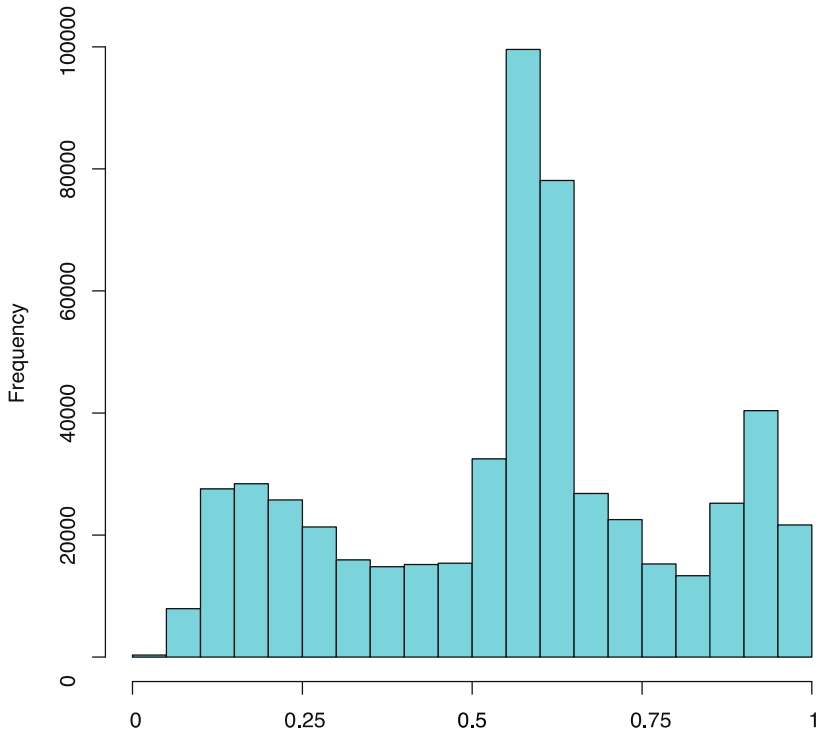
1.2.2.4 GBS of the 40 strawberry recombinants

We applied our pipeline for reconstructing the parental mosaic structure of the 40 samples. We used the same threshold of 0.025x as described in the reconstruction of *A. thaliana* recombinants. Using this threshold we removed one sample of the late flowering type from the analysis, resulting in genotyping of 20 early and 19 late flowering samples. Analysis of the raw allele frequencies of the sample data we encountered a skewed distribution towards one of the parental alleles (H4x4), indicating either there exists a bias for a parental strain during sequencing or errors in the reference sequences (Figure 16A). TIGER´s design automatically took this bias into account during the estimation of the transition probabilities of the HMM (see Method 1.1.4). To quickly verify the predicted genotypes we counted the resulting genotype frequencies of H4x4, *tlf1* and heterozygous in the samples. The resulting frequencies are as expected for a selected extreme F_2 population considering that the sample size is only 39. Therefore a smaller drift in the frequencies was expected and can be tolerated (Table 5).

After genotyping we observed the typical “islands-structures”, but compared to the genotyping of *A. thaliana* the frequency was higher (Figure 16B). The abundant type of island patterns were short double COs (< 400kb), which mostly formed heterozygous regions (92%). After applying a

threshold of 400kb for removing false positive double CO, the majority of the remaining islands were based on one of the parental genotypes.

A



B

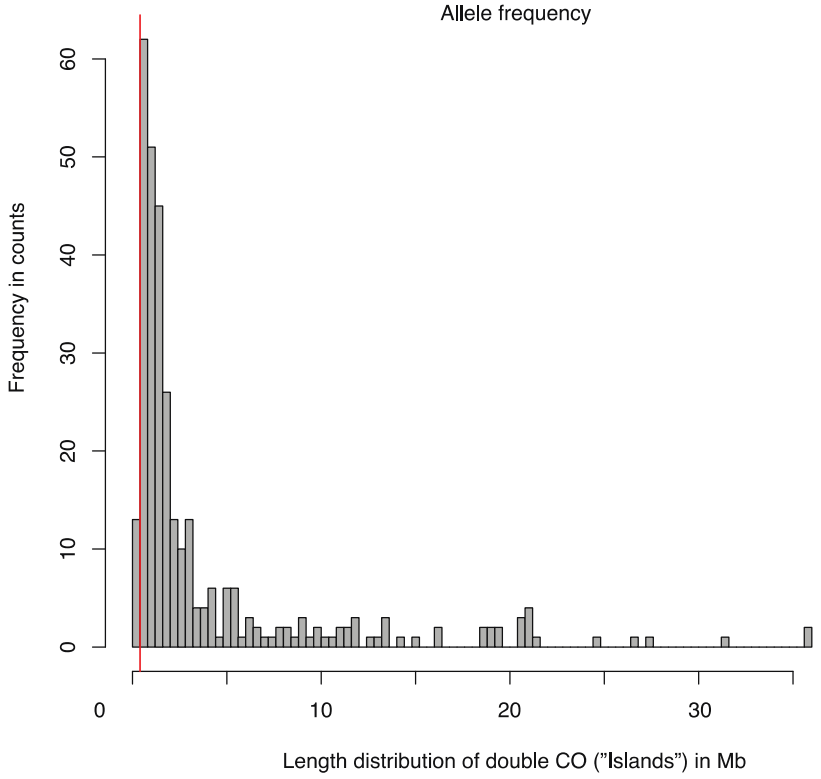


Figure 16. A) Histogram of the allele frequency bias (x-axis) between TLF1 and H4x4 (0 and 1) as assessed on markers before imputation with TIGER, indicating a bias towards the H4x4 allele. B) Length of double COs, where the red line is the threshold of 400 kb applied for filtering.

Table 5 Genotype frequencies after applying TIGER

Genotypes	Predicted (median)	Expected
H4x4 (WT)	0.26	0.25
TLF1	0.21	0.25
Heterozygous	0.57	0.50

1.2.2.5 Evaluation and breakpoint resolution

Before applying QTL-analysis for identifying candidate regions for flowering time, we estimated the recombination resolution using 15,225 SNPs: average resolution was 21,732 bp with a median average resolution of 6,684 bp. To validate the predicted genotypes we selected 28 samples based on availability of genetic material for genotyping with 21 selected SSR markers which have been analyzed by the group of Daniel James Sargent (University Fondazione Edmund Mach) (Table 6). SSR markers located on chromosome four were removed, as they showed globally no conformation with the predicted genotype in all samples. In our approach each chromosome is genotyped independently and there is no logical explanation why TIGER should have a bias for one special chromosome. A possible explanation could be an error in the reference sequence, leading to a wrong genotype prediction. For example reads might be mapping to the correct reference sequence but the position of the reference sequence could be wrong. We further removed three SSR markers, which show similar patterns, as those removed from chromosome four, located at chromosome two and two of them at chromosome three. In the final set we used 15 SSR markers for validation of our predicted genotypes, which results in 97% agreement.

Table 6 SSR markers fore genotyping validation of the predicted results of TIGER in *F. vesca* mapping populations

Chr.	Position in cM	Name
2	0	CFVCT020
2	12,2	CFV-3099
2	37,3	EMFn134
3	0	UFFxa02H04
3	60,2	EMFn207
3	72,2	CFVCT011
4	0	UDF007
4	11,6	EMFV007
4	34,1	FvH91
5	0	FvH93
5	15,6	UDF006
5	19,9	CFVCT024
5	20,1	EMFn010
5	22,6	CFV-3821
5	32,1	UDF009
6	30,1	EMFn117
6	101	EMFv160BC
7	0	EMFn201
7	19,4	EMFVi008
7	33,8	CFVCT023
7	35,2	BFACT44

1.2.2.6 Genotype frequency and QTL detection

We applied a QTL-analysis regarding flowering time for 39 samples. To reduce the data amount without reducing the quality of the result, we identified sequence blocks, which were not interrupted by a CO event in any of the samples and identified a representative marker for this region. We further applied filters to remove markers showing a segregation distortion or being nearly identical towards other markers until we reached a final set of 1,547 markers. After this reduction step we applied the multiple QTL mapping (MQM)-analysis with co-factors (Arends, Prins, Jansen, & Broman, 2010; Jansen, 1994). Using MQM reduces the appearance of ghost QTLs by using the standard composite interval mapping (CIM) (Jansen & Stam, 1994; Zeng, 1994) with a generalized linear model regression. We found four significant QTL regions with LOD score > 3.9, at the bottom of chromosome four, two in the middle of chromosome six and at the end of chromosome seven (Figure 17A). Three of four QTLs where stable by using a MQM-permutation test, where the marker data is shuffled for 100 times and each outcome was tested if the QTL was appearing or not. The QTL at chromosome 6 disappeared.

Afterwards we analyzed the region under the QTLs (Table 7) and searched for genes of interest in the particular regions. For that, we used the current annotation of the reference *F .vesca* from Genome Database for Rosaceae (GDR) (Jung et al., 2008) and filtered for known flowering time related transcript genes. Using this approach we could identify six genes, which have been annotated as flowering time related (Centroradiales (CEN), Apetala1 (AP1), Dormancy Associated MADS-Box (DAM), Constant-like (COL13 and COl16) and TLF1).

Table 7 Significant identified QTL regions

Chromosome	Start (bp)	End (bp)
4	24,610,434	25,478,187
6	9,523,164	12,044,254
7	17,495,980	22,554,356

By comparing the genotypes of each sample for our candidate genes we identified a pattern for early and late flowering phenotypes. For the late flowering phenotype each of the samples carries at least the candidate genes with one of the H4x4 allele. This was not the case for the early flowering phenotypes plants. For that class the parental allele of the TLF1 mutant is either on chromosome four or on chromosome seven or on both. Only two samples were not in agreement with the observed pattern.

Additionally we tested for interacting QTLs using scantwo (Broman, Wu, Sen, & Churchill, 2003) and identified two possible interactions: chromosome 7 and chromosome 6 interacting with chromosome 4 (Figure 17B). Including the interaction the QTLs could explain 84% of the observed phenotypic variation. In comparison, the single effects only explained 77%.

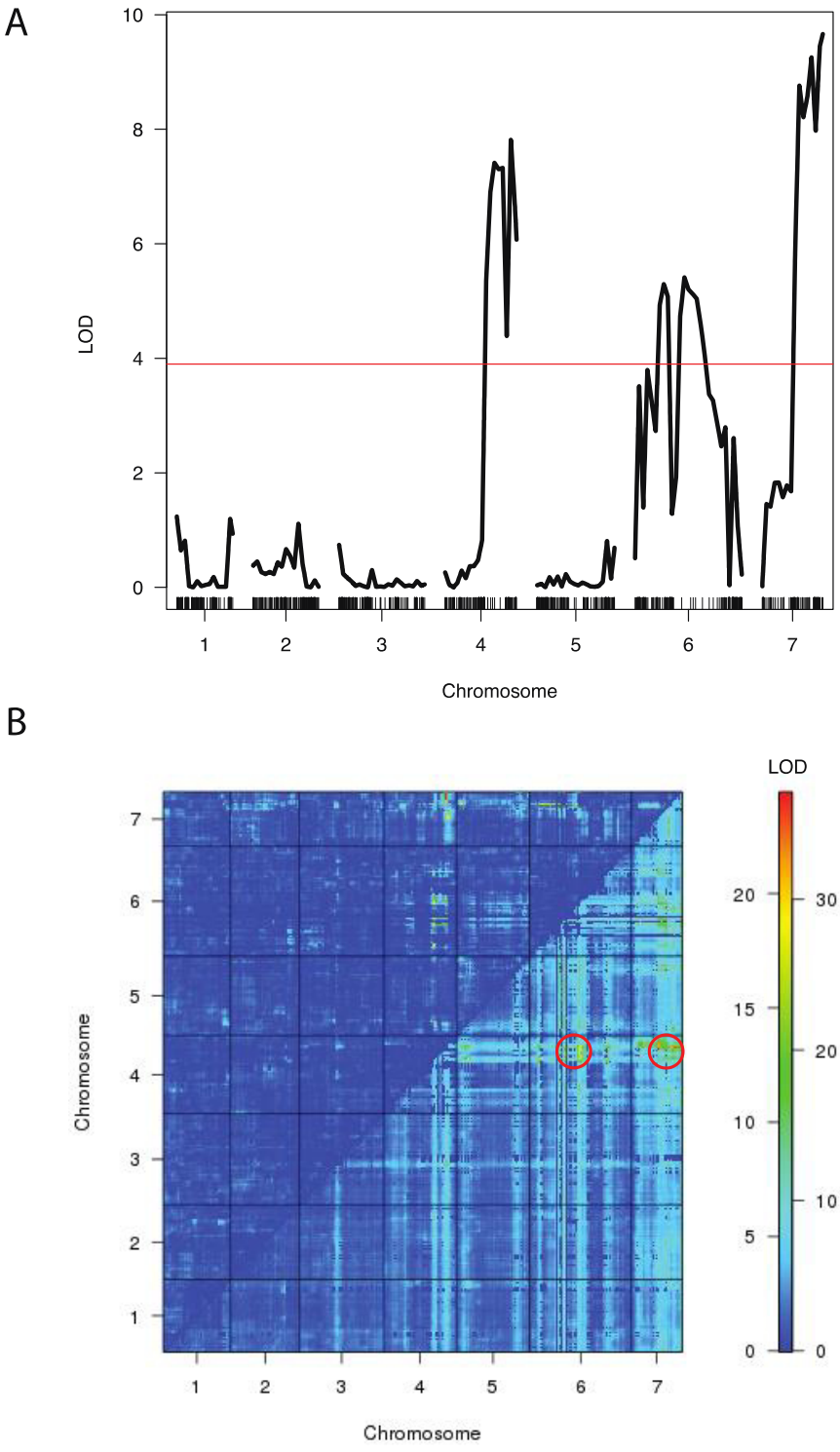


Figure 17. Single QTL-analysis and interaction between QTLs

Three significant QTLs have been detected for flowering time phenotype, the x-axis the chromosomes are listed and the vertical small black lines represent the genotyped markers. The red horizontal line defines the threshold LOD score of 3.9. Every QTL line above this line is significant (A). (B) A heat map was constructed using scantwo, where the upper triangle reports LOD scores for an additive and the lower triangle a full

1.2.3 GBS applied to a *Sorghum bicolor* mapping population

In this section we will present the genotyping on a *Sorghum bicolor* mapping population for investigating potential genes for chilling tolerance. Sequencing and phenotypic data were provided from Wubishet A. Bekele, Ph.D. Student of the group of Rod Snowden located at University Gießen. The analysis of the sequencing data, genotyping and additional QTL analysis will be described in the following sub sections.

1.2.3.1 The S. bicolor genome

S. bicolor is the world's fifth most important grain crop plant in the world. It is commercially used in northeast Africa and southern plains of the United States (Paterson et al., 2009). *S. bicolor* is a diploid plant and has a relative small genome size ~730 Mb distributed along 10 chromosomes. A draft genome was assembled using whole genome shotgun sequencing in 2009. The analyzed *S. bicolor* genome is 4.9x larger than the one of *A. thaliana*. The repetitive content was estimated to be 61% (Paterson et al., 2009). To increase the area for production towards the northern part of the world (Europe, America and north Asia) new varieties have to be developed to cope with the colder climate environment. Hence a project was established to investigate the underlying genetic network regarding the chilling tolerance of *S. bicolor*.

1.2.3.2 Plant material, SNP marker estimation and sequencing results

A mapping population was established by crossing cold resistant M71 (grain *Sorghum*) and cold sensitive SS79 (sweet *Sorghum*) both lines were already described by Shiringani, Frisch, & Friedt, 2010. The mapping population was propagated into the F_6 generation and afterwards selections of samples were selected for sequencing. Samples were selected and distinguished in respect to their response to cold. For each type 30 samples were collected, phenotyped and sequenced using a HiSeq Illumina sequencer.

The parental lines were resequenced with genome coverage of 5x for M71 and 6x for SS79 using a HiSeq Illumina sequencer. We used the published reference sequence SBi version 1.0 from PlantGDB (Duvick et al., 2008)

(ftp://ftp.jgi-psf.org/pub/JGI_data/Sorghum_bicolor/v1.0/Sbi/assembly/Sbi1/), with an assembly size of 659,229,367 bp. We applied our marker filtering pipeline (see Methods 1.1.2) for both parental lines, resulting in a final set of 143,166 SNPs markers (Table 8). The markers were nearly uniformly distributed along the ten chromosomes, except for chromosome seven were we observed a decrease of marker density which could not be explained by chromosome size (Table 9).

We obtained resequenced data from 60 samples containing 8,325,874 100 bp reads on average. On average 7,086,200 were aligned against the reference sequence where on average 4,970,151

of these reads were aligned uniquely (Figure 18A) resulting in an average coverage rate of 0.78x per sample. Only one sample did not receive enough sequenced reads for genotyping. The highest coverage was 0.98x (Figure 18B).

Table 8 SNP marker filtering results

	S. bicolor accession	
	M71	SS79
Raw SNPs after alignment against reference sequence	2,951,433	4,603,750
After removement of chloroplast and mitochondria SNPs	2,783,472	4,337,632
After InDel removement	2,605,321	4,106,571
After quality check	222,540	368,096
After TE check	212,363	349,269
After vicinity check	188,837	314,597
After merging and segregation check	143,166	

Table 9 Number of SNPs per chromosome

Chromosome	Chromosome length in bp	Number of SNP markers
1	73,840631	18,122
2	77,932,606	18,514
3	74,441,160	14,476
4	68,034,345	16,549
5	62,,352,331	15,6865
6	62,208,784	11,386
7	64,342,021	7,653
8	55,460251	10,854
9	59,635,592	15,253
10	60,981,646	14,673

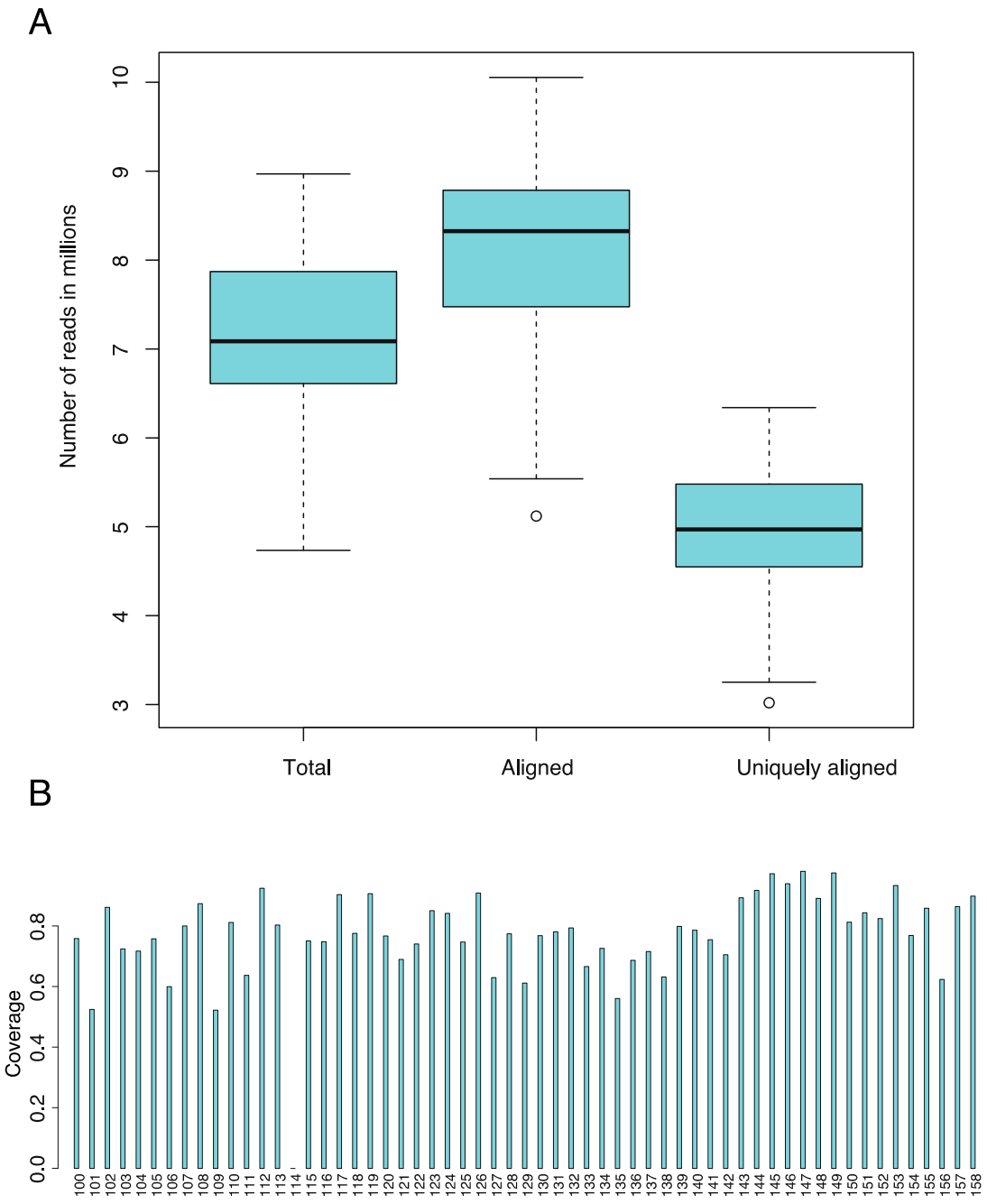


Figure 18. A) shows the result of sequencing 60 sample using multiplexing NGS regarding total number of reads, total aligned reads towards the reference sequence and how many reads were unique. B) shows the for each sample the achieved coverage rate after alignment to the reference genome.

1.2.3.3 Reconstruction of the mosaic structure for each sample

We applied our pipeline TIGER to reconstruct the parental genotypes as described in the Methods part. TIGER could successfully genotype 39 of 40 samples after applying a coverage threshold. Additionally, compared to the other two projects we had to adjust the expected Mendelian ratios as they were set up towards a segregating F_2 population. The generated data was derived from F_6 lines, which represent almost complete inbreed lines. The expected amount of heterozygous regions is 3.13% across all genomes and 48.44% for either of the homozygous parental regions.

By analyzing the overall local parental allele frequencies in our data before applying TIGER, we found a much higher level of heterozygosity of ~14.62% (Figure 19), indicating that there might be a high rate of misalignments generating false heterozygous calls. After genotypes were predicted using TIGER we estimated the genotype frequencies from the reconstructed samples (Table 10). Inter-marker distance was of 13,701 bp on average using 143.166 SNP (median: 2007 bp) (Figure 20A). The genotyping revealed a higher number of double CO compared to the previously described mapping populations, 20% were <400kb (Figure 20B). The average length of double COs was 3,005,925 bp. In this analysis we did not filter out small islands as within RIL (F_6) populations small double COs are expected to appear and additionally those small islands could indicate problems with the reference sequence if the sequencing data was supporting them.

Table 10 Genotype distribution after genotyping with TIGER

Genotypes	Frequency in %
M71	52.96
Heterozygous	0.06
SS79	46.98

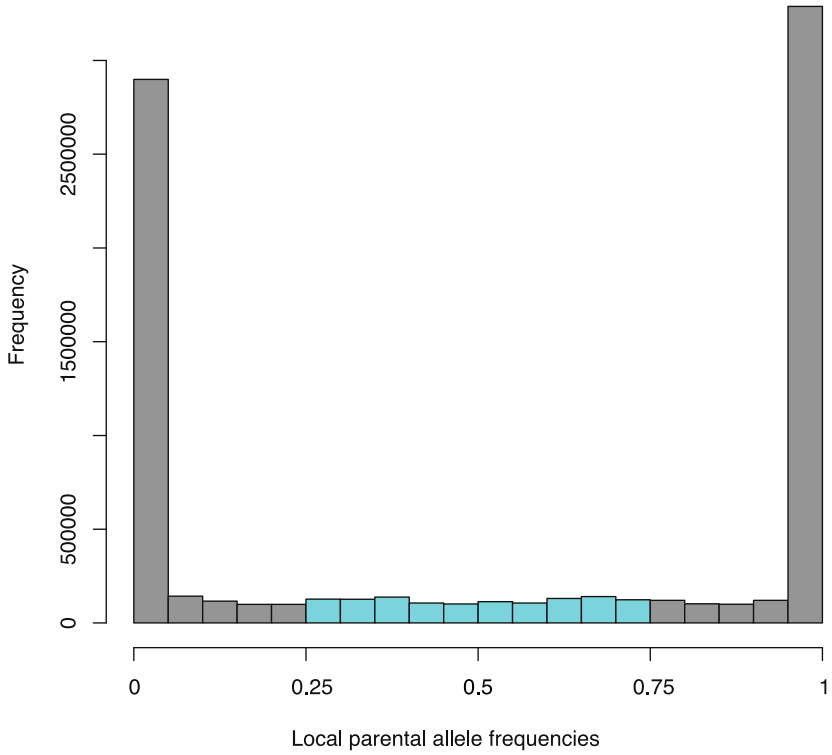


Figure 19. Histogram of allele frequencies as assessed on markers.
Heterozygous blocks are blue (based on the rough sliding window labeling), x-axis are the supported allele frequency for either of the parents (Left M71 and right SS79) and y-axis are the counts.

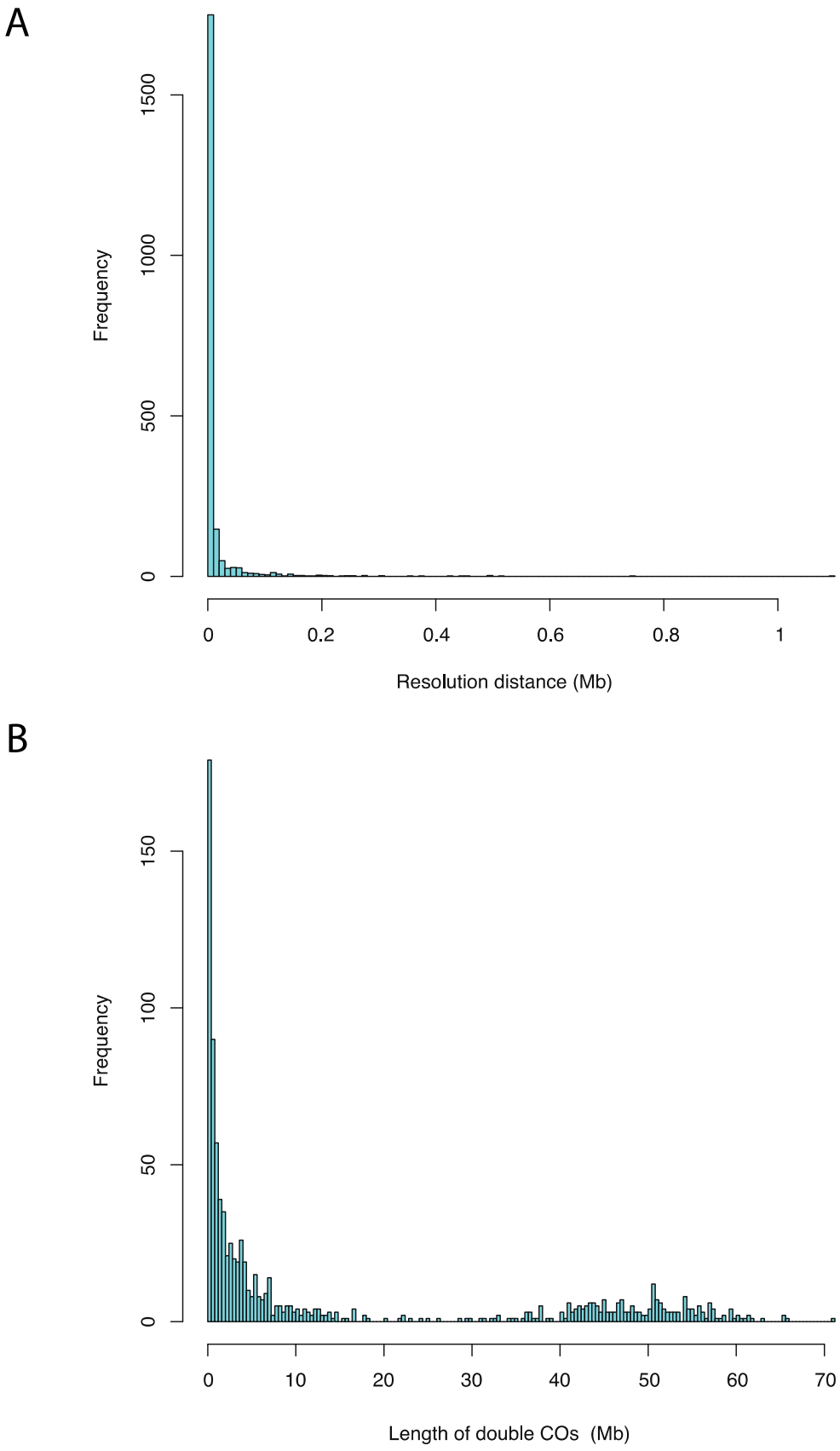


Figure 20. A) Inter marker distance between predicted breakpoints, x-axis in Mb and y-axis frequency in counts. B) Length distribution of double COs, x-axis in Mb and y-axis in counts.

1.2.3.4 Detection of selection pattern

During the selection of the SNP markers with our pipeline and after genotyping we observed a pattern of selection at the end of the chromosome arm four (starting from 65Mb) increasing the SS79 allele within the population. Normally a parental allele frequency of around 50% would be expected if the population were not selected for any trait. However, we have here a population which has been selected for both extreme phenotypes against cold. Therefore we would expect that allele frequency vary around 50% but not at 75 or 25 % allele frequency. We found an atypical pattern of selection where the allele frequency reaches nearly 80% for SS79 (Figure 21), which could represent a selection or segregation distortion. Those positions have been removed for further downstream analysis as it most likely introduced by selection for increased sugar content based on the information given from our collaborators.

1.2.3.5 QTL-analysis

Phenotypes for cold tolerance fo r the resequenced samples and with the combination of the predicted genotypes a QTL-analysis were performed. We applied the same strategy as described in the previous sections to filter the marker. This resulted in 1438 genotyped markers. By applying a composite interval mapping (CIM), we were able to identify QTLs at chromosome two and six and additionally one new QLT at chromosome eight (Figure 22A). Using the more robust MQM method (Arends et al., 2010) we identified three QTLs, one at chromosome two (56,370,532 – 57,243,381 bp) and two at chromosome four (47,981,352 – 49149112 bp and at 51,371,726 – 51,677,869 bp) (Figure 22B).

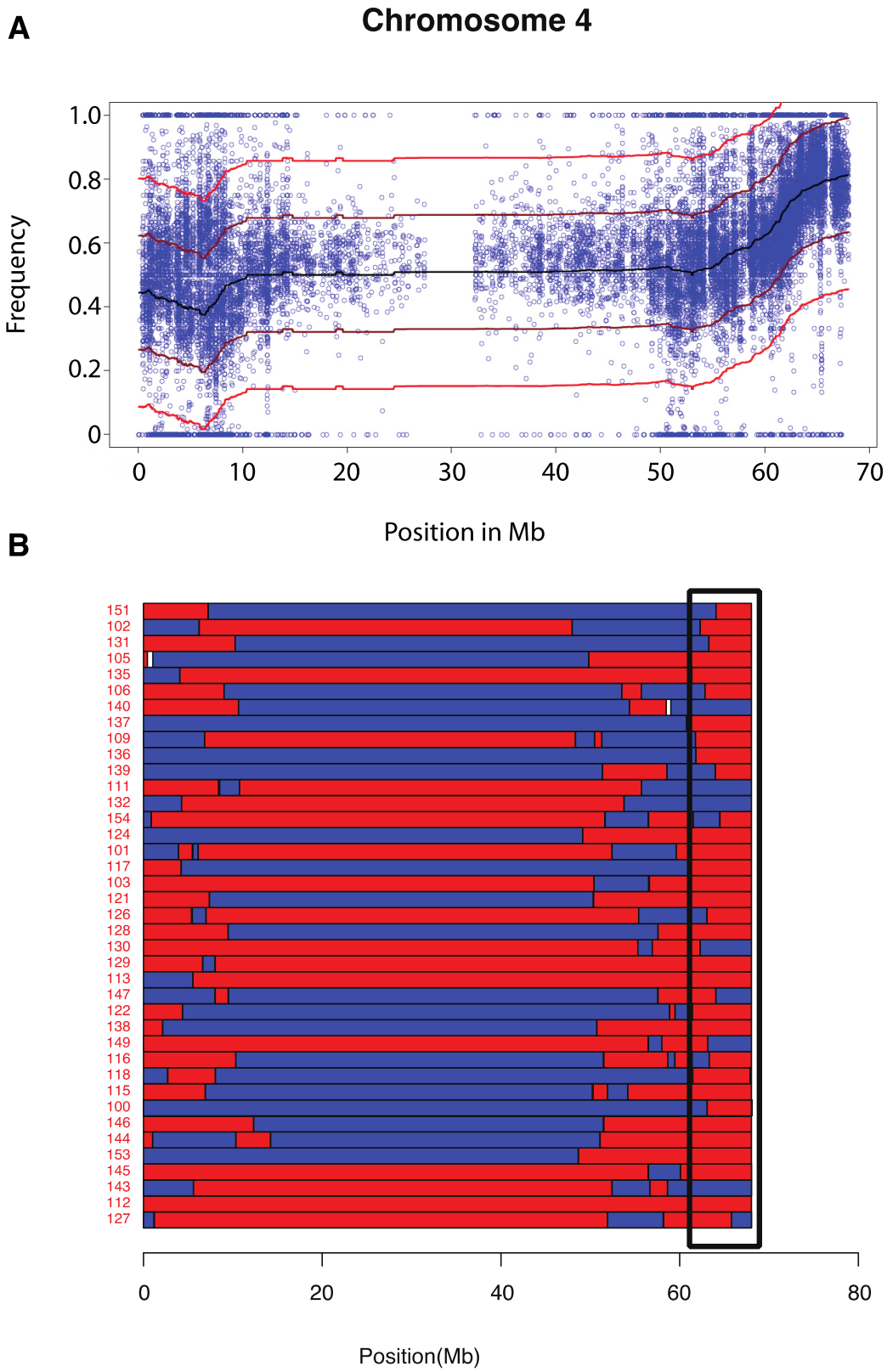


Figure 21. A) Allele frequencies for each SNP marker on chromosome 4 indicating a selection/segregation distortion at the end of the chromosome arm for M71, x-axis in Mb, y-axis the frequency in ratio for M71. The black line is the mean value for the distribution and brown and red the first or second standard deviation. B) This shows the individual genotypes of our samples, red SS79, blue M71, where we can observe the same enrichment for SS79 at the end of the chromosome arm. (x-axis in Mb and y-axis are the samples).

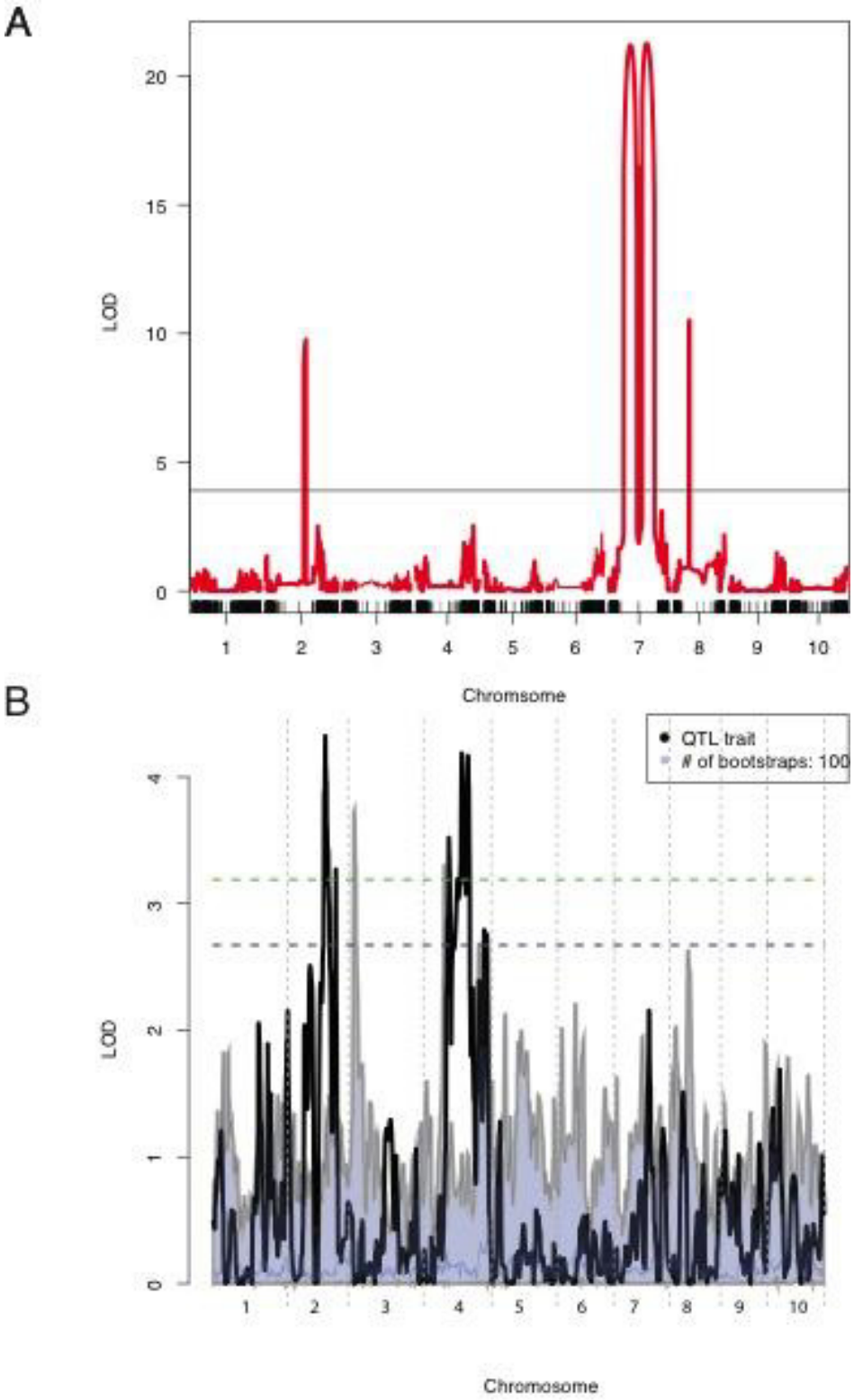


Figure 22. QTL analysis using CIM and MQM

CIM method detected four QTLs (at chromosome 2, two at chromosome 7 and one at chromosome 8). A). Applying MQM we find two QTLs (at chromosome 2 and two for chromosome 4). The QTLs were significant (tested against random bootstraps of the data). B). The x-axis shows the number of chromosomes and the y-axis the LOD score.

1.3 Discussion and conclusion

This chapter will discuss and conclude the quality of applying the TIGER pipeline using sparse sequenced data of bi-parental mapping populations based on the previous presented projects.

1.3.1 Genotyping by sequencing pipeline TIGER

We presented the design of TIGER and showed applications for genotyping by sequencing for three different taxa having sparse sequencing data. Imputing missing genotypes allowed us to apply QTL analysis with higher resolution, where we found candidate genes, which now can be followed up by fine-mapping.

We developed TIGER to genotype any bi-parental mapping population with a single constraint: the crossing depth of the sample. Our method calculates HMM transition probabilities sample-wise, hence each sample is imputed with its own sample specific error rate. We showed and proved the quality and stability regarding the coverage rate of the genotype prediction by using simulation data. The overall error rates were quite low (less than 3%) even at a simulated coverage of 0.1x, indicating that genotype and CO predictions were robust. By analyzing the errors we identified the source of wrong genotypes. In general, we speculate that the observed error types are common for all imputation based on HMMs but have simply not been reported for the other tools. The regions, which showed a tendency for errors, were located near the ends of the chromosome and at the centromeric regions, most likely because the chromosome ends miss additional information at one side of the prediction. For the centromere regions we have two possible explanations for the increased error rate. First we do not expect any recombination at these heterochromatic regions and secondly a drop of marker density, a high amount or repetitive elements and unassembled regions make these regions complicated to access with short read alignments. Misleading SNP markers at the border of these regions can cause the erroneous COs.

Another type of error is the misinterpretation of heterozygous regions as homozygous regions due to low coverage rates, because the second allele has not been observed. Those errors could lead to wrong interpretation of the data for further downstream analysis e.g. introducing a false segregation distortion or in the worst case a wrong QTL region. Those errors get even more frequent if the information content were further reduced by not considering all possible read information per marker position e.g. only one read per marker would be used (Andolfatto et al., 2011). We showed that TIGER could correct for sequencing biases towards one parental allele. By applying the TIGER pipeline, the best result was achieved for *A. thaliana* to a resolution of two kb. Especially this CO resolution could be resulting from a combination of a less-biased representation of the genome and more accurate sample wise genotype predictions. In general we observed that

the quality of the reference sequence has an impact on the quality of the correct genotype prediction, as shown for *F. vesca* and *S. bicolor*.

1.3.2 Appearance of "islands"

The highly repetitive content of *F. vesca* can be one reason why we have obtained an increased amount of island structures as compared to *A. thaliana*. Another reason could be the quality of the reference sequence. The *A. thaliana* reference sequence was published in 2000 (The Arabidopsis Genome Initiative, 2000) and was since further improved (Lamesch et al., 2012). The reference sequences of the Strawberry and *Sorghum* are based on short reads and have been not updated regularly. Hence, they could still contain wrong assembled regions with respect to length or ordering. The errors in the reference sequence of H4x4 could result from the molecular marker-based scaffolding (Shulaev et al., 2011), which was the possible cause for the unexpected high number of rearrangements seen earlier and could be the reason for our "island" problem as well..
A similar observation was made during the validation using SSR markers for the genotyping of the strawberry mapping population, where a whole chromosome was not matching our genotyped data. As TIGER has no bias for particular chromosomes we speculate that there must be certain problems with the reference sequence.

Errors in the reference sequence are a problem for GBS prediction as it could lead to a false genotype prediction which could affect downstream genome-wide association study (GWAS) or QTL analyses. To avoid such effects, potentially false double COs were removed using fixed thresholds. However, this approach is not optimal because for each data set as the threshold was arbitrarily selected and for double CO sensitive experiments real COs might have been removed. Nevertheless such island structures can have several reasons: As we already have explained above such structures result from errors in the assemblies. But additionally there are also biological reasons for such an appearance like rearrangements or unusual recombination events. To identify real islands from false positive results a local realignment including a local assembly can be done. These steps could also answer if the used reference sequence might contain an error. Further having long read sequences spanning the whole island region could help to decide the result if the observed island structure is real or an error.

1.3.3 Future improvements

To improve the resolution one might think about replacing the simple gap filling between the markers next to a recombination break by applying a focused HMM. Such an additional HMM takes the filtered markers at such locations and tries to find the best position to place a breakpoint.
The prospect of having longer reads, i.e. produced from third generation sequencers combined with sparse sequenced data allows to genotype more accurately and improve the detection of

smaller structural variations. Using longer reads also allows to determine if an island structure results from an error in the reference sequence, a translocation or a true recombination event.

2. Genotyping Multi-parental RIL populations

This chapter describes the work of genotyping plant individuals, where the individual genome can be inherited from up to four parental genotypes. The work is based on the *Arabidopsis* multi-parental RIL (AMPRIL) (Xueqing Huang et al., 2011) population, which have been previously genotyped with 300 markers. The goal of this project is to genotype the population with a higher density of markers by using a NGS approach and the concept of TIGER for genotyping. The final goal is to use the dense information of genotyped markers including the phenotypic values for each individual to apply a GWAS. For that purpose the whole population have been replanted, phenotyped and prepared for sequencing by Petra Pecinkova. In this work the resulting NGS data was used for genotyping and my colleague Jonas Klasen used the genotyped markers for GWAS.

2.1 Introduction

In general a bi-parental population has a high potential for identifying QTLs because of its balanced minor allele frequency of 0.5. However, the resolution of the QTL is low. It can vary between several Mb or half of a chromosome because of the low rate of recombination events (Kover et al., 2009). In contrast to bi-parental mapping populations, natural populations have a higher resolution of the QTL region through higher number of ancestral recombination. The disadvantage of these populations is that the minor allele frequency converges towards 0 by carrying many rare alleles. This reduces the likelihood of detecting QTLs. Furthermore, population structures appear, which leads to false positive predictions of QTLs. Population structure is introduced by sub-populations in the population. To reduce the prediction of false positive QTLs population structures have to be considered during the analysis (Kang et al., 2008; Price, Zaitlen, Reich, & Patterson, 2010; Reich & Goldstein, 2001).

To achieve high resolution for QTL detection and QTL positioning multi-parental RILs were introduced. This allows combining ancestral and recent recombination events (Cavanagh, Morell, Mackay, & Powell, 2008) and the minor allele-frequency is 1/ number of parents of the population. Thus the likelihood of detecting QTL is high only if the number of parents is reasonably small (Cavanagh et al., 2008). The emergence of population structures is less likely compared to natural populations. Multi-parental RIL populations have been already established in different taxa e.g. mice (Talbot et al., 1999), *Drosophila melanogaster* (Macdonald & Long, 2007), *A. thaliana* (Kover et al., 2009) and in wheat (B. E. Huang et al., 2012) and are now developed for many other taxa.
For multi-parental populations genotyping is more challenging as compared to bi-parental populations, as bi-allelic SNP markers can only distinguish two parental alleles, and thus the alleles in a line derived from multiple parents might not always reveal a unique parental genotype.

The AMPRIL parental population is founded by eight *Arabidopsis* accessions Antwerp (An-1), C24, Colombia-0 (Col-0), Cape Verde Islands (Cvi), Eringsboda (Eri), Kyoto (Kyo), Landsberg erecta (L*er*) and Shahdara (Sha), which have been selected from different climatic regions to build up a genetic pool based on their genetic diversity and phenotypic variation (Figure 23A). The population itself is divided into 12 sub-populations, where each sub-population is based on a cross of four parents. F_1s have been crossed with other hybrid F_1s to generate a F_2 population, containing the mixture of four parental lines. The F_2 populations were further selfed multiple times (Figure 23B). The sub-populations were constructed using a crossing scheme, which separates the entire AMPRIL population into two broad populations (population I and II) (
Table 11). The difference between population I and II are the combinations of parents for each sub-population and the number of inbreeding generations of the respective population and as well population I was already been genotyped with a low number of markers (Xueqing Huang et al., 2011). Population I was propagated until F_5 and population II until F_7. Additionally for the AMPRIL population phenotypic data (i.e. flowering time) were collected for each individuals of each sub-population.

For genotyping the AMPRIL population we could not apply the previously presented method TIGER, as we cannot estimate the probabilities using the beta mixture-model. The allele distribution is only accounting for two parental genotypes, whereas for each of the AMPRIL sub-population we have to consider 10 possible genotypes (four parental and all heterozygous genotypes). Therefore we developed a two stage HMM approach building up on the ideas implemented in the earlier approach. We describe in the methods sections the construction and analysis of the AMPRIL populations, and afterwards the outcome of the prediction of the genotypes of the AMPRIL population.

Table 11 Crossing scheme for generating the AMPRIL population

The first column shows the parental combination for each of the populations and the second column the combination of the resulting F_1 based on the cross of column one.

Population I	X	
A	Col-0	Kyo
B	Cvi	Sha
C	Eri	An-1
D	Ler	C24

Crossing scheme	A	B	C	D
A	-	BA	CA	DA
B	AB	-	CB	DB
C	AC	BC	-	DC
D	AD	BD	CD	-

Population II	X	
E	Col-0	Cvi
F	Sha	Kyo
G	Ler	An-1
H	Eri	C24

Crossing scheme	E	F	G	H
E	-	FE	GE	HE
F	EF	-	GF	HF
G	EG	FG	-	HG
H	EH	FH	GH	-

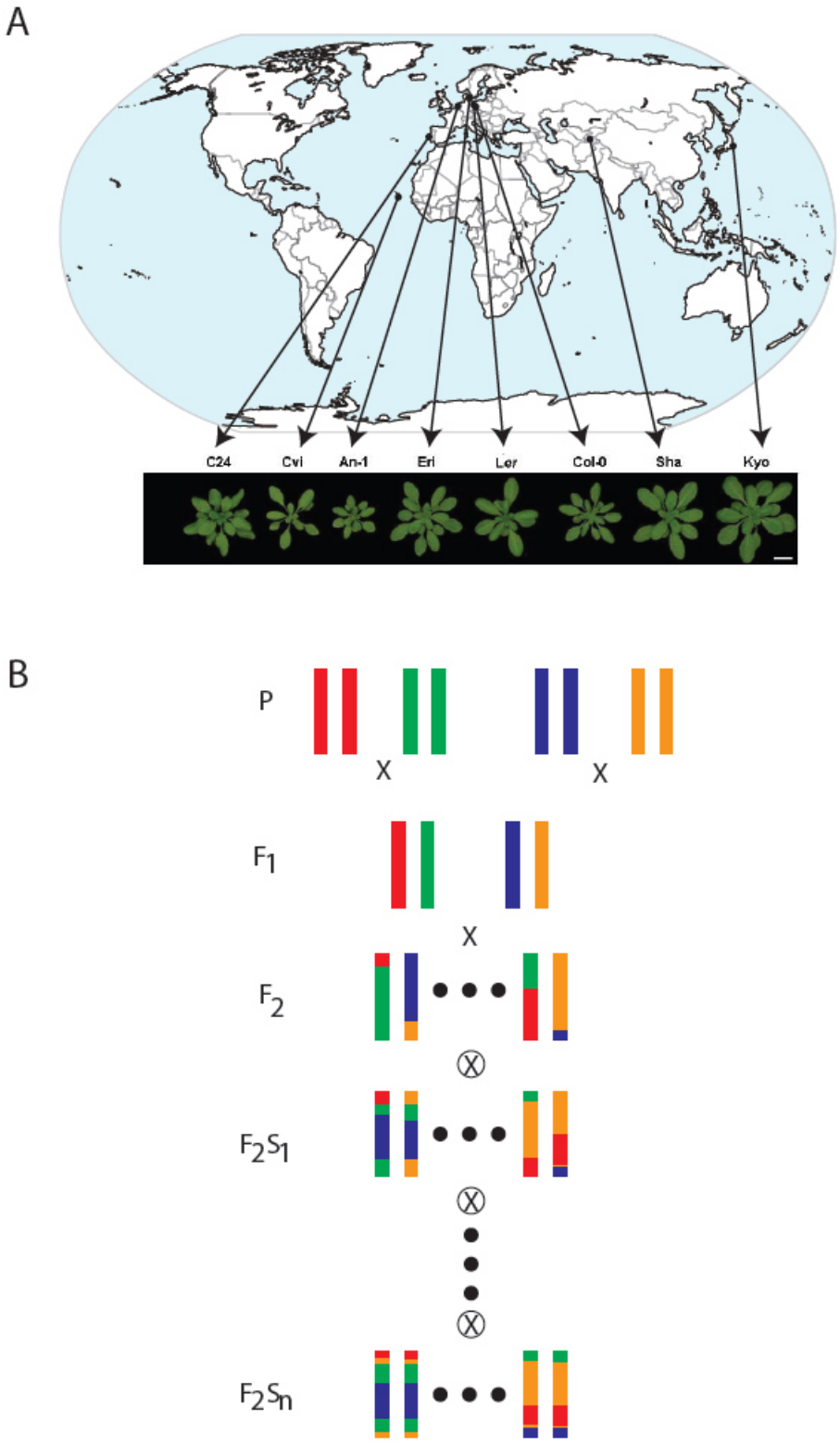


Figure 23. A) Shows the origin of the eight founder lines of the AMPRIL population. B) Shows the crossing scheme of one sub-population of the AMPRIL population representative for all subpopulations.

2.2 Method

2.2.1 Resequencing the samples from the AMPRIL population using RAD-seq

The construction and preparation of sequencing library was done by Petra Pecinkova, member of the group of Ales Pecinka in the Department of Maarten Koornneef at the Max Planck Institute for Plant Breeding Research.

For resequencing of the progeny of the AMPRIL population, we decided to use RAD-sequencing. RAD-sequencing is based on restriction enzyme digestion of DNA. After digesting fragments are selected and further used to prepare a sequencing library (Baird et al., 2008). RAD-sequencing reduces the resolution for genotyping, as only regions with the restriction site will be sequenced. However at the same time it enriches reads coming from the restriction site to have a higher coverage at each restriction site allowing for multiplexed sequencing of hundreds of individuals. Higher read counts allow for more accurate genotyping, in particular for heterozygous allele frequency present in the subpopulations. We selected the restriction enzyme CviQL, a four cutter with cutting pattern G'TAC and an expected cutting frequency of 235,933 restrictions sites based on the reference sequence TAIR10.

To resequence 1,100 samples from the AMPRIL in a cost and time effective way we used a multiplexing system with 210 barcodes. Each barcode was 12 bp long. The sequences of the barcodes were selected to have not a particular nucleotide bias to avoid sequencing bias during short read generation using Illumina sequencing. The final sample reads were de-multiplexed and aligned using the Shore pipeline (Ossowski et al., 2008; Schneeberger et al., 2009)

2.2.2 Assignment of genotypes at each marker positions

Based on the experience how to genotype sparse sequencing data for individuals from bi-parental mapping populations we developed a similar pipeline for genotyping the AMPRIL populations. The first step is to identify all possible SNPs markers, which can be used for genotyping. Hence, SNP markers from the parental lines were obtained by resequencing them and aligning them to the reference sequence of *A. thaliana* (The Arabidopsis Genome Initiative, 2000). Afterwards all possible SNPs candidates have been filtered accordantly to the same standard as already described in the method section of TIGER (section 1.1.2). From the resulting parental SNP marker information we created the combined SNP list for each sub-population based on the combination of the four parents. For not observed SNPs marker positions in either one of the four parents, the reference allele was used. If the marker allele at a certain position was equal for all four parents, the position was removed from the SNP list. From here on we only consider bi-allelic SNP makers.

Based on the SNP marker list, each sample from each sub-population was pre-genotyped using the aligned short reads and the extended version of the equation 1 (Equation 1). To apply genotypes at a marker position based on the short read data we test initially if five or more reads

are supporting only one of the two possible alleles. We used here the same threshold as already introduced in the method part of TIGER. The SNP position for that sample will be labeled homozygous for that particular allele, otherwise the probabilities for homozygous or heterozygosity alleles are calculated. Even having four parents we can only observe two alleles at each marker position. Therefore we can reuse the same calculation as already previous described (section 1.1.3). After consideration of the correct allele, we have to genotype that marker position. The correct genotype is the combination whose parents supporting the same allele at that marker position. To resolve non-informative SNP markers or to correct wrong genotyped markers two HMMs are used afterwards.

2.2.3 Two stage Hidden-Markov-Models

For imputing and correcting genotypes of homozygous and heterozygous regions we developed two different HMMs specialized either for the homozygous or the heterozygous part of the genome. The HMM predicting homozygous regions consist of five hidden states, which are fully connected. The hidden states are A, B, C, D and Z. The states from A-D represent the four parents of a sub-population and Z stands for uncertainty between homozygous or real heterozygous regions (Figure 24A). The emission states represent the possible observation values, which are the results of the genotyped marker position from the pre-genotype step.

Regions labeled by the model to be Z are further analyzed by a second HMM. The second HMM contains 10 hidden nodes, where each hidden states has 10 emission states. The 10 hidden states summarize all possible outcomes of a cross between four parents, four homozygous and six heterozygous genotypes. The hidden states are also fully connected (Figure 24B). The second model was used to resolve heterozygous regions. The second model only considers SNPs markers, which are unique for one of the parents at that marker position. Removing such markers in possible heterozygous regions facilitate the prediction of the correct genotype. Additionally, to reduce errors at the border of the labeled Z regions, we included the flanking regions of to 100 markers allowing placing a possibl e CO breakpoint more accurately but it was not force to start or end with these particular genotypes. Further the second HMM tests if there really exists a CO at that block label as Z.

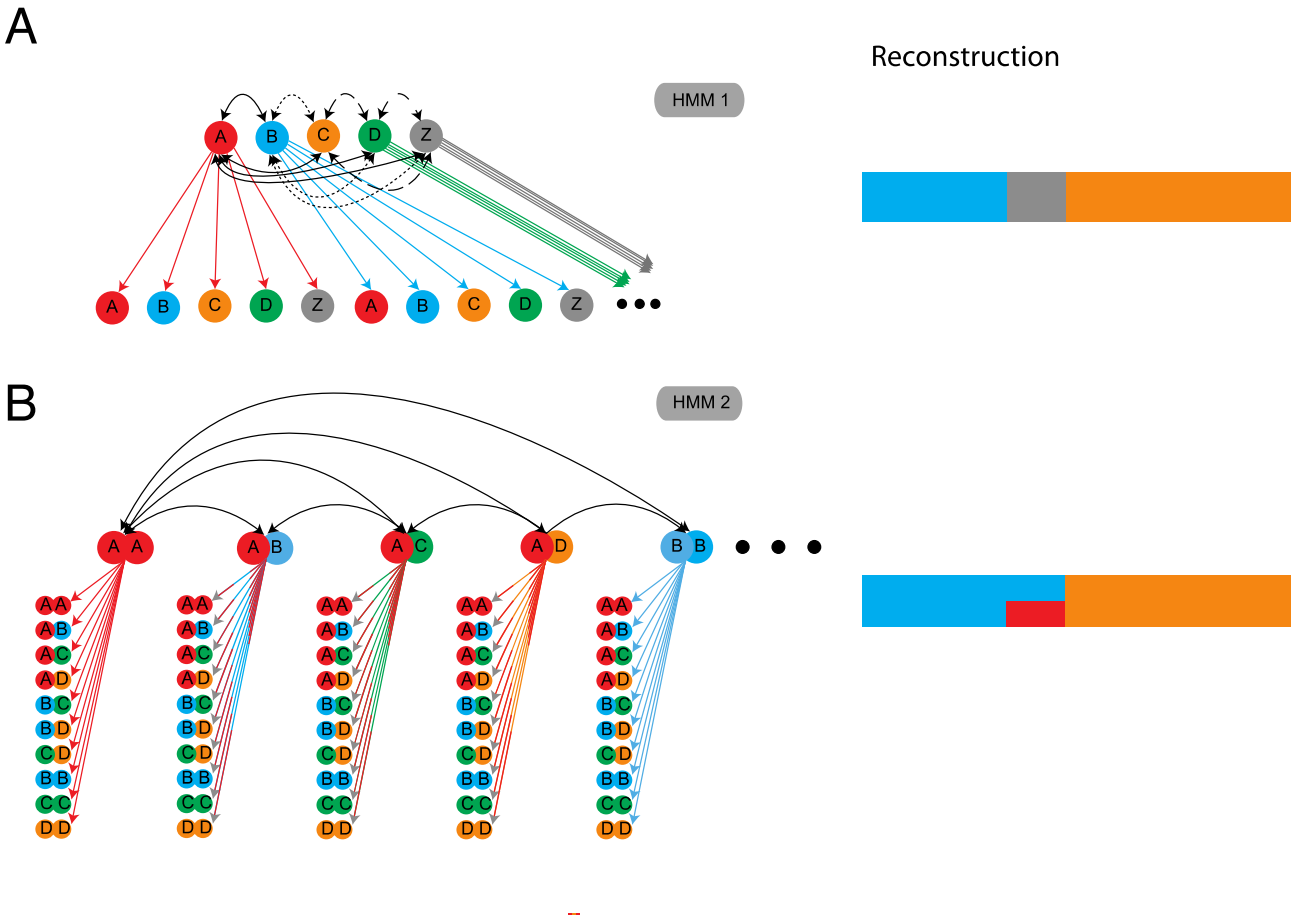


Figure 24. HMMs for predicting and correcting genotypes in the AMPRIL population
Two HMMs are used for predicting the genotypes using sparse sequencing data. A) The first model assigned blocks of homozygous genotypes and used Z as a label to mark not resolved regions. B) The second HMM used only unique alleles (only supported by one of the parental genotypes) to resolve heterozygosity.

2.2.4 Visualization of the allelic support of four parents

Sliding windows were used to visualize the result for the support for each possible parent of a sample for each SNP marker. Eight sliding windows (Figure 25) were used to represent each of the eight parents. And for each sliding window only unique SNPs supporting clearly only one of the parent's alleles were selected, no allele sharing was allowed. In principle the combination of the sliding window approaches can be used for genotyping but it would lack resolution where exactly the genotype might change. With these eight sliding windows we were able to compare the prediction from the HMM models. It was in particular quite useful for identifying outcross events between different sub-population as the HMMs were designed to predict genotypes coming only from four parental lines. Hence, the HMM reported in such cases very small blocks of heterozygous genotypes leading to an increased CO rate.

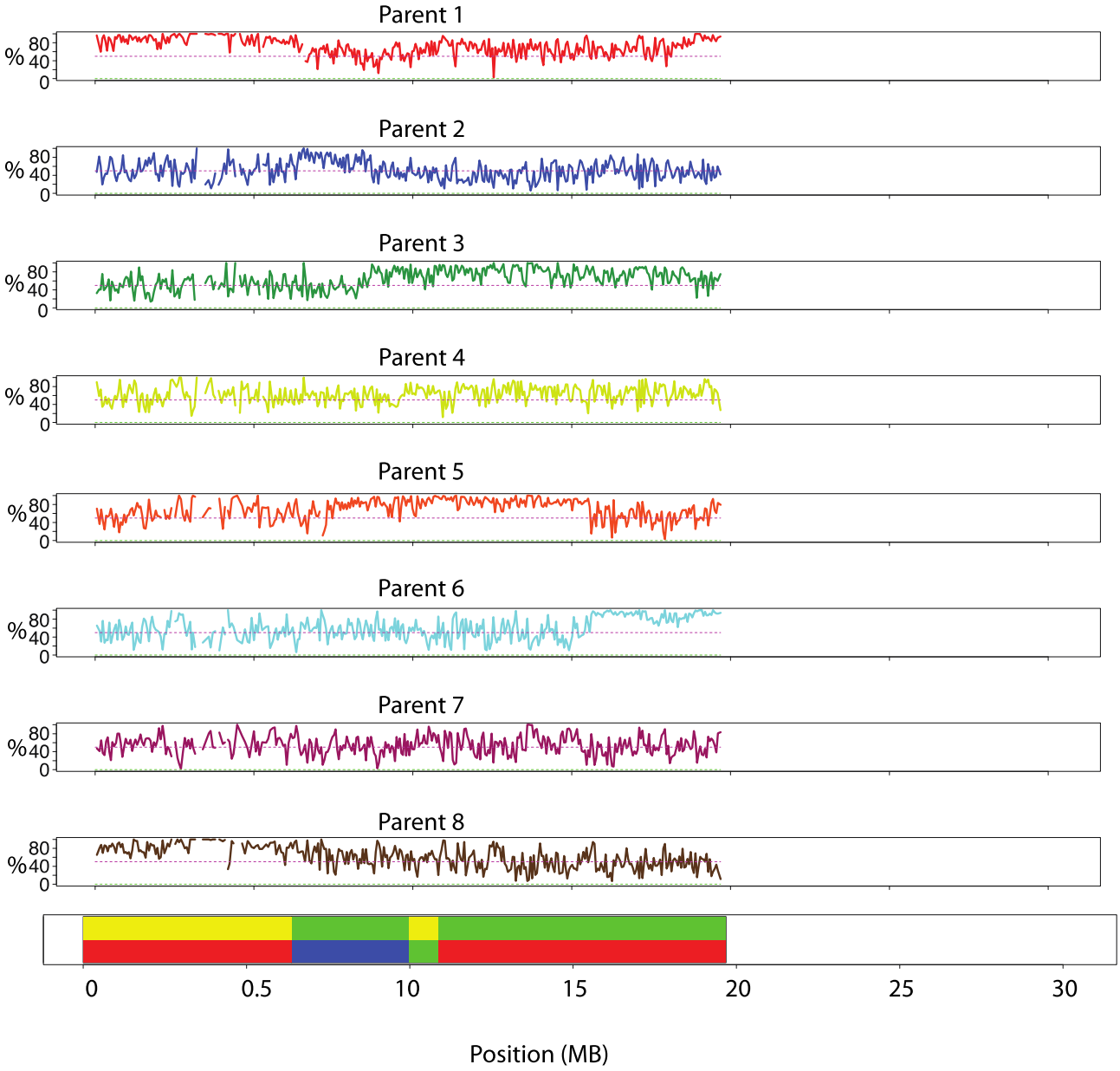


Figure 25. Sliding window of eight parents for chromosome 1 for an outcrossed individual
Each sliding window shows marker positions supporting one of the eight parents. The last panel shows the output after genotyping with the HMM model. This example shows how outcross events can produce wrong genotype patterns/predictions. The x-axis measured in Mb and the y-axis shows the percentage support of that parent at that position.

2.2.5 Simulations and training of the HMM

Simulated genotypes were used for training and validation of our approach. We simulated a segregating mapping population with the same crossing scheme as the AMPRIL subpopulations and with 5,000 individuals with three different coverage rates (0.1x, 1x, and 10x) including 1.2 million markers and the default recombination rate using an extended version of the tool Pop-seq (James et al., 2013; Salomé et al., 2012). The extended version was necessary to allow for simulating multi-parental mapping populations.

To train the HMMs we estimated the probability matrix for the transition and emission using first a supervised learning strategy by selecting 1,000 randomly simulated samples in-silico sequenced with an average coverage rate of 1x. After the probabilities were calculated we applied the resulting HMM to 10 randomly sequenced samples and changed slightly the emission probabilities manually by adding or removing values in the range from 0.01 – 0.001 until the genotype prediction were nearly correct with respect to the short read data. To avoid over fitting based on the small testing set, we ran the new model on all samples again and select again 10 new randomly selected samples. For the new 10 samples again we compared the genotype prediction having the support from the short read data. If the prediction was overall not correct, the emission probabilities were further stepwise changed and a new model was applied to the total sample set again. The procedure was performed until the chancing furthers the emission values resulting in worse genotype predictions. To test for possible over-fitting, the HMMs were applied to a 1,000 new random simulated samples sequence with 1x coverage. To validate the prediction of our model a comparison between the predicted genotypes of AMPRIL data and previously genotyped 300 SNP markers was performed.

For the second HMM the probabilities were trained based on the simulation population used for training the first HMM but only selecting regions which were heterozygosity. Afterwards the emission probabilities were similar manual improved and validated as described above.

2.2.6 Genetic incompatibilities

Having high-resolution genotyping data and a high sample size gave us the possibility to analyze the genotype frequency across the entire genome. To identify regions in which particular allele combinations were underrepresented. The idea is to compare the expected genotype frequency based on the presented crossing scheme against the observed genotype frequency, assuming that no bias where introduced based on selection. This analysis can be done using a Chi-square test (X^2) testing for independency. To apply the X^2 test, data reduction has to be applied, as comparing a two million x two million matrix with a X^2 test for each cell is expensive in computer time and memory. We can reduce the data by only comparing marker positions, which are located next to a CO. After data reduction we have to handle heterozygous genotypes. Heterozygous genotypes can mask potential bias for certain parental combinations. Therefore we consider only

homozygous genotypes. Instead removing marker position having a heterozygous genotype we transformed them into two homozygous genotypes for each of the heterozygous genotypes. This procedure we done for each marker position leading to doubling our sample size and resolve heterozygous genotypes without losing any information. This is possible as we only compare parental counts at each marker position in the following static test, which is independent for each position. Afterwards for each sub-population a X^2-test is applied to genotype marker combinations. The resulting scores are corrected for multiple testing using false discovery rate (FDR) converting the X^2 values into q-values and selected those values fulfilling our selection criteria of a p-value of < 0.05 (Figure 26).

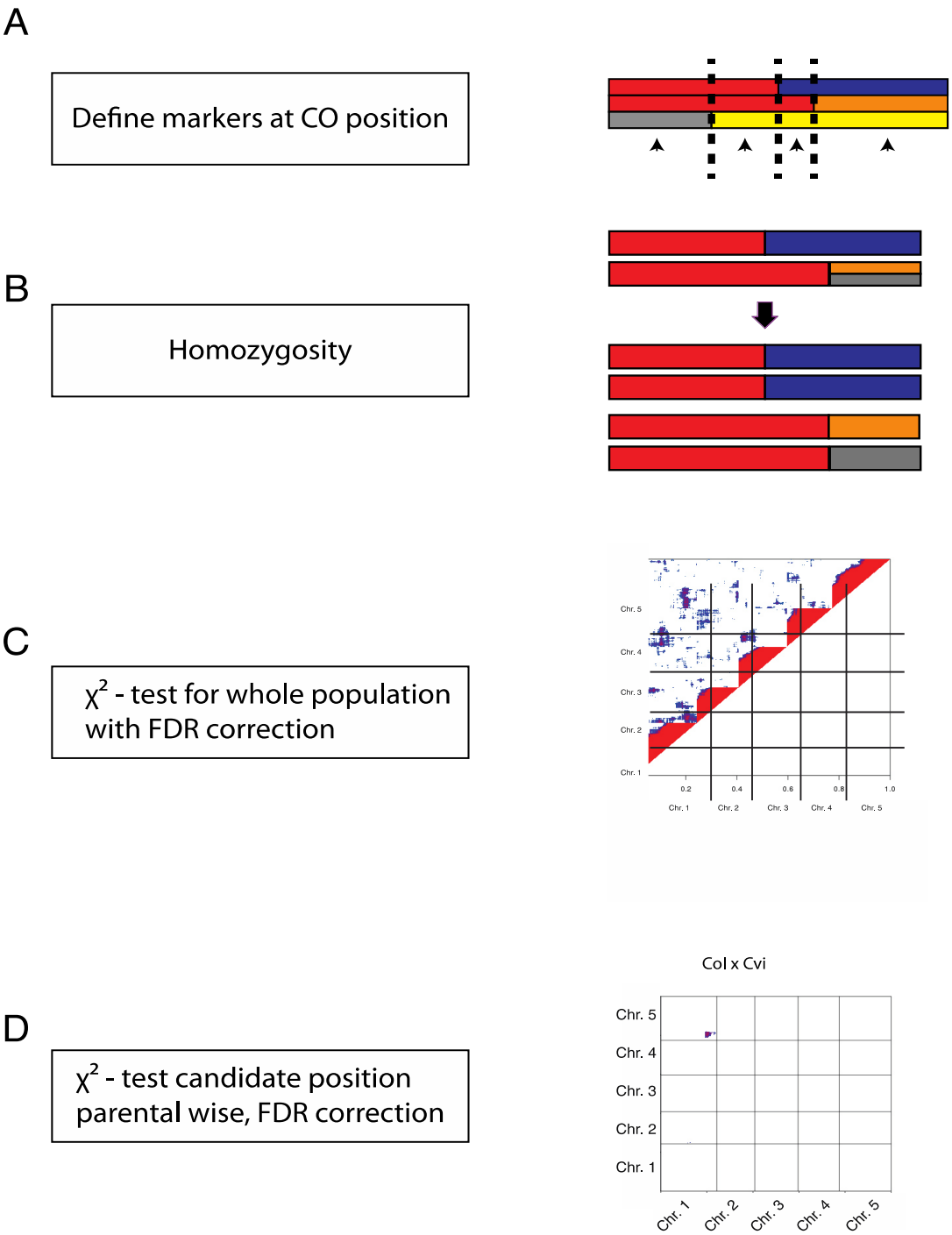


Figure 26. Testing for independence of unlinked markers to identify genetic incompatibilities

A) Representative marker position were defined based (black arrows) for each segment defined by the global observed CO blocks. Colours represent different genotypes. B) Doubling the sample size to convert heterozygous marker positions into homozygous representations at each marker position. C) Apply for each marker position a X^2 test, where the result has to be corrected by FDR. White to red colour indicates strengthens of correlation, where white has no and red has strong correlations. D) Screening for each sub-population for those marker which haven been found to be correlated, to resolve the affected genotypes

2.3 Results

This section summarized the results of the introduced methods. It covers the analysis from resequencing of the individuals and genotyping of the whole AMPRIL population.

2.3.1 Resequencing results of the founder lines of the AMPRIL population

Before genotyping, SNPs markers have to be determined for our eight founder lines. Deep resequencing of the parental lines using Illumina short read technology was done with an average coverage of 44.5x per accession.

The high coverage rate allows obtaining high quality SNPs after an alignment towards the reference sequence TAIR10 (Lamesch et al., 2012; The Arabidopsis Genome Initiative, 2000). For each parent we identified on average 300,000 SNPs (Table 12) after applying quality filtering, where SNPs were removed having either low mapping quality or located in repetitive regions.

Table 12 SNPs per parental line using high resequenced short read data

Accessions	SNPs
An-1	320,412
C24	364,553
Col-0	324
Cvi-0	464,009
Eri	339,142
Kyo	339,551
Ler-1	342,766
Sha	375,944

As each subpopulation is derived from four parents, we estimated the total number of SNPs per subpopulation. We removed SNPs shared along all four parents. Applying this rule we identified on average 1,318,564 SNPs (≈ 11 SNPs per kb) per sub-population (Table 13). For the downstream genome-wide association study analysis (GWAS), which acts on the entire population, a high-density SNP-marker list was generated containing 2,002,751 non-redundant SNPs (≈17 SNPs per kb) by combining all filtered parental SNPs. The final SNP markers were equally distributed across the genome. However, at the centromeric regions a drop was obtained and at the pericentromeric regions SNP density was increased, as expected (Figure 27).

Table 13 SNPs number per sub-population

SNP markers	Sub-population
1,401,305	AB
1,144,373	AC
1,215,206	AD
1,401,305	BA
1,444,973	BC
1,466,295	BD
1,144,373	CA
1,444,973	CB
1,239,342	CD
1,215,206	DA
1,466,295	DB
1,239,342	DC
1,401,305	EF
1359694	EG
1,386,591	EH
1,401,305	FE
1,253,633	FG
1,270,708	FH
1,359,694	GE
1,253,633	GF
1,239,342	GH
1,386,591	HE
1,270,708	HF
1,239,342	HG

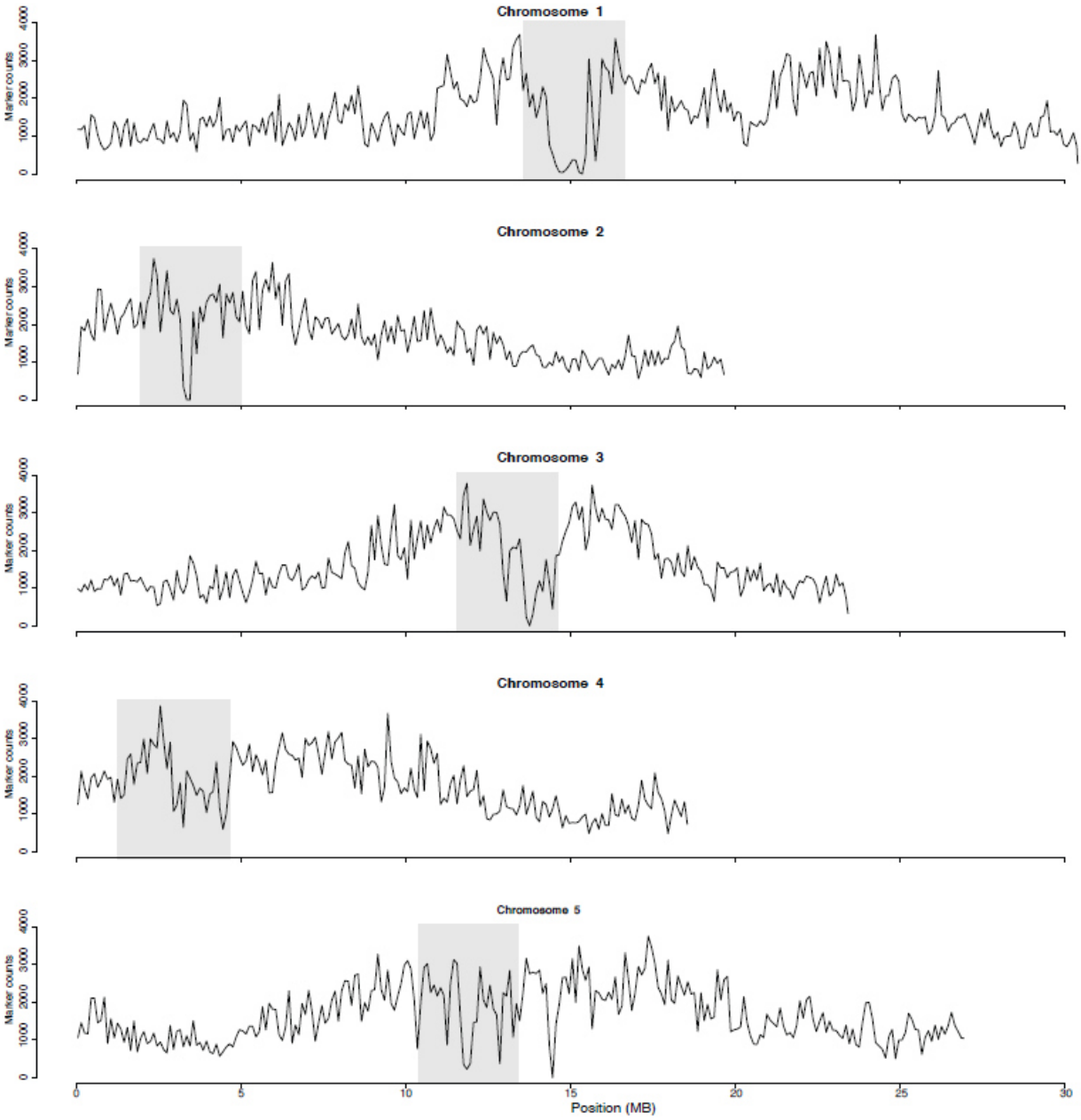


Figure 27. SNP marker density for all five chromosomes based on a sliding window of 100 kb. Around 2 million SNPs are distributed along the five chromosomes (an increase of SNPs can be observed at the pericentromeric regions and a drop at the centromeric regions (grey)).

2.3.2 Resequencing the AMPRIL population

Around 1,100 individuals were resequenced from the 12 different sub-populations following a RAD-seq library preparation (Baird et al., 2008). Nine Illumina sequencing lanes were used for paired end sequencing on the Illumina HiSeq 2000 system to produce short reads for all samples including six samples with biological replicates. In total 2,143,621,998 reads were produced, with an average length of 100 bp. After applying the Shore pipeline (Ossowski et al., 2008) for de-multiplexing, we observed that the numbers of reads per barcode were not uniformly distributed (Figure 28). It has been observed before that there could appear a barcode bias (Alon et al., 2011;

Andolfatto et al., 2011), which we also observed in our case. Certain barcodes were not sequenced well, resulting in a sparse sequencing representation of individual genomes.

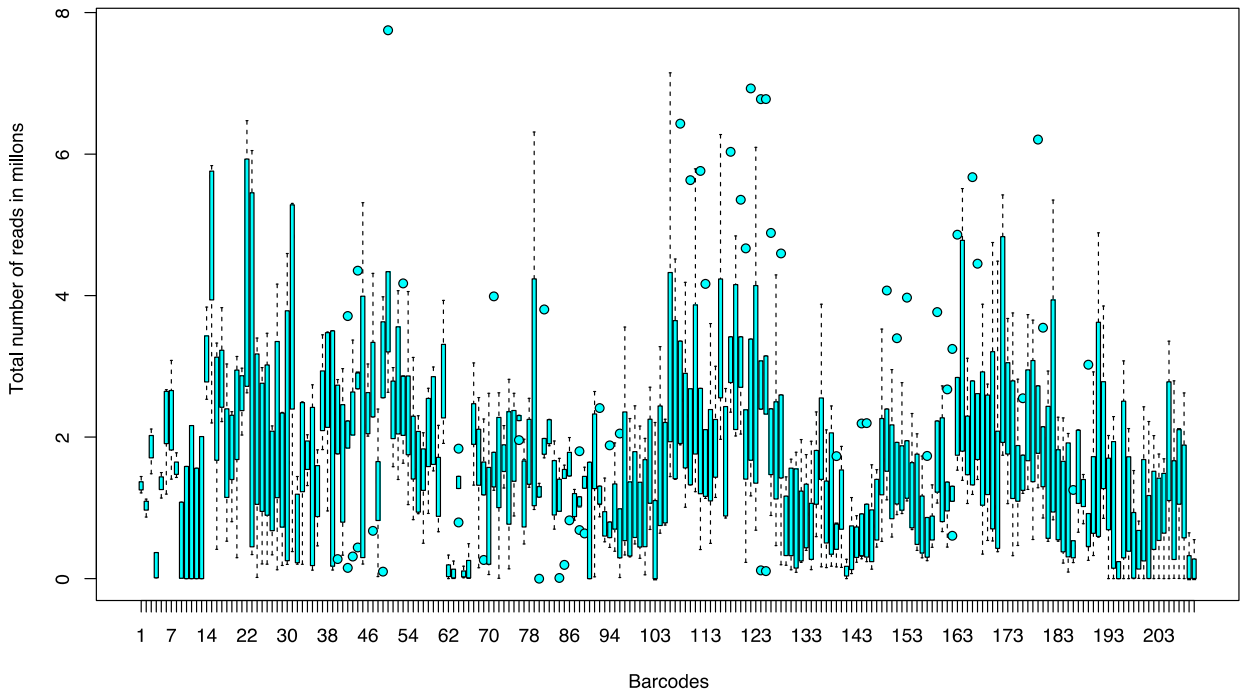


Figure 28. Barcode bias for multiplex sequencing of 1100 samples
Read number variation for each of the 210 barcodes reused in the nine lanes.

After the alignment of the paired end reads against the reference genome, we got for each sample on average 1,846,116 reads aligned. For genotyping we only considered uniquely aligned reads, reducing the number of read per sample to 1,056,582 reads on average.

In theory RAD-seq should give a high number of reads aligned to the restriction site, resulting an accurate genotyping for SNPs in these regions. In our case we could not observe an enrichment of reads. The majority of the markers had a read coverage of zero or one (Figure 29). The low coverage rate per marker makes distinguishing between homozygous and heterozygous regions challenging.

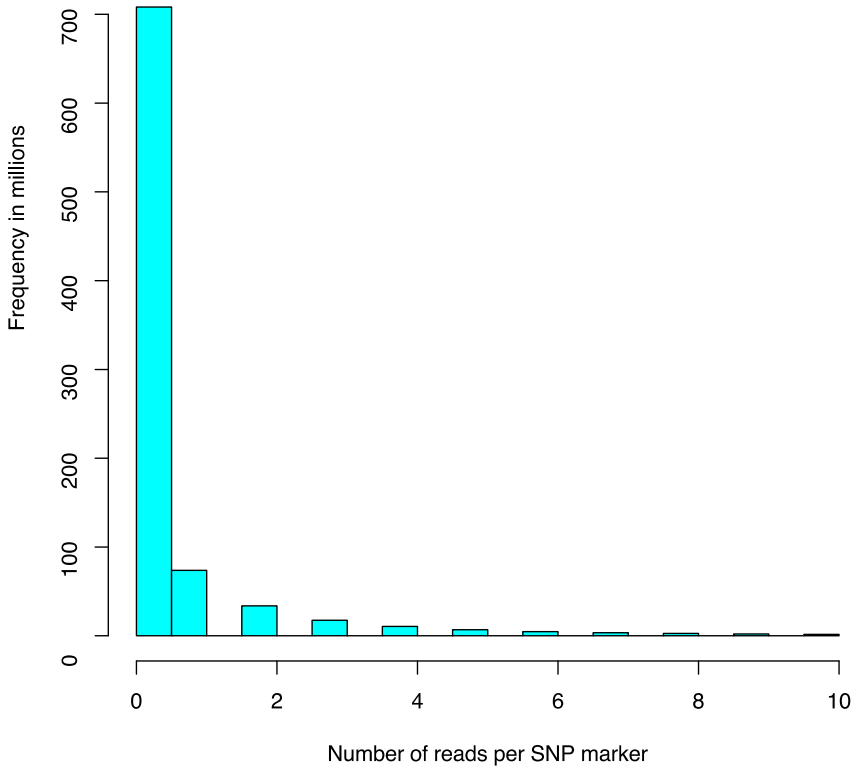


Figure 29. Number of reads aligned to SNP markers

Most of the SNPs did not have any read aligned. X-axis: number of reads per SNP marker and y-axis is the frequency in millions.2.2.3 Validation per simulation

2.3.3 Validation with simulated data

To evaluate the prediction efficiency and accuracy of our approach, we applied the genotyping pipeline to simulated data. We simulated 5,000 samples with relationships similar to one sub-population, and simulated sequencing with three different coverage rates (0.1x, 1x and 10x), an error rate from 1-3% and 1,239,342 SNPs for genotyping. A ten-fold cross validation was applied with ten separate genotyping runs. Each run contained 500 randomly selected samples. We counted the number of COs from the predicted and expected results. The difference indicates if the approach is over- or underestimating the total number of CO. By comparing the results, for each of the three simulated coverage rates, we concluded that our model slightly underestimates the number of CO (Figure 30).

We use a receiver operating characteristic (ROC) -like curve representation for estimating the distance between expected and predicted CO positions, applied for physical distances and marker counts (Figure 30). As expected increasing the coverage reduces the regions of wrong predicted genotypes in physical and marker distance. In our case the best result was achieved with a coverage rate of 1x. This was to be expected as our HMMs were initially trained on a coverage rate of 1x, based on the low sequencing results of the AMPRIL population. Having a different coverage rate i.e. 10x or 0.1x produced outliers. Not considering outliers the average resolution was 1,582 bp for crossing over distances between expected and predicted position at a coverage

rate of 0.1x, which represent in our set 12 wrongly genotyped markers per CO. On average 219 bp (2 markers) were mis-genotyped given a coverage rate of 1x at the CO site.

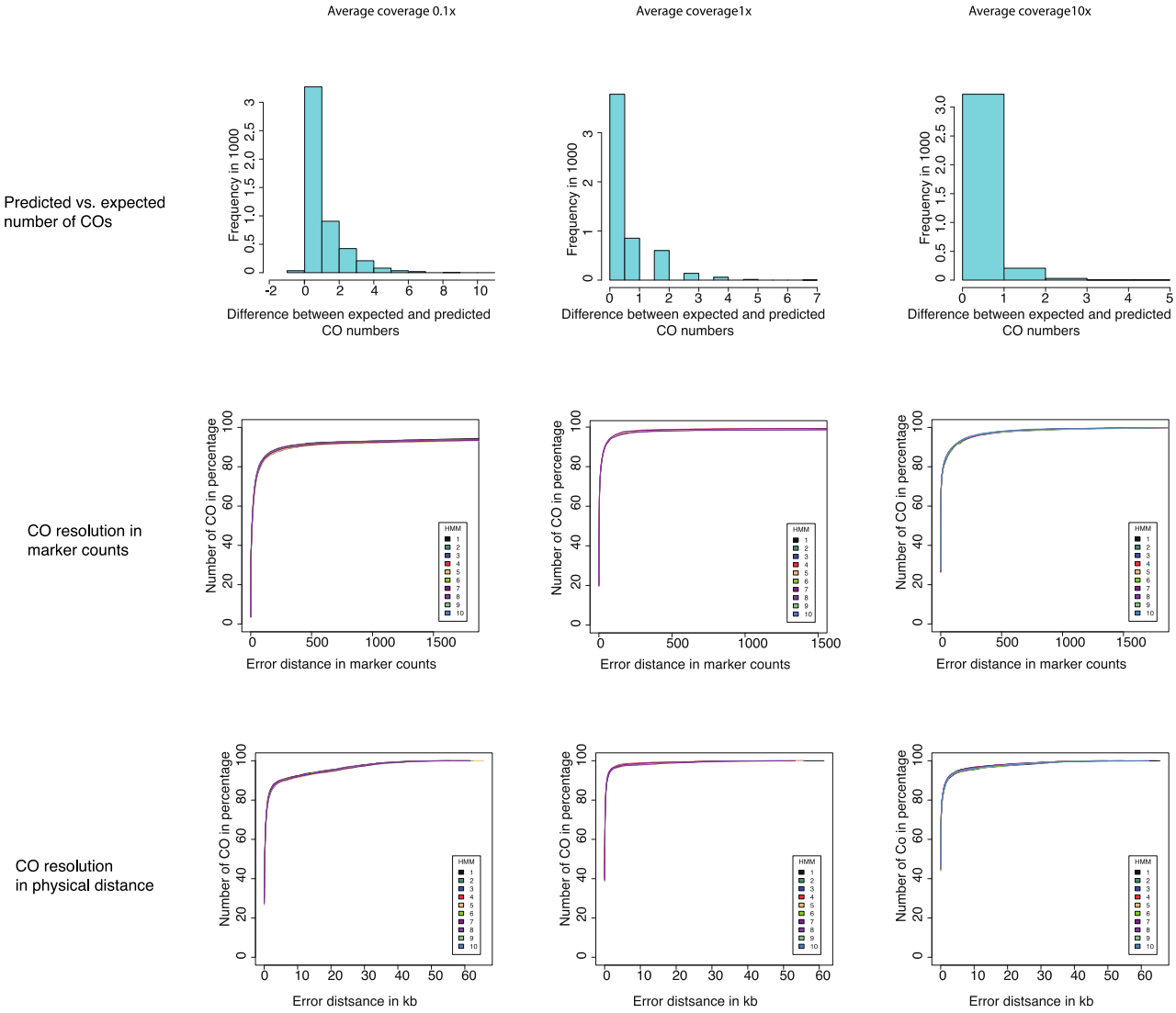


Figure 30. Validation based on simulations of 5000 individuals

The columns show the effects for three different coverage rates (0.1x, 1x and 10x). The first row estimates the differences between predicted and expected COs (not taking the position into account). The second row estimates the number of marker as a distance between expected and predicted CO positions and the last row shows the same information but as physical distance of predicted and expected CO.

2.3.4 Error position and type

Using simulation data for validation allowed us to classify problematic regions. The errors produced were similar to the pattern already been described for genotyping bi-parental mapping populations using TIGER. We encountered errors at the beginning and the end of the chromosome arms and near the centromeric regions (Figure 31). We concluded previously that these patterns came with the usage of a HMM as it needs information for the beginning and end for calling the correct genotype and the drop of markers increases the rate of selecting the wrong genotype. The error rate itself is quite low, below 0.02% for 1x coverage rate. The dominant error type were

predicting the wrong homozygous genotype (on average 62 %) or predicting homozygous blocks in heterozygous regions (28%). Strong haplotype sharing between the parents could explain some of the errors, when predicting the wrong homozygous genotype. The lack of sequencing information could additionally lead to not encountering all heterozygous genotypes.

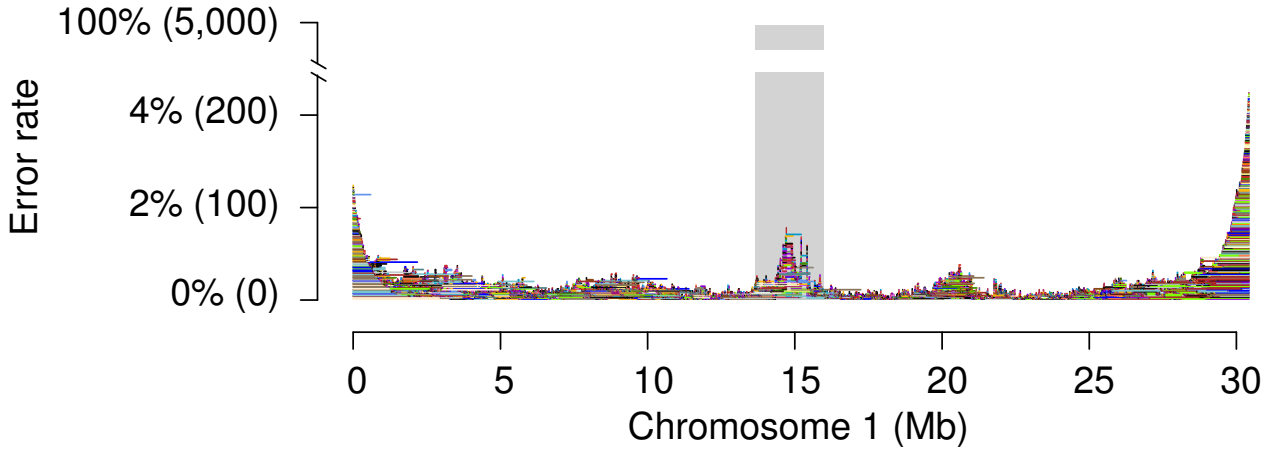


Figure 31. Examples for types and errors as occurred on chromosome 1. The typical patterns are errors at the beginning and end of the chromosome and near to the centromeric regions. Coloring bars are regions which have been wrongly genotyped for 5000 individuals

2.3.5 Using technical replicates for testing of reproducibility

Applying our pipeline to technical replicates allows us to analyze the stability of the prediction of the CO position on real data.

By comparing six samples, for which we had technical replicates, we found on average 0.18% of the markers with different genotypes assigned (Figure 32A). By comparing 89 CO sites we observed a median shift of the predicted COs of 960 bp (Figure 32B) and the majority of disagreeing regions were between homozygous and heterozygous predictions (83.33%). The median of shift of predicted COs was used instead of the mean, as it is more robust against outliers as we have two samples having a higher error rate of over 1%. The high error rate was related to a higher difference in read coverage between the individual samples.

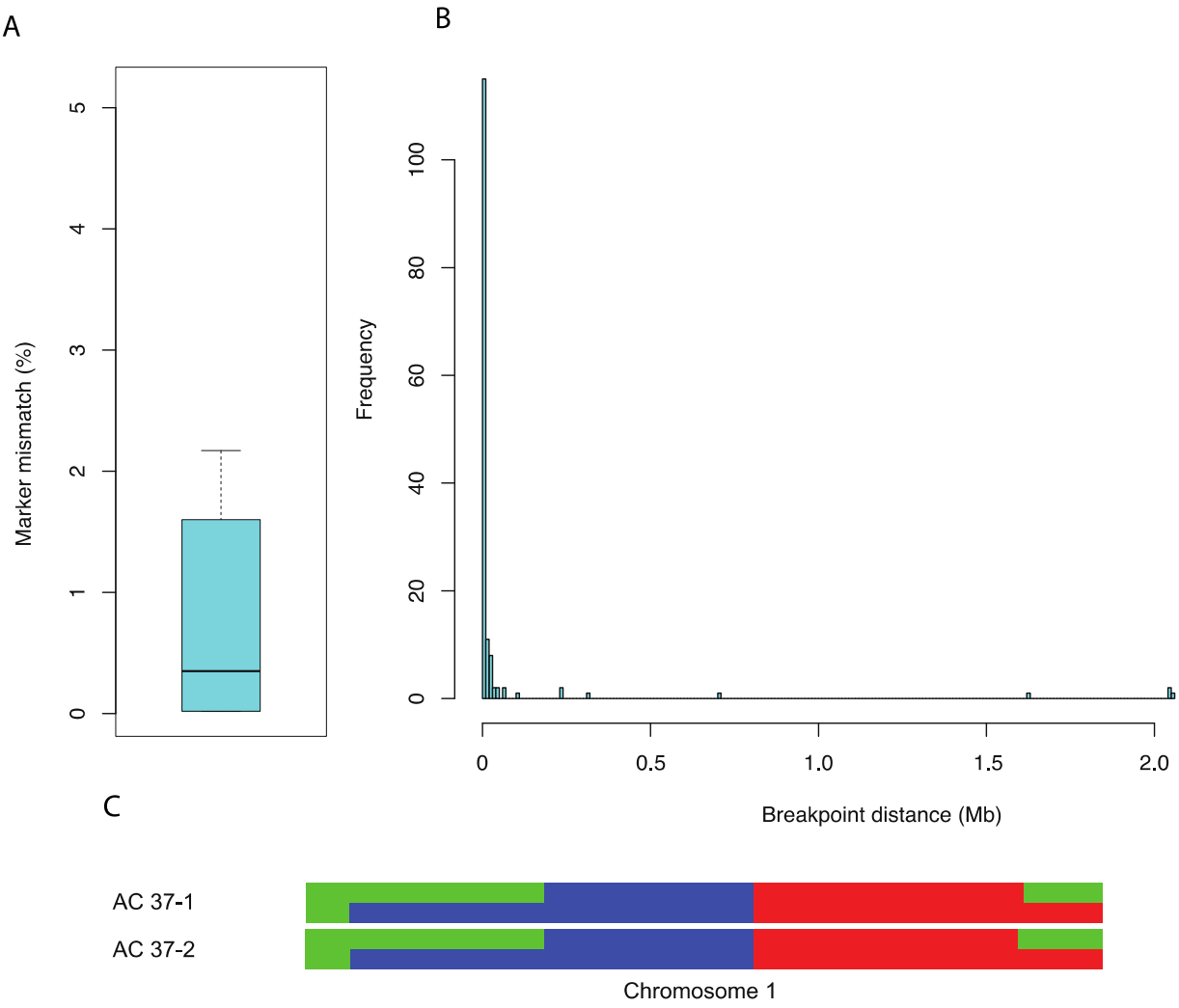


Figure 32. Technical replicate analysis

A) The HMM labeled on average 0.18% of on average 131,318,564 SNP markers differently within the 12 samples of six replicates. A total of predicted 89 CO sites showed an average shift of 960bp. C) Graphical representation of one chromosome of a replicate pair, where the colors represent (green, blue and red) different parental genotypes.

2.3.6 Genotype validation using 300 previously genotyped SNP markers

We used genotyping data from a 300 SNP markers assay applied to the same population which was previously released (Xueqing Huang et al., 2011) to estimate the error-rate of our genotyping data. During the comparison we observed different levels of differences between the two genotype sets within subpopulation I and II. For population II we estimated a homogenous error rate lower than 2% for all subpopulations, whereas for population I we achieved a mixture of different error rates (Figure 33).

We compared the error rates with the percentage of heterozygosity of the predicted genotypes for each sub-population and observed that the error rate correlates with an increased content of predicted heterozygosity (Figure 34) In particular for population I we observed that the differences between the previous markers and our prediction could be clustered into three different clusters. A

possible explanation besides contamination could be different level of inbreeding as we see sub-population where the mean heterozygosity is similar to that of F_3, F_4 and F_5 populations, which is in strong contrast to the expected level of heterozygosity of F_6 and F_7.

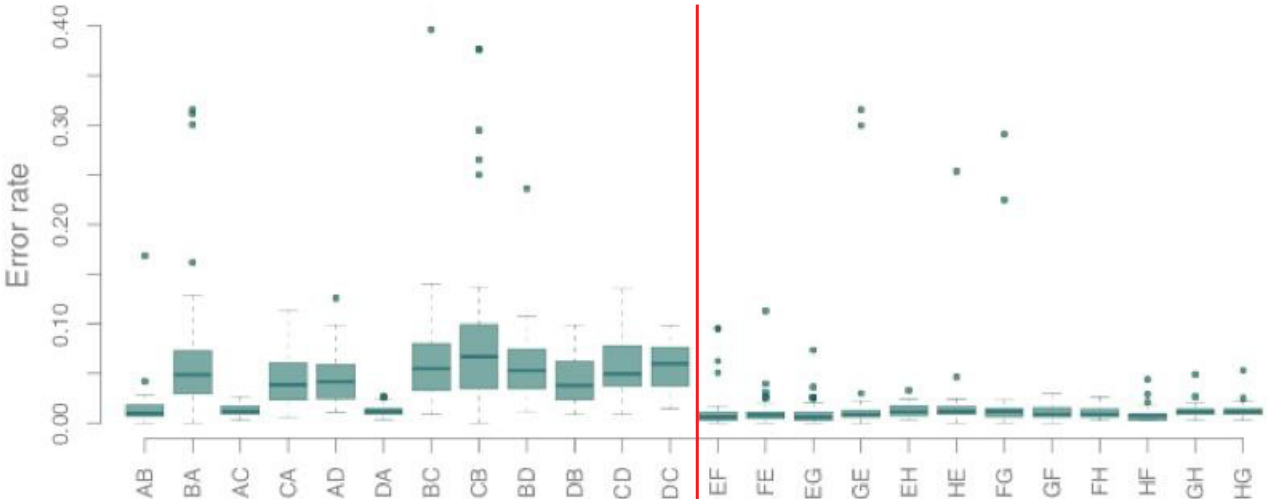


Figure 33. Percentage of disagreeing markers between of previously 300 genotyped SNPs as compared to our predicted genotypes

The comparison showed a problem with correctly genotyping population I, where as in population II we observed an average error rate of 1-2%. Figure was modified from Klasen, 2014.

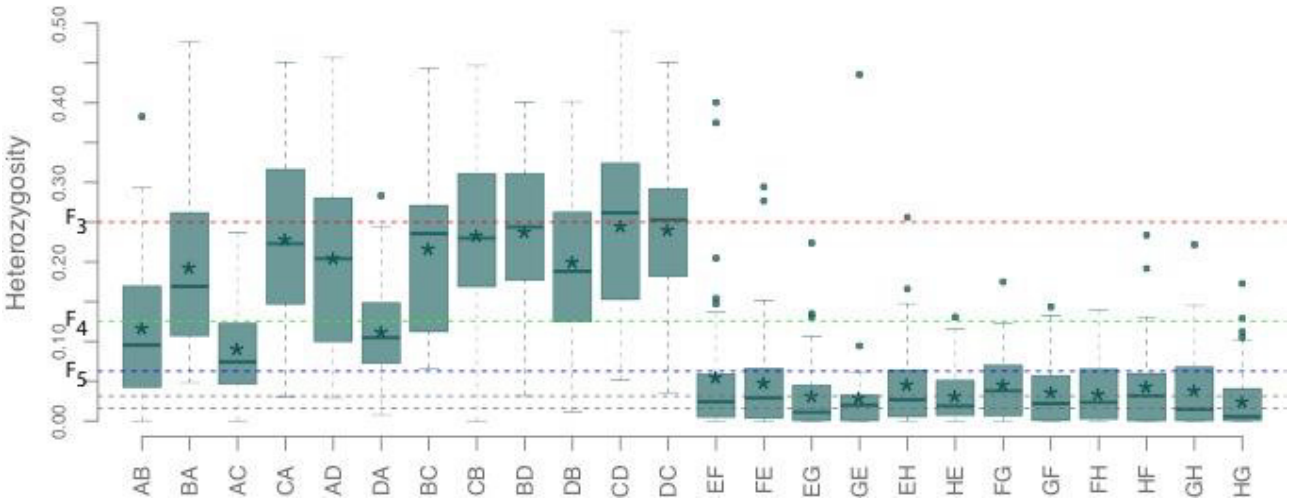


Figure 34. Comparison of heterozygosity content for each sub-population of the AMPRIL population

An analysis of the heterozygosity reveals that for population I we have an unexpected high levels of heterozygosity. Figure was modified from Klasen, 2014.

2.3.7 Outcross events

For some samples we observed an extremely high content of heterozygosity of over 70% of the genome. The appearance of high heterozygosity appears distributed along all sub-populations. By analyzing enriched heterozygosity samples we observed in some cases combinations of genotypes that were not in agreement with the crossing scheme or our model predicted unusual number of COs in close proximity range i.e. 1000 COs, as the model can not decide which

genotype combination is the correct one. Further, we resolved for some samples the appearance of more than four parents. Both events could be explained by outcrossing events, either inside the same or between different sub-populations (Figure 35). Overall we identified 36 (~4%) samples with an obvious outcross footprint within the same sub-populations as estimated by heterozygosity level (>50%) and 29 (~3%) samples with footprints from outcrossing between different sub-populations (as identified by allele combinations that were not possible following the crossing scheme of any of the subpopulations).

2.3.8 CO landscapes and genotype frequencies per sub-populations

On average the total AMPRIL population contains 13 COs per sample (2.6 COs per chromosome) (Figure 36A). From the CO landscape we received the typical decrease of CO near the centromeres and higher rates in the chromosomes arms (Figure 36B). In between we have some hot and cold spots. The CO landscapes per sub-population are following in general the global CO landscape. We analyzed the genotype frequencies from our predicted genotype. The expectation of the genotype frequency should be 25% for each of the parents and each chromosome within each of the sub-populations. We observed exceptions from this rule for different sub-populations. In general, we observed a lower frequency for the Cvi-0 genotype, most strikingly observable on chromosomes three and five (Figure 37).

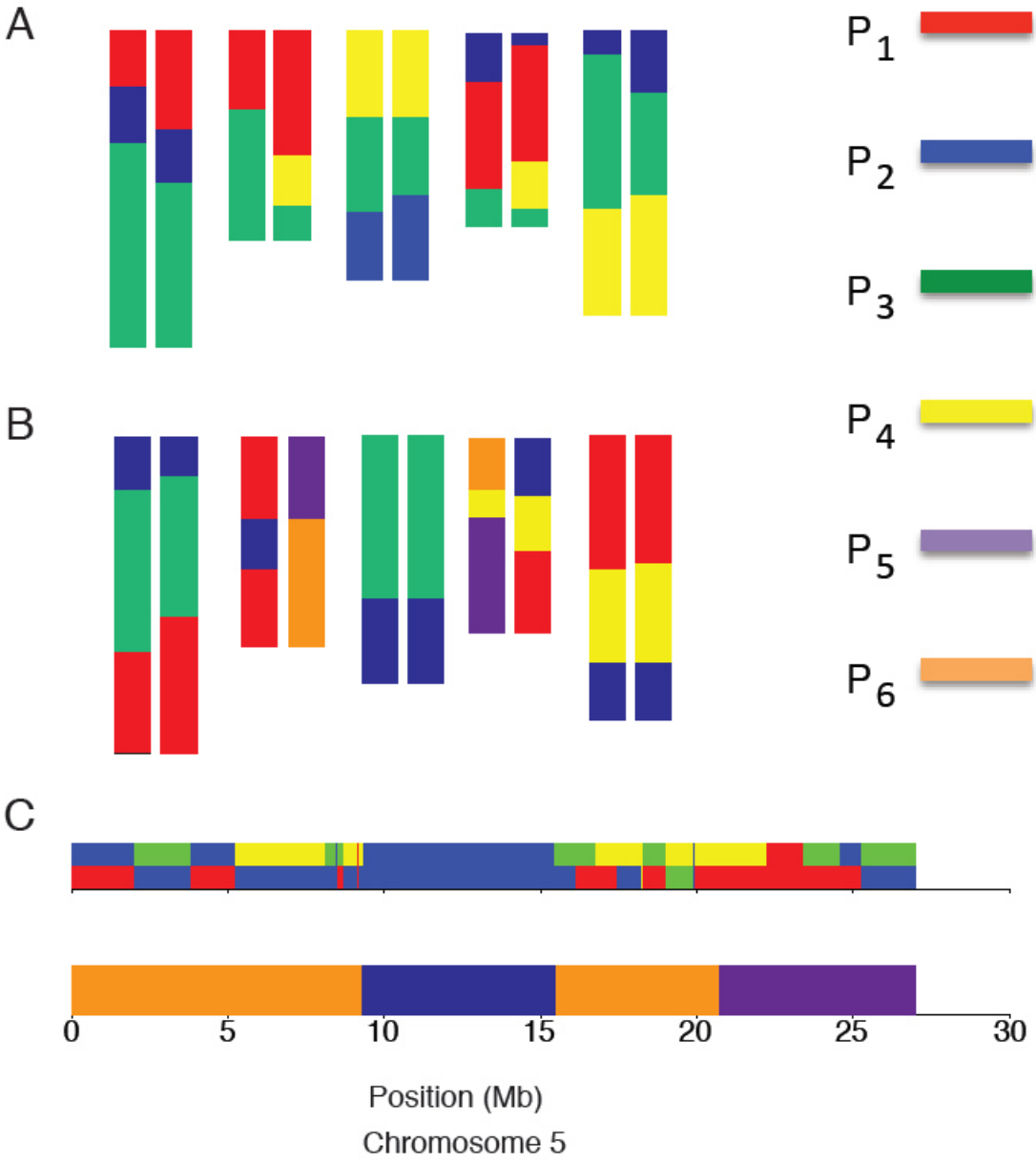


Figure 35. Examples for genotypes that reveal outcross events during their generation
We observed two different footprints evidencing outcross events during the generation of the AMPRIL population. A) Outcross events in between sub-population increasing the heterozygosity content or B) outcross events between different sub-populations introducing additionally unexpected genotypes. For an outcross between different sub-populations we see an increased number of COs and combination of unexpected genotypes (red and blue in heterozygosity regions). By including the additionally parental genotypes into the reconstruction of the parental haplotypes the number of CO decrease and we observe more reasonable genotype haplotype blocks. (P = parental lines).

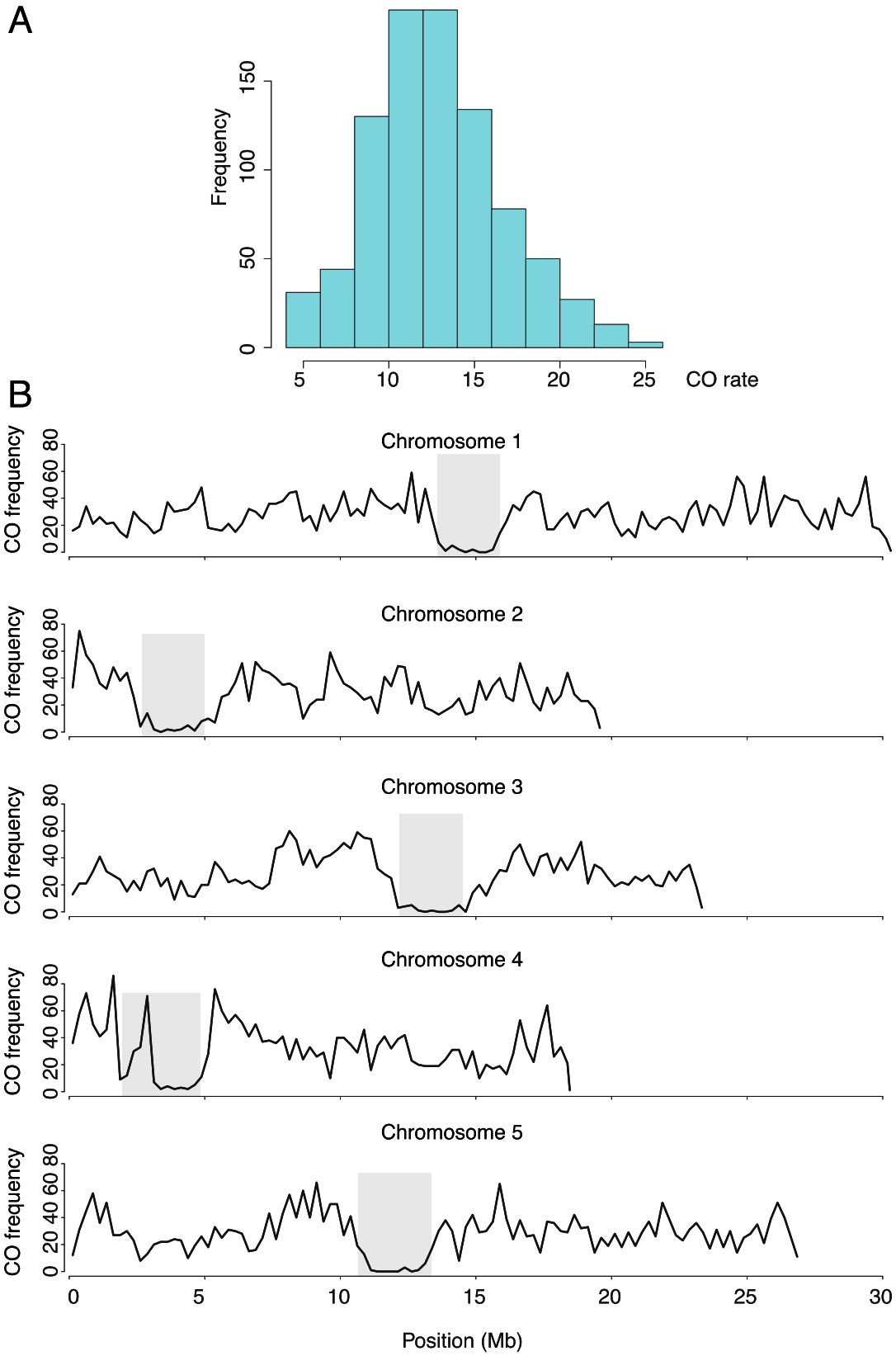


Figure 36. CO frequency distribution of the AMPRIL populations

A) Frequency distribution of COs per sample. B) CO landscape of the AMPRIL population. The grey box shows the location of the centromers.

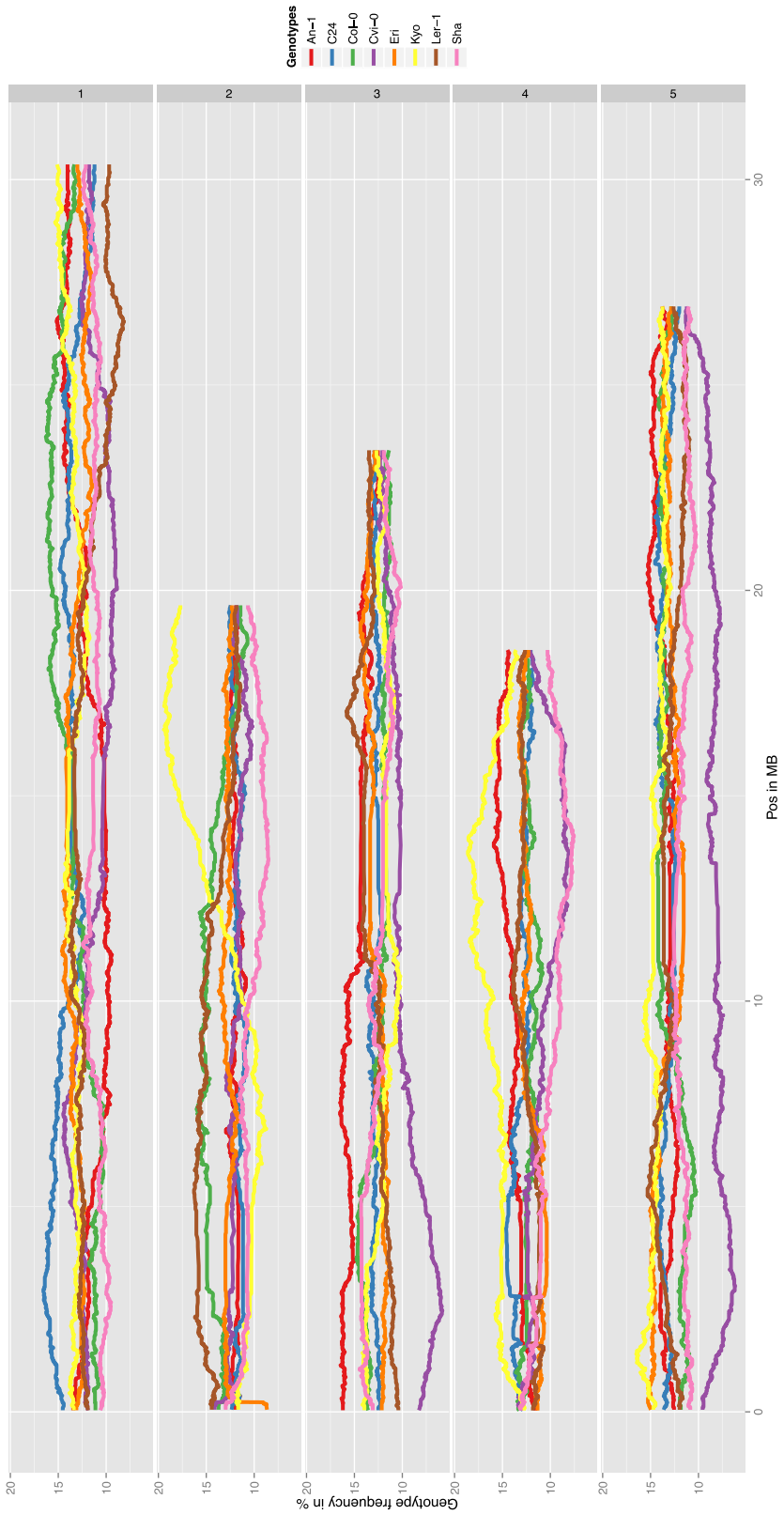


Figure 37. Genotype frequencies for each sub-population for each chromosome

The genotype frequencies were normalized based on the number of expected parental frequency per subpopulation. Frequencies that deviate from the expected frequency could happen by change (drift) or by selection of certain loci.

2.3.9 Genetic incompatibilities

Different genotype frequencies differed from their expectation, which could be explained by drift, selection or genetic incompatibilities. To find incompatibilities, we applied an X^2 test on the observed genotypes. To reduce the computation time we reduced the data set from around 2 million markers to 13,254 positions. Those positions represent regions, where a CO was observed (Figure 38).

In general genotypes at two different sites at the same chromosome are linked. The genotype dependency between two markers decreased towards the chromosome arms due the higher rate of recombination events. However, analyzing the combination of markers on different chromosomes, which are entirely unlinked and thus can freely segregate, revealed eight combinations of pair-wise markers where the genotypes were significantly dependent on each other (p-value <=0.0001 after correction for multiple testing). Among these eight combinations was one combination at the end of chromosome one and at the beginning of chromosome five, which was already described as genetic incompatibility by Bikard et al., 2009. They described the incompatibility between the genotypes Col-0 and Cvi at this particular locus and other incompatibilities in their populations. Chromosome one and five share the same gene due to gene duplication, whereas in background of Col-0 this gene is not functional at chromosome one and Cvi exists a deletion of this region at chromosome five. The combination of both non-functional alleles is lethal. Besides the cluster described by Bikarad et al., the data showed other combinations, which require further validation.

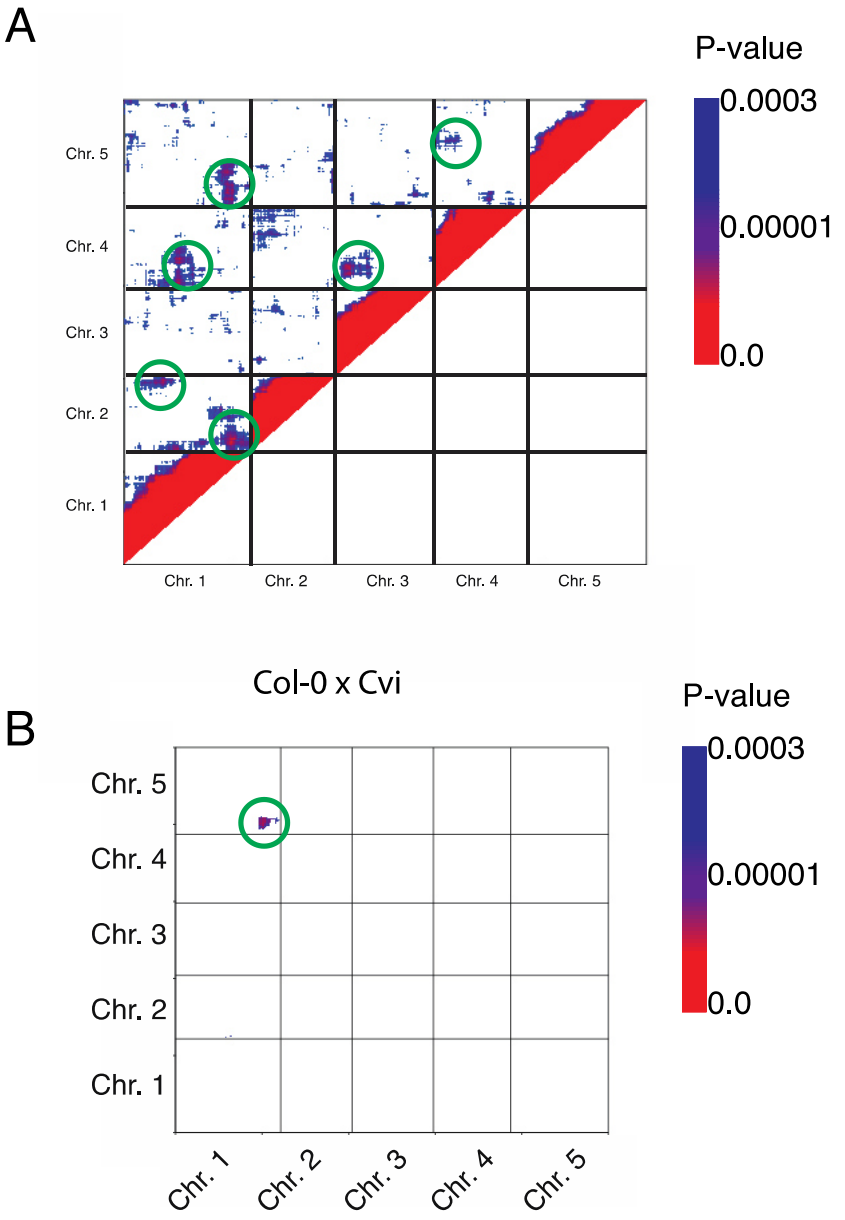


Figure 38. Genotype incompatibilities

After obtaining the genotype information for each sample a test of independency (X^2-test) between the all pairs of markers were performed. Each colored dot describes a significant marker combination after correction for multiple testing (p<=0.05), red and blue peaks are most significant p <0.00001 and p = {0.00001<x<=0.0003}, respectively. As expected a strong dependency of markers on the same chromosome can be observed, which decreases with distance as expected. We observed several small significant clusters (green circle) between loci from independent chromosomes. B) By analyzing the independency between the parental lines using our genotype data we obtained a cluster (chromosome 1 and 5) already been described to be a genetic incompatibility between Col-0 and Cvi.

2.4 Discussion

2.4.1 Summary

Published tools for genotyping by sequencing and imputation of sparse genotyping data e.g. like BEAGLE (Browning & Browning, 2007), IMPUTE2 (Howie, Donnelly, & Marchini, 2009) or TIGER can not be used for genotyping the AMPRIL population, as they do not base their reconstruction on the genotypes of four parents. There exist another intercross population similar to the AMPRIL population named Multiparent Advanced Generation Inter-Cross (MAGIC), which also have been genotyped with a tool that is aware of the parental haplotypes (Richard et al., 2014). Nevertheless this tool could not be applied to the AMPRIL population as it expects a mixture of 19 parental genotypes for each sample. Hence a new tool was developed, where two HMMs are predicting the genotypes. The validation of the prediction was done using simulation studies and by comparing the predictions with 300 previously genotyped markers for each of the samples. The analysis from simulation studies showed a similar error profile as observed by using TIGER with a bi-parental mapping population. The regions of errors were located at the arms of the chromosome and at the centromeric regions. Therefore we speculate that this error profile is a common outcome when applying an HMM approach for genotyping.

The estimation of the error rate by using existing marker data revealed two contrasting results for population I and II. The difference between the error rates from both populations was explained by the different rates of heterozygosity. To investigate whether this really represents the actual heterozygosity in the samples we manually checked whether the predicted genotypes had a possible high content of heterozygosity. We could not observe any problems, which would have interfered with predicting the correct genotypes, as the predicted genotypes were supported by the alignment of the short read data. We could not determine any reason why the level of heterozygosity as proposed by the short read analysis would not be reflecting the real levels homozygosis, despite the fact that the crossing schemes recorded for both populations indicated that they should be the same.

The high content of observed heterozygosity could have an influence on the analysis based on the genotype data frequency i.e. the detection of genetic incompatibilities, which might be not visible through allele sharing at certain locus. High heterozygosity could be masking a possible incompatibility, as heterozygosity might rescue lethal phenotypes with non-lethal alleles introduced by higher diversity. Therefore, we used only population II for testing of the dependency of haplotypes. With this data eight incompatible genotype combinations were identified. One of those cluster has already been validated by Bikard et al., 2009

Besides the problem of population I containing different genotypes from different generations the entire AMPRIL population revealed footprints of putative outcross events between and within sub-populations. 3% of the samples had an outcross event across different sub-population and 4% within the same sub-population estimated from their high rate of heterozygosity or not allowed appearance of parental genotype combinations. Samples with an outcross footprint suggesting an outcross event between different sub-populations were identified due to the combination of unexpected parental genotypes. Hence, the conclusion of 3% can be seen as the true estimation for the AMPRIL population. Nevertheless this is not true for outcross events within the sub-population. Here sample having more than 50% heterozygosity content were labeled as an outcross event. Therefore the estimation of 4% is only true for a recent outcross events and the total amount of outcross events within the sub-populations is possibly higher. The fact that outcross events between different sub-populations have been observed by increased density of COs can be explained by the limitation of the model. The models were designed to predict the genotypes based on the background of four parental genotypes. Data not fitting this assumption will lead to wrong prediction e.g. for many observed changes between genotypes, those genotypes could share alleles with the true parental genotype.

2.4.2 Improvements

The presented method for genotyping the AMPRIL population showed high accuracy based on the comparison with simulations and with previously genotyped markers. Nevertheless the method can be further improved. The improvements could be done in a similar way as already described for TIGER. The major disadvantage of this approach is not considering insertion and deletion as an additionally source of information for accurate genotyping e.g. to account for translocation and rearrangement. The appearance of any structural variation can produce wrong genotype prediction based on the short sequencing alignments. Structural variants are identifiable as abnormally mapped reads (Wijnker et al., 2013). The allele information from rearrangements could influence the interpretation of the correct representative genotype at that locus. It is getting more complicated if in a heterozygosity region where one parental chromosome has a structural variation and the other parental not. These cases make it quite difficult to estimate the correct allele frequencies that can be interpreted into the correct genotype call. For correct assessment of structural variations a local realignment or local assembly at a region with unusually high abnormal aligned paired reads could be done. Even better would be to align longer sequencing information directly as short read data could stack if the rearrangement contains repetitive elements. To facilitate the prediction of genotypes in a background of four parents, we omitted the case of rearrangement by removing SNP markers containing InDel information.

Another improvement could be to increase the accuracy towards the CO positions between two homozygous genotypes. To estimate the COs between homozygous and heterozygous an

additional HMM could be used based on markers supporting only unique markers to resolve the correct heterozygous genotypes and the position of the CO. A similar approach could be implemented for CO between homozygous genotypes.
The usage of pedigree information could be another improvement for genotyping in general, in particular for complex combination e.g. such as multi-parental populations. Having the genotypes for the previous generation allows to identify false predicted genotypes or to pinpoint the influence of a certain outcross event.

2.4.3 Outlook

The purpose of this work was to genotype ~1,100 sample of the AMPRIL population with nearly two million markers, which is the basis for QTL and epistatic interaction mapping. Klasen, 2014 presented a new method to approach these tasks. Both results were compared to the newly developed method. The new method used a hierarchical clustered on the input SNPs based on linkage disequilibrium information. By combining a penalized regression method for population structure during parameter estimation steps the new method was able to associate phenotypes to certain SNPs cluster. They tested it for the phenotype flowering time, where the new method reported known QTLs also 10 so far unknown QTLs. (see Klasen 2014 for more information). This shows that the genotyping data are sufficient to unravel even so far unknown correlations between phenotype and genotype.

3. Error correction for long reads generated with Pacific Biosciences sequencing technology

3.1 Introduction

Current NGS technologies e.g. Ilumina MiSeq offer a maximal read length of up to 300 bp. Short reads might therefore not be long enough to access long repeats or rearrangement as they those can be several kb in length. Therefore since the introduction of NGS, consumers and scientists have been demanding longer reads as it would simplify the analysis of such features (Lee et al., 2014).

To provide longer reads based on single molecule sequencing, Pacific Biosciences commercially released the PacBio RS platform in 2011. It was the first platform using single molecule real time sequencing (SMRT) for high throughput sequencing. In the early days SMRT allowed sequencing of reads up to 8,000 bp with an average of 2-3 kb (in 2011), hence the PacBio RS became the third generation of next generation sequencing. SMRT is based on monitoring individual DNA polymerase molecules in the process of replicating a template strand. The polymerase is located in a nanophotonic structure called zero-mode waveguide (ZMW), constructed on a plate of glass (Figure 39A). Each ZMW is 100 nm wide, which allows occupancies of 0.01 - 1 μM labelled nucleotides per ZMW, and contains a Φ 29 DNA polymerase fixed to the glass bottom, labeled nucleotide (a dye-linker-pyrophosphate group) and a molecule of template DNA. Upon integration of the labeled nucleotide in the active site of the polymerase, a light pulse is emitted, followed by the cleavage of the dye-linker-pyrophosphate group (source of the emitted the light pulse), and the next nucleotide can be incorporated (Figure 39B). Each of the four nucleotides contains a unique dye-linker-pyrophosphate group. A high-multiplex confocal fluorescence detection system is used to observe the emitted light impulse during the replication of the template string. By translating the light pulse back into the nucleotide alphabet, the sequence of nucleotides of the template DNA can be obtained. The template DNA is designed as a circular molecule. It is made up from the two strands of the template DNA (forward and reverse strands) including two linker sequences, which are used for primer annealing. Primers initiate the replication. The circular template design can be used for circular consensus sequencing (CSS). The CSS method allows reporting high quality reads, by overlapping the sequences produced during several rounds of replicating the same template DNA. However, CSS is only applicable if the raw read includes multiple rounds of sequencing of the template (Figure 39C).

The drawback of Pacific Biosciences is the high error rate of the reads, on average 12-15% (English et al., 2012; Quail et al., 2012;Koren et al., 2012; Lee et al., 2014), demanding for

correction methods. The major error type of Pacific Biosciences is introducing InDels (Eid et al., 2009).

We ordered Pacific Biosciences reads as guidance for on going assembly of the genome of *A. thaliana* accession Ler based on short read data produced by Illumina sequencing technology. As the long reads are used as guidance, we required high quality reads to prevent false positive results of the final assembly.

In the following section, we describe the Pacific Biosciences data and the new approach we developed to correct Pacific Biosciences reads and we compare the outcome with an existing algorithm.

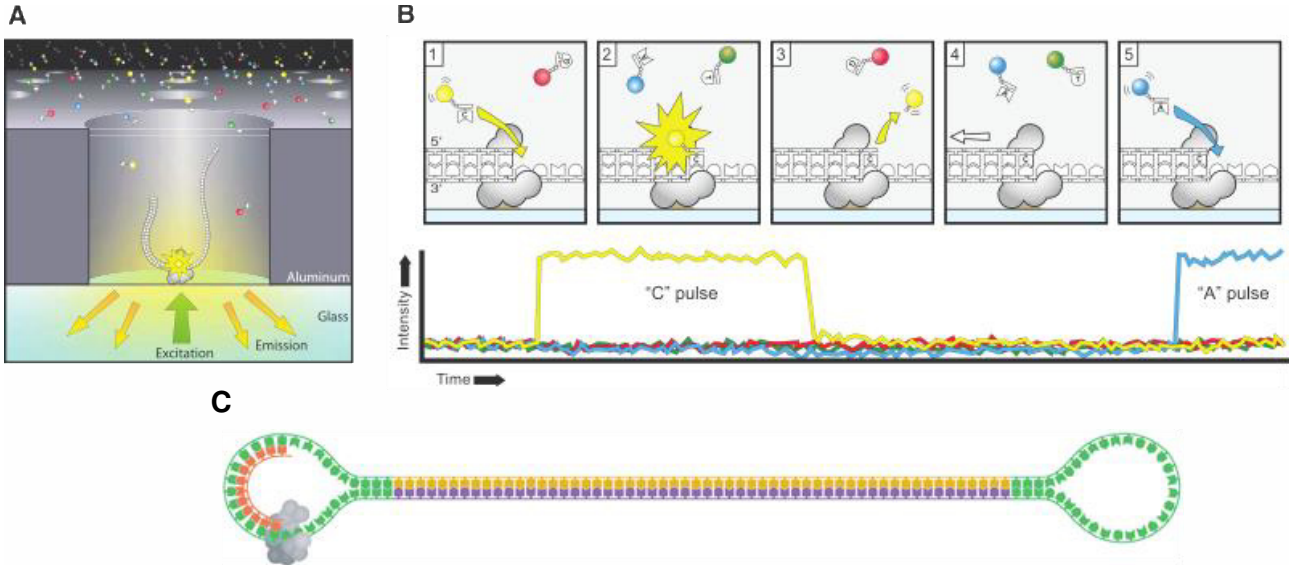


Figure 39. A zero-mode waveguide containing a Φ 29 DNA polymerase replicating a template string. A light pulse is emitted while incorporating a new nucleotide. B) The Φ 29 DNA polymerase waits for the correct nucleotide to match the template strand, incorporating a light impulse and afterwards the dye-linker-pyrophosphate group is cleaved out and the template strand moves and the cycle is repeated. The light pulse can be recorded and subsequently be used to reconstruct the template sequence. C) Shows how a template is built up that can be used for circular consensus sequencing. In green are the linker sequence and in between the template sequence and in orange a primer sequence to start the replication (Travers, Chin, Rank, Eid, & Turner, 2010)

.

3.2 Method

3.2.1 Correcting Pacific Biosciences reads by using second generation sequencing data

The second generation sequencing technologies are advanced in terms of low error-rates, typically lower than 0.1% (Ross et al., 2013). Hence, short reads could be used to correct the errors in Pacific Biosciences reads. An intuitive way to do so would be to align the short reads against the Pacific Biosciences reads and to use the resulting consensus sequence for correction.

A weakness of this approach is that the consensus strategy requires a large number of short reads aligning uniformly to the Pacific Biosciences reads, to produce an accurate consensus sequence. Further repetitive elements in the long sequence including the high error rate could increase the rate of short read stacking, which can increase the problem of correct base calling. Therefore we developed a different strategy using short read data for correction by applying a k-mer graph approach. A k-mer string can represents a possible substring of a sequence with a given length. To represent a sequence with a k-mer approach first all k-mers have to be built. The overlapping of such a set can represent the original substring. To start with the k-mer approach first a k-mer distribution has to be established, which represent the frequency of all possible k-mers from a given sequence. In our case we used high density sequenced Illumina data (>50x) and 17 kmers. The resulting 17-mers are counted and as well the frequencies of the counts are recorded. Both values are used to define a k-mer distribution (Figure 40). The k-mer distribution contains three parts: k-mers produced by sequencing errors, k-mers resulting from the unique part of the genome and k-mers built up from repetitive sequences. The parts can be identified based on the counts and their frequencies. The k-mers, which have been introduced by sequencing errors, are those with a low number of counts but high frequency. K-mers representing repetitive elements have a high number of counts but are low in their frequencies and unique k-mers are in-between both cases. The repetitive k-mers are the right tail of the k-mer distribution the sequencing errors are the left end and the unique k-mers are located near the mean value between count/frequency (by excluding the beginning of the left tail). The pipeline uses only unique k-mers which are defined by manually assessment of the k-mer distribution. These unique k-mers are used for identifying candidate Pacific Biosciences reads for correction.

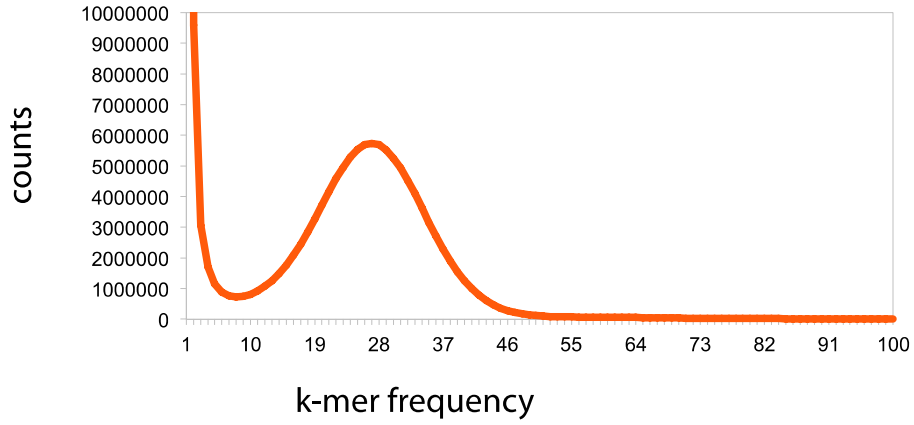


Figure 40. k-mer distribution of Illumina short read data

K-mer distribution of 31 k-mers of *Arabidopsis* deep sequenced data (50x). X-axis counts the number of times the k-mer was found in the short read data and y-axis indicates how often such k-mer counts have been observed (frequency).

3.2.2 Workflow

The algorithm presented here for correcting Pacific Biosciences reads can be divided into five parts:

1) Identifying and removing the linker sequence from the Pacific Biosciences reads.

The linker sequences are usually automatically removed from the output sequence but errors during the sequencing step prevent their recognition. To recover all the remaining linker sequences we used first BLASTN (Altschul, Gish, Miller, Myers, & Lipman, 1990) searches. Each region, which matched at least 90% of the linker sequence was removed and resulted in a split of the read into smaller sub-reads.

2) Using sub-reads for self-error-correction.

If sub-reads were generated, those can be used to correct each other as each sub-string sequence is coming from the same molecule but from a different sequencing reaction. Having the sequence of each sub-string in the same strand, a consensus sequence can be constructed. We observed for some cases that the sub-reads were not complementary to each other, because the sub-reads came from different regions in the genome and were accidently fused during library preparation building chimeric reads. Therefore a test was applied to find clusters of reads aligning to each other. For that purpose we used a combination of two multi-alignment tools CLUSTALW (Larkin et al., 2007) and MAFFT (Katoh & Frith, 2012). CLUSTALW was used to cluster the sub-reads based on their pairwise alignment score and MAFFT was applied to generate the consensus sequence for each cluster (Figure 41A,B).

3) Using short read data for building the k-mer distribution.

A k-mer distribution of the short read data was done by using the tool Jellyfish (Marçais & Kingsford, 2011). For the correction we use only unique k-mers. Unique k-mers are mostly located between the first and second valley of the kmer distribution (having sufficient genome coverage, e.g. ≈20x) (Figure 40). According to this rule we removed k-mers having sequencing errors (left site of the valley) as well as highly repetitive k-mers (right side of the second valley) (Figure 41C). This description is only possible if such a distribution is clearly visible. If such an ideal distribution is not visible, especially for the right site the k-mer value has to be reduced and the distribution has to be plotted again. The threshold for the right site would be than the point after passing the second peak (unqiue k-mers) (Liu et al., 2013).

4) Identifying optimal start points for error correction.

We defined seeds as regions in the Pacific Biosciences reads where a perfect match with a unique k-mers exists. The first step is to locate all possible seeds position and then fill the gap between the seeds or to the end of the sequence using unique k-mers. To avoid wrong “seeds” leading to wrong correction we introduce a weighting score, based on the entropy of the seed. The entropy scores the information content on its likeness of random appearance e.g. less likely events contains more useful information (as it occurs rarely) compared to frequent events. Translated to our purpose low complexity seeds get a lower score than unique seeds, and as low complexity seeds tend to have a higher likelihood to produce wrong corrections, we used this property to filter them out. The entropy is calculated as:

Equation 2 Entropy

$$H = -\sum_{i \epsilon Z^n} p_i \log_b p_i \text{ , where } Z = \{A, C, G, T\} \text{ and } n = \text{ length of sequence}$$

The probability that a given nucleotide occurs is p. We set the probabilities to ¼ for each of the nucleotides as we assume no bias for a certain nucleotide. H is the resulting best entropy, which has to be compared to the H value of the observed nucleotide frequencies. To discriminate between unique and low complexity seeds, we applied threshold entropy of 0.5 based on the distribution (Figure 42).

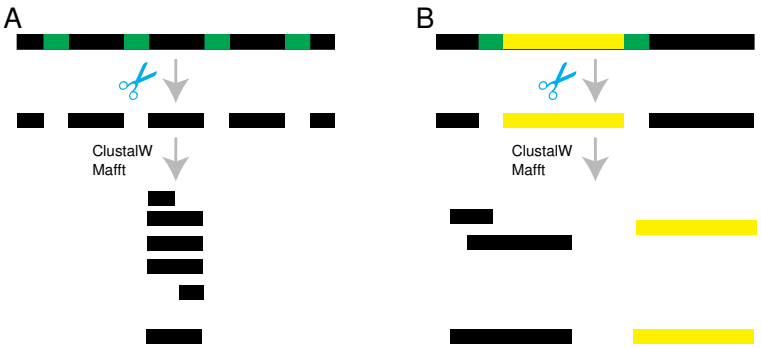
5) Building up all possible sequence representations starting with the given seed.

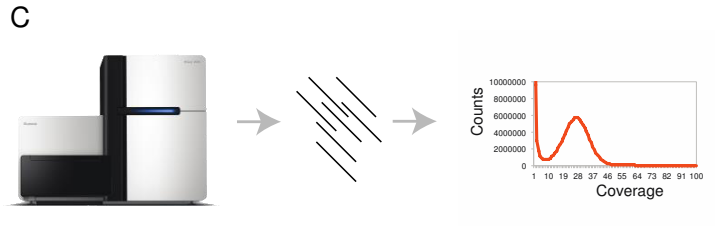
Given a list of seeds from the read sequence and the k-mer distribution, the task is to find all possible sequences which contains all observed seeds. Those representing possible candidates containing the error free representation of the read. From the candidates a representative sequence has to be chosen. How the extension from a seed is done is being explained in the following.

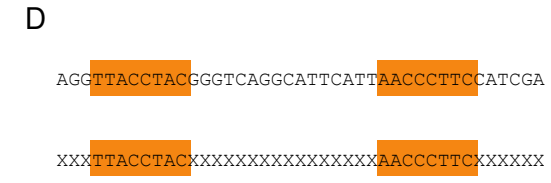
Each seed from the seed list is extended with k-mer that overlap k-1 with the actual k-mer. This is done until one of the finish criteria are fulfilled: no more k-mers are found in the k-mer distribution

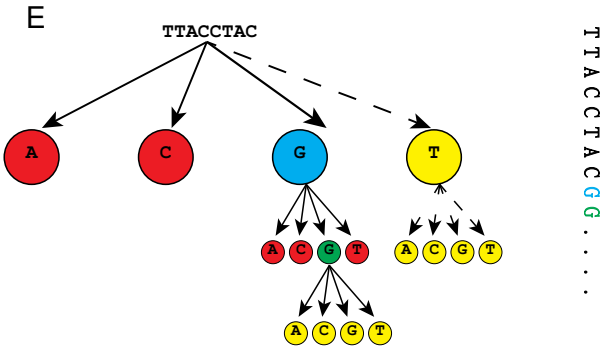
that actually do overlap with the last k-mer, until a loop has been observed where the resulting sequence contains the same k-mer more than four times, until the thereby generated sequence is too long to fit to the seed pattern given by the Pacific Bioscience read or the next seed pattern in the list has been observed. The sequence represented by this extension is a candidate representation of the input read sequence.

The algorithm for extending the k-mers is a recursive function which explores the solution space of all possible sequences. Therefore a distance threshold is necessary to stop the extension algorithms after reaching a certain distance in case we have to search the beginning- and end-sequence from only having one seed. The same holds for the case of having seeds where the partner seed sequence cannot be found. The used distance is added up with 20% of the real calculated distance to account for small insertions or deletions. That also has the effect that for the majority of the reads a longer error corrected candidate sequence is produced.
To ensure that all possible candidate sequence are found the extension is done as well in reverse complement mode as it could happened that one of the starting seed already fulfil the ending criteria to early compared by staring from its neighbouring seed.
After collecting all possible candidate solutions the correct solution is resolved using the Pacific Biosciences reads as a template sequence for aligning all obtained solutions. The algorithm reports the sequences with the highest similarity score as a correction of the Pacific Biosciences reads (Figure 41D,E and F).


A
ClustalW
Mafft
B
ClustalW
Mafft


C
Counts
10000000
8000000
6000000
4000000
2000000
0
1 10 19 28 37 46 55 64 73 82 91 100
Coverage


D
AGGTTACCTACGGGTCAGGCATTCATTAACCCTTCCATCGA
XXXTTACCTACXXXXXXXXXXXXXXXXAACCCTTCXXXXXX


E
TTACCTAC
A
C
G
T
A C G T
A C G T
A C G T
T T A C C T A C G G . . . .

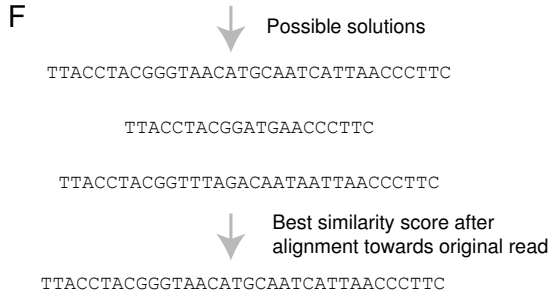

F
Possible solutions
TTACCTACGGGTAACATGCAATCATTAACCCTTC
TTACCTACGGATGAACCCTTC
TTACCTACGGTTTAGACAATAATTAACCCTTC
Best similarity score after
alignment towards original read
TTACCTACGGGTAACATGCAATCATTAACCCTTC

Figure 41. Workflow for correcting the Pacific Biosciences reads.
A) BLAST search was applied to identify and remove linker sequences (green) in the data (black). Afterwards the resulting sub-reads were clustered and a consensus of the read clusters was build. B) Yellow indicates a sequence coming from a different region or has been produced artificially. Sequences containing a linker sequence at its border can be identified and kept apart during the cluster based approach.
Both clusters are corrected independently. C) Short read data from the same sample are used to generate a k-mer-distribution, where only unique k-mers are used for correction (between both black lines). D) Using the unique k-mers seeds (orange) identified in the Pacific Biosciences reads, the seeds sequences are then used as starting points to assemble the missing information between the seeds using a depth-first assembly approach. The seed sequence are extended nucleotide-by-nucleotide and the new resulting k-mer was tested against the unique k-mers (red: no match, blue: node accepted path, green: new extension and yellow: to explore). F) After the generation of all possible solutions, all possible sequences bridging the gap between two seeds were aligned against the original Pacific Biosciences reads and the most similar solution was selected as correct sequence.

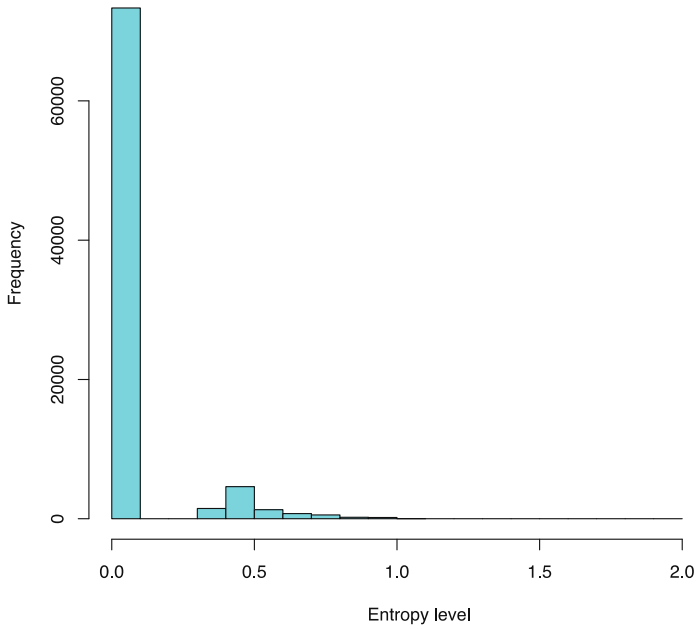


Figure 42. Seed entropy level
Entropy level of all candidate seeds is plotted against their frequency showing a bimodal distribution: a distribution for seeds having no entropy level and one for those containing entropy information. The threshold was set at the center point of the second distribution based (entropy level 0.5).

3.3 Results:

In this part we present the analysis and the correction of the Pacific Biosciences reads generated by the Pacific Bioscience RS system for a new assembly of the L*er* genome.

3.3.1 Pacific Biosciences reads statistics

We obtained around 1,400,000 Pacific Biosciences reads. The read length distribution showed that the data contained long reads with more than 9,000 bp but those were rare as the median length is around 454 bp (Figure 43A).

From the total amount of reads 8,449 (0.6%) were CCS labeled reads. The benefits of CCS reads are that the molecule is sequenced several times and each resulting sub-read can be used to create a consensus sequence representing an error free long sequence. The drawback is that the resulting sequences are shorter than the long sequence. Our data consisted of a median read length of 759 bp and the mean was 795.35 bp. The longest sequence was about 1,647 bp and the shortest 94 bp (Figure 43B).

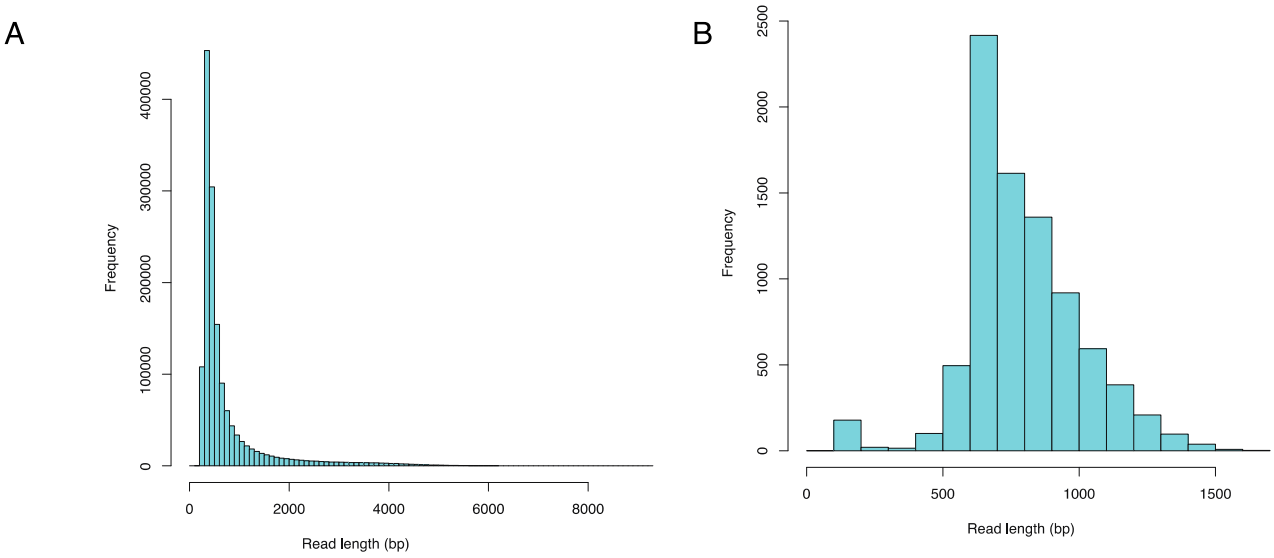


Figure 43. Pacific Biosciences read length distribution

A) Read length distribution of all 1,400,000 reads, indicating the presence of long reads up to 8 kb. However the median is around 756.41 bp. B) Read length distribution of CSS reads showing that the maximum read length was 1,6 kb and the mean was in a similar range (795.35 bp) compared to the total reads.

We have run a BLAST search against the TAIR10 database (Huala, 2001), to estimate the number of informative reads actually coming from our sample. We retrieved only 77,642 informative reads by applying the default BLASTN parameters, which represent only 5.5% of our total data. Besides using BLASTN for filtering we applied an additional strategy to estimate the number of informative reads by aligning short reads obtained from the same sample to the Pacific Biosciences reads, to validate the finding of BLASTN.

We resequenced the same sample with 30x using Illumina sequencing generating 93,109,450 paired-end reads. We aligned the short reads using Genomemapper (Schneeberger et al., 2009)

with standard alignment parameters and revealed that only 44,431 Pacific Biosciences reads were successfully aligned by the short reads data, implying that only 3.17% of the long reads were useful. For both types of analysis (BLAST and short read alignment) we plotted the distribution of the frequencies of the length of the respective long reads. This showed that mostly short Pacific Bioscience reads were identified and the average read length (≈1.2 kb) for both distributions (Figure 44).

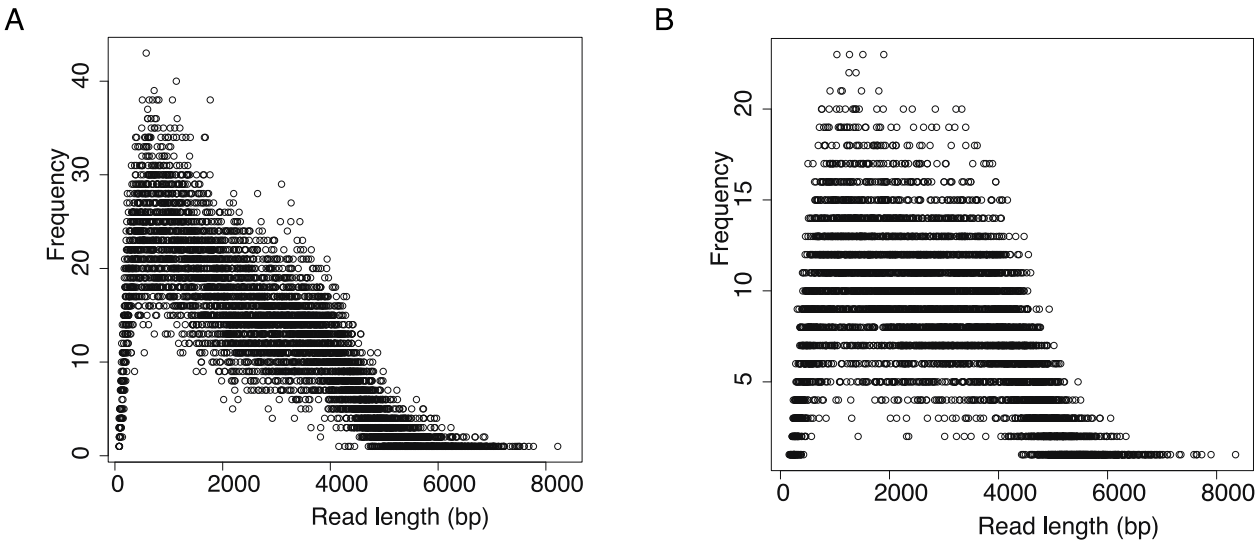


Figure 44. Read length distribution after filtering with BLAST (A) or with an alignment of short reads onto the Pacific Biosciences reads (B). Both distributions are similar in shape and of average length (≈ 1.2 kb) contrary to the total number of reads.

3.3.2 Linker removal

During the first handling of the data we recurrently detected the occurrence of linker sequences in the reads. Hence, before applying the correction pipeline we located the linker sequences, removed them and combined the resulting sub-reads, when possible, into longer reads. We identified linker sequences in 102,962 reads, corresponding to 7% of the total dataset. We calculated the number of linkers identified in such reads and plotted the resulting number of linkers per read, where at least one linker was found. The average number of linkers per read was 5.01 (mean) or three linkers (median) and the maximum was 62.

3.3.3 Correction evaluation

From the alignment and BLAST analysis we could only recover 5.5 % or 3.17 % from the total Pacific Biosciences reads for downstream analysis. By applying the new pipeline described here we could identify 170,583 (12%) candidate reads. A candidate read contains at least one high entropy seed (entropy > 0.5), which was used for the correction (see Method). After applying our pipeline we divided the resulting reads into two classes, depending on the amount of not corrected

nucleotides content. “N”s in the sequence refers to nucleotides, which could not be corrected, as they contained repetitive elements for example. We succeeded in correcting 64,403 reads partially (with less than 50% of Ns), and around 10,000 reads completely. We compared these results with the published PacBioToCa software (Koren et al., 2012), which as well corrects Pacific Biosciences, reads using short read data. The PacBioToCa tries to align the short reads and builds a consensus sequence (Koren et al., 2012) (Table 14).

Table 14 Comparison between the results of our correction approach and those of PacBioToCa

We divided the corrected reads obtained by our method into two major classes depending on their “Ns content”. PacBioToCa were able to find more informative reads from our data, but we received a higher read length for the median and maximum in both classes.

	Number of not correct Ns		PacBioToCa
	Up to 50%	> 50%	
Total number of reads	64,403	106,180	199,579
Read length(bp)			
Minimum	31	82	316
Median	1,088	1,004	745
Mean	1,406	1,452	850
Maximum	10,050	9,953	4,991

To assess the quality of the correction method the error corrected sequences were aligned against an earlier version of an assembly of the L*er* genome, assembled from deep short read data, and the coverage along the genome was estimated. For the coverage estimation, we only took into account the reads which had a high-scoring segment pair (HSP) (Altschul et al., 1990) between the assembled genome and the aligned long reads. By combining both read classes obtained by our approach, we achieved a higher coverage rate of 39 Mb compared to 36 Mb from PacBioToCa (Table 15).

Table 15 Output comparison between our correction and PacBioToCa based on the coverage rate.

	Amount of not corrected sequence (Ns)		PacBioToCa
	Up to 50	From 50%	
Coverage (bp)	36,227,350	5,962,907	36,599,573
Number of regions	44,403	15,158	49,722
Average length (bp)	815.88	393.38	736.08

3.3.4 Simulation studies

We also used simulated data to evaluate our approach. We randomly selected sub-sequence from the TAIR10 reference sequence based on the expected average size of 1.2 kb and added errors to the sequence based on average error rate of 14%. We used short read data from a resequencing of the reference line (Hartwig et al., 2012) to generate a k-mer distribution. We were able to find and correct all simulated errors except for errors located in repetitive regions.

3.3.5 Error distribution and type of errors

After correction of the Pacific Biosciences reads we were interested in studying the error distribution, we used completely corrected reads (10,000) and compared them to their original sequence. We applied a global alignment to have an end-to-end comparison. Before counting the errors, we clipped the alignments until the first 5 nucleotides in series were found, as the corrected sequences were extended We observed that the error rate in the original reads was high at the beginning of the reads (82% of the nucleotides were incorrect) then dropped to nearly 30% incorrect nucleotides at around 500 bp into the read and later increased towards 45-55% (at around 4000bp). After 4,000 bp, an interpretation was not possible anymore, as too few reads reach that length (only few single cases) (Figure 45A).

From the error distribution, we additionally analyzed the type of errors corrected by our approach. In our dataset, the major error type were InDels (Figure 45), as expected from this sequencing technology (Eid et al., 2009).

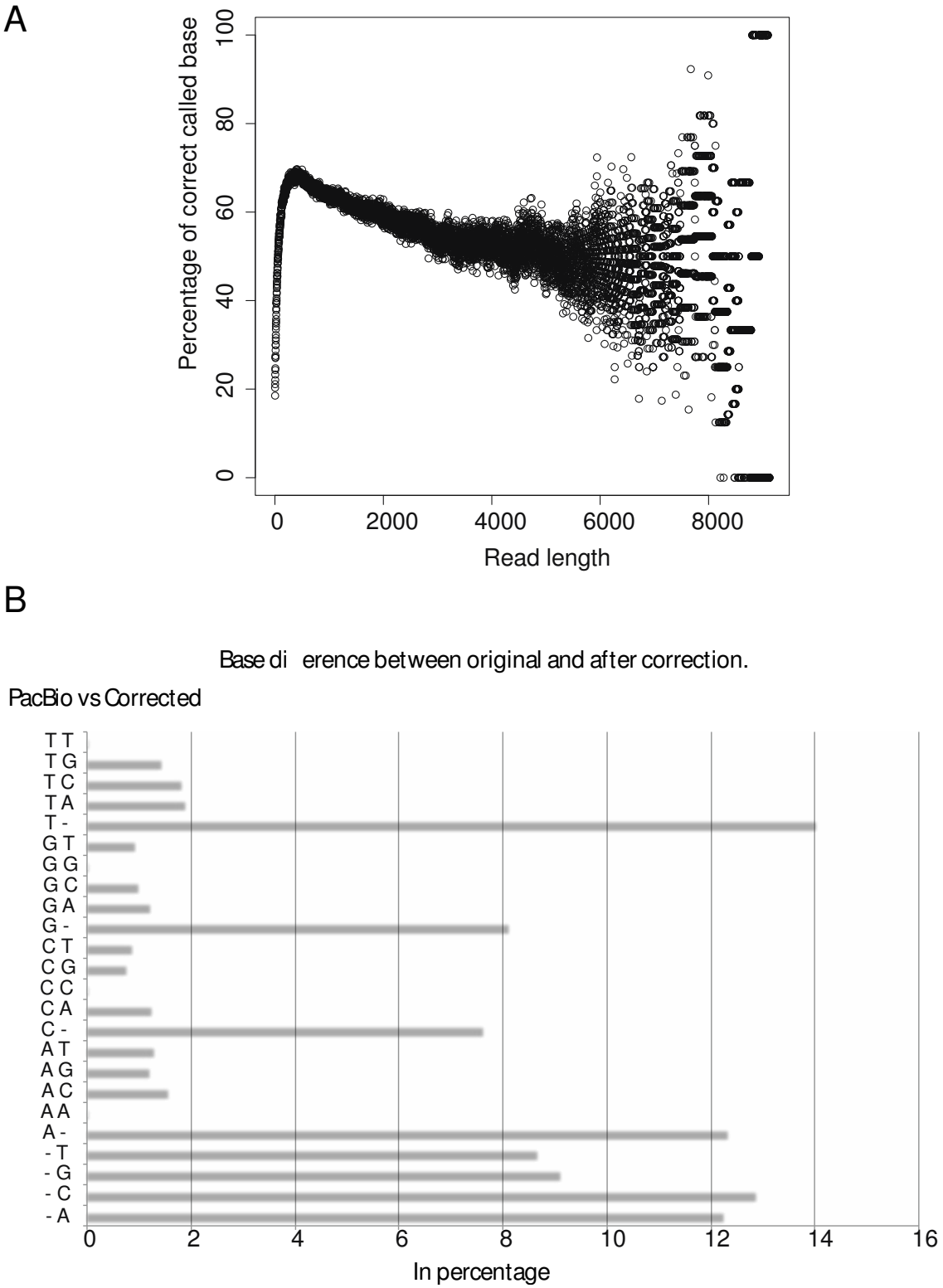


Figure 45. Error distribution and type

A) Error distribution measured by comparing 10,000 fully corrected reads against their uncorrected counterparts, showing that the error rate in the data set was overall high but dependent on the location in the read. B) Base differences between the original and after correction showed that InDels are the major errors in the Pacific Biosciences reads, resulting from the absence of signal detection or the detection of a false double impulse of the same light spectrum.

3.4 Discussion

Here we presented an algorithm using short read data for correcting long reads produced from Pacific Biosciences sequencing technology. We showed that Pacific Biosciences can indeed deliver long reads of up to 8 kb, while the average length was 1.2kb which compared to current short reads of 300 bp (Illumina HiSeq technology) is extremely long. However we observed a high error rate, which could not be resolved by simply aligning short reads. In principle this could work, but as we showed only a limited number of long reads showed an alignment with a high dense (50x) short read data. The number of errors of the long Pacific Bioscience reads were too high. By increasing the acceptance of more mismatches the number of aligned short reads could be increased but as well the number of not unique alignments would increase as well. That would lead to false corrections. To address the high error rate we used another approach based on k-mer structure.

Our approach gave similar results to the existing PacBioToCa software, nevertheless our pipeline generated on average longer reads and covered more of the target genome. The longer reads result from of the internal assembly process during the correction step, which can be compared to a simple de Bruijn graph approach (de Bruijn, 1946). A strict filtering removing short corrected reads would increase the read length distribution compared to PacBioToCa, without decreasing the coverage too much.

Another difference between our approach and PacBioToCa is that we do not correct sequences with low entropy seeds. For these our algorithm will by design consume too much computational and memory resources. To correct for repetitive regions the resulting search tree would contain repetitive loops, and each loop would produce further loops until a threshold is reached. Revisiting already seen path (loops) is a known issue during the assembly steps and is currently mostly solved by collapsing such regions into one region. We avoid this complication by excluding these sequences from the analysis.

During our analysis we also discovered chiasmic reads, which are sequences coming from different regions but have been connected by a linker structure or by accident with each other (no linker sequence in between) (Eid et al., 2009). The current PacBio RS II with a new chemistry aims to reduce this issue. Additionally the observed error profiles should be reduced using higher velocity during signal capture. Nevertheless as InDels can also occur in the polymerization process itself the need of correction will remain.

3.4.1 Improvements

The data used in this project was generated in 2011. Since then Pacific Bioscience improved their technology based on faster capture methods and more stable chemicals. The current used machine is currently the Pacific Bioscience RS II. The Pacific Bioscience RS II produces on average read length of 10 to 15 kb and a maximum read length of up to 64,500 bp (“PacBio Blog:

A Closer Look at Accuracy in PacBio Sequencing," 2014). Furthermore new algorithm did appear MHap (Berlin et al., 2014) using the fact that the quality of Pacific Bioscience improved. Instead of aligning short read data to the long reads, short Pacific Bioscience reads are used to correct ultra long Pacific Bioscience reads.

Nevertheless our algorithm could be also adapted in future. The major drawback of our approach is not considering repetitive regions due to computational and memory usage. An approach could be done for repetitive regions by having guidance for correction based e.g. a known repeat database to overcome the problem of revisiting infinite loops. This would reduce the correct search path and could allow for corrections in repetitive region.

4. Outlook

4.1 The future perspective of GBS

Genotyping by sequencing is an approach, which is increasing in popularity the last years as it becomes more and more a standard for sequencing mapping populations with and without reference sequences. The term GBS is nowadays connected with sequencing with additional genome complexity reduction like RAD-seq and not with whole genome sequencing as we presented it in chapter two. Due to the complexity reduction these methods allow genotyping of high numbers of individuals in a cost effective way. There are reports covering the potential of GBS even in complex plant genomes i.e. tetraploids and hexaploids (Endelman, 2015; Y.-F. Huang, Poland, Wight, Jackson, & Tinker, 2014; Huihui Li et al., 2015). Especially if the repetitive content is high, it does not make sense to sequence the entire genome for genotyping, therefore the RAD-seq approach is here favored by selecting an enzyme that does not cut in such regions.

It depends on the number of SNPs that are available, if other technologies for genotyping are cheaper compared to classical GBS (Burghel et al., 2015). Some of those technologies are based on target sequencing of low numbers of predefined SNPs allowing for high coverage at the SNP position for accurate genotyping (Zavodna, Grueber, & Gemmell, 2013). GBS approaches for mapping populations can be used for improving and guiding the assembly of difficult genomes (Glazer, Killingbeck, Mitros, Rokhsar, & Miller, 2015).

4.2 Would increasing the read length of short read data have an impact on GBS?

As seen in the introduction, read length of short read data is increasing during the last years. Thus the question arises if this could be an advantage for GBS analysis. The answer depends on the applied case of GBS. If GBS is applied on individuals which are highly heterozygous more coverage at the restriction site would be preferable, to accurately call the correct allele combination instead of having longer reads and low coverage. Then there is no need for imputation of missing markers at the restriction site. This is mostly important for genomes, which are not diploid but have different multi allelic combinations at this site. Longer reads are always needed to correctly identify rearrangements. Given the task that GBS has to identify such events, or to identify new loci for the discovery e.g. traits, read length matters are more favorable.

4.3 Will imputation be needed in the future?

As read number and their length increases, sequencing cost is dropping as well allowing for higher coverage, which could lead to the conclusion that in future imputing of missing markers is not necessary. But currently the opportunity is taken to sequence more individuals from a population by multiplexing to increase the resolution of QTL or GWAS analysis instead of sequencing with a higher coverage rate. Hence, imputation of missing markers will still be needed. As well having a

non-sparse sequencing representation of an individual does not mean that all possible markers have gotten the same coverage rate, due to existing sequencing bias and the complex repetitive structure of the sequenced genome. Those features are even more problematic of heterozygous cases. Not only for such cases good imputation and error correction of genotypes are necessary.

4.4 Third generation sequencing and their potential.

From the presented work in chapter 3 the impression could arise that long reads are not as useful as expected. This might be true for our data generated with the Pacific Bioscience RS I but not for the machine RS II. There are already genomes or transcriptome assemblies published based on Pacific Bioscience reads only. (English et al., 2012; Lee et al., 2014; Martin et al., 2014; Pendleton et al., 2015; Stadermann, Weisshaar, & Holtgräwe, 2015). More publications are expected to appear soon. Further, next to Pacific Bioscience another competitor appeared in Oxford Nanopore technology, using nanopores for generating even longer reads (Feng, Zhang, Ying, Wang, & Du, 2015; Karlsson, Lärkeryd, Sjödin, Forsman, & Stenberg, 2015; Laszlo et al., 2014) allowing to assemble small genomes, however, still at a high error rate 38.2% (Laver et al., 2015). Nevertheless both technologies are driving each other for better quality and longer reads, which in the end will help for detection of rearrangements, as well as genome or transcriptome assemblies in the future.

Abbreviations

AFLPs	Amplified fragment length polymorphisms
AMPRIL	Arabidopsis multi-parental RIL
An-1	Antwerp
AP1	Apetala1
bp	Basepair
CAPS	Cleaved amplified polymorphic sequences
CEN	Centroradiales
CIM	Composite interval mapping
cM	Centi-morgan
CO	Crossing over
COL	Constant-like
Col-0	Colombia-0
CSS	Circular consensus sequencing
Cvi	Cape Verde Islands
DAM	Dormancy Associated MADS-Box
DNA	Deoxyribonucleic acid
EM	Expectation-maximization
Eri	Eringsboda
FDR	False discovery rate
GBS	Genotyping by sequencing
GWAS	Genome-wide association study
H4x4	Hawaii-4
HMM	Hidden Markov Model
HSP	High-scoring segment pair
InDels	Insertions and deletions
kb	Kilobase
Kyo	Kyoto
Ler	Landsberg erecta
LOD	Logarithm (base 10) of odds
MAGIC	Multiparent Advanced Generation Inter-Cross
Mb	Megabase
MQM	Multiple QTL mapping
NGS	Next-Generation-Sequencing
nm	Nanometer
PCR	Polymerase chain reaction

QTL	Quantitative trait loci
RAPD	Random amplification of polymorphic DNA
RFLP	Restriction fragment length polymorphism
RIL	Recombinant inbreed line
ROC	Receiver operating characteristic
Sha	Shahdara
SNP	Single nucleotide polymorphisms
SSR	Simple sequence repeats
TFL1	TERMINAL FLOWER1
TIGER	Trained Individual GenomE Reconstruction
Wt	Wild-type
X^2	Chi-square test
ZMW	Zero-mode waveguide
µM	micromolar

References

Alon, S., Vigneault, F., Eminaga, S., Christodoulou, D. C., Seidman, J. G., Church, G. M., & Eisenberg, E. (2011). Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Research*, *21*(9), 1506–1511. doi:10.1101/gr.121715.111

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–10. doi:10.1016/S0022-2836(05)80360-2

Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T. T., Mast, J., Sunayama-Morita, T., & Stern, D. L. (2011). Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, *21*(4), 610–7. doi:10.1101/gr.115402.110

Arends, D., Prins, P., Jansen, R. C., & Broman, K. W. (2010). R/qtl: high-throughput multiple QTL mapping. *Bioinformatics (Oxford, England)*, *26*(23), 2990–2. doi:10.1093/bioinformatics/btq565

Bagherieh-Najjar, M. B., de Vries, O. M. H., Hille, J., & Dijkwel, P. P. (2005). Arabidopsis RecQl4A suppresses homologous recombination and modulates DNA damage responses. *The Plant Journal : For Cell and Molecular Biology*, *43*(6), 789–98. doi:10.1111/j.1365-313X.2005.02501.x

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One*, *3*(10), e3376. doi:10.1371/journal.pone.0003376

Berglund, E. C., Kiialainen, A., & Syvänen, A.-C. (2011). Next-generation sequencing technologies and applications for human genetic history and forensics. *Investigative Genetics*, *2*(1), 23. doi:10.1186/2041-2223-2-23

Berlin, K., Koren, S., Chin, C.-S., Drake, J., Landolin, J. M., & Phillippy, A. M. (2014). *Assembling Large Genomes with Single-Molecule Sequencing and Locality Sensitive Hashing*. *bioRxiv*. Cold Spring Harbor Labs Journals. doi:10.1101/008003

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, *32*(3), 314–31. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1686077&tool=pmcentrez&rendertype=abstract

Broman, K. W., Wu, H., Sen, S., & Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics (Oxford, England)*, *19*(7), 889–90. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12724300

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*(5), 1084–97. doi:10.1086/521987

Burghel, G. J., Hurst, C. D., Watson, C. M., Chambers, P. A., Dickinson, H., Roberts, P., & Knowles, M. A. (2015). Towards a Next-Generation Sequencing Diagnostic Service for Tumour Genotyping: A Comparison of Panels and Platforms. *BioMed Research International*, *2015*, 478017. doi:10.1155/2015/478017

Bystrykh, L. V. (2012). Generalized DNA barcode design based on Hamming codes. *PloS One*, *7*(5), e36852. doi:10.1371/journal.pone.0036852

Cavanagh, C., Morell, M., Mackay, I., & Powell, W. (2008). From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. *Current Opinion in Plant Biology*, *11*(2), 215–21. doi:10.1016/j.pbi.2008.01.002

Conti, L., & Bradley, D. (2007). TERMINAL FLOWER1 is a mobile signal controlling Arabidopsis architecture. *The Plant Cell*, *19*(3), 767–78. doi:10.1105/tpc.106.049767

Coyne, J. A., Aulard, S., & Berry, A. (1991). Lack of underdominance in a naturally occurring pericentric inversion in Drosophila melanogaster and its implications for chromosome evolution. *Genetics*, *129*(3), 791–802. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1204747&tool=pmcentrez&rendertype=abstract

Craig, D. W., Pearson, J. V, Szelinger, S., Sekar, A., Redman, M., Corneveaux, J. J., … Huentelman, M. J. (2008). Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods*, *5*(10), 887–93. doi:10.1038/nmeth.1251

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews. Genetics*, *12*(7), 499–510. doi:10.1038/nrg3012

de Bruijn, N. G. (1946). A combinatorial problem . Retrieved October 1, 2014, from http://www.dwc.knaw.nl/DL/publications/PU00018235.pdf

Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M. D., Lawrence, C. J., … Brendel, V. (2008). PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Research*, *36*(Database issue), D959–65. doi:10.1093/nar/gkm1041

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., … Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, *323*(5910), 133–8. doi:10.1126/science.1162986

Endelman, J. (2015). Genotyping-By-Sequencing of a Diploid Potato F2 Population. In *Plant and Animal Genome XXIII Conference*. Plant and Animal Genome. Retrieved from https://pag.confex.com/pag/xxiii/webprogram/Paper15683.html

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., … Gibbs, R. A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS One*, *7*(11), e47768. doi:10.1371/journal.pone.0047768

Feng, Y., Zhang, Y., Ying, C., Wang, D., & Du, C. (2015). Nanopore-based fourth-generation DNA sequencing technology. *Genomics, Proteomics & Bioinformatics*, *13*(1), 4–16. doi:10.1016/j.gpb.2015.01.009

Glazer, A. M., Killingbeck, E. E., Mitros, T., Rokhsar, D. S., & Miller, C. T. (2015). Genome Assembly Improvement and Mapping Convergently Evolved Skeletal Traits in Sticklebacks with Genotyping-by-Sequencing. *G3 (Bethesda, Md.)*, *5*(7), 1463–72. doi:10.1534/g3.115.017905

Hanano, S., & Goto, K. (2011). Arabidopsis TERMINAL FLOWER1 is involved in the regulation of flowering time and inflorescence development through transcriptional repression. *The Plant Cell*, *23*(9), 3172–84. doi:10.1105/tpc.111.088641

Hartung, F., Suer, S., & Puchta, H. (2007). Two closely related RecQ helicases have antagonistic roles in homologous recombination and DNA repair in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(47), 18836–41. doi:10.1073/pnas.0705998104

Hartwig, B., James, G. V., Konrad, K., Schneeberger, K., & Turck, F. (2012). Fast isogenic mapping-by-sequencing of ethyl methanesulfonate-induced mutant bulks. *Plant Physiology*, *160*(2), 591–600. doi:10.1104/pp.112.200311

Higgins, J. D., Ferdous, M., Osman, K., & Franklin, F. C. H. (2011). The RecQ helicase AtRECQ4A is required to remove inter-chromosomal telomeric connections that arise during meiotic recombination in Arabidopsis. *The Plant Journal : For Cell and Molecular Biology*, *65*(3), 492–502. doi:10.1111/j.1365-313X.2010.04438.x

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, *5*(6), e1000529. doi:10.1371/journal.pgen.1000529

Huala, E. (2001). The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Research*, *29*(1), 102–105. doi:10.1093/nar/29.1.102

Huang, B. E., George, A. W., Forrest, K. L., Kilian, A., Hayden, M. J., Morell, M. K., & Cavanagh, C. R. (2012). A multiparent advanced generation inter-cross population for genetic analysis in wheat. *Plant Biotechnology Journal*, *10*(7), 826–39. doi:10.1111/j.1467-7652.2012.00702.x

Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., … Han, B. (2009). High-throughput genotyping by whole-genome resequencing. *Genome Research*, *19*(6), 1068–1076.

Huang, X., Paulo, M.-J., Boer, M., Effgen, S., Keizer, P., Koornneef, M., & van Eeuwijk, F. A. (2011). Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(11), 4488–4493.

Huang, Y.-F., Poland, J. A., Wight, C. P., Jackson, E. W., & Tinker, N. A. (2014). Using genotyping-by-sequencing (GBS) for genomic discovery in cultivated oat. *PloS One*, *9*(7), e102448. doi:10.1371/journal.pone.0102448

James, G. V., Patel, V., Nordström, K. J., Klasen, J. R., Salomé, P. A., Weigel, D., & Schneeberger, K. (2013). User guide for mapping-by-sequencing in Arabidopsis. *Genome Biology*, *14*(6), R61. doi:10.1186/gb-2013-14-6-r61

Jansen, R. C. (1994). Controlling the type I and type II errors in mapping quantitative trait loci. *Genetics*, *138*(3), 871–81. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1206235&tool=pmcentrez&rendertype=abstract

Jansen, R. C., & Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, *136*(4), 1447–55. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1205923&tool=pmcentrez&rendertype=abstract

Ji, Y., Wu, C., Liu, P., Wang, J., & Coombes, K. R. (2005). Applications of beta-mixture models in bioinformatics. *Bioinformatics (Oxford, England)*, *21*(9), 2118–22. doi:10.1093/bioinformatics/bti318

Jung, S., Staton, M., Lee, T., Blenda, A., Svancara, R., Abbott, A., & Main, D. (2008). GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. *Nucleic Acids Research*, *36*(Database issue), D1034–40. doi:10.1093/nar/gkm803

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, *178*(3), 1709–23. doi:10.1534/genetics.107.080101

Karlsson, E., Lärkeryd, A., Sjödin, A., Forsman, M., & Stenberg, P. (2015). Scaffolding of a bacterial genome using MinION nanopore sequencing. *Scientific Reports*, *5*, 11996. doi:10.1038/srep11996

Katoh, K., & Frith, M. C. (2012). Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics (Oxford, England)*, *28*(23), 3144–6. doi:10.1093/bioinformatics/bts578

Klasen, J. R. (2014). *Development and application of statistical algorithms for the detection of additive and interacting loci underlying quantitative traits*. Cologne.

Knoll, A., & Puchta, H. (2011). The role of DNA helicases and their interaction partners in genome stability and meiotic recombination in plants. *Journal of Experimental Botany*, *62*(5), 1565–79. doi:10.1093/jxb/erq357

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., … Adam M Phillippy. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, *30*(7), 693–700. doi:10.1038/nbt.2280

Koskela, E. A., Mouhu, K., Albani, M. C., Kurokura, T., Rantanen, M., Sargent, D. J., … Hytönen, T. (2012). Mutation in TERMINAL FLOWER1 reverses the photoperiodic requirement for flowering in the wild strawberry Fragaria vesca. *Plant Physiology*, *159*(3), 1043–54. doi:10.1104/pp.112.196659

Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., … Mott, R. (2009). A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS Genetics*, *5*(7), e1000551. doi:10.1371/journal.pgen.1000551

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., … Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, *40*(Database issue), D1202–10. doi:10.1093/nar/gkr1090

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., … Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, *23*(21), 2947–8. doi:10.1093/bioinformatics/btm404

Laszlo, A. H., Derrington, I. M., Ross, B. C., Brinkerhoff, H., Adey, A., Nova, I. C., … Gundlach, J. H. (2014). Decoding long nanopore sequencing reads of natural DNA. *Nature Biotechnology*, *32*(8), 829–33. doi:10.1038/nbt.2950

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, *3*, 1–8. doi:10.1016/j.bdq.2015.02.001

Lee, H., Gurtowski, J., Yoo, S., Marcus, S., McCombie, W. R., & Schatz, M. (2014). *Error correction and assembly complexity of single molecule sequencing reads. bioRxiv*. Cold Spring Harbor Labs Journals. doi:10.1101/006395

Lex, N. (2014). Developments in Next-Generation-Sequencing. Retrieved from https://github.com/lexnederbragt/developments-in-next-generation-sequencing

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–60. doi:10.1093/bioinformatics/btp324

Li, H., Vikram, P., Singh, R. P., Kilian, A., Carling, J., Song, J., … Singh, S. (2015). A high density GBS map of bread wheat and its application for dissecting complex disease resistance traits. *BMC Genomics*, *16*(1), 216. doi:10.1186/s12864-015-1424-5

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., … Fan, W. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects, 47. Genomics. Retrieved from http://arxiv.org/abs/1308.2012

Macdonald, S. J., & Long, A. D. (2007). Joint estimates of quantitative trait locus effect and frequency using synthetic recombinant populations of Drosophila melanogaster. *Genetics*, *176*(2), 1261–81. doi:10.1534/genetics.106.069641

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics (Oxford, England)*, *27*(6), 764–70. doi:10.1093/bioinformatics/btr011

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews. Genetics*, *11*(7), 499–511. doi:10.1038/nrg2796

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, *9*, 387–402. doi:10.1146/annurev.genom.9.081307.164359

Martin, J. A., Johnson, N. V, Gross, S. M., Schnable, J., Meng, X., Wang, M., … Wang, Z. (2014). A near complete snapshot of the Zea mays seedling transcriptome revealed from ultra-deep sequencing. *Scientific Reports*, *4*, 4519. doi:10.1038/srep04519

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews.*

Genetics, *11*(1), 31–46. doi:10.1038/nrg2626

Mir, K., Neuhaus, K., Bossert, M., & Schober, S. (2013). Short barcodes for next generation sequencing. *PloS One*, *8*(12), e82933. doi:10.1371/journal.pone.0082933

Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., Warthmann, N., & Weigel, D. (2008). Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Research*, *18*(12), 2024–33. doi:10.1101/gr.080200.108

PacBio Blog: A Closer Look at Accuracy in PacBio Sequencing. (2014). Retrieved November 17, 2014, from http://blog.pacificbiosciences.com/2013/01/a-closer-look-at-accuracy-in-pacbio.html

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., … Rokhsar, D. S. (2009). The Sorghum bicolor genome and the diversification of grasses. *Nature*, *457*(7229), 551–6. doi:10.1038/nature07723

Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., … Bashir, A. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*, *12*(8), 780–786. doi:10.1038/nmeth.3454

Poland, J. a., & Rife, T. W. (2012). Genotyping-by-Sequencing for Plant Breeding and Genetics. *The Plant Genome Journal*, *5*(3), 92. doi:10.3835/plantgenome2012.05.0005

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews. Genetics*, *11*(7), 459–63. doi:10.1038/nrg2813

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., … Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*(1), 341. doi:10.1186/1471-2164-13-341

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *PROCEEDINGS OF THE IEEE*, *77*(2), 257–286.

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology*, *5*(2), 94–100. doi:10.1016/S1369-5266(02)00240-6

Ratcliffe, O. J., Bradley, D. J., & Coen, E. S. (1999). Separation of shoot and floral identity in Arabidopsis. *Development (Cambridge, England)*, *126*(6), 1109–20. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10021331

Reich, D. E., & Goldstein, D. B. (2001). Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology*, *20*(1), 4–16. doi:10.1002/1098-2272(200101)20:1<4::AID-GEPI2>3.0.CO;2-T

Richard, M., Paula, K., Clark, R., Raetsch, G., Toomajian, C., Stegle, O., & Gan, X. (2014). Imputation Analysis of MAGIC Arabidopsis thaliana recombinant inbred lines. Retrieved from http://mus.well.ox.ac.uk/19genomes/magic.html

Rowan, B. A., Patel, V., Weigel, D., & Schneeberger, K. (2015). Rapid and Inexpensive Whole-Genome Genotyping-by-Sequencing for Crossover Localization and Fine-Scale Genetic Mapping. *G3 (Bethesda, Md.)*, g3.114.016501–. doi:10.1534/g3.114.016501

Salomé, P. A., Bomblies, K., Fitz, J., Laitinen, R. A. E., Warthmann, N., Yant, L., & Weigel, D. (2012). The recombination landscape in Arabidopsis thaliana F2 populations. *Heredity*, *108*(4), 447–55. doi:10.1038/hdy.2011.95

Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O., & Weigel, D. (2009). Simultaneous alignment of short reads against multiple genomes. *Genome Biology*, *10*(9), R98. doi:10.1186/gb-2009-10-9-r98

Shiringani, A. L., Frisch, M., & Friedt, W. (2010). Genetic mapping of QTLs for sugar-related traits in a RIL population of Sorghum bicolor L. Moench. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *121*(2), 323–36. doi:10.1007/s00122-010-1312-y

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., … Folta, K. M. (2011). The genome of woodland strawberry (Fragaria vesca). *Nature Genetics*, *43*(2), 109–16. doi:10.1038/ng.740

Smit, AFA, Hubley, R & Green, P. (2010). RepeatMasker. Retrieved October 14, 2014, from http://www.repeatmasker.org

Sonah, H., Deshmukh, R. K., Singh, V. P., Gupta, D. K., Singh, N. K., & Sharma, T. R. Genomic resources in horticultural crops: status, utility and challenges. *Biotechnology Advances*, *29*(2), 199–209. doi:10.1016/j.biotechadv.2010.11.002

Stadermann, K. B., Weisshaar, B., & Holtgräwe, D. (2015). SMRT sequencing only de novo assembly of the sugar beet (Beta vulgaris) chloroplast genome. *BMC Bioinformatics*, *16*(1), 295. doi:10.1186/s12859-015-0726-6

Talbot, C. J., Nicod, A., Cherny, S. S., Fulker, D. W., Collins, A. C., & Flint, J. (1999). High-resolution mapping of quantitative trait loci in outbred mice. *Nature Genetics*, *21*(3), 305–8. doi:10.1038/6825

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, *408*(6814), 796–815. doi:10.1038/35048692

Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S., & Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, *38*(15), e159–e159. doi:10.1093/nar/gkq543

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2002). *Current Protocols in Bioinformatics*. (A. Bateman, W. R. Pearson, L. D. Stein, G. D. Stormo, & J. R. Yates, Eds.)*Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* (Vol. 11). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/0471250953

Wijnker, E., Velikkakam James, G., Ding, J., Becker, F., Klasen, J. R., Rawat, V., … Schneeberger, K. (2013). The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. *eLife*, *2*, e01426. doi:10.7554/eLife.01426

Wong, K. H., Jin, Y., & Moqtaderi, Z. (2013). Multiplex Illumina sequencing using DNA barcoding. *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]*, *Chapter 7*, Unit 7.11. doi:10.1002/0471142727.mb0711s101

Xie, W., Feng, Q., Yu, H., Huang, X., Zhao, Q., Xing, Y., … Zhang, Q. (2010). Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(23), 10578–83. doi:10.1073/pnas.1005931107

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, *25*(21), 2865–71. doi:10.1093/bioinformatics/btp394

Zavodna, M., Grueber, C. E., & Gemmell, N. J. (2013). Parallel tagged next-generation sequencing on pooled samples - a new approach for population genetics in ecology and conservation. *PloS One*, *8*(4), e61471. doi:10.1371/journal.pone.0061471

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics*, *136*(4), 1457–68. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1205924&tool=pmcentrez&rendertype=abstract

Danksagungen

In diesem Abschnitt möchte ich die Möglichkeiten ergreifen mich bei jedem zu bedanken, die einen besonderen Einfluss während meiner Promotion auf mich gehabt haben.

Als erstes würde ich mich bei Prof. Dr. George Coupland und Korbinian Schneeberger bedanken. Ich danke George Coupland das ich in seinem Abteilung aufgenommen wurde und natürlich das er Teil meiner Prüfungskomitee ist. Einen großen Dank geht natürlich an Dr. Korbinian Schneeberger, meinem Ansprechpartner und Gruppenleiter. Ich danke ihm für die vielen Gespräche, Anregungen, für die interessanten Projekt und die tolle Zeit in der Gruppe.

Weiterhin bedanke ich mich bei der Arbeitsgruppe Schneeberger für die tollen Momente, Gespräche, Anregungen, Kaffeepausen und vieles mehr. Hervorheben möchte ich besonders Geo James Velikkakam, Eva-Maria Willing, Jonas Klaasen, Vimal Rawat und Hequan Sun. Ich hoffe dass wir uns weiterhin begegnen werden.

Auch gilt mein Dank Prof. Dr. Achim Tresch, welcher Teil meines Prüfungskomitee ist, aber auch für die vielen interessanten offenen Gespräche während meiner Zeit am Max-Planck-Institute für Pflanze-Züchtung Forschung.

Ich danke Prof. Dr. Marcel Bucher dass er sich bereit erklärt hat die Rolle als Prüfungsvorsitz zu besetzten.
Weiterhin bedanke ich mich bei Armin Walter, Florian Battke, Sebastian Bender und Bernd Schaffeld für die Durchsicht und Korrekturvorschläge.

Wie immer am Ende einer Danksagung sollte auch ich meinen Eltern Manubhai & Ramabend Patel, meinen beiden Geschwistern Azadi & Divya und meiner Verwandtschaft, für deren Geduld und unendliche Unterstützung die mir immer gewährt wurde, danken.
Auch darf natürlich meine Lebensgefährten Florence Jacob nicht fehlen. Ich danke dir besonders für deinen Beistand in guten und besonders auch in den nicht so guten Phasen.

Am Ende danke ich jedem den ich in Köln und Umgebung in dieser Zeit kennen und schätzen lernen durfte! VIELEN DANK

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keener anderen Fakultät oder Universtiät zur Prüfung vorgelegt worden ist, sowie das ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde.
Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. George Coupland und Prof. Dr. Achim Tresch betreut worden.

Ich versichere, dass ich alle Angaben wahrheitsgesmäß nach besten Wissen und Gewissen gemacht habe und verpflichte mich, jedemögliche, die obgen Angaben betreffenden Veränderungen, dem Dekanat unverzüglich mitzuteilen

Datum Unterschrift