Whole genome analysis of *Arabidopsis thaliana* using Next Generation Sequencing

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Dipl.-Bioinf. Korbinian Schneeberger

aus München

Tübingen

2010

Tag der mündlichen Qualifikation:	16.03.2011
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Detlef Weigel
2. Berichterstatter:	Prof. Dr. Daniel Huson

ERKLÄRUNG

Hiermit erkläre ich, dass ich diese Schrift selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die im Wortlaut oder dem Sinne nach anderen Werken entnommen sind durch Angaben der Quellen kenntlich gemacht wurden. Die Doktorarbeit besteht teilweise aus Auszügen eigener Publikationen oder Publikationen, die derzeit in Vorbereitung sind.

Tübingen, Oktober 2010
Korbinian Schneeberger

ACKNOWLEDGEMENTS

I would like to thank Detlef Weigel for his unrestricted support within the last four years and for various, different lessons I learned under his supervision. I am further indebted to Daniel Huson for supporting my graduation.

The are many, too many to list all of them, reasons to thank my outstanding colleagues Beth Rowan, Christa Lanz, Claude Becker, Dani Bezdan, Felipe Felippes, Felix Ott, Gunnar Rätsch, Jörg Hagmann, Joffrey Fitz, Juliane Damaris Klein, Jun Cao, Lisa Smith, Marco Todesco, Markus Schmid, Norman Warthmann, Patrice Salmone, Richard Clark, Roosa Laitinen, Sascha Laubinger, Sandra Gesing, Sebastian Bender, Stig Uggerhøj Andersen and Vinicius Galvao. But in particular, I would like to thank Stephan Ossowski for our outstanding collaboration on the development of Next Generation Sequencing applications and our numberless discussions inside and outside the lab.

Last but truly not least I would like to thank my family Günter, Irmgard and Sebastian Schneeberger for all their support, and my lovely girl-friend Stephanie Bürger for her trust, love, and for not giving up in hard times.

Table of Contents

Abstract

As-of the end of 2010 it has become commonplace that Next Generation Sequencing (NGS) has revolutionized biology. Despite this it remains true that the advent of NGS reduced the costs of whole-genome sequencing tremendously. Soon even small labs will be able to afford sequencing of every genome of every species they like.

Nonetheless sequencing of genomes might be cheap, resultant sequence reads alone are mostly not informative. It requires an army of new methods to handle the requirements that Next Generation Sequencing data brings along.

In this thesis I present parts our efforts of the last four years to implement NGS technologies in plant whole-genome sequencing. First, I will outline how NGS read alignments against multiple reference sequences simultaneously can be performed efficiently and how this affects the outcome of whole-genome sequencing. Afterwards I will show how reference sequence guided assembly can further improve the reconstruction of genomic sequences. And finally I will summarize how we adapted our computational methods to plant genetics to pinpoint genomic disruptions that were causal for previously identified phenotypes.

Abstract (deutsch)

Gegen Ende des Jahres 2010 ist es schon zum Allgemeinplatz geworden, dass Next Generation Sequencing (NGS) die Biologie revolutioniert hat. Doch das ändert nichts an der Tatsache, dass seit der Einführung von NGS die Kosten für Genom-Sequenzierungen drastisch gesunken sind. Schon bald werden auch kleine Labore in der Lage sein jedes Genom von jeder beliebigen Art zu sequenzieren.

Aber obwohl das Sequenzieren von Genomen billig ist, sind die resultierenden Sequenzen allein nicht hilfreich. Es bedarf schon einer ganzen Armee an vornehmlich neuen Methoden, um mit den Anforderung, die NGS Daten mit sich bringen zu Rande zu kommen.

In dieser Arbeit präsentiere ich Auszüge aus den Bemühungen aus den letzten vier Jahren, in denen wir versucht haben NGS Technologien für Pflanzengenom-Sequenzierungen nutzbar zu machen. Zuerst beschreibe ich, wie wir NGS Daten gleichzeitig, und trotzdem effizient, gegen mehrere Referenzsequenzen aligniert haben und warum so ein Vorgehen das Ergebnis von Genom-Sequenzierung verbessert. Danach zeige ich, wie man unbekannte DNA Regionen mit Hilfe von Homologie-basiertem Assembly rekonstruieren kann. Und zum Schluss fasse ich zusammen, wie wir unsere Methoden verändert haben, um genomische Modifikationen, die phänotypische Veränderungen zur Folge hatten, Ding fest zu machen.

1 Introduction

Apart of being the introduction for my thesis Detlef Weigel and I used the first chapter of this thesis as a working draft for a review article that was submitted to *Trends in Plant Sciences* at the end of 2010 and is currently under review.

1.1 Bridging the phenotype/genotype divide with millions of short reads.

In 1944 Oswald T. Avery concluded from his experiments on transformation of *Streptococcus pneumoniae* that DNA is the molecule that encodes the information that is passed on to future generations and therefore determines performance of the offspring [1]. This information destines the design, development and performance of cells, components and body plans of whole organism. Where some of the simple traits are exclusively determined by its underlying genotype, genotypes are not all-embracing determining complex traits and the subsequent fait of an organism. It is commonly accepted that genotype and environment orchestrate the shape of a phenotype, though it is still questioned to what extend.

Bridging the phenotype/genotype divide will help to understand this complex relationships that finds its practical importance in many different aspects of research and industry. Currently there are three methods dominating this field: genome-wide association studies (GWAS), quantitative trait locus (QTL) mapping and genetic mapping. The former is performed on natural populations; it utilizes millions of generations introducing recombination into this population and tries to correlate phenotypic expression to the genotypes. The latter are performed on artificial mapping populations that are usually only designed for this particular purpose. All three methods share the need for segregating polymorphisms, i.e. genetic loci that can act as markers to distinguish between the (parental) phenotypes.

Thus, the first challenge of bridging the phenotype/genotype divide is identifying variation that segregates between the genomes of individuals of a species. As a matter of course this task is tremendously simplified with the advent of Next Generation Sequencing (NGS) technologies that allows for whole-genome resequencing of hundreds of individuals. It can be seen as pure consequence that in the year 2008 the 1000 Genomes Project was launched, and followed by a similar project in plants, the 1001 *Arabidopsis thaliana* Genomes Project [2]. As-of 2010 the resequencing results of over 1,000 human genomes [3] and more than 100 plant genomes are publically available already (1001genomes.org).

By means of some exemplary selected work we will first review the efforts of the last years trying to decipher the complement of genomic variations in plant species and how technological advancements improved this filed. And in the second part we will then outline how whole-genome sequencing can be utilized to support plant genetics.

1.1.1 DNA Sequencing of *Arabidopsis thaliana*: From the beginnings until now

In 1986, only 14 years before release of the whole-genome reference sequence of *A. thaliana*, the first DNA sequence of one of its genes, ADH, was published [4]. Though detected by sequence homology to its maize paralog the final steps of gene cloning included tedious *de novo* determination of the actual gene sequence. This was tremendously simplified with the advent of the whole-genome reference sequences. Within the year 2000 the first plant genome was published and by then straightforward PCR cloning of essentially all genes of this species became possible [5].

This reference sequence was derived from a mono-genic sample of the most commonly used lab strain Col-0, and further allowed insights in whole-genome duplications, the consequential gene losses and the extensive local gene duplication rate. Interestingly, there was already some appreciation of the high levels of genetic differences between different strains. At the same time as the reference assembly was constructed a whole-genome shotgun sequencing of a second strain, Landsberg *erecta* (L*er)*, was performed and alignments against 82 Mb of the finished reference sequence were scanned for single nucleotide polymorphisms (SNPs), insertion and deletions (indels). In total the Arabidopsis Genome Initiative reported one SNP every 3.3 kb and nearly 15,000 indels sized from 2 bp to 38 kb. A significant proportion of the large indels contained complete gene sequences (not related to transposons) that were found elsewhere in the L*er* sequences, illustrating the high dynamics in the genomes of *A. thaliana* strains. Using the reference sequence and this set of L*er* sequences Jander and co-workers demonstrated the usefulness of this resources for map-based cloning and the enormous gain in speed that these two resources enable, reducing the total effort from three to five person-years to less than one person-year [5]. Ziolkowski and colleagues later studied the mechanistic origin of around ~8,500 of the large indels found within these sequences, distinguishing different mechanisms leading to indel polymorphisms [6]. Multiple hundreds of these indels overlapped with genes that were found to be expressed in Col-0.

The reference sequence allowed targeting genomic regions and thus the generation of more data focusing the aspects of natural variation. Primers throughout the whole genome allowed for targeted PCR studies of many individuals. Magnus Nordborg and colleagues analyzed 876 multiple alignments (covering 0.48 Mb of the reference sequence) scattered throughout the whole genome that were generated from a set of 96 *A. thaliana* individuals sampled throughout the northern hemisphere. For the first time they were able to use genomic sequence difference (revealed by these alignments) for a systematic survey investigating the genome-wide haplotype structure and allowing for population genetics across the genome rather than studying genome-wide averages. They demonstrated that the patterns of polymorphisms mostly agree with the expected distribution but they also highlighted that the presence of population structure shared worldwide genomic distribution of summary statistics deviate significantly from what was assumed by standard population genetics models. Their gene biased set of alignment revealed 1 SNP every ~190 bp [7].

A very comprehensive study of natural variation in *A. thaliana* was performed with resequencing microarrays at single base resolution. By hybridizing 20 accessions Clark and co-workers found more than 1 million SNPs [8]. The comprehensive design of the microarray included all theoretical possible single nucleotide variations. Tough tightly linked SNPs, however, suppress hybridization and SNP detection was confounded in highly diverged regions. The resultant absence of information was used to identify polymorphic regions without resolving the underlying sequence [9]. These studies were a major break through in polymorphisms detection but also stated the dawning of the era of microarray-based resequencing. With the advent of the first Next Generation Sequencing (NGS) platforms in the year 2006 a more detailed view on these complex regions became possible. The first resequencing studies in plants were performed with sequence reads that were not longer than ~40 bp but were sufficient for unambiguous alignments and a per-base readout of sequence polymorphisms. In the first resequencing study performed on plants Ossowski and colleagues reported one SNP every 200 bp [10]. But it required a more sophisticated local assembly to reveal over 10,000 diverged regions harbored indels as long as 600 bp. However, comparison with microarrays showed great overlap between the detection of deleted and/or diverged regions of both technologies [11]. The first part of the work that is presented here we utilized the presence of polymorphisms that are segregating in the population for the resequencing of the genome of *Arabidopsis thaliana* Est-1. In addition to the reference sequence we used known sequence variation as alignment target for the short read data [12]. Using a novel data structure we could show that detection of complex variation is not limited by alignment constraints but by the variation data included in the alignment targets.

Nevertheless assembly approaches hold the promise to be even more informative in higly diverged regions. As a second part of this thesis I will describe our efforts to assemble *A. thaliana* genomes. Using alignments against the reference sequence they reduced the complexity of the whole-genome shotgun data and assembled these subsets separately. Compared to the reference sequence the assemblies are of lower quality, nevertheless 50% of the genome assembly resides in contigs that are ~200kb or larger. These assemblies compromise the first *Arabidopsis thaliana* assemblies after the release of the reference sequence in the year 2000.

After NGS analysis was established the data yield of single sequencing runs increased steadily allowing for population-scale sequencing projects. A large scale whole-genome resequencing project for *Medicago truncula* (medicagohapmap.org) is on the way and recently the resequencing results for 30 soy-beans varieties [13] and six elite maize inbred lines [14] were reported. Resequencing in maize has revealed more than 1.2 million SNPs and over 30,000 indels. Sequence comparisons between lines revealed presence/absence variations (PAVs) of hundreds of expressed genes. The authors speculate that PAV might play an important role in heterosis in maize — outlining the immediate impact for breeders. Currently the most ambiguous community effort might be the 1001 *Arabidopsis thaliana* Genomes Project (1001genomes.org) describing its goal as the discovery of whole-genome sequence variation of 1001 strains. The first set of 80 genomes is released and allows new interpretations of the genome of this species. Unexpectedly high proportions of the genes harbor coding sequence disrupting polymorphisms leading to over 1,000 genes with deleterious changes per accessions.

1.1.2 Using NGS for computational genetics

After whole-genome sequencing was established the most straightforward application for mutant identification should be direct sequencing of the mutant genome. Nevertheless, as-of today we are not aware of any published analysis reporting a plant mutant identification by merely screening the mutant genome without exploiting any additional resources. Whole-genome mutant resequencing suffers from the confounding variations (either natural, mutagen-induced or both) that hitchhike in the background of the mutant genome or are introduced by the other accessory parental genotypes [15]. These changes need to be distinguished from the very limited number of causal changes (in mutagen-induced mutants this typically only one). Thus, the first attempts of whole-genome sequencing of mutant genomes were performed on mutants that were roughly mapped before hand. In 2008, Sarin et al were among the first to report eukaryotic mutant identification by whole-genome sequencing in a 4 Mb mapping interval of an ethyl methanesulfonate (EMS) induced *Caenorhabditis elegans* mutant [16], similar studies were reported for *Drosophila melanogaster* [17] and yeast [18], [19]. Even though EMS mutagenesis is common in plants too, one of the first direct whole-genome sequencing of a plant mutant was reported on an spontaneous mutation in a non-reference strain of *A. thaliana* [20]. This sequencing effort displays the first section of the last chapter of this thesis, which describes how whole-genome sequencing can be used to find the one causal change: After eliminating all natural variations found in the background genome only one change remained and that was found to be causal for the phenotype.

Except of work in yeast, mutant identification was based on rough prior knowledge about the location or architecture of the mutation. Within the second part of the third chapter I will illustrate how we developed a method that allows to simultaneous map and identify mutagen-induced changes without any prior knowledge about the mutation [21]. As this method drastically reduces the amount of work needed to map mutations, and as it is not specific for *A. thaliana* nor plants this method was later introduced to *C. elegans* as well [22]. Using whole-genome sequencing of a pool, so-called bulk segregant analysis (BSA), of phenotypically striking plants implemented genotyping into the sequence analysis and significantly reduced the time for mapping from months to days. Nevertheless this method requires generation of mapping populations and for some phenotypes this might be tedious or even impossible. Backcrossing mutant to the wild-type genome is typically less challenging and is usually performed to eliminate mutagen-induced mutations that blur the phenotype. In this process mutagen-induced nucleotide changes that are genetically linked to the causal mutation and physically surround it will remain, whereas unlinked nucleotide changes will be out-crossed. Consequently, mutagen-induced variants will be enriched in the vicinity of the causal mutation. It has been shown in *C. elegans*, though should be applicable to other systems as well, that such an enrichment generated by 4-6 back-cross generations can generate a significant pattern in whole-genome sequencing allowing to define mapping intervals as small as a couple of Mb and that this reduction of the search-space is enough to pinpoint causal mutations in reference strains [23]. Obviously combining these two methods and using the mutagen-induced changes not only as indication for linkage to the causal region but also as markers (like I will outline it in the last chapter), should reveal the skew in the frequency of the parental lines (the mutagenized and the non-mutagenized parents) in addition. This method would combine the advantages of both methods: no need for out-crossing and only two rounds of backcrossing. Nevertheless it still needs to be proven to be feasibly, especially the expected low number of mutagen-induced changes serving as markers could confound the analysis.

Many interesting and agricultural important traits deviate from Mendelian segregation but rather show phenotypic expression at continuous levels. The complexity of the underlying genetic architecture requires large mapping population to dissect all loci with varying effects on the expression of the trait. Though in many cases genotyping and phenotyping on a sufficient scale will not be feasible without pooling individuals [24]. Pooling of those plants with similar expression of phenotypes in order to fix major effect QTLs, bulk segregant analysis (BSA), was already introduced around 20 years ago and was already successfully used for mapping of major effect QTL and Mendelian loci, e.g. [25]. This method was advanced and tedious genotyping of individual plants was replaced by rapid genotyping of thousands of markers using microarrays [26], [27]. Though pooling of individuals allows handling of hundreds of individuals this still displays a limited sample size and implicates the lack of sufficient statistical power for the detection of small effect loci [28]. Recently Ehrenreich and co-workers overcome this problem by drastically increasing the sample sizes in yeast [24]. By further extending BSA of extreme traits with whole-genome sequencing to measure parental allele frequencies throughout the whole genome they dissected quantitative traits with 20 major and minor effect loci.

It remains to be proven if such large pools of recombinants can established in crop species as well. Like in conventional genetics mapping whole-genome sequencing individuals would ease dissection of complex traits as single recombination events can be correlated with individual (quantitative) phenotypes. Besides its feasibility whole-genome sequencing is still rather expensive for large numbers of individuals. Recent advancements in the complexity reduction of such genomes and technological advancements in sequencing library preparation carry promises to apply whole-genome sequencing on hundreds of individuals. Sequencing restriction-site associated DNA (RAD) tags reduces the representation of genomes and thus allows not only the for absence/presences variation detection of restriction sites but also for polymorphism mining in homologous regions of multiple genomes - without sequencing the whole genome [29]. Combined with bar-coding, i.e. labeling the DNA sequence reads for their donor genome in order to divide the sequencing power of NGS to multiple individuals hundreds of genomes can individually genotyped [30], [31]. Additional genotyping of the parental lines the genotyping information of RAD sequencing can immediately be used for genetic mapping even in complex genomes, as demonstrated for stickleback and *Neurospora crassa* [29]. Interestingly, this method can be extended to systems without reference sequence and thus introduces an eased way to genetic mapping in non-model species such as many of the crop are. Rather than aligning the sequence reads to a reference sequence clustering based on sequence similarity between the reads was shown to be sufficient to identify markers with complete linkage to one of the parental lines [29].

Instead of analyzing homologous regions Huang et al reconstructed the parental mosaics of 150 rice recombinant inbred lines (RIL) by whole-genome shotgun sequencing using bar-coded samples [32]. Using a sliding window approach Huang et al assessed the genotypes of each of the rice RILs. With only ~0.02x genome coverage per RIL on average a resolution of recombination breakpoints of 40 kb was reported. Using NGS improved this analysis compared to similar projects performed with array technologies for *A. thaliana* [33] and rice [34]. Microarrays genotype at predefined sites that might not work to distinguish every given pair of individuals, whereas NGS technologies analyzes a genome without assumption and additionally allows for bar-coding and thus a dramatic decrease in costs compared to arrays that require one hybridization per individual. Xie and co-workers extended this method to be parent-independent by assessing the parental genotypes from the RILs based on a maximum parsimony method that prefers lower number of recombination [35].

So far the lack of reference sequences, polyploidy and mere genome size prohibited large-scale resequencing of many of the crop species. Possibilities to reduce the sequenced genomic space using target enrichment sequencing promise nothing less but the application of NGS-based genetics even for very large genomes. Various different ways to enrich sequencing libraries have been introduced (see [36] for a detailed review) and they all share the characteristics that they can reduce the sequenced space of a genome to predefined loci. Noteworthy, in order to prepare the enrichment targets it is not necessary to have near-to-complete knowledge about the whole genome sequence, already partial knowledge about the genomic sequence, e.g. EST or RNA-seq assemblies, can be accessed with target enrichment sequencing. Combined with bar-coded sequencing [37] target enrichment sequencing is expected to facilitate simple genetic mapping even in species with large genomes without reference sequences – as it is case for many of the important crop species.

Genome-wide association studies (GWAS) are another way to uncover the genetic variation that determines phenotypic differences. GWAS utilize the natural variation and the accumulation of recombination introduced in natural population rather than looking at artificial mapping population generated from pre-selected parental lines. Where later methods suffer from the little number of recombination that is introduced only by a very limited number of generations, GWAS is statically challenged by population structure, genetic heterogeneity and – so far – by the lack of resources for development of high-density haplotype maps and the drastic effort needed to generate such genome-wide variation data.

Introduced in human genetics [38] and recently applied to plants for the first time [39], GWAS in other than model species lack resources for development of high-density haplotype maps. Besides the community-wide efforts in maize [40], soybean [13], *M. truncula* and *A. thaliana* [2] NGS can overcome the costly generation of individual genotypes and clearly constitutes a step forward compared to microarray-based generation of HapMap type of resources [8], [41].

Huang and colleagues were first to demonstrate the practicability of GWAS to crop species by studying 14 agronomic traits in over 500 rice genomes [42]. Bar-coded sequencing to an average coverage of 1-fold revealed around one forth of the genome of each individual. To compensate for missing genotypes Huang et al applied a novel imputation method to fill missing genotypes and generate a detailed haplotype map using approximately 3.6 million SNPs that represent ~80% of the worldwide genetic diversity. Out of 80 loci that were found to be associated with 14 traits six were closely linked to traits that recently had been identified. On average the genetic diversity explained 36% of the phenotypic variation – more than recently reported for human. Surprisingly, Huang and his co-workers observed SNPs linked to causal loci featuring higher association than causal SNPs themselves as similarly described in *A. thaliana* before [39], [42]. They speculate that this might be reflected by multiple causal alleles of the same gene code for identical phenotypes. Hence, association studies combining the genotypic variation by their predicted effect on gene function rather than associating single variations is presumably worth exploring. Prerequisite of this, the access to structural variations and near-to-complete genomic sequences of all individuals can only be achieved with whole-genome sequencing of all individuals, rather than microarray-based genotyping. Advancement in methods for genome assembly and sequencing technologies promise this for the near future.

2 Aligning short reads against multiple reference sequences

This chapter describes our efforts to design and implement a short read alignment tool that can align short reads against multiple references simultaneously. This work was published in Genome Biology in 2009 [12] and has been labeled as "highly accessed" after only a month past publication. I implemented GenomeMapper together with Jörg Hagmann, a diploma student under my guidance. Stephan Ossowski and I preformed the resequencing analysis of *A. thaliana* Est-1 as a proof-of-principle, for which Norman Warthmann had prepared the sequencing sample. Sandra Gesing and Oliver Kohlbacher implemented a parallel version of GenomeMapper. Gunnar Rätsch, Geraldine Jean, Fabio de Bona, Uta Schulze, Bettina Hepp, Sören Sonnenburg, Lisa Thalheim, Dominik Diesch, Andre Kahles, Jörg Hagmann and I merged the source code of GenomeMapper with that of QPLAMA [43] introducing spliced alignments into GenomeMapper.

2.1 Resequencing depends on similarity between the reference sequence and the focal genome

Genome resequencing with short reads generally relies on alignments against a reference sequence. Even though this reference sequence might be generated of genomes of multiple individuals it will always suffer of its linear characteristic and its inability to capture and display allelic variation present within the respective species. We have developed a new alignment algorithm *GenomeMapper* that supports simultaneous alignments of short reads against multiple genomes by integrating related genomes (e.g., individuals of the same species) into a single graph structure. This constitutes the first alignment algorithm for handling multiple references.

To by-pass this problem of having no alignment in regions of high divergence partial de novo assemblies of targeted regions have been attempted [10]. Even though this helped to bridge some of the regions that could not be revealed by alignments alone the recall of such regions was limited and regions longer than a couple of hundred bp could not be dissected at all. Over 80% of the targeted regions could not be analyzed with this approach, i.e. the largest indel was not more than 641bp [10]. Thus short read analysis of complex genomes is greatly aided by using a sequence backbone that features most of the alleles variations inherent in the focal genome. But, of course, as the genome sequence of the focal genome is usually unknown, it would be desired to align all short reads against all known variations simultaneously.

Further, we note that the information derived from resequenced individual genomes is in itself useful for subsequent resequencing efforts, especially when the latter are at lower sequence coverage than the earlier efforts. Incorporating known polymorphisms increases the genome space against which the sample reads are aligned, which should greatly improve the mapping results. For example, an alignment suggesting a string of deleted bases in the focal genome becomes much more reliable if this deletion is known to exist in the population. The incorporation of such missing or inserted bases in the reference sequence would not only decrease the complexity of the alignments, but also reduce sequencing costs, as more reads can be placed on the genome.

Apart of such practical reasons, aligning against only a single reference biases the analysis towards a comparison within the sequence space conserved with the reference. Taking into account all known genome variants would reduce this bias. Aligning reads against multiple genomes separately increases computation time and storage space and introduces new problems of merging and interpreting redundant results.

2.2 Multiple genomes in one index

One way to decrease runtime for the generation of sequence alignments is to build index structures of either the reads or the reference sequence. We implemented later strategy in GenomeMapper. To allow for simultaneous alignments against multiple genome sequences, all target sequences have to be combined into one data structure. GenomeMapper achieves this goal by building a joint index of all genomes that are alignment targets. This index will be persistently stored and, once compiled, the index does not need to be rebuilt for future alignment tasks.

The index is a simple hash-based mapping of k-mers (sequence signatures of 5 - 13 bp) to their locations within the target sequences. Each k-mer present in target sequences is unambiguously converted into a single integer, applying a two-bit representation of the four DNA nucleotides. Each hash key points to one hash value consisting of a list of all genome locations of the k-mer. While this rather simplistic hash indexing approach has some disadvantages compared to more recently developed strategies, e.g., Burrows-Wheeler indexing [44], the latter are usually geared towards ungapped alignments and are not easily extendable to non-linear structures imposed by multiple genomes. Further, spaced-seed approaches, implemented in tools such as SHRiMP or ZOOM, can be more sensitive [45]. However, when these approaches are applied to real data, they do not result in a substantial increase in the number of alignments compared to an approach with contiguous seeds followed by a complex alignment, because contiguous seeds are usually chosen short enough, i.e., 9 to 12 bp, for anchoring and subsequent aligning of reads (see below for comparison with other mapping tools).

Mapping indices tend to require a large amount of random access memory (RAM). Current compute servers usually allow multiple processors to share physical RAM. To avoid the unnecessary overhead of loading the same index multiple times, GenomeMapper makes use of memory-mapped files, allowing computer processes to share the same index structure within the memory. This reduces the overall memory footprint when running several instances of GenomeMapper in parallel.

The input for GenomeMapper's index creation step consists of the sequence of one of the genomes and a list of differences in the other genomes compared to the first one, i.e., one FASTA file and a list of single nucleotide polymorphisms (SNPs) and indels of every additional genome. Each position not explicitly annotated as different is assumed to be identical in all of the genomes, and will therefore only be stored once. This is important to avoid redundant alignments to several genomes. Divergent sequences are stored separately for each of the genomes. Identical regions, which are represented once, need to be connected with polymorphic regions, which are represented by branches in the index. Hence the reference looses its linear/sequential characteristic, but rather forms a sequence graph. Note that none of the genomes represents "the reference" anymore (Figure 2-1A).

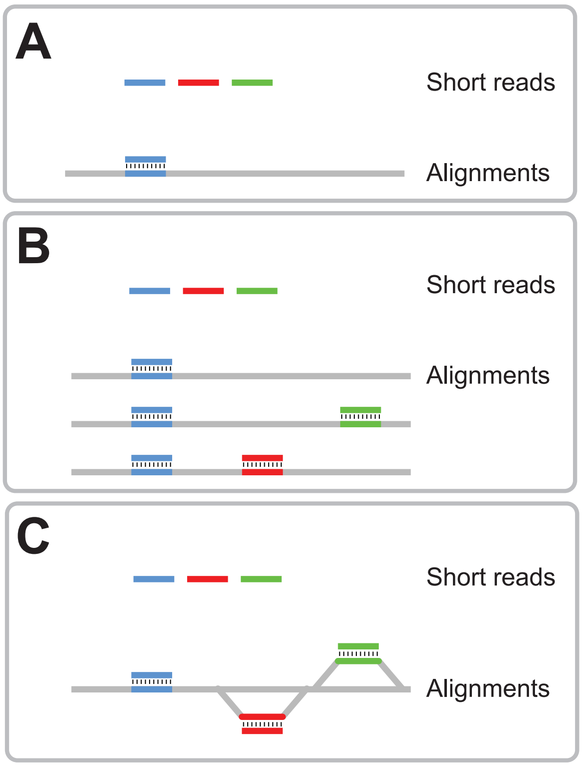

Figure 2-1: **Efficient alignments against multiple genomes.** (A) Only reads that are sufficiently similar can be aligned against a single reference. (B) Separate alignment against multiple genomes allows access to divergent regions, but result in redundant alignments of reads that match all targets (blue). (C) Alignments against a graph index representing multiple genomes provide access to divergent regions without redundant alignments.

In order to store this information efficiently, each of the genomes is partitioned into non-overlapping sequence blocks of up to 256 bp, which represent the genomic sequence of all genomes. The connections of blocks to their neighbors allow for continuous reconstruction of each genome. Invariant regions will be represented by one block only. Every variant, including all SNPs, will trigger the formation of branches, which constitute the parallel blocks that account for the non-linearity of the genome graph (Figure 2-2AB). Since complex differences such as inversions or duplications can always be defined as combinations of deletions and insertions, they can be readily incorporated into a graph index.

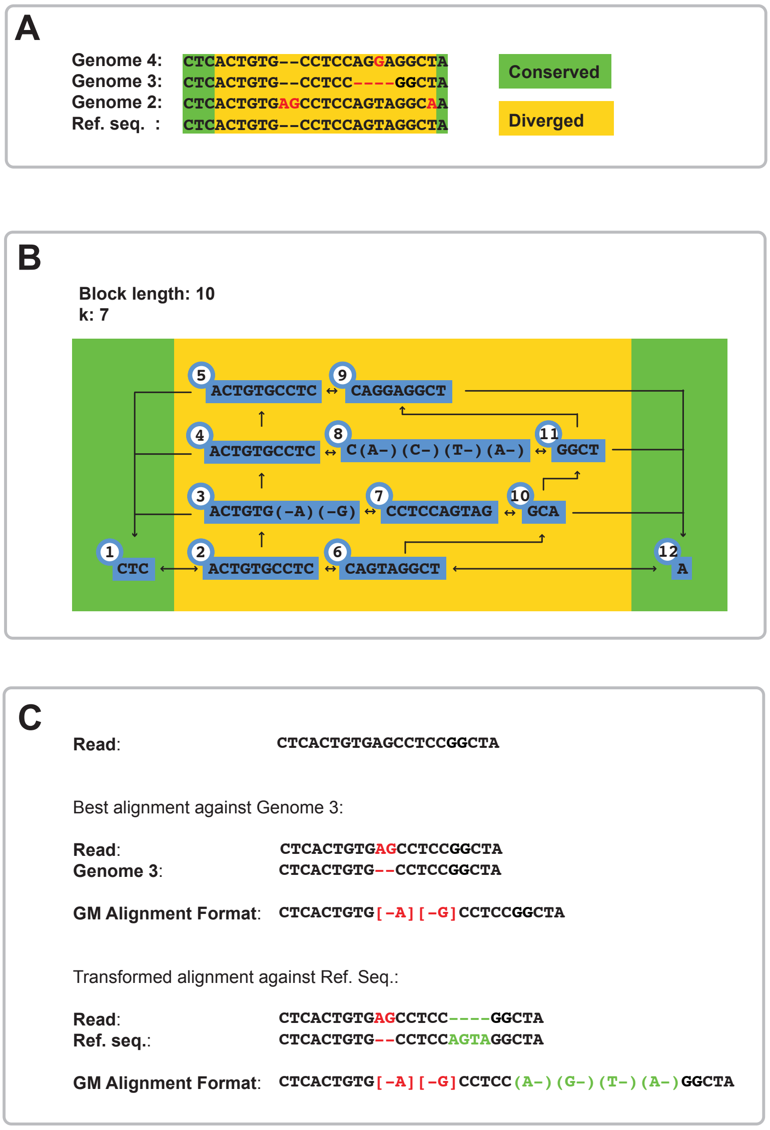

Figure 2-2: **GenomeMapper's graph index structure.** (A) Examples of orthologous sequences in four divergent genomes. Sequences at the beginning and end of each fragment are shared (underlain with green boxes). Divergent regions start k-1 positions (in this case 6 positions) before the first true variable position, to account for the k-mer length used for the hash key calculation. (B) Graph structure created by these sequences, with k-mer length 7, and maximal block length of 10 (instead of 256) for reasons of illustration. The number attached to each block is its unique identifier. Note that blocks do not occupy their maximal block length after an indel, exemplified by blocks 3 and 8. Blocks 1 and 12 correspond to sequences identical in all four genomes, and are present only once in the index structure. Arrows between the blocks visualize the edges between the nodes in the genome graph, as they are stored in the block table. (C) Alignment of a read against the most similar genome, Genome 3, with a 2 bp insertion. Although the insertion is also observed in Genome 2, the 4bp deletion downstream in Genome 3 makes the read more similar to it than Genome 2 does. The transformed alignment of the read against the original reference sequence (Ref. seq.) includes the 4bp deletion (as supported by Genome 3) given in parentheses (green), whereas the 2bp insertion (which is neither supported by Genome 3 nor the reference sequence) is annotated like a mismatch using square brackets.

A unique identifier for each block allows for a constant look-up time in a table that stores all relevant block information. In addition to referring to the genomes, which it is a part of, each block encodes for its sequence, the connections to its neighboring blocks and the position within the genome. Each block thus harbors the genome sequence of all or a subset of genomes with identical sequences within the respective region. The block table is the implementation of a sequence graph, where the blocks represent the nodes and the connections between them the edges (Figure 2-2B). From now on we refer to this table as genome graph. A comprehensive list of all features stored in the block table is given in Table 2-1.

Feature	Size in bytes	Description
genome_pos	4	Start position of the block with respect to the sequence it is derived from.
ref_pos	4	Start position of the block with respect to the reference sequence.
Chr	4	Chromosome identifier
Strain	4	Genome identifier
Indel_offset	2	Stores the position of the indel within the block
ins_pos	4	Stores the number of inserted bases upstream.
Seq	256	Genome sequence, stored as transformed alignment.
prev_block	4	Block identifier of preceding block
next_block	4	Block identifier of succeeding block
next_strain_front	4	Block identifier of the first block of the next diverged strain (in diverged regions only). This is the interconnection between the strains in Fig 2 and 4.
next_strain_end	4	Block identifier of the last block of the next strain (in diverged regions only)

Table 2-1: **Features stored in block structure.** Each block is represented by one such block structure in the block table. Since the block table is an ordered list, the blocks do not need to store their identifier, as this is encoded as their positions in the list (block table).

In order to show that storing genomes separately is more memory consuming than building a genome graph we generated genome graphs incorporating the information of one, two, or three strains in addition to the reference. A graph featuring only the reference genome has 100% of conserved sequence. Incorporating a single additional *Arabidopsis thaliana* strain reduced the conserved genome space by about 5%, namely to 94.3% (for Bur-0) or 95.1% (for Tsu-1).

Simultaneously adding both Bur-0 and Tsu-1 resulted in a further decrease of conserved genome space, but by less than 4%, to 91.6%. Adding a third genome, Est-1, reduced conserved genome space by less than 2%, to 89.9%.

Since all relevant information is stored in the genome graph, the positional information attached to each k-mer in the hash described above (linking each k-mer to its locations in the genome), must merely store the block identifier (represented by 3 bytes) and the position within the block (1 byte). Based on this information the position of every base within each of the genomes can be inferred. The 4 byte encoding accommodates a combined length of all unique sequences of up to 4 Gb.

Efficient read mapping requires that each k-mer generated from one of the sequences in the genome graph can be queried for its locations in a time linear to the number of hits. This is achieved by building a hash table connecting the k-mer (hash key) to its positional information in the genomes (hash value). Each hash key refers to a list of entries. Each of these entries stores a block identifier and a block position, allowing for a unique positioning of each k-mer.

2.3 Need for complex alignments

Earlier studies showed that in a random comparison of two natural Arabidopsis strains, there is typically one SNP every 200 bp. In addition, using early-generation Illumina single reads, over 60,000 small indels (1-3 bp) and 10,000 indels of up to several hundred base pairs have been detected in two strains, presenting a lower bound for the degree of polymorphism in this species [10].

Mismatches in alignments result not only from sequence differences, but also from sequencing errors. The error probability of Illumina sequence reads has been shown to be less than 1% for most, but not all parts of the read [10]. In comparison to the rate of natural variation in Arabidopsis, mismatches from errors in individual reads outnumber true SNPs approximately 17 to 1, while true gaps are almost as frequent as gaps resulting from sequencing errors. In order to distinguish between sequencing errors and real polymorphisms, we separated all mismatches (and gaps) at positions resulting in SNP calls (or indel calls) from the mismatches (and gaps) at positions with a high confidence reference call (based on the single reference alignments of the Est-1 reads against the reference sequence). The outcome of this analysis depends on the divergence between sample and reference sequence, as well as on the quality of the sequencing run. It does not allow for general conclusion about the Illumina GA sequencing technology.

To avoid misplacement of individual reads, some mapping tools favor alignments where the cumulative base quality of mismatching bases is low [46]. With respect to the high level of natural differences in Arabidopsis, such a strategy could bias alignments away from polymorphic regions. GenomeMapper instead performs for each read an alignment based on dynamic programming similar to the Needleman-Wunsch alignment algorithm [47].

Our method ensures that all alignments within a given number of mismatches and gaps are reported, provided that they share at least one identical substring of length k when using a k-mer index. No other constraints are imposed on the number of mismatches, gaps or base call quality. By default, GenomeMapper aligns against all instances of a repeat, but it also can be instructed to align only against a subset of them.

In our experience, resequencing projects of bacterial or medium-sized eukaryotic genomes such as those of Arabidopsis strains do not benefit from utilizing alignments other than the optimal ones. Nonetheless, GenomeMapper can be configured to report not only the best scoring alignments, but also all hits within the specified range of mismatches and gaps (all-hits instead of best-hits strategy). As expected, this comes with an increase in runtime, especially for highly repetitive genomes.

2.4 Aligning sequences against the graph

GenomeMapper's alignment procedure is partitioned into three steps including speed optimization. The optimization bypasses the costly calculation of alignment matrices without decrease in sensitivity and is based on two observations: first, a dynamic programming alignment is only required if the best alignment involves gaps; and, second, the frequency of gaps is lower than that of mismatches. This is the case for both sequencing errors in Illumina reads and true polymorphisms. To cope with this, GenomeMapper applies a higher penalty for gaps than for mismatches. Therefore alignments with a penalty lower than the gap penalty do not require dynamic programming. The optimization cannot be applied in an all-hits strategy including gapped alignments, and will not increase speed if the best alignment features gaps.

In the first step of the alignment procedure, GenomeMapper scans the hash index for k-mers identical between read and genome graph to quickly detect all genomes and locations with nearly identical alignments. In the second step GenomeMapper determines the location and sequence of nearly identical maximal substrings (NIMS) between read and genome graph. GenomeMapper will finally perform a k-banded alignment applying dynamic programming to ensure a consistent gap placement.

The following describes these three steps in detail.

GenomeMapper starts by calculating the hash keys for a subset of non-overlapping k-mers of the read sequence and retrieves their genomic positions from the hash index. The selected k-mers are non-overlapping k-mers at the read positions 1, k+1, 2k+1… The last k-mer will be calculated starting at read length-k positions. The pair consisting of a k-mer along with one of its positions in the genome will be referred to from now on as *hit*. Each hit will then become target of a simple alignment algorithm comparing the read and genome sequence surrounding the hit, not allowing for any gaps. The rationale behind this approach is to identify all alignments, which feature one mismatch less than non-overlapping k-mers fit into the actual read (compare to q-gram lemma).

If the best alignment of a read contains up to one mismatch less than the number of non-overlapping k-mers fitting into the read, at least one hit within this alignment can be computed [48].

If the first step does not reveal a valid alignment, which is always optimal due to the pre-requisite that one mismatch is less penalized than one gap, GenomeMapper starts calculating hits not only for a subset, but for each of the k-mers within the read sequence. If two hits are adjacent in the read and in the genome graph, they will be merged resulting in so-called *extended hits*. If a single mismatch between read and genome sequence is adjacent to extended hits on either side, GenomeMapper can bridge this mismatch by merging the extended hits, now harboring this mismatch. Once all hits are maximally extended (they now constitute NIMSs), the read has to be aligned against the regions determined by each of the

NIMS, aborting as soon as the best possible alignment will be worse than the mismatch and gap constraints.

To retrieve the genomic sequence for the alignments, GenomeMapper needs to follow the links between blocks. Starting from the block harboring the hit or NIMS, respectively, GenomeMapper follows the edges of the genome graph to generate a target sequence for the alignment. If multiple blocks reside next to one of these blocks, each of the branches will generate a separate target sequence for an independent alignment. Note that GenomeMapper will not concatenate sequences from different genomes. The alignment phase is implemented with an efficient parallelization, which substantially reduces runtime. It is distributed in a master-slave model on shared-memory architecture. All alignment threads can access the genome data and the read data. The master thread distributes individual hits by signaling each alignment thread and collects the results. The number of threads used by the parallel implementation is a user-defined parameter that can be adjusted to the hardware. The parallel version of GenomeMapper relies on POSIX threads to efficiently manage the individual compute threads. POSIX threads are available for all relevant platforms (including Linux, Mac OS, and Windows).

GenomeMapper employs a k-banded dynamic programming alignment. The alignment constraints of the upper limit of gaps implicates that all valid traceback paths will not involve alignment matrix cells which are more than k vertical or horizontal trajectories distant from the main diagonal. Thus, only k cells on either side of the diagonal have to be computed shaping a band with k being the number of allowed gaps.

GenomeMapper's alignment method is similar to the Needleman-Wunsch algorithm [47]. The only differences are modified traceback rules and the introduction of an abortion criterion. Since reads are aligned not only against the genomic region determined by NIMSs, but against an enlarged genomic region in order to allow for gaps, the start and end of the traceback routine is limited to the cells in the last and first row of the k-band and not only to the bottom-right and top-left matrix cell, respectively, as required by the Needleman-Wunsch algorithm.

If the number of allowed edit operations is exceeded before the alignment is finished, GenomeMapper will stop the computation. As the alignment matrix is filled column-wise and penalties are non-negative, its computation is aborted as soon as the best score of all cells in each column is worse than the score resulting from the maximal number of allowed gaps and mismatches.

Furthermore, if GenomeMapper runs in best-hit mode, the maximal number of edit operations is not only defined by the global adjustments as set from the command line, but are further restricted by the best alignment found so far. Every best alignment will update the restrictions for the upcoming alignments of the same read. Every alignment meeting the constraints of mismatches, gaps and edit operations is stored and ranked by its score. The best-hit mode will only report the alignments with the highest score.

2.5 Alignment representation

Independent of the algorithm used to detect the best alignments GenomeMapper will report two different representations of the alignment. The first one constitutes the alignment of the read against the genome it is most similar to (reference-free alignment). Because commonly used tools for alignment consensus analysis such as MAQ [49], Mosaik [50], SHORE [10], VAAL [51] report base calls based on the location relative to one reference sequence, GenomeMapper implements a second alignment representation, which transforms the strain alignment into an alignment against the reference sequence. This reference-based alignment can then be used as input for one of the tools mentioned above. Which of the genomes constitutes the 'reference sequence' is defined in the index creation. As the reference sequence is not necessarily the most similar sequence to the read, the reference-based alignment can feature more mismatches and gaps than the strain alignment and can exceed the user defined constraints.

This transformation generates two categories of mismatches in the reference-based alignment. The first category contains mismatches that are unique to the read sequence. The second consists of mismatches identical between the read and the strain it was aligned to, but different from the reference sequence. Such mismatches are more likely to represent true polymorphisms, since they have already been previously observed. GenomeMapper indicates the different types of mismatches using round and square brackets (Figure 2-2C).

An alignment is typically anchored by the position of the 5' nucleotide in the target sequence at which the alignment starts. Since different genomes may feature indels of different lengths, however, even for identical sites positional information can become ambiguous. The decision for one of the locations only, e.g., that of the reference genome, would overvalue the reference.

Currently the sole community-wide accepted description of a genomic location is the corresponding nucleotide within the reference sequence, which easily accommodates gaps, but not insertions relative to the reference. We therefore implemented two position descriptors into GenomeMapper. The first refers to the particular genome against which the alignment was performed (the strain alignment). The second represents the position of the alignment against the reference (the reference alignment). Insertions are annotated using the upstream reference position followed by the position of the inserted nucleotide within the insertion, separated by a decimal point (e.g., "80359.12" describes the 12th nucleotide within the insertion after position 80359 of the reference). Strain alignments transformed to reference alignments lose their reference-free characteristic and therefore are immediately comparable to conventional mapping results.

SHORE's consensus algorithm works on a proprietary input format (called map.list) representing read ID, alignment and alignment locus, repetitiveness as well as read and alignment qualities. We have extended this format to store both the alignment of a read against the reference and against the best matching strain.

The consensus prediction including reference-like, SNP and indel positions is performed identical to the decision tree algorithm used for single reference alignments described in an earlier work [10], with two exceptions: First, SNPs and deletions that are known from previously sequenced genomes have a decreased threshold for minimum coverage (1 instead of 3). See later section why this still allows for fair comparisons in the quality performance comparisons. Second, the limit for the length of deletions and insertions (equal to maximum allowed gaps) is removed completely. For insertions longer than the read length this means that covering reads can have no anchor in the reference sequence anymore but are completely contained within the insertion. In order to approximate the concordance of the consensus call for long insertions (reads supporting the insertion divided by all reads covering a position), we compare the maximum read coverage within an insertion to the total number of reads not supporting the insertion. This issue does not apply to deletions, because deletions are always spanned by a read and therefore concordance can be accurately calculated similar to reference or SNP calls.

'SHORE consensus' can be configured to run in graph mode with a single parameter. The output format is identical to the one of the 'single reference' consensus.

2.6 Performance and quality comparison to other mapping tools

A major requirement for the practical relevance of short read alignment tools is their runtime. Different approaches for fast mapping of short reads have been suggested, including methods for indexing substrings of either the short reads or the reference sequence with the use of k-mers or spaced seeds (academic tools such as Bowtie, BWA, CloudBurst, MAQ, MOM, MosaikAligner, mrFAST, mrsFAST, Pash, PASS, PatMaN, RazorS, RMAP, SeqMap, SHRiMP, SliderII, SOAP, SOAP2, ssaha2 [10], [52], [53], [54], [55], [56], [44], [57], [49], [58], [59], [46], [60], [61], [62], [63], [64], [65], [66] and commercial tools such as ZOOM [67]). Further, it has been reported that the current demand for rapid alignments can be met with new indexing strategies [44]. However, this is normally at the cost of not allowing complex alignments including gaps. For natural inbred strains of *Arabidopsis thaliana*, the high level of individual differences constitutes a substantial challenge. It has been estimated that several percent of the reference genome are either missing or very divergent in other strains of this species, which features homozygous genomes that are 25 times smaller than a haploid human genome [8], [9]. This results in regions inaccessible to simple short read alignments, in particular for alignment algorithms that do not accommodate many mismatches and gaps. Based on these requirements we have set up a performance and quality comparison of GenomeMapper and other short read alignment tools.

As GenomeMapper can also be used for alignments against a single target genome, we could compare runtime and sensitivity of GenomeMapper (version 0.3.1s) to SOAP (version 1.11 [58]), soap2 (version 2.01 [59]), bowtie (version 0.9.8 [44]) and MAQ (version 0.7.1 [49]). SOAP and MAQ have previously been compared to bowtie

[44], but with a human target. Here we aligned against the *Arabidopsis thaliana* Col-0 reference genome [68] with seed length set to 12. All tests were performed on 10 independent read sets, each consisting of 500,000 reads randomly sampled from reads generated in this work for the *Arabidopsis thaliana* Est-1 strain. We tried to run all alignment tools with optimal parameters to achieve the best possible sensitivity and runtime (Table 2-2). To be directly comparable with GenomeMapper, we set SOAP, soap2 and MAQ to report all repetitive best hits rather than a random subset of them, even though this comes with an additional investment in runtime. All tests were performed on a compute server with 8 cores (two AMD Opteron quad core processors) and 32 GB RAM. Figure 3 compares average runtimes, measured as the wall clock, as well as sensitivity of all alignments and of gapped alignments, both measured as the number of reads which could be aligned. As this analysis is based on real data for which no gold-standard sequence information is available, nothing is known about the true origin of the DNA reads. We therefore took the fraction of aligned reads as a proxy for sensitivity.

GenomeMapper	serial: -E mm -M mm -G gaps serial with NIMS length 13: -E mm -M mm -G gaps -l 13 4 cores: -E mm -M mm -G gaps -t 3 4 cores and NIMS length 13: -E mm -M mm -G gaps -l 13 -t 3
SOAP	-s 12 -v mm -g gaps -w 10000 -c 0 -r 2
Bowtie	Allowing for 0 mismatches: -v 0 --time --seed 8526367 Allowing for 2 mismatches: -v 2 --time --seed 8526361
soap2	Allowing for 0 mismatches: -M 0 -r 2 Allowing for 2 mismatches: -M 3 -r 2
MAQ	Allowing for 0 mismatches: -n 1 –e 0 -C 513 -N Allowing for 2 mismatches: -n 2 –e 80 -C 513 –N

Table 2-2: **Command lines used for the different alignment tools.** mm and gaps denote the maximal number of allowed mismatches and gaps, respectively. Not given, options for input, output and index files.

Without allowing any mismatches, there was little difference in runtime or in sensitivity between the alignment tools, with GenomeMapper being slower than bowtie and soap2, but faster than SOAP and MAQ. Allowing two mismatches caused similar increases in runtime for all tools. With respect to sensitivity, over 99% of the differences in the reads that could be aligned with up to two mismatches resulted from different strategies in aligning ambiguous base calls (Ns). SOAP, for example, aligns Ns without an alignment penalty.

Different from SOAP, GenomeMapper's runtime was drastically affected by allowing additional gaps (which are not accommodated by the other tools tested) (Figure 2-3A). The first reason for this disparity is the different alignment strategy. SOAP does neither allow for gaps combined with mismatches nor for multiple gaps in the same alignment, while the dynamic programming alignment in GenomeMapper supports any combination of gaps and mismatches. Secondly, even though SOAP was set to run on one processor (option –p was set to 1), we found it running in parallel on up to four CPUs, and therefore using more computational power than the other tools.

By applying GenomeMapper's parallelization set to run on four cores, runtime was reduced significantly. Parallelization is geared towards complex alignments and did not reduce runtime for ungapped alignments. Another way to lower runtime is offered by skipping alignments triggered by NIMS/hits of length 12 (seeds that could not be extended by at least one base, option –l, indicated by "NIMS 13" in Figure 3A), but this came at a cost of sensitivity being reduced by 0.6%.

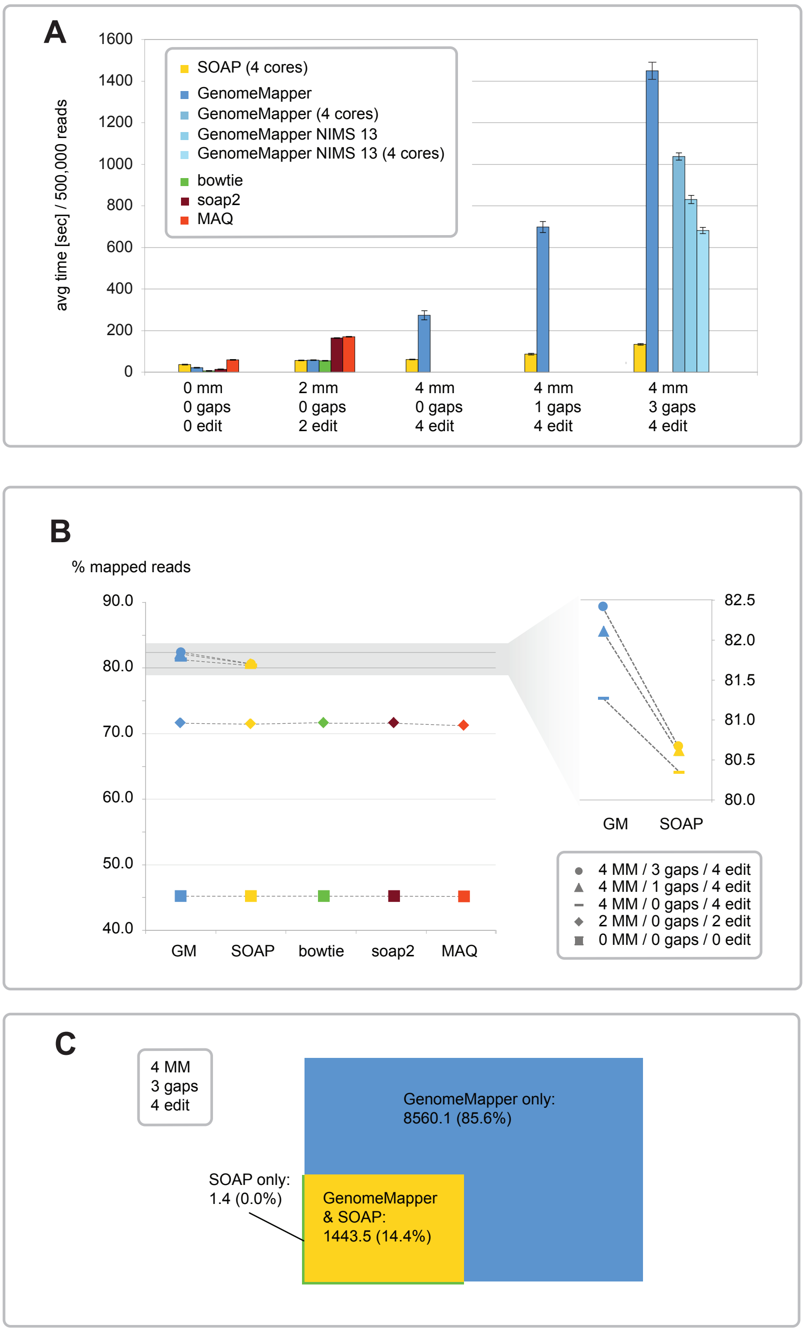

Figure 2-3: **Performance of GenomeMapper compared to other short read alignment tools.** (A) Runtime, measured as wall clock time between invocation and termination of the program, averaged from 10 independent tests with different random sets of 500,000 short reads from Est-1. The worst test was excluded from average calculations. Error bars indicate standard deviation. mm, gaps and edit refer to the maximal number of mismatches, gaps and edit operations allowed. GenomeMapper was run with four different parameter settings: the serial version; the parallel version on four cores; the serial version merely aligning NIMS of length 13 or longer; and the parallel version aligning only NIMS of length 13 or longer. SOAP was found running on up to four CPUs instead of only 1 CPU as configured with the command line (option -p). (B) Average sensitivity, measured as the percentage of aligned reads. Only GenomeMapper and SOAP can perform gapped alignments. (C) Average sensitivity of alignments, allowing for three gaps and four mismatches with a combined maximum of four edit operations, measured as number of reads with gapped alignments. Fractions refer to the number of all reads with gapped alignments.

Compared to SOAP, GenomeMapper's more accurate alignment method resulted in higher sensitivity (Figure 2-3B; compare results for 4 MM/1 gap and 4 MM/3 gaps). Considering only gapped alignments, GenomeMapper aligned over five times as many reads as SOAP (Figure 2-3C), whereas only 1 out of 500,000 reads was aligned by SOAP, but not GenomeMapper. This difference showcases GenomeMapper's ability to combine multiple gaps with mismatches in the same alignment.

Note that the reads used for benchmarking had been quality trimmed. This removes the common trend of read endings having increased chances of harboring mismatches because of higher error rates. Untrimmed reads with additional mismatches would have almost completely prohibited SOAP from performing gapped alignments. This is expected to be even more of an issue with longer reads.

GenomeMapper's relatively high runtime when allowing a large number of gaps and mismatches is mostly explained by the enormous number of alignments performed once optimizations could not reveal the best alignment. Nonetheless accurate alignments are important for correct read placement in regions of high divergence and therefore justify the performance loss. While aligning against a genome graph comes with additional computational costs, it greatly increases sensitivity. One can compensate for increased runtime with computing power, but reads that are never correctly aligned in the first place are lost for further analyses.

2.7 Proof-of-principle analysis

To examine the practical relevance of graph based alignments against multiple genomes we compared its performance to a conventional single reference approach using reads from the genome of *Arabidopsis thaliana* strain Est-1 from Estonia, generated in the Arabidopsis thaliana 1001 Genomes Project. 47.7 million alignable single-end high quality reads were produced on an Illumina Genome Analyzer. After quality trimming of the reads to 36 to 42 bp, the average depth of genome coverage was 13 fold.

First, we used the reference sequence (TAIR8 [68]) as alignment target. In the second analysis, we included two more *Arabidopsis thaliana* genomes, Bur-0 and Tsu-1 (Figure 2-4). Previous Illumina single-read sequencing and comparison against the Col-0 reference had revealed 570,100 and 502,036 SNPs, as well as 48,999 and

47,765 indels of up to 3 bp, respectively [10]. In addition, 16,463 and 3,007 longer indels of up to 641 bp had been discovered from targeted de novo assembly of highly polymorphic regions [10]. These two genomes differ from the reference by 0.5 to 0.6%, which reflects a lower bound of sequence divergence, given the limitations of short read analyses.

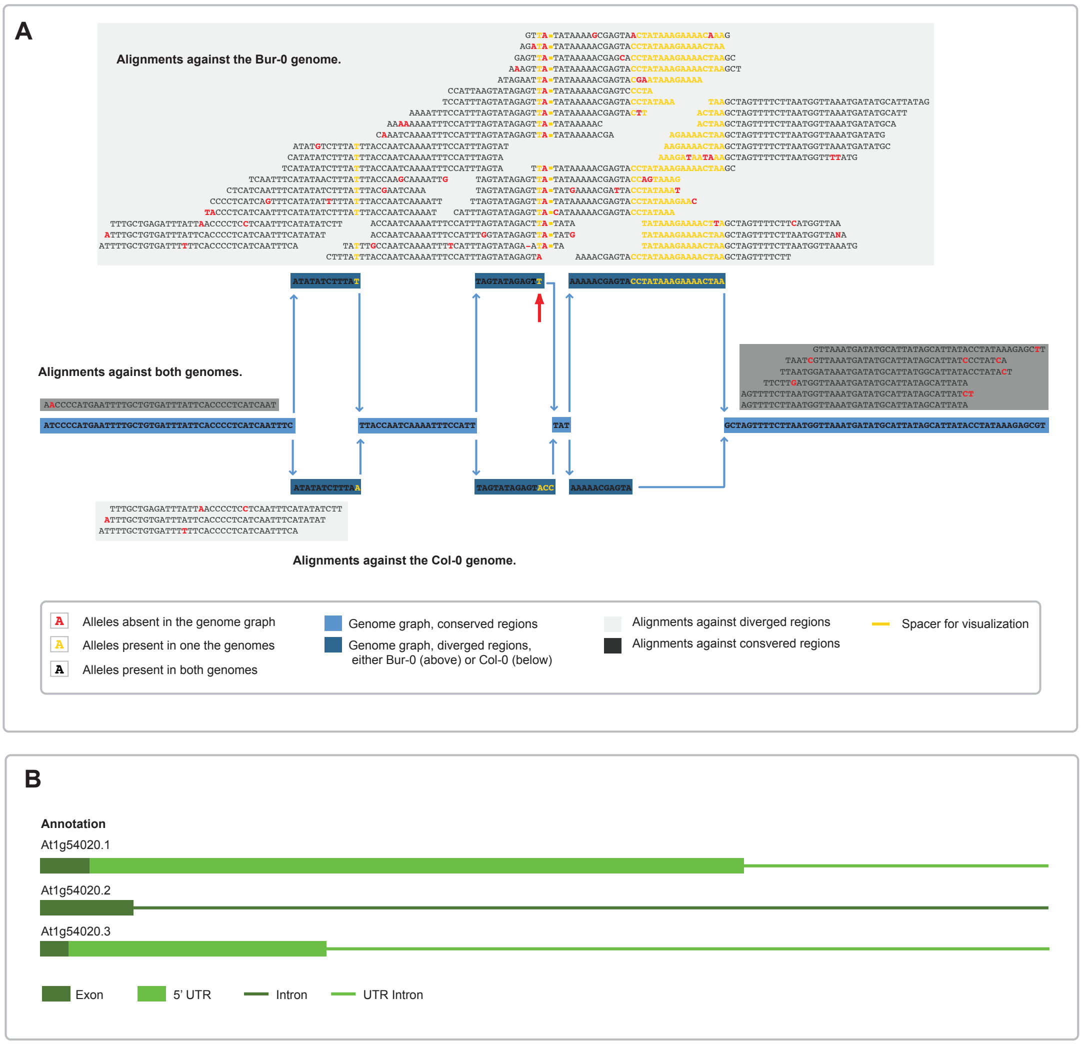
A
Alignments against the Bur-0 genome.
Alignments against both genomes.
Alignments against the Col-0 genome.
A Alleles absent in the genome graph
A Alleles present in one the genomes
A Alleles present in both genomes
Genome graph, conserved regions
Genome graph, diverged regions, either Bur-0 (above) or Col-0 (below)
Alignments against diverged regions
Alignments against consvered regions
Spacer for visualization
B
Annotation
At1g54020.1
At1g54020.2
At1g54020.3
Exon
5' UTR
Intron
UTR Intron

Figure 2-4: Alignments against a 17 bp insertion present in a non-reference genome. (A) Alignments of Est-1 reads against the graph of Arabidopsis chromosome 1, reference positions 20,166,584 to 20,166,747. Alignments against both the Col-0 reference and the Bur-0 variant genomes are highlighted in dark gray, alignments of reads aligning best against a single genome are highlighted in light gray. Most reads align against the Bur-0 allele, suggesting that Est-1 is more similar to Bur-0 at this locus. In particular, the 17 bp insertion found in Bur-0 is supported by the Est-1 reads. Due to the alignment constraints (maximum of four edit operations), these alignments could not have been performed against the Col-0 sequence only. Within the second divergent region, indicated by a red arrow, Bur-0 has a complex change, ACC->T, relative to Col-0, with Est-1 featuring a third allele, ACC->TA. Since this change is near the 17 bp insertion, only a subset of the alignments would have been found with single reference alignments only. For simplicity, Tsu-1, which is also included in the graph target, is not shown here. (B) Annotation of this region with respect to the Col-0 reference genome.

The Bur-0 and Tsu-1 genomes together with the Col-0 reference genome were used to build a multiple genome graph. To take advantage of the additional information produced by the graph based alignments, and to make it comparable to a single reference analysis, we updated SHORE [10], our genome resequencing analysis pipeline. This included incorporation of GenomeMapper's transformed alignment representation, different scoring schemes for previously known and newly discovered polymorphisms, and the support of indels up to any length, restricted only by the maximal indel length within the known genome space.

More than 1% of all reads, 0.51 million reads, could be aligned to the genome graph, but not to the single reference. These additional alignments resemble highly divergent regions of Est-1, which are particularly interesting, but also constitute the regions that are least accessible to conventional methods. Compared to the "reference only" alignments, the graph alignments increased the number of recovered SNPs by 15%, of deletions by 22.6%, and of insertions by 37.2% (Table 2-3). In particular, 1,551 deletions and 1,841 insertions longer than 3 bp, with a maximum length of 641 bp and 281 bp, known from previous de novo assembly of larger indels in Bur-0 and Tsu-1 [10], were detected. Only a small subset of the long indels was represented in the "reference only" analysis (two 3 bp deletions can modify the sequence in the same way as one 6 bp deletion). Due to the limitation of three gapped positions per alignment, the vast majority of long indels could not be discovered with the conventional "reference only" alignment. These observations illustrate that indel detection is not limited by alignment constraints, but only by the data included in the genome graph.

		Predicted by both analyses	Private to genome graph analysis	Private to reference-only analysis	Total gain in genome graph analysis
SNPs		401,158	66,264	5,423	15.0%
Deletions	all	25,926	6,807	778	22.6%
	1-3 bp	25,865	5,256	778	16.8%
	≥ 4 bp	61	1,551	0	2,542%
Insertions	all	22,305	9,220	678	37.2%
	1-3 bp	22,285	7,379	678	29.2%
	≥ 4 bp	20	1,841	0	9,205%

Table 2-3: **Recovery of Est-1 variants using SHORE.** Predictions made by both analysis include includes variants predicted by graph-analysis that have been found in the single reference analysis in the same sequence context, though with a differing position resulting from ambiguous alignments. Some of the variants longer than 3 bp could be reassembled in the single reference analysis, by combining shorter indels.

The reliability of variant detection was improved as well, with 244,101 SNP calls made in the “reference only” analysis having additional support from one of the additional genomes in the graph (11,382 and 16,958 for deletions and insertions, respectively). Similarly, recall rates for 1-3 bp indels were drastically increased. We reduced the minimum coverage requirement for SNPs and short deletions supported by other genomes from 3 to 1.

This is valid assumption; first, as the reliability of SNPs increases drastically if prior knowledge about its existence is available. This prior knowledge is given through the incorporation of known SNPs into the alignment target and mismatches of reads against the reference supported by matches against the sequence of other genomes are annotated in the alignment in a different format than non-supported mismatches (e.g. a mismatch of A to T, non-supported: [39], supported: (AT)).

Second, validation results for single reference and genome graph analysis based on 600 kb of dideoxy sequences distributed throughout the Est-1 genome [7] are not worse for the graph based analysis (Table 2-4). In a average *Arabidopsis thaliana* strain, about 85% of SNPs are accessible with 36 bp single end short reads, with the remainder being located in repetitive regions [10]. Of 2,316 SNPs in the validation set, 85.2% were called using genome graph analysis, an increase of over 7% compared to the single reference analysis at a similar error rate of less than 0.5%. Recall rates for indels were increased even more, by 14.8% for insertions and 8.4% for deletions.

			Graph analysis		Single reference analysis	
		N*	Recall§	FDR†	Recall§	FDR†
SNPs		2,316	85.2%	0.4%	77.5%	0.4%
Deletions	all	183	53.6%	2.0%	38.8%	2.7%
	1-3 bp	132	68.2%	2.2%	53.8%	2.7%
	≥ 4 bp	51	15.7%	0.0%	0	n/a
Insertions	all	167	53.9%	2.2%	45.5%	1.3%
	1-3 bp	128	66.4%	2.3%	59.4%	1.3%
	≥ 4 bp	39	12.8%	0.0%	0	n/a

Table 2-4: **Validation of polymorphism predictions in Est-1.** * Number of known variants in 600 kb of dideoxy sequence data from [7]. § Ratio of confirmed to the sum of confirmed and missed predictions of the respective kind; indicates sensitivity of method. † False discovery rate, percentage of erroneous calls.

For a final comparison, we aligned all Est-1 reads against the three known genomes separately, with the Bur-0 and Tsu-1 genome sequences generated by introducing all known variations into the reference Col-0 genome. As expected, nearly the same set of reads could be aligned, but the graph alignments were 21.3% faster than the serial alignments. This improvement would be even greater, if one took into account the additional analyses needed for merging and filtering of separate and redundant alignments.

The results of the graph analysis of Est-1 can be downloaded from the 1001 Genomes portal (1001genomes.org).

2.8 Discussion

The first goal for short read mapping tools was the design of efficient alignment algorithms that were faster than the speed with which raw data were produced.

Considering that intraspecific sequence differences are often more substantial than previously anticipated, a major challenge is the requirement not to disregard or misplace too many reads. With the rapidly increasing knowledge of variants, one could simply align against all known genomes for a species separately. This would not require any new methods, but it comes with the overhead of redundant alignments in conserved regions. We have shown that graph alignments are already superior with information from only two divergent genomes added to the first genome sequence produced for Arabidopsis. This advantage should become much more drastic once hundreds of genomes are incorporated into the graph structure. In addition, this should improve the workflow, as the separate handling of hundreds of separate references would become increasingly impractical.

We have demonstrated that short read alignment against a complex graph representing multiple genomes is not only possible and produces meaningful results, but also provides access to regions that are highly divergent from the first reference. In addition, our approach reduces the number of false positive SNP calls caused by misalignments near indels [10]. To our knowledge this constitutes the first approach that efficiently incorporates multiple references and solves resultant problems. We note in addition that the representation of multiple genomes in a complex graph structure is not restricted to short read mapping or intraspecific analyses. Other applications are easily conceivable, e.g., accurate local and global alignments of longer reads (up to whole genomes) against all known genomes of a species or even against a structure representing groups of related species, enabling analysis of metagenomic samples in one step. Likewise, read alignments against splice graphs representing known isoforms with differing exon-intron junctions would be beneficial for mRNA analysis.

Once the species-wide genome graph of Arabidopsis covers most common variants (see the Arabidopsis thaliana 1001 Genomes Project [2]), resequencing of newly collected material will become easier, as fewer inaccessible regions remain. A prerequisite for this are universal and community-wide accepted positional descriptors of insertions.

3 Reference-based assembly of four *Arabidopsis thaliana* genomes

This chapter describes our efforts to design and implement a homology-guided assembly tool used to assemble four diverse genomes of *Arabidopsis thaliana*. These assemblies display the first four assemblies of this species after the release of the reference sequencing in the year 2000. As-of today this work was submitted to PNAS and send out for review.

Together with Stephan Ossowski I implemented this method that was designed in collaboration with Juliane Klein, Daniel Huson and Detlef Weigel. Additional work includes analysis of expression studies performed with the help of Lisa Smith, Stefan Henz and Felix Ott. Norman Warthman and Christa Lanz performed the sequencing of the Sanger validation data and the Illumina sequencing, respectively.

3.1 About the differences of assembly, homology-guided assembly and resequencing

The original papers reporting the first complete eukaryotic genome assemblies were published around the turn of the millennium; many of them had the words “the genome of” in the title, and this was also true for the first assembled plant genome released in 2000: “Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*” [68]. In the past decade, there has been a growing appreciation that individuals of the same species are not only distinguished by small scale differences such as single nucleotide polymorphisms (SNPs), but that copy number variants often account for an even greater difference in genetic material, both within and between closely related species (e.g., [69]).

Since the advent of Next Generation Sequencing (NGS) technologies, the main challenge of whole genome assembly has no longer been the generation of sufficient amounts of sequencing data or the costs per sequenced base, but the complexity and size of genomes, which are difficult to reconstruct with short reads produced in a whole-genome shotgun approach. Where high-quality reference sequences are available, they can be used as alignment targets for short sequence reads (see earlier chapters), followed by analysis of overlapping alignments and consensus base calling [70], [71], [50], [10]. While these methods provide good sensitivity and specificity in regions that are conserved between the reference sequence and the focal genome, regions of high variability permit short alignments and this makes their analysis unreliable or even impractical [10].

We have previously estimated that up to 7% of the *A. thaliana* non-centromeric genome comprises highly diverged regions [10], [8], [9] and other plant species can be even more polymorphic [40], [72]. Moreover, with the arrival of each new sequencing technology, the estimated level of variation in *A. thaliana* has increased: whole genome shotgun sequencing with Sanger dideoxy reads disclosed 1 SNP every 1,000 bp in a single divergent genome [68], while array hybridization of DNA from 19 divergent genomes identified 1 SNP every 800 bp, with an estimated false negative rate of 75% based on targeted Sanger dideoxy resequencing of selected regions [8], [7]. In agreement, the first NGS study revealed a density of at least 1 SNP every 200 bp between random pairs of strains [10]. The latter study followed the alignment-consensus approach and thus excluded most regions of high divergence or repetitiveness. Therefore the estimated divergence within and between populations of *A. thaliana* is expected to be even higher than this recent estimation. All of these studies were based on sequence differences that have been accessible through the conservation between reference sequence and focal genome.

The general need for advanced whole genome assemblies is particularly acute in light of efforts such as the Human 1000 Genomes Project [73] or the *Arabidopsis thaliana* 1001 Genomes Project [2], both of which aim to identify all non-private variations within the pan-genome of a single species.

Prediction methods for structural variants (SVs) have been developed to annotate diverged regions based on paired-end sequencing. By comparing expected and observed distance and orientation of the alignments of the two sequenced ends of a single fragment from the focal genome to the reference genome, these methods can reveal various types of variations between the structure of the reference sequence and the sample genome [74], [75], [55], [76], [77], [78], [79], [80], [81]. Unfortunately, these predictions do not include the actual sequence of the variants, and they often miss larger rearrangements, complex changes, and small insertion/deletions. The reasons include missing data and statistically insignificant deviation from the expectation as well as complicated situations caused by multiple, overlapping events that cannot be easily inferred from paired-end alignments only. Further, regions similar in length to the reference but dissimilar in sequence content will not form alignments with unexpected distance or orientation to each other. To overcome these shortcomings, it has been suggested to locally assemble regions of high dissimilarity between sample and reference sequence, but no study published to date has applied this idea to all diverged regions or to the entire genome [71], [49], [10]. The last chapter already outlined one way to reduce reference bias, i.e. the efficient usage of multiple references as alignment target.

But perhaps the simplest way to by-pass all problems specific to reference-based approaches is de novo assembly, being independent of any extrinsic homology information, and therefore not biased towards the sequence of any other genome. Even though this approach has been applied to complex genomes analyzed with short NGS reads only [82], [83], the resulting contigs and scaffolds tend to be rather short and are known to lack a substantial portion of the genome. Part of the problem is the complexity inherent in whole genome shotgun data, even when the reads are reasonably long. Different studies have tried to reduce the complexity by introducing reduced-representation libraries through digestion of the genome prior to sample preparation, leading to scaffolds that are dozens of kb long [84].

In this chapter I present our work on the assemblies of four homozygous *A. thaliana* genomes, from the divergent strains L*er*-1, C24, Kro-0, and Bur-0, based on a new, multi-tiered approach of de novo assembly guided by homology to a reference genome. Using 2 Mb of Sanger shotgun reads, we find that the per-base error is less than 0.01%, with very few substantial mis-assemblies. These genome sequences greatly expand our knowledge of the *A. thaliana* pan-genome. I further show how this new information can be used for accurate estimation of strain-specific expression differences by correcting the probe set definition for Affymetrix tiling arrays, and by improving the accuracy of expression quantification from sRNA-seq experiments.

3.2 Homology-guided assembly

3.2.1 Overview

Our whole genome homology-guided assembly approach is outlined in Figure 3-1.

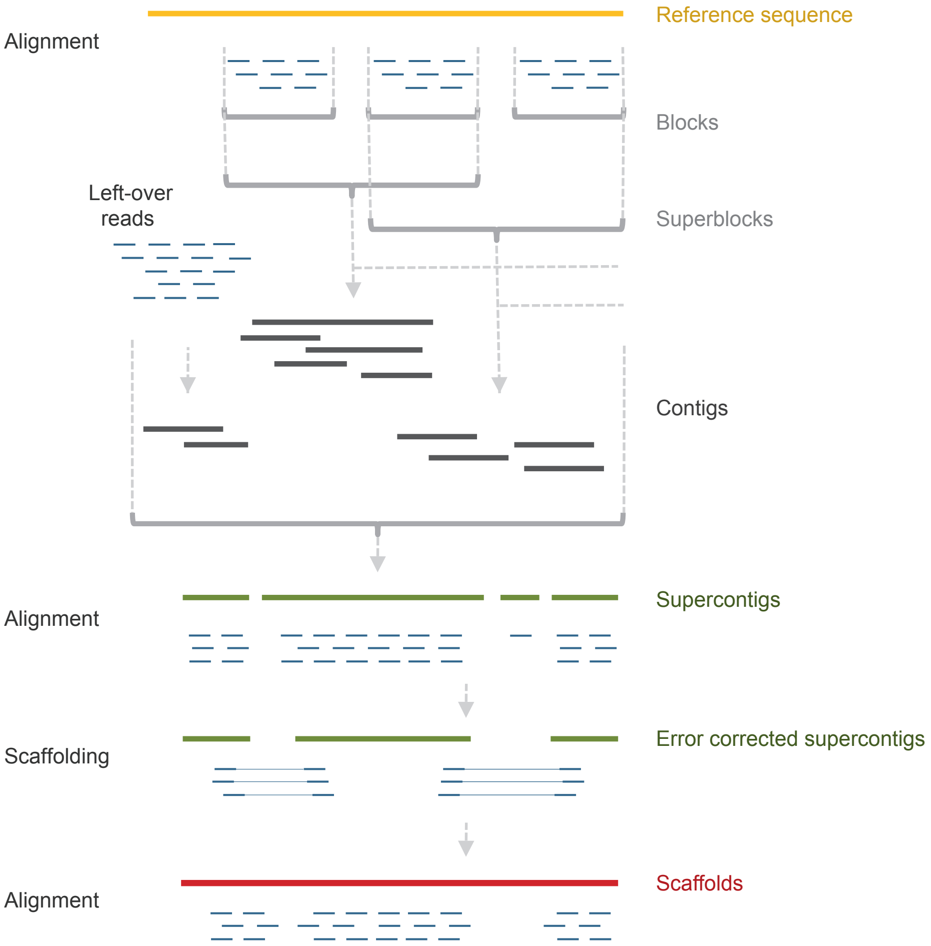

Figure 3-1: **Illustration of homology-guided assembly workflow.** Reads and their alignments are shown in blue. Regions of constant coverage were defined as blocks. Adjacent blocks were combined into superblocks until they reached a minimal length of 12 kb. Superblocks were defined in an overlapping fashion, such that blocks could belong to several superblocks. All reads of a superblock were assembled with reads that had not been aligned. Resulting contigs (dark blue) were merged into a non-redundant set of supercontigs (green). Short read alignments against the supercontigs allowed for error correction and scaffolding. Short read alignments against the scaffolds (red) allowed for final quality assessment and filtering.

The data we used had been produced on the Illumina Genome Analyzer platform as paired-end data, with individual reads of 36 to 80 bp, and average library inserts from 177 to 4,700 bp. Read pairs coming from different ends of the same fragment are subsequently referred to as mates. We started with filtering the raw reads and aligning them against the *A. thaliana* reference sequence TAIR8 (ftp://ftp.arabidopsis.org) followed by consensus calling using the short read analysis pipeline SHORE [10] and GenomeMapper [12]. We partitioned the short reads based on their alignment locations, i.e. we defined regions with constant coverage or neighbored regions that were connected by the alignments of mate pairs as *blocks* and combined adjacent blocks to superblocks, such that neighboring superblocks shared at least one block. Each of these superblocks represented one subset of reads, i.e. the reads that aligned to the constituent blocks. In addition we included 'dangling' reads that were not align-able to the reference but which had mates that aligned to one of the constituent blocks.

All such read subsets were assembled separately using three de Bruijn-graph based tools, ABYSS [85], VELVET [86] and EULER-SR [87]. As optimal parameter settings of these tools highly depends on the input data, we executed each tool eight times per read set using eight different kmer-sizes for calculation of the de Bruijn graph (Table 3-1).

	Command line	Comment
Assembly of partitioned read sets		
Velvet (0.7.60)	velveth outDir X -fastq -short <readfile> -shortPaired <readfile> velvetg outDir -scaffolding no -min_contig_lgth 100 -cov_cutoff <dependent> -max_coverage <dependent> -exp_cov auto -ins_length <dependent> -ins_length_sd <dependent>	X = eight different values per sample (17 to 51)
Abyss (1.0.14)	abyss-pe -j j=3 k=X n=10 name=outDir lib=’’ se=’<readfile>’	X = eight different values per sample, ranging from 17 to 51
Euler-SR (1.0)	qualityTrimmer -minQual 6 -span 3 -maxTrim 8 -fastq reads.fq -outFasta reads.fa filterIlluminaReads reads.fa reads_filtered.fa Assemble.pl reads_filtered.fa X library.rules joinsas shore_filtered.fa 8 shore_filtered.j printContigs shore_filtered.j	X = eight different values per sample, ranging from 17 to 28
Superlocas (0.0.1)	superlocas -I <readfile> -O <output> -LO <readfile left overs> -F fastq -C 100 3 -P pos 15 -K 15 -Kmerg 15 -Llo 25 -Lm 25 -Lt 21 -Slo 1 -Sm 1 -St 2 -Ltn 12 -Ltd 25 -Stn 2 -Std 0 -DR 15 1000	Used for Bur-0, C24, Kro-0 (36 to 80bp reads, 60x coverage)
Superlocas (0.0.1)	superlocas -I <readfile> -O <output> -LO <left overs> -F fastq -C 100 8 -P pos 31 -K 31 -Kmerg 31 -Llo 60 -Lm 60 -Lt 50 -Slo 1 -Sm 1 -St 2 -Ltn 50 -Ltd 60 -Stn 2 -Std 1 -DR 15 1000	Used for Ler-1 (80bp reads only, 200x coverage)
Assembly of left over reads		
Velvet (0.7.60)	velveth outDir 25 -fastq -short <readfile> -shortPaired <readfile> -shortPaired2 <readfile> velvetg outDir -scaffolding no -min_contig_lgth 100 -cov_cutoff 4 -max_coverage <sample_dependent> -exp_cov auto -ins_length <sample_dependent> -ins_length_sd <sample_dependent>	

Table 3-1: **Short read assembly tools, versions and command line calls.**

Excluding read pairs with both reads not aligning to the reference would introduce a bias towards regions conserved between reference and focal genome, and would not reveal larger insertions. We therefore made use of the SUPERLOCAS assembler, which allows for efficient incorporation of all left-over reads, as it builds the assembly graph for the left-over reads only once and subsequently anchors each of the local assembly graphs into the persistent left-over graph. VELVET was used to assemble all unmapped reads (including read pairs with a dangling read) de novo, in order to separately reconstruct long stretches of non-reference sequence.

About 14 Mb of the reference sequence corresponds to highly repetitive peri-centromeric and centromeric sequences [8]. Because they attract many erroneous mappings [10], we excluded all superblocks overlapping with these regions.

Our assembly pipeline introduces high levels of redundancy into the combined set of contigs of the block assemblies, contigs of the SUPERLOCAS run, and contigs from the VELVET run performed on the left-over reads. As this redundancy generates overlaps between the contigs we used the homology guided Sanger assembler AMOScmp to merge all contigs of each chromosome arm into a set of non-redundant supercontigs.

To validate the supercontigs, we aligned all original short reads against these. Consistent differences between supercontigs and short reads were taken as indications of mis-assemblies. With this information we corrected or, in the most extreme cases, removed supercontigs. Read pairs with ends that aligned to different supercontigs were employed for scaffolding with BAMBUS [88].

Resulting scaffolds were used as alignment target for a third and final round of short read mapping and consensus analysis. We developed a simple metric to assign per-base quality values according to the base qualities of the consensus analysis. Scaffolds shorter than 500 high-quality basepairs were discarded, and low quality positions were masked, additionally we ran a more stringent base masking to produce high quality though less informative assembly. Both assemblies can be downloaded at 1001genomes.org.

	Bur-0			C24		
	CA*	AS*	sAS*	CA*	AS*	sAS*
N50 (intrinsic)	6,563	193	185	6,154	109	105
L50 (kb)	3.7	147.3	147.1	4.0	273.2	273.7
N50 (target)	7,788	208	216	7,265	117	119
L50 (kb)	3.3	139.7	135.0	3.5	260.4	251.2
# Scaffolds	145,683	2,526	2,143	138,438	2,052	1,740
Total length (Mb)	96.7	101.0	96.5	96.8	101.3	98.1
Longest scaffold	59 kb	1.12 Mb	1.12 Mb	64 kb	2.18 Mb	2.18 Mb
# Ambiguous bases	0.0%	4.03%	8.30%	0.0%	3.60%	6.81%
	Kro-0			L*er*-1		
	CA*	AS*	sAS*	CA*	AS*	sAS*
N50 (intrinsic)	6,831	161	154	4,405	113	108
L50 (kb)	3.6	163.5	167.3	5.7 kb	272.5	270.8
N50 (target)	8,011	178	181	5,016	121	126
L50 (kb)	3.2	151.8	145.6	5.2	261.9	246.5
# Scaffolds	160,535	2,670	2,408	104,403	1,528	1,261
Total length (Mb)	97.3	99.9	96.7	98.6	100.8	96.3
Longest scaffold	51 kb	1.48 Mb	1.48 Mb	88 kb	1.09 Mb	1.09 Mb
# Ambiguous bases	0.0%	5.10%	8.12%	0.0%	1.3%	8.53%

Table 3-2: **Assembly statistics comparison of alignment-consensus and assembly derived contigs.** CA = Consensus-alignment approach, AS = Assembly, sAS= stringently masked Assembly.

3.2.2 Combing the contigs with AMOScmp

The block assemblies, the SUPERLOCAS run and the VELVET assembly on the left-over-reads introduced high levels of redundancy into the combined set of contigs. This is evident at five levels: (I) different runs of different assembly tools performed on identical sets of input reads; (II) reads used two times if their respective block has been allocated to two overlapping superblocks; (III) unmapped reads can contribute to multiple superblock assemblies as they are not removed from the SUPERLOCAS' left-over-graph once incorporated; (IV) reads re-used due to repetitive alignments to multiple blocks; (V) assembling the left-over pairs was performed including the dangling pairs that were also included in the block assemblies.

This redundancy generates overlaps between contigs. We use AMOScmp [89] in order to assemble the contigs and at the same time purge the redundancy. AMOScmp applies an alignment-layout-consensus approach using alignments against the reference to guide the overlap calculation of contigs. In order to reduce complexity and hardware requirements we ran AMOScmp on each chromosome arm separately.

We used AMOScmp (version 2.0.8) to assemble the contigs that were produced by the short read assembly tools. This allows removing redundancy inherent in the contig assemblies. Contigs were separated by chromosome arm and assembled using the respective chromosome arm reference sequence as homology target. Within the AMOScmp script we executed all programs with default values except of casm-layout using parameter –t 3500 (maximum ignorable trim length) and *make-consensus* using parameter -o 10 (minimum overlap bases).

3.2.3 Correcting for mis-assemblies

We aligned all short reads against the set of supercontigs. Differences between aligned reads and reference sequence reveal mis-assemblies in the supercontigs.

Any supercontig shorter than 100 bp, featuring an average coverage below 4 or below 1% of expected coverage as well as supercontigs with an average repetitiveness of read alignments of 1.8 or more were removed. Afterwards all remaining supercontigs were split at any region where variant predictions indicate mis-assemblies including uncovered regions, local clustering of differences (i.e. two or more predicted differences in less than 10 bp) and regions with mate pairs that do not align in the expected order and orientation. It has been shown that distance and/or orientation of the alignments of the two reads of a read pair reveal difference between focal genome and reference sequence [76], [79], [90], [91]. The likeliness that observed distances of such read alignments reflect the sequenced clone length is determined by the size distribution of all sequenced fragments that can be estimated based on unique alignments in conserved regions. In order to distinguish real mis-assemblies from wrong read placement we required that at least 10% of the positions that are spanned by the discordant read alignment not to be covered by any unique alignment.

3.2.4 Scaffolding

Read pairs which reads aligned to two different supercontigs defined a connection (bridge) between the respective supercontigs. Bridges suggest that two supercontigs are in local vicinity and have a defined order in the focal genome. We also used homology of supercontigs to the reference sequence to infer additional connections, as described in the BAMBUS [88] manual.

Any supercontig providing at least 5 bridges to other supercontigs is classified as essential. Next, supercontigs are filtered according to the following rules: (I) essential supercontigs smaller than 50 bp and non-essential supercontigs smaller than 100 bp and (II) essential supercontigs with more than 1 error per 200 bp and non-essential supercontigs with more than 1 error per 1,000 bp are removed.

After running BAMBUS with the set of filtered bridges and connections based on homology as input the final scaffolding graph was plotted, manually evaluated and suspicious connections were removed.

By default BAMBUS connects contigs within scaffolds using a fixed number of 60 Ns, which does not necessarily correspond to the real distance. We therefore predicted the most likely distance between connected contigs based on the observed alignment locations of read pairs mapped to two connected contigs and the insert size distributions of the respective sequencing library and used this estimation as the number of Ns to be inserted between contigs in a scaffold.

3.2.5 Base quality assessment and masking

In order to assign a final per-base quality value to the assembled scaffolds we aligned the short reads used for assembly against the respective genome and used SHORE's resequencing pipeline for consensus analysis. Based on SHORE's position-wise quality values q_{ref} (reference) and q_{var} (variation) we assigned a per-base quality q_{ass} to the residues of the assembly by the following scheme. If there was only a SHORE call supporting the reference, then q_{ass} equals q_{ref}. If only a variation quality was given then q_{ass} equals 0. If there is evidence for two allele calls, q_{ass} was assigned the maximum of 0 and the difference of q_{ref} and q_{var}. Every base that was assigned a quality value of less than 10 was masked. Every scaffold that featured less than 500 unmasked bases was discarded. Ns at the beginning of scaffolds were chopped. In order to produce a second, more sensitive though shorter assembly, we masked all bases with a quality of less than 15. Additionally we masked all unmasked regions that were shorter 100bp.

3.2.6 Short read mapping and consensus analysis

Short read alignments followed by a consensus analysis were used at three different stages within the assembly. First, the read partitioning was based on short read alignments against the reference sequence. Second, short reads were aligned against supercontigs for assembly correction and scaffolding. Third, short reads were aligned against the final scaffolds for per-base quality assessment and filtering.

For each such alignment-consensus analysis we used the short read analysis pipeline SHORE with GenomeMapper as alignment tool. Within the alignment we allowed for at most 10% of the positions of a read to mismatch, incl. 7% being involved in gaps. Repetitive alignments were removed if another alignment of the same read in combination with an alignment of the read pair was more likely to resemble the sequenced clone (paired-end correction).

Base calling was performed using SHORE's quality metric for homozygous variation. For the third analysis we additionally allowed base calling in repetitive positions.

3.2.7 Assembly statistics

We used a mix of single-end, paired-end and mate-pair libraries for all four genomes, with different contributions of single- and paired-end data, and different insert sizes. Total coverage was greater than 70x for all and greater than 320x for L*er*-1.

Two common metrics to assess assembly quality are N50 and L50, which indicate the total number and minimum length, respectively, of all scaffolds that together account for 50% of the genome. After exclusion of centromeric regions, we had targeted for assembly sequences that correspond to around 105 Mb of the reference. Based on this value, N50 and L50 for our assemblies ranged from 117 to 208, and 140 kb to 262 kb, respectively, with the longest scaffold in each assembly being between 1.1 and 2.2 Mb. For comparison, contigs derived from a standard alignment-consensus with additionally concatenating consecutively called positions yielded N50 and L50 values of 6,147 and 4.1 kb, respectively, for L*er*-1 (Table 3-2). The cumulative length of all scaffolds in each assembly was about 5% shorter than the target of 105 Mb; we assume that this was mainly caused by repetitive sequences. Indeed, the non-pericentromeric segments of the reference sequence not covered by our assemblies were largely repetitive, with a 36 to 41% repeat content, compared to an average of 8% of repetitive positions in non-centromeric regions (as assessed with 36-mers [10]).

3.2.8 Comparison with a standard alignment-consensus approach

We performed standard resequencing analyses on all four strains in order to analyze the difference between the homology-guided assembly and the alignment-consensus methods, comparing both the contig sizes, genome coverage and the resultant polymorphism calls. We used the same set of reads as for the assembly and again applied SHORE's resequencing pipeline using GenomeMapper as alignment tool. We allowed for 10% and 7% of the nucleotide of a reads to mismatch and or to gap, respectively. Concatenating adjacent base calls (including reference, SNP and micro-indel calls) generated the alignment-consensus contigs. Table 3-2 shows that homology-guided assemblies out-perform a standard analysis.

3.2.9 Assembly validation with 2 Mb of dideoxy data

To assess the quality and error rate of the assemblies, we used 955 shotgun Sanger reads of the Bur-0 genome generated for this project, and a published set of 3,388 fragments of the C24, L*er*-1 and Bur-0 genomes produced by targeted Sanger resequencing [7]. We refer to these sets as "shotgun" and "MN2010", respectively. MN2010 is enriched for unique and genic sequences, whereas the shotgun set results from a random sampling. All Sanger reads were aligned against the respective assembly and the peri-centromeric portions of the reference using BLAST [92] (Table 3-3). Between 4.2% (48) and 4.4% (49) of MN2010 fragments aligned to organelles or peri-centromeric regions. In the uncurated shotgun set , 28.0% (267) of the reads aligned against organelles or peri-centromeric regions.

	L*er*-1 (MN2010)	C24 (MN2010)	Bur-0 (MN2010)	Bur-0 (shotgun)
Sanger reads	1,139	1,139	1,110	955
Organelle/centromere hits	48	48	49	267
No significant hits	12	4	6	52 (30)*
Euchromatic hits	1,079	1,087	1,055	658
1) Identical	1,069	1,074	1,046	629
2) with mismatching bases	6	9	4	17
3) with indels in simple repeats	2	4	4	4
4) with indels (up to 476 bp)	2	0	1	8
Nucleotides queried	580 kb	584 kb	563 kb	285 kb
# Mismatching bases	11	14	8	22

Table 3-3: **Assembly validation.** For Bur-0 52 reads were blasted against NCBI non-redundant database, 21 reads featured alignments related to rDNA, one to human DNA (see cell with asterisk).

We found very high agreement between our assemblies and the MN2010 data. Of all reads that aligned to euchromatic regions, at least 98.8% aligned uniquely and without any mismatch, and only 0.4% to 0.8% had mismatches. An additional 0.2% to 0.4% revealed short indel errors, all of which were associated with low sequence complexity including simple repeats. There were only three MN2010 reads that revealed long indel errors not associated with simple repeats, of up to 476 bp. The total per-base error measured across all MN2010 alignments (excluding the three with long indel errors) was less than 1 in 40,000 bp.

The per-base error estimate with the shotgun set for Bur-0 was higher, but still less than 1 in 10,000 bp. Eight reads out of 658, compared to three out of 3,388 for the MN2010 set, revealed long indel errors. This was not unexpected, as the shotgun set was randomly sampled from the genome, and included more intergenic and repetitive sequences that should be more difficult to assemble. In addition, the shotgun reads had not been subjected to similarly extensive manual curation as the MN2010 set, and were thus likely to contain more errors themselves.

We compared all reads of the shotgun set without significant BLAST hit (*E* value < e^{-10}) against NCBI's non-redundant database [92]. Twenty-one of 52 reads corresponded to rDNA, and one was apparently the result of contamination with human DNA. The remaining 30 reads, or 4.4% of all reads excluding organelles, centromeres and contamination, present an upper boundary for the "unassembled space". This is in agreement with the total scaffold length of 96.2% of the size of the reference (Table 3-2), and less than what had been estimated to be inaccessible using alignment-consensus analysis [10].

3.3 Whole Genome Alignment

One disadvantage of assemblies in general compared to the resequencing approaches is the lack of a one-to-one relation of the bases of one genome to the homologous bases of the other assembly. Though this becomes necessary if one needs to annotate the difference between them.

We used the MUMmer whole genome alignment tool to align all scaffolds of each strain to the reference sequence and followed the instructions for "Mapping a draft sequence to a finished sequence" (http://mummer.sourceforge.net/manual/#mappingdraft). For this we ran *nucmer* using a parameter setting favoring specificity over sensitivity ("*nucmer --mum -b 100 -g 90 -l 35 -c 80 -f --prefix=outputFolder referenceSquence assemblySequence*"). Thus, we only allowed for alignment anchors that were unique in both the reference and query. Further we allowed nucmer to extend alignments across poor scoring regions by maximally 100 edit distance, while longer diverged regions or indels larger than 50bp always lead to an alignment break. Finally we increased nucmer's default values for minimum length of a single match and a cluster of matches and restricted the alignment to matches of the forward strand of the query.

The reasoning behind using strict alignment parameters is that relaxed alignments tend to produce false positives due to aligning regions that are not orthologous to each other. Long indels can nonetheless be accurately defined by annotating the alignment breakpoints and the distance between high-scoring segment pairs (HSPs).

Resultant scaffold to reference alignments were parsed in order to retrieve SNPs, insertions and deletions without any further filtering except that ambiguous insertions featuring more then 10% Ns were removed. Additionally we analyzed alignments with multiple HSPs by annotating the alignment breaks (gaps between

HSPs) to distinguish between simple deletions or insertions, highly diverged regions and spurious alignments in repetitive regions. Therefore a deletion was defined if more than 20bp of the reference sequence are not matched by scaffold sequence, while the scaffold sequence could be fully aligned to the HSPs upstream and downstream of the break. Vice versa an insertion is defined if more than 20bp of scaffold sequence is not matched by reference sequence. Finally we defined a highly diverged region (HDR) if more than 20bp from both reference and scaffold could not be aligned against each other, thus the break between the HSPs represents diverged but not deleted alleles in the reference and the analyzed strain.

3.4 Sequence assemblies capture large scale variations

The major advantage of genome assembly compared to resequencing followed by short read alignment-consensus analysis is the ability to detect large-scale rearrangements. We used the whole genome alignments introduced in the last chapter in order to annotation differences between the genomes. The parameter setting used for the whole genome alignment favored correct alignments over sensitivity. Because not all regions where forced to align to the reference sequence, some large-scale structural differences as well as differences in repetitive regions are likely to have remained un-annotated. The portion of the reference genome that cannot be aligned against our assemblies was as low as 3.7%, while the lowest non-aligned fraction with an alignment-consensus approach was almost three fold higher, 10.3% (Table 3-4). Based on the regions that were accessible through the whole-genome alignments, we annotated SNPs, deletions, insertions as well as HDRs.

	Accessibility (MB)	SNPs	Micro-deletion	Micro-insertion
Assembly				
Bur-0	101.0 (96%)	541,713	52,429	49,421
C24	101.3 (96.3%)	552,177	53,157	50,596
Kro-0	99.9 (94.9%)	451,928	43,847	40,659
L*er*-1	100.8 (95.8%)	530,081	50,230	49,025
Consensus Q25				
Bur-0	93.9 (89.2%)	487,550	37,231	38,136
C24	94.1 (89.4%)	484,757	37,340	37,035
Kro-0	94.4 (89.7%)	391,301	32,203	31,271
L*er*-1	93.7 (89.1%)	478,925	47,902	47,731
Overlap				
Bur-0	n/a	440,254	31,815	30,553
C24	n/a	439,990	32,457	31,002
Kro-0	n/a	355,170	27,159	26,005
L*er*-1	n/a	426,107	36,247	35,658

Table 3-4: **Comparison of accessibility, SNPs, deletions and insertions.** Differences were obtained by assembly and alignment-consensus approaches, respectively.

There was good concordance between SNPs and micro-indels (one to three bp) predicted either based on the whole genome alignments, or by the alignment-consensus approach (Table 3-4). This overlap greatly depended on the quality cutoff used for a set of SNPs, the parameter settings used in the whole genome alignment or in the consensus calling. The assemblies, though, revealed more small-scale changes: On average, an additional 12% SNPs were called, and 29% and 23% more micro-deletions and micro-insertions, respectively (Table 3-4).

We also analyzed the length distributions of apparent deletions and insertions relative to the reference and HDRs (Table 3-5, Table 3-6, Table 3-7 and Table 3-8). Over 1.7 Mb of reference sequence was missing from the L*er*-1 assembly, with the majority in deletions over 2 kb. As expected, deleted regions were significantly enriched for transposable elements (63.5%, compared to 13.7% of all positions in non-centromeric regions).

	Deletions		Insertions		HDRs > ~30 bp[†]	
Variation length	*n*	Cumulative length (bp)*	*n*	Cumulative Length (bp)*	*n*	Cumulative Length (bp)*
1	35,370	35,370	34,261	34,261		
2	9,861	55,092	10,060	54,381		
3-4	8,305	83,313	7,963	81,529		
5-8	5,816	120,122	5,677	117,295		
9-16	3,757	163,795	3,505	157,730		
17-32	1,824	205,347	1,238	185,530	66	1,752
33-64	663	235,657	579	211,943	165	9,885
65-128	296	261,847	340	241,753	379	45,063
129-256	219	302,672	127	263,429	406	121,191
257-512	204	376,717	63	286,029	359	250,682
513-1,024	240	553,208	20	298,852	217	406,617
1,025-2,048	160	776,910	2	302,228	138	599,170
>2,048	208	1,773,452	4	318,357	99	1,137,349

Table 3-5: **Variants of different sizes between L*er*-1 and Col-0.** [†]Length in the reference genome. *Cumulative length of all variants shorter and including the class in that row.

	Deletions		Insertions		HDRs > ~30 bp[†]	
Variation length	*n*	Cumulative length (bp)*	*n*	Cumulative length (bp)*	*n*	Cumulative length (bp)*
1	36,694	36,694	34573	34,573		
2	10,423	57,540	10135	54,843		
3-4	8,858	87,660	7859	81,566		
5-8	6,354	127,727	5438	115,700		
9-16	4,334	178,050	3274	153,534		
17-32	1,827	218,583	1166	180,073	70	1,756
33-64	762	253,147	481	202,186	180	10,483
65-128	350	283,895	291	227,851	358	44,313
129-256	241	328,747	85	242,314	352	108,845
257-512	210	404,027	50	259,966	282	210,541
513-1024	234	578,935	13	267,508	199	350,250
1025-2048	163	807,032	1	269,546	106	502,459
>2048	189	1,894,724	3	301,511	64	829,494

Table 3-6: **Variants of different sizes between Bur-0 and Col-0.** [†]Length in the reference genome. *Cumulative length of all variants shorter and including the class in that row.

	Deletions		Insertions		HDRs > ~30 bp[†]	
Variation length	*n*	Cumulative length (bp)*	*n*	Cumulative length (bp)*	*n*	Cumulative length (bp)*
1	31,032	31,032	28571	28,571		
2	8,592	48,216	8031	44,633		
3-4	7,082	72,321	6677	67,284		
5-8	5,141	104,635	4583	96,108		
9-16	3,713	148,204	2789	128,140		
17-32	1,971	191,726	946	149,716	54	1,397
33-64	554	217,174	479	172,047	154	8,877
65-128	279	241,148	236	192,416	310	37,666
129-256	215	281,291	92	207,723	254	85,067
257-512	174	345,001	21	214,713	232	168,391
513-1024	188	484,314	6	218,032	145	274,326
1025-2048	112	641,926	5	223,883	80	387,252
>2048	162	1,591,653	3	335,397	60	713,869

Table 3-7: **Variants of different sizes between Kro-0 and Col-0.** [†]Length in the reference genome. *Cumulative length of all variants shorter and including the class in that row.

	Deletions		Insertions		HDRs > ~30 bp[†]	
Variation length	*n*	Cumulative length (bp)*	*n*	Cumulative length (bp)*	*n*	Cumulative length (bp)*
1	37,595	37,595	35,206	35,206		
2	10,355	58,305	10,457	56,120		
3-4	8,714	87,954	8,346	84,571		
5-8	6,367	128,071	5,633	120,093		
9-16	4,225	177,409	3,496	160,648		
17-32	1,851	219,350	1,216	188,266	71	1,815
33-64	720	252,104	601	215,796	176	10,362
65-128	401	287,031	306	242,921	396	47,240
129-256	251	333,547	153	268,914	415	123,628
257-512	209	409,796	64	291,084	337	242,964
513-1024	248	590,609	20	303,683	258	430,356
1025-2048	186	852,734	4	309,567	124	604,152
>2048	185	1,764,217	10	496,155	119	1,409,065

Table 3-8: **Variants of different sizes between C24 and Col-0.** [†]Length in the reference genome. *Cumulative length of all variants shorter and including the class in that row.

The lengths of HDR alleles were strongly correlated (Figure 3-2), even though they were too divergent to be aligned directly. Additionally there was an overrepresentation of HDRs where the divergent accession had a shorter allele than the reference Col-0 strain. That we did not find the opposite case might again reflect the imperfectness of the assembly of long insertions.

Inversions constitute a class of rearrangements that should be included in the HDRs, as suppressed recombination could lead to greater sequence differences. Careful manual curation of regions with reverse complimentary alignments compared to flanking sequences revealed eighteen inversions.

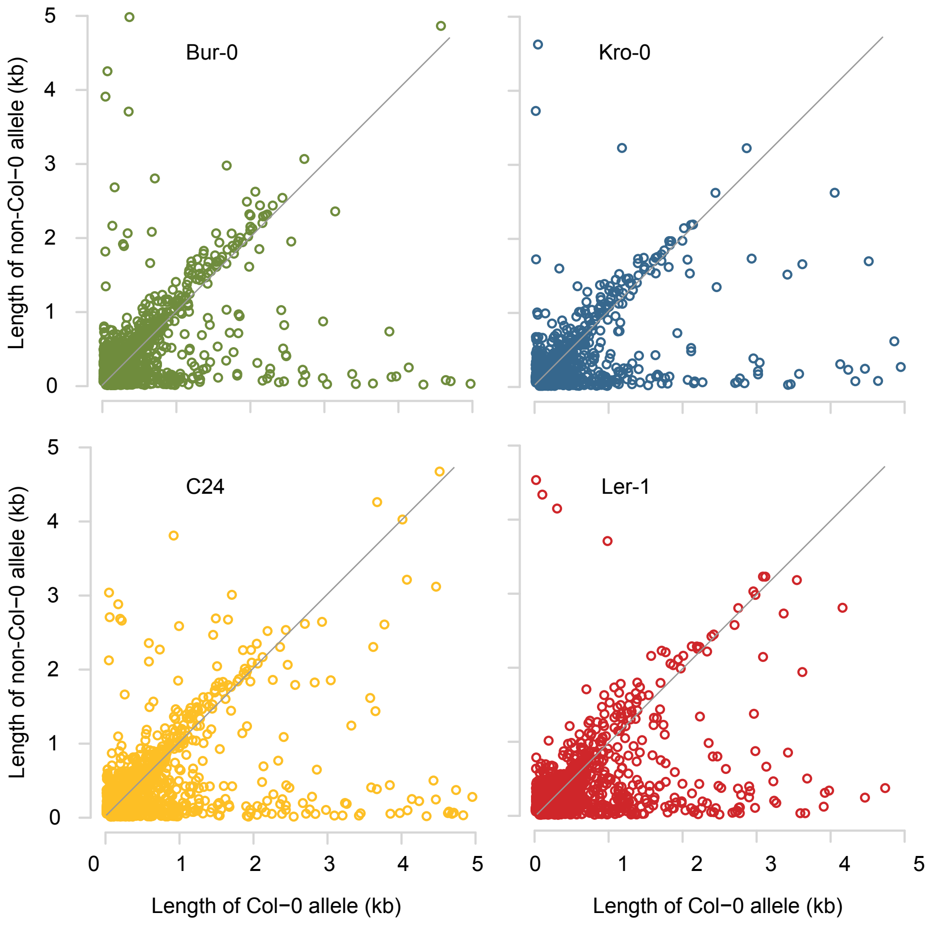

Figure 3-2: **Allele length comparisons of highly diverged regions (HDRs).**

3.4.1 Annotation of polymorphisms

All polymorphisms overlapping exons were characterized as either major (deleterious) or minor changes. Deleterious changes encompass long indels and HDRs as well as micro-indels causing a frame-shift or SNPs introducing or removing a stop codon. Micro-indels changing the length of the coding sequence by a factor of three (including multiple compensating indels in the same gene) are classified as minor changes as are any amino-acid changes except for stop mutations. Genes not featuring any mutation or only synonymous SNPs are classified as conserved.

3.4.2 Shared polymorphisms and their effect on genes

When comparing only four individuals, a large fraction of polymorphisms is expected to be found in only a single strain [10], [8], [9], [7] and this expectation is met in our accessions. Kro-0 has overall the fewest variants, both relative to the Col-0 reference and to the other three accessions (Figure 3-3). This could reflect closer relationship between these two genomes, though might also be affected by different overall qualities of the assembly. Of the 27,929 genes within the TAIR 8 annotation (excluding TEs and pseudogenes) that are present in the 105 Mb target reference genome, more than 95% could be at least partially detected in our assemblies. Slightly less than half, 45%, of the protein-coding genes had no non-synonymous change (Table 3-9). In each accession, over 3% of the genes with completely aligned sequences featured large disruptions in their coding sequence, with Kro-0 having the fewest changes (Table 3-9). Among genes with partial alignments representation, between 1,212 and 1,540 were interrupted by a HDR.

	Bur-0	**C24**	**Kro-0**	**L*er*-1**
Accessible genes	26,842	26,823	26,673	26,727
Fully aligned	23,220	23,262	23,448	23,770
Conserved	7,986	7,918	10,354	8,897
Minor change	14,320	14,438	12,306	14,007
Non-synonymous	14,224	14,350	12,237	13,904
Deletion*	379	398	311	380
Insertion*	315	342	305	378
Major change	914	906	788	866
Deletion*	342	317	291	311
Insertion*	338	325	283	319
Stop	300	336	271	314
Stop “reversion”	99	92	73	83
Partially aligned	3,622	3,561	3,225	2,957
HDR in genes	1,369	1,540	1,212	1,461
HDR in exons	374	422	314	365

Table 3-9: **Functional annotation of polymorphisms.** Annotations were made in respect to 27,929 non-centromeric genes (TAIR8 annotation). *Minor-effect indels have a length that is multiple of 3 bp.

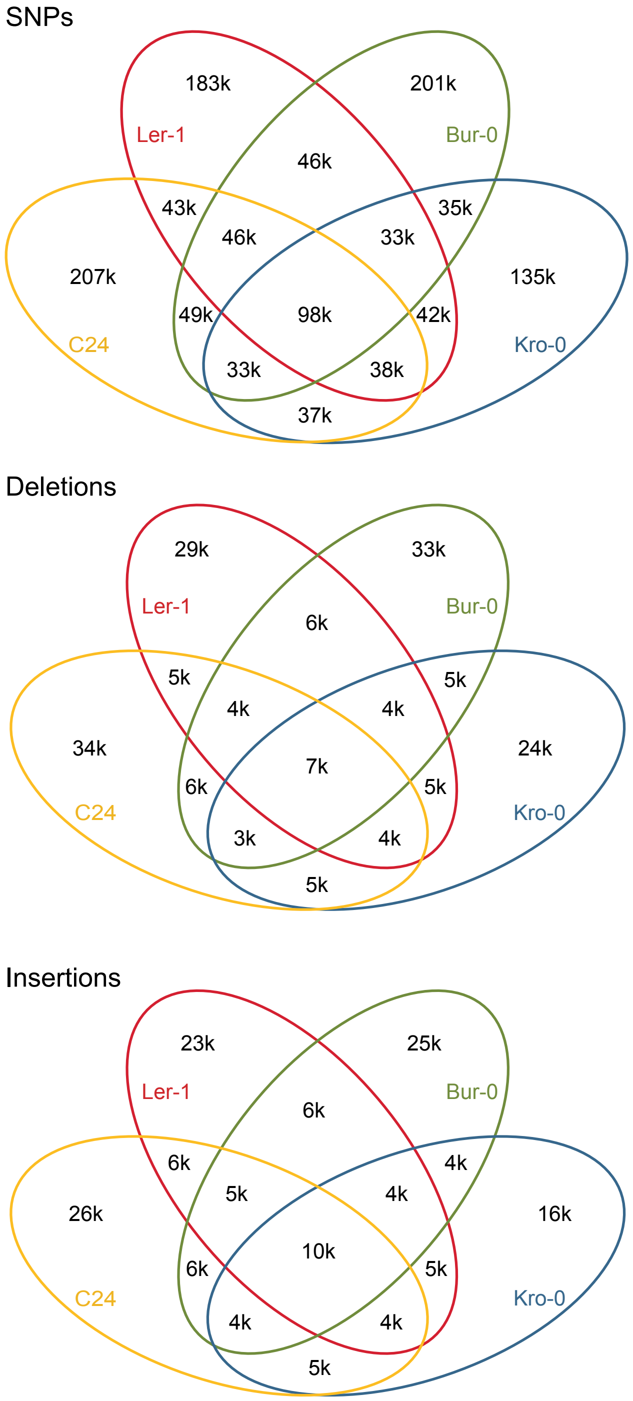

Figure 3-3: **Number of sharded polymorphisms by types.**

In humans, indels occur preferentially in multiples of 3 bp, consistent with such indels not causing frame shifts [93]. In our assemblies, 1 bp deletions were the most prevalent group, although there were distinct peaks at multiples of 3 bp that were not seen in intergenic sequences. When considering all indels in the coding sequence of a gene, the amount of deleterious changes of 1 bp was reduced, though this class was still the second most frequent in coding sequences. Intriguingly, we found that more than 20% of the genes with a length variation of a multiple of 3 bp were assembled from multiple indels with individual lengths that are not a multiple of 3 bp. This is in agreement with the total variation in coding sequence length, which showed more pronounced peaks at multiples of 3 bp (Figure 3-4).

The pairwise alignments of all assembled genes can be accessed through our web tool *POLYMORPH* (http://1001genomes.org/projects/assemblies.html).

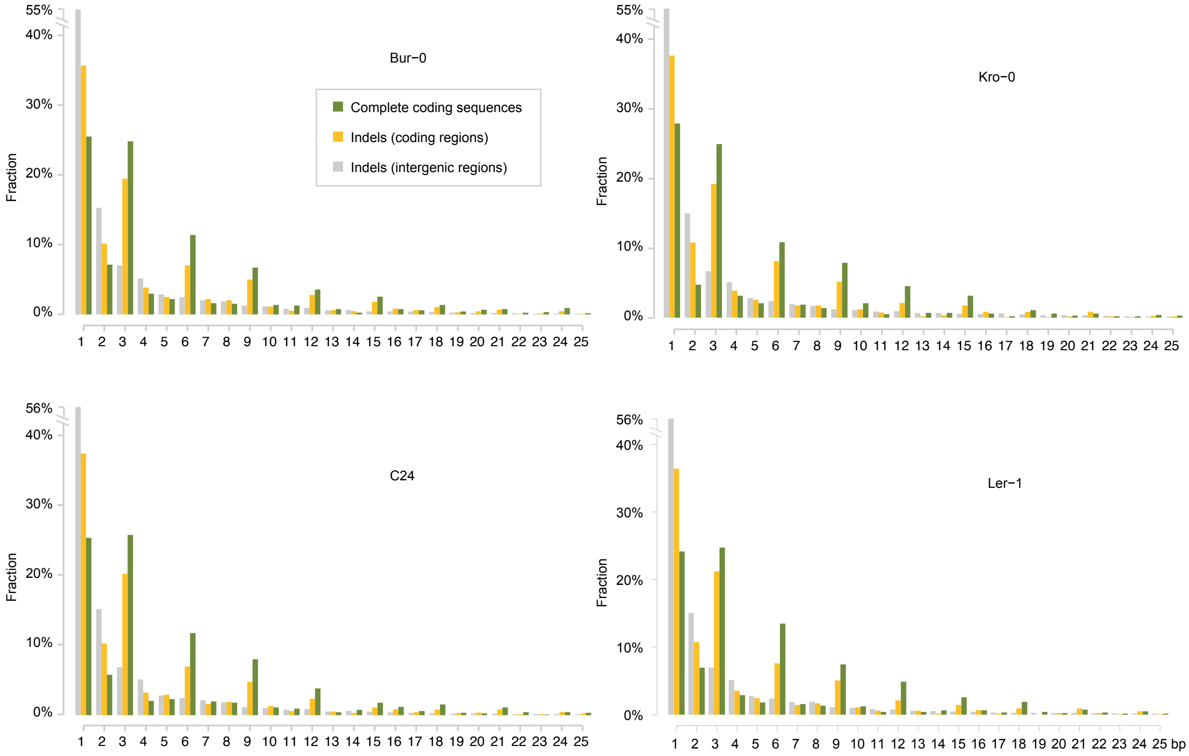

Figure 3-4: **Length variation in coding sequences.**

3.5 Using the assemblies for accurate expression analyses

3.5.1 Correcting expression estimates for protein-coding genes

Although RNA-seq is starting to eclipse microarray-based investigations of genome-wide expression profile, like arrays it suffers from typically relying on reference sequences; these are the basis of probe design for arrays, and alignment targets for RNA-seq analysis. Lack of sequence conservation between individuals easily confounds expression estimates [94].

We first investigated whether our genome assemblies would improve the interpretation of results from hybridization of RNA-derived probes to tiling arrays. We synthesized probes from RNA extracted from whole inflorescences of the Col-0 reference strain along with Bur-0 and C24, and applied these to the Affymetrix Arabidopsis Tiling 1.0R Array [95], [96].

We removed probes from about 90% (27,607) of genes, because the sequence was not identical between the focal and the reference genome. After probe removal, 8% (2,432) of genes were no longer considered, because fewer than three probes had been retained. Overall, average estimates of expression levels increased slightly and were changed for many loci, especially for genes where half or more of the probes targeted polymorphic sequences (Figure 3-5A). We also noticed that the variance in expression estimates for conserved genes (i.e. all genes where less than 2.5% of exonic positions differed between Col-0, Bur-0 and C24) was substantially lower than for polymorphic genes, even though the average estimate was the same (Figure 3-5B,C).

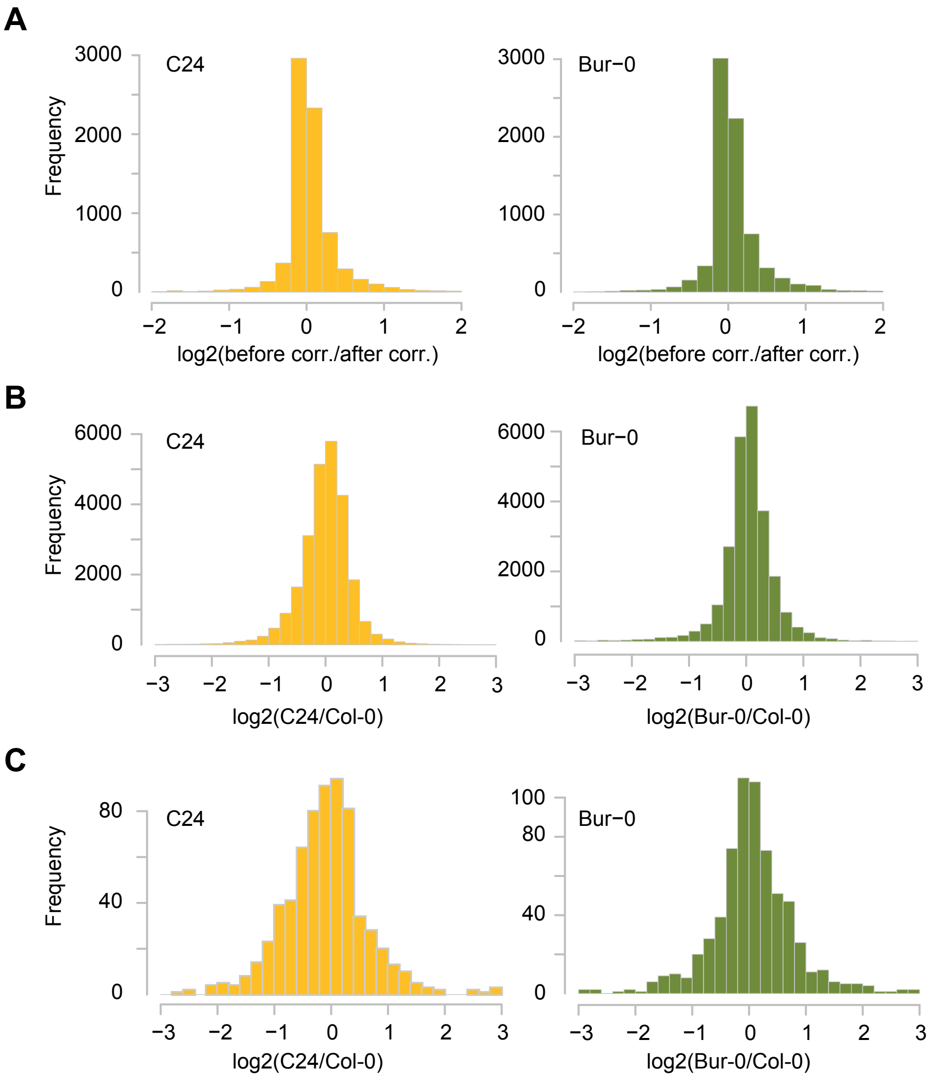

Figure 3-5: **Tilling array expression analysis.** (A) Effect of probe correction on expression estimates for 7,056 genes for which half or more of all probes were removed. Note that the distribution is skewed toward the estimates being higher after correction. (B) Expression of conserved genes (at least 97.5% of exonic nucleotides conserved between Col-0, Bur-0 and C24) (C) Expression of polymorphic genes (at least 2.5% of exonic nucleotides differ between Col-0, Bur-0 and C24).

3.5.2 Correcting expression estimates for small RNA loci

Loci that spawn populations of small RNAs (sRNAs) are much more difficult to define than mRNA producing loci, because they are defined by a collection of molecules. Because sRNAs are short, typically 20 to 24 nucleotides long, it is self-evident that even small-scale differences between the focal accession and the reference will greatly affect the number of correctly mapped sRNAs. We defined sRNA loci by consecutive and overlapping alignments of sequence reads from a sRNA library, and used the normalized number of reads in such segments to estimate expression of the entire locus.

We sequenced sRNA libraries from C24 and Bur-0 inflorescences with two biological replicates each. The resulting 6.5+5.6 and 5.8+6.8 million reads for Bur-0 and C24, respectively, were aligned against the reference genome allowing for one mismatch base using SHORE and GenomeMapper. Reads that did not align against the reference were further aligned against the Bur-0 or C24 assemblies, allowing for one mismatch. For Bur-0 (C24), 4.5+3.6 (4.6+5.4) million reads could be aligned against the reference, and 0.33+0.25 (0.28+0.38) million reads only against the new assembly.

Based on the reference alignments, we defined 30,787 segments with continuous coverage of at least 10 reads from each replicate for Bur-0, and 28,174 segments for C24. Taking not only the reads that aligned against the reference, but also those that aligned against the respective assembly into account, significantly changed the expression estimates of 348 segments (1.1%) for Bur-0, and 284 (1.0%) for C24 (Figure 3-6). In addition, 1,283 (4.0%) of Bur-0 segments, and 1,184 (4.0%) of C24 segments, could only be revealed by alignments against the non-reference genome. Finally, 579 (1.9%) of Bur-0 segments, and 556 (2.0%) of C24 segments that had been defined by reference alignments alone were merged with neighboring segments by adding the alignments to the strain assemblies.

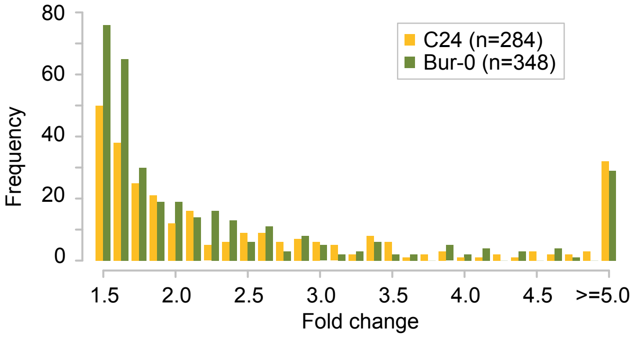

Figure 3-6: **sRNA expression analysis.** Increase in expression estimates for distinct sRNA loci resulting from incorporating information from genome assemblies.

3.6 Discussion

Since the release of the *A. thaliana* reference genome sequence ten years ago [68], no other whole-genome assemblies have been reported for this species or any other member of the genus. The reference genome was generated for some 70 million US dollars with a Bacterial Artificial Chromosome (BAC)-by-BAC strategy using dideoxy sequencing [97]. Differently from the first human genomes, a single individual, Col-0, the most commonly used lab strain, was assembled. This reference has not only been a boon to functional studies of *A. thaliana*, but it also revealed many new insights into how plant genomes evolve [98], [99].

The reference was also essential for efforts to record sequence variation between natural strains of *A. thaliana* [100]. However, only recently has there been broad realization how variable individual genome sequences in many eukaryotic species are [69], [10], [8], [9], [40], [72]. Thus, simple alignment-based methods do not provide the complete picture to mine and exploit species-wide sequence diversity. De novo assembly of individual genomes would be the obvious alternative. However, with our off-the-shelf computational resources, none of the common short read assembly tools were able to produce assemblies from the entire read set of the genomes we investigated. Causal factors might be the complexity of the *A. thaliana* genome, the length and quality of NGS reads as well as the limited insert sizes of our libraries.

The only previous work comparable to ours is that of Huang and colleagues [101], who reported a whole-genome assembly of Illumina short read data from the 367 Mb cucumber genome. While their L50 values are not too dissimilar from ours, a larger fraction of the genome was completely missing, and there was no validation of the Illumina-only genome assembly. Thus, a direct comparison is difficult.

Our homology-guided assemblies combine the advantages of having a reference genome of very high quality and being able to assemble short reads de novo. These assemblies consist of scaffolds that can be over 2 Mb long (Table 3-2). An obvious next application would be to use this approach for improving draft genome sequences, through an iterative process, in which the initial draft genome is used as a reference for homology-guided assembly of NGS reads from the same strain.

A comparison of our four assemblies confirms the value of having a well-balanced mixture of libraries with different insert sizes. Conventional paired-end libraries yield more data and longer reads than mate-pair libraries, and supply most of the sequence information for the initial sequence assemblies. Mate-pair libraries are generated from larger, circularized DNA fragments, they are limited to shorter reads, yield less data and they suffer from a higher level of clonal events than paired-end libraries. Thus, they alone are not sufficient for assembling a genome, but their insert size of several kb is advantageous for scaffolding.

Notably, there was a limit to improving assembly statistics with additional short read data. We do not know whether this reflects an inability of the assembly tools to exploit more than about 70x coverage, or whether this is an intrinsic property of read lengths and library insert sizes used and genomic repeat content. Assemblies also did not improve significantly with longer reads. For example, the C24 and L*er*-1 assemblies have almost the same N50 and L50 and genome coverage, despite C24 being sequenced mostly with 40 bp reads, and L*er*-1 with 80 bp reads. Again, the assembly tools used might not be optimized for increasing read lengths, or there are not many repeats that can be spanned by 80 bp, but not 40 bp reads.

While there was near-symmetry of deletions and insertions up to about 16 bp, larger insertions were underrepresented, likely reflecting a weakness of the homology-guided assembly approach. Still, over 300 kb of novel sequence were recovered. In comparison, in a recent study where we applied local assemblies of single-end reads to bridge gaps in homology alignments we could only reveal a fifth of that amount [10]. Notably, based on partial shotgun dideoxy data, a very rough estimate has been that on the order of 1,500 genes would be affected by medium- to large indels in a comparison between L*er* and Col-0 [6], which is in broad agreement with the almost 3,000 genes that can only be partially aligned between our L*er* assembly and the Col-0 reference (Table 3-5).

In addition to comparing genome sequences, there is great interest in studying individual patterns of DNA methylation, chromatin modifications and RNA expression.

We have already demonstrated how our assemblies improve mRNA and sRNA expression studies. We expect a similar impact on DNA methylation analyses. In humans, it has already been shown that more than half of methylation differences can be due to mutations in the underlying DNA sequence, so that it cannot be methylated anymore [102]. Thus, having knowledge of the genome sequence is essential if one wants to interpret changes in DNA methylation.

Our assemblies have also shed new light on the functional consequences of sequence variants. For example, we have shown that a substantial fraction of 1 and 2 bp indels in coding regions are compensated by nearby indels that restore the coding frame (Figure 3-4). Current genome-wide association studies generally do not consider the nature of a variant, because not all variants are analyzed or even known. With complete information, it would be possible to annotate the predicted effect of the combination of sequence variants in an allele, and subsequently base genome-wide association studies on classes of alleles with reduced or increased activity, rather than ignoring such information.

Finally, the availability of several reference sequences should improve the identification of variants in the 1001 Genomes projects that is underway for *A. thaliana*, by exploiting all known variants as targets for mapping of short reads [12] as suggested in an earlier chapter.

4 Mutant detection by deep sequencing

This chapter describes two projects published in Plant Physiology and Nature Methods, respectively [20], [21]. First we used our SHORE pipeline to analyze the genome of a mutant. Conventional genetic mapping revealed a 530 kb mapping interval harboring a spontaneous change causal for small growth and purplish leaves. The task was to distinguish all the natural occurring changes from the one that was new and causal within the given interval. For this we additionally sequenced the parental background in which the mutation occurred as well. This project was performed with Roosa Laitinen who had identified the mutant plant and performed all experiments together with Noemie Jelly. I performed the SHORE analysis of the mutant and the background genome, as well as the final mutation identification.

In the second part of this chapter, I will introduce a novel method to use the deep sequencing data not only to detect the mutation but also to perform the genetic mapping. The ultimate advantage of this method is the speed in which it can be performed. Where conventional mapping could take months, this method was finished within eight working days after DNA was extracted. Ryan Lister already introduced this idea [103], but we provided the proof-of-concept study and software allowing every biologist worldwide to map his or her mutant with deep sequencing. My part of this project was the design and implementation of the analysis software, the mutation annotation pipeline and the application of this pipeline to our sequencing data. All of this has been done together with Stephan Ossowski. Stig U. Andersen had performed all experiments expect of the Illumina sequencing, which was done by Christa Lanz.

4.1 Identification of a mutation in a non-reference background

Within this chapter we show that short read sequencing is also suitable for the analysis of new mutations in a non-reference inbred accession that differs from the reference genome in about 0.5% of all positions.

4.1.1 A spontaneous mutation

Roosa Laitinen crossed two normal appearing, green individuals of Arabidopsis thaliana accessions, Kro-0 and Anh-1, to each other. The F1 plants were all normal, but the F2 population segregated purplish, small and non-flowering plants. Plants could be prompted to flower in high humidity, but the resulting seeds were not viable. Leaves were about 10 times smaller than in wild type, but leaf cell number was reduced only about three fold, indicating that both decreased cell expansion and division contributed to the dwarf phenotype.

Using conventional mapping with almost 1,900 F2 plants of the Kro-0 x Anh-1 cross, we identified a 530 kb interval, between 21.36 and 21.88 Mb on chromosome 1, that was linked to the dwarf phenotype. The mapping interval contained 116.5 kb of repetitive DNA, which is often polymorphic and may suppress recombination [104], possibly explaining the failure to further reduce the final mapping interval.

Based on the mapping data we concluded that plants showed the dwarf phenotype had inherited both alleles from Kro-0 genotype. Since the original Kro-0 line did not exhibit the dwarf phenotype, and other Kro-0 x Anh-1 crosses did not produce abnormal F2 progeny, we concluded that a spontaneous mutation had occurred in the germline of the particular Kro-0 individual used for the original cross to Anh-1.

4.1.2 Sequencing and analysis of the background and mutant genomes

We sequenced the entire Kro-0 parental genome at 25-fold coverage, with 36 to 42 bp paired-end reads generated on Illumina's Genome Analyzer. In parallel, we produced 25-fold coverage of the haploid genome from F3 mutant plants. The reason not to directly sequence one individual but a pool from 100 plants was to obtain sufficient material for sequencing. SNPs and indels were called for both the parent and mutant pool, by independently comparing them to the Col-0 reference using the SHORE pipeline [10]. For background cleaning we made use of all variants detected in the Kro-0 parent. To predict mutations private to the dwarf sample, only those with a SHORE quality value of at least 25 were considered.

Within the 530 kb mapping interval, we identified 5,691 single nucleotide differences in the dwarf pool relative to the Col-0 reference sequence. Of these, 4,023 were predicted with high confidence. This level of polymorphism is similar to that found in other accessions in this region, with 4,036 and 3,511 found in the genomes of Bur-0 and Tsu-1, respectively [10]. Of the 4,023 high-quality polymorphisms, 531 were predicted to change the coding potential of 63 genes. All but one were shared with the normal Kro-0 parent. The one remaining mutation in the dwarf pool, a 1-bp deletion, resided in the seventh exon of the gene At1g58440, located in the middle of the mapping interval at 21.718 Mb. The deletion disrupted the At1g58440 open reading frame (Figure 4-1). Dideoxy sequencing confirmed that the mutation was specific to F3 individuals with the dwarf phenotype. A Col-0 line with a T-DNA insertion in At1g58440 (N522763) showed the same purplish, dwarf and abnormal root phenotype as these plants. At1g58440 encodes SQUALENE EPOXIDASE 1 (SQE1), which catalyzes a key step in sterol metabolism, and the morphological phenotypes of sqe1 mutants are very similar to the ones seen in our dwarfs, including partial rescue by growing plants in 90% humidity [105], [106].

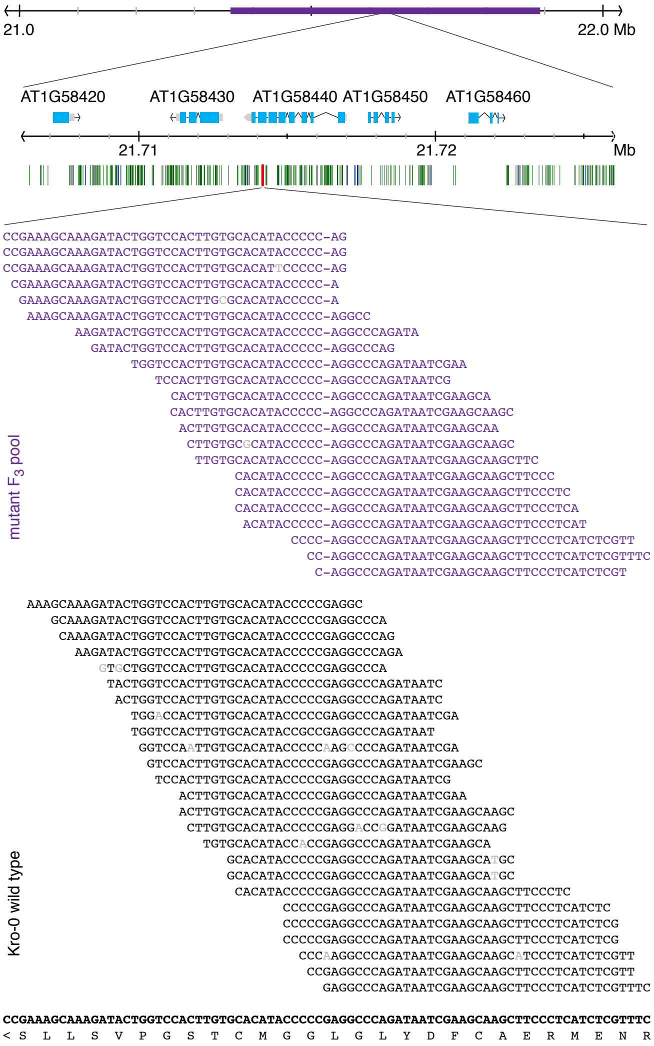
21.0
22.0 Mb
AT1G58420
AT1G58430
AT1G58440
AT1G58450
AT1G58460
21.71
21.72
Mb
mutant F_3 pool
CCGAAAGCAAAGATACTGGTCCACTTGTGCACATACCCCC-AG
CCGAAAGCAAAGATACTGGTCCACTTGTGCACATACCCCC-AG
CCGAAAGCAAAGATACTGGTCCACTTGTGCACATTCCCCC-AG
CGAAAGCAAAGATACTGGTCCACTTGTGCACATACCCCC-A
GAAAGCAAAGATACTGGTCCACTTGCGCACATACCCCC-A
AAAGCAAAGATACTGGTCCACTTGTGCACATACCCCC-AGGCC
AAGATACTGGTCCACTTGTGCACATACCCCC-AGGCCCAGATA
GATACTGGTCCACTTGTGCACATACCCCC-AGGCCCAG
TGGTCCACTTGTGCACATACCCCC-AGGCCCAGATAATCGAA
TCCACTTGTGCACATACCCCC-AGGCCCAGATAATCG
CACTTGTGCACATACCCCC-AGGCCCAGATAATCGAAGCA
CACTTGTGCACATACCCCC-AGGCCCAGATAATCGAAGCAAGC
ACTTGTGCACATACCCCC-AGGCCCAGATAATCGAAGCAA
CTTGTGCGCATACCCCC-AGGCCCAGATAATCGAAGCAAGC
TTGTGCACATACCCCC-AGGCCCAGATAATCGAAGCAAGCTTC
CACATACCCCC-AGGCCCAGATAATCGAAGCAAGCTTCCC
CACATACCCCC-AGGCCCAGATAATCGAAGCAAGCTTCCCTC
CACATACCCCC-AGGCCCAGATAATCGAAGCAAGCTTCCCTCA
ACATACCCCC-AGGCCCAGATAATCGAAGCAAGCTTCCCTCAT
CCCC-AGGCCCAGATAATCGAAGCAAGCTTCCCTCATCTCGTT
CC-AGGCCCAGATAATCGAAGCAAGCTTCCCTCATCTCGTTTC
C-AGGCCCAGATAATCGAAGCAAGCTTCCCTCATCTCGT
Kro-0 wild type
AAAGCAAAGATACTGGTCCACTTGTGCACATACCCCCGAGGC
GCAAAGATACTGGTCCACTTGTGCACATACCCCCGAGGCCCA
CAAAGATACTGGTCCACTTGTGCACATACCCCCGAGGCCCAG
AAGATACTGGTCCACTTGTGCACATACCCCCGAGGCCCAGA
GTGCTGGTCCACTTGTGCACATACCCCCGAGGCCCA
TACTGGTCCACTTGTGCACATACCCCCGAGGCCCAGATAATC
ACTGGTCCACTTGTGCACATACCCCCGAGGCCCAGATAATC
TGGACCACTTGTGCACATACCCCCGAGGCCCAGATAATCGA
TGGTCCACTTGTGCACATACCGCCGAGGCCCAGATAAT
GGTCCAATTGTGCACATACCCCCAAGCCCCAGATAATCGA
GTCCACTTGTGCACATACCCCCGAGGCCCAGATAATCGAAGC
TCCACTTGTGCACATACCCCCGAGGCCCAGATAATCG
ACTTGTGCACATACCCCCGAGGCCCAGATAATCGAA
ACTTGTGCACATACCCCCGAGGCCCAGATAATCGAAGCAAGC
CTTGTGCACATACCCCCGAGGACCGGATAATCGAAGCAAG
TGTGCACATACCACCGAGGCCCAGATAATCGAAGCA
GCACATACCCCCGAGGCCCAGATAATCGAAGCATGC
GCACATACCCCCGAGGCCCAGATAATCGAAGCATGC
CACATACCCCCGAGGCCCAGATAATCGAAGCAAGCTTCCCTC
CCCCCGAGGCCCAGATAATCGAAGCAAGCTTCCCTCATCTC
CCCCCGAGGCCCAGATAATCGAAGCAAGCTTCCCTCATCTCG
CCCCCGAGGCCCAGATAATCGAAGCAAGCTTCCCTCATCTCG
CCCAAGGCCCAGATAATCGAAGCAAGCATCCCTCATCTCGTT
CCGAGGCCCAGATAATCGAAGCAAGCTTCCCTCATCTCGTT
GAGGCCCAGATAATCGAAGCAAGCTTCCCTCATCTCGTTTC
CCGAAAGCAAAGATACTGGTCCACTTGTGCACATACCCCCGAGGCCCAGATAATCGAAGCAAGCTTCCCTCATCTCGTTTC
< S L L S V P G S T C M G G L G L Y D F C A E R M E N R

Figure 4-1: **Detection of the causal mutation with natural occurring ones.** Mapping interval (purple) on chromosome 1, and polymorphisms in the vicinity of the causal mutation (red). Green and blue lines indicate single nucleotide changes and deletions, respectively, shared with the parental Kro-0 strain. Bottom shows alignments of Illumina DNA sequence reads against the reference genome sequence, positions 21,714,424 to 21,714,504 (TAIR9). The amino acid sequence encoded by the reverse strand is given below.

This study provides a proof of concept for identifying mutations in a background other than a high-quality reference genome using direct whole genome sequencing. Different from conventional studies aimed at identifying causal mutations, we took an unbiased approach, and did not use any prior information on candidate genes associated with the phenotype in question.

In summary, our work indicates that short-read sequencing is sensitive enough for mutation identification, as long as a high-quality reference sequence from close relatives is available.

4.2 Simultaneous mapping and identification of mutants

Identification of causative point mutations after phenotypic mutant screens typically begins with genetic mapping, followed by transformation rescue or candidate gene sequencing. We present a one step alternative: performing hundreds of thousands of genotyping assays while sequencing all candidate genes. This is accomplished by deep sequencing of a pool of F2 progeny obtained from a cross to a polymorphic strain and does not require prior knowledge of mapping markers.

4.2.1 State-of-the-art in genetic mapping

Identification of sequence polymorphisms causing mutant phenotypes typically begins with genetic mapping. Not only is this laborious, but a major limitation is often the final size of the mapping interval, which may be hundreds of kb and contain dozens of candidate genes [107], [16], [108]. Two different approaches have been taken to overcome this drawback. The first is to use high-throughput genotyping assays and a dense marker map, allowing rapid genotyping of many recombinants. Thereby only a narrow candidate region needs to be analyzed for the presence of causative mutations by candidate gene sequencing or transformation rescue [109], [110]. The second approach involves performing few genotyping assays to arrive at a wide candidate region (several Mb), followed by whole-genome sequencing to pinpoint potential causative mutations within this region [16]. The first method requires multiple, successive rounds of genotyping, whereas the second is less well suited when mutation density is high, such as that commonly encountered in ethyl methane sulphonate (EMS)-mutagenized Arabidopsis populations [111]. Moreover, both approaches involve two discrete steps – recombinant genotyping and candidate gene sequencing or transformation rescue. Array hybridization or transcriptome profiling have been employed to rapidly identify larger deletions causing mutant phenotypes, but these methods are generally not applicable to point mutations [110], [112], [113]. Comprehensive whole-genome detection of single nucleotide polymorphisms (SNPs) using array technology is feasible for yeast [114], but challenging for more complex genomes. To dramatically speed up the identification of causative point mutations, we combine genome-wide genotyping and candidate gene sequencing to a single step by deep sequencing a large pool of recombinants, as outlined in a recent review [103].

4.2.2 Proof-of-principle experiment

For a proof-of-principle experiment, we used an allelic group of recessive EMS-induced Arabidopsis mutants displaying abnormally slow growth and light green leaves due to a lesion in an unknown gene. One of the mutants, in the Columbia (Col-0) background, was crossed to a wild-type Landsberg erecta (Ler-1) plant and the offspring were allowed to self-fertilize to produce a genetic mapping population consisting of Col-0/Ler-1 recombinants, following the principle of bulk segregant analysis [25], [115]. Five hundred recombinants displaying the mutant phenotype, and therefore all harboring the causative mutation within a homozygous Col-0 genomic region, were selected and a single genomic DNA sample was prepared from pooled plants. Sequencing a single-end Illumina library yielded 2.6 Gb of high-quality data that could be mapped to the Col-0 reference genome using SHORE and GenomeMapper [10], providing 22-fold genome coverage. By combining information from adjacent makers in a sliding window approach, one can effectively interrogate many more recombinant chromosomes than represented at individual positions. For example, with a density of 160 markers/200 kb, and 22-fold coverage, almost 4,000 chromosomes are analyzed within a 200-kb window, since the Illumina reads are much shorter than the inter-marker distance and thus constitute random sampling of independent chromosomes. Thus, the great majority of the 1,000 recombinant chromosomes represented in our DNA sample is predicted to contribute to the definition of the mapping interval. In conventional mapping of Arabidopsis mutations with individual recombinants, the final mapping interval from 1,000 chromosomes would be 0.1 cM, or on average 20 kb.

We developed the software package SHOREmap to allow simultaneous genetic mapping and identification of causal mutations, based on SHORE output of aligned reads from pools of recombinants. A map of 82,127 high-quality Col-0/Ler-1 SNPs is available (almost 1 SNP/kb in non-repetitive regions), as is a set of 1,219 annotated reference sequence errors [10], [8].

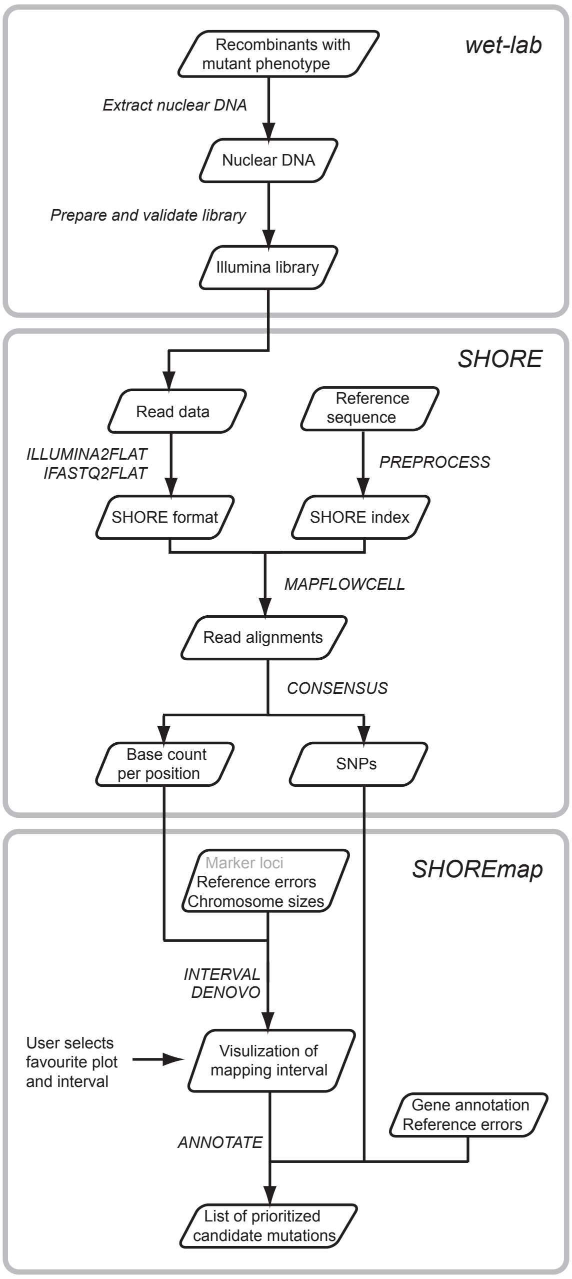

Figure 4-2: **Method workflow.** In the wet-lab, the Illumina library was prepared using genomic DNA extracted from a pool of recombinants. SHORE was used to align Illumina reads to the reference sequence. Based on the alignments, base counts per position and SNPs were defined. The candidate region was then delimited using SHOREmap, with (INTERVAL) or without (DENOVO) marker position information. Finally, SNPs corresponding to candidate mutations were prioritized and annotated using SHOREmap ANNOTATE to allow identification of the causal mutation.

First, reads were filtered, masked and trimmed using SHORE. Read alignments to the reference sequence were performed with GenomeMapper, considering alignments of up to four mismatches including gaps. Final consensus calling produced the input files for SHOREmap. The homozygous SNP calls, the base counts per position, together with the positions of Ler-1 markers [110] and reference errors [111] were used as input for SHOREmap INTERVAL. See Figure 4-2 and Table 4-1 for command line calls.

	Program	Parameters	Run time
SHORE			
1	PREPROCESS	-f TAIR8.fa –i IndexFolder -r -l 42	6 h
2	ILLUMINA2FLAT (the -c parameter requires that the GAPipeline was run with the --with-sig2 flag)	-a genomic -i 1001 -b BustardOutputFolder -o Run_01 -c -m 42 -k 36	2 h
3	MAPFLOWCELL	-n 4 -o Run_01 -i IndexFolder/TAIR8.fa.shore	6 h (8 CPUs)
4	MERGE	-p Run_01 -d AlignmentFolder	30 min
5	CONSENSUS	-n Mutant -f IndexFolder/TAIR8.fa.shore -o AnalysisFolder -i map.list -v -r	1 h
SHOREmap			
6	INTERVAL	--consensus=consensus_summary.txt --marker=ler-1.marker_pos.txt	1 h

		--chrsizes=At.chrsizes.txt	
		--referrors=At.ref.errors.txt	
7	ANNOTATE	--snp=homozygous_snps.txt	3 min
		--dist=SHOREmap_INTERVAL.output.txt	
		--chrom=4	
		--start=15,000,000	
		--end=18,000,000	
8	DENOVO	--snp=minor_allele_frequency.txt	1 h
		--refseq=reference.txt	
		--chrsizes=At.chrsizes.txt	
		--support=4 (recommended 2 to 4)	
		--freq=0.15 (recommended 0.1 to 0.2)	

Table 4-1: **Command line calls.** Parameters and run time used. Run time is estimated on one core of an Intel Xeon processor with 2.3 GHz except for MAPFLOWCELL. Memory requirements depend on genome size and range approximately from 1 to 16 GB.

SHOREmap INTERVAL determines base frequencies of the Col-0 and Ler-1 alleles at each marker position and plots the parameter r:

$$r = \begin{cases} 0, & if\ ler = col \\ -ler, & if\quad col = 0 \\ col, & if\quad ler = 0 \\ col / ler, & if\ \ col > ler \\ -(ler / col), & if\ \ ler > col \end{cases}$$

where ler and col are the sums of reads supporting the Ler-1 and Col-0 alleles at each marker position. Ten plots with different sliding window parameters are automatically generated by SHOREmap INTERVAL, based on r values.

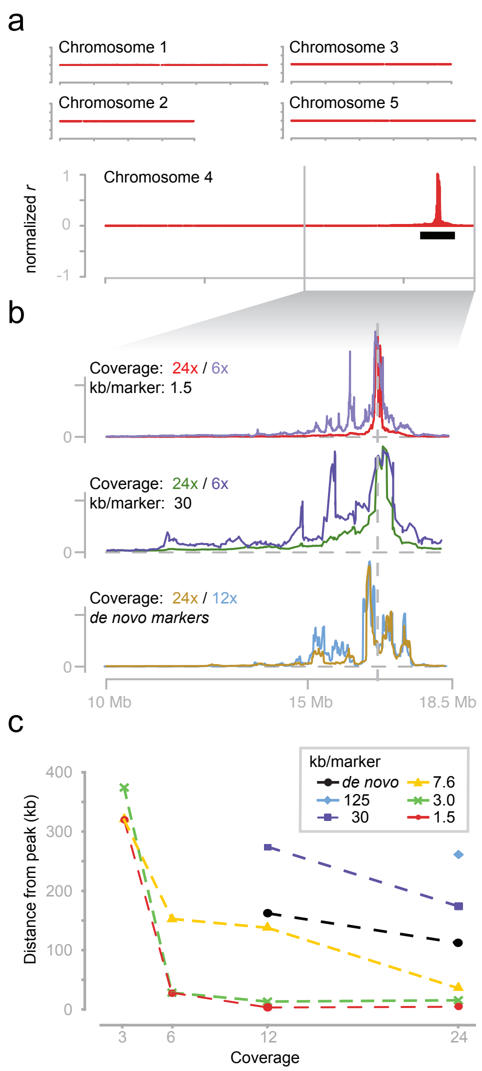

Figure 4-3: **Delimiting the candidate region**. (A) Visual output from SHOREmap INTERVAL. Red lines indicate r values in a sliding window of 200 kb. The black line delimits the candidate region analyzed using SHOREmap ANNOTATE. (B) Close-ups of SHOREmap INTERVAL and DENOVO plots. The dashed vertical line indicates the position of the causal mutation. (C) Mapping accuracy is displayed as the distance in kb between the peak and the causal mutation. Coverage: fold genome coverage. kb/marker: genome size in kb divided by the total number of markers used. de novo: markers were solely based on de novo prediction from the mutant sequence data generated on the Illumina platform.

A mapping interval for the causative mutation was readily apparent with a 200 kb sliding window (Figure 4-3). Mutations within this region were then used as input for SHOREmap ANNOTATE, which ranks base pair substitutions according to their distance from the highest r value (the peak in Figure 4-3A,B), and outputs the effect of base changes according to feature annotation (Table 4-2). Any General Feature Format (GFF) file can be applied; in this case, Arabidopsis TAIR8 annotation (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR8_genome_release/) was used. A mutation causing a serine to asparagine non-synonymous codon change in the AT4G35090 gene was positioned very close to the peak, within less than 10 kb. The second closest mutation was more than 150 kb away and the next closest mutation causing an amino acid change was found at a distance of 300 kb (Table 4-2). All three were G/C-to-A/T transitions, as is typical for the majority of EMS-induced mutations in Arabidopsis [111]. AT4G35090 was therefore unambiguously the prime candidate gene.

Chr	Position (bp)	Base	Mutant base	Distance peak	Annotation	Gene AGI ID	Effect
4	16,702,262	C	T	4,035	Coding	AT4G35090	Nonsyn.
4	16,940,438	C	T	242,211	Intergenic		
4	17,005,131	C	T	306,904	Coding	AT4G35900	Nonsyn.
4	16,287,342	C	T	410,885	Intergenic		
4	16,287,340	C	T	410,887	Intergenic		
4	17,129,041	C	T	430,814	Intronic	AT4G36195	
4	17,240,494	A	G	542,267	Intronic	AT4G36520	
4	17,245,055	A	G	546,828	3' UTR	AT4G36540	
4	16,091,082	C	T	607,145	Intergenic		
4	17,317,842	G	T	619,615	Intergenic		

Table 4-2: **Candidate ranking.** Top 10 ranked mutations from the SHOREmap ANNOTATE output.

Conventional dideoxy sequencing validated the mutation found in this mutant line. Four additional allelic mutants all contained lesions in the AT4G35090 open reading frame as well, confirming that the mutation responsible for the altered phenotype in this line had been correctly identified (Table 4-3).

Allele	CDS	Codon change	AA change	Identified by
at4g35090-3	545	AGT -> AAT	Ser -> Asn	Illumina sequencing
at4g35090-4	751	GCG -> ACG	Ala -> Thr	dideoxy sequencing
at4g35090-5	801	TGG -> TGA	Trp -> Stop	dideoxy sequencing
at4g35090-6	879	TGG -> TGA	Trp -> Stop	dideoxy sequencing
at4g35090-7	1350	TGG -> TGA	Trp -> Stop	dideoxy sequencing

Table 4-3: **Identification of additional AT4G35090 mutant alleles.** The AT4G35090 open reading frame was sequenced in four mutants allelic to that used for Illumina sequencing. This revealed one single base pair change within the coding sequence (CDS) of each mutant. Three of the changes cause premature stop codons, and one results in an amino acid substitution.

We varied marker density and genome coverage, to determine the minimal thresholds for accurate mutation identification. Six-fold genome coverage was sufficient at high marker densities (less than 3 kb/marker) and little accuracy was gained at coverage above 11-fold. Likewise, even a low density of markers (30 kb/marker) was sufficient to identify the causal mutation, as long as more than 20-fold sequencing depth was used. Little gain in accuracy was seen for marker densities above 3 kb/marker (Figure 4-3). We also note that bulk segregant analysis makes the method robust to occasional misphenotyping of recombinant individuals, which can be a considerable source of error in conventional mapping. If Ler-1 reads do not disappear near the peak of r values, this would indicate the presence of misphenotyped individuals in the mapping population. This information can in turn be taken into account when identifying potential causative mutations.

For many organisms, a dense map of SNP markers is not yet available. We probed the usefulness of our method in such cases by performing de novo marker prediction using the same data set. Markers were defined at positions with significant support (here we used 15% minor allele frequency) for two alleles based on the SHORE base count per position output. A sliding window analysis of the position-wise average distance to the closest identified marker divided by the local marker density was performed with SHOREmap DENOVO (Figure 4-3B). This value is highest at the site of the causal mutation, since it resides in a pure Col-0 background in all selected recombinants. Running SHOREmap ANNOTATE on the candidate region defined by the peak ranked the causal mutation as the top candidate. As expected, the mapping accuracy suffered slightly using de novo predicted markers because of a drop in marker density in the candidate region. High genome coverage was critical for de novo marker prediction mapping (Figure 4-3B,C), since supporting two alleles during marker discovery demands a large number of reads per position. Subsampling of genome coverage revealed the minimum requirement for SHOREmap DENOVO to be approximately 11 fold, while 6-fold coverage moved the peak 1.2 Mb away from the causal mutation.

SHOREmap DENOVO determines marker positions based on the base count from the SHORE output. Each position featuring four or more reads (22-fold coverage) or two or more reads (11-fold coverage) from two alleles and a frequency of at least 15% for both of these alleles will be recorded as a marker position. The number of such marker positions in and near the mapping interval will be reduced due to the homozygous nature of the mapping interval. SHOREmap DENOVO determines the density within a sliding window as the average of the sum of position-wise distance to the nearest marker (dist_m) multiplied with the inverse sum of frequencies of the Ler-1 allele at the predicted marker positions (ler):

$$density_{window} = \frac{\sum \text{dist}_{\text{m}} * \frac{1}{\sum ler}}{\text{window size}} * ref^2,$$

where

$$ref = \frac{ref\ \ calls}{window\ \ size}$$

is the percentage of reference calls within the sliding window used to normalize for the accessibility of the genomic region to base calling mostly influenced by repetitiveness. The output plot from SHOREmap DENOVO is shown in Figure 4-4, and the commands used in Table 4-1. A target interval of 3 Mb was chosen based on the SHOREmap DENOVO visualization for which all included base changes were ranked using SHOREmap ANNOTATE. The causal mutation was ranked number one for both tested datasets (22-fold and 11-fold).

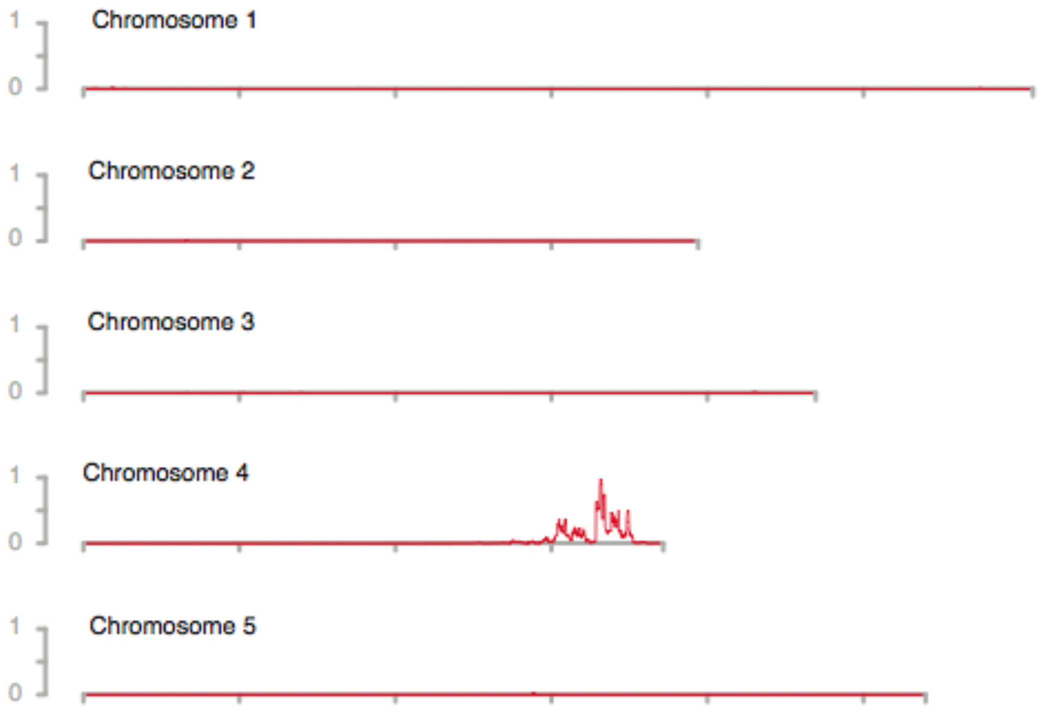

Figure 4-4: **SHOREmap DENOVO output.** Red lines visualize scarcity-values in a sliding window of 200 kb.

Compared to other approaches, our method combines mapping by recombinant genotyping and candidate gene sequencing into a single seamless step. This shortens the overall time required for genetic mapping from months to weeks and, importantly, greatly reduces investigator hands-on time. The steps requiring investigator input are: DNA isolation (1 day), library preparation and validation (4 days), Illumina cluster generation and sequencing (2 days), and data analysis (1 day). Once the mapping population has been established, the present method therefore allows a single investigator to identify a causative mutation within only eight working days – approximately an order of magnitude faster than with conventional methods. Since high-quality reference sequences and dense SNP maps are also available for other model organisms such as *Caenorhabditis elegans* and *Drosophila melanogaster* [109], [116], the method described here is broadly applicable across species. Furthermore, the *de novo* marker prediction mapping, as implemented in SHOREmap DENOVO, opens a door to advanced genetics for an even wider community working with species not yet considered genetically tractable.

4.2.3 Further application of SHOREmap

4.2.3.1 Mapping large deletions

Small deletions (approximately 1-10 bp) are annotated in the SHOREmap ANNOTATE output table. In case larger deletions are expected, for example after fast neutron mutagenesis, they should be considered as potential causal mutations within the candidate region. Since the number of large deletions within the candidate region will usually be relatively small, the relevant entries can easily be inspected manually.

4.2.3.2 Mapping recessive lethal or dominant mutations

Mapping of fully-penetrant gain-of-function mutations and recessive lethal mutations are two very interesting possible applications of our method. Consider two polymorphic strains, “A” and “B”. The mutant allele resides in the “A” background and is designated “a”.

For dominant mutations, the mapping cross could be:

aA x BB → aB aB AB AB or

aa x BB → aB aB aB aB

Select individuals presenting the mutant phenotype (aB) and self:

aB x aB → aa aB aB BB

Select individuals presenting mutant phenotype (aa and aB) and sequence pooled DNA. Since BB individuals are discarded, 2/3 of the alleles in the candidate region will be of “A” origin - a significant overrepresentation. Candidate causal mutations would be identified as positions where 2/3 of the bases at a given position do not match the “A” or “B” reference sequence.

For lethal recessive mutations, the mapping cross would be:

aA x BB → aB aB AB AB

Select aB individuals, based on segregation of dead aa offspring, and self:

aB x aB → aa aB aB BB

Select viable individuals (aB and BB) and sequence pooled DNA. Since aa individuals are discarded, 2/3 of the alleles in the candidate region will be of “B” origin. Candidate causal mutations would be identified as positions where 1/3 of the bases at a given position do not match the “A” or “B” reference sequence.

We expect that high genome coverage, at least 22x, will be necessary for these approaches to allow accurate determination of the candidate region and reliable identification of candidate mutations.

4.2.3.3 QTL mapping

We consider SHOREmap well suited for QTL mapping, but its success will depend on genetic architecture (how much variation is explained by one QTL) and, perhaps even more importantly, the number of sequence changes in the QTL region relative to the reference genome.

The strategy would be to use bulk segregant analysis and sequence two pools of recombinants at each extreme of the phenotypic distribution. SHOREmap analysis of these data would then provide accurate information about the relative representation of parental alleles at each genomic locus as well as data on the differences between parents in candidate regions. This approach is similar to eXtreme array mapping implemented by [26], which suffered from peaks being very broad. We expect that the digital signal from sequencing combined with SHOREmap analysis will lead to an improved signal to noise ratio and much sharper peaks.

Literature cited

1 Avery OT, Macleod CM, McCarty M (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. Clin Orthop Relat Res 379: S3-S8.

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

3 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

4 Chang C, Meyerowitz EM (1986) Molecular cloning and DNA sequence of the Arabidopsis thaliana alcohol dehydrogenase gene. Proc Natl Acad Sci U S A 83: 1408-1412.

5 Jander G, Norris SR, Rounsley SD, Bush DF, Levin IM, Last RL (2002) Arabidopsis map-based cloning in the post-genome era. Plant Physiol 129: 440-450.

6 Ziolkowski PA, Koczyk G, Galganski L, Sadowski J (2009) Genome sequence comparison of Col and Ler lines reveals the dynamic nature of Arabidopsis chromosomes. Nucleic Acids Res 37: 3189-3201.

7 Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, et al (2005) The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol 3: e196.

8 Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, et al (2007) Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science 317: 338-342.

9 Zeller G, Clark RM, Schneeberger K, Bohlen A, Weigel D, Rätsch G (2008) Detecting polymorphic regions in Arabidopsis thaliana with resequencing microarrays. Genome Res 18: 918-929.

10 Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res

18: 2024-2033.

11 Santuari L, Pradervand S, Amiguet-Vercher AM, Thomas J, Dorcey E, et al (2010) Substantial deletion overlap among divergent Arabidopsis genomes revealed by intersection of short reads and tiling arrays. Genome Biol 11: R4.

12 Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, et al (2009) Simultaneous alignment of short reads against multiple genomes. Genome Biol 10: R98.

13 Lam HM, Xu X, Liu X, Chen W, Yang G, et al (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet .

14 Lai J, Li R, Xu X, Jin W, Xu M, et al (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet 42: 1027-1030.

15 Sarin S, Bertrand V, Bigelow H, Boyanov A, Doitsidou M, et al (2010) Analysis of multiple ethyl methanesulfonate-mutagenized Caenorhabditis elegans strains by whole-genome sequencing. Genetics 185: 417-430.

16 Sarin S, Prabhu S, O'Meara MM, Pe'er I, Hobert O (2008) Caenorhabditis elegans mutant allele identification by whole-genome sequencing. Nat Methods 5: 865-867.

17 Blumenstiel JP, Noll AC, Griffiths JA, Perera AG, Walton KN, et al (2009) Identification of EMS-induced mutations in Drosophila melanogaster by whole-genome sequencing. Genetics 182: 25-32.

18 Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, et al (2008) Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res 18: 1638-1642.

19 Irvine DV, Goto DB, Vaughn MW, Nakaseko Y, McCombie WR, et al (2009) Mapping epigenetic mutations in fission yeast using whole-genome next-generation sequencing. Genome Res 19: 1077-1083.

20 Laitinen RA, Schneeberger K, Jelly NS, Ossowski S, Weigel D (2010) Identification of a spontaneous frame shift mutation in a nonreference Arabidopsis

accession using whole genome sequencing. Plant Physiol 153: 652-654.

21 Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, et al (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Methods 6: 550-551.

22 Doitsidou M, Poole RJ, Sarin S, Bigelow H, Hobert O (2010) C. elegans mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy. PLoS One 5: e15435.

23 Zuryn S, Le Gras S, Jamet K, Jarriault S (2010) A Strategy for Direct Mapping and Identification of Mutations by Whole Genome Sequencing. Genetics .

24 Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, et al (2010) Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature 464: 1039-1042.

25 Michelmore RW, Paran I, Kesseli RV (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.

26 Wolyn DJ, Borevitz JO, Loudet O, Schwartz C, Maloof J, et al (2004) Light-response quantitative trait loci identified with composite interval and eXtreme array mapping in Arabidopsis thaliana. Genetics 167: 907-917.

27 Lai CQ, Leips J, Zou W, Roberts JF, Wollenberg KR, et al (2007) Speed-mapping quantitative trait loci using microarrays. Nat Methods 4: 839-841.

28 Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.

29 Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3: e3376.

30 Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, et al (2008) Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods 5: 887-893.

31 Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res 36: e122.

32 Huang X, Feng Q, Qian Q, Zhao Q, Wang L, et al (2009) High-throughput genotyping by whole-genome resequencing. Genome Res 19: 1068-1076.

33 Singer T, Fan Y, Chang HS, Zhu T, Hazen SP, Briggs SP (2006) A high-resolution map of Arabidopsis recombinant inbred lines by whole-genome exon array hybridization. PLoS Genet 2: e144.

34 Edwards JD, Janda J, Sweeney MT, Gaikwad AB, Liu B, et al (2008) Development and evaluation of a high-throughput, low-cost genotyping platform based on oligonucleotide microarrays in rice. Plant Methods 4: 13.

35 Xie W, Feng Q, Yu H, Huang X, Zhao Q, et al (2010) Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. Proc Natl Acad Sci U S A 107: 10578-10583.

36 Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, et al (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods 7: 111-118.

37 Nijman IJ, Mokry M, van Boxtel R, Toonen P, de Bruijn E, Cuppen E (2010) Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. Nat Methods .

38 Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, et al (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42: 348-354.

39 Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, et al (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465: 627-631.

40 Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, et al (2009) A First-Generation Haplotype Map of Maize. Science 326: 1115.

41 McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, et al (2009)

Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proc Natl Acad Sci U S A 106: 12273.

42 Huang X, Wei X, Sang T, Zhao Q, Feng Q, et al (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42: 961-967.

43 De Bona F, Ossowski S, Schneeberger K, Rätsch G (2008) Optimal spliced alignments of short sequence reads. Bioinformatics 24: i174-i180.

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

45 Ma B, Tromp J, Li M (2002) PatternHunter: faster and more sensitive homology search. Bioinformatics 18: 440.

46 Malhis N, Butterfield YS, Ester M, Jones SJ (2009) Slider--maximum use of probability information for alignment of short sequence reads and SNP detection. Bioinformatics 25: 6-13.

47 Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443-453.

48 Ukkonen E (1992) Approximate string-matching with q-grams and maximal matches. Theoretical Computer Science 92: 191-211.

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

50 Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, et al (2008) Whole-genome sequencing and variant discovery in C. elegans. Nat Methods 5: 183-188.

51 Nusbaum C, Ohsumi TK, Gomez J, Aquadro J, Victor TC, et al (2009) Sensitive, specific polymorphism discovery in bacteria using massively parallel sequencing. Nat Methods 6: 67-69.

52 Campagna D, Albiero A, Bilardi A, Caniato E, Forcato C, et al (2009) PASS: a program to align short sequences. Bioinformatics 25: 967-968.

53 Coarfa C, Milosavljevic A (2008) Pash 2.0: scaleable sequence anchoring for

next-generation sequencing technologies. Pac Symp Biocomput : 102-113.

54 Eaves HL, Gao Y (2009) MOM: maximum oligonucleotide mapping. Bioinformatics 25: 969-970.

55 Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, et al (2010) Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics 26: i350-i357.

56 Jiang H, Wong WH (2008) SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics 24: 2395-2396.

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

58 Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24: 713-714.

59 Li R, Yu C, Li Y, Lam TW, Yiu SM, et al (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966-1967.

60 Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11: 1725-1729.

61 Prüfer K, Stenzel U, Dannemann M, Green RE, Lachmann M, Kelso J (2008) PatMaN: rapid alignment of short sequences to large databases. Bioinformatics 24: 1530-1531.

62 Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M (2009) SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol 5: e1000386.

63 Schatz MC (2009) CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 25: 1363-1369.

64 Smith AD, Xuan Z, Zhang MQ (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 9: 128.

65 Weese D, Emde AK, Rausch T, Döring A, Reinert K (2009) RazerS--fast read mapping with sensitivity control. Genome Res 19: 1646-1654.

66 Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, et al (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41: 1061-1067.

67 Lin H, Zhang Z, Zhang MQ, Ma B, Li M (2008) ZOOM! Zillions of oligos mapped. Bioinformatics 24: 2431-2437.

68 Initiative TAG (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796-815.

69 Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7: 85-97.

70 Ahn SM, Kim TH, Lee S, Kim D, Ghang H, et al (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res 19: 1622-1629.

71 Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53-59.

72 Springer NM, Ying K, Fu Y, Ji T, Yeh CT, et al (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 5: e1000734.

73 Kaiser J (2008) DNA sequencing. A plan to capture human diversity in 1000 genomes. Science 319: 395.

74 Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, et al (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 40: 722-729.

75 Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, et al (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods 6: 99-103.

76 Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318:

420-426.

77 Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, et al (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol 10: R23.

78 Lam HY, Mu XJ, Stütz AM, Tanzer A, Cayting PD, et al (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat Biotechnol 28: 47-55.

79 Lee I, Ajay SS, Yook JI, Kim HS, Hong SH, et al (2009) New class of microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites. Genome Res 19: 1175-1183.

80 Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, et al (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res 20: 623-635.

81 Wang K, Li M, Hadley D, Liu R, Glessner J, et al (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17: 1665-1674.

82 Li R, Fan W, Tian G, Zhu H, He L, et al (2010) The sequence and de novo assembly of the giant panda genome. Nature 463: 311-317.

83 Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, et al (2010) Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. PLoS Biol 8: e1000475.

84 Young AL, Abaan HO, Zerbino D, Mullikin JC, Birney E, Margulies EH (2010) A new strategy for genome assembly using short sequence reads and reduced representation libraries. Genome Res 20: 249-256.

85 Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19: 1117-1123.

86 Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821-829.

87 Chaisson MJ, Pevzner PA (2008) Short read fragment assembly of bacterial genomes. Genome Res 18: 324-330.

88 Pop M, Kosack DS, Salzberg SL (2004) Hierarchical scaffolding with Bambus. Genome Res 14: 149-159.

89 Pop M, Phillippy A, Delcher AL, Salzberg SL (2004) Comparative genome assembly. Brief Bioinform 5: 237-248.

90 Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al (2005) Fine-scale structural variation of the human genome. Nat Genet 37: 727-732.

91 Lee S, Cheran E, Brudno M (2008) A robust framework for detecting structural variations in a genome. Bioinformatics 24: i59-i67.

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

93 Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, et al (2010) The characterization of twenty sequenced human genomes. PLoS Genet 6: e1001111.

94 Plantegenet S, Weber J, Goldstein DR, Zeller G, Nussbaumer C, et al (2009) Comprehensive analysis of Arabidopsis expression level polymorphisms with simple inheritance. Mol Syst Biol 5: 242.

95 Naouar N, Vandepoele K, Lammens T, Casneuf T, Zeller G, et al (2009) Quantitative RNA expression analysis with Affymetrix Tiling 1.0R arrays identifies new E2F target genes. Plant J 57: 184-194.

96 Laubinger S, Zeller G, Henz SR, Sachsenberg T, Widmer CK, et al (2008) At-TAX: a whole genome tiling array resource for developmental expression analysis and transcript identification in Arabidopsis thaliana. Genome Biol 9: R112.

97 Theologis A (2001) Goodbye to 'one by one' genetics. Genome Biol 2: comment2004.1-comment2004.9.

98 Chory J, Ecker JR, Briggs S, Caboche M, Coruzzi GM, et al (2000) National Science Foundation-Sponsored Workshop Report:" The 2010 Project" functional genomics and the virtual plant. A blueprint for understanding how plants are built and

how to improve them. Plant Physiol 123: 423.

99 McCourt P, Benning C (2010) Arabidopsis: a rich harvest 10 years after completion of the genome sequence. Plant J 61: 905-908.

100 Rounsley SD, Last RL (2010) Shotguns and SNPs: how fast and cheap sequencing is revolutionizing plant biology. Plant J 61: 922-927.

101 Huang S, Li R, Zhang Z, Li L, Gu X, et al (2009) The genome of the cucumber, Cucumis sativus L. Nat Genet 41: 1275-1281.

102 Shoemaker R, Deng J, Wang W, Zhang K (2010) Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res 20: 883-889.

103 Lister R, Gregory BD, Ecker JR (2009) Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. Curr Opin Plant Biol 12: 107-118.

104 Fu H, Zheng Z, Dooner HK (2002) Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc Natl Acad Sci U S A 99: 1082-1087.

105 Rasbery JM, Shan H, LeClair RJ, Norman M, Matsuda SP, Bartel B (2007) Arabidopsis thaliana squalene epoxidase 1 is essential for root and seed development. J Biol Chem 282: 17002-17013.

106 Posé D, Castanedo I, Borsani O, Nieto B, Rosado A, et al (2009) Identification of the Arabidopsis dry2/sqe1-5 mutant reveals a central role for sterols in drought tolerance and regulation of reactive oxygen species. Plant J 59: 63-76.

107 St Johnston D (2002) The art and design of genetic screens: Drosophila melanogaster. Nat Rev Genet 3: 176-188.

108 Page DR, Grossniklaus U (2002) The art and design of genetic screens: Arabidopsis thaliana. Nat Rev Genet 3: 124-136.

109 Chen D, Ahlford A, Schnorrer F, Kalchhauser I, Fellner M, et al (2008) High-resolution, high-throughput SNP mapping in Drosophila melanogaster. Nat Methods

5: 323-329.

110 Hazen SP, Borevitz JO, Harmon FG, Pruneda-Paz JL, Schultz TF, et al (2005) Rapid array mapping of circadian clock and developmental mutations in Arabidopsis. Plant Physiol 138: 990-997.

111 Greene EA, Codomo CA, Taylor NE, Henikoff JG, Till BJ, et al (2003) Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in Arabidopsis. Genetics 164: 731-740.

112 Gong JM, Waner DA, Horie T, Li SL, Horie R, et al (2004) Microarray-based rapid cloning of an ion accumulation deletion mutant in Arabidopsis thaliana. Proc Natl Acad Sci U S A 101: 15404-15409.

113 Mitra RM, Gleason CA, Edwards A, Hadfield J, Downie JA, et al (2004) A Ca2+/calmodulin-dependent protein kinase required for symbiotic nodule development: Gene identification by transcript-based cloning. Proc Natl Acad Sci U S A 101: 4701-4705.

114 Gresham D, Ruderfer DM, Pratt SC, Schacherer J, Dunham MJ, et al (2006) Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. Science 311: 1932-1936.

115 Giovannoni JJ, Wing RA, Ganal MW, Tanksley SD (1991) Isolation of molecular markers from specific chromosomal intervals using DNA pools from existing mapping populations. Nucleic Acids Res 19: 6553-6558.

116 Shen Y, Sarin S, Liu Y, Hobert O, Pe'er I (2008) Comparing platforms for C. elegans mutant identification using high-throughput whole-genome sequencing. PLoS One 3: e4012.