Large SNP arrays for genotyping in crop plants

Martin W Ganal*, Andreas Polley, Eva-Maria Graner, Joerg Plieske, Ralf Wieseke, Hartmut Luerssen and Gregor Durstewitz

TraitGenetics GmbH, Am Schwabeplan 1b, D-06466 Gatersleben, Germany

**Corresponding author (Fax, +49-39482-799718; Email, ganal@traitgenetics.de)*

Genotyping with large numbers of molecular markers is now an indispensable tool within plant genetics and breeding. Especially through the identification of large numbers of single nucleotide polymorphism (SNP) markers using the novel high-throughput sequencing technologies, it is now possible to reliably identify many thousands of SNPs at many different loci in a given plant genome. For a number of important crop plants, SNP markers are now being used to design genotyping arrays containing thousands of markers spread over the entire genome and to analyse large numbers of samples. In this article, we discuss aspects that should be considered during the design of such large genotyping arrays and the analysis of individuals. The fact that crop plants are also often autopolyploid or allopolyploid is given due consideration. Furthermore, we outline some potential applications of large genotyping arrays including high-density genetic mapping, characterization (fingerprinting) of genetic material and breeding-related aspects such as association studies and genomic selection.

[Ganal MW, Polley A, Graner E-M, Plieske J, Wieseke R, Luerssen H and Durstewitz G 2012 Large SNP arrays for genotyping in crop plants. *J. Biosci.* **37** 821–828] **DOI** 10.1007/s12038-012-9225-3

1. Introduction

Large-scale genotyping with many molecular markers has been originally spearheaded for the detailed analysis of the human genome with respect to the identification of loci affecting quantitatively inherited traits. Due to the lack of specifically designed populations, in humans it is essentially the only possible approach towards the identification of genes and loci underlying such traits in which a phenotypically defined group of individuals (e.g. such that have a specific disease) is compared with a control population that lacks the respective trait. Specifically, in this process it is attempted to identify molecular markers that have a clear bias for specific alleles in only one of the two groups. It is clear that in this process very large numbers of genetic markers are needed (McCarthy *et al.* 2008). Initially such large numbers have been identified through the comparative sequencing of individual genes in a panel of individuals in large consortia (International HapMap Consortium 2007). With the development of more sophisticated sequencing technologies (also called next-generation sequencing, NGS), the generation of very large sets of sequence data and even the complete sequencing of entire genomes has become possible (Metzker 2010). The comparison of such DNA sequences between individuals or with a reference genome made it feasible to identify many millions of SNPs.

The identification of large numbers of molecular markers in human has been paralleled by the simultaneous development of high-throughput technologies that permit the genotyping of many thousands to millions of such markers using highly miniaturized genotype calling arrays (also called chips). Array-based genotyping methods are either based on the use of solid phase bound oligonucleotide probes diagnostic for the respective alleles and subsequent hybridization of genomic DNA onto such arrays or the use of single base primer extension technologies to determine the specific allelic state for a given SNP (McGall and Christians 2002; Gunderson *et al.* 2006; Steemers *et al.* 2006). Currently in human genetic analysis, genotyping of several millions of SNP markers using these arrays for individuals is routine.

Keywords. Molecular marker; plant breeding; plant genetics; single nucleotide polymorphism

Abbreviations used: GBS, genotyping by sequencing; LD, linkage disequilibrium; NGS, next-generation sequencing; PIC, Polymorphism Information Content; QTL, quantitatively inherited traits; SNP, single nucleotide polymorphism

Published online: 15 October 2012

In crop plants, the development of large genotyping arrays started much later than in humans due to a number of factors. One of them is probably the fact that specific segregating populations can be developed easily and in relatively large numbers. Another factor is that until the establishment of NGS technologies, SNP identification through DNA sequencing has been expensive and complex (Ganal *et al.* 2009). A further factor is that the genomes of many important crop plants have a large genome size, which is in some cases (e.g. barley, wheat and maize) at least as large as or significantly larger as the human genome. Finally, a considerable number of crop plant species are not diploid but polyploids or ancestral polyploids. This makes SNP identification and SNP calling much more difficult and complex than in a diploid organism such as human since SNPs between the different genomes have to be discriminated from SNPs between individuals (Durstewitz *et al.* 2010).

2. Large-scale SNP identification

NGS technologies have enabled the identification of large numbers of SNP markers in basically any crop plant via comparative sequencing of individuals (Varshney *et al.* 2009). Over the last years, this process started in crop plants with the comparative sequencing of the transcriptome of different individuals after reverse transcription of messenger RNA (Barbazuk *et al.* 2007; Novaes *et al.* 2008; Hasenmeyer *et al.* 2011; Hiremath *et al.* 2011). The advantage of this approach is that the identified SNPs are mostly located in genes and genes mostly occur in a single copy in the genome. SNPs present in single-copy sequences are a prerequisite for SNP marker analysis. Since the number of SNPs in genes is limited due to selection constraints in coding regions, this approach frequently results in only a few thousand useful markers, and thus alternative approaches have been developed. These approaches use NGS technologies in combination with complexity reduction technologies. These complexity reduction technologies have the advantage that, since they are DNA-based, they are not limited to mostly protein coding sequences. Thus, other single-copy sequences can be surveyed for SNPs as well. Complexity reduction technologies are based, for example, on the selective sequencing of a DNA fraction derived from the digestion with methylation-sensitive restriction enzymes (Deschamps *et al.* 2010; Gore *et al.* 2009a, b), the pre-amplification with specific AFLP (amplified fragment length polymorphisms) primer combinations (van Orsouw *et al.* 2007) or the use of the RAD (restriction-site associated DNA) technology (Davey *et al.* 2011). As for the transcriptome-based approach, these complexity reduction technologies have the advantage that they can be used more or less independently of the genome size. The most comprehensive approach towards the identification of SNP markers for genotyping is the comparison of fully sequenced genomes from individuals of a given species. Ideally, this approach requires a complete genome sequence as reference for SNP identification although other approaches without a full reference sequence have been described as well (You *et al.* 2011). In the more recent years, a considerable number of plant species (especially species that have a relatively small genome) have become fully sequenced so that suitable reference genomes are available (Feuillet *et al.* 2010). Comparative genome sequencing also termed genome re-sequencing has the advantage that essentially all SNPs in single-copy sequences between individuals can be identified. This has been demonstrated for crop plants such as maize, rice and soybean (Huang *et al.* 2009; Lai *et al.* 2010; Lam *et al.* 2010; Arai-Kichise *et al.* 2011).

3. SNP marker selection for genotyping arrays

The selection of SNP markers for a genotyping array requires a number of considerations that can be separated in the categories technological aspects, marker information content, distribution of the markers within the genome and total marker number. Technological aspects concerning the selection of SNP markers are, for example, the suitability of an identified SNP for a specific platform. In case of the Illumina Infinium technology, at least 50 bases upstream or downstream of the investigated SNP should be available and devoid of other SNPs since the marker could otherwise fail in specific germplasm subgroups that contain the opposite allele of such a flanking SNP than the one defined in the assay design. In highly diverse species such as maize, with one SNP every 44 base pairs in some germplasm (Gore *et al.* 2009a), this results in a significant number of markers that cannot be used for a genotype array, while in other species with a low level of polymorphism, such as tomato, this problem is hardly relevant. Furthermore, the sequence context (e.g. GC-content) that flanks an SNP could also determine the functionality of an SNP assay and influence the technical suitability of a given SNP which is reflected in the assay design score (for the Infinium technology the score should be >0.5), although this is usually only a minor constraint. Also, markers in highly repetitive sequences should be eliminated since they can impair the entire assay procedure (e.g. in the Golden Gate procedure). Another constraint in the marker selection is the marker information content reflected by the PIC (Polymorphism Information Content) value (Anderson *et al.* 1993). SNPs in close proximity occur usually in haplotypes (i.e. essentially equivalent to alleles). SNPs specific for a given haplotype are in full linkage disequilibrium (LD), and thus the analysis of more than one of these SNPs does not provide additional information. Due to this, in humans, on SNP genotyping arrays there are nowadays predominantly haplotype-specific SNPs, which are also called tagSNPs (Pfeiffer and Gunderson 2009). In

plants, the concept of using haplotype-specific SNPs has not yet been fully appreciated in current genotype array development. Especially in organisms with a low level of polymorphism and if only a few lines have been sequenced for SNP identification, there is a high probability that a number of SNPs selected from a gene are in full LD and thus provide no additional information. In some cases, this can result in up to 50% of the markers providing redundant information in large sets of analysed germplasm. A third constraint is the distribution of the SNP markers in the respective genome. If transcriptome sequencing is performed for marker identification, it is clear that markers present only in genes can be analysed. If markers are available from complexity reduction approaches or whole genome sequencing, it has to be decided which markers should be included in the array design. Polymorphisms in genes are often the basis of genetic variation leading to different phenotypes, so that SNPs in genes might have a higher priority. On the other hand, this disregards flanking sequences with regulatory functions or regions where no genes are present (even if they appear only devoid of genes due to problems with annotation). If a sufficient number SNPs is available for an array, a compromise could be to put SNPs in as many different genes as possible onto the array and add additional SNPs based on their chromosomal position in a reference sequence. This was the approach that has been used for the development of a large maize genotyping array (Ganal *et al.* 2011). A final constraint is the marker number that can be put on a genotyping array. This limitation is simply due to the fact that the larger a genotyping array gets, the higher are the costs per sample so that the number of markers on an array can be an economic limitation. The final constraint in the marker number on a genotyping array is the combination of all previously mentioned aspects in the way that there is frequently only a limited set of markers left over after the previously mentioned factors have been considered. Based on our array development experience in a number of different crop plants, a good rule is to include initially around 5 to 10 times the number of high quality SNPs into the selection procedure, as will ultimately be placed on the final genotyping array.

4. Genotyping with a large array

Once a large genotyping array has been developed for a species, the quality of the array has to be tested by genotyping a first set of individuals or lines. This set has to be carefully chosen to represent samples from the entire genetic range of the species or germplasm that shall be analysed with the respective array. Furthermore, all possible allelic configurations should be represented in the samples. This means in crop plants, where material is often inbred, that heterozygous samples (e.g. F1 hybrids) also need to be analysed (Ganal *et al.* 2011). Only in this way it is possible to define the individual cluster limits for routine genotyping of large sample numbers since this requires in a first step the careful inspection of the intensity clusters generated with each marker. If this is not done carefully, it might be necessary to go through the allele calling again for each batch of analysed samples marker by marker, especially if a different type of material or different population types (e.g. doubled haploids or F2 populations) are being analysed. Figure 1A shows an example for the cluster pattern that is generated with a high-quality marker in a diploid species.

While the cluster definition is relatively simple in diploid species, it is more difficult in crop plants that are polyploid or have gone relatively recently through a polyploidization step. In autotetraploid species such as potato, instead of the typical three clusters (AA, AB and BB), five clusters can be observed (AAAA, AAAB, AABB, ABBB and BBBB). This

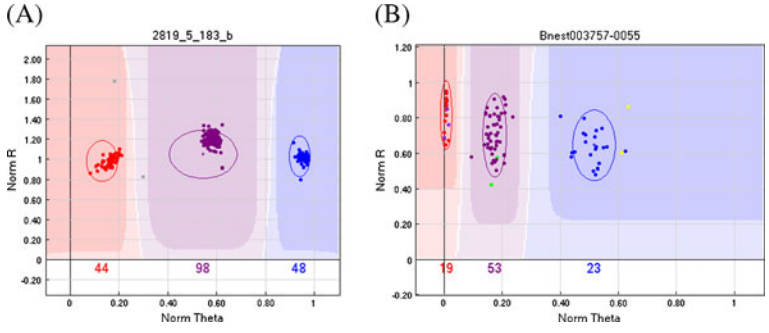

Figure 1. SNP analysis in species with different ploidy levels (Illumina Infinium platform). (**A**) Pattern for a high-quality SNP marker in a diploid organism such as tomato. In the red area are signals that are from individuals which are homozygous for allele 1. In the purple area are signals from heterozygous plants and in the blue area are signals from individuals that are homozygous for allele 2. (**B**) Pattern of a high-quality SNP marker in an allotetraploid species (e.g. oilseed rape) where one of the two genomes is polymorphic and the second genome is monomorphic in the background. Three clearly defined clusters shifted to one side are detected and the three clusters represent for the G/A polymorphism the situation GGGG, GGGA and GGAA.

makes the definition of the allelic state more difficult since the clusters need to be defined more carefully and mostly in a manual fashion. In allopolyploid species such as the allotetraploid *Brassica napus* (oilseed rape), with a typical marker, there are only three clusters, but a second monomorphic genome is usually in the background so that the actual clusters are defined as AAAA, AAAB and AABB and are shifted significantly closer together (figure 1B). In allopolyploid species, the scoring and analysis of markers that are polymorphic in more than one genome is not advisable (Durstewitz *et al.* 2010). In maize as a more recently polyploidized species, a considerable number of markers behave also as in an allotetraploid species especially if the respective marker is located in a conserved duplicated area of the genome. The described shift of clusters is even more pronounced in the allohexaploid wheat, where two genomes are in the background so that actually the situations AAAAAA, AAAAAB and AAAABB are assayed. In such a species, it is frequently difficult to clearly separate the three clusters and a considerable number of markers have to be eliminated from the analysis due to that problem (Akhunov *et al.* 2009).

5. Use of large genotyping arrays in genetic research and breeding of crop plants

Large genotyping arrays can be used in genetic research and plant breeding for a variety of aspects. Generating high-density genetic maps containing thousands of markers permits the localization of single gene traits to a very precise point in the genome. For breeding, this enables the development of very tightly linked markers for marker-assisted breeding that only rarely show recombination with the respective trait. Very tightly linked markers being less than one cM (centi-Morgan) distant from a trait are also an indispensable tool for map-based isolation of the underlying gene. Maps with many markers permit also the localization of quantitatively inherited traits (QTLs) to a more precise position. Furthermore, with high-density maps it is easier to compare genetic maps of different populations and to identify genomic differences or chromosomal rearrangements with high precision and resolution. Genetic maps with many thousands of markers can also be used for the independent validation of sequence assemblies generated during genome sequencing of a given organism. In maize, two high-density maps have been used to compare the genetic marker order with the physical marker order on the genome sequence and a number of potential inconsistencies in the genome sequence could be identified (Ganal *et al.* 2011).

The analysis of genetic material with large genotyping arrays is another application. Large genotyping arrays permit a detailed analysis of lines or individuals (Yan *et al.* 2009). This enables the description of genetic relationships or allelic constitutions at a genome-wide level or in specific genomic regions at a much higher resolution than previously possible. If the marker density on genotyping arrays is much higher than the average extent of haplotypes, it permits a comprehensive description of genetic material. Especially in inbreeding species that have passed through a recent genetic bottleneck during breeding, relatively low marker densities (a few thousand markers) can be sufficient for a precise description of the genome. Figure 2 shows an example for the comparative analysis of maize lines from different genetic origins with respect to their diversity in a specific genomic region. In the area of plant breeding, large genotyping arrays also provide a tool for the very precise comparison und description of varieties within the breeding process (e.g. through the localization of critical crossover events or allelic status in specific genomic regions). The prediction can be made that with large genotyping arrays, variety identification will also be put to a level where much more precise and unanimous statements on identity or relationship can be made than with the previous marker technologies.

As outlined in the introduction, large genotyping arrays are used in human genome analysis predominantly for the identification of genomic regions that are associated with specific quantitatively inherited traits. There it has resulted in the identification of many genes that have an effect on specific complex diseases. In crop plants, this association genetics approach is just at its beginning with very little data being published using genome-wide marker sets (Rafalski 2010; Zhao *et al.* 2011). However, the prediction can be made that in the near future, we will see many more data on such experiments, especially since the most interesting traits that are relevant for plant breeding, including the pivotal yield trait, are quantitatively inherited.

Since many traits that are important in plant breeding are controlled by numerous QTLs with small effects, recently a breeding method developed in animal breeding has gained considerable interest in crop plant breeding. This approach is called genomic selection (Meuwissen *et al.* 2001), and for this breeding method large genotyping arrays with many markers are pivotal. Carefully phenotyped populations (training populations) of mostly unrelated individuals from a breeding pool are genotyped at very high marker density and for each marker its individual effect on the respective phenotype is determined. In a next step, the cumulative effects of all or selected subsets of markers are used as predictors for the actual phenotype in related or derived material without directly phenotyping these individuals. Expectations are high that this will result in faster breeding cycles and an increased genetic gain per generation or year of breeding (Hamblin *et al.* 2011).

6. Summary and future trends in the development and use of large genotyping arrays

Genotyping arrays are currently being developed for a large number of important crop plants in order to get more precise

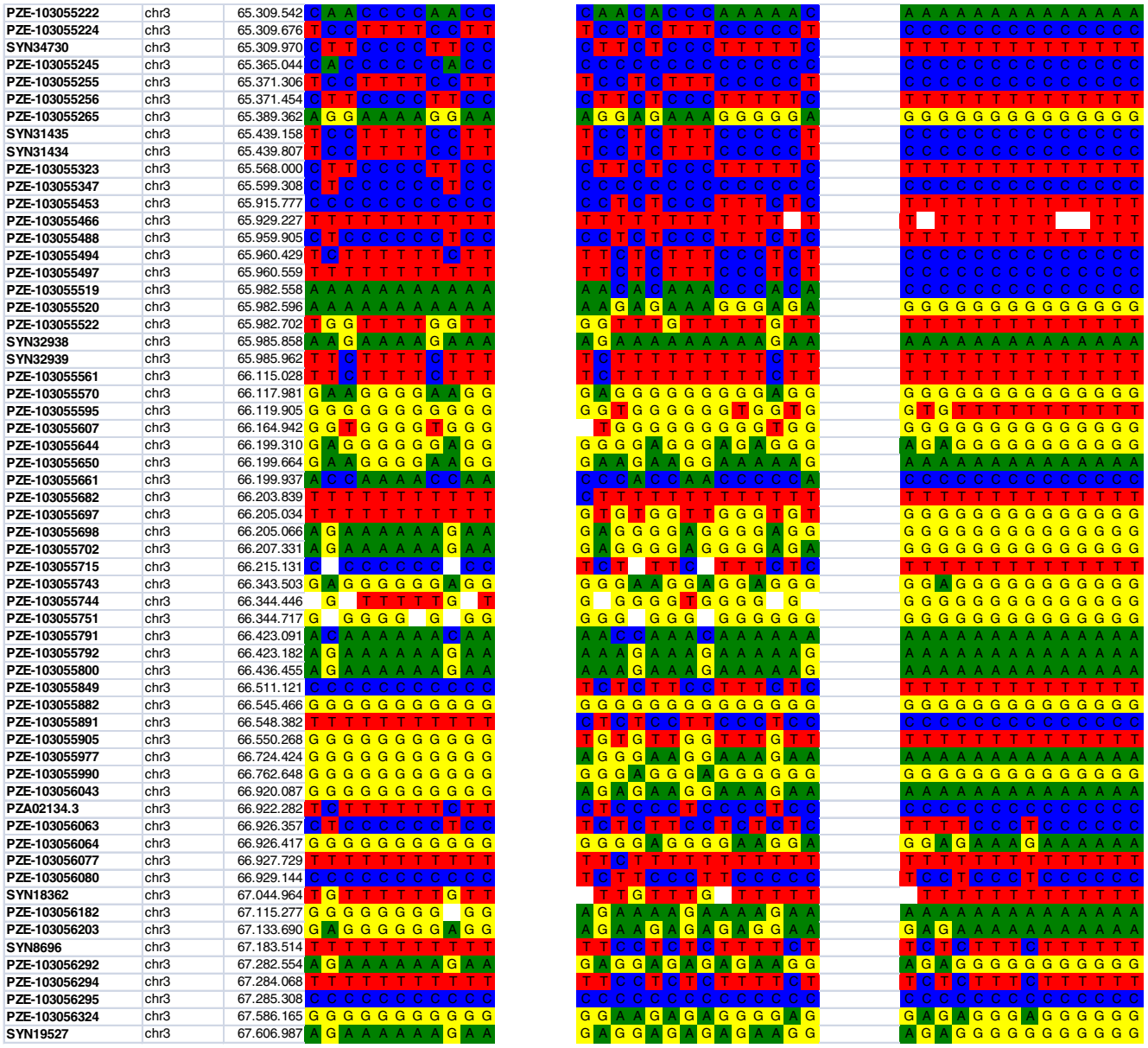

Figure 2. Linkage disequilibrium in different types of maize germplasm. Markers, chromosomal assignment and physical position (in kb) of each marker are presented in the first three columns. In the following the genotype data (in the nucleotide code, blank = no allele called) from the lines of three different germplasm pools are shown. Note the different numbers of haplotypes and different levels of LD.

insights into their genetic constitution and for the improvement of plant breeding. For many crop plants (table 1) currently first-generation arrays with marker numbers between 5000 and 100000 have been or are being developed. Through the use of high-throughput sequencing technologies and the increased availability of high-quality genomic reference sequences, it can be expected that genotyping arrays with large numbers of SNPs will be available within the next 2 years for essentially all major crop plants. In the longer term, it is very likely that these first-generation arrays will be replaced by improved arrays with more markers and specifically more haplotype-defining markers. Especially in the area of plant breeding for association genetics approaches and genomic selection (Hamblin *et al.* 2011), such arrays will be extensively used.

On the other side in plant breeding as in animal breeding the large arrays will be too expensive to be used in routine applications. Because of that another set of customized arrays will probably be also developed for many crop plants that contain smaller numbers (a few thousand) of markers

Table 1. Overview on large genotyping array development in major crop plants

Crop species	Array size*	Reference
Apple (*Malus domestica*)	8 K	Chagné *et al.* (2012)
Grape (*Vitis vinifera*)	9 K	Myles *et al.* (2011)
Maize (*Zea mays*)	60 K	Ganal *et al.* (2011)
Rice (*Oryza sativa*)	44 K	Zhao *et al.* (2011)
Rye (*Secale cereale*)	5 K	Hasenmeyer *et al.* (2011)
Sunflower (*Helianthus annuus*)	10 K	Bachlava *et al.* (2012)
Barley (*Hordeum vulgare*)	9 K	unpublished
Cherry (*Prunus spec.*)	6 K	unpublished *http://www.illumina.com/applications/agriculture.ilmn#ag_consortia*
Grape (*Vitis vinifera*)	20 K	in preparation *http://www.illumina.com/applications/agriculture.ilmn#ag_consortia*
Peach (*Prunus persica*)	8 K	unpublished *http://www.illumina.com/applications/agriculture.ilmn#ag_consortia*
Potato (*Solanum tuberosum*)	10 K	unpublished *http://solcap.msu.edu/*
Tomato (*Solanum lycopersicum*)	10 K	unpublished *http://solcap.msu.edu/*
Wheat (*Triticum aestivum*)	9 K	unpublished *http://www.triticeaecap.org/*
Oilseed rape (*Brassica napus*)	60 K	in preparation *http://www.illumina.com/applications/agriculture.ilmn#ag_consortia*
Wheat (*Triticum aestivum*)	90 K	in preparation *http://www.illumina.com/applications/agriculture.ilmn#ag_consortia*

* Number of features on the array. Note that this does not necessarily correspond to the number of SNPs that can be analysed (e.g. with the 60 K maize array, approximately 50000 functional SNPs can be analysed).

that are linked to interesting genes and QTLs and/or have a high PIC value so that they are optimized for marker-assisted breeding and backcrossing.

With the advent of fully sequenced genomes, other genotyping methods not based on arrays appear to be promising as well. Specifically, genotyping by sequencing (GBS) methods (Elshire *et al.* 2011) have the potential to either being used directly for genotyping many thousands of loci simultaneously in many lines or at least for the identification of many additional SNP markers for genotyping arrays (Davey *et al.* 2011). Currently, it is not clear which of the technologies (array-based genotyping or GBS) will prevail in the next couple of years since both technologies have their advantages and disadvantages. In routine analyses, array-based genotyping has the advantage that the data are highly reproducible within and between laboratories and they can be easily stored in and retrieved from databases since the same markers are always used. On the other side, array-based genotyping is relatively inflexible since the SNPs on an array cannot easily be replaced or added, and thus arrays-based genotyping is prone to ascertainment bias when samples are being analysed that are from different gene pools (e.g. wild and domesticated material) and this could result in artificially low number of polymorphic markers. GBS methods have the potential of detecting many more and unbiased loci in a genome than genotyping arrays, especially with the continuously increasing sequence output from the novel sequencing technologies, and the fact that the technology does not require the initial efforts necessary for the development of large genotyping arrays. Thus, it can be expected that the GBS technologies will be intensively used for high-density genetic mapping in the near future. However, currently it is not clear how effectively GBS can be used in the characterization of unrelated genetic material, since not in all lines are the same loci being investigated and many data points have to be imputed which becomes difficult when the haplotype situation is unclear. Furthermore, GBS technologies still provide tremendous bioinformatical challenges regarding the standardization of data from different sources, platforms and technologies, and so it will be difficult to use them effectively and in a cost-efficient manner in routine plant-breeding processes that require large standardized datasets in a short timeframe.

In the long term and with continuously decreasing cost in full genome sequencing, it will be likely that in many crop plants (especially those with a relatively small genome) within approximately 10 years, whole genome sequencing of many lines will be the ultimate method of choice for a comprehensive genotyping effort.

Acknowledgements

The authors acknowledge the assistance of the technical staff at TraitGenetics during SNP marker development and the analysis of many samples using genotyping arrays. Research in the area of large-scale genotyping at TraitGenetics has in part been funded by grants from the German Federal Ministry of Education and Research (BMBF).

References

Akhunov E, Nicolet C and Dvorak J 2009 Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina Golden Gate assay. *Theor. Appl. Genet.* **119** 507–517

Anderson JA, Churchill GA, Autrique JE, Tanksley SD and Sorrells ME 1993 Optimizing parental selection for genetic linkage maps. *Genome* **36** 181–186

Arai-Kichise Y, Shiwa Y, Nagasaki H, Ebana K, Yashikawa H, Yano M and Wakasa K 2011 Discovery of genome-wide DNA polymorphisms in a land race cultivar of Japonica rice by whole-genome sequencing. *Plant Cell Physiol.* **52** 274–282

Bachlava E, Taylor CA, Tang S, Bowers JE, Mandel JR, Burke JM and Knapp SJ 2012 SNP discovery and development of a high-density genotyping array for sunflower. *PLoS ONE* **7** e29814

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

Chagné D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C, Vanderzande S, Hellens RP, *et al.* 2012 Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. *PLoS ONE* **7** e31745

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

Deschamps S, la Rota M, Ratashak JP, Biddle P, Thureen D, Farmer A, Luck S, Beatty M, *et al.* 2010 Rapid genome-wide single nucleotide polymorphism discovery in soybean and rice via deep resequencing of reduced representation libraries with the Illumina genome analyzer. *Plant Genome* **3** 53–68

Durstewitz G, Polley A, Plieske J, Luerssen H, Graner EM, Wieseke R and Ganal MW 2010 SNP discovery by amplicon sequencing and multiplex SNP genotyping in the allopolyploid species *Brassica napus*. *Genome* **53** 948–956

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

Feuillet C, Leach JE, Rogers J, Schnable PS and Eversole K 2010 Crop genome sequencing: lessons and rationales. *Trends Plant Sci.* **16** 77–88

Ganal MW, Altmann T and Röder MS 2009 SNP identification in crop plants. *Curr. Opin. Plant Biol.* **12** 211–217

Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A, Clarke JD, Graner E-M, *et al.* 2011 A large maize (*Zea mays* L.) SNP genotyping array: Development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS ONE* **6** e28334

Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Pfeiffer JA, McMullen MD, *et al.* 2009a A first-generation haplotype map of maize. *Science* **326** 1115–1117

Gore MA, Wright MH, Ersoz ES, Bouffard P, Szekeres ES, Jarvie TP, Hurwitz BL, Narechania A, *et al.* 2009b Large-scale discovery of gene-enriched SNPs. *Plant Genome* **2** 121–133

Gunderson KL, Steemers FJ, Ren H, Ng P, Zhou L, Tsan C, Chang W, Bullis D, *et al.* 2006 Whole-genome genotyping. *Methods Enzymol.* **410** 359–376

Hamblin MT, Buckler ES and Jannink JL 2011 Population genetics of genomics-based crop improvement methods. *Trends Genet.* **27** 98–106

Hasenmeyer G, Schmutzer T, Seidel M, Zhou R, Mascher M, Schön C-C, Taudien S, Scholz U, Mayer KF and Bauer E 2011 From RNA-seq to large-scale genotyping: genomics resources for rye (*Secale cereale* L.). *BMC Plant Biol.* **11** 131

Hiremath PJ, Farmer A, Cannon SB, Woodward J, Kudapa H, Tuteja R, Kumar A, Bhanuprakash A, *et al.* 2011 Large-scale transcriptome analysis in chickpea (*Cicer arietinum* L.), an orphan legume crop of the semi-arid tropics of Asia and Africa. *Plant Biotechnol. J.* **9** 922–931

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

International HapMap Consortium 2007 A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449** 851–61

Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, *et al.* 2010 Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.* **42** 1027–1030

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, *et al.* 2010 Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* **42** 1053–1059

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP and Hirschhorn JN 2008 Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9** 356–369

McGall GH and Christians FC 2002 High-density genechip oligonucleotide probe arrays. *Adv. Biochem. Eng. Biotechnol.* **77** 21–42

Metzker ML 2010 Sequencing technologies – the next generation. *Nat. Rev. Genet.* **11** 31–46

Meuwissen THE, Hayes BJ and Goddard ME 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157** 1819–1829

Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, *et al.* 2011 Genetic structure and domestication history of the grape. *Proc. Natl. Acad. Sci. USA* **108** 3530–3535

Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D and Sedoroff RR 2008 High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9** 312

Peiffer DA and Gunderson KL 2009 Design of tag SNP whole genome genotyping arrays *Methods Mol. Biol.* **529** 51–61

Rafalski JA 2010 Association genetics in crop improvement. *Curr. Opin. Plant Biol.* **13** 174–180

Steemers FJ, Chang W, Lee G, Barker DL, Shen R and Gunderson KL 2006 Whole-genome genotyping with the single-base extension assay. *Nat. Methods* **3** 31–3

Van Orsouw NJ, Hogers RCJ, Janssen A, Yalcin F, Snoeeijers S, Verstege E, Schneiders H, van der Poel H, *et al.* 2007 Complexity reduction of polymorphic sequences (CRoPS): A novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE* **2** e1172

Varshney RK, Nayak SN, May GD and Jackson SA 2009 Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* **27** 522–530

Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD and Crouch J 2009 Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS ONE* **24** e8451

You FM, Huo N, Deal KR, Gu YQ, Luo M-C, McGuire PE, Dvorak J and Anderson OD 2011 Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. *BMC Genomics* **12** 59

Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, *et al.* 2011 Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* **2** 467