Method development and application of Next Generation Sequencing in forward genetics

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von

Geo Velikkakam James
aus Kerala, India

Köln, November 2013

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

Max-Planck-Institut für
Pflanzenzüchtungsforschung

Berichterstatter:

Prof. Dr. George Coupland

Prof. Dr. Thomas Wiehe

Prüfungsvorsitzender:

Prof. Dr. Martin Hülskamp

Tag der Disputation: Januar 2014

“Well done is better than well said”

Dedicated to my love and family

Content

List of abbreviations

Bp	Base pair(s)
kb	Kilobases
Mb	Megabases
Gb	Gigabases
CAMs	Candidate mutations
dCARE	Deep candidate resequencing
PcG	Polycomb group
FNR	Fast-neutron radiated
Chr	Chromosome
Pos	Position
QTL	Quantitative trait loci
RIL	Recombinant inbred line
NIL	Near isogenic lines
MAGIC	Multiparent advanced generation inter-cross
AMPRIL	*Arabidopsis* multiparent RIL
EMS	Ethyl-methanesulfonate
NGS	Next generation sequencing
BSA	Bulk segregant analysis

Abstract

Forward genetic screens remain one of the main genetic tools to characterize gene functions in plants. Recent advances in Next Generation Sequencing (NGS) technology have greatly reduced the time required for mutant identification in forward genetic screening. The major advantage of NGS enabled mapping, known as mapping-by-sequencing, is the simultaneous marker identification and genotyping and identification of the genomic loci causing phenotypes. We have been among the first to show that mapping-by-sequencing can be performed even within the same genetic background using mutagen-induced changes as segregating markers. As a proof of this concept, we mapped a previously unknown suppressor of *like heterochromatin protein1 (lhp1)* mutant. We developed a computational pipeline for the same and integrated it into an existing mapping-by-sequencing pipeline called SHOREmap.

Though mapping-by-sequencing is now being routinely used, less effort has been put in optimizing the experimental set-up. Therefore, we developed new computational pipeline called Pop-Seq simulator that can simulate different mapping populations and sequencing experiments. It simulates recombinant genomes by following empirical determined recombination frequency and landscape, which make simulations close to reality. Using Pop-Seq simulator we simulated different mapping-by-sequencing scenarios and created guidelines for mapping-by-sequencing experiments in *Arabidopsis*. Although mapping-by-sequencing has already become a standard method in *Arabidopsis*, the application in crops is hindered by the large genome sizes and the lack of complete reference genomes. Therefore, we have used the Pop-Seq simulator to extend our analysis on the experimental design of mapping-by-sequencing to two crop model species, rice and barley, in which next generation sequencing-based mapping becomes tangible reality. Besides, we have developed a reference-free method called NIKS (needle in the *k*-stack) that enables mapping-by-sequencing in species without pre-assembled reference sequence, gene annotation, or genetic map. NIKS directly compares genomes using k-mers from whole genome sequencing data to identify homozygous mutations and extend the sequence associated with mutation site by local *de novo* assembly. We have used *ab initio* gene structural prediction to annotate the effect of mutations, which led us to the

identification of causal mutation. This method will facilitate mapping-by-sequencing in non-model species.

Zusammenfassung

Vorwärts genetische Verfahren sind in der Pflanzengenetik eine der wichtigsten Methoden zur Identifizierung von Genen und ihrer Funktion. Jüngste Fortschritte in der Sequenziertechnik der nächsten Generation (engl. Next Generation Sequencing (NGS)) haben den Zeitaufwand für die Kartierung von Mutationen mittels vorwärts genetischer Verfahren drastisch reduziert. Eine neue Methode, die auch als Kartierung durch Sequenzierung (engl. Mapping-by-sequencing) bezeichnet wird, ermöglicht nun die simultane Identifizierung und Genotypisierung von Markern. Diese werden benötigt um die genomische Region, die einem Phänotypen zugrunde liegt, zu bestimmen. Wir waren unter den Ersten die gezeigt haben, dass Kartierung durch Sequenzierung im selben genetischen Hintergrund mittels Mutagen induzierter Marker durchgeführt werden kann. Dies konnten wir anhand der bereits bekannten Suppressions Mutante *like heterochromatin protein1* (*lhp1)* nachweisen. Für die Kartierung haben wir eine Pipeline zusammengestellt welche nun in SHOREmap integriert ist.

Obwohl Kartierung durch Sequenzierung ein routinemäßig eingesetztes Verfahren ist, wurde der Optimierung des experimentellen Aufbaus bisher wenig Aufmerksamkeit geschenkt. Aus diesem Grund haben wir eine Simulationssoftware (Pop-Seq Simulator) entwickelt, welche empirisch bestimmte Rekombinationsfrequenzen und –landkarten verwendet und somit realitätsnahe Simulation ermöglicht. Mittels der Simulation von verschiedenen Szenarien, bei denen Kreuzungsschemata und Sequenziertiefe variiert wurden, konnten wir Leitlinien für verschiedene experimentelle Setups in Arabidopsis erstellen. Auch wenn Kartierung durch Sequenzierung mittlerweile in Arabdopsis Standard ist, ist die Verwendung dieser Methode in Kulturpflanzen durch vielfach größere Genome und das Fehlen vollständiger Referenzgenomsequenzen erschwert. Aus diesem Grund haben wir unsere Analysen auf zwei Kulturpflanzen in denen Kartierung durch Sequenzierung schon jetzt möglich ist, Gerste und Reis, erweitert, um auch in diesen optimale experimentelle Setups zu bestimmen. Darüberhinaus haben wir mit NIKS (engl. Needle In the K-Stack) eine Methode entwickelt, die nicht auf einer Referenzgenomsequenz, Genannotation oder genetische Karte beruht. NIKS vergleicht Genome mittels k-*mers* aus NGS Daten, wobei homozygote Mutationen mittels lokalen Assemblies der Region gefunden werden. Im Anschluss werden

Genstruktur und Annotation vorhergesagt, welche die Bestimmung der kausalen Mutation ermöglichen. Durch diese Verallgemeinerung der Methode wird die Anwendung von Kartierung durch Sequenzierung über die Grenzen von Modellorganismen ermöglicht.

Chapter 1. Introduction

The genetic screens to identify the gene responsible for phenotypic variation have been a common task in genetics. In plants, identification of genes contributing to the variations in phenotype has great deal of implications not only in understanding fundamental processes but also for the betterment of crop (Rafalski, 2010). Genetic screens systematically associate observable characteristics or traits (known as phenotype) and the genetic make-up (known as genotype). During the course of time, different strategies have been deployed and these strategies have largely been classified into two major groups, forward and reverse genetics. Forward genetic screens select for a phenotype associated with a biological process and identify the genetic region contributing to the phenotype. Whereas reverse genetics screen select a gene of interest and analyze mutant of the gene in order to identify the process it has been involved with (Page and Grossniklaus, 2002; Alonso and Ecker, 2006).

The environment where the screen been conducted has influence on traits and higher the heritability of a trait, lesser is the influence of environment on phenotype (Paterson *et al.*, 1991; Mauricio, 2001; Collard *et al.*, 2005). Therefore, accurate genetic screen requires simultaneous recording of environment, phenotype and genotype. The fact that forward genetics allows the direct analysis of a biological process of interest without any prerequisite knowledge. The process of forward genetics screens start with random mutagenesis to introduce genetic variations that occasionally lead to phenotypes of interest. Subsequently, mapping localizes genetic element responsible for the phenotype.

Mapping experiments can be summarized as; first, identification of polymorphic markers between parental lines. Second, generation of segregating mapping population by crossing a parent with phenotype of interest to a suitable wild type parent. Finally, genotyping the mapping population at each marker position and associating phenotype to genotype in order to identify causal genetic region, known as mapping interval. This may further require fine mapping to reduce the genetic region under probe by using more segregant and even more markers, if available. If genome-wide information such as whole genome sequence and gene annotation is available, this could be utilized to pinpoint the candidate genes within the mapping interval.

Short life cycle, self-compatibility and relatively easy to be grown in a greenhouse has made *Arabidopsis thaliana*, a member of Brassicaceae family, a widely accepted model species in plant science (Meinke *et al.*, 1998; Somerville and Koornneef, 2002; Koornneef and Meinke, 2010). Moreover, features like foremost sequenced genome with a stable assembly, comparatively well annotated and characterized genes and amenability to forward genetic screen has made *Arabidopsis* a model system of choice for plant biologist.

Traditionally, mapping populations are derived from biparental cross between phenotypically diverged parents in order to map phenotype of interest. The primary mapping population derived from such a cross is F_2 progenies, facilitating uneven phenotype. The underlying genetic segregation causes such an uneven trait in F_2, which is utilized during mapping by associating the fixed allele within a group of progeny having phenotype under selection. In other words, by classifying the mapping populations into groups with same phenotype and by pooling DNA from group with phenotype of interest, mutation underlying the phenotype, as well as the closely linked genetic region will be selected. Thus, in case of a recessive phenotype the mutant region will be homozygous. On the other hand, due to recombination and independent chromosome assortment in each pooled plant, regions unlinked to mutation have equal likelihood to have both alleles, hence heterozygous in the pool. Thus, the allele frequency at and near the mutation will be one (all alleles are the same), and this frequency will gradually decline to the random expected frequency of 0.5 with increasing genetic distance from the mutation. This way of analysis is known as Bulk Segregant analysis (BSA) (Michelmore *et al.*, 1991).

BSA was first utilized in mapping resistant gene in lettuce by grouping mapping population into resistance and susceptible groups. Later this method became common for the phenotype with discrete groups. On the other hand, when the phenotype are complex, continuous and contributed from multiple loci, method called quantitative trait locus (QTL) mapping is used (Falconer and Mackay, 1996; Kearsey, 1998). In QTL mapping, each segregant plants are genotyped and phenotyped separately to associate genomic loci and their contributions to phenotype (reviewed by (Collard *et al.*, 2005)). Depending on the objective, advanced mapping populations were created by either selfing or backcrossing to produce recombinant inbred line (RIL) or near-isogenic lines (NIL), respectively. In RILs, heterogeneous homozygous genome state is obtained by several round of selfing. Whereas in NILs, repeated

backcrosses are done to introgress homozygous alleles from one parent to second parental background. Recently in *Arabidopsis* and in other plant species, advanced mapping populations, such as MAGIC (Multiparent Advanced Generation Inter-Cross) and AMPRIL (*Arabidopsis* multiparent RIL), were introduced by using multiple parental lines and complex crossing schemes to increase the power of mapping (Kover *et al.*, 2009; Huang *et al.*, 2011). AMPRIL crossing scheme represents the founder lines evenly, whereas MAGIC consists of more recombination events per lines (Weigel, 2012). As both designs consists of multiple parents that aids to study epistatic effects as well as is likely to generate unnatural allelic combinations. Together with these new methods of generating mapping populations, revolution in both genotyping and up to certain extent phenotyping made the mapping process faster and powerful than ever. Due to revolution in DNA sequencing, marker discovery and genotyping has gained tremendous speed and throughput in last decade and brought up new framework of Next Generation Sequencing (NGS) integrated approaches in mapping (Varshney *et al.*, 2011).

1.1. Changes brought by Next Generation Sequencing

During last century, plant breading has made remarkable progress in crop improvement by utilizing molecular markers and appropriate statistical methods, 21st century has even more to contribute with new developments in sequencing and associated methods (Collard and Mackill, 2008). Last decade has witnessed the revolution in DNA sequencing method from Sanger chain-termination technology to pyrosequencing or sequencing by synthesis methods by Roche and Illumina (Sanger *et al.*, 1977; http://www.454.com/; http://www.illumina.com; reviewed by Wall *et al.*, 2009; L., Liu *et al.*, 2012) The transition was massive, as the technology which served for decades had only few kb per run compared to more than million-throughput in new technology. Moreover the cost per base pair got reduced during the course of time, encouraging more sequencing project than ever before (Figure 1). Roche and Illumina have now several platforms with different throughputs to serve various needs. Illumina platform is known for the high throughput of ~6 billion per run with relatively short read length of 100- 300 bp (http://www.illumina.com). On the other hand, Roche has longer read length up to 1000 bp but comparatively lower throughput of ~1 million reads per run (http://www.454.com/). Recently introduced Ion Torrent has a read length and throughput intermediate to previously discussed technologies,

therefore apt for specific uses (http://www.lifetechnologies.com). Recent development of sequencing from a single molecule without any pre-amplification step provides longer reads with an average size of 3 kb but with lower throughput (http://www.pacificbiosciences.com). While longer reads as well as signal from an unbiased molecule is promising, current error rate of 11-14% hamper the utility of this method alone and requires support information from another platform (Roberts *et al.*, 2013). The other exciting but non-commercialized technologies like Nanopore sequencing, suggests that the advancement of DNA sequencing is still on its peak, and is projected to grow further. There are many more technologies, which are not mentioned above. Further comparisons between technologies have been published in various comparative studies (Liu *et al.*, 2012; Quail *et al.*, 2012; Jünemann *et al.*, 2013).

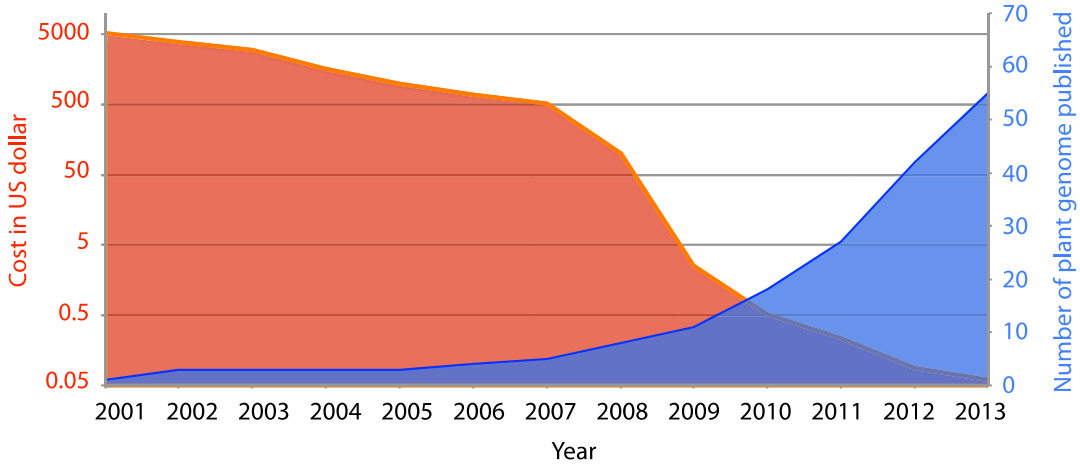

Figure 1: Improvement in DNA sequencing over the years. With NGS, cost per Mb of DNA sequencing (Red y-axis, Data from www.genome.gov/sequencingcosts) has decreased unprecedentedly and together with increase in throughput, the number of published plant species genome has increased (Blue y-axis).

The throughput of NGS has revolutionized many biological fields, including genetics and genomics. The growing sequencing throughput, has not only opened the door for the sequencing more samples but has also changed the way forward genetics has been conducted and practiced. The speed and throughput of NGS, has not only stimulated the whole genome sequencing projects but also resequencing projects, to investigate natural variations between accessions/ecotype. The number of plant species being sequenced is growing in a great speed as the cost per base pair is reducing (Figure 1). This momentum to generate more genomic resources with

respect to sequencing has got attention in non-model organisms as well (Tautz and Domazet-Lošo, 2011; Varshney *et al.*, 2011). Simultaneously, resequencing projects like 1001 *Arabidopsis* genome project (http://1001genomes.org), where the whole genome sequencing of multiple accessions/ecotypes has given rise to rich data sets of natural variations that have direct implication in genomic studies such as whole genome profiling (WSP) and genome wide association studies (GWA) (Nordborg and Weigel, 2008; Weigel, 2012).

Other than whole genome sequencing and resequencing, NGS has improved other biological study fields such as expression analysis, direct identification of DNA binding site (Chip-seq) or DNA methylation pattern identification, just to name a few. Most of the mentioned techniques are still under optimization and have recently been the subject of extensive studies to explore the utility of each or combined methods, which were not possible to carry out at all or in genome-wide scale before NGS.

1.2. NGS enabled forward genetics

In forward genetics, genetic mechanism of a phenotype is studied by introducing random mutations artificially by means of either chemicals or radiation, and plants showing phenotype of interest are selected to raise a mapping population. The aim of mutagenesis is to introduce maximal genomic variation with minimal reduction in viability and by screening this population, traits that are almost impossible to identify by conventional breeding are being developed and characterized at the molecular level (Sikora *et al.*, 2012). In order to pinpoint the molecular identity of the mutant, positional cloning is commonly employed with the help of available genomic information (Lukowitz *et al.*, 2000). However, the positional cloning of mutants can be time consuming because of lower recombination rates in the region. Feasibility of resequencing and recent developments in marker discovery and genotyping due to the up-rise in DNA sequencing methods has brought up new framework of NGS integrated mapping approaches. The major advantage of so called, mapping-by-sequencing, is the simultaneous discovery of segregating markers in the mapping population and genotype to identify mapping interval that contribute to phenotype. Finally, using the same data together with gene annotation, effect of mutation on gene function is predicted to pinpoint candidate genes. Hence, mapping-by-sequencing merges three steps that were required previously to one, thus speeding up the whole process.

Mapping-by-sequencing consists of multiple intermediate bioinformatic analysis, mainly comprising of resequencing analysis. Soon after the sequencing, base calling is done by converting the raw images of sequencing to bases with quality score. Different sequencing platforms use separate in-build pipelines for base calling. Phred scoring is the generally accepted quality scale in sequencing and represents log-transformed error probability at every base pair.

$$Q = -10 \log_{10} P\,(\text{error})$$

where Q is phred quality score and P is estimated error probability (Ewing and Green, 1998). Shotgun sequencing is prone to have sequencing error and vary in magnitude and profile based on the sequencing platform. For example, base qualities towards the end of reads get reduced in case of Illumina platform (Nakamura *et al.*, 2011). Therefore, it is necessary to remove the sequencing-specific artifacts such as poor quality reads, low base call and adaptor contamination, before starting resequencing analysis. The quality filtered reads are aligned to the reference genome in order to identify the genetic variations in the sequenced genome. There are multiple alignment tools available for general and specific use, depending on the sequencing platform and the type of reads. Most of the alignment tools use supplementary data structure called indices, for fast and memory efficient alignment. Majority of alignment tools implement based on either hash table (seed based) or suffix/prefix tree and use it for read sequence or reference sequence, or both. Alignment step is followed by identifying variant in sequenced genome. Different tools use different statistics to identify variation. Generally, there is overlap between the outcomes, though tools provide minor portion of unique variations. Along with other reviews, Pabinger *et al*. compiled a comprehensive list of tools and compared their performances (Nielsen *et al.*, 2011; Pabinger *et al.*, 2013). Apart from individual specific tools for specific tasks, pipelines such as SHORE, ngs_backbone, GATK and HugeSeq are few to name, that perform all the tasks sequentially (Ossowski *et al.*, 2008; Blanca *et al.*, 2011; DePristo *et al.*, 2011; Lam *et al.*, 2012).

Despite being slower, "seed" based read alignment methods are preferably used for such analysis for aligning short reads to reference genome sequence due to their robustness in identifying polymorphism (Jimenez-Gomez, 2011). Removal of multiple hit reads as well as duplicate reads helps in the realistic estimation of allele frequency in a pool. Selection of high quality markers improves the precision in mapping interval and assists to remove false mapping intervals (Galvão *et al.*, 2012;

Lindner *et al.*, 2012). Filtering variations between sequenced line and reference genome introduces deceptive variations called background mutations. Since background mutations are artifacts of analysis protocol, these are uninformative and needs to be filtered out. Depending on experimental setup, different approaches have been suggested consisting of liberal approach like filtering common markers present in multiple mutants and more conservative approach such as removing mutations present in non-mutagenized progenitor. The disadvantage of last method is the obligatory sequencing of progenitor that is avoidable in first approach (Uchida *et al.*, 2011; Nordström *et al.*, 2013).

The method of integrating BSA with NGS, was introduced in plants by mapping *Arabidopsis* recessive mutation in AT4G35090 gene from a forward genetic screen using an outcrossed mapping population (Schneeberger, Ossowski, *et al.*, 2009). In this case, segregating markers that were used for mapping, consisted of the natural variations between parents along with mutagen induced mutations. The method developed for this study is called as SHOREmap. Similar computational tools have been successfully developed in other studies to extend the analysis based on web-interface and cloud computing (Austin *et al.*, 2011; Minevich *et al.*, 2012).

In contrast to the initial studies done on an outcrossed population where considerable variation in phenotype may occur, mapping was done in a backcrossed population of Rice and *Arabidopsis* by crossing mutant plant to non-mutagenized progenitor and mutagen induced variations were used as segregating markers (Abe *et al.*, 2012; Hartwig *et al.*, 2012) (Chapter 2). However, in population derived from backcross, depending on mutagen, the number of segregating markers was typically lower and the average number of short reads at each marker position was usually lower than the number of plants pooled in bulk DNA. This impedes accurate allele frequency estimation, thus fictitiously including nearly fixed mutations as fixed ones. The identification of causal mutation was possible by identifying the fixed mutation with the help of deep candidate resequencing (dCARE); the true allele frequency estimation in a large bulk DNA by targeted resequencing. This approach displays how different mapping stages can benefit from different sequencing platforms (Chapter 2). After considerable number of backcrosses, it is even possible to directly resequence the mutant genome to identify the genetic causal region. However, this approach is highly time consuming and even after four rounds of backcross, one may end-up with large number of candidate genes spinning on more than one chromosome (Ashelford

et al., 2011). Nonetheless, direct sequencing approach has the advantage of having mutant genome sequenced for further characterization studies. Availability of multiple alleles for a phenotype makes direct individual resequencing a better option by identifying commonly disturbed genes from both allelic groups (Uchida *et al.*, 2011). In short, the key factors involved in mapping-by-sequencing are A) availability of reference genome, B) size of the genome, C) availability of genetic material such as alleles or mutants, and D) prior knowledge about mapping interval.

1.3. Guidelines for mapping-by-sequencing

All the above different strategies are now routine in forward genetic screening and are becoming replacement of traditional mapping procedure. As the cost of sequencing is going down and the number of organisms having a stable reference genome is increasing, more and more mutant identification by mapping-by-sequencing is being reported (Cuperus *et al.*, 2010; Golas *et al.*, 2013; Schreiber *et al.*, 2012; Tabata *et al.*, 2013). Practical application of mapping-by-sequencing requires decisions on the experimental setup right from generating mapping populations to the adjustment of next generation sequencing reaction. As both, the composition of mapping populations and the amount of sequencing are directly related to time and financial effort, thus it is important to optimize each step of mapping-by-sequencing experiments. However, there is only limited effort to optimize mapping-by-sequencing procedure by suggesting the best practical experimental design for such experiments (Austin *et al.*, 2011). It usually remains unclear what the expected outcome of mapping-by-sequencing experiments could be. Therefore, this usually leads to conservative decisions resulting in an excess of mutants and sequencing data. Although, different studies commented on the experimental design by sub-setting the data analyzed, these conclusions were either incomplete and may be specific to the given data, or failed to compare different mapping scenarios (Austin *et al.*, 2011; Abe *et al.*, 2012). A comprehensive study to suggest guidelines for mapping-by-sequencing should consider the effect of mapping population as well as sequencing depth on mapping outcome of the experiment along with other parameters like phenotyping error and availability of allelic groups. We have studied the effect of crossing scheme and mapping population with help of a newly developed simulation tool called Pop-Seq simulator, which considers empirical recombination frequency and landscape. From over 400,000 mapping-by-sequencing

simulations in *Arabidopsis*, we have studied the expected outcome of given experimental setups (Chapter 3). The utility of Pop-Seq simulator was further showed by realistic mapping-by-sequencing experiment setups in Rice and Barley.

1.4. Reference free NGS enabled mapping methods

However, all the above mentioned methods prerequisite the availability of a reference genome sequence. Though, the number of species having available reference genome sequence is increasing substantially, currently this requirement impedes the applicability of mapping-by-sequencing to larger portion of plant species. Comparative genomics approaches, like utilizing the genome sequence of closely related species is an alternative or even utilization of partial syntenic blocks as reference genome (Galvão *et al.*, 2012). Nonetheless, the evolutionary distances between species may become critical and are subject to failure due to lack of homology or even absence of genomic region in the closely related genome.

Apparently, a reference free method is needed in order to directly compare short reads from two samples. An algorithmic framework has been introduced in plants by mapping genes in *Arabis alpina* without genetic maps and reference sequences using k-mers (Chapter 4). The short DNA reads with a length of k (k-mer) was used to compare between mutant and parent to identify mutations. Subsequently building up a local assembly followed by *ab initio* gene structural prediction in order to predict the effect of mutation at gene level. This method also succeeded in identifying more mutations than a comparative genomics approach using *Arabidopsis* genome as reference.

Chapter 2. Fast Isogenic Mapping-by-Sequencing of Ethyl Methanesulfonate-Induced Mutant Bulks

This chapter explains the method for mutant identification by mapping-by-sequencing in an isogenic population. We demonstrate how mapping-by-sequencing and candidate gene identification can be performed within the same genetic background using mutagen-induced changes as segregating markers. As a proof-of-principal, we mapped the previously unknown suppressor of *like heterochromatin protein1* (*lhp1*) mutant, from ethyl methanesulfonate (EMS) forward screen by using mutagen-induced mutations as markers. *lhp1* in its functional form is involved in chromatin-mediated gene repression. As a method to identify the causal mutation from candidates, we introduced deep candidate resequencing (dCARE) using Ion Torrent Personal Genome Machine to resolve three linked candidate mutations in the mapping interval. dCARE reduced the number of causal candidate mutations to one, which was further confirmed by complementation studies. This study was published under Break Through Technologies in Plant Physiology 2012 (Hartwig *et al.*, 2012). Appropriate contents for this chapter are taken from the manuscript. This project was conceived together by Korbinian Schneeberger and Franziska Turck. Ben Hartwig performed the mutant screening as well as wet lab confirmation. I performed the analysis of short read sequences by developing SHOREmap backcross pipeline that was later integrated to SHOREmap.

2.1. Introduction

Recent advances in NGS technology have greatly reduced the time required for mutant identification in forward genetic screening. The major advantage of NGS enabled mapping, known as mapping-by-sequencing, is the simultaneous marker identification and genotyping to identify the genomic region causing phenotype. In plants, mapping-by-sequencing was introduced by mapping a mutant in an outcrossed population between *Arabidopsis* reference accession Columbia (Col-0) and a diverged accession Landsberg erecta (L*er*). Plants with phenotype of interest were bulked and sequenced to identify markers followed by allele frequency to detect mapping region (Schneeberger, Ossowski, *et al.*, 2009). For a recessive phenotype, pooled DNA from plants with phenotype shows homozygous mutation at the causal site. This principal was followed by other studies for successful mapping (Cuperus *et al.*, 2010; Austin *et al.*, 2011).

Although the integrated approach of BSA and resequencing is powerful and extremely fast, crossing with diverged accession to generate mapping population impairs the recognition of mutant with subtle phenotype. Moreover, for genetic screening of suppresser or enhancer of a preexisting mutation, availability of primary mutation in another suitable accession becomes inevitable. If not, this screening has to be first probed for the initial mutation, thus adding more complex and lengthy procedure to mutant identification. An alternative is to remain within the same accession background by backcrossing mutant to parental line. However, this will eliminate the opportunity to use natural variation among parents as markers. Thus, in a population generated by backcrossing mutant to parental line, unknown variations that are introduced by mutagen remain as markers.

Recently in rice, mapping was done using pooled mutants that were backcrossed to parental line and used provisional reference genome in order to filter out natural variations (Abe *et al.*, 2012). However, in this case prior knowledge about candidate gene was used to pinpoint the causal change. Alternatively, in *Arabidopsis*, direct sequencing of four times backcrossed mutant genome to parental line produced 103 putative casual mutations that had potential to change 48 putative proteins (Ashelford *et al.*, 2011). Moreover, these candidate mutations were clustered to two separate regions, demanding additional mapping information to prioritize candidates.

Two major tasks in analyzing mapping-by-sequencing data from a backcross population are; A) identification of mutagen induced variation as markers B) precise estimation of allele frequency in the pool as the number of plants pooled are typically higher than the average number of read count. This concern increases when the mutations are physically close by therefore has lower likelihood of conceiving recombination in between, thus results to delicate difference in allele frequency.

We have developed a backcross analysis pipeline and have integrated into existing SHOREmap tool. The identification of mutagen-induced variations was done by filtering out the markers identified in the genome of non-mutagenized parent. The putative candidate casual mutations were further analyzed in detail with deep candidate resequencing (dCARE) (Figure 2.1). dCARE involves targeted sequencing of bulk segregant DNA. As a proof-of-concept, we applied this method to screen suppressor of *like heterochromatin protein1* (*lhp1*) mutant. The pleiotropic phenotype of *lhp1* mutant plants differs quantitatively between accessions such as Wassilewskija-2 and Col-0, making it difficult to create a robust mapping population for subtle modifiers. Therefore *antagonist of lhp1-1 (alp1;lhp1)* double mutant was backcrossed to original *lhp1* allele and F_2 offspring of this cross gave 3:1 ratio for suppresser phenotype, indicating that a single mutation was responsible for the suppression.

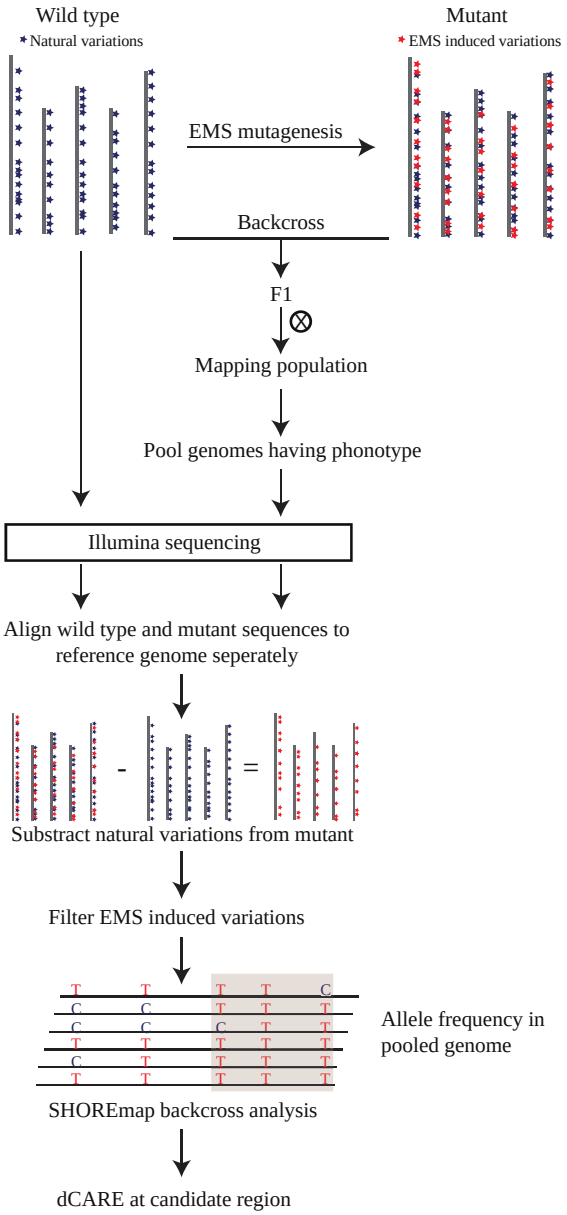

Figure 2.1: Schematic illustration of the fast isogenic mapping approach. Chemical mutagens typically introduce hundreds of novel mutations. Within the M_2 generation, mutants are screened for phenotypes. Selected plants are backcrossed to the nonmutagenized progenitor. The F_2 offspring of such a cross forms an isogenic mapping population, as only novel mutations are segregating. Backcrossed individuals that display the mutant phenotype are selected, bulked, and their DNA is prepared as a pool and whole-genome is sequenced. If the parental line is genetically different from the reference line Col-0, it needs to be resequenced in order to filter naturally occurring differences that need to be differentiated from novel mutations.

Thus, novel EMS-induced mutations can be selected for SHOREmap analysis by filtering for mutations that do not reside in the parental line (adopted from Hartwig *et al.*, 2012).

2.2. Materials and Methods

2.2.1. Sequenced material and SNP identification

For the mapping of *alp1;lhp1*, we have sequenced pooled DNA from leaf samples of 270 BC_2F_2 double mutant plants. In parallel, DNA from *lhp1* single mutant that was used as parent for this screen, was sequenced. Each reaction of sequencing was done on Illumina Genome Analyzer IIx. Each DNA clone was sequenced twice from both the ends to generate interconnected sequences reads called paired-end. We have generated paired-end reads having 2x 96 bp length for *alp1;lhp1* double mutant as well as parental line. We applied short read analysis pipeline, SHORE, for identification of SNPs and short INDELs. Using the function SHORE *import*, raw reads were trimmed based on quality values with a cutoff Phred score of +38. After *import*, 43.4 and 42.2 million high-quality reads from *lhp1* mutant and *alp1;lhp1* double mutant, respectively, were independently aligned to the Col-0 reference genome using GenomeMapper as an alignment tool (Ossowski *et al.*, 2008; Schneeberger *et al.*, 2009) (*Arabidopsis* Genome Consortium; The *Arabidopsis* Information Resource 10). Out of total high-quality reads, 93% and 94% of reads were aligned to the TAIR10 Col-0 reference sequence and yielded an average nucleic genome coverage of 41 and 49 fold for *lhp1* and *alp1;lhp1*, respectively (Table 2.1). The alignments were corrected for the expected paired-end distance of 300 bp by SHORE *correct4pe*. We applied SHORE *consensus* to both sequence sets to identify variations between the mutant and reference genome sequence. The minimum minor allele frequency for SNP calling was kept to 20%. Inbuilt SHORE heterozygous SNP configuration was used for SNP calling. Since the SNPs from *lhp1* were the natural variation between mutant line and reference sequence, these SNPs were filtered out from the double mutant to get mutagen-induced mutations. EMS changes (GC:AT) with a SHORE quality score > 24 and supported by more than seven reads, were used in SHOREmap *backcross* for allele frequency analysis. Allele frequency was calculated as

$$Allele\ frequency = \frac{C_m}{C_t}$$

where C_m and C_t are coverage of mutant allele and total coverage at locus, respectively. Sequence changes in the region that featured evidence for selection were

annotated for their effect on gene identity using TAIR10 gene annotation. See Appendix note I for further details on command line calls for the resequencing and mapping-by-sequencing analysis.

2.2.2. Targeted deep resequencing of individual mutations

Later on the putative candidate mutations were amplified for dCARE analysis by designing primers with the help of Primer3 (version 0.4.0) to amplify 80 to 150 bp amplicons. These amplicons contained the candidate mutations at a distance from +1 to +50 from the 3' end of the primer that contained the A-type extension required for Ion Torrent PGM sequencing. DNA was amplified from the same pool of DNA as used for whole-genome resequencing. Amplicons were purified and sequenced in an Ion Torrent PGM (Life Technologies) using a 316K chip to a depth of 5,000 to 20,000 reads per amplicon. Allele frequencies of both the wild type and mutant were estimated from raw reads. Using a 21-mer around the mutation site, an ad-hoc script was used to count the allele occurrence with perfect match or one mismatch. Coverage at each locus was calculated by the sum of satisfying reads from the above criteria.

Table 2.1: Resequencing summary of mutants. Resequencing output of each mutant (adopted from Hartwig *et al.*, 2012).

	Chromosome	Unique bases sequenced	Mean depth of sequencing	Coverage of genome
alp1:lhp1				
	1	1167674295	38.38	99.94
	2	827866858	42.03	99.96
	3	1025324459	43.71	99.96
	4	773853919	41.64	99.96
	5	1081475692	40.09	99.96
Mean			41.17	99.96
lhp1				
	1	1393315836	45.79	99.96
	2	994873293	50.51	99.98
	3	1229419565	52.41	99.98
	4	912048405	49.07	99.97
	5	1283372543	47.58	99.97
Mean			49.07	99.97

2.3. Results

2.3.1. Analysis of isogenic mapping population for mutant identification

After quality trimming and SNP identification from short read analysis using SHORE, we have identified 14225 and 13721 variations from *alp1;lhp1* double mutant and *lhp1* mutant respectively compared to Col-0 reference genome. Mutagen induced changes in double mutant *alp1;lhp1* was identified by removing all sequence differences identified in *lhp1* mutant. These variations identified in *lhp1* mutant are the natural variations between the line used for mutagenesis and Col-0 along with sequencing error and resequencing artifacts. *alp1;lhp1* specific SNPs having at least eight reads of support and a SHORE quality score greater than 24 were retained that made 1351 mutations for further allele frequency analysis. As EMS introduce mainly G/C:A/T mutations, we filtered for those and were left with a set of 412 novel EMS changes (Appendix Table 1). Allele frequency of each EMS mutation in the pool was plotted against chromosomal position to identify the fixed genomic region. Selection for the lower arm of chromosome 3 became apparent through an allele frequency distortion in this region (Figure 2.2). Across the five chromosomes, there were only three mutations that had a mutant allele frequency higher than 80% and clustered on the lower arm of chromosome 3. Functional prediction of these three mutations based on the TAIR10 gene annotation was that two mutations were located in exons of AT3G57940 and AT3G63270 and one in an intron of AT3G61130. Moreover, the first two mutations caused missense mutation leading to amino acid changes of Val→Ile and Gly→Glu, respectively (Figure 2.3). The script used for EMS induced mutant identification as well as frequency analysis and visualizations were compiled for download under SHOREmap *backcross* analysis package. The complete package is downloadable from (http://shoremap.org).

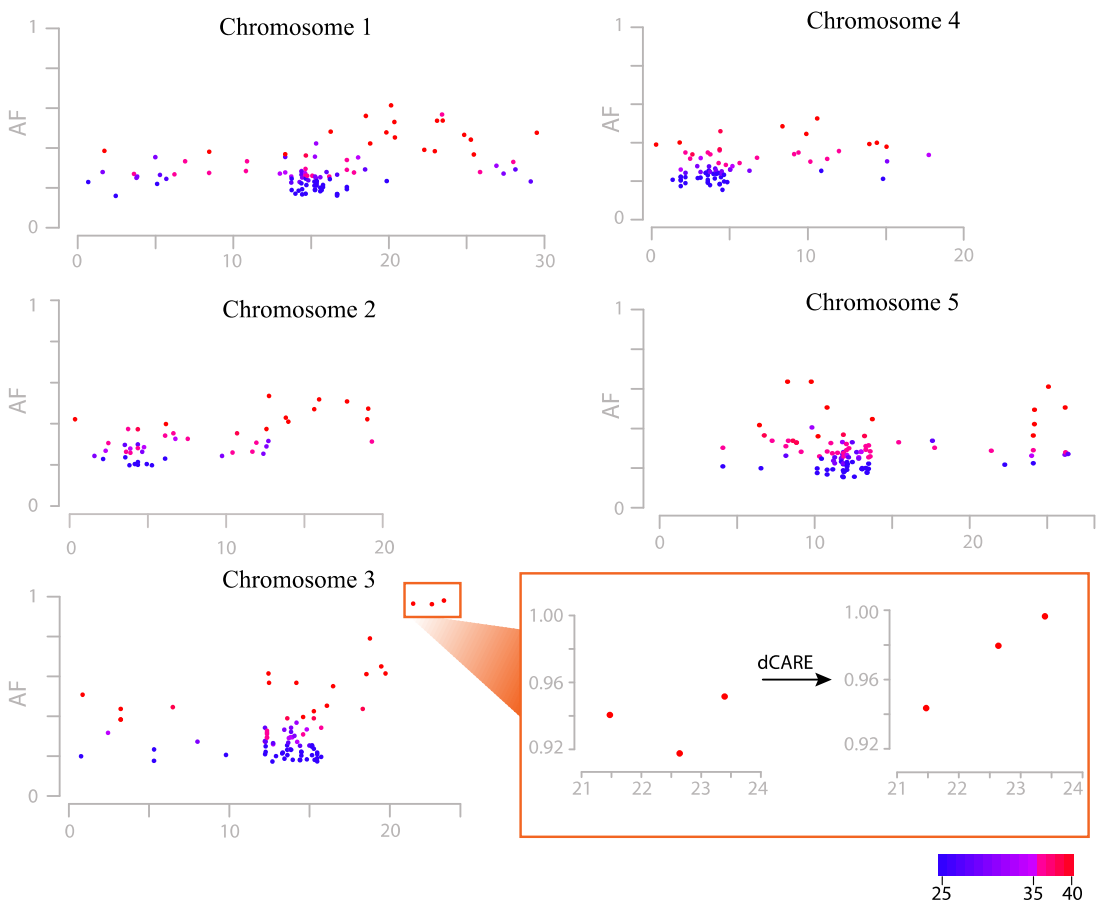

Figure 2.2: Allele frequency estimations at EMS changes. Allele frequency estimations at EMS-induced mutations of *alp1;lhp1* across all five chromosomes were shown (x-axis: Chromosomal position in Mb). Allele frequencies (AF; y-axis) were estimated as fractions of short reads supporting the mutant allele divided by the number of all reads aligning to a given marker. The color indicates the resequencing consensus (SHORE) score, and only base calls with a quality score of more than 25 have been considered. The long arm of chromosome 3 was found to be under selection, as local allele frequencies appeared to be higher as compared to other regions in the genome. This region was magnified to show the allele frequency difference in detail (orange box) and the estimated allele frequency from dCARE was shown on right side (adopted from Hartwig *et al.*, 2012).

2.3.2. dCARE identifies causal change

Although the putative candidate mutations were spaced over 2 Mb apart, nearly complete linkage between the three candidate mutations was apparent in the pooled DNA. Based on *Arabidopsis* genetic maps, this physical distance corresponds

to approximately 7 to 8 centimorgan, suggesting that several recombination events between these mutations are expected in a pool of 270 recombinants (Giraut *et al.*, 2011). Our analysis of the raw reads from illumina sequencing covering the three mutations revealed two Col-0 wild-type reads out of 50 and 48 reads, for the mutation in AT3G57940 and AT3G61130, respectively. Whereas, the mutation in AT3G63270 had one wild-type read out of 41 reads. Although the mutation in AT3G63270 could therefore act as main candidate, the disparity was too minor to reliably exclude the other mutations. As usually the number of individuals pooled in bulk segregant analyses is considerably larger than the average whole genome resequencing coverage, thus not powerful enough to resolve the real allele frequency accurately. Therefore, an increased number of short-read alignments at the mutation sites by a targeted deep resequencing of mutant locus would enable to determine the allele frequency in a bulked DNA much more precisely. In order to generate more sequencing data for the mutated regions, we amplified regions across the mutations by PCR using the pooled DNA from bulked segregant as template and sequenced the amplicons with the Ion Torrent Personal Genome Machine (Rothberg *et al.*, 2011). This dCARE analysis generated 20,111, 4,390, and 19,203 reads across the candidate mutations in AT3G57940, AT3G61130, and AT3G63270, respectively. For the changes in AT3G57940 and AT3G61130, we found 5.7% and 2.1% reads supporting wild-type allele, whereas only less than one percentage of the reads at AT3G63270 supported the wild-type allele. The presence of Col-0 wild-type reads at all candidate mutations can be explained by contamination of the segregant bulk, possibly due to mis-scoring of mutants or by sequencing errors that occur at a low rate. Both types of error affect mutations independent of their linkage to the causative change and represent background noise. In fact, the rate of wild type allele at AT3G63270 is even slightly lower than the rate of sequencing errors reported for Ion Torrent PGM sequencing (Rothberg et al., 2011). As a consequence, we could not reliably identify any wild-type alleles for the mutation affecting AT3G63270, whereas the wild-type allele was clearly apparent for both linked mutations (Table 2.2). Thus, dCARE reduced the list of candidates to AT3G63270. dCARE demonstrates the apt utilization of different NGS platforms on mapping-by-sequencing experiment. Utility of dCARE resembles the traditional fine mapping procedure in a mapping-by-sequencing context.

Table 2.2: Raw reads and allele frequency calculations at three putative candidate mutation locus. Illumina platform was used for whole genome resequencing, whereas dCARE used Ion Torrent sequencing platform for targeted deep resequencing (adopted from Hartwig *et al.*, 2012).

Seq-run	Chr	Pos	AGI	EMS	WT	Cov	A		C		G		T		N	
							Cov	Fre	Cov	Fre	Cov	Fre	Cov	Fre	Cov	Fre
Illumina	3	21455099	AT3G57940	A	G	50	47	94.0	0	0.0	2	4.0	0	0.0	1	2.0
	3	22622352	AT3G61130	T	C	48	0	0.0	2	4.2	0	0.0	44	91.7	2	4.2
	3	23376305	AT3G63270	T	C	41	0	0.0	1	2.4	0	0.0	39	95.1	1	2.4
dCARE	3	21455099	AT3G57940	A	G	20111	18966	94.3	0	0.0	1145	5.7	0	0.0	0	0
	3	22622352	AT3G61130	T	C	4390	0	0.0	90	2.1	0	0.0	4300	97.9	0	0
	3	23376305	AT3G63270	T	C	19203	0	0.0	86	0.4	0	0.0	19117	99.6	0	0

Independent of dCARE analysis, AT3G63270 was established as the *antagonist of lhp1 (alp1)* by complementation study and test cross between independent alleles. ALP1 encodes a gene related to Harbinger-like transposases. From phylogenetic study of available homologous, ALP1 is likely to be derived from an ancient Harbinger transposon but seems to have acquired a plant-specific function over time. However, ALP1 is an expressed gene that is not directly regulated by LHP1 and the Polycomb Group (PcG) pathway, thus required further study to reveal its function and interaction with *lhp1* mutant (Hartwig *et al.*, 2012).

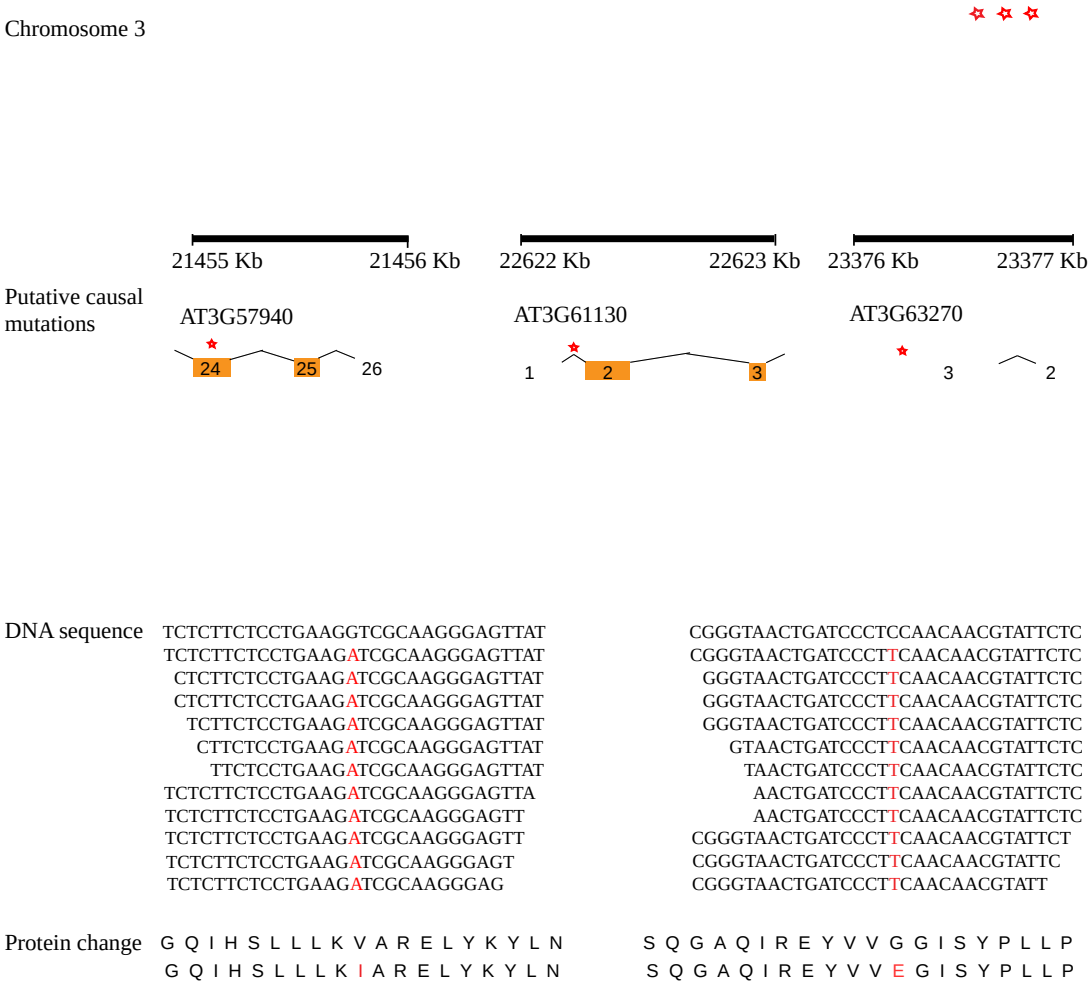

Figure 2.3: Annotation of putative causal mutations. The genomic regions of candidate EMS mutations (red asterisks) along with gene annotations are shown (top). Only partial gene structure is shown where orange boxes indicate exons. Locations of EMS mutations that have putative effects on amino acid sequences are shown in red letters; for clarity, the DNA sequences in the graph do not reflect the actual number of reads at these locations (coverage was shown in table 2.2) (adopted from Hartwig *et al.*, 2012).

Chapter 3. User guide for mapping-by-sequencing in *Arabidopsis thaliana*

Though mapping-by-sequencing accelerates mutant identification by combining genetic mapping with whole-genome sequencing, less effort has been put in optimizing the experimental set up. Moreover, different strategic approaches reported so far has not compared comprehensively. This chapter explores the different strategies and optimal experimental design for each of the mapping-by-sequencing scenarios. The guidelines are formulated based on simulations of different experimental setups mainly the type of mapping population, sequencing coverage and sequencing methods by following empirically determined recombination frequency and landscape of *Arabidopsis thaliana*. Using a newly developed simulation tool called Pop-Seq simulator, different mapping populations and sequencing experiments were simulated to replicate different mapping-by-sequencing scenarios in-silico. This study was published recently in a special edition of Genome Biology for Plant Genomics (Velikkakam James *et al.*, 2013). Within few weeks' span of time, this paper was designated as 'Highly accessed' and made to top two in the list of most popular recently viewed articles. Appropriate contents for this chapter are taken from the published manuscript. Korbinian Schneeberger and myself designed this study. The simulation tool; Pop-Seq simulator was designed and implemented by me together with Vipul Patel and Korbinian Schneeberger. Karl J.V. Nordstrom and Jonas R. Klasen helped with ad-hoc external scripts for SHOREmap analysis and collecting recombination frequency for rice and barley, respectively. I simulated different scenarios and performed analysis to formulate results. Patrice Salome and Detlef Weigel provided empirical recombination data for *A. thaliana*.

3.1. Introduction

Forward genetic screens remain one of the major genetic tools to discover gene function in plants as well as in other organisms. Soon after the realization of applicability of NGS in mapping, several analysis pipelines have been introduced and were already applied to various model species, including plants, yeast, nematodes, mammals, and invertebrates (Schneeberger, Ossowski, *et al.*, 2009; Birkeland *et al.*, 2010; Doitsidou *et al.*, 2010; Austin *et al.*, 2011; Abe *et al.*, 2012; Hartwig *et al.*, 2012). The two main parts of mapping-by-sequencing is first, to generate a mapping population, and second, selection of mutant and whole genome shotgun resequencing.

Different types of crossing schemes for mapping-by-sequencing have been suggested to develop mapping population. The very first mapping-by-sequencing experiments were performed on pooled genomes of mutant recombinants that were generated by crossing the mutants to diverged strains followed by one round of selfing (Schneeberger, Ossowski, *et al.*, 2009; Cuperus *et al.*, 2010). Recently, several groups suggested use of backcrossed instead of outcrossed individuals as mapping population, as mutagen-induced changes segregate like natural polymorphisms. Even though there is no prior knowledge about their distribution or location, mutagen-induced changes can be identified within whole-genome sequencing data and subsequently used for mapping (Abe *et al.*, 2012; Hartwig *et al.*, 2012; Lindner *et al.*, 2012). Similarly, direct sequencing of an individual mutant recombinant, as suggested for *Caenorhabditis elegans* and later for *A. thaliana*, will allow for a rough mapping of the causal mutation (Zuryn *et al.*, 2010; Ashelford *et al.*, 2011). Although multiple rounds of backcrossing are usually not sufficient to considerably minimize the size of linked regions around causal mutations, this strategy has the advantage to characterize the complete genome of a mutant recombinant. Alternatively, direct sequencing of two or more independently generated alleles of the same mutant followed by a subsequent search for genes that carry mutations in all mutant alleles is powerful enough to unambiguously identify the causal mutation (Uchida *et al.*, 2011; Nordström *et al.*, 2013).

Irrespective of the actual strategy, application of mapping-by-sequencing involves decisions on the experimental makeup, for instance the size of the mapping population, as well as the amount of next generation sequencing data. Since both are directly related to time and financial effort, it is important to optimize the setup of

mapping-by-sequencing experiments. The lack of general guidelines describing an optimal design might lead to conservative decisions that prime an unnecessarily high number of individuals and sequencing coverage.

Within this study, we established guidelines for mapping-by-sequencing for *A. thaliana* by simulation. Simulation studies are powerful and well utilized to study the power of linkage map and crossing scheme with respect to QTL identification in different mapping population (Slate, 2008; Klasen *et al.*, 2012). However, simulations solely dependent on theoretical calculation may differ from reality. Thus, we developed a simulation tool called Pop-Seq simulator that follows an experimentally established recombination landscape. We simulated more than 400,000 mapping-by-sequencing experiments to analyze the differences in the design of mapping populations in relation to the number of candidate mutations identified in the course of such an experiment. Pop-Seq simulator consists of two parts, first, Pop simulator which simulates virtual genotype from a given cross of two parents. Successively, these virtual genomes that are represented in genotypes are passed to Seq simulator, which generate the sequence read count per allele per locus. Furthermore, we evaluated the impact of technical aspects, such as read length and read pairing, on mapping-by-sequencing.

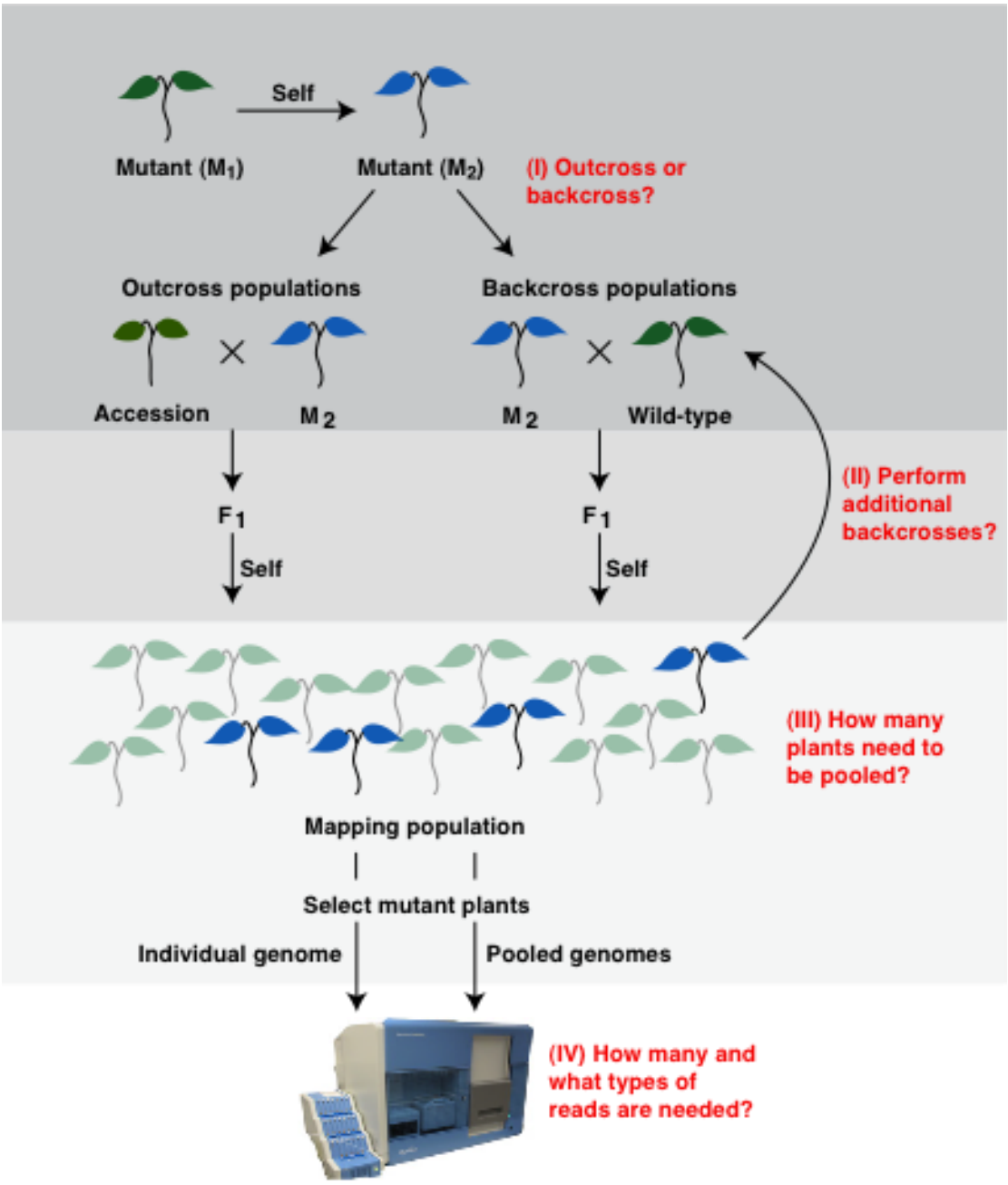

Figure 3.1: Overview of different strategies in mapping-by-sequencing. Various questions during experimental setup of mapping-by-sequencing are shown in red. (I) Mutants can be crossed to diverged accessions or backcrossed to the wild-type. (II) The number of backcrosses and number of plants used as parents contribute to the outcome of mapping-by-sequencing. (III) The number of mutant plants sampled from mapping population greatly impacts the mapping results. (IV) Finally, the sequencing coverage as well as type of sequencing (single-end or paired-end) affects the outcome of mapping-by-sequencing (adopted from Velikkakam James *et al.*, 2013).

Even though our simulations were focused on *A. thaliana*, our simulation pipeline is generic and can be applied to other species as well as other mapping or sequencing strategies. In the last section, we describe the extension of our analysis on the experimental design of mapping-by-sequencing to two crop model species; rice and barley, in which next generation sequencing-based mapping becomes tangible reality.

3.2. Materials and Methods

3.2.1. Simulation of recombinant populations by Pop simulator

We implemented Pop-Seq simulator in Perl 5.14.2. Both simulation tools could stand-alone and were based on object-oriented programming. In case of Pop simulator, initial parental genomes were generated using user specified number of homozygous makers placed randomly or at the specified locus. If the marker under selection was not specified, then one of those markers was randomly selected as causal mutation and used for selection at the end of each population stage if specified. Similarly, a wild-type genome was simulated except marker loci with wild-type allele. Throughout the Pop simulator, each parental allele was coded internally with parental name and decoded back at the end of the simulation. In order to simulate offspring genomes, we combined recombined haploid genomes from one or two virtual parents. Offspring genomes were used as parents for further crosses. During each cross, the virtual gametes were generated from each genome by determining the number and location of recombination involved. The actual number of recombination per meiosis for each chromosome was randomized based on the distribution of recombination events in *Arabidopsis*; these empirical determined recombination frequencies were derived from a cross between *Arabidopsis* Col-0 and Fei-0 (Salome *et al.*, 2011). It was calculated by

$$X \sim Trinomial(n, [p_1, p_2, p_3])$$

where p1, p2 and p3 are the observed frequencies of none, one and two or more recombination per chromosome per meiosis. The location of each recombination was selected after the observed frequencies over each marker along the chromosome and placed in-between two adjacent markers. The probability of a recombination at position x_{ij} between two adjacent markers was calculated by

$$p_{\left(x_{j_i}\right)} = \frac{p_{(m_i)}}{l_i},\ i = \{2,3,..,\mathrm{k}\}\ ,\ j_i = \{1,2,..,l_i\}$$

where i, j, k and l are the marker, base pair, total number of markers and length between adjacent markers, respectively. $p_{(m_i)}$ is the observed probability of recombination in between marker m_i and m_{i-1}. The location of additional recombination events was modeled after a gamma distribution in order to take crossing over interference into account. Both gamma distribution parameters scale and shape were chosen such that the resulting distribution followed the empirical data. One gamete genome from each parent was randomly selected to make offspring genome. This step was repeated to generate user specified number of mutagenic plants. Depending on the user specification, mutant phenotype was classified as recessive or dominants, and mutant plants were selected accordingly. As a parameter, user can either specify the number of mutant plants or total number of segregant in population. The crossing scheme can be defined by simple encoding where backcross and selfing are represented by "B" and "S", respectively. For example, F2:B1:F1:B1:F1 for generating BC_2F_2 by crossing F_2 and recurrent parent to make BC_1F_1 followed by one round of selfing and repeating the backcross and selfing cross to make BC_2F_2. Moreover, multiple parent crosses are possible, but current version is limited to only four parents with limited option to generate recombinant inbred lines. Empirical configuration data about the species recombination frequency and rate per marker are specified in Configuration file. Along with recombination information, species specific information such as chromosome number and size are also specified in this file. The complete options of Pop simulator are explained in Appendix note III.

3.2.2. Simulation of whole-genome sequencing by Seq simulator

Accurate simulation of whole-genome sequencing of bulks and individual genomes needs to consider the total number of alignments per marker, the parental allele frequencies and sequencing errors. To incorporate the variation in the number of alignments per marker, we assigned a prior normalization *n* value to each marker position based on the observed coverage in real resequencing experiments of *Arabidopsis* wild-type (Schneeberger *et al.*, 2011). The value describes the ratio of observed coverage at single marker in relation to the genome-wide average. Actual number of reads at each marker position c_i per sequencing simulation was then calculated by

$$c_i \sim Multinomial(m, [n_1, n_2, .. n_k])$$

where m, n and k are the total number of reads, normalized coverage probability per marker and the total number of markers, respectively. Then, we used the allele frequency a_1, and a_2 within the population under investigation and assigned each read r_i to one of the parental alleles by

$$r_i \sim Trinomial(c_i, [a_1, a_2, s]),$$

where a1, a2 and s are the allele frequency of mutant, allele frequency of wild-type and sequencing error respectively. We obtained the allele frequency at each marker position from the virtual genome generated by Pop simulator. It is also possible to simulate individual mutant genome where allele frequency is 0, 0.5 or 1. In both, pooled or individual, it is possible to adopt the results from any other source to Seq simulator by following the format of Pop simulator. We used a constant sequencing error rate of 0.3% (Galvão *et al.*, 2012). The frequency of different types of sequencing errors in Illumina sequencing is non-randomly distributed, however as this would have a limited impact on our simulations, we did not address this fact here (Ossowski *et al.*, 2008).

3.2.3. Selection of homozygous mutations

Definition of homozygous mutations are influenced by pool size and sequencing coverage. In order to define a uniform threshold for the detection of homozygous mutations across all deeply and shallowly sequenced pools with a few or many plants, we introduced two thresholds representing pool size and sequencing coverage. First, we calculated the mutant allele frequency at loci where one single wild-type chromosome is present, defined as

$$g_f = 1 - \left(\frac{[m * 2] + 1}{n * 2}\right)$$

where m and n are the number of mis-scored and total mutants in the pool, respectively. For the second threshold, we calculated the mutant allele frequency as estimated by the short read alignments, where one alignment is sampled from a non-mutant chromosome, defined as

$$r_f = 1 - \left(\frac{\lceil c_p * e \rceil + 1}{c_p} \right)$$

where c_p is the actual coverage at position p and e is the estimated sequencing error frequency of 0.3% (Galvão *et al.*, 2012). Only mutations with mutant allele frequencies greater than g_f and r_f have been considered as homozygous mutations.

3.2.4. High quality marker selection for outcross simulation

For all simulations based on outcross populations we defined a high quality marker set based on resequencing data of *A.* thaliana L*er* compared to TAIR10 reference assembly (Schneeberger *et al.*, 2011). SNPs having low score as well as being in a regions which is hard to resolve through resequencing, may mislead and negatively influence the mutant identification, therefore we filtered this SNP sets based on the quality score and the local vicinity. All SNPs with a resequencing quality score below 25 were discarded, as well as SNPs that overlap with regions with different copy numbers between the parents as predicted by the resequencing. Further we iteratively removed SNPs, which were closer than 50 bp. This yielded 291,973 high quality markers.

3.2.5. Simulation of mapping-by-sequencing

Simulation of in-silico mutant genomes was done by Pop-Seq simulator and was started by creating an initial mutant genome with 700 or 1,400 randomly placed, homozygous mutations. Depending upon the population under study, single, double or three rounds of backcross was made followed by one round of selfing. After each round of selfing, the homozygous mutant plant was selected to proceed for next generation. Contrastingly, in outcross populations, preselected 291,973 markers were used to generate mapping populations. Only one round of selfing was made to generate F_2 segregating populations and various mutant plants were pooled for respective experiments.

3.2.6. Comparison of single-end and paired-end sequencing

Single and paired-end sequencing was simulated with reads ranging from 50 to 750 bp in length. Insert length for paired-end sequencing was simulated with three times the read length. For each combination of read length and sequencing type 100,000 random alignment locations were chosen. The read length defined the end of

the alignment. The actual location of the read pair was defined by read length and insert size. If the simulated alignments overlapped with one or more markers the alignment was scored as informative.

3.2.7. Availability of simulation pipeline

Our pipeline; Pop-Seq simulator, for simulating recombinant genomes and emulate output of a NGS analysis is available at https://sourceforge.net/projects/popseq/. Recombination frequency and landscape are specified by configuration files, which we provide for all simulations performed in this study.

3.3. Results

3.3.1. Paired-end versus single-end sequencing

Many of the new sequencing technologies allow sequencing of one or both ends (paired-end) of DNA clones. However, it is not yet clear, which kind of sequencing is most appropriate for mapping-by-sequencing (Figure 1). Paired-end sequencing enables access to (the borders of) repetitive sequences, which increases the number of markers and mutations that can be analyzed. Even though single-end sequence reads might not be able to explore the same genomic space as paired-end sequence reads, they are independent of each other. In bulk segregant sequencing, independent reads are counted to estimate allele frequencies. If both reads of a pair align to different markers, they cannot contribute twice to the estimation of allele frequencies as they carry the same genetic background (ignoring the very rare cases, where read pairs span recombination events). If two single-end reads overlap with markers, both contribute to the estimation of allele frequencies as they have been sampled independently. It is thus not obvious whether paired-end or single-end sequencing is advantageous for mapping-by-sequencing.

We have compared the efficiency of single and paired-end reads by counting the number of randomly generated read or read pair alignments that overlap with predefined marker. A read, respectively a pair, was scored as informative if it was uniquely aligned to at least one or more markers. The length of the simulated reads ranged from 50 to 750 bp to cover a wide range of next generation sequencing read length (Figure 2.1). Reads, which align equally well to multiple regions in the genome, are excluded for further analysis. Increased read length span some of the short repeats and thus allows aligning more reads uniquely (Cahill *et al.*, 2010; Koehler *et al.*, 2011).

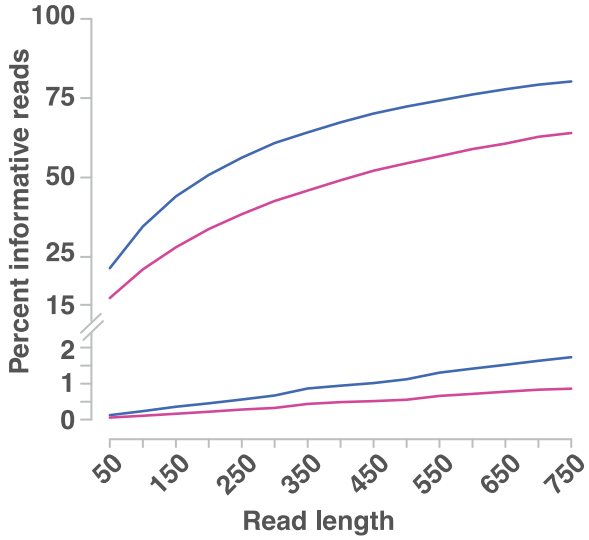

Figure 3.2: Percentage of informative reads for different sequencing read lengths and types. Only informative reads or read pairs that overlap with at least one marker or mutation can contribute to mapping-by-sequencing. The number of informative reads from single-end and paired-end sequencing are shown in purple and blue, respectively. The lower part of the graph refers to resequencing of backcross population that has a lower mutation density (here, 1,400 mutations per mutant genome). While the upper graph refers to markers in outcross populations (281,668 and 291,973 for single-end and paired-end sequencing, respectively) (adopted from Velikkakam James *et al.*, 2013) .

For the analysis of mapping-by-sequencing with outcross populations, we defined two sets of 291,973 and 281,668 markers for paired-end and single-end sequencing, respectively, in order to take the different mapping properties into account. Depending on the read length, paired-end sequencing featured between 25 and 78% more informative read pairs. Consequently, it would require between 25 and 78% more single-end reads in order to end up with the same number of informative reads. This calculation allows for a cost comparison of mapping-by-sequencing for single and paired-end sequencing based on actual sequencing costs. However, as paired-end sequencing enables the analysis of parts of the otherwise inaccessible DNA, it might be advantageous to sequence both ends, even if this would be more expensive. In particular if combined with mutation identification, paired-end sequencing has a higher chance not to miss the causal mutation. We repeated this exercise for mapping-by-sequencing based on backcross populations that were

simulated with 1,400 mutations in the genome. Here, paired-end sequencing featured between 95 and 119% more informative read pairs.

3.3.2. In-silico mapping-by-sequencing experiments

Assessing different types of mapping-by-sequencing experiments require establishment and sequencing of thousands of mapping populations, which is practically not feasible in plants. In contrast, in-silico simulations do allow for the generation of many experiments, with the potential caveat that they rely on prior assumptions. In particular, genuine simulations of mapping-by-sequencing experiments require realistic assumptions about mutation load, next generation sequencing and meiotic recombination.

The most commonly used mutagen for *Arabidopsis* is ethyl methanesulfonate (EMS), a chemical mutagen that predominantly introduces C to T and G to A changes. There are various reports about the frequency of EMS-induced mutations, including one change in 112 kb to one change in 171 kb, indicating a dosage dependency of the mutation rate, which suggests that the actual frequency range is likely to be much wider (Jander *et al.*, 2003; Ashelford *et al.*, 2011). In order to explore the effects of different mutation rates, we simulated low (700 changes) and high (1,400 changes) rates of mutations that were randomly introduced into the genome.

Similarly, realistic simulations of next generation sequencing rely on correct assumptions about the number of short read alignments per reference position (from here on referred to as coverage) and sequencing errors. As we were only interested in coverage at marker loci, we simulated whole-genome sequencing by randomizing the number of read alignments at each marker. The absolute number of alignments per marker followed a coverage distribution assessed on real resequencing experiments using Illumina sequencing. Deriving the coverage distribution from real sequencing experiments has the advantage that it considers all factors that contribute to the variation in sequence coverage. Perhaps most prominently, several different groups have demonstrated that local GC content is correlated with sequence coverage which is consequently also represented in our coverage landscape (Ossowski *et al.*, 2008; Aird *et al.*, 2011). Moreover, in a recent study we rigorously assessed the sequencing error rate of sequence reads aligned to marker positions, where the actual per base sequencing error rate was between 0.09 and 0.21% after quality filtering (Galvão *et*

al., 2012). In order to avoid overly optimistic simulation we assumed a sequencing error rate of 0.3% in our simulations. Based on these assumptions each of the simulated read alignments was then assigned to a parental allele, following a multinomial distribution based on local allele frequencies within the bulked segregant and Illumina sequencing-specific error rate (Materials and methods).

Most important, however, might be realistic simulations of recombinant genomes that greatly rely on frequency and location of recombination. Thus, we based our simulations on experimentally determined recombination frequencies derived from F_2 population established by crossing two diverged *Arabidopsis* accessions (Salome *et al.*, 2011). These data reveal the number of recombination in a single cross as well as their distribution over the physical range of the chromosomes. We used the frequency of recombination events along the chromosomes as a probability function after which recombination location and frequency were simulated (Materials and methods).

This method for in-silico simulation of recombination breakpoint events can be applied to any type of crossing regime. In this study, we focused on three different types of mapping-by-sequencing scenarios (Figure 3.1). First, we simulated F_2 mapping populations generated by crossing a mutant plant to a non-mutagenized accession with a diverged background followed by selfing of the F_1 hybrid (as performed by (Schneeberger, Ossowski, *et al.*, 2009; Cuperus *et al.*, 2010; Austin *et al.*, 2011; Schreiber *et al.*, 2012)). We refer to these classical mapping populations as "outcross populations". In outcross populations, natural variations along with mutagen-induced changes serve as genetic markers. A second type of population was simulated by backcrossing the mutant plant to the non-mutagenized progenitor, followed by selfing of the hybrid (as performed by (Abe *et al.*, 2012; Hartwig *et al.*, 2012)). We refer to these mapping populations as "backcross populations", in which only mutagen-induced changes serve as markers.

In contrast to the previous two methods, which makes use of recombination, the third type of simulation constitutes direct sequencing of individual mutant genomes selected from the backcross populations (as performed by (Zuryn *et al.*, 2010; Ashelford *et al.*, 2011)). In the next sections, we explore the consequences of different crossing schemes and the effect of pool size and coverage on the extent of the resulting mapping interval and on the number of candidate mutations (CAMs).

3.3.2.1. Mapping-by-sequencing with outcross populations

Mapping-by-sequencing with outcross populations is based on mutant allele frequencies assessed at large-scale marker sets leading to the identification of mapping intervals. Such regions can then be screened for novel mutagen-induced changes using the same whole-genome sequencing data (see (I) in Figure 3.1). Usually a rough identification of linked region suffices, as even in larger region sequencing data can easily be screened for CAMs. In order to evaluate this process we used the mapping-by-sequencing analysis pipeline SHOREmap, which implements a likelihood ratio test statistics that converts mapping-by-sequencing data into confidence-mapping intervals (Galvão *et al.*, 2012). These mapping intervals represent the region in which causal candidates reside at a given confidence level p (here p=0.99). As we assume that mutations are randomly introduced into the genome, the number of CAMs is linearly correlated with the length of mapping intervals, which we used to quantify the outcome of a mapping-by-sequencing experiment. Though marker density positively impacts on mapping resolution, inclusion of markers that cannot be accessed with the actual sequencing methods or that have been falsely included can have severe local effects on the precise determination of mapping intervals (Galvão *et al.*, 2012). The marker set we used consisted of 291,973 markers, after discarding closely linked polymorphisms and those in repetitive regions from the complete set of differences between *Arabidopsis* accessions Columbia (Col-0) and Landsberg erecta (L*er*) (Schneeberger *et al.*, 2011) (Material and methods).

3.3.2.1.1. Interplay of pool size and genome-wide coverage

Outcross populations were simulated with 40 to 400 mutant genomes. Next generation sequencing was simulated at various genome-wide coverage levels ranging from 5 to 200x. Each combination of pool size and coverage was independently repeated for 500 times. For each data set we performed a SHOREmap analysis and assessed the size of the final mapping intervals (Figure 3.3). Overall, the sizes of the mapping intervals were remarkably variable. This variation was lower for pools with more recombinants as compared to pools with fewer recombinants. As expected, the number of recombinants also strongly influenced mapping resolution. For example, at an average genome-wide coverage level of 15x, pools with 200 recombinants yielded an average interval size of 381 (± 222) kb, whereas pools with 50 recombinants

generated interval sizes of 783 (± 567) kb on average. Like in conventional mapping experiments, the decrease in the size of the mapping interval was not linear. The first indication of saturation was observed at a sequencing coverage of 5 to 15x, where increasing the pool size beyond 350 recombinants did not improve the interval size.
In contrast to pool size, coverage alone had only a small effect on size and variation of mapping intervals. Pools of 100 recombinants, which were sequenced at 15x, yielded an average interval size of around 500 (± 310) kb, as compared to 419 (± 298) kb at a coverage of 200x. The reason for the weak impact of coverage on the size of the mapping interval is the large number of markers, which are distributed throughout the genome allowing for an accurate assessment of allele frequencies even at low coverage levels.

Assuming 1,400 mutagen-induced mutations per genome, the average number of CAMs was around five for pools of more than 100 recombinants sequenced at an average genome-wide coverage of 25x. In practical application, additional prioritization by functional annotation and location of mutations in the interval has the potential to reduce this low number of CAMs to one outstanding candidate only (Schneeberger, Ossowski, *et al.*, 2009).

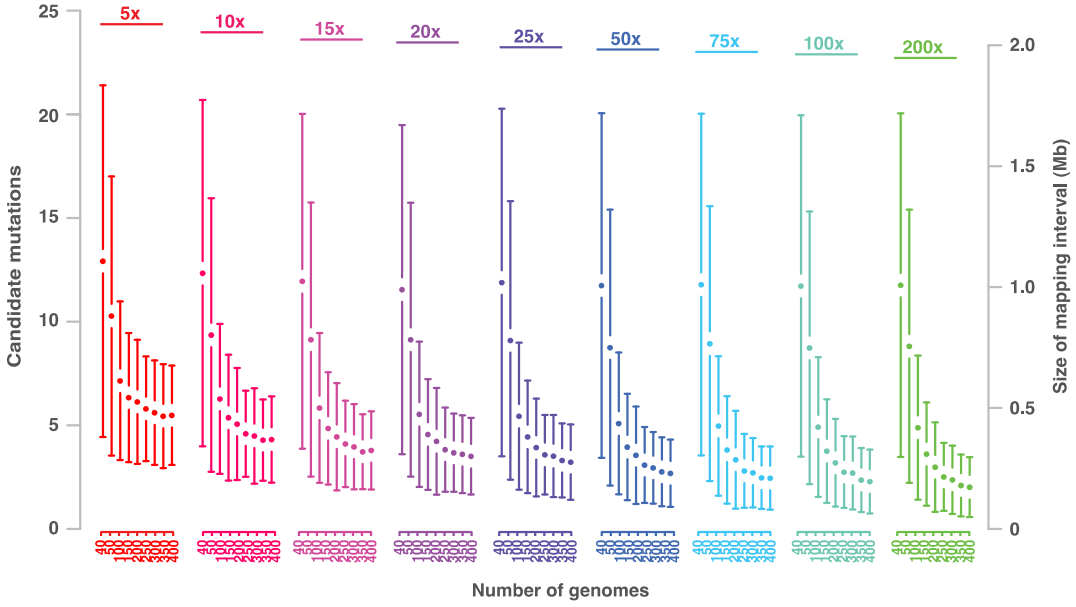

Figure 3.3: Results of mapping-by-sequencing with outcross populations. Pools of 40 to 400 individuals (colored blocks) were sequenced with increasing coverage ranging from 5 to 200x. For each combination of pool size and coverage we simulated 500 independent populations and performed a mapping-by-sequencing analysis on each of them. Average mapping interval size with one standard deviation as well as

the imputed number of candidate mutations within mapping region are shown on the right and left y-axis, respectively. The initial number of mutations per genome was 1,400 (adopted from Velikkakam James *et al.*, 2013).

3.3.2.2. Mapping-by-sequencing with backcross populations

Conventional genetic mapping requires a cross of the mutant to a diverged genome. In addition to genetic variation, this introduces phenotypic variation, which can interfere with the recognition of subtle phenotypes. Moreover, if the mutagenesis was performed in a complex (transgenic or otherwise mutagenized background) this needs to be introgressed into the diverged genome, if tedious genotyping of all recombinants for the presence of first site mutations should be avoided.

In order to bypass these obstacles, it has been suggested to use F_2 populations derived from backcrossing the mutant plant to the non-mutagenized progenitor as mapping populations (Abe *et al.*, 2012; Hartwig *et al.*, 2012). Within backcross populations all mutagen-induced mutations segregate, except for the causal and closely linked mutations, which are fixed in the mutant pool by selecting the mutant phenotype. Thus, selection for fixed differences between the mutant pool and its genetic background considerably reduces the number of putative causal changes. To quantify results of each simulation, we used the number of homozygous differences between the mutant pool and the background. However, the absolute number of homozygous mutations greatly depends on the definition and settings of parameters used for their identification. As sequencing errors can introduce wild-type alleles at otherwise homozygous loci, selecting only those positions without reads that support the wild-type allele excludes some of the real homozygous mutations. On the other hand, including positions, which support wild-type alleles, will introduce false positives. In order to allow comparison across samples, we defined and applied thresholds, which are adjusted to pool size and sequencing coverage (Materials and methods). Backcrossing was simulated by crossing a single mutant plant to its isogenic parent followed by one generation of inbreeding to establish a BC_1F_2 mapping population (see (I) in Figure 3.1).

3.3.2.2.1. The interplay of pool size and genome-wide coverage in BC1F2 populations

We simulated BC1F2 populations with 3 to 70 mutants for high and low mutation rates separately. Sequencing was simulated at different coverage levels, ranging from 5x to 200x. For each combination of pool size and coverage level, we simulated 500 independent mapping populations and scored the number of homozygous mutations (Figure 3.4). Mutations that are not fixed, but are close to fixation have a high probability to appear as fixed in the sequencing data. This effect becomes stronger at low coverage levels, where the reduced number of reads does not allow identifying low frequencies of wild-type alleles. As expected, more recombinants reduced the average number of homozygous candidate mutations. Sequencing pools with 30 recombinants at coverage of 25x revealed 43 (± 18) CAMs on average. Like for outcross populations, the variation of CAMs was high in pools with few recombinants, but got reduced in larger pools.

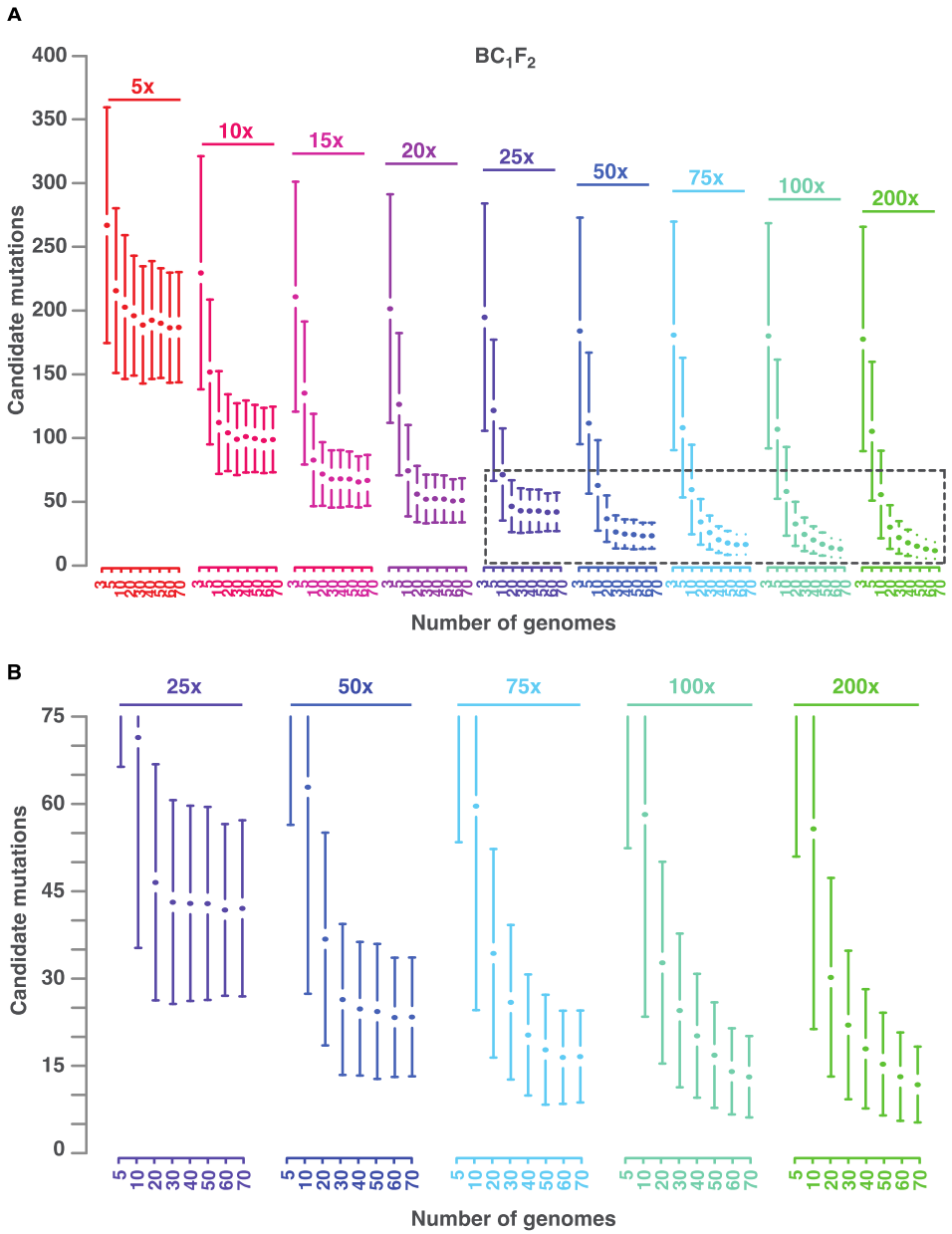

Figure 3.4: Results of mapping-by-sequencing with backcross populations. A) Pools of 3 to 70 BC_1F_2 individuals (colored blocks) were sequenced with increasing coverage ranging from 5 to 200x. For each combination of pool size and coverage we simulated 500 independent populations and performed a mapping-by-sequencing analysis on each of them. Average number of candidate mutations with one standard deviation is shown on the y-axis. The initial number of mutations per genome was 1,400. B) Zoom in on the framed region in panel A. Pools with three recombinants are not shown (adopted from Velikkakam James *et al.*, 2013).

In great contrast to outcross populations, we observed immediate saturation of the number of CAMs with increasing pool size. For example, pools with 20 mutants sequenced at a coverage level of 20x revealed 56 (± 22) CAMS on average. Pools

with 70 mutants, which were sequenced with the same sequencing effort, revealed almost the same number. In general, for coverage levels having less than 25x, we observed no reduction in the number of CAMs when the pool size is increased beyond 20 recombinants. This suggests that low-fold sequencing lacks the power to make use of the compliment of recombination in the pool and more sequencing is required to exploit all recombination events. In agreement, we still observed a decrease in CAMs for deeply sequenced samples (200x) when pool size is increased from 60 to 70. This illustrates the mutual importance of both pool size and coverage.

3.3.2.2.2. Effects of successive backcrossing

In a series of simulations, we increased the number of backcross generations up to three before establishing a mapping population (see (II) in Figure 3.1). In total, mapping-by-sequencing of 81,000 BC_2F_2 and BC_3F_2 populations were compared to the prior analysis of BC_1F_2 pools. As expected, additional backcrosses reduced the variation of CAMs in pools with a few plants (Figure 3.5). In particular, when genome-wide coverage or the number of mutants was limited, additional rounds of backcrossing helped to reduce the number of CAMs. However, pools with a reasonable number of recombinants sequenced with sufficient coverage did not improve with additional backcrosses.

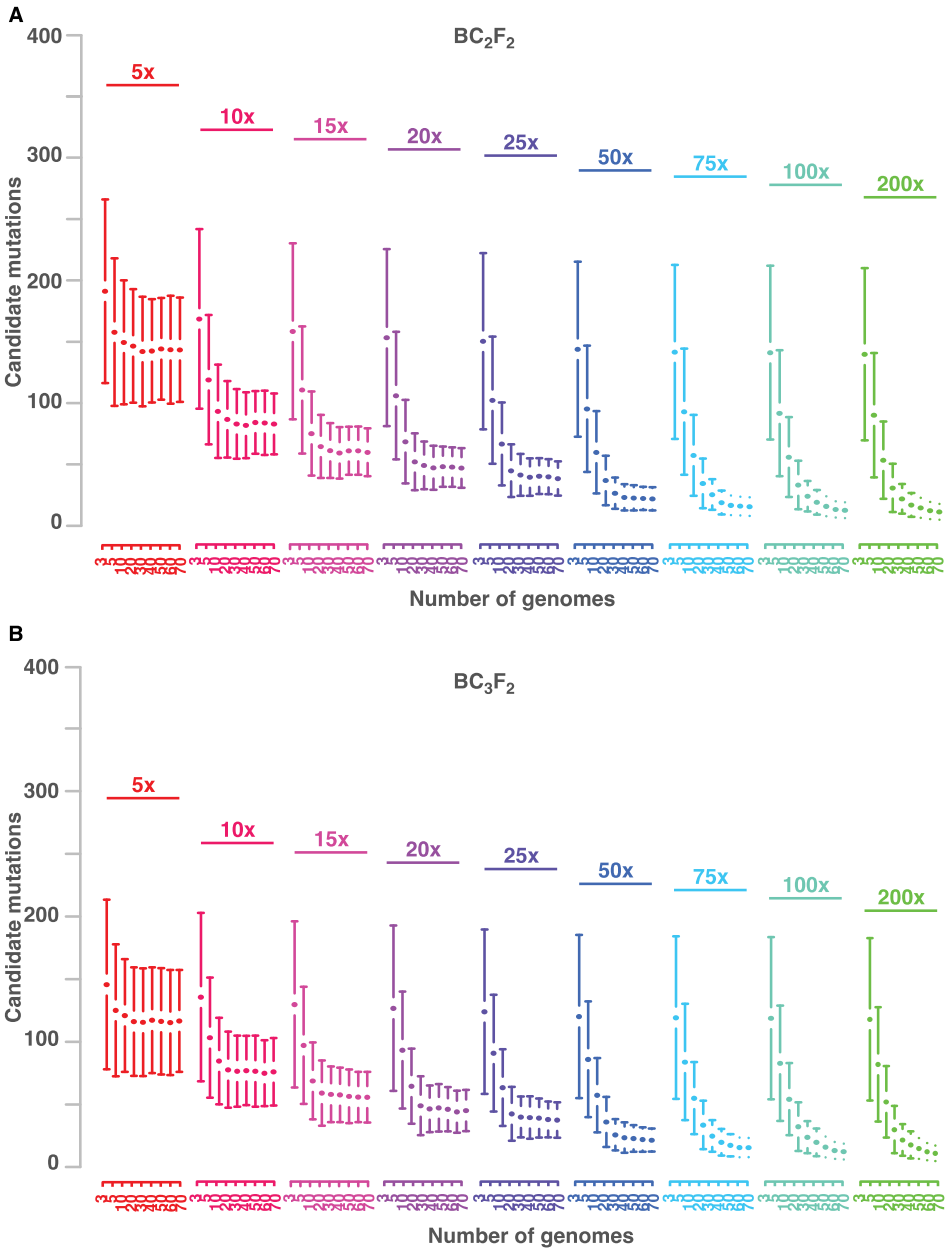

Figure 3.5: Effect of coverage and pool size on BC2F2 and BC3F2 backcross populations. A) Pools of 3 to 70 BC2F2 individuals (colored blocks) were sequenced with increasing coverage ranging from 5 to 200x. For each combination of pool size and coverage we simulated 500 independent populations and performed a mapping-by-sequencing analysis on each of them. Average number of candidate mutations with one standard deviation is shown on the y-axis. The initial number of mutations per genome was 1,400. B) Outcome of the same analysis with BC3F2 recombinants (adopted from Velikkakam James *et al.*, 2013).

In order to test the influence of mutation load, we simulated whole backcross simulations explained above with different initial 700 mutations per genome (Figure

3.6). This only reduced the CAM proportional to the change in the mutation load and persisted the trend observed in previously tested mutation load. Both mutation loads together mimic the realistic mutation load which one could expect from a chemical treatment such as EMS (Jander *et al.*, 2003; Ashelford *et al.*, 2011). Within this magnitude, mutation load has least influence on experimental setup.

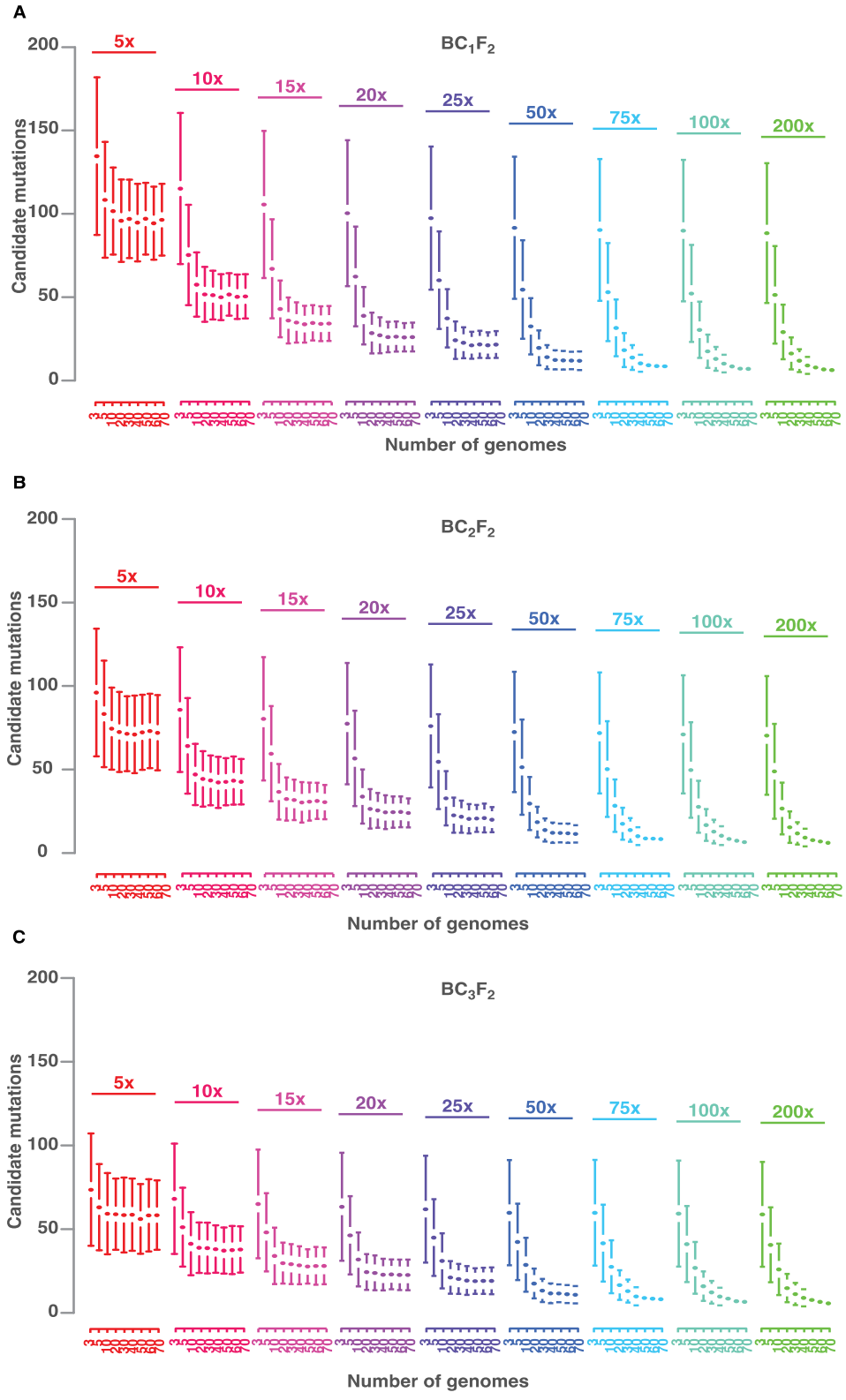

Figure 3.6: Effect of coverage and pool size on BC_1F_2, BC_2F_2 and BC_3F_2 backcross populations. Outcome from different populations, BC_1F_2, BC_2F_2 and BC_3F_2 are shown on panel A, B and C respectively. On each panel different number of recombinants are pooled ranging from 3-70 (each block on x-axis) and the same

pools are sequenced with multiple coverage levels, ranging from 5-200x (shown on top of each blocks) to illustrate the effect of both pool size and coverage. Mean number of candidate mutations with one SD is on y-axis. Mutation load of the genome is 700 mutations (adopted from Velikkakam James *et al.*, 2013).

Backcross populations are usually derived from one single mutant plant and its wild type parent. However, generation of a backcross population is based on multiple mutant siblings, all of which are crossed to their wild type parent, may bring additional variation around the causal locus. Here, we simulated the generation of backcross populations using three mutant siblings and compared the mapping outcome to our previous results, which were based on one mutant parent only (Figure 3.7). The improvement in mapping resolution was very limited and restricted to pools with few mutants only.

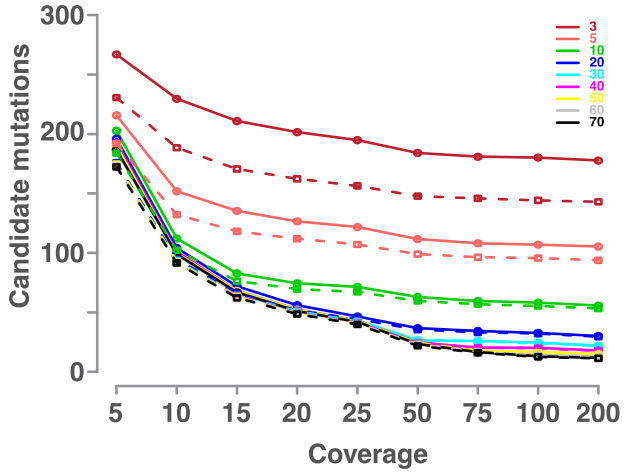

Figure 3.7: Effect of number of individuals used during backcross. The average mapping outcome of BC_1F_2 population are developed by crossing wild type parent to individual mutant (bold line) or three mutants (dotted lines) as parent. Mutation load of simulation is 1400 mutation per genome. Different colors; dark red, light red, green, blue, cyan, purple, yellow, gray and black indicate the size of recombinants in pool 3, 5, 10, 20, 30, 40, 50, 60 and 70 respectively (adopted from Velikkakam James *et al.*, 2013).

3.3.2.3. Effects of mis-scored plants

Complex or subtle phenotypes can lead to mis-scored plants. Such plants introduce wild-type alleles at the causal candidate locus and severely interfere with genetic mapping. In order to study the effect of mis-scored recombinants, we simulated different rates of mis-scored plants ranging from 1 to 6% within a

population of 50 BC_1F_2 mutants sequenced at 50x (Materials and methods). Compared to previous results, pools with 1 to 2% false scored plants yielded 82% and 145% more CAMs, respectively (Figure 3.8). This illustrates that even small errors in the phenotyping can have severe effect on mapping-by-sequencing based on backcross populations.

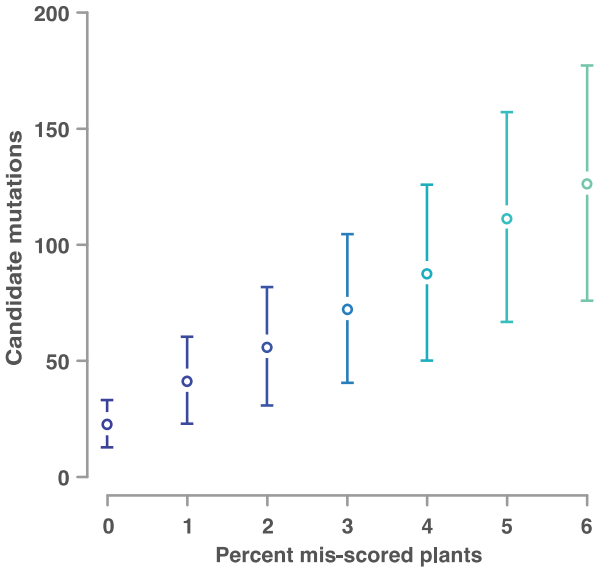

Figure 3.8: Effect of phenotyping error on BC1F2 population. The effect of phenotyping error ranging from 0 to 6% and respective observed candidate mutations are on y-axis (adopted from Velikkakam James *et al.*, 2013).

3.3.2.4. Direct sequencing of mutant genomes

As an alternative to bulk segregant analysis, individual mutant genomes can be sequenced directly (see (III) in Figure 3.1). However, the large number of background mutations interferes with the unambiguous identification of causal mutations. Backcrossing removes some of these background mutations (Zuryn *et al.*, 2010; Ashelford *et al.*, 2011). Here, we analyzed mutant genomes after one to three rounds of backcrossing. Mutants that are selected from backcross populations will generally yield fewer CAMs. The theoretical fraction of the recurrent parental genome after n rounds of backcrossing is (2^(n+1)-1)/2^(n+1) (Collard *et al.*, 2005). Our simulated populations closely followed the expected percentage and showed an average reduction of foreground genome by 12.8% and 6.8% in BC2 and BC3 respectively. As expected, direct sequencing yielded more CAMs than in our bulk segregant analyses. For example, across all coverage levels, pools with no more than three

BC1F2 mutant individuals showed less than half of the CAMs as compared to direct sequencing of BC1F2 individuals, illustrating the power of bulk segregant analysis.

3.3.3. Application of mapping-by-sequencing simulations in model crop species

Mapping-by-sequencing has already been successfully applied to crop species, like rice and polyploid wheat (Abe *et al.*, 2012; Trick *et al.*, 2012; Nordström *et al.*, 2013). As the size of some of the crop genomes can be as large as multiple Gb, an informed decision on the experimental design of mapping-by-sequencing seems even more important for such species. Here, we explored the power of Pop-Seq simulator to address questions about the experimental design of mapping-by-sequencing experiments in rice and barley, where mapping-by-sequencing has started to become a part of standard molecular toolbox.

3.3.3.1. Mapping-by-sequencing in the crop model species rice

First, we estimated the recombination frequency and landscape of rice by combining two publically available rice RIL populations (Harushima *et al.*, 1998; Huang *et al.*, 2009). Further, we selected a publically available set of 139,244 markers for the simulation of outcross populations (McNally *et al.*, 2009). Similar to *Arabidopsis*, we randomly introduced 2,222 mutations (1 every 171 kb), of which one was selected to be causal. Based on this, we simulated mapping-by-sequencing using both outcross and BC_1F_2 backcross populations with 50 to 400 and 10 to 80 mutant genomes, respectively (Figure 3.9). Sequencing of these pooled genomes was simulated at various genome-wide coverage levels ranging from 10 to 100x. Each combination of pool size and coverage was simulated for 300 times.

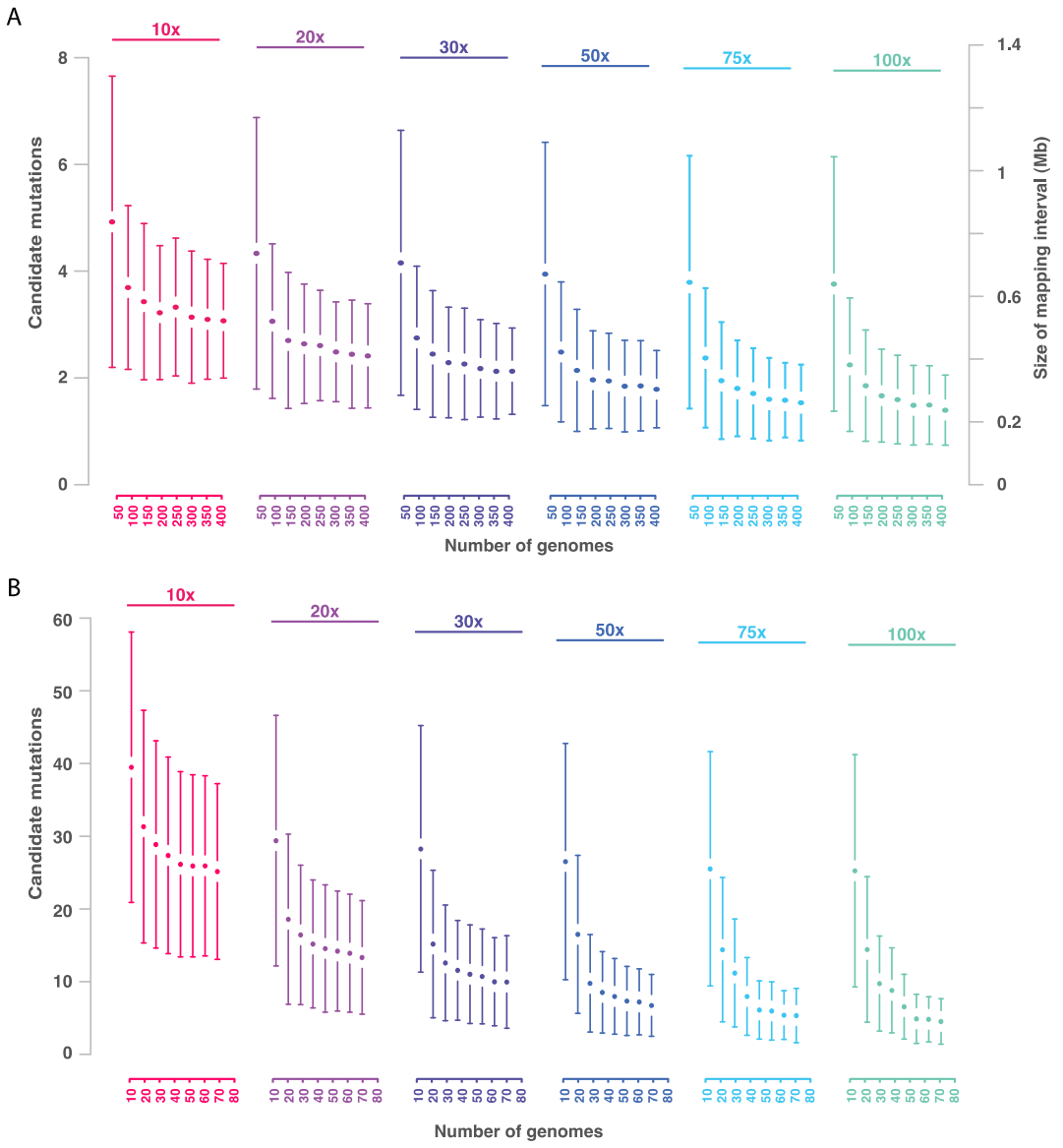

Figure 3.9: Simulated Mapping-by-sequencing outcome from rice. A) and B) show the simulated outcome of mapping-by-sequencing in outcross and backcross rice populations, respectively (adopted from Velikkakam James *et al.*, 2013).

Overall, we observed very similar trends for mapping-by-sequencing in rice as compared to *Arabidopsis*. Changes in the genome-wide coverage affected the outcome of backcross populations more than outcross populations and pools with very low number of recombinants drastically suffered from the lack of recombination. Outcross populations with 150 mutant recombinants sequenced with not more than 20x featured less than 3 CAMs on average in our simulations. In contrast, backcross

populations consisting of 50 mutants, which were sequenced at a genome-wide coverage of 50 yielded around 10 CAMs on average.

In general, the greater genome size of rice as compared to *Arabidopsis*, was counteracted by an enriched recombination frequency allowing for similar conclusions on the experimental design in rice as in *Arabidopsis*.

3.3.3.2. Mapping-by-sequencing based on targeted enrichment sequencing

As-of-today, the large genome sizes of crop species like the one of Hordeum vulgare (barley) make whole-genome resequencing as part of mapping-by-sequencing an expensive and risky approach. To address this general problem genome-complexity reduction methods, like transcriptome sequencing, restriction site associated DNA sequencing or targeted enrichment sequencing, have been proposed (Baird *et al.*, 2008; Gnirke *et al.*, 2009; Turner *et al.*, 2009; Elshire *et al.*, 2011). For example, targeted enrichment sequencing has been already proven to be suitable for mapping-by-sequencing (Galvão *et al.*, 2012).

Here, we simulated targeted enrichment sequencing of ~60 Mb of the barley genome. This included the simulation of deep sequencing at selected regions of the genome, but at the same time the simulation excluded the rest of the genome from sequencing. Even though enrichment sequencing has a high chance to exclude the causal mutation from the actual sequencing data, mapping-by-sequencing based on enrichment sequencing will guide subsequent fine-mapping efforts.

The design of the enrichment reduced the set of genome-wide marker as defined between the two cultivars Morex and Barke from 11,371,643 to 164,492 markers, which are accessible through our enrichment sequencing (Mayer *et al.*, 2012). Mapping populations were simulated with 50 to 600 mutant plants selected from F2 outcross populations and were based on the recombination frequency and landscape for barley as observed in the Oregon Wolfe Barley mapping population (Cistué *et al.*, 2011). Sequencing was simulated at coverage levels of 100 to 1,000x reflecting the high coverage gained in enriched regions. Each combination of pool size and coverage was simulated for 300 times.

Overall, the reduced recombination frequency in barley as compared to the other species resulted in large mapping intervals (Figure 3.10). Similar to the observations for the other two species, increased coverage had only a minor effect on the results of outcross populations-based mapping-by-sequencing, but an increase in

the number of mutants can have a strong effect on the size of the mapping interval. Simulation of mapping populations with 400 mutants that were sequenced with an average coverage of 200x at the enriched regions resulted in mapping intervals with an average size of 3.2 Mb.

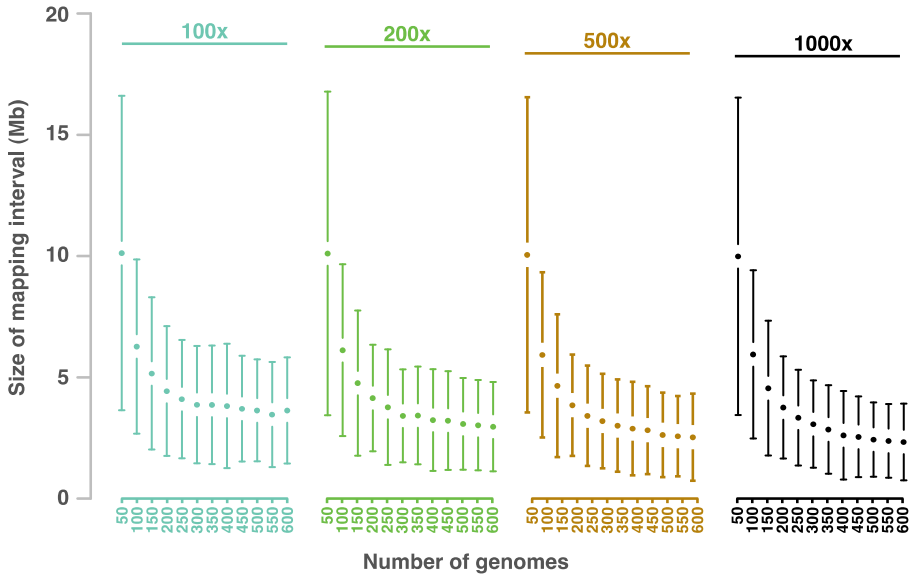

Figure 3.10: Targeted enrichment sequencing in barley. Simulation outcome of ~60 Mb targeted enrichment sequencing in barley (adopted from Velikkakam James *et al.*, 2013).

Chapter 4. Mapping-by-sequencing in non-model organism

Previously discussed mapping-by-sequencing strategies and proposed computational methods are prerequisite for a reference genome and gene annotation; therefore it is limited to species with well-characterized genomes. The constraint of characterized genome can be abolished either by having a close relative genome sequence or a method that compares wild-type and mutant directly without aligning the short reads to a reference genome. We propose two approaches, first, a comparative genomics approach where a close relative genome is being used as intermediate to identify mutations. Second, we introduce a reference-free algorithm called NIKS (needle in the k-stack) based on comparing k-mers in whole-genome sequencing data for identification of homozygous mutations. We applied both approaches to two mutants of non-model species *Arabis alpina*. NIKS successfully identified causal mutation, whereas the approach with a mediator genome was hampered due to lack of conservation. In case of NIKS, the effect of mutation was characterized by both *ab initio* and homology based annotation. This study was published in Nature Biotechnology 2013 (Nordström *et al.*, 2013). Korbinian Schneeberger and George Coupland designed this study. Karl J V Nordström implemented NIKS and applied to mutant samples. Karl J V Nordström and myself performed the analysis. Maria C Albani, Caroline Gutjahr, Benjamin Hartwig, Franziska Turck, Uta Paszkowski provided the biological material.

4.1. Introduction

With various suggested strategies in crossing as well as multiple computational frame works, mapping-by-sequencing has become a routine procedure for characterization of mutant from forward genetic screens. However, so far the utility of this fast and powerful method is limited to species with complete or draft genome. This is because the availability of the reference genome is indispensible from the mapping-by-sequencing procedure, therefore eliminating the majority of plant species. Many genome sequence assembly projects are currently in progress including major crops. However, this may take time to get to a complete mature genome. It is important to have a mature genome assembly, instead of unordered scaffolds, as genome assembly is used as genetic maps during mapping-by-sequencing, in addition to as a target for short read alignment. Even in case of the finished genome assemblies, resequencing different accessions cofound higher level of translocation which will hinder the reconstruction of mutant genome based on the alignment of reads to reference genome (Long *et al.*, 2013). As the mutation identification starts with aligning short reads to reference genome, regions having copy number variations, translocation or loci having higher local divergences may experience difficulty in mutant identification. Therefore an unbiased method to compare genomes without utilizing the prior information of reference genome will improve mapping-by-sequencing. This is further important if the trait under selection is known to be rapidly evolving, such as resistance, therefore the reference genome from one accession may not necessarily represent another accession (Cai *et al.*, 1997; Song *et al.*, 2003). The local deletions in reference genome may hinder identification of putative causal region. Nevertheless this could be solved by creating local assembly around the mapping interval (Takagi *et al.*, 2013)

Utilization of reference sequence from close relative species gave an alternative approach for mutant mapping. However, this approach utterly relies on the homology between two genomes and additionally inherits all above-mentioned drawbacks. Recently, methods for identification of SNPs by direct comparison of genomes were introduced, but none has proved to be accurate enough for identification of mutagen induced mutations (Ratan *et al.*, 2010; Iqbal *et al.*, 2012). Challenge in characterization of mutants without reference genome involves identification of homozygous mutations followed by annotation of the effect of

mutation in coding sequence. We introduced a method called Needle in the k-stack (NIKS) to compare isogenic genomes. This reference-free method utilizes the occurrence of each substring of size k (hereafter referred to as k-mers) within whole genome sequencing data of one sample and identifies the homozygous mutations by comparing sample-specific k-mers between samples. A homozygous mutation can be identified by using sample specific k-mer from one sample and the similar k-mer in the second sample with mismatch representing mutation. Multiple levels of data reduction were done before searching for sample specific k-mer. The initial step is to filter out k-mers having sequencing error. Genome sequencing with a decent coverage expects to produce a Gaussian distribution while plotting the occurrences of sufficiently large k-mers vs. frequency of k-mers, with a peak representing the average k-mer coverage (Pevzner *et al.*, 2001; Kelley *et al.*, 2010). However, a sequencing error converts the frequent k-mer to a k-mer with low representation in the genome. These k-mers will disturb Gaussian distribution by having a peak at left (having few occurrences in the sequence) that can be detectable in case of sequencing with a decent coverage. Mutagen-induced mutations introduce sample specific k-mers in the genome but are represented in higher magnitude in the sequences. Therefore, comparing k-mers between two samples for unique k-mers per sample can identify homozygous mutagen induced mutations (Figure 4.1). NIKS identifies mutations and creates local assembly around mutation that provides usually multiple hundreds of bp in length.

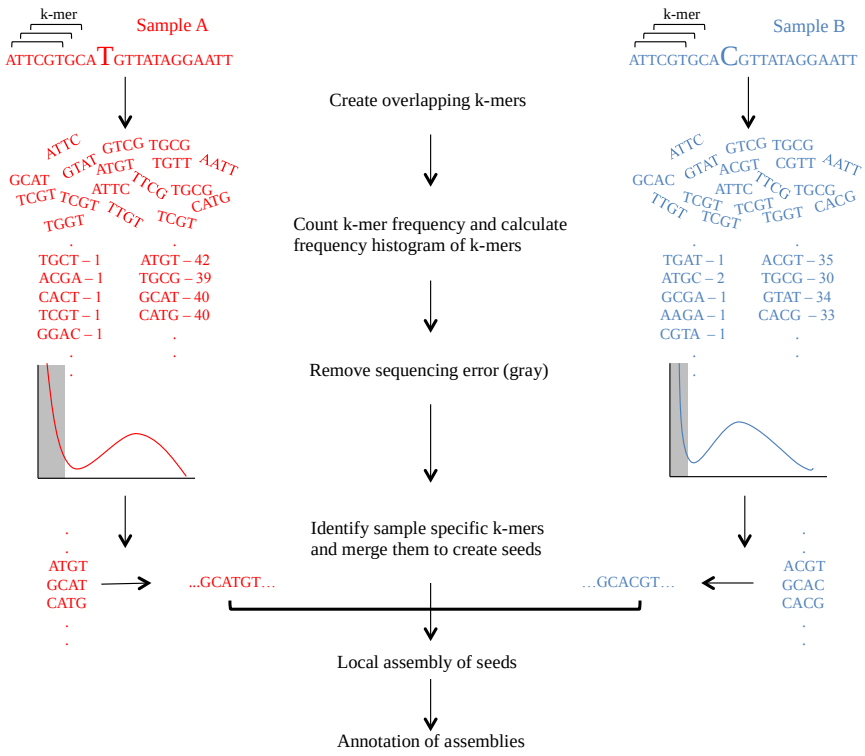

Figure 4.1: Workflow of NIKS. Two related genomes distinguished by mutagen-induced changes are sequenced. Raw reads from whole genome sequencing are analyzed for the frequency of k-mers separately. K-mers having sequencing error tent to have lower frequency, therefore k-mers with low frequency can be removed (shown in gray background). Comparing k-mers from two samples identifies sample-specific k-mers that harbor mutations. Sample-specific k-mers are merged to longer sequences called seeds. In case of small-scale differences, each seed may have a counterpart in other genome with subtle difference. Seeds are extended by local assembly using reads that share at least one k-mer with one of the seeds. These local assemblies containing the mutagen-induced mutations are used for gene prediction.

Once the putative mutations are identified, classification and prioritization of candidates could only be possible by annotating the putative effect of mutation. As the genome sequences and thus gene annotations are not available, this can be done either by homology-based annotation or by *ab initio*. Homology based annotation requires the availability of close relative homologous sequences. The success rate of homology-based annotation is essentially depending on the percentage of homology and differs between difference loci in the genome. Nonetheless, as the objective in forward genetic screen is to identify causal non-synonymous mutation, thus in gene

space, the likelihood of greater conservation is expected around the mutation site. On the other hand, *ab initio* annotation is independent and hardly requires any prior information, but suffer from low specificity and sensitivity (Reese and Guigó, 2006; Yandell and Ence, 2012). Moreover based on genome construction, different *ab initio* tools have been known to perform differently (Coghlan *et al.*, 2008; Yandell and Ence, 2012). Therefore, we used four *ab initio* prediction tools such as AUGUSTUS, GENEID, GENSCAN, FGENESH to compare the *ab initio* prediction tools in order to select the best prediction tool for *A. alpina* (Guigó *et al.*, 1992; Burge and Karlin, 1997; Salamov and Solovyev, 2000; Stanke and Waack, 2003; Blanco *et al.*, 2007). We made a set of 745 highly conserved genes using Program to Assemble Spliced Alignments (PASA), which served as a benchmark to access the performance of each *ab initio* prediction tool (Haas *et al.*, 2003). We sequenced two *A. alpina* mutants and identified causal mutation by NIKS. We applied NIKS as well as homology based mutant identification by using close relative *Arabidopsis* genome assembly to identify mutations. These mutations were annotated using homology as well as *ab initio* approach to identify putative candidates

4.2. Materials and Methods

4.2.1. *A. alpina* mutant sequencing

Two of the *A. alpina* mutants were selected from a recent EMS screen where one mutant, *perpetual flowering 1-1 (pep1-1),* was previously characterized through a homology based candidate gene approach to carry a splice-site mutation in the PEP1 gene that is responsible for the phenotype (Wang *et al.*, 2009). Whereas the second mutant *floral defective 1 (fde1)* displayed floral homeotic defects, in which underlying genetic cause was not known. We used both mutant genomes in this study and aimed to use *pep1-1* mutation to reconfirm the approach whereas *fde1* to identify the unknown causal mutation. The *fde1* mutant was backcrossed once to wild-type followed by selfing to generate segregating population of BC_1F_2, whereas *pep1-1* mutant was backcrossed twice to make BC_2F_2. Plants having mutant phenotype from *fde1* BC_1F_2 and *pep1-1* BC_2F_2 mutant families were selected and DNA from 86 and 97 mutant plants were pooled separately for sequencing. Both mutant pools were sequenced in a 2x100 bp paired-end sequencing using Illumina HiSeq2000. The *pep1-1* mutant pool was sequenced in one lane of Illumina instrument whereas the *fde1* was sequenced in two lanes. The sequencing generated data accounting for 51 to 158 times of *A. alpina* genome.

Table 4.1: Summary of sequencing data generated. Summary of Illumina HiSeq2000 sequencing for each mutant samples. The estimated genome coverage is based on assumption that *A. alpina* genome size is 375 Mb (adopted from Nordström *et al.*, 2013).

Mutant	Sample	Number of individual pooled	Number of read pair generated	Estimated genome coverage
pep1-1	BC_2F_2	97	125,167,649	~67.4
fde1	BC_1F_2	86	293,395,860	~158.0

4.2.2. Resequencing analysis of *A. alpina* mutants and SNP calling using mediator genome

All raw reads from each mutant samples were quality filtered and trimmed if the ends of the reads were of low quality. *Arabidopsis thaliana* Col-0 reference genome (TAIR10) was used as mediator genome for alignment of short reads and SNP calling. These reads were aligned to mediator genome using GenomeMapper in the short read analysis pipeline SHORE (Ossowski *et al.*, 2008; Schneeberger *et al.*, 2009). Relaxed criteria of 10 mismatches and 7 gaps were allowed for the aligning of 100bp long reads. Aligned reads were corrected for the expected insert size between read pairs. Using default heterozygous parameters and a minimum allele frequency of 20%, SHORE *consensus* was used to identify SNPs.

4.2.3. SNP calling using NIKS in *A. alpina* mutants

We applied NIKS pipeline to identify mutations in all two *A. alpina* mutant samples. NIKS does not require any prior information of reference genome and directly compares two samples using whole genome shotgun raw sequence. Therefore we used segregating population samples, *pep1-1* and *fde1* and compared them to identify mutations in each sample. Comparing the two genomes using NIKS, we aimed to identify the unknown lesion in *fde1* and simultaneously confirm the *pep1-1* mutation, which was characterized as PEP1 gene by homology based candidate gene approach (Wang *et al.*, 2009). NIKS pipeline starts by generating k-mers. We generated 31-mers using jellyfish and assessed the frequency of each k-mer within the raw reads (Marçais and Kingsford, 2011). This generated 17.7, 41.6 billion k-mers from *pep1-1, fde1* samples respectively. Multiple rounds of data reduction were made to identify mutation sites, which includes unique and sample specific k-mers in each sample set compared to counter sample set, was selected. This reduced the k-mer count to 3.4, 0.7 million k-mer from *pep1-1, fde1* samples respectively. Sequencing errors can produce unique k-mers in a sample. However, if the sample coverage in the sequencing is decent enough then these k-mer frequencies will be minimum and distinguishable from the expected Gaussian distribution. If errors are introduced during PCR, this could lead to illusive k-mers with sufficient frequency. Therefore it is advised to remove reads produced from single PCR template by filtering out reads with same starting sequences. Sample specific k-mers were merged with overlapping k-mers in order to increase the length, which in ideal case is k*2-1 and called as

seeds. The counterpart of seed (seed pair) from the counter genome was identified to call SNPs. In total we found 29 mutations in *pep1-1* and *fde1* mutant analysis and classified them to each group by taking advantage of EMS biased mutation spectrum. All 29 mutations were canonical EMS mutations that converted C->T (G->A), thus 13 and 16 mutant alleles were assigned to *pep1-1* and *fde1* samples, respectively.

4.2.4. Annotation of mutant SNPs

Two strategies were implemented to annotate the effect of mutations on protein coding. First, homology based annotation of seed sequences by using BLAST to identify the orthologous sequence followed by imputing the effect of mutation on protein sequence. Second, *ab initio* annotation of seed sequences for which no prior information is required. To identify the best tool for *ab initio* annotation in *A. alpina*, we used four annotation tools, namely AUGUSTUS 2.4, FGENESH 2, GENEID 1.3 and GENSCAN 1 and tested the sensitivity of each tool (Guigó *et al.*, 1992; Burge and Karlin, 1997; Salamov and Solovyev, 2000; Stanke and Waack, 2003; Blanco *et al.*, 2007). We generated a set of test genes using available cDNA sequence from *A. alpina*. We used cDNA as input for Program to Assemble Spliced Alignments (PASA) and followed *pasa_asmbls_to_training_set* pipeline for generating training set (Haas *et al.*, 2003). This training set was further filtered for complete genes and reduced the representation of protein from similar family by filtering out proteins having similarity of greater than or equal to 70%. CD-HIT was used to cluster the proteins to avoid the over representation of protein family in training set (Li and Godzik, 2006; Fu *et al.*, 2012). This reduced the initial set of 29661 genes to 12106 genes. In order to get high confident genes with less annotation mistakes, we used Blat to compare the protein sequence to *Arabidopsis* protein database and selected genes which had full coverage and >90% identity. This produced 745 genes in total, which were later used to identify the accuracy of *ab initio* prediction. The output from all four prediction tools was converted to Gene Transfer Format (GTF). With the help of Eval tool, we estimated the accuracy of each prediction tool (Keibler and Brent, 2003). This was calculated by the mean of sensitivity and specificity, and was estimated in three different levels, such as transcript, exon and nucleotide level.

4.3. Results

4.3.1. Comparison of NIKS and comparative genomic approach with mediator genome for *pep1-1* and *fde1* mutations

Both *A. thaliana* and *A. alpina* are from the same family Brassicaceae. Both genomes have different chromosome numbers of five and eight, which makes up to a genome size of 119 and 375 Mb, for *A. thaliana* and *A. alpina* respectively. Nonetheless, being the model plant organism for decades and the most well studied closest plant genome to *Arabis, A. thaliana* was selected as the mediator genome to identify SNP from *pep1-1* and *fde1* mutant genome. We used *A. thaliana* Col-0 reference genome (TAIR10) as a mediator genome for alignment of short reads from both mutant samples separately. Using very relaxed criteria for short read alignment such as 10% mismatch and 7% gaps of the total length of the read, it yet had poor alignment of raw reads. Only 8% raw reads were aligned to reference sequence, from which we identified 2,062,177 variations to mediator genome from *pep1-1* genome without any quality filtering. Similarly, *fde1* genome had 8% of raw reads aligned to mediator reference genome and identified 2191156 variations. Since we were using a mediator genome, it was expected that majority of identified mutations will be the difference between *A. thaliana* and *A. alpina* and will not be interesting for mutant identification. Therefore, these >90% of shared mutations between two mutants were filtered out along with mutation having low SHORE quality score of <24. As *pep1-1* mutation was previously identified as a splice-site lesion in PEP1 gene, we examined at homologous gene in *Arabidopsis*. Unfortunately this site was not covered with any short read alignment. On average 18% of *Arabidopsis* genome was covered with short mutant reads. Compared to mutations identified by NIKS, only two and four mutations were shared between NIKS and comparative approach from *pep1-1* and *fde1*, respectively. Scarcity of short read alignment at mutation loci caused the missing of rest of the mutations. In general, mediator genome approach was hampered by low homology. Though, in this particular case, comparative genomics was abortive, we checked whether majority of the short read aligned regions cover the mutations in coding region or not. Indeed 83% of covered mutations were in coding region. This anticipated output was encouraging as in mutant mapping; SNPs within coding regions are predominantly interesting as putative candidates.

Table 4.2: Fixed genomic differences between bulked F_2 individuals of *pep1-1* and *fde1*. Contigs and related information were derived from NIKS analysis. *ab initio* annotation of these contigs were done and the effect of mutations were predicted. Causal mutations are shown in bold letters. [a]Short read alignment coverage at each mutation locus in homology based alignment approach (adopted from Nordström et al., 2013).

Allele		Contig assoc. with mutation					*ab initio* annotation		Coverage[a]
pep1-1	f*de1*	Mutant genome	Length (bp)	Mutation position	Chr	Position	Gene	Effect	
T	C	*pep1-1*	549	437	2	5,482,633		none	NO
T	C	*pep1-1*	625	236	5	2,627,760	AT5G08160	syn (L>L)	NO
T	C	*pep1-1*	807	410	5	2,754,082	AT5G08510	intronic	NO
A	G	*pep1-1*	829	354	5	~2,998,250	AT5G09670	none	NO
A	**G**	***pep1-1***	**889**	**408**	**5**	**3,175,363**	**AT5G10140**	**splice-site change**	NO
T	C	*pep1-1*	812	451	5	~3,219,708		none	NO
T	C	*pep1-1*	653	220	5	3,333,724	AT5G10550	syn (R>R)	YES
A	G	*pep1-1*	783	437	5	3,336,193		none	NO
T	C	*pep1-1*	780	348	5	3,818,093	AT5G11850	nonsyn (G>D)	YES
A	G	*pep1-1*	882	445	5	10,116,108		nonsyn (F>S)	NO

A	G	*pep1-1*	732	368	5	17,725,725	AT5G44050	intronic	NO
A	G	*pep1-1*	772	341	-	-	-	none	
A	G	*pep1-1*	850	448	-	-	-	nonsyn (E>K)	
G	A	*fde1*	828	410	4	16,422,853	AT4G34320	nonsyn (Q>STOP)	YES
G	A	*fde1*	745	361	4	16,756,818	AT4G35230	intronic	YES
G	A	*fde1*	637	261	4	~17,051,245		none	NO
G	A	*fde1*	806	388	4	17,135,887		none	NO
C	T	*fde1*	819	388	4	17,178,292	AT4G36360	nonsyn (G>E)	YES
C	T	*fde1*	863	427	4	~17,286,500	AT4G36660	nonsyn (E>K)	NO
C	T	*fde1*	764	313	4	17,357,762		none	NO
C	**T**	***fde1***	**880**	**454**	**4**	**17,401,794**	**AT4G36920**	**nonsyn (D>N)**	NO
G	A	*fde1*	798	385	4	17,460,182		none	NO
C	T	*fde1*	789	353	4	17,475,571	AT4G37080	nonsyn (A>T)	YES
G	A	*fde1*	863	429	4	~17,729,980		none	NO

4.3.2. Annotation of candidate mutations in *pep1-1* and *fde1*

We tested the *ab initio* gene prediction accuracy of AUGUSTUS 2.4, FGENESH 2, GENEID 1.3 and GENSCAN 1.0 in *A. alpina* sequence. We utilized transcriptome data from *A. alpina* in order to access the accuracy of different gene

prediction tools. In contemplation of generating a set of genes for bench marking, PASA training gene set prediction pipeline was used to generate gene structures for protein coding genes with a minimum of 100 amino acids (Haas *et al.*, 2003). Genes with either partial 5' or 3' end were filtered out. Gene prediction tool are known to have biased prediction accuracy rate towards certain family of proteins. Therefore, over representation of any protein family in benchmark gene set would bias the accuracy rate of prediction. Hence, we applied CD-HIT to remove proteins that were 70% or more similar to others within the group (Li and Godzik, 2006; Fu *et al.*, 2012). This reduced the number of benchmark gene set to 12106 genes. Remaining gene sets were aligned to *A. thaliana* and genes that were fully covered and had homology greater than 90%, were selected as highly conserved gene set for benchmarking.

We used ad-hoc scripts to convert outputs from all four-gene structural prediction tools to unique Gene Transfer Format (GTF). This helped to compare the predicted genes from different tools against the benchmark genes and calculate accuracy of each prediction tool at transcript, exon and nucleotide level (Figure 4.2).

Though at nucleotide level all four prediction tools showed high accuracy, both sensitivity and specificity decreased at exon and transcript level. At exon level AUGUSTUS and FGENESH outperformed other two gene prediction tools and produced an accuracy of ~90%. While considering the complete transcripts, the accuracy declined and remarkable differences were showed between prediction tools. In all three level of accuracy check, AUGUSTUS outperformed FGENESH, GENEID and GENSCAN. GENSCAN gave the least accuracy level among all four prediction tools in all three level of accuracy test.

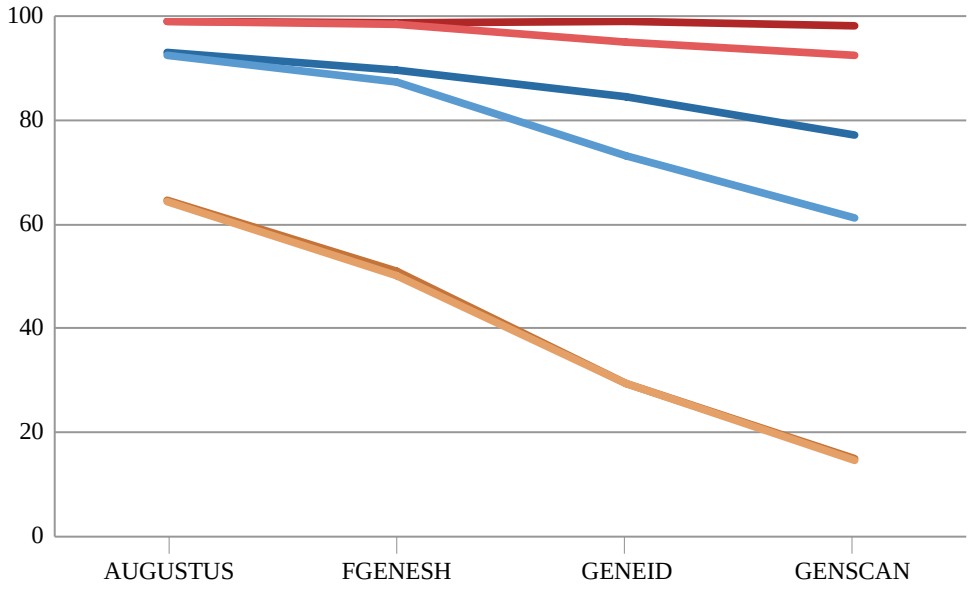

Figure 4.2: Accuracy of gene prediction tools. Sensitivity and specificity of four *ab initio* prediction tools based on the 745 benchmark genes. The sensitivity (dark color) and specificity (light color) were calculated at three different levels of prediction, such as nucleotide (red), exon (blue) and transcriptome (orange) .

AUGUSTUS was further used to annotate mutations identified by NIKS. Genes were predicted on contigs having mutations and in addition to full-length gene models AUGUSTUS predicted gene models that were partially present in the sequence. Mutations were introduced into the predicted gene models and annotated by their putative effect on the coding sequence (Table 4.2). Together with other putative candidates, AUGUSTUS predicted the known causal splice site mutation in *pep1-1* sample. Out of 24 mutations, nine mutations containing contigs did not produce any gene model. However, majority (89%) of these contigs appeared to be intergenic regions by homology search to *Arabidopsis*.

Chapter 5. Discussion

The revolution in DNA sequencing methods have changed the way genetic and genomic experiments been performed. NGS mediated new computational frameworks help to identify the causal mutation in forward genetics faster than ever before. Though this field is undergoing rapid changes, recent developments have shown more potential to be unfolded.

5.1. Mutant mapping using isogenic background

Chapter 2 illustrates one of the possibilities in the modification of crossing scheme that was not possible without the help of NGS. Conventional genetic mapping requires outcrossing to a diverged accession for the establishment of the mapping population. However, differences in phenotypes that segregate between *Arabidopsis* accessions are likely to mask subtle phenotypes that are caused by mutations. On the other hand, isogenic background has the advantage of eliminating possible artifacts in phenotypes caused by the introduction of new genomic background. As the identification of segregating markers as well as genotyping has been simultaneously done in mapping-by-sequencing, mapping has become possible even in an isogenic background. Cases like suppressor or enhancer screens of a previously identified mutant line, it is critical to keep the genome intact in order to avoid additional possible steps of genotyping. We demonstrated this approach in Chapter 2 by doing a suppressor screen of the *lhp1* mutant. The *lhp1* mutant phenotype differs quantitatively between accessions such as Col-0 and Wassilewskija-2, making it difficult to create a robust outcross mapping population for subtle modifiers. Therefore, we backcrossed *alp1;lhp1* double mutant plant to single mutant parent *lhp1*, generating an isogenic mapping population. Consequently, conventional markers were absent in the population and cannot be used to distinguish between parental alleles. We performed whole-genome sequencing and identified mutagen-induced changes by selecting mutant specific markers that were absent in parental genome. This way mutant mapping was done in an isogenic background that was not possible without sequencing. It is important to notice that as the number of segregating markers were comparable to the mutation rate of mutagen, which typically is one change in 112 to 171 kb in case of EMS, the number of segregants

required to resolve the linkage disequilibrium was rather low (Jander *et al.*, 2003; Ashelford *et al.*, 2011). This was further acknowledged by the simulation study in Chapter 3 in which comparable results could be obtained by using less segregants from backcross population as compared to outcross population.

Differences between reference sequence and the sequenced accession were identified as mutations. These mutations were referred as background mutations. Mutagen induced mutations were identified by filtering out background mutations from mutations identified in mutant line. Depending on the experimental setup, background mutations were defined in different ways. Conservative approach is to sequence the non-mutagenized progenitor in order to identify background mutations. However, this comes with the cost of sequencing progenitor. If two or more mutants are available from different mutagenic events on same progenitor, then mutants may be used reciprocally as background mutations. The assumption here is that, the probability of having similar mutation at the same locus of genome in two independent mutagenic events, is low. Both cases have the cavity of having non--sequenced regions in background genome leading to partial filtering of background mutations, thus producing false positive mutations. This can be avoided by considering mutations, when same locus has sufficient non-mutagenized allele in background genome, thus ensuring that the genome is being sequenced at this locus. However, this may lead to false negative markers. Therefore, these strategies need to be fine-tuned depending on the sequencing coverage and the expected number of markers in the genome. As false markers indulge in the mapping interval identification, it is advised to use a strict background mutation filtering to identify the mapping interval and then revisit the marker definition with more relaxed criteria (Galvão *et al.*, 2012). We extended the SHOREmap tool by integrating backcross analysis pipeline (http://shoremap.org). Appendix note II illustrates detailed option list of SHOREmap backcross .

However, whole-genome resequencing of pooled DNA from bulked segregant, usually results in a list of linked candidate changes. Mutations that are physically closer to causal mutation are only influenced by a minor number of recombination. And the typical coverage of whole-genome resequencing is incompatible to distinguish between homozygous and nearly homozygous changes. As closely linked candidate mutations may only have few recombination in the pool, non-causative mutations can be excluded by quantitative detection of rare wild-type

alleles. This is achievable by dCARE, a method that facilitates deep but targeted sequencing, thus reflects the true allele frequency in bulked DNA. Different NGS platforms have varying throughput, read length and cost per base and dCARE showcases how the power of different NGS platforms could benefit to the different stages of the mapping process. dCARE utilizes comparatively low throughput sequencing platform, Ion torrent, but is suitable for targeted resequencing. Often confounded but still improvable problem is, how to sequence larger genomic parts having multiple candidate loci spinning over more than few Mb. This becomes more important when these methods are transferred to crops and cereals with higher genome sizes and often tend to have larger mapping interval. Though most platforms provide method either by hybridization or capturing, for targeted sequencing, higher cost and lack of custom made arrays prevent the utility.

5.2. Simulating virtual genomes and mapping-by-sequencing: Tool and lessons learned.

We implemented two simulation programs, Pop simulator and Seq simulator collectively known as Pop-Seq simulator. Pop-Seq simulator simulates simplified virtual genotype and marker frequency by NGS genotyping. Pop-Seq simulator is implemented in Perl and follows Object Oriented Programming (OOP). Internally, Pop simulator starts with defining initial stage of homozygous parents at user defined marker positions. The parameters defined in the current version of our simulator configuration file are chromosome number, their respective sizes and ploidy level. In addition to that, the configuration file also contains a recombination landscape, probabilities for number of recombination and parameters for a gamma distribution to simulate crossover interference. By modifying respective values in configuration file, this tool can be applied to different species. However, if empirical data are not available for the species under consideration, then rather simplified simulation is also possible by defining equal recombination probability throughout the genome. Current implementation of the Pop simulator can handle variable number of chromosomes but limited to a ploidy level of two. The user is empowered to design crossing scheme by combining common crossing activities such as selfing, outcrossing or backcrossing, and is able to select dominant or recessive marker position in the genome to progress for next generation. Current version of Pop simulator can even handle crossing scheme with four founder parents, enabling to simulate AMPRIL lines (Huang *et al.*,

2011).

Whereas in Seq simulator, true allele frequency at each marker position is calculated from the virtual genotypes generated by Pop simulator. User is able to define the resequencing accessibility at each marker positions. Based on this expected local coverage, total reads are arbitrarily distributed on the estimated true allele frequency to generate reads per allele as output. By using Pop-Seq simulator, Chapter 3 formulates the optimal experimental design and the opportunities to be explored in mapping-by-sequencing experiments in *Arabidopsis*. Other than the options of different crossing to create segregation populations, direct sequencing of individual mutant genome is also possible given that allelic group is available for simultaneous analysis. Mapping-by-sequencing experiments have different layers of decisive steps. Possibilities in mapping-by-sequencing are primarily dependent on the starting biological material, available genomic resources; mainly reference genome sequence and the sequenced genomic material such as DNA or transcriptome. Chapter 3 primarily focuses on the crossing scheme of mapping population and the effect brought by this on whole experiment, particularly on the pool size and the required depth of sequencing. Compared to outcross populations, backcross populations require higher coverage for optimal mapping results. This is predominantly due the difference in genetic composition of both populations. *Arabidopsis* outcross population typically contains hundreds of thousands of natural variations, which are much denser than the expected recombination frequency. Thus, sliding-window-like approaches can combine the information from neighboring markers, and establish precise allele frequency in the pooled DNA. Whereas, backcross population consists only of mutagen induced mutations that are typically in the magnitude of hundreds across the genome, thus reduces the power of statistics or even treated markers independently in backcross analysis. Thus backcross population demands higher coverage but require low number of segregants pooled compared to an outcross population (Table 5.1).

As an alternative to bulk segregant analysis, we also analyzed direct sequencing of individual genomes of backcross populations. Each successive backcross reduces the foreground genome and the number of putative candidates. However, it requires multiple backcross generations before the number of putative candidates is as low as in bulk segregant analyses. Our study suggests that multiple rounds of backcrosses can be avoided by pooling multiple genomes. The genome-

wide mutation rate of radiation mutants is reported to be significantly lower as compared to chemically induced mutants (Belfield *et al.*, 2012). Direct sequencing of mutants with fewer, but putatively more severe mutations can simplify the interpretation of whole-genome analysis of directly sequenced mutant genomes.

Table 5.1: Suggestions for the design of mapping-by-sequencing experiments. Suggestions for the experimental set-up in different crossing scenario summarized from simulation study (adopted from Velikkakam James *et al.*, 2013).

	Outcross populations	Backcross populations	Direct sequencing	Deep candidate resequencing (dCARE)
Generation	F_2	BC_1F_2	$BC_{1\text{-}3}F_2$	n/a
Number of mutants	~150	~50	1	as many as possible
Optimal coverage	>25	~50	>25	n/a
Sequencing type	Paired-end	Paired-end	Paired-end	Single-end

As the mis-scored plants can have severe effects on mapping result, clarity of phenotype in a segregation population is very important and complex phenotypes may benefit from backcrossed mapping populations as the genetic background stays isogenic. From the simulation study in *Arabidopsis*, we came to the conclusion that having lower number of segregant is beneficial compared to accommodation of wrong segregant in the pool.

Paired end sequencing is beneficial in accessing the boards of repeat rich regions, thus may increase the number of markers been analyzed. Though, less repeat rich genomes like *Arabidopsis* may have low influence, crop genomes, known for their repeat content, may provide access to higher marker numbers with such reads. However, in crops other than the repeat content, genome size itself is a challenge. Though sequencing bigger genome is feasible, resequencing mutant lines for

identifying mapping region may need other strategies that are cost effective (Mayer *et al.*, 2012; Brenchley *et al.*, 2012). Chapter 3 explores one of the similar strategies by simulating target enriched sequencing in Barley. Though the chance of missing the causal mutation from the targeted sequence data is high, such approaches will lead and help fine-mapping efforts.

5.3. Mapping-by-sequencing in Crops

Advances in technologies have increased the easiness and made it more feasible than ever before to perform functional genetics studies(Kakioka *et al.*, 2013). Homology based approaches such as ordering incomplete reference genome (scaffolds) based on synteny or even using the closely related species' reference genome to align short reads, can be rewarding. As the number of sequenced genome or the transcriptome assembly is increasing, utilization of these incomplete but useful information in mapping is advantageous. For example, significant macro-collinearity between grass genomes encourage the synteny based mapping in these genomes (Pfeifer *et al.*, 2012). The probability of success increases when the mutant genome and the reference genome assembly are closely related and from the same genus (Wurtzel *et al.*, 2010). As the coding regions that are arguably conserved between genus and the main focus of forward genetic screening is to identify non-synonymous mutations, the amount of mutations undetected due to the lack of homology, should be minimum. However, this approach still needs to sequence the whole genome, moreover, even within the same species, structural variations or even absence of genomic regions in reference genome may cause difficulties in identifying causal mutation. Local assembly around the candidate mapping interval could resolve this, and subsequently could identify the causal mutation (Takagi, Uemura, *et al.*, 2013). Such approaches will certainly help the utilization of incomplete reference genome in mapping-by-sequencing.

Chapter 4 introduces a new computational framework called NIKS. NIKS enables comparison of isogenic genomes directly without the help of reference genome to identify homologous mutagen induced changes. The ability of NIKS to identify more mutations than a comparative approach as well as successful *ab initio* annotation of mutation for functional characterization indicates the power of NIKS for mapping in non-model organisms. In general *ab initio* prediction was successful in

annotating candidate mutations. However, an accuracy benchmarking is advised, if the species under study is getting annotated for the first time. Along with other sequencing data, these sequences generated for mutant identification can be utilized further for creating genomic resources like partial reference sequences and marker discovery. Sequencing backcrossed mutant as well as parental genome or independent multiple alleles of the phenotype will help in removing background mutations. The major advantage of NIKS is to enable mutant characterization without any prior knowledge of genetic map, reference sequences and even without segregation population. NIKS led NGS empowered mapping to non-model organisms and resolved one of the major hurdles. The remaining major obstacle in the application of mapping-by-sequencing to crops is the genome size that substantially increases the cost of experiment.

Recently, transcriptome sequencing became an obvious choice for the development of markers in species with larger genome size (Bancroft *et al.*, 2011; Barbazuk and Schnable, 2011; Dutta *et al.*, 2011; Edwards *et al.*, 2011; Margam *et al.*, 2011; Zhou *et al.*, 2012). As major portion of such genomes is non-coding, transcriptome sequencing helps in reducing the area under probing. The advantage of probing coding regions makes this approach even more suitable for mapping-by-sequencing experiments. Apart from the effective reduction in the genome representation, transcriptome mediated mapping experiments have several advantages: First, the reduced representation of genome directly reduces the cost of experiment. This has major impact when the genome has size is in few Gb and only minority of the genome is coding. Second, effect of mutation on transcript splicing can be directly assessed. A direct identification of mutation affecting splicing can be identified from the data on both annotated and un-annotated transcripts. Finally, comparing mutant and wild-type transcriptome can identify the alteration in expression level due to regulatory mutations. Altogether, these attributes make RNA-seq enabled mapping-by-sequencing, an efficient and cost effective means in larger genome mapping. Three different studies have applied this approach to map genes from Maize and Zebra fish (Liu *et al.*, 2012; Hill *et al.*, 2013; Miller *et al.*, 2013). However, variable expression level of genes across genome makes RNA-seq data noisy, thus demanding more statistical driven approach in analysis (Hill *et al.*, 2013).

Though RNA-seq offers alternative method for performing mapping-by-sequencing, this method has several cavities as well; First, low or no expression of

candidate transcript may or may not be due to causal mutation, obstructs the identification of casual mutations. If lack of expression is genuinely due to time or is tissue specific, having sampling done at various time points and on various tissues could solve this problem. This will explode the magnitude of samples to be sequenced if pre-knowledge or educated guess is not possible. At least for some phenotypes, it is possible to reduce the time between the emergence of the mutant phenotype and the extraction of RNA. If lack of expression is due to nonsense mutation, leading to a nonsense mediated decay (NMD), then mapping interval can still be identified without knowing causal mutation (Chang *et al.*, 2007; Miller *et al.*, 2013). The identification of casual mutation can be done by subsequent scrutiny of the mapping interval using suitable methods such as targeted sequencing or chromosome walking (Liu *et al.*, 2012; Trick *et al.*, 2012). Secondly, the possibility of utilizing the expression level in order to find the causal mutation comes with the demand to have multiple replicates in order to have a significant conclusion. It is also critical to have parental samples (control in this case) extracted in similar manner as mutations. By providing a direct comparison of expression levels, RNA-seq empowers the identification of the effect of noncoding regulatory mutations, but not the mutation itself. Third, mutations that influence the regulation of allele-specific expression may generate false positive SNPs (Main *et al.*, 2009; Pastinen, 2010). Finally, in RNA-seq, increasing coverage does not propositionally increase the coverage of low expressed genes. This can be as severe as 50% of the reads derived from 1% of genes (Trick *et al.*, 2012). Normalized RNA-seq would be an alternative that cleaves the highly abundant transcripts from the sample but comes with the cost of lack of expression level (Christodoulou *et al.*, 2011). Though this has been applied in marker discovery project in new species, this has not been tried yet in mapping-by-sequencing context.

5.4. Further challenges in mapping-by-sequencing

During last century, identification of a wide range of mutagen-induced phenotypes founded the basis of genetic research in *Arabidopsis* (Page and Grossniklaus, 2002). Different strategies are adopted in mapping-by-sequencing; indicating the availability of more than one optimum way in mapping. Figure 5.1 summaries the possibilities and proven strategies in mapping-by-sequencing (Figure 5.1). In future, given the available genomic resource and the specific limitations of the species under study, one has to decide optimum strategy on a case-by-case basis.

Three major stakeholders in this strategic planning are plant material, genomic resource used for study and computational method used in analysis.

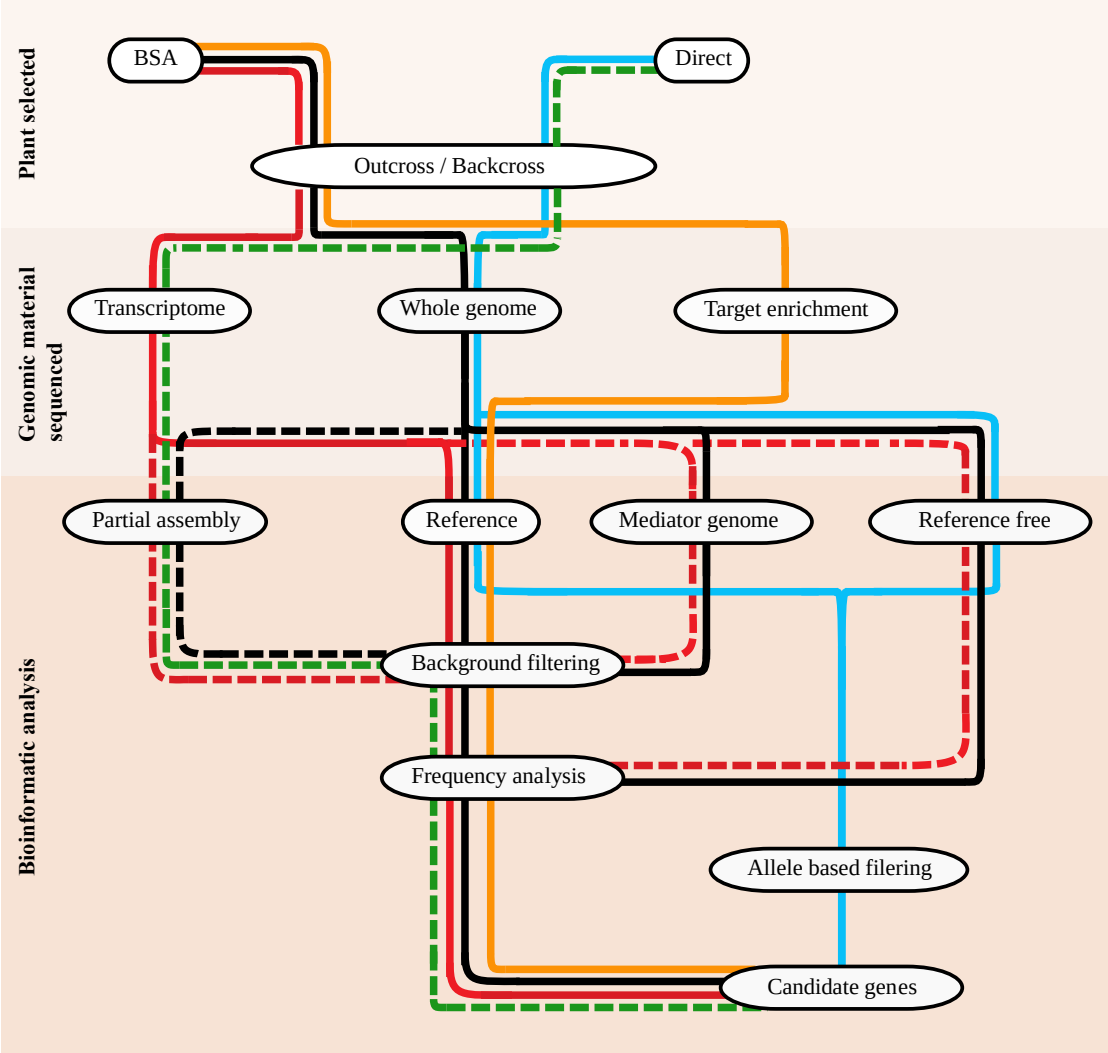

Figure 5.1: Roadmap of different strategies in mapping-by-sequencing. Three levels of options, such as mutant selected, genomic material sequenced and bioinformatics analysis are shown. Each colored continuous lines indicate proven strategies, whereas dotted lines indicate strategies yet to be established.

However, beyond the recessive phenotype, NGS enabled mapping could enable for traits, which are dominant and quantitative. In *Arabidopsis*, dominant mutant phenotypes are less common than recessive phenotypes (Meinke, 2013). At least in some cases, this is mainly due to the lethality. Unlike recessive mutant alleles, whose presence can be masked by the presence of a functional wild-type allele, dominant mutant alleles can be found in both heterozygotes and homozygotes state.

Currently, the suggested solution is to sequence mutant and non-mutant pools from BC_2F_1, and expects a region at 0.5 allele frequency versus the background frequency of 0.25 (Lindner *et al.*, 2012). However, the difference between expected and random frequency is subtle and needs extra information to identify the causal fixation. An alternative approach for dominant mutant mapping is to utilize the Mendelian segregation ratio to identify homozygous mutants in $F_{2\text{-}3}$ population. If F_3 family (each family is derived from selfing single F_2) is fixed for mutant phenotype, indicates homozygous mutant F_2 progenitor. Thus, by pooling mutant and wild-type phenotype plants separately produce homozygous allele in respective pools. Therefore, allele frequency analysis has higher leverage difference between two pools. This method has yet to be applied in dominant mutant mapping-by-sequencing.

Speeding-up in genetic mapping now opens new avenues even for more complex phenotype. As NGS enabled mapping of alleles that are naturally present in population, and which quantitatively contributes to complex phenotypes, will be of great interest. Such alleles can be identified by genome wide association studies (GWAS). GWAS utilize natural populations and NGS based genotyping that provides simultaneous marker discovery and genotyping (Atwell *et al.*, 2010; Li *et al.*, 2010; Witte, 2010). Alternatively, mapping of natural allele can be done by creating mapping population from distinct parents and pooling plants with extreme phenotype. In this case, genomic loci contributing to phenotype will show difference in allele frequency between two extreme pools. This principal was initially adopted in yeast to map major QTLs (Ehrenreich *et al.*, 2010). Later, NGS enabled QTL mapping was done in species with even higher genome size, such as rice and Drosophila (Turner *et al.*, 2011; Takagi, Abe, *et al.*, 2013). Though mapping was successfully done in major QTLs, improvement in the algorithm to reduce the noise from sequencing, in order to identify the minor QTLs is still needed. Current simplified approach of subtraction of allele frequency between extreme pools obstructs the identification of QTLs present in close physical vicinity. As in Claesen et al., further improvement of resolution in QTL peak detection methods, either powered by statistics or by modified crossing scheme or even both, is much needed (Claesen *et al.*, 2013).

Currently, mapping studies end with functional annotation in gene space. However, this can be extended with other ‘omics’ data to get a unified global functional interpretation. With advance in genomics, it is possible to study the molecular phenotypes such as transcription/translational rate, chromatin accessibility

and methylation rate, just to name a few (Boyle *et al.*, 2008; Cokus *et al.*, 2008; Ingolia, 2010; Churchman and Weissman, 2011). Future direction of studies must examine different layers of evidence to provide important links between genomic information and organismic functions, in order to postulate major mechanisms, if not complete, of complex traits.

Literatures cited:

Abe, A., Kosugi, S., Yoshida, Kentaro, *et al.* (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat Biotech.*, **30,** 174-178

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C. and Gnirke, A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18..

Alonso, J.M. and Ecker, J.R. (2006) Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in Arabidopsis. *Nat. Rev. Genet.*, **7**, 524–536.

Ashelford, K., Eriksson, M.E., Allen, C.M., *et al.* (2011) Full genome re-sequencing reveals a novel circadian clock mutation in Arabidopsis. *Genome Biol.*, **12**, R28.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., *et al.* (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, **465**, 627–631.

Austin, R.S., Vidaurre, D., Stamatiou, G., *et al.* (2011) Next-generation mapping of Arabidopsis genes. *Plant J. Cell Mol. Biol.*, **67**, 715–725.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. and Johnson, E.A. (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers J. C. Fay, ed. *PLoS ONE*, **3**, e3376.

Bancroft, I., Morgan, C., Fraser, F., *et al.* (2011) Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. *Nat Biotech*, **29**, 762–766.

Barbazuk, W.B. and Schnable, P.S. (2011) SNP discovery by transcriptome pyrosequencing. *Methods Mol. Biol. Clifton NJ*, **729**, 225–246.

Belfield, E., Gan, X., Mithani, A., *et al.* (2012) Genome-wide analysis of mutations in mutant lineages selected following fast-neutron irradiation mutagenesis of Arabidopsis thaliana. *Genome Res,* **22,** 1306-1315.

Birkeland, S.R., Jin, N., Ozdemir, A.C., Lyons, R.H., Weisman, L.S. and Wilson, T.E. (2010) Discovery of Mutations in Saccharomyces cerevisiae by Pooled Linkage Analysis and Whole-Genome Sequencing. *Genetics*, **186**, 1127–1137.

Blanca, J.M., Pascual, L., Ziarsolo, P., Nuez, F. and Cañizares, J. (2011) ngs_backbone: a pipeline for read cleaning, mapping and SNP calling using Next Generation Sequence. *BMC Genomics*, **12**, 285.

Blanco, E., Parra, G. and Guigó, R. (2007) Using geneid to identify genes. *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al*, **Chapter 4**, Unit 4.3.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. and Crawford, G.E. (2008) High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*, **132**, 311–322.

Brenchley, R., Spannagl, M., Pfeifer, M., *et al.* (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, **268**, 78–94.

Cahill, M.J., Köser, C.U., Ross, N.E. and Archer, J.A.C. (2010) Read length and repeat resolution: exploring prokaryote genomes using next-generation sequencing technologies. *PloS One*, **5**, e11518.

Cai, D., Kleine, M., Kifle, S., *et al.* (1997) Positional Cloning of a Gene for Nematode Resistance in Sugar Beet. *Science*, **275**, 832–834.

Chang, Y.-F., Imam, J.S. and Wilkinson, M.F. (2007) The nonsense-mediated decay RNA surveillance pathway. *Annu. Rev. Biochem.*, **76**, 51–74.

Christodoulou, D.C., Gorham, J.M., Herman, D.S. and Seidman, J.G. (2011) Construction of normalized RNA-seq libraries for next-generation sequencing

using the crab duplex-specific nuclease. *Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al*, **Chapter 4**, Unit4.12.

Churchman, L.S. and Weissman, J.S. (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, **469**, 368–373.

Cistué, L., Cuesta-Marcos, A., Chao, S., *et al.* (2011) Comparative mapping of the Oregon Wolfe Barley using doubled haploid lines derived from female and male gametes. *TAG Theor. Appl. Genet. Theor. Angew. Genet.*, **122**, 1399–1410.

Claesen, J., Clement, L., Shkedy, Z., Foulquié-Moreno, M.R. and Burzykowski, T. (2013) Simultaneous Mapping of Multiple Gene Loci with Pooled Segregants. *PLoS ONE*, **8**, e55133.

Coghlan, A., Fiedler, T.J., McKay, S.J., Flicek, P., Harris, T.W., Blasiar, D., $author.lastName, $author firstName and Stein, L.D. (2008) nGASP – the nematode genome annotation assessment project. *BMC Bioinformatics*, **9**, 549.

Cokus, S.J., Feng, S., Zhang, X., *et al.* (2008) Shotgun bisulfite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, **452**, 215–219.

Collard, B.C.. and Mackill, D.J. (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos. Trans. R. Soc. B Biol. Sci.*, **363**, 557 –572.

Collard, B.C.Y., Jahufer, M.Z.Z., Brouwer, J.B. and Pang, E.C.K. (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, **142**, 169–196.

Cuperus, J.T., Montgomery, T.A., Fahlgren, N., Burke, R.T., Townsend, T., Sullivan, C.M. and Carrington, J.C. (2010) Identification of MIR390a precursor processing-defective mutants in Arabidopsis by direct genome sequencing. *Proc. Natl. Acad. Sci.*, **107**, 466 –471.

DePristo, M.A., Banks, E., Poplin, R., *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.

Doitsidou, M., Poole, R.J., Sarin, S., Bigelow, H. and Hobert, O. (2010) C. elegans Mutant Identification with a One-Step Whole-Genome-Sequencing and SNP Mapping Strategy. *PLoS ONE*, **5**, e15435.

Dutta, S., Kumawat, G., Singh, B.P., *et al.* (2011) Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [Cajanus cajan (L.) Millspaugh]. *BMC Plant Biol.*, **11**, 17.

Edwards, C.E., Parchman, T.L. and Weekley, C.W. (2011) Assembly, Gene Annotation and Marker Development Using 454 Floral Transcriptome Sequences in Ziziphus Celata (Rhamnaceae), a Highly Endangered, Florida Endemic Plant. *DNA Res.* **19,** 1-9

Ehrenreich, I.M., Torabi, N., Jia, Y., Kent, J., Martis, S., Shapiro, J.A., Gresham, D., Caudy, A.A. and Kruglyak, L. (2010) Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature*, **464**, 1039–1042.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE*, **6**, e19379.

Ewing, B. and Green, P. (1998) Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Res.*, **8**, 186–194.

Falconer, D.S. and Mackay, T.F.C. (1996) *Introduction to quantitative genetics*, Essex, England: Longman.

Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinforma. Oxf. Engl.*, **28**, 3150–3152.

Galvão, V.C., Nordström, K.J.V., Lanz, C., Sulz, P., Mathieu, J., Posé, D., Schmid, M., Weigel, D. and Schneeberger, K. (2012) Synteny-based Mapping-by-Sequencing enabled by Targeted Enrichment. *Plant J. Cell Mol. Biol.*, **71**, 517-526

Giraut, L., Falque, M., Drouaud, J., Pereira, L., Martin, O.C. and Mézard, C. (2011) Genome-Wide Crossover Distribution in Arabidopsis thaliana Meiosis Reveals Sex-Specific Patterns along Chromosomes. *PLoS Genet*, **7**, e1002354.

Gnirke, A., Melnikov, A., Maguire, J., *et al.* (2009) Solution Hybrid Selection with Ultra-long Oligonucleotides for Massively Parallel Targeted Sequencing. *Nat. Biotechnol.*, **27**, 182–189.

Golas, T.M., Geest, H. van de, Gros, J., Sikkema, A., D'Agostino, N., Nap, J.P., Mariani, C., Allefs, J.J.H.M. and Rieu, I. (2013) Comparative next-generation mapping of the Phytophthora infestans resistance gene Rpi-dlc2 in a European accession of Solanum dulcamara. *Theor. Appl. Genet.*, **126**, 59–68.

Guigó, R., Knudsen, S., Drake, N. and Smith, T. (1992) Prediction of gene structure. *J. Mol. Biol.*, **226**, 141–157.

Haas, B.J., Delcher, A.L., Mount, S.M., *et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.*, **31**, 5654–5666.

Hartwig, B., Velikkakam James, G., Konrad, K., Schneeberger, K. and Turck, F. (2012) Fast Isogenic Mapping-by-Sequencing of Ethyl Methanesulfonate-Induced Mutant Bulks. *Plant Physiol.*, **160**, 591–600.

Harushima, Y., Yano, M., Shomura, A., *et al.* (1998) A High-Density Rice Genetic Linkage Map with 2275 Markers Using a Single F2 Population. *Genetics*, **148**, 479 –494.

Hill, J.T., Demarest, B.L., Bisgrove, B.W., Gorsi, B., Su, Y.-C. and Yost, H.J. (2013) MMAPPR: Mutation Mapping Analysis Pipeline for Pooled RNA-seq. *Genome Res.*, **23**, 687–697.

Huang, X., Feng, Q., Qian, Q., *et al.* (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res.*, **19**, 1068 –1076.

Huang, X., Paulo, M.-J., Boer, M., Effgen, S., Keizer, P., Koornneef, M. and Eeuwijk, F.A. van (2011) Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 4488–4493.

Ingolia, N.T. (2010) Chapter 6 - Genome-Wide Translational Profiling by Ribosome Footprinting. In Jonathan Weissman; Christine Guthrie and Gerald R. Fink, ed. *Methods in Enzymology*. Guide to Yeast Genetics: Functional Genomics, Proteomics, and Other Systems Analysis. Academic Press, pp. 119–142.

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

Jander, G., Baerson, S.R., Hudak, J.A., Gonzalez, K.A., Gruys, K.J. and Last, R.L. (2003) Ethylmethanesulfonate Saturation Mutagenesis in Arabidopsis to Determine Frequency of Herbicide Resistance. *Plant Physiol.*, **131**, 139 –146.

Jimenez-Gomez, J.M. (2011) Next Generation Quantitative Genetics in Plants. *Front. Plant Sci.*, **2**.77

Jünemann, S., Sedlazeck, F.J., Prior, K., *et al.* (2013) Updating benchtop sequencing performance comparison. *Nat. Biotechnol.*, **31**, 294–296.

Kakioka, R., Kokita, T., Kumada, H., Watanabe, K. and Okuda, N. (2013) A RAD-based linkage map and comparative genomics in the gudgeons (genus Gnathopogon, Cyprinidae). *BMC Genomics*, **14**, 32.

Kearsey, M.J. (1998) The principles of QTL analysis (a minimal mathematics approach). *J. Exp. Bot.*, **49**, 1619–1623.

Keibler, E. and Brent, M.R. (2003) Eval: A software package for analysis of genome annotations. *BMC Bioinformatics*, **4**, 50.

Kelley, D.R., Schatz, M.C. and Salzberg, S.L. (2010) Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.*, **11**, R116.

Klasen, J.R., Piepho, H.-P. and Stich, B. (2012) QTL detection power of multi-parental RIL populations in Arabidopsis thaliana. *Heredity*, **108**, 626–632.

Koehler, R., Issac, H., Cloonan, N. and Grimmond, S.M. (2011) The uniqueome: a mappability resource for short-tag sequencing. *Bioinformatics*, **27**, 272–274.

Koornneef, M. and Meinke, D. (2010) The development of Arabidopsis as a model plant. *Plant J.*, **61**, 909–921.

Kover, P.X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I.M., Purugganan, M.D., Durrant, C. and Mott, R. (2009) A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS Genet.*, **5**, e1000551.

Lam, H.Y.K., Pan, C., Clark, M.J., *et al.* (2012) Detecting and annotating genetic variations using the HugeSeq pipeline. *Nat. Biotechnol.*, **30**, 226–229.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Li, Y., Huang, Y., Bergelson, J., Nordborg, M. and Borevitz, J.O. (2010) Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. *Proc. Natl. Acad. Sci.*, **107**, 21199–21204.

Lindner, H., Raissig, M.T., Sailer, C., Shimosato-Asano, H., Bruggmann, R. and Grossniklaus, U. (2012) SNP-Ratio Mapping (SRM): Identifying Lethal Alleles and Mutations in Complex Genetic Backgrounds by Next-Generation Sequencing. *Genetics*, **191**, 1381–1386.

Liu, K., McCormack, M. and Sheen, J. (2012) Targeted parallel sequencing of large genetically-defined genomic regions for identifying mutations in Arabidopsis. *Plant Methods*, **8**, 12.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. and Law, M. (2012) Comparison of Next-Generation Sequencing Systems. *BioMed Res. Int.*, **2012**.

Liu, S., Yeh, C.-T., Tang, H.M., Nettleton, D. and Schnable, P.S. (2012) Gene Mapping via Bulked Segregant RNA-Seq (BSR-Seq). *PLoS ONE*, **7**, e36406.

Long, Q., Rabanal, F.A., Meng, D., *et al.* (2013) Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat. Genet.*, **45**, 884–890.

Lukowitz, W., Gillmor, C.S. and Scheible, W.-R. (2000) Positional Cloning in Arabidopsis. Why It Feels Good to Have a Genome Initiative Working for You. *Plant Physiol.*, **123**, 795 –806.

Main, B.J., Bickel, R.D., McIntyre, L.M., Graze, R.M., Calabrese, P.P. and Nuzhdin, S.V. (2009) Allele-specific expression assays using Solexa. *BMC Genomics*, **10**, 422.

Marçais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinforma. Oxf. Engl.*, **27**, 764–770.

Margam, V.M., Coates, B.S., Bayles, D.O., *et al.* (2011) Transcriptome Sequencing, and Rapid Development and Application of SNP Markers for the Legume Pod Borer Maruca vitrata (Lepidoptera: Crambidae). *PLoS ONE*, **6**, e21388.

Mauricio, R. (2001) Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. *Nat. Rev. Genet.*, **2**, 370–381.

Mayer, K.F.X., Waugh, R., Langridge, P., *et al.* (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*. **491**:711-6

McNally, K.L., Childs, K.L., Bohnert, R., *et al.* (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc. Natl. Acad. Sci.* **106**:12273-8.

Meinke, D.W. (2013) A survey of dominant mutations in Arabidopsis thaliana. *Trends Plant Sci.*, **18**, 84–91.

Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D. and Koornneef, M. (1998) Arabidopsis thaliana: A Model Plant for Genome Analysis. *Science*, **282**, 662–682.

Michelmore, R.W., Paran, I. and Kesseli, R.V. (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. U. S. A.*, **88**, 9828–9832.

Miller, A.C., Obholzer, N.D., Shah, A.N., Megason, S.G. and Moens, C.B. (2013) RNA-seq–based mapping and candidate identification of mutations from forward genetic screens. *Genome Res.*, **23**, 679–686.

Minevich, G., Park, D.S., Blankenberg, D., Poole, R.J. and Hobert, O. (2012) CloudMap: A Cloud-Based Pipeline for Analysis of Mutant Genome Sequences. *Genetics*, **192**, 1249–1269.

Nakamura, K., Oshima, T., Morimoto, T., *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.*, **39**, e90.

Nielsen, R., Paul, J.S., Albrechtsen, A. and Song, Y.S. (2011) Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*, **12**, 443–451.

Nordborg, M. and Weigel, D. (2008) Next-generation genetics in plants. *Nature*, **456**, 720–723.

Nordström, K.J.V., Albani, M.C., Velikkakam James, G., Gutjahr, C., Hartwig, B., Turck, F., Paszkowski, U., Coupland, G. and Schneeberger, K. (2013) Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. *Nat. Biotechnol.*, **31**, 325–330.

Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N. and Weigel, D. (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res.*, **18**, 2024 –2033.

Pabinger, S., Dander, A., Fischer, M., *et al.* (2013) A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.*

Page, D.R. and Grossniklaus, U. (2002) The art and design of genetic screens: Arabidopsis thaliana. *Nat. Rev. Genet.*, **3**, 124–136.

Pastinen, T. (2010) Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.*, **11**, 533–538.

Paterson, A.H., Damon, S., Hewitt, J.D., Zamir, D., Rabinowitch, H.D., Lincoln, S.E., Lander, E.S. and Tanksley, S.D. (1991) Mendelian Factors Underlying Quantitative Traits in Tomato: Comparison across Species, Generations, and Environments. *Genetics*, **127**, 181–197.

Pevzner, P.A., Tang, H. and Waterman, M.S. (2001) An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci.*, **98**, 9748–9753.

Pfeifer, M., Martis, M., Asp, T., Mayer, K.F.X., Lübberstedt, T., Byrne, S., Frei, U. and Studer, B. (2012) The Perennial Ryegrass GenomeZipper – Targeted Use of Genome Resources for Comparative Grass Genomics. *Plant Physiol.*, **161**:571-82

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. and Gu, Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.

Rafalski, J.A. (2010) Association genetics in crop improvement. *Curr. Opin. Plant Biol.*, **13**, 174–180.

Ratan, A., Zhang, Y., Hayes, V., Schuster, S. and Miller, W. (2010) Calling SNPs without a reference sequence. *BMC Bioinformatics*, **11**, 130.

Reese, M.G. and Guigó, R. (2006) EGASP: Introduction. *Genome Biol.*, **7**, S1.

Roberts, R.J., Carneiro, M.O. and Schatz, M.C. (2013) The advantages of SMRT sequencing. *Genome Biol.*, **14**, 405.

Rothberg, J.M., Hinz, W., Rearick, T.M., *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–352.

Salamov, A.A. and Solovyev, V.V. (2000) Ab initio Gene Finding in Drosophila Genomic DNA. *Genome Res.*, **10**, 516–522.

Salome, P.A., Bomblies, K., Fitz, J., Laitinen, R.A.E., Warthmann, N., Yant, L. and Weigel, D. (2011) The recombination landscape in Arabidopsis thaliana F2 populations. *Heredity*. **108**:447-455

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, **74**, 5463–5467.

Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O. and Weigel, D. (2009) Simultaneous alignment of short reads against multiple genomes. *Genome Biol.*, **10**, R98.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jørgensen, J.-E., Weigel, D. and Andersen, S.U. (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods*, **6**, 550–551.

Schneeberger, K., Ossowski, S., Ott, F., *et al.* (2011) Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 10249–10254.

Schreiber, K.J., Austin, R.S., Gong, Y., Zhang, J., Fung, P., Wang, P.W., Guttman, D.S. and Desveaux, D. (2012) Forward chemical genetic screens in Arabidopsis identify genes that influence sensitivity to the phytotoxic compound sulfamethoxazole. *BMC Plant Biol.*, **12**, 226.

Sikora, P., Chawade, A., Larsson, M., Olsson, J. and Olsson, O. (2012) Mutagenesis as a Tool in Plant Genetics, Functional Genomics, and Breeding. *Int. J. Plant Genomics*.

Slate, J. (2008) Robustness of linkage maps in natural populations: a simulation study. *Proc. R. Soc. B Biol. Sci.*, **275**, 695–702.

Somerville, C. and Koornneef, M. (2002) A fortunate choice: the history of Arabidopsis as a model plant. *Nat. Rev. Genet.*, **3**, 883–889

Song, J., Bradeen, J.M., Naess, S.K., *et al.* (2003) Gene RB cloned from Solanum bulbocastanum confers broad spectrum resistance to potato late blight. *Proc. Natl. Acad. Sci.*, **100**, 9128–9133.

Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**, ii215–ii225

Tabata, R., Kamiya, T., Shigenobu, S., Yamaguchi, K., Yamada, M., Hasebe, M., Fujiwara, T. and Sawa, S. (2013) Identification of an EMS-induced causal mutation in a gene required for boron-mediated root development by low-coverage genome re-sequencing in Arabidopsis. *Plant Signal. Behav.*, **8**, 38–44.

Takagi, H., Abe, A., Yoshida, K., *et al.* (2013) QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J.*, **74**, 174–183.

Takagi, H., Uemura, A., Yaegashi, H., *et al.* (2013) MutMap-Gap: whole-genome resequencing of mutant F2 progeny bulk combined with de novo assembly of gap regions identifies the rice blast resistance gene *New Phytol.*, **200**:276-83

Tautz, D. and Domazet-Lošo, T. (2011) The evolutionary origin of orphan genes. *Nat Rev Genet*, **12**, 692–702.

Trick, M., Adamski, N., Mugford, S.G., Jiang, C.-C., Febrer, M. and Uauy, C. (2012) Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. *BMC Plant Biol.*, **12**, 14.

Turner, E.H., Lee, C., Ng, S.B., Nickerson, D.A. and Shendure, J. (2009) Massively parallel exon capture and library-free resequencing across 16 individuals. *Nat. Methods*, **6**, 315–316.

Turner, T.L., Stewart, A.D., Fields, A.T., Rice, W.R. and Tarone, A.M. (2011) Population-Based Resequencing of Experimentally Evolved Populations Reveals the Genetic Basis of Body Size Variation in Drosophila melanogaster. *PLoS Genet*, **7**, e1001336.

Uchida, N., Sakamoto, T., Kurata, T. and Tasaka, M. (2011) Identification of EMS-Induced Causal Mutations in a Non-Reference Arabidopsis thaliana Accession by Whole Genome Sequencing. *Plant Cell Physiol.*, **52**, 716 –722.

Varshney, R.K., Chen, W., Li, Y., *et al.* (2011) Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**:83-9.

Velikkakam James, G., Patel, V., Nordström, K.J., Klasen, J.R., Salomé, P.A., Weigel, D. and Schneeberger, K. (2013) User guide for mapping-by-sequencing in Arabidopsis. *Genome Biol.*, **14**, R61.

Wall, P.K., Leebens-Mack, J., Chanderbali, A.S., *et al.* (2009) Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics*, **10**, 347.

Wang, R., Farrona, S., Vincent, C., Joecker, A., Schoof, H., Turck, F., Alonso-Blanco, C., Coupland, G. and Albani, M.C. (2009) PEP1 regulates perennial flowering in Arabis alpina. *Nature*, **459**, 423–427.

Weigel, D. (2012) Natural Variation in Arabidopsis: From Molecular Genetics to Ecological Genomics. *Plant Physiol.*, **158**, 2–22.

Wetterstrand K.A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed 2 July 2013

Witte, J.S. (2010) Genome-Wide Association Studies and Beyond. *Annu. Rev. Public Health*, **31**, 9–20.

Wurtzel, O., Dori-Bachash, M., Pietrokovski, S., Jurkevitch, E. and Sorek, R. (2010) Mutation Detection with Next-Generation Resequencing through a Mediator Genome. *PLoS ONE*, **5**, e15628.

Yandell, M. and Ence, D. (2012) A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.*, **13**, 329–342.

Zhou, Y., Gao, F., Liu, R., Feng, J. and Li, H. (2012) De novo sequencing and analysis of root transcriptome using 454 pyrosequencing to discover putative genes associated with drought tolerance in Ammopiptanthus mongolicus. *BMC Genomics*, **13**, 266.

Zuryn, S., Gras, S. Le, Jamet, K. and Jarriault, S. (2010) A Strategy for Direct Mapping and Identification of Mutations by Whole-Genome Sequencing. *Genetics*, **186**, 427–430.

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

[http://www.illumina.com**]**

[http://www.pacificbiosciences.com]

[http://1001genomes.org]

Appendix Notes:

Note 1: Exemplary command line calls for resequencing analysis using SHORE and successively by SHOREmap

Task	Command
SHORE	
Preprocessing the reference sequence	shore preprocess -f TAIR10_chr_all.fa -i TAIR10 -W
Importing raw reads into SHORE	shore import -v Fastq -a genomic -i 1001 -x s_1_1.fq -y s_1_2.fq -o sampleA --rplot --disable-illumina-filter -k 75
Evoking short read alignments	shore mapflowcell -f sampleA -i TAIR10/TAIR10_chr_all.fas.shore -n 10% -g 7% -p
Correcting alignments for paired-end information	shore correct4pe -l sampleA/1 -x 300 -e 1 -p
Merge alignment files	shore merge -p sampleA -d merge
SNP calling	shore consensus -n sampleID -f TAIR10/TAIR10_chr_all.fas.shore -o consensus -i merge/map.list.gz -g 4 -a scoring_matrix_het.txt -v -r
SHOREmap	
Analyze frequencies of novel mutations	SHOREmap.pl backcross --marker mutant_sample/consensus/ConsensusAnalysis/quality_variant.txt --out SHOREmap_out --chrsizes TAIR10_chrsizes.txt --bg background_sample/consensus/ConsensusAnalysis/quality_variant.txt --marker-score 25 --marker-freq 0 –marker-cov 8 --bg-freq 20

Note 2: Detailed option list of SHOREmap backcross

Resequencing of pooled backcrossed segregant can be analyzed using *SHOREmap backcross* in order to identify the putative candidate mutations. SHOREmap *backcross* analysis filter out the natural variation between reference sequence and mutant accession and visualize the frequency of mutagen-induced mutations. The usage of *SHOREmap backcross* with minimum parameter as follows:

SHOREmap.pl backcross --chrsizes Chromsome.txt --out SHOREoutput --marker quality_varitent.txt

Mandatory:

--chrsizes	*File:* Tabbed file with chromosome name and size of each chromosome
--out	*Characters:* Output folder name.
--marker	*File:* Marker file. Output file from SHORE (quality_varitent.txt). If the list of marker is from different source then convert to SHORE marker format

Optional:

--marker-score	*Numeric:* Minimum SHORE score cutoff for filtering marker. Default is 25
--marker-freq	*Numeric:* Minimum concordances for filtering marker. Default is 80
--marker-cov	*Numeric:* Minimum read support for filtering marker.
--bg	*File:* File with background mutations. Usually this are the markers derived from the non-mutagenized sample. These mutations represent the natural variation between sequenced accession and reference sequence. If more than one file list them with comma-separation.
--bg-score	*Numeric:* Minimum SHORE score cutoff for filtering background markers
--bg-freq	*Numeric:* Minimum concordance for background markers. Default is 20
--bg-cov	*Numeric:* Minimum read support for background markers.

--bg-ref *File:* File with details of background reference allele calls. If this file is given, then to qualify as mutagen induced marker, a marker must not only be absent in --bg file/s but also the reference allele must be present in --bg-ref file.

Plotting options:

-no-summary *Flag:* Turn off plotting all chromosome in single page as summary

-no-filter *Flag:* Plot all markers after background correction. No marker score, frequency and coverage cutoff been used during plotting.

-non-EMS *Flag:* Plot non-canonical EMS (marked as "x") mutations also.

-other-mutagen *Flag:* No differentiation between EMS and Non-EMS markers

-verbose *Flag:* Verbose. Be talkative and report what is going on while analyzing.

Note 3: Detailed option list of Pop simulator and Seq simulator

Pop simulator simulates virtual genomes produced by following user specified crossing scheme and represented in genotype. Minimal option to run Pop simulator includes the requirement of population size, marker positions and crossing scheme. The usage is as follows:

perl simulate_F2_seq.pl -n 20 -f F2 -m Marker_file.txt

This command will simulate 20 F_2 segregants with genotype information at specified marker positions given by *Marker_file*. Elaborated parameter options are briefly explained below:

Mandatory:

-n / -a *Numeric*: Number of segregant or mutant plants, respectively. Either one of the option is mandatory.

-f *Characters*: Crossing scheme in "litral words" superated by ":" . For example: F5 for five times selfing and F2:B1:F1 for generating BC_1F_2 by crossing F_2 and recurrent parent to make BC_1F_1 followed by one round of selfing. Three plants are used during each round of backcross.

-m *File*: Marker file in SHORE marker output format.

Optional:

-o *Characters:* Output file name. Default is genotyping.txt in the local directory. Rewrite if the file already excites.

-g *Flag:* Write graphical outputs files for each chromosome. Default name of files are Chromosome_1-9.txt

-l *Flag:* Print complete genomes from every generation, including the intermediate populations

-s *Character:* Mutation site. Example: 1:100 for chr1 and 100bp. Default is 1:100000

-i *File*: Output file from previous Pop simulation. Utilize the genomes simulated earlier.

-p *Flag*: Flag for 4 parent cross. Two F_1 are made out of four parents and further selfed to user specified times to simulate recombinant inbred lines.

-e *Float*: Expected phenotyping error in percentage

-c *File:* Config file for species. Default is Arabidopsis_config.txt, which is provided with package. Rice and Barley config files are also provided in package. To create config file for other species, please followe the guidelines given in the README of the package.

-d *Flag:* Use single genome till the last stage of simulation and create last population with specified plants from it. This simulate single seed descent

-j *Flag:* Use only single plant in each backcross. Default is three.

-h *Flag:* Help

-V *Flag:* Verbose. Be talkative and report what is going on while analyzing.

The output of Pop simulator serves as an input for Seq simulator. The minimum required parameters to run Seq simulator are genome file, marker file, normalized coverage file and required coverage. The usage as follows:

perl simulate_seq.pl -c Coverage.txt -g Genotype.txt –m Marker_file.txt -x 50

Following are the details of each parameter:

-h *Flag:* Help

-c *File*: Coverage file with normalized expected coverage at each marker position

-g *File:* Output file from Pop simulator having simulated genomes.

-m *File:* Marker file in SHORE marker output format

-x *Numeric:* Required coverage of simulation

-o *File:* Name of output file . Default is Output.txt

-b *Flag:* Instead of pooling all the genome from the input file, each genome sequencing will be simulated separately

-a *Flag:* Faster simulation method. Not usable below 1x coverage

Appendix Tables:

Table SI: List of markers induced by EMS in *alp1;lhp1 mutant*

Chr	Locus	Wild-type	Mutant	Score	Cov	Concordance	Repetitiveness
1	528089	C	T	25	8	0.22	1.00
1	1455002	C	T	30	9	0.27	1.00
1	1567367	C	T	40	11	0.38	1.00
1	2296104	C	T	25	9	0.15	1.00
1	3467889	C	T	36	10	0.26	1.00
1	3629253	C	T	30	9	0.24	1.00
1	3675554	C	T	30	10	0.25	1.00
1	4838651	C	T	30	8	0.35	1.00
1	4954896	G	A	25	10	0.21	1.00
1	5137993	C	T	28	8	0.26	1.00
1	5544941	C	T	27	10	0.24	1.00
1	6069148	C	T	36	12	0.26	1.00
1	6752291	C	T	36	17	0.33	1.00
1	8294426	G	A	36	11	0.27	1.00
1	8301873	G	A	40	15	0.38	1.00
1	10661932	G	A	36	10	0.28	1.00
1	10728593	G	A	36	17	0.33	1.00
1	12840240	C	T	34	9	0.26	1.00
1	13180730	C	T	28	23	0.35	1.60
1	13181778	G	A	40	33	0.36	1.12
1	13187030	G	A	34	13	0.27	1.25
1	13572421	C	T	34	9	0.25	1.21
1	13572438	C	T	28	9	0.22	1.48
1	13598810	G	A	25	11	0.24	1.00
1	13598853	G	A	25	8	0.18	1.04
1	13840908	C	T	25	9	0.16	1.06
1	14044723	G	A	25	10	0.18	1.19
1	14105701	C	T	30	10	0.22	1.17
1	14223413	G	A	25	11	0.22	1.00
1	14227448	C	T	34	15	0.28	1.28
1	14227454	C	T	28	14	0.26	1.23
1	14237345	C	T	28	14	0.27	1.23
1	14270990	C	T	25	8	0.19	1.17
1	14271005	G	A	25	8	0.16	1.14
1	14271927	C	T	30	12	0.24	1.09
1	14464852	G	A	34	8	0.27	1.00
1	14486841	C	T	36	14	0.25	1.00
1	14486851	C	T	36	15	0.29	1.00
1	14497035	C	T	25	8	0.16	1.00
1	14499162	C	T	25	11	0.24	1.02
1	14509881	C	T	38	21	0.36	1.45

Chr	Locus	Wild-type	Mutant	Score	Cov	Concordance	Repetitiveness
1	14544115	G	A	25	12	0.21	1.00
1	14592376	G	A	36	16	0.25	1.01
1	14666764	G	A	28	12	0.23	1.50
1	14922405	C	T	36	16	0.25	1.00
1	15060309	G	A	25	10	0.20	1.02
1	15060369	G	A	25	13	0.22	1.02
1	15091990	C	T	32	14	0.35	1.67
1	15096904	C	T	25	8	0.18	1.20
1	15098643	C	T	25	9	0.18	1.18
1	15103278	C	T	30	18	0.25	1.13
1	15148451	C	T	30	8	0.24	1.11
1	15169937	G	A	32	25	0.42	1.73
1	15196913	C	T	25	9	0.20	1.04
1	15196933	C	T	25	9	0.21	1.02
1	15198938	C	T	25	32	0.22	1.04
1	15208584	G	A	25	10	0.23	1.00
1	15437072	C	T	25	26	0.18	1.02
1	15437359	G	A	30	34	0.25	1.01
1	15437371	G	A	25	32	0.19	1.01
1	15522690	C	T	25	8	0.18	1.00
1	15612103	G	A	25	11	0.21	1.00
1	15612109	G	A	25	11	0.20	1.00
1	15839181	G	A	28	9	0.27	1.51
1	16012283	C	T	36	9	0.25	1.07
1	16052952	C	T	25	9	0.24	1.00
1	16103829	G	A	40	20	0.48	1.00
1	16513124	C	T	25	21	0.16	1.04
1	16514381	G	A	25	18	0.15	1.15
1	16519806	C	T	27	28	0.26	1.59
1	16522410	G	A	25	21	0.16	1.13
1	17143580	C	T	25	14	0.20	1.03
1	17144868	G	A	25	8	0.19	1.11
1	17145062	G	A	36	18	0.33	1.05
1	17145106	G	A	36	19	0.28	1.11
1	17602952	G	A	36	10	0.27	1.00
1	17866864	C	T	34	9	0.35	1.00
1	18316593	G	A	28	10	0.29	1.00
1	18364642	G	A	40	15	0.56	1.00
1	18641548	G	A	40	20	0.42	1.00
1	19680096	G	A	40	17	0.47	1.00
1	19706004	C	T	25	10	0.23	1.09
1	19983954	G	A	40	28	0.61	1.00
1	20196679	G	A	40	21	0.53	1.00
1	20227699	G	A	40	21	0.45	1.00

Chr	Locus	Wild-type	Mutant	Score	Cov	Concordance	Repetitiveness
1	22119003	G	A	40	15	0.38	1.00
1	22795495	G	A	40	17	0.38	1.00
1	22937848	G	A	40	17	0.53	1.00
1	23257047	G	A	36	9	0.56	1.00
1	23304045	G	A	40	25	0.53	1.00
1	24681784	G	A	40	23	0.46	1.00
1	25109150	G	A	40	17	0.44	1.00
1	25299556	G	A	40	13	0.36	1.00
1	25699757	C	T	36	9	0.27	1.00
1	26750832	C	T	32	14	0.30	1.00
1	27211891	G	A	30	9	0.26	1.00
1	27834832	G	A	36	11	0.32	1.00
1	27971974	G	A	28	8	0.29	1.00
1	28956103	C	T	28	9	0.23	1.00
1	29345246	C	T	40	8	0.47	1.00
2	48429	C	T	25	18	0.18	1.05
2	368426	G	A	40	18	0.43	1.00
2	1606662	C	T	30	15	0.25	1.16
2	1606664	G	A	25	9	0.15	1.16
2	2163852	G	A	25	8	0.24	1.00
2	2324978	C	T	34	11	0.28	1.00
2	2496417	G	A	36	15	0.31	1.00
2	3548843	G	A	25	12	0.15	1.08
2	3550221	C	T	28	17	0.30	1.53
2	3580120	G	A	25	9	0.24	1.00
2	3631586	G	A	36	10	0.27	1.10
2	3758100	C	T	36	8	0.38	1.00
2	3845310	G	A	25	9	0.20	1.02
2	3899518	C	T	34	14	0.29	1.21
2	3906321	C	T	36	13	0.27	1.11
2	3948585	C	T	25	8	0.19	1.12
2	4190014	G	A	25	12	0.21	1.00
2	4380286	G	A	28	15	0.31	1.29
2	4380291	G	A	38	19	0.38	1.28
2	4381592	G	A	25	12	0.21	1.11
2	4381601	C	T	36	19	0.29	1.10
2	4383398	G	A	25	13	0.22	1.18
2	4516286	G	A	25	8	0.18	1.02
2	4660276	G	A	30	10	0.27	1.00
2	4682855	C	T	25	10	0.18	1.14
2	4776727	G	A	34	12	0.29	1.33
2	4950351	G	A	25	12	0.21	1.10
2	5282409	C	T	25	9	0.20	1.00
2	5475183	C	T	25	10	0.18	1.17

Chr	Locus	Wild-type	Mutant	Score	Cov	Concordance	Repetitiveness
2	5475189	C	T	25	10	0.16	1.16
2	6124027	G	A	25	9	0.24	1.07
2	6124266	C	T	36	15	0.35	1.24
2	6176791	G	A	40	15	0.41	1.00
2	6652425	C	T	36	9	0.36	1.00
2	6778878	C	T	34	8	0.33	1.00
2	7568835	G	A	36	17	0.33	1.00
2	9760661	G	A	30	12	0.25	1.00
2	10431148	G	A	36	12	0.27	1.00
2	10720689	G	A	38	9	0.36	1.00
2	11701603	G	A	36	13	0.27	1.00
2	11954694	G	A	36	11	0.31	1.00
2	12402711	G	A	30	12	0.26	1.00
2	12600322	G	A	40	8	0.38	1.00
2	12600328	G	A	30	8	0.30	1.00
2	12722481	C	T	30	10	0.32	1.00
2	12761295	G	A	40	26	0.54	1.00
2	13836108	G	A	40	17	0.44	1.00
2	13995863	G	A	40	15	0.42	1.00
2	15652847	C	T	40	21	0.48	1.00
2	15965780	G	A	40	21	0.53	1.00
2	17746783	C	T	40	17	0.52	1.00
2	19039666	C	T	40	18	0.43	1.00
2	19100239	C	T	40	12	0.48	1.00
2	19332588	C	T	36	8	0.32	1.00
3	730961	C	T	25	11	0.19	1.00
3	826933	C	T	40	9	0.50	1.00
3	2413273	G	A	34	8	0.31	1.00
3	3205070	C	T	40	9	0.38	1.14
3	3205071	C	T	40	9	0.38	1.14
3	3205085	C	T	40	9	0.43	1.09
3	5280990	G	A	25	11	0.17	1.19
3	5285703	G	A	25	13	0.22	1.10
3	6456060	G	A	38	14	0.44	1.00
3	7995702	G	A	30	10	0.26	1.20
3	9773917	C	T	25	11	0.20	1.00
3	12210366	C	T	36	10	0.26	1.17
3	12210533	C	T	28	11	0.33	1.71
3	12210537	G	A	28	8	0.27	1.74
3	12214444	C	T	25	9	0.20	1.00
3	12248419	G	A	25	11	0.21	1.02
3	12249639	G	A	25	12	0.24	1.02
3	12252193	G	A	28	15	0.26	1.74
3	12334101	C	T	36	14	0.32	1.00

Chr	Locus	Wild-type	Mutant	Score	Cov	Concordance	Repetitiveness
3	12334106	C	T	36	13	0.31	1.00
3	12334114	C	T	36	13	0.30	1.00
3	12334193	C	T	36	16	0.29	1.00
3	12334254	G	A	36	15	0.28	1.00
3	12425773	C	T	40	17	0.61	1.00
3	12459357	C	T	40	14	0.56	1.00
3	12672117	C	T	25	8	0.16	1.09
3	12715128	G	A	34	11	0.25	1.38
3	12715161	C	T	28	11	0.26	1.28
3	12893634	C	T	25	8	0.19	1.08
3	13280245	G	A	25	11	0.20	1.17
3	13315273	C	T	25	10	0.21	1.00
3	13413001	C	T	30	11	0.32	1.72
3	13421587	G	A	25	11	0.18	1.02
3	13586890	C	T	25	11	0.24	1.19
3	13595615	G	A	25	10	0.22	1.15
3	13598303	G	A	38	16	0.38	1.25
3	13604965	C	T	25	9	0.20	1.00
3	13658611	G	A	28	16	0.26	1.67
3	13693460	G	A	34	9	0.28	1.33
3	13774798	C	T	28	11	0.22	1.49
3	13794584	G	A	34	14	0.29	1.49
3	13835007	G	A	28	25	0.31	1.65
3	13835026	G	A	30	21	0.33	1.74
3	13840268	G	A	25	8	0.22	1.17
3	13912473	G	A	34	14	0.29	1.22
3	13938604	C	T	25	11	0.17	1.06
3	14049194	G	A	30	12	0.29	1.13
3	14164972	C	T	40	14	0.56	1.00
3	14174981	C	T	32	14	0.36	1.73
3	14216604	C	T	34	16	0.26	1.25
3	14392427	C	T	25	9	0.17	1.18
3	14394603	C	T	25	10	0.17	1.19
3	14394625	C	T	28	16	0.24	1.23
3	14475646	C	T	25	10	0.21	1.00
3	14476306	C	T	25	8	0.21	1.03
3	14476740	C	T	25	11	0.19	1.00
3	14587126	C	T	40	19	0.39	1.00
3	14587602	C	T	36	9	0.30	1.00
3	14814713	C	T	34	13	0.33	1.41
3	14820146	G	A	25	8	0.17	1.20
3	14980218	G	A	28	10	0.24	1.39
3	15141273	C	T	30	9	0.24	1.17
3	15143316	C	T	25	8	0.20	1.02

Chr	Locus	Wild-type	Mutant	Score	Cov	Concordance	Repetitiveness
3	15143716	G	A	25	11	0.24	1.02
3	15144644	C	T	25	10	0.24	1.00
3	15253456	C	T	38	24	0.38	1.21
3	15258016	C	T	40	15	0.42	1.00
3	15258203	C	T	25	15	0.23	1.03
3	15305302	C	T	25	17	0.17	1.05
3	15465812	G	A	25	10	0.20	1.10
3	15465823	G	A	25	10	0.21	1.08
3	15466756	C	T	25	9	0.16	1.07
3	15466762	C	T	25	9	0.18	1.00
3	15717631	C	T	25	11	0.19	1.00
3	15717797	G	A	36	9	0.33	1.00
3	16071224	G	A	40	8	0.44	1.00
3	16458036	C	T	40	19	0.54	1.00
3	18310153	C	T	38	18	0.43	1.31
3	18536084	C	T	40	32	0.60	1.00
3	18757858	C	T	40	29	0.78	1.00
3	19462997	C	T	40	27	0.64	1.00
3	19728658	C	T	40	17	0.61	1.00
3	21455099	G	A	40	47	0.96	1.00
3	22622352	C	T	40	44	0.96	1.00
3	23376305	C	T	40	39	0.98	1.00
4	158374	C	T	40	16	0.37	1.00
4	1231684	G	A	25	8	0.20	1.18
4	1683886	C	T	40	13	0.38	1.00
4	1751956	C	T	25	12	0.21	1.02
4	1753451	G	A	34	16	0.25	1.25
4	1753487	C	T	25	12	0.21	1.15
4	1753889	C	T	25	11	0.20	1.18
4	1753991	C	T	25	13	0.20	1.00
4	1753994	C	T	25	11	0.17	1.00
4	2057261	C	T	36	18	0.33	1.00
4	2057798	C	T	25	10	0.21	1.00
4	2058440	G	A	25	9	0.18	1.00
4	2062991	G	A	28	11	0.23	1.35
4	2362823	G	A	36	17	0.30	1.00
4	2509201	G	A	38	12	0.32	1.00
4	2824240	C	T	30	8	0.27	1.00
4	2856794	C	T	25	10	0.21	1.00
4	3039701	G	A	30	15	0.24	1.00
4	3040240	C	T	25	12	0.21	1.05
4	3048933	C	T	34	15	0.31	1.33
4	3377359	G	A	25	14	0.23	1.00
4	3377364	C	T	25	15	0.23	1.00

Chr	Locus	Wild-type	Mutant	Score	Cov	Concordance	Repetitiveness
4	3505971	C	T	25	10	0.22	1.08
4	3506033	C	T	36	16	0.33	1.02
4	3506228	G	A	25	11	0.20	1.00
4	3506608	C	T	25	8	0.19	1.00
4	3506664	C	T	25	10	0.19	1.05
4	3574581	G	A	30	12	0.24	1.04
4	3574587	G	A	25	12	0.24	1.07
4	3587289	G	A	28	17	0.27	1.24
4	3653212	G	A	30	11	0.24	1.00
4	3654292	C	T	25	8	0.17	1.00
4	3671480	C	T	36	12	0.32	1.00
4	3871575	C	T	25	30	0.23	1.19
4	3963607	C	T	25	13	0.21	1.09
4	3991189	G	A	28	13	0.25	1.44
4	4030446	G	A	30	12	0.24	1.13
4	4052720	G	A	30	8	0.24	1.00
4	4066510	C	T	25	8	0.23	1.00
4	4209915	G	A	36	20	0.28	1.13
4	4229695	C	T	25	10	0.23	1.10
4	4274255	C	T	36	22	0.34	1.00
4	4274306	G	A	38	22	0.35	1.09
4	4274777	G	A	25	8	0.18	1.19
4	4283066	G	A	25	9	0.18	1.00
4	4326437	C	T	38	14	0.44	1.25
4	4362007	C	T	25	10	0.21	1.00
4	4409809	C	T	30	14	0.23	1.09
4	4409818	G	A	30	14	0.24	1.10
4	4459412	C	T	25	8	0.15	1.00
4	4559212	G	A	25	11	0.22	1.04
4	4565539	C	T	25	9	0.19	1.04
4	4678150	C	T	36	9	0.27	1.00
4	4770181	C	T	25	10	0.19	1.05
4	4958658	G	A	30	9	0.25	1.00
4	4958660	G	A	30	9	0.25	1.00
4	5105463	G	A	34	12	0.27	1.27
4	5568907	C	T	36	11	0.28	1.00
4	6210123	C	T	30	11	0.24	1.00
4	6700268	C	T	36	16	0.31	1.00
4	8347080	C	T	40	12	0.46	1.00
4	9106512	C	T	36	14	0.33	1.00
4	9390545	C	T	36	15	0.33	1.00
4	9894974	C	T	40	14	0.42	1.00
4	10158115	C	T	36	13	0.29	1.00
4	10601154	C	T	40	18	0.50	1.00

Chr	Locus	Wild-type	Mutant	Score	Cov	Concordance	Repetitiveness
4	10865356	C	T	25	10	0.24	1.00
4	11249251	C	T	36	13	0.30	1.00
4	12010179	C	T	36	16	0.34	1.00
4	13981827	C	T	40	12	0.38	1.00
4	14467913	C	T	40	16	0.38	1.00
4	14868086	C	T	25	9	0.20	1.00
4	15093833	C	T	40	17	0.36	1.00
4	15132107	C	T	30	9	0.29	1.00
4	17834574	G	A	34	9	0.32	1.00
5	4004237	G	A	25	9	0.20	1.00
5	4005950	C	T	36	14	0.30	1.00
5	6362038	C	T	40	14	0.41	1.00
5	6461731	C	T	25	8	0.20	1.00
5	6676861	C	T	38	18	0.36	1.00
5	7179925	C	T	36	9	0.33	1.00
5	8067959	G	A	28	8	0.26	1.00
5	8068125	C	T	36	11	0.31	1.00
5	8177879	C	T	40	17	0.63	1.00
5	8225565	C	T	36	9	0.33	1.00
5	8524929	C	T	38	11	0.33	1.00
5	8748204	C	T	36	12	0.32	1.00
5	8750589	C	T	38	11	0.32	1.00
5	9054815	C	T	36	10	0.28	1.00
5	9704823	G	A	40	17	0.63	1.00
5	9751375	G	A	32	8	0.40	1.00
5	10087331	G	A	25	18	0.17	1.00
5	10087385	G	A	25	20	0.19	1.05
5	10142959	G	A	40	16	0.36	1.10
5	10234889	C	T	36	13	0.25	1.00
5	10368670	C	T	25	9	0.24	1.00
5	10724651	G	A	40	16	0.50	1.00
5	10730334	C	T	25	12	0.19	1.03
5	10730336	G	A	25	10	0.16	1.03
5	10731992	G	A	36	19	0.28	1.00
5	10969465	C	T	36	9	0.32	1.00
5	11066800	C	T	36	10	0.27	1.00
5	11093834	C	T	25	11	0.19	1.00
5	11142826	G	A	30	14	0.23	1.00
5	11217460	G	A	28	15	0.25	1.31
5	11217696	C	T	30	13	0.22	1.16
5	11402073	G	A	36	9	0.27	1.00
5	11646111	C	T	30	14	0.25	1.05
5	11646138	C	T	30	13	0.24	1.05
5	11646174	G	A	25	13	0.23	1.03

Chr	Locus	Wild-type	Mutant	Score	Cov	Concordance	Repetitiveness
5	11718415	G	A	32	48	0.33	1.32
5	11729523	C	T	28	42	0.26	1.61
5	11733052	C	T	25	10	0.15	1.17
5	11768393	G	A	25	12	0.18	1.03
5	11773359	C	T	25	17	0.22	1.15
5	11774499	G	A	28	24	0.28	1.72
5	11775330	G	A	36	19	0.27	1.10
5	11776375	C	T	30	9	0.22	1.18
5	11776690	C	T	38	12	0.36	1.45
5	11778486	C	T	25	11	0.15	1.19
5	11803108	C	T	25	21	0.18	1.00
5	11803883	C	T	36	14	0.25	1.00
5	11851810	C	T	25	11	0.18	1.10
5	11973646	C	T	36	12	0.29	1.08
5	11973650	C	T	36	13	0.32	1.09
5	12009107	G	A	25	12	0.19	1.11
5	12009131	G	A	34	17	0.26	1.36
5	12010174	G	A	25	21	0.23	1.02
5	12011690	C	T	25	9	0.20	1.08
5	12041420	C	T	25	11	0.22	1.10
5	12110748	G	A	28	16	0.25	1.52
5	12133821	G	A	36	13	0.30	1.06
5	12350141	G	A	28	14	0.33	1.80
5	12359375	G	A	25	12	0.23	1.00
5	12487823	G	A	25	9	0.15	1.10
5	12633254	G	A	28	16	0.28	1.51
5	12768493	C	T	34	11	0.25	1.34
5	12768521	G	A	30	10	0.24	1.18
5	12918997	C	T	34	15	0.28	1.44
5	12918998	C	T	34	15	0.28	1.44
5	12986017	G	A	25	9	0.19	1.00
5	12986020	G	A	25	9	0.19	1.00
5	12986027	G	A	25	9	0.20	1.00
5	12986028	C	T	25	9	0.19	1.00
5	13143226	G	A	38	10	0.36	1.07
5	13226313	G	A	36	17	0.29	1.06
5	13226341	C	T	36	15	0.31	1.18
5	13226880	G	A	25	11	0.20	1.00
5	13304517	C	T	25	10	0.18	1.19
5	13304524	C	T	25	10	0.17	1.16
5	13399435	C	T	25	10	0.19	1.05
5	13406258	G	A	36	22	0.31	1.18
5	13406880	C	T	25	12	0.22	1.00
5	13406895	C	T	36	12	0.25	1.08

Chr	Locus	Wild-type	Mutant	Score	Cov	Concordance	Repetitiveness
5	13505301	C	T	36	14	0.25	1.13
5	13505322	G	A	36	14	0.28	1.14
5	13647781	C	T	40	19	0.44	1.00
5	15347294	C	T	36	14	0.33	1.00
5	17521625	C	T	28	9	0.33	1.00
5	17664078	C	T	36	14	0.30	1.00
5	21320859	G	A	36	15	0.28	1.00
5	22187873	C	T	25	13	0.21	1.00
5	23920865	G	A	34	8	0.26	1.00
5	24029249	C	T	36	18	0.29	1.10
5	24029259	C	T	25	13	0.22	1.16
5	24029468	C	T	40	23	0.36	1.04
5	24102801	G	A	40	10	0.42	1.00
5	24106528	C	T	40	22	0.49	1.00
5	25008363	C	T	40	26	0.60	1.00
5	26069971	G	A	30	15	0.26	1.20
5	26086756	G	A	40	22	0.50	1.00
5	26113581	G	A	36	11	0.28	1.00
5	26272128	G	A	28	8	0.27	1.00

Table SII: List of plant genomes published till April 2013

Organisum	**Year**	**Name**	**Assembled / estimated genome size**	**Citation**
Arabidopsis thaliana	2000	Arabidopsis	119 Mb	(Arabidopsis Genome Initiative, 2000)
Oryza sativa L. ssp. japonica	2002	Rice	420 Mb	(Goff, 2002)
Oryza sativa L. ssp. indica	2002	Rice	466 Mb	(Yu *et al.*, 2002)
Populus trichocarpa	2006	Black cottonwood	~485 Mb	(Tuskan *et al.*, 2006)
Vitis vinifera	2007	Grapevine	475 Mb	(Jaillon *et al.*, 2007)
Lotus japonicus	2008	Lotus	472 Mb	(Sato *et al.*, 2008)
Carica papaya	2008	Papaya	372 Mb	(Ming *et al.*, 2008)
Physcomitrella patens	2008	Physcomitrella	480 Mb	(Rensing *et al.*, 2008)
Sorghum bicolor	2009	Sorghum	~730 Mb	(Paterson *et al.*, 2009)
Cucumis sativus	2009	Cucumber	367 Mb	(S., Huang *et al.*, 2009
Zea mays	2009	Maize	2.3 Gb	(Schnable *et al.*, 2009)
Ricinus communis	2010	Castor bean	~320 Mb	(Chan *et al.*, 2010)
Malus × *domestica*	2010	Apple	742 Mb	(Velasco *et al.*, 2010)

Fragaria vesca	2010	Strawberry	~240 Mb	(Shulaev *et al.*, 2010)
Theobroma cacao	2010	Cacao	430 Mb	(Argout *et al.*, 2011)
Brachypodium distachyon	2010	Brachypodium	~272 Mb	(Vogel *et al.*, 2010)
Glycine max	2010	Soybean	~1.1 Gb	(Schmutz *et al.*, 2010)
Glycine soja	2010	Soybean	~1.1 Gb	(Kim *et al.*, 2010)
Arabidopsis lyrata	2011	Arabidopsis	207 Mb	(Hu *et al.*, 2011)
Brassica rapa	2011	Chinese cabbage	~283 Mb	(Wang *et al.*, 2011)
Thellungiella parvula	2011	Thellungiella	160 Mb	(Dassanayake *et al.*, 2011)
Solanum tuberosum	2011	Potato	844 Mb	(Xu *et al.*, 2011)
Selaginella moellendorffii	2011	Selaginella	~106 Mb	(Banks *et al.*, 2011)
Phoenix dactylifera	2011	Date palm	~658 Mb	(Al-Dous *et al.*, 2011)
Cajanus cajan	2011	Pigeonpea	833 Mb	(Varshney *et al.*, 2011)
Cannabis sativa	2011	Cannabis	534 Mb	(Bakel *et al.*, 2011)
Medicago truncatula	2011	Medicago	375 Mb	(Young *et al.*, 2011)
Solanum lycopersicum	2012	Tomato	900 Mb	(Consortium, 2012)
Linum usitatissimum	2012	Flax	350 Mb	(Z., Wang *et al.*, 2012)

Manihot esculenta	2012	Cassava	770 Mb	(Prochnik *et al.*, 2012)
Triticum aestivum	2012	Wheat	~17 Gb	(Brenchley *et al.*, 2012)
Cucumis melo	2012	Melon	450 Mb	(Garcia-Mas *et al.*, 2012)
Setaria italica	2012	Foxtail millet	~423 Mb	(G., Zhang *et al.*, 2012)
Hordeum vulgare	2012	Barley	5.1 Gb	(Mayer *et al.*, 2012)
Prunus mume	2012	Prunus mume	280 Mb	(Q., Zhang *et al.*, 2012)
Gossypium raimondii	2012	Cotton	~775 Mb	(K., Wang *et al.*, 2012)
Azadirachta indica	2012	Neem	364 Mb	(Krishnan *et al.*, 2012)
Thellungiella salsuginea	2012	Thellungiella salsuginea	243 Mb	(Wu *et al.*, 2012)
Musa acuminata	2012	Banana	523 Mb	(D'Hont *et al.*, 2012)
Pyrus bretschneideri	2013	Pear	527 Mb	(Wu *et al.*, 2013)
Citrullus lanatus	013	Watermelon	~425 Mb	(Guo *et al.*, 2013)
Betula nana	013	Dwarf birch	~450 Mb	(Wang *et al.*, 2013)
Hevea brasiliensis	013	Rubber tree	~2.15 Gb	(Rahman *et al.*, 2013)
Cicer arietinum	2013	Chickpea	~738 Mb	(Varshney *et al.*, 2013)
Triticum urartu (A genome)	2013	Wheat	4.94 Gb	(Ling *et al.*, 2013)

Phyllostachys heterocycla	2013	Moso bamboo	2.05 Gb	(Peng *et al.*, 2013)
Utricularia gibba	2013	Bladderwort	82 Mb	(Ibarra-Laclette *et al.*, 2013)
Lupinus angustifolius	2013	Lupin	960 Mb	(Książkiewicz *et al.*, 2013)
Capsella rubella	2013	Capsella rubella	135 Mb	(Slotte *et al.*, 2013)

Supplementary literatures cited:

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

Argout, X., Salse, J., Aury, J.-M., *et al.* (2011) The genome of Theobroma cacao. *Nat. Genet.*, **43**, 101–108.

Bakel, H. van, Stout, J.M., Cote, A.G., Tallon, C.M., Sharpe, A.G., Hughes, T.R. and Page, J.E. (2011) The draft genome and transcriptome of Cannabis sativa. *Genome Biol.*, **12**, R102.

Banks, J.A., Nishiyama, T., Hasebe, M., *et al.* (2011) The Selaginella Genome Identifies Genetic Changes Associated with the Evolution of Vascular Plants. *Science*, **332**, 960–963.

Brenchley, R., Spannagl, M., Pfeifer, M., *et al.* (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.

Chan, A.P., Crabtree, J., Zhao, Q., *et al.* (2010) Draft genome sequence of the oilseed species Ricinus communis. *Nat. Biotechnol.*, **28**, 951–956.

Consortium, T.T.G. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

D'Hont, A., Denoeud, F., Aury, J.-M., *et al.* (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*, **488**, 213–217.

Dassanayake, M., Oh, D.-H., Haas, J.S., *et al.* (2011) The genome of the extremophile crucifer Thellungiella parvula. *Nat. Genet.*, **43**, 913–918.

Al-Dous, E.K., George, B., Al-Mahmoud, M.E., *et al.* (2011) De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). *Nat. Biotechnol.*, **29**, 521–527.

Garcia-Mas, J., Benjak, A., Sanseverino, W., *et al.* (2012) The genome of melon (Cucumis melo L.). *Proc. Natl. Acad. Sci.*, **109**, 11872–11877

Goff, S.A. (2002) A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica). *Science*, **296**, 92–100.

Guo, S., Zhang, J., Sun, H., *et al.* (2013) The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions. *Nat. Genet.*, **45**, 51–58.

Hu, T.T., Pattyn, P., Bakker, E.G., *et al.* (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.*, **43**, 476–481.

Huang, S., Li, R., Zhang, Z., *et al.* (2009) The genome of the cucumber, Cucumis sativus L. *Nat. Genet.*, **41**, 1275–1281.

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., *et al.* (2013) Architecture and evolution of a minute plant genome. *Nature*, **498**, 94–98.

Jaillon, O., Aury, J.-M., Noel, B., *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.

Kim, M.Y., Lee, S., Van, K., *et al.* (2010) Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. *Proc. Natl. Acad. Sci.*, **107**:22032-7.

Krishnan, N.M., Pattnaik, S., Jain, P., *et al.* (2012) A draft of the genome and four transcriptomes of a medicinal and pesticidal angiosperm Azadirachta indica. *BMC Genomics*, **13**, 464.

Książkiewicz, M., Wyrwa, K., Szczepaniak, A., Rychel, S., Majcherkiewicz, K., Przysiecka, Ł., Karlowski, W., Wolko, B. and Naganowska, B. (2013) Comparative genomics of Lupinus angustifolius gene-rich regions: BAC library exploration, genetic mapping and cytogenetics. *BMC Genomics*, **14**, 79.

Ling, H.-Q., Zhao, S., Liu, D., *et al.* (2013) Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature*, **496**, 87–90.

Mayer, K.F.X., Waugh, R., Langridge, P., *et al.* (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*. **491**:711-6

Ming, R., Hou, S., Feng, Y., *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature*, **452**, 991–996.

Paterson, A.H., Bowers, J.E., Bruggmann, R., *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.

Peng, Z., Lu, Ying, Li, L., *et al.* (2013) The draft genome of the fast-growing non-timber forest species moso bamboo (Phyllostachys heterocycla). *Nat. Genet.*, **45**, 456–461.

Prochnik, S., Marri, P.R., Desany, B., *et al.* (2012) The Cassava Genome: Current Progress, Future Directions. *Trop. Plant Biol.*, **5**, 88–94.

Rahman, A.Y.A., Usharraj, A.O., Misra, B.B., *et al.* (2013) Draft genome sequence of the rubber tree Hevea brasiliensis. *BMC Genomics*, **14**, 75.

Rensing, S.A., Lang, D., Zimmer, A.D., *et al.* (2008) The Physcomitrella Genome Reveals Evolutionary Insights into the Conquest of Land by Plants. *Science*, **319**, 64–69.

Sato, S., Nakamura, Y., Kaneko, T., *et al.* (2008) Genome Structure of the Legume, Lotus japonicus. *DNA Res.*, **15**, 227–239.

Schmutz, J., Cannon, S.B., Schlueter, J., *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Schnable, P.S., Ware, D., Fulton, R.S., *et al.* (2009) The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*, **326**, 1112–1115.

Shulaev, V., Sargent, D.J., Crowhurst, R.N., *et al.* (2010) The genome of woodland strawberry (Fragaria vesca). *Nat. Genet.*, **43**, 109–116.

Slotte, T., Hazzouri, K.M., Ågren, J.A., *et al.* (2013) The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.*, **45**, 831–835.

Tuskan, G.A., DiFazio, S., Jansson, S., *et al.* (2006) The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, **313**, 1596–1604.

Varshney, R.K., Chen, W., Li, Y., *et al.* (2011) Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**:83-9

Varshney, R.K., Song, C., Saxena, R.K., *et al.* (2013) Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. *Nat. Biotechnol.*, **31**, 240–246.

Velasco, R., Zharkikh, A., Affourtit, J., *et al.* (2010) The genome of the domesticated apple (Malus × domestica Borkh.). *Nat. Genet.*, **42**, 833–839.

Vogel, J.P., Garvin, D.F., Mockler, T.C., *et al.* (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, **463**, 763–768.

Wang, K., Wang, Z., Li, F., *et al.* (2012) The draft genome of a diploid cotton Gossypium raimondii. *Nat. Genet.*, **44**, 1098–1103.

Wang, N., Thomson, M., Bodles, W.J.A., Crawford, R.M.M., Hunt, H.V., Featherstone, A.W., Pellicer, J. and Buggs, R.J.A. (2013) Genome sequence of dwarf birch (Betula nana) and cross-species RAD markers. *Mol. Ecol.*, **22**, 3098–3111.

Wang, X., Wang, H., Wang, J., *et al.* (2011) The genome of the mesopolyploid crop species Brassica rapa. *Nat. Genet.*, **43**, 1035–1039.

Wang, Z., Hobson, N., Galindo, L., *et al.* (2012) The genome of flax (Linum usitatissimum) assembled de novo from short shotgun sequence reads. *Plant J.*, **72**, 461–473.

Wu, H.-J., Zhang, Z., Wang, J.-Y., *et al.* (2012) Insights into salt tolerance from the genome of Thellungiella salsuginea. *Proc. Natl. Acad. Sci.*, **109**, 12219–12224.

Wu, J., Wang, Z., Shi, Z., *et al.* (2013) The genome of the pear (Pyrus bretschneideri Rehd.). *Genome Res.*, **23**, 396–408.

Xu, X., Pan, S., Cheng, S., *et al.* (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189–195.

Young, N.D., Debellé, F., Oldroyd, G.E.D., *et al.* (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, **480**, 520–524.

Yu, J., Hu, S., Wang, J., *et al.* (2002) A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica). *Science*, **296**, 79–92.

Zhang, G., Liu, X., Quan, Z., *et al.* (2012) Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. *Nat. Biotechnol.*, **30**, 549–554.

Zhang, Q., Chen, W., Sun, L., *et al.* (2012) The genome of Prunus mume. *Nat. Commun.*, **3**, 1318.

Acknowledgement

First and foremost, I would like to thank Korbinian Schneeberger for his boundless support and trust on me. I thoroughly enjoyed our all informal discussions during late evenings and learned many new lessons. All those ‘small questions’ were big lessons for my PhD. I would also like to thank George Coupland for seeing potential in me and trusting and selecting me for this position. His valuable comments helped me in progressing through the entire phase of my PhD. I thank, Maarten Koornneef for his support as co-supervisor and moreover, as a mentor to guide me through future opportunities. I thank my PhD committee members, Thomas Wiehe and Martin Hülskamp .

Karl, you have been a very good friend and have helped me throughout the years, especially during my starting days. My wonderful colleagues, Jonas, Vipul, Eva, Vimal, Hequan, Xue, Ben and Jia, thanks for creating such a supportive and pleasant working atmosphere. Our coffee breaks as well as lunchtime were truly relaxing. In fact, some of our coffee break discussions were really inspiring (Of course, this is only referring to those who never say no to coffee..). From you all, other than science, I also learned some new skills such as skiing and speedminton and of course, improved my badminton skills. It is hard to name all, because there are many if not too many, therefore, I would collectively thank all my good colleagues at MPIPZ, specially my fellow IMPRS friends. I would also like to thank my all collaborators and co-authors for the pleasant cooperation and successful projects.

During all these five years of my Europe stay, Animesh, Sharmistha, Aniruddho, Soumi and Surender, you all provided me a second home away from Koln. In Koln, Deepak, Vivek, Vimal, Bala, Shachi, Aishwarya, Arun, Luci, Ganga and Vid, you all kept my home sickness away by providing homely atmosphere. I won’t forget our “Pandeyji’s” get together for some spicy food and drinks blended with scientific discussions. Along with my degree, I will also be taking a lot of friendships from Cologne. Because of all of you, I will always cherish my stay in Cologne.

I thank all my old good friends back in India and around the world, especially Imran, Atul, Neha, Amit and Ramesh. Their faith and encouragement have always increased my self-confidence.

My family, Chachen, Ammachi, Machayan, Jose, Antony, Mebi, Smitha, Joyal, Isabella and Angelina supported me and backed me whenever I needed. Their trust and confidence always wonder me and push me to give the best of me. Last but not the least, I would like to mention my friend, who then became my girlfriend and now my lovely better half (truly in all sense), Shushimita. Her support during the most difficult days of my life is speechless. I won't use the word thank to express my gratitude to my family, its more than that. I dedicate this small gift to them.

Yours,

Geo Velikkakam James

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 keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; noch nicht veröffentlicht worden ist sowie, dass 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 Dr. Korbinian Schneeberger betreut worden.

Date:

Koln: Geo Velikkakam James

Curriculum Vitae

Geo Velikkakam James
Keldermansstraat 44
3067XL, Rotterdam
The Netherlands
E-mail: geovjames@gmail.com

EXPERIENCE SUMMARY:

Company	:	Rijk Zwaan Zaadteelt en Zaadhandel B.V., The Netherlands
Designation	:	Researcher
Duration	:	November 2013 – present

Institute	:	Max Planck Institute for Plant Breeding Research, Cologne, Germany
Designation	:	Ph.D. Scholar
Duration	:	May 2010 – October 2013
Lab	:	Genome Plasticity and Computational Genetics

Company	:	KeyGene N. V., Wageningen, The Netherlands.
Designation	:	Trainee (Master thesis)
Duration	:	6 months
Lab	:	Field Crops & Bioinformatics lab

Company	:	MAHYCO Life science research centre, Jalna, India.
Designation	:	Research Associate
Lab	:	Molecular Breeding and Applied Genomics lab
Duration	:	June 2006 to July 2008

PUBLICATIONS:

Hartwig B., **Velikkakam James G.**, Konrad K., Schneeberger K., Turck F.: Fast Isogenic Mapping-by-Sequencing of Ethyl Methanesulfonate-Induced Mutant Bulks. Plant Physiol. 2012, 160:591–600.

Nordström K.J.V.*, Albani M.C.*, **Velikkakam James G.**, Gutjahr C., Hartwig B., Turck F., Paszkowski U., Coupland G., Schneeberger K.: Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. Nat. Biotechnol. 2013, 31:325–330.

Velikkakam James G., Patel V., Nordström K.J.V., Salome P., Weigel D., Schneeberger K.: User guide for mapping-by-sequencing in *Arabidopsis thaliana.* Genome Biol. 2013, 14, R61.

Wijnker E.*, **Velikkakam James G.***, Ding J., Becker F., Klasen R.J., Rawat V., Rowan B.A., de Jong D.F., Bastiaan C.S., Zapata L., Huettel B., de Jong H., Ossowski S., Weigel D., Koornneef M., Keurentjes J.J.B., Schneeberger K.: The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. In press at eLife.

* Authors contributed equally

ACADEMIC CREDENTIALS:

Year	Degree	University/Board	Class	Remark
2010-present	Ph.D.	Max Planck Institute for Plant Breeding Research, Cologne		IMPRS fellow
2008-2010	M.Sc. Bioinformatics	Wageningen University, Netherlands	First	Avg 8/10
2003-2006	B.Sc. (Biotechnology, Microbiology, Chemistry)	Nagpur University, MH, India	First	College topper
2001-2003	H.S.C (Computer science).	Department of public examination, Kerala, India	First	
1991-2001	S.S.L.C.	Board of public examination, Kerala, India	Distinction	School topper

AWARDS AND HONORS:

- ✓ International Max Planck Research School (IMPRS) fellow 2010 - 2013
- ✓ Anne van den Ban Fund (ABF) scholarship 2009 - 2010
- ✓ 2009-2010 board member of Wageningen Student Organisation (WSO)
- ✓ 2008 member of OpCie (Biotechnology), Wageningen university
- ✓ Twice PCM scholarship fellow

PERSONAL PROFILE:

Name	:	Geo Velikkakam James
Father	:	James V. J.
Date of birth	:	15/01/1986
Gender	:	Male
Marital status	:	Married
Spouse name	:	Shushimita
Nationality	:	Indian
Languages known	:	English, Hindi & Malayalam
Permanent address	:	Velikkakam House Pariyaram P. O. Chalakudy, Thrissur Kerala, India. PIN 680721

Cologne **Geo Velikkakam James**