Review

Fast-forward genetics enabled by new sequencing technologies

Korbinian Schneeberger[1,2] and Detlef Weigel[1]

[1] Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany
[2] Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany

New sequencing technologies are dramatically accelerating progress in forward genetics, and the use of such methods for the rapid identification of mutant alleles will be soon routine in many laboratories. A straightforward extension will be the cloning of major-effect genetic variants in crop species. In the near future, it can be expected that mapping by sequencing will become a centerpiece in efforts to discover the genes responsible for quantitative trait loci. The largest impact, however, might come from the use of these strategies to extract genes from non-model, non-crop plants that exhibit heritable variation in important traits. Deployment of such genes to improve crops or engineer microbes that produce valuable compounds heralds a potential paradigm shift for plant biology.

A brief history of genotype analysis since 1869

It was above the city of Tübingen, Germany, in Hohentübingen Castle, that DNA was isolated for the first time in 1869, by the young Swiss biochemist Friedrich Miescher [1]. It was another 75 years before DNA was demonstrated to be the material from which genes are made [2]. Genes contain the information for the development and performance of organisms, and they interact with the environment to produce an individual's phenotype. In some cases, the influence of the environment can be ignored, whereas in others, it largely obscures the contribution of the genotype. Bridging the phenotype–genotype divide thus requires us to simultaneously record environment, genotype and phenotype. On the genotype side, whether one is performing genome-wide association studies (GWAS) in distantly related individuals or linkage mapping in experimentally created populations; one first needs to determine genetic polymorphisms that segregate among phenotypically differentiated individuals. The identification of genetic variants has become easier with the advent of next generation sequencing (NGS) methods [3], which support the rapid and inexpensive whole genome resequencing of hundreds or even thousands of individuals. Here, we will briefly discuss how technological advances during the past few years have improved the detection of sequence polymorphisms. We will then outline how deep sequencing can be exploited in the near to medium term to support new forward genetic approaches. Several practical suggestions and considerations are included in this section. We are aware that NGS methods also have a large impact on GWAS applications; for a discussion of this, the reader is referred to other sources [4–6].

Corresponding author: Weigel, D. (weigel@weigelworld.org).

A brief history of resequencing

The first sequence of a fragment of *Arabidopsis thaliana* genomic DNA, 3176 bp surrounding the *ADH* locus, was published in 1986 [7]; only 14 years later, almost the entire genome had been decoded. The release of the 115,409,949 bp genome assembly comprising ten contigs, one for each of the chromosome arms, was a milestone for biology [8] (Figure 1). It not only accelerated the positional cloning of genes identified from mutant phenotypes, but also provided often surprising insights into the evolution of plant genomes [9].

Although this landmark achievement was introduced as an "analysis of the completed *Arabidopsis* genome" [8], the authors recognized that the reference sequence, produced from the commonly used laboratory strain Columbia (Col-0), would serve as a platform for the "parallel analysis of genome-wide polymorphisms and quantitative traits". This was reinforced by comparing the Col-0 genome, which had been assembled primarily from individually analyzed bacterial artificial chromosomes, with a low-coverage, whole-genome shotgun sequence of another strain popular with geneticists, Landsberg *erecta* (L*er*). Alignments against 82 Mb of the finished reference sequence were scanned for single nucleotide polymorphisms (SNPs), as well as insertions and deletions (indels). In total, 25,274 SNPs, or one SNP every 3.3 kb, and 14,570 indels, ranging in size from 2 bp to 38 kb, were identified. The resulting resource eliminated the tedious step of polymorphism discovery, and it has been estimated that, together with the reference genome sequence and improved genotyping methods, this reduced the total effort required for positional cloning of a gene from up to five person-years to less than one person-year [10]. It was realized early on that a polymorphism distinguishing Col-0 and L*er* had an approximately 50% chance of being informative in a cross of Col-0 with another strain [11]; thus, the L*er* polymorphisms were useful in other mapping crosses. The reference sequence was also exploited for targeted discovery and evaluation of polymorphisms across many *Arabidopsis* strains, using PCR amplification of discrete loci. These studies greatly increased the estimates of polymorphism density, to about 1 SNP per 200 bp, when comparing a randomly chosen strain with the reference [12–14].

The first attempt to describe sequence diversity in the entire *Arabidopsis* genome made use of large-scale microarrays. Hybridizing genomic DNA from 20 divergent strains to tiling arrays with almost one billion different oligonucleotides increased the number of known SNPs in

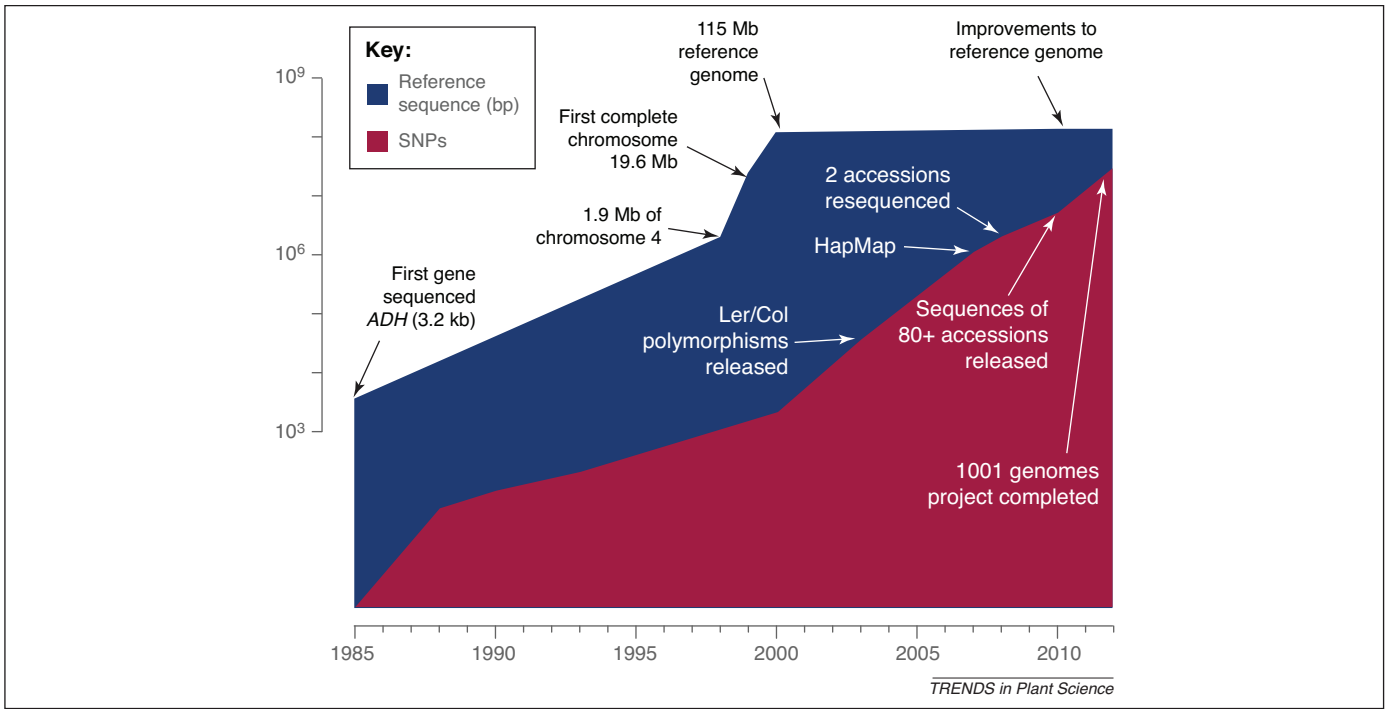

Figure 1. Timeline of *Arabidopsis* sequencing and resequencing efforts. SNPs are used as a proxy for all types of polymorphisms. The numbers of publicly available SNPs before 2003 and after 2010 are estimated.

Arabidopsis by two orders of magnitude, and provided the foundation for the first haplotype map outside of mammals [15,16]. Despite its success, this study, along with a similar one in rice (*Oryza sativa*) [17], marked the twilight of the era of microarray-based resequencing as NGS technologies became available. The first NGS analyses were performed with reads that were not longer than approximately 40 bp, but more than ten times as many SNPs could be identified in a single *Arabidopsis* genome, at a fraction of the cost of the microarray approach [18]. NGS methods also greatly reduced the false discovery rate, which could be as high as 6% in array data, to below 1% [18]. In addition, methods for robust detection of copy number and structural variants from short read data were introduced. Perhaps the most significant advance was that local assemblies, pioneered in *Arabidopsis*, enabled the discovery of over 10,000 diverged regions that harbored indels as long as 600 bp [18]. Aberrant signals on tiling arrays [15,19,20] supported the assignment of such regions. *Arabidopsis* also provided the first example of exploiting previously discovered polymorphism to improve the short read analysis of additional individuals [21]. We believe that a consequent further step will be the use of both a reference sequence and known polymorphisms to generate, in an iterative manner, a homology-guided assembly of diverged genomes.

For species other than *Arabidopsis* and rice, large-scale polymorphism discovery began only with NGS methods; published examples include soybean (*Glycine max*) [22] and maize (*Zea mays*) [23,24]. In the latter species, the extremely high rates of presence–absence polymorphisms that had already been inferred from array hybridization [25] were confirmed in an impressive manner. Similarly, a much improved haplotype map has been generated for rice [26], and many more such efforts are under way. Currently the most ambitious effort might be the 1001 Genomes Project for *Arabidopsis* (http://1001genomes.org), but it is likely that we will soon see much larger efforts for crop and tree species. A major unsolved issue is how to deal with the original reference sequences, which are not always optimal for annotating variants in divergent genomes. Despite this challenge, we believe that, for the community to take full advantage of these resources, the reference genome sequences must remain stable, because the reference positions are the basis for interpreting and comparing resequencing data.

A concern with NGS methods has often been the accuracy of sequence determination, given that reads are relatively short and that the per bp accuracy of individual reads is generally much lower than in reads generated by Sanger dideoxy sequencing. However, this limitation can easily be overcome with the increased coverage that is possible with NGS technologies. In one of the first applications, using reads of only approximately 40 bp, it was possible to comprehensively detect spontaneous single base pair mutations that were present at a frequency of 1 in 5 million bp in the genomes of different *Arabidopsis* strains [27]. The false positive rate with a simple consensus approach was around 1%, and could be lowered further by using a more sophisticated maximum likelihood approach [28]. Importantly, the false negative rate was estimated to be only 10%. In this specific case, it helped that the spontaneous mutations had accumulated in a strain that was closely related to the one used to produce the reference genome sequence for *Arabidopsis*, and that *Arabidopsis* can be easily selfed. Recent successes with human genomes demonstrate that the challenges posed by larger and heterozygous genomes can also be met [29,30]. It must be cautioned that the identification of more complex mutations, such as insertions, deletions and inversions, is more demanding, but given the improvements that have been

made with current NGS methods over the past few years, and in anticipation of the new methods already on the horizon, it is clear that many of these challenges will disappear [3].

Current status of NGS-enabled genetics

NGS methods support the comprehensive, high confidence discovery of very rare, spontaneously arising mutations [27,31,32]; it should thus be possible to identify induced variants responsible for a mutant phenotype directly. Unfortunately, a chemical mutagen such as ethyl methanesulfonate (EMS) generates several mutations per Mb [33], and direct sequencing of one mutant genome will not suffice. The most straightforward approach – for which there are, however, no published reports – is to identify lesions in two or more allelic mutants (Figure 2). The alternative is to first narrow down the region containing the causal mutation by mapping. Initial attempts with *Caenorhabditis elegans*, which has a genome similar in size to that of Arabidopsis, used prior mapping information so only a small fraction of the genome had to be analyzed; nevertheless, the authors were left with several candidates for the causal variant [34]. Using the principle of bulked segregant analysis, which was invented by plant geneticists [35], it has been shown that one can map a mutation and pinpoint the causal change in a single sequencing reaction (Figure 3). The SHOREmap pipeline developed

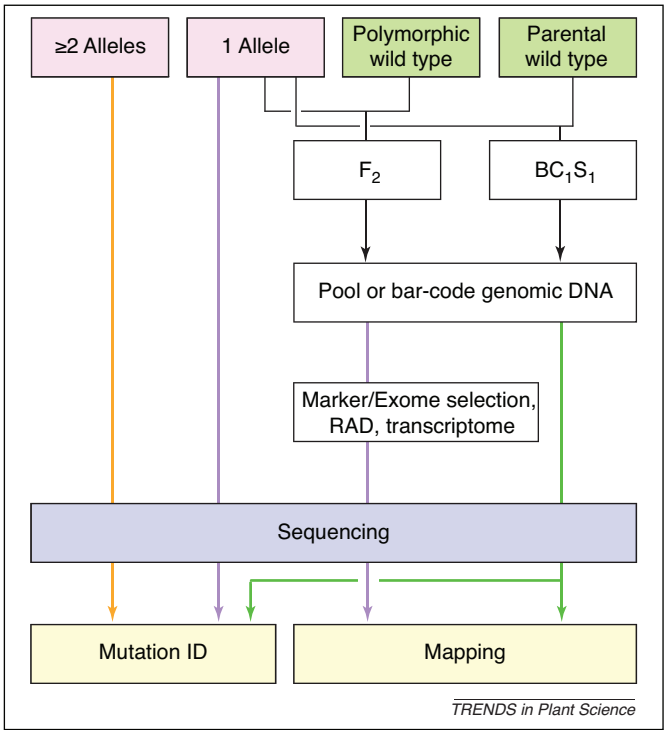

Figure 2. Three main strategies for mutation identification. If the purple path is taken, a second sequencing reaction might be required to discover candidate mutations in the final mapping interval. BC_1S_1, backcross followed by selfing.

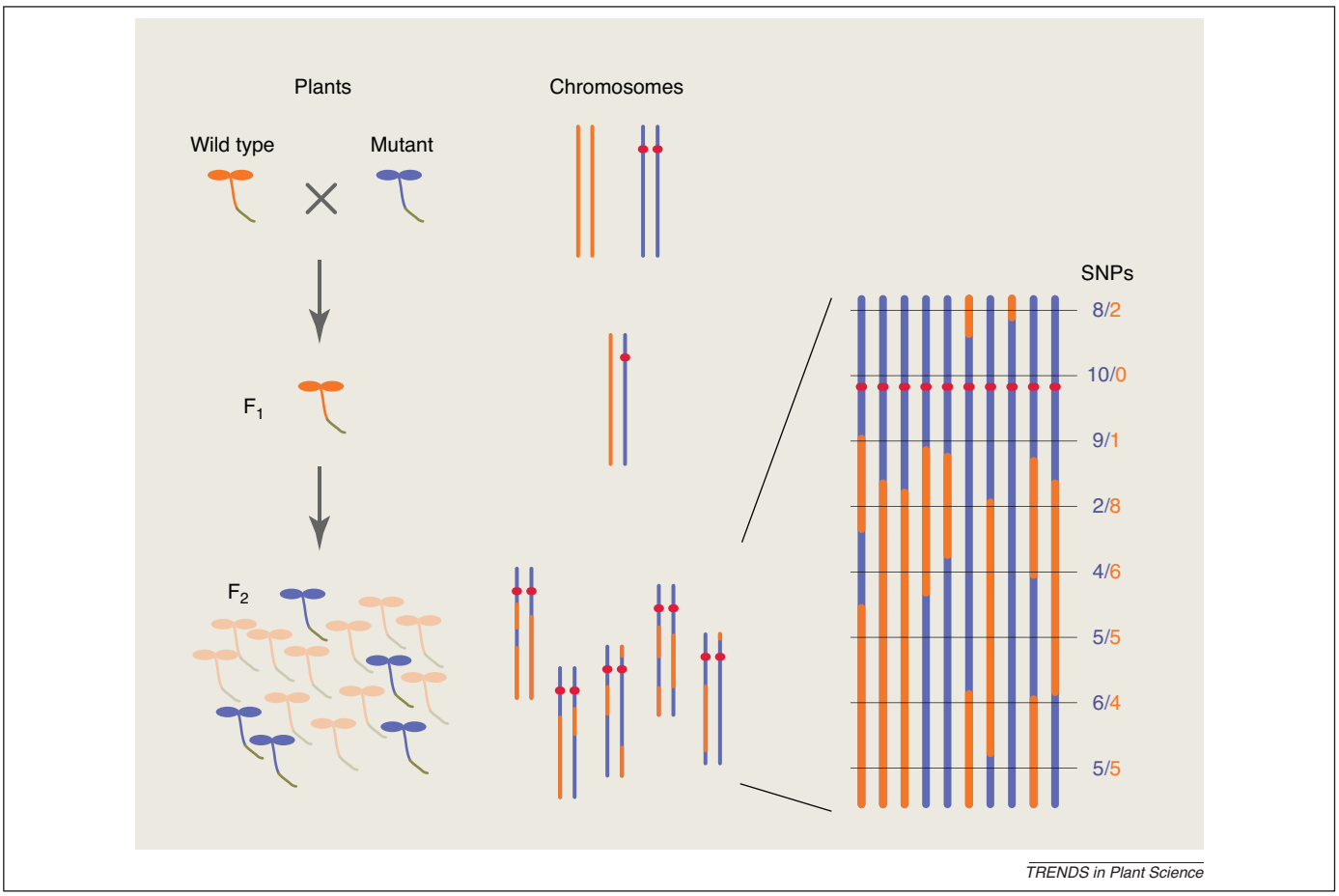

Figure 3. Diagram of SHOREmap [36] bulked segregant sequencing. Red indicates a mutation that is causal for the mutant phenotype. Note that a metric that reflects deviation from the expected 1:1 SNP ratio is needed to robustly identify the final mapping interval, rather than merely plotting raw SNP ratios.

for this purpose incorporates several modules, from mapping to *de novo* marker identification during the sequencing process, and finally annotation of candidate mutations [36].

How can the SHOREmap method perform so well, although only a small fraction of the genome of each individual recombinant plant is sequenced? The secret is that one can combine information from adjacent markers. This approach was pioneered in rice for the characterization of recombinant inbred lines, which were sequenced as multiplexed, bar-coded samples at merely 0.02× coverage per line. Despite this very low coverage, a resolution of recombination breakpoints of 40 kb was achieved [37]. Thus, one can effectively interrogate many more recombinant chromosomes than are analyzed at individual marker positions. For example, with a density of 1 marker/kb, and 20-fold coverage, the equivalent of 2000 chromosomes is analyzed within a 100 kb window, because the short reads constitute random sampling of independent chromosomes. When using large bulked segregant populations, containing hundreds of plants, all individuals contribute to the definition of the mapping interval. In conventional mapping of *Arabidopsis* mutations with individual F_2 recombinants, the final mapping interval from 1000 chromosomes would be 0.1 cM, or an average of 20 kb, which is much less than the predicted density of EMS mutations. An important question is how many individuals are required for this approach, because some mutant phenotypes are difficult to score in hundreds of plants. As with conventional methods, the higher the number of recombinants analyzed, the smaller the final mapping interval. Nevertheless, because the density of induced mutations is normally only of the order of one per 100–200 kb, which translates to not much less than a centiMorgan in *Arabidopsis*, dozens of individuals should in principle suffice for SHOREmap analysis. Moreover, most point mutations affect either the open reading frame or splice sites, enabling prioritization of candidate mutations for subsequent validation.

The steps requiring investigator input in this scheme are: DNA isolation (1 day), library preparation and validation (4 days), sequencing (2 days) and data analysis (1 day). Once the mapping population has been established, this method allows a single investigator to identify a causative mutation within only eight working days – approximately an order of magnitude faster than previous methods.

NGS-based mutation identification has been applied to other genetic model systems, including *Drosophila melanogaster*, *C. elegans* and *Saccharomyces cerevisiae* [38–40]. Experience with these organisms has shown that mutation identification can be confounded by spontaneous mutations or background polymorphisms, which is a particular issue in species that are not easily inbred [34]. It is therefore advisable also to sequence individuals from the generation that was used for mutagenesis. Importantly, it has been demonstrated that the mutant individual does not have to come from the reference strain. For example, even rare *de novo* mutations causal for a genetic defect can be identified in a non-reference *Arabidopsis* background [41], as can recently mobilized transposons [42]. Similar analyses in humans indicate that these methods are applicable to outbred species with much larger genomes [29,30], although this becomes more challenging when there are many paralogs, especially in plants with extremely high rates of presence–absence polymorphisms such as maize [23–25]. Methods for *de novo* assembly of short read data, which are becoming increasingly powerful [43], will be helpful in this regard. Such methods will also aid the identification of more complex mutations than SNPs or small indels.

Near-term prospects for NGS-enabled genetics

The SHOREmap approach discussed above (Figure 3) requires the generation of mapping populations by outcrossing to a polymorphic wild-type strain. However, some phenotypes are very sensitive to the genetic background and unambiguous identification of mutant plants is difficult. The simplest solution to this problem would be direct sequencing of two or more independently generated alleles, and a subsequent search of the genome for genes that carry unique new mutations in the same gene. If only one allele is at hand, an alternative is to use the non-causal mutations induced by EMS as novel markers. Although in principle this could be exploited by sequencing pooled progeny of the original mutant, we note that in such an individual, one-third of EMS-induced changes will be homozygous (which follows from the 1:2:1 segregation of new mutations and the ancestral alleles). If possible, at least one backcross to the original wild-type strain should be performed, to ensure that all EMS-induced changes segregate in the next generation (Figure 2). Such a strategy, which removes mutagen-induced nucleotide changes that are not linked to the causal mutation, has already been applied to *C. elegans* [44]. The mutants were sequenced after four to six rounds of backcrossing (although this many backcrosses might not be necessary). Performing independent backcrosses and then sequencing pooled, selfed progeny of backcrossed plants should reduce the number of novel variants that are homozygous but unlinked to the causal mutation.

It might also be desirable to improve the mapping resolution. In the original SHOREmap approach, similar to other methods in which individual chromosomes are only lightly sequenced [37], the entire genome is analyzed. The advantage is that no separate sequencing reaction is required for mutation identification. The disadvantage is that most reads are not informative, because they are from regions that are not polymorphic. Mapping resolution could be improved by analyzing only polymorphic markers. In one of the simplest implementations, one would enrich for sequences at known polymorphic sites using sequence capture technology [45–47]. In addition, information could probably be gained by multiplex sequencing of bar-coded individuals, instead of sequencing pooled genomic DNA [37,48,49]. This would also be helpful when phenotyping is difficult, because it would flag rare individuals that carry wild-type alleles in the most probable mapping interval.

Because polymorphism density is usually much higher than the frequency of recombination breakpoints, it is not necessary to analyze all polymorphic sites, which is of particular interest when dealing with large genomes. If no polymorphisms are known, sequence complexity can be

reduced by the analysis of size-selected restriction fragments [50] or restriction site associated DNA fragments [51], capture of the exome (the transcribed portion of the genome) [45–47] or sequencing of the transcriptome. Exome capture and transcriptome sequencing both have the advantage that many mutations reside in open reading frames, and the causal mutation could thus potentially be discovered without an additional whole-genome sequencing step. Transcriptome sequencing has the further advantage that it does not require species-specific enrichment strategies, the disadvantage being that the total amount of sequencing data that has to be generated could be large, because of the greatly varying representation of individual transcripts in the sequencing library. We note that beside the causal mutation, the genome is homozygous, and that no matter what strategy is being used, information from non-affected individuals or the wild-type parent is needed to recognize novel variants.

Extending NGS-enabled genetics to quantitative traits

Two important yet challenging opportunities for NGS-enabled genetics are mapping genes for quantitative phenotypes and the identification of genes conferring specific attributes in plants with unsequenced genomes. Many agriculturally important traits do not have simple genetics, but are determined by collections of alleles that act in a quantitative manner. The conventional way of identifying such alleles has been through quantitative trait locus (QTL) mapping, which statistically associates genetic markers with specific phenotypes. Even more so than for the mapping of Mendelian traits, it is the case that the larger the population, the more precisely QTL will be mapped. This is because only large populations contain a sufficient number of individuals in which different QTL alleles are assorted in an informative manner. Similarly, if independently acting QTL are linked on the same chromosome, many individuals are required to recover a sufficient number of individuals in which linkage between these QTL is broken. Nevertheless, contrary to conventional wisdom [52], it has been shown that the increased resolution of recombination breakpoints afforded by high-density genotyping with NGS methods greatly improves QTL mapping in accurately phenotyped populations [37,53]. With barcoding methods, it is now possible to simultaneously analyze at least 96 samples in a single NGS reaction [54], and one can expect that this will be further improved. Thus, for small populations of up to a few hundred individuals, the analysis of individual, bar-coded DNA samples is probably preferable to that of pooled DNAs. Even when focusing on phenotypic extremes, DNA pools provide only limited mapping resolution [55–57].

This situation changes with an increase in the number of phenotyped individuals. The size of the population available for phenotypic selection then becomes the main limitation. In an impressive example from *S. cerevisiae*, it was possible to identify 14 loci that explained 70% of the genetic variance for the trait of interest. Several of these loci were mapped to single-gene resolution, which was achieved by starting with a population of several million individuals and selecting 0.5% of all individuals at either end of the phenotypic distribution [58]. Although this is beyond what is ordinarily feasible for plants, simulations suggest that, depending on the number and effect size of the QTL, the heritability of the trait and the number of chromosomes in the genome, as few as 10,000 individuals are sufficient for highly accurate QTL mapping using pools of phenotypically extreme individuals [58].

The resolution of QTL analysis can also be increased by combining the mapping in populations derived from controlled crosses with GWAS. GWAS provides very high resolution because it can take advantage of the many recombination events that have occurred since the different individuals in natural populations diverged from a common ancestor.. However, in contrast to mapping in populations derived from controlled crosses, the individuals might not be equally related (or unrelated), and this can greatly confound GWAS results. Mapping in experimental populations can help to distinguish spurious associations from causal ones, making the combination of the two strategies – known as joint linkage–linkage disequilibrium (LD) or joint linkage-association mapping – particularly powerful, even in highly polymorphic species with large genomes and complex histories such as maize [6,59–62].

Extending NGS-enabled genetics to unsequenced genomes

Although genome sequencing projects are under way for all major crops, not all alleles and genes of interest to breeders can be found in the gene pool of domesticated species. Notable examples are genes that encode pathways for valuable metabolites [63] or for disease resistance [64–66]. Ploidy barriers often prevent introgression of such alleles from near relatives through conventional means, and even where wide crosses are possible, this is a tedious strategy. In other cases, the goal is not to transfer these genes to other plants, but to microbes [63]. Before the advent of NGS methods, developing the necessary resources required for fine mapping and gene identification would have generally been prohibitive. Because of the much lower costs of NGS approaches, the only prerequisites now required are that there is heritable variation for the trait of interest and that one can make experimental crosses.

Even without a genome sequence, it might be possible to identify genes underlying traits with simple genetics by bulked segregant analysis only, especially when they are likely to belong to a specific family, as is the case for many disease resistance genes [67]. There are many variations regarding how one can proceed, but a good strategy will often be to first define the transcriptome, which can now be assembled from NGS reads [68,69], followed by exome sequencing [45–47]. Apart from ploidy changes, the number of genes encoded in a genome, and thus the total length of the transcriptome, does not vary greatly.

With a more complex genetic architecture, it will generally be necessary to first produce a genome sequence. The increase in read lengths expected from the third generation of single-molecule sequencing technologies [3] will greatly facilitate accurate and complete genome assembly. Ultimately, we envision that genome sequencing will entail reading an entire chromosome from one telomere to the

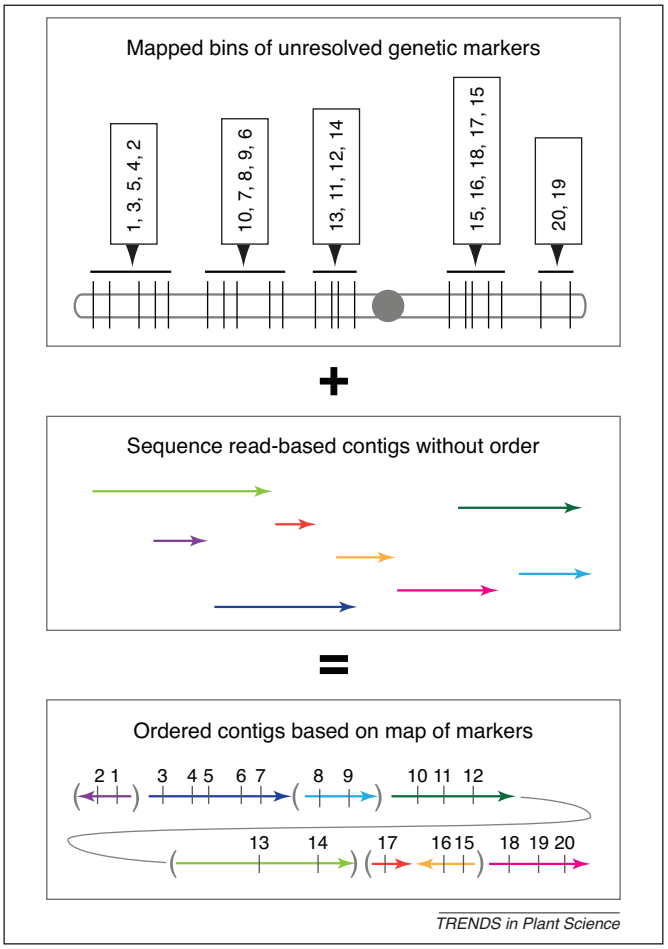

Figure 4. Synergistic use of different NGS methods for assembling genome sequences. An ultra-high-resolution genetic map generated with the help of NGS analysis can be used to order relatively short contigs produced by assembly of NGS genomic data on a chromosomal scale. The density of polymorphic markers usually exceeds the density of crossover events, leading to bins of markers that cannot be resolved by genetic mapping alone (top). By contrast, assembled sequence contigs or scaffolds generally lack long-distance information (middle). Assigning markers to contigs (i.e. by aligning marker sequences to contigs) enables linking of the physical and genetic maps (bottom). Parentheses enclose contigs of unknown orientation.

other. Until then, we believe that high-resolution genetic maps will play an increasingly important role in supporting genome assemblies. In most species, each chromosome experiences on average one crossover in meiosis. Thus, in a diploid species with a 1 Gb genome and 10 chromosomes, 1000 F_2 individuals afford an average distance between crossovers of about 50 kb. A tenfold excess of SNPs would require only 200,000 informative markers, which is achievable with current NGS methods. A distance of 50 kb is close to the contig size that should soon be within reach of NGS short read technologies such as the one from Illumina (Figure 4).

Conclusion and outlook

New sequencing methods have already had a measurable impact on forward genetics. In *Arabidopsis*, conventional mapping of mutations with major phenotypic effects is rapidly becoming obsolete. Application of NGS-enabled methods such as SHOREmap [36] to the cloning of mutant genes from species with larger genomes is straightforward, and can be easily extended to natural variants, the main constraint being the mapping resolution, which now is limited only by the number of phenotyped individuals. Similarly, we foresee that NGS-enabled methods will dramatically accelerate QTL cloning. Although the joint linkage-LD or linkage association mapping approach [6] eliminates the need to generate very large pedigrees from crosses, it is applicable only when the alleles of interest are suitably common in natural populations. For rare alleles, there is no simple alternative to linkage mapping.

We are most excited about the use of NGS-enabled genetics in non-model, non-crop plant species. Plant phenotypes and traits are diverse, and plants harbor many genes that are valuable for breeding or biotechnological applications. NGS-enabled genetics is providing a universal and generic platform for the isolation of these genes, even more so when we consider the next generation of NGS methods, with higher sequence output and longer read lengths [3]. We foresee that the much larger gene pool that is now becoming accessible will be the foundation for a paradigm shift not only in plant breeding, but for other practical uses of plant genes.

Conflict of interests

The authors declare no conflicts of interests.

Acknowledgments

We thank Jörg Hagmann, Dan Koenig, Jonas Müller and Beth Rowan for comments on the manuscript. The development of methods for next-generation genetics in the Weigel laboratory has been generously supported by FP6 IP AGRON-OMICS (contract LSHG-CT-2006-037704), a Gottfried Wilhelm Leibniz Award of the DFG, and the Max Planck Society.

References

1 Miescher, F. (1871) Ueber die chemische Zusammensetzung der Eiterzellen. *Med.-Chem. Unters.* 441–460
2 Avery, O.T. *et al.* (1944) Studies of the chemical nature of the substance inducing transformation of pneumococcal types Induction of transformation by a deoxyribonucleic acid fraction isolated from Pneumococcus Type III. *J. Exp. Med.* 79, 137–158
3 Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46
4 Nordborg, M. and Weigel, D. (2008) Next-generation genetics in plants. *Nature* 456, 720–723
5 Bergelson, J. and Roux, F. (2010) Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. *Nat. Rev. Genet.* 11, 867–879
6 Myles, S. *et al.* (2009) Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21, 2194–2202
7 Chang, C. and Meyerowitz, E.M. (1986) Molecular cloning and DNA sequence of the *Arabidopsis thaliana* alcohol dehydrogenase gene. *Proc. Natl. Acad. Sci. U.S.A.* 83, 1408–1412
8 The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815
9 McCourt, P. and Benning, C. (2010) *Arabidopsis*: a rich harvest 10 years after completion of the genome sequence. *Plant J.* 61, 905–908
10 Jander, G. *et al.* (2002) Arabidopsis map-based cloning in the post-genome era. *Plant Physiol.* 129, 440–450
11 Peters, J.L. *et al.* (2001) A physical amplified fragment-length polymorphism map of *Arabidopsis*. *Plant Physiol.* 127, 1579–1589
12 Törjék, O. *et al.* (2003) Establishment of a high-efficiency SNP-based framework marker set for *Arabidopsis*. *Plant J.* 36, 122–140
13 Schmid, K.J. *et al.* (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res.* 13, 1250–1257
14 Nordborg, M. *et al.* (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* 3, e196

15 Clark, R.M. *et al.* (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* 317, 338–342
16 Kim, S. *et al.* (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* 39, 1151–1155
17 McNally, K.L. *et al.* (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc. Natl. Acad. Sci. U.S.A.* 106, 12273–12278
18 Ossowski, S. *et al.* (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* 18, 2024–2033
19 Zeller, G. *et al.* (2008) Detecting polymorphic regions in the *Arabidopsis thaliana* genome with resequencing microarrays. *Genome Res.* 18, 918–929
20 Santuari, L. *et al.* (2010) Substantial deletion overlap among divergent *Arabidopsis* genomes revealed by intersection of short reads and tiling arrays. *Genome Biol.* 11, R4
21 Schneeberger, K. *et al.* (2009) Simultaneous alignment of short reads against multiple genomes. *Genome Biol.* 10, R98
22 Lam, H.M. *et al.* (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* 42, 1053–1059
23 Gore, M.A. *et al.* (2009) A first-generation haplotype map of maize. *Science* 326, 1115–1117
24 Lai, J. *et al.* (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.* 42, 1027–1030
25 Springer, N.M. *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
26 Huang, X. *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–967
27 Ossowski, S. *et al.* (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327, 92–94
28 Lynch, M. (2008) Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Mol. Biol. Evol.* 25, 2409–2419
29 Roach, J.C. *et al.* (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328, 636–639
30 Ng, S.B. *et al.* (2010) Exome sequencing identifies the cause of a Mendelian disorder. *Nat. Genet.* 42, 30–35
31 Denver, D.R. *et al.* (2009) A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proc. Natl. Acad. Sci. U.S.A.* 106, 16310–16314
32 Keightley, P.D. *et al.* (2009) Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res.* 19, 1195–1201
33 Comai, L. and Henikoff, S. (2006) TILLING: practical single-nucleotide mutation discovery. *Plant J.* 45, 684–694
34 Sarin, S. *et al.* (2008) *Caenorhabditis elegans* mutant allele identification by whole-genome sequencing. *Nat. Methods* 5, 865–867
35 Michelmore, R.W. *et al.* (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
36 Schneeberger, K. *et al.* (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods* 6, 550–551
37 Huang, X. *et al.* (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res.* 19, 1068–1076
38 Blumenstiel, J.P. *et al.* (2009) Identification of EMS-induced mutations in *Drosophila melanogaster* by whole-genome sequencing. *Genetics* 182, 25–32
39 Doitsidou, M. *et al.* (2010) *C. elegans* mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy. *PLoS ONE* 5, e15435
40 Birkeland, S.R. *et al.* (2010) Discovery of mutations in *Saccharomyces cerevisiae* by pooled linkage analysis and whole genome sequencing. *Genetics* 186, 1127–1137
41 Laitinen, R.A. *et al.* (2010) Identification of a spontaneous frame shift mutation in a nonreference *Arabidopsis* accession using whole genome sequencing. *Plant Physiol.* 153, 652–654
42 Mirouze, M. *et al.* (2009) Selective epigenetic control of retrotransposition in *Arabidopsis*. *Nature* 461, 427–430
43 Gnerre, S. *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U.S.A.* 108, 1513–1518
44 Zuryn, S. *et al.* (2010) A strategy for direct mapping and identification of mutations by whole-genome sequencing. *Genetics* 186, 427–430
45 Albert, T.J. *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* 4, 903–905
46 Turner, E.H. *et al.* (2009) Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat. Methods* 6, 315–316
47 Gnirke, A. *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* 27, 182–189
48 Craig, D.W. *et al.* (2008) Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* 5, 887–893
49 Cronn, R. *et al.* (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.* 36, e122
50 Van Tassell, C.P. *et al.* (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* 5, 247–252
51 Baird, N.A. *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3, e3376
52 Kearsey, M.J. and Farquhar, A.G. (1998) QTL analysis in plants; where are we now? *Heredity* 80, 137–142
53 Xie, W. *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
54 Smith, A.M. *et al.* (2010) Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Res.* 38, e142
55 Wolyn, D.J. *et al.* (2004) Light-response quantitative trait loci identified with composite interval and eXtreme array mapping in *Arabidopsis thaliana*. *Genetics* 167, 907–917
56 Werner, J.D. *et al.* (2005) *FRIGIDA*-independent variation in flowering time of natural *Arabidopsis thaliana* accessions. *Genetics* 170, 1197–1207
57 Lai, C.Q. *et al.* (2007) Speed-mapping quantitative trait loci using microarrays. *Nat. Methods* 4, 839–841
58 Ehrenreich, I.M. *et al.* (2010) Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature* 464, 1039–1042
59 Buckler, E.S. *et al.* (2009) The genetic architecture of maize flowering time. *Science* 325, 714–718
60 Kump, K.L. *et al.* (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat. Genet.* 43, 163–168
61 Tian, F. *et al.* (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* 43, 159–162
62 Brachi, B. *et al.* (2010) Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. *PLoS Genet.* 6, e1000940
63 Graham, I.A. *et al.* (2010) The genetic map of *Artemisia annua* L. identifies loci affecting yield of the antimalarial drug artemisinin. *Science* 327, 328–331
64 Cai, D. *et al.* (1997) Positional cloning of a gene for nematode resistance in sugar beet. *Science* 275, 832–834
65 Song, W.Y. *et al.* (1995) A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*. *Science* 270, 1804–1806
66 Song, J. *et al.* (2003) Gene *RB* cloned from *Solanum bulbocastanum* confers broad spectrum resistance to potato late blight. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9128–9133
67 van der Linden, C.G. *et al.* (2004) Efficient targeting of plant disease resistance loci using NBS profiling. *Theor. Appl. Genet.* 109, 384–393
68 Robertson, G. *et al.* (2010) De novo assembly and analysis of RNA-seq data. *Nat. Methods* 7, 909–912
69 Mizrachi, E. *et al.* (2010) De novo assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. *BMC Genomics* 11, 681