The Plant Journal (2007) **51**, 727–737 doi: 10.1111/j.1365-313X.2007.03194.x

TECHNICAL ADVANCE

Indel arrays: an affordable alternative for genotyping

Neeraj Salathia[1], Hana N. Lee[1], Todd A. Sangster[2,3], Keith Morneau[1], Christian R. Landry[4,†], Kurt Schellenberg[1], Aditi S. Behere[5,‡], Kevin L. Gunderson[6], Duccio Cavalieri[7], Georg Jander[5] and Christine Queitsch[1,*]

[1]*FAS Center for Systems Biology, Harvard University, Cambridge, MA 02138, USA,*
[2]*Committee on Genetics, University of Chicago, Chicago, IL 60637, USA,*
[3]*Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA,*
[4]*Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA,*
[5]*Boyce Thompson Institute for Plant Research, Ithaca, 14853, NY, USA,*
[6]*Illumina Inc. San Diego, CA 92121, USA, and*
[7]*Department of Pharmacology, University of Florence, Viale Pieraccini 6, 50139 Florence, Italy*

Received 26 December 2006; revised 24 April 2007; accepted 14 May 2007.
[*]For correspondence (fax +1 617 495 2196; e-mail cqueitsch@cgr.harvard.edu).
[†]Present address: Département de Biochimie, Université de Montréal, C.P. 6128, Succ. Centre-Ville, H3C 3 J7, Montréal, QC, Canada.
[‡]Present address: Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA.

Summary

Natural variation and induced mutations are important resources for gene discovery and the elucidation of genetic circuits. Mapping such polymorphisms requires rapid and cost-efficient methods for genome-wide genotyping. Here we report the development of a microarray-based method that assesses 240 unique markers in a single hybridization experiment at a cost of less than US$50 in materials per line. Our genotyping array is built with 70-mer oligonucleotide elements representing insertion/deletion (indel) polymorphisms between the *Arabidopsis thaliana* accessions Columbia-0 (Col) and Landsberg *erecta* (L*er*). These indel polymorphisms are recognized with great precision by comparative genomic hybridization, eliminating the need for array replicates and complex statistical analysis. Markers are present genome-wide, with an average spacing of approximately 500 kb. PCR primer information is provided for all array indels, allowing rapid single-locus inquiries. Multi-well chips allow groups of 16 lines to be genotyped in a single experiment. We demonstrate the utility of the array for accurately mapping recessive mutations, RIL populations and mixed genetic backgrounds from accessions other than Col and L*er*. Given the ease of use of shotgun sequencing to generate partial genomic sequences of unsequenced species, this approach is readily transferable to non-model organisms.

Keywords: microarray, indel, genotyping, mutant mapping.

Introduction

Despite great advances in high-throughput technologies and computational annotations, the functions of many of the approximately 30 000 *Arabidopsis thaliana* genes, especially the effect of genetic interactions on phenotype and responses to environmental stimuli, remain unknown. Currently, large-scale mutagenesis screens and analysis of naturally occurring genetic variation are most commonly used to further dissect complex molecular traits and discover causative genes and epistatic interactions (Koornneef *et al.*, 2004). In particular, quantitative trait locus (QTL) mapping has the potential to identify allelic variants fixed in disparate natural populations by differential selective pressures, thereby placing biological pathways in an ecological and evolutionary framework (Erickson *et al.*, 2004).

Both QTL mapping and forward genetics approaches with F_2 populations rely on determination of linkage to a genome-wide panel of genetic markers. In particular, a daunting challenge is the genotyping of recombinant inbred line (RIL)

and F_2 populations, typically containing several hundred lines or individuals, to a resolution limited primarily by the occurrence of recombination events. Traditional PCR and multiplex PCR mapping approaches are routinely used for genotyping (Bell and Ecker, 1994; Jander, 2006; Ponce *et al.*, 2006). These methods are time- and labor-intensive, and typically assess relatively few markers. QTL studies, in particular of non-model species, are also limited by the non-trivial correlation of maps based on anonymous markers, such as AFLPs, to physical maps for subsequent fine-mapping and comparison of QTL across divergent RIL sets.

Thus, affordable and easily accessible genotyping techniques will significantly aid future identification of gene function by the plant research community. Several companies have developed methods for genome-wide SFP (single feature polymorphism) discovery and analysis. Hybridization of genomic DNA to short oligonucleotides in either single-channel (Affymetrix) or comparative hybridization (NimbleGen) platforms allows detection of indels and some point mutations (Borevitz, 2006; Borevitz *et al.*, 2003; Selzer *et al.*, 2005; West *et al.*, 2006). In addition, high-density tiled DNA microarrays (NimbleGen, Perlegen) may be used for resequencing targeted regions and detecting SFPs (Albert *et al.*, 2005). These array-based techniques excel in genome-wide SFP discovery and assessment but are prohibitively expensive for routine mapping requirements (Table S1).

Illumina's bead array technology allows parallel analysis of many SNP polymorphisms using allele-specific primer extension and signal amplification (Gunderson *et al.*, 2005; Gunderson *et al.*, 2006). Although Illumina SNP genotyping services are cost-competitive with our approach for large populations, they are not cost-efficient when genotyping small sample sizes. Another company, Sequenom, offers high-throughput SNP genotyping based on allelic mass differences detected by MALDI–TOF MS (matrix-assisted laser desorption/ionization–time of flight mass spectrometry) analysis (Storm *et al.*, 2003). However, Sequenom genotyping services are currently not available for sample sizes below 24 lines, and genotyping for larger sample sizes is more expensive compared with our approach (Table S1).

Here, we present a novel genotyping method that combines the advantages of high-resolution microarray-based genotyping with cost-efficiency for both large and small sample sizes while avoiding extensive statistical analysis. Our method is based on comparative genomic hybridization (CGH) of DNA to spotted microarrays of 70-mer oligonucleotide markers, representing insertion/deletion (indel) polymorphisms between Landsberg *erecta* (L*er*) and the reference accession Columbia-0 (Col; described by Jander *et al.*, 2002). Although far fewer indel markers than SNPs exist, we demonstrate that sufficient indels are available to limit the mapping resolution of a typical Col/L*er* segregating plant population by recombination events rather than the number of genotyped markers.

Results

Array design

We reasoned that indel polymorphisms, particularly longer indels, might result in more faithful differentiation in comparative hybridizations than individual SNP, allowing fewer array replicates. Indel polymorphisms between Col and L*er* were identified by BLAST sequence comparison of both genomes, available from the TAIR and Monsanto databases respectively (see Experimental procedures). Based on previous studies, several parameters were considered in designing the array elements: oligonucleotide length, cross-hybridization due to sequence similarity, deletion length, sequence composition and melting temperature. For the significantly smaller genomes of *Bacillus subtilis* and *Saccharomyces cerevisiae*, spotted arrays with 50-mer oligonucleotides have a tendency to non-specific cross-hybridization with other regions of the genome that show sequence identity >75% (Kane *et al.*, 2000; D. Cavalieri, unpublished data). Our preliminary tests indicated that the performance of 70-mers was superior to that of 50-mers (data not shown). Therefore, we designed 70-mer oligonucleotides with less than 70% overall sequence identity to other regions of the sequenced Col genome. Whenever possible, deletions were centered within the 70-mer. We avoided indels with stretches of more than 12 consecutive adenine or thymine nucleotides, which may bind non-specifically to common repeat sequences. To allow uniform hybridization conditions, we chose oligonucleotides with a melting temperature between 60 and 76°C. In total, 374 unique indels, representing 342 and 32 insertion alleles present in the Col and L*er* genomes, respectively, matched our computational criteria and were used to build a pilot array for marker verification. The greater abundance of L*er* deletions is probably an artifact of the short reads produced by L*er* shotgun DNA sequencing, and does not represent a fundamental difference between the two genomes. CGH of differentially labeled Col and L*er* genomic DNA should result in a relatively high Col-specific fluorophore signal to indel elements present only in the Col genome, and vice versa for elements present only in the L*er* genome. L*er*- and Col-specific signals should be equally present for non-polymorphic control markers (Figure 1, see Experimental procedures).

Array and marker verification

We used a multi-pronged approach to assess which of the 374 unique markers robustly differentiated between Col and L*er* genomic DNA. First, we performed eight independent CGHs of Col and L*er* genomic DNA to the array. Out of the 374 markers, 298 showed an average L*er*/Col signal intensity ratio of less than 0.5 and were therefore potentially

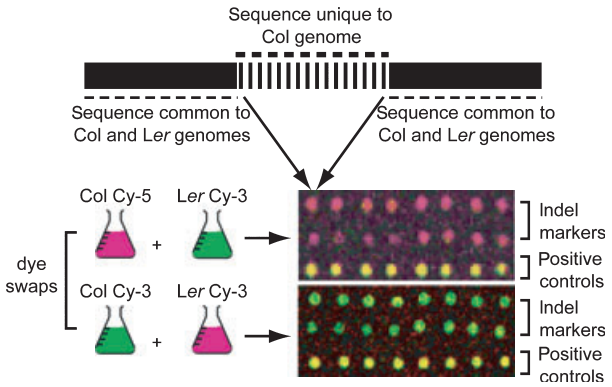

Figure 1. Basic principle of the indel array.
70-mer indel oligo markers unique to the Col genome are spotted on the array surface. Col and L*er* genomic DNA are labeled with a fluorophore and competitively hybridized to the array. Fluorescence signal from the hybridized oligo should correspond to the fluorophore with which Col genomic DNA is labeled, as illustrated by the dye-swap experiment. Positive control elements present in both the L*er* and Col genomes fluoresce equally in the Cy-5 and Cy-3 fluorescence channels.

polymorphic (data not shown). Markers that scored consistently as non-polymorphic or were highly inconsistent between replicate arrays (standard deviation of signal ratio >0.6) were excluded from further consideration.

In parallel, we used PCR to confirm which indel markers represented genuine deletions. Primers were designed for 325 unique markers whose indel polymorphism size allowed easy PCR amplification. Of these, 274 unique markers were PCR-verified as polymorphic between Col and L*er*. In addition to aiding the design of the indel array, these markers represent a valuable resource for traditional mapping, quick confirmation of recombination breakpoints, and subsequent fine mapping with near-isogenic lines (NILs; suggested primer sequences are given in Table S2). Combining PCR and array data, we retained markers that were polymorphic by PCR and whose ratio average across the array experiments indicated their polymorphic status. We also retained markers that we were unable to amplify by PCR, but which were consistently identified as polymorphic on the array. We excluded markers that were polymorphic by PCR but did not show differential binding in microarray experiments (suggested primer sequences are given in Table S3). After filtering by these criteria, we retained 318 unique markers.

Lastly, we compared indel array-generated maps with PCR-generated maps to empirically determine the robustness and accuracy for each marker. We performed both PCR and array-based genotyping of 44 novel Col × L*er* RILs (T. Sangster *et al.*, unpublished) and compared marker status for individual markers. Markers were retained only if the marker could both be informatively scored in over 25% of array experiments and correctly assess Col and L*er* genotypes in more than 80% of the experiments (Table S4). As before, ratio signals <0.5 indicated L*er* marker status and those greater than 0.5 indicated Col.

Taken together, we defined 277 robustly performing markers: 240 unique markers, 30 indel replicates (different array elements representing the same indels), three technical replicates, two markers that are both technical and indel replicates, and 16 positive controls (Table S5). All but one of the markers represent deletions in the L*er* genome relative to Col. The indel polymorphisms range in size from 25 to 7260 bp (Figure 2a), and the mean and median indel sizes are 372 and 55 bp, respectively. The majority (53%) are located in intergenic regions (Figure 2b). The average spacing of markers is approximately 500 kb, with little coverage in centromeric regions (Figure 2c).

Technical and biological replicates were used to further judge the performance of our final marker set. Here, technical replicates assess the performance of elements printed in replicate on the same chip in a single hybridization experiment, whereas biological replicates assess the reproducibility of marker performance from independent hybridization experiments using batches of independently prepared DNA. Both technical and biological replicates were highly correlated (R^2 = 0.98 and 0.86, respectively, Figure 3). Due to this robust repeatability, single hybridizations were deemed sufficient for generating reliable maps.

Analysis of marker status

We aimed to create simple analysis tools that would be usable without extensive statistical and computational knowledge. We developed a Microsoft Excel Visual Basic macro to automate conversion of raw signal intensity to fully normalized ratios (see Experimental procedures). A graphical representation of data output for the published Col × L*er* RIL CS1982 (Lister and Dean, 1993) is shown in Figure 4. Marker ratios almost always fall into distinct bins of either Col or L*er*, highlighting the ease and power of our genotyping approach.

To assess the likelihood of a particular signal ratio corresponding to a particular marker genotype, we compared indel array ratios from four RI lines (CS1982, CS1990, CS1992 and CS1994; Lister and Dean, 1993) with high-resolution SNP hybridization maps (see Experimental procedures). From this comparison, we calculated the likelihood that a marker of known genotype would fall into a discrete array ratio bin, which ranged from 0 to 1.3 in 0.1 increments. Such likelihoods allow estimation of the false call rates of various signal ratios. For example, if the measured ratio of a marker of previously unknown genotype falls between 0.1 and 0.2, its likelihood of being falsely called L*er* while truly being Col is 0.9% (Figure 5).

The genotype of most markers can be determined with a high statistical accuracy using such false call probabil-

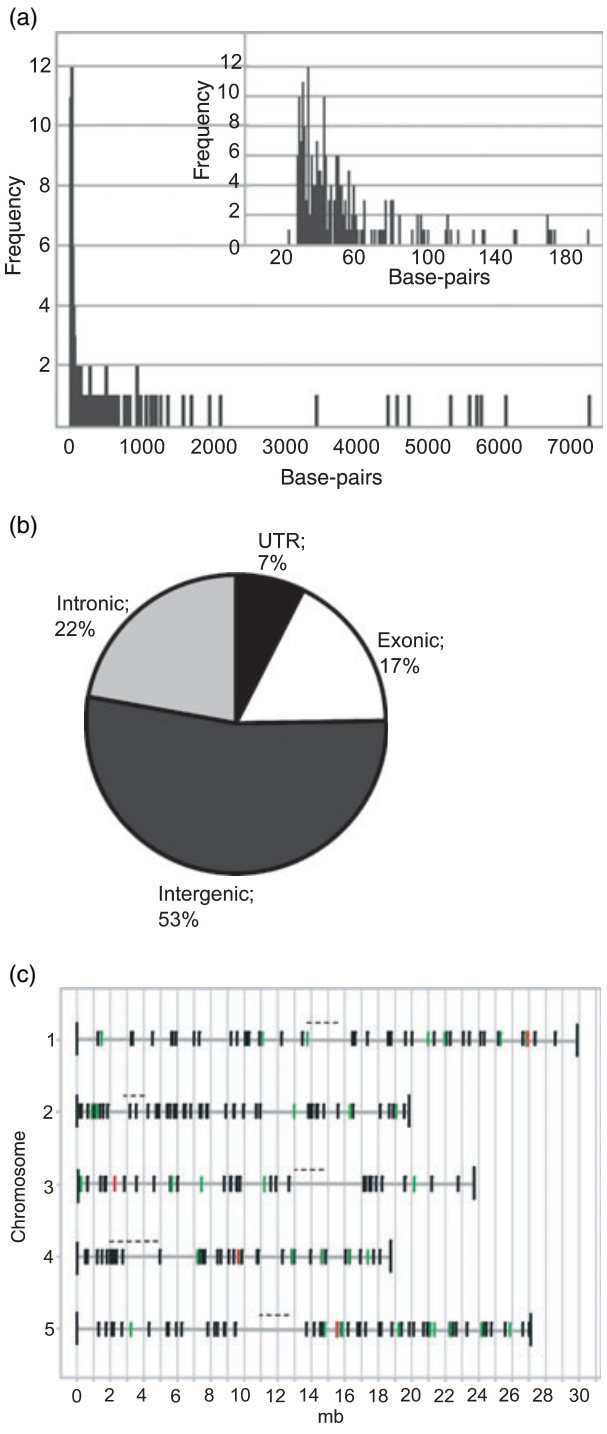

Figure 2. Indel marker properties.
(a) Size distribution of indel markers. Inset, size distribution of indels < 200 bp.
(b) Distribution of marker characteristics.
(c) Distribution of marker locations in the Arabidopsis genome. Locations of indel markers are indicated by marks along chromosomes represented by horizontal grey lines. Black marks indicate locations of unique markers, green marks represent locations where two array elements recognize the same indel, and red marks represent technical repeat markers. Horizontal dashed lines indicate approximate locations of the centromeres (Copenhaver *et al.*, 1999).

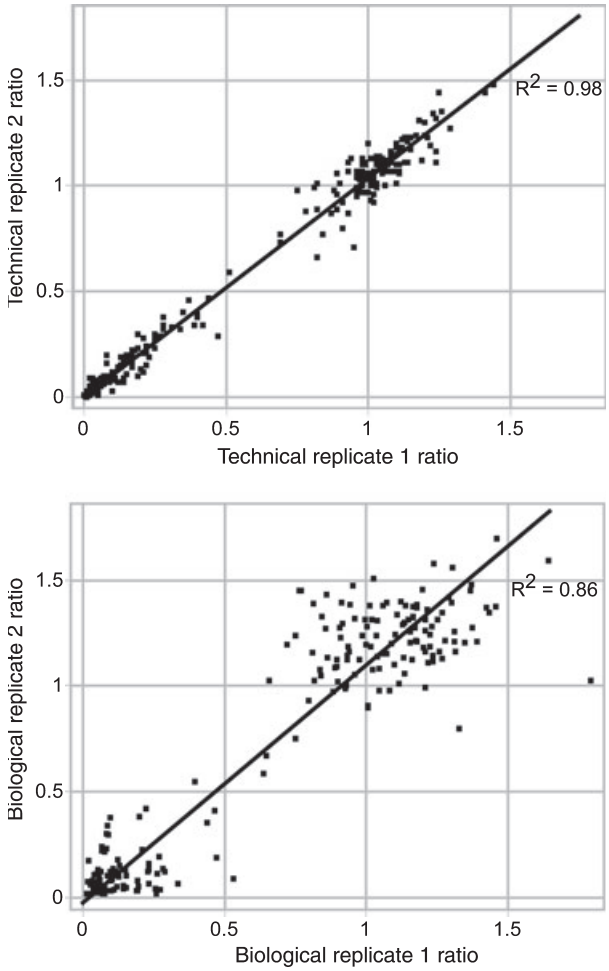

Figure 3. Indel array repeatability analysis.
Randomly selected Col × L*er* RIL versus Col hybridization. Upper panel, technical replicate ratios. Lower panel, biological replicate ratios. Tha data fall into two clusters: lower ratios represent the L*er* genotypic state in the RIL, higher ratios represent Col.

ities. Confidence in the status of other markers can be improved by Bayesian inference. Here, the prior probability of each genotype at a questionable marker is combined with information on flanking marker genotypes to create a posterior probability (see Experimental procedures).

Following this analysis strategy, a graphical map of RIL CS1982 (Figure 6) was generated from normalized ratio data (Figure 4). We compared such calculated maps of four Col × L*er* RILs with high-density SNP genotyping maps (Figure 6 and Table S6). We observed no conflicts between indel and SNP data, highlighting the accuracy of the indel array (Figure 6). Thus, hybridization and map generation can be conducted in a high-throughput manner with a robust and simple statistical method for data analysis.

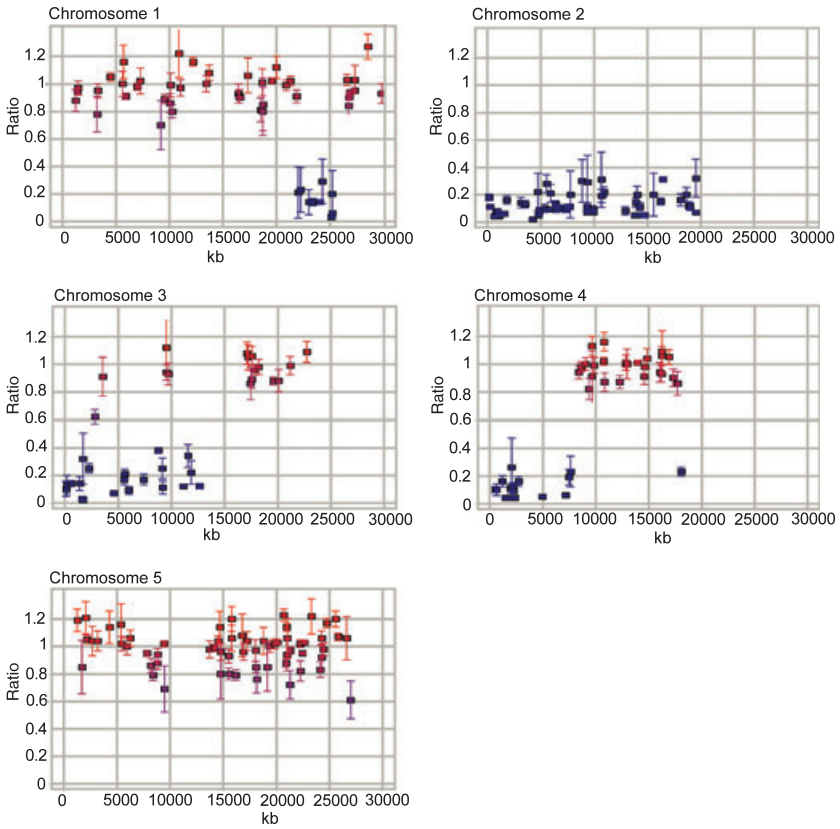

Figure 4. Raw ratio data of the Col × L*er* RIL CS1982 hybridization versus Col.
Markers are colored in gradation from blue, representing low ratio values scored as L*er*, to red, representing ratio values higher than 0.5 scored as Col.

Array versatility

Mapping Mendelian mutations in segregating populations. We investigated the utility of the indel array for mapping a fully penetrant recessive mutation in a segregating population, a time- and labor-intensive endeavor in Arabidopsis. Recent studies have shown promise in using Affymetrix and other SNP-based platforms to map mutations (Borevitz *et al.*, 2003; Hazen *et al.*, 2005a; Törjék *et al.*, 2003). The *erecta* mutation is present in the L*er* genotype and is easily scored. We used bulk segregant analysis to map the *erecta* mutation in a segregating Col × L*er* F_2 population, analogous to experiments performed by Borevitz *et al.* (2003). Of 606 F_2 plants, 457 were phenotypically wild-type and 149 showed the *erecta* mutant phenotype, a ratio very close to the expected 3:1. Single leaves were taken from each plant, and the two phenotypic classes were pooled. Genomic DNA from both pools was competitively hybridized to the indel array. Unlinked markers should not show a bias between pools, whereas markers linked to *erecta* should be enriched for the L*er* genotype in the mutant pool compared with the wild-type pool. As the L*er* genome should not bind to most markers on the array, low fluorescence signal is expected in the *ERECTA* region from the *erecta* pool. Indeed, low ratio signals formed a cluster on chromosome 2, centered at 11 Mb (Figure 7). *ERECTA* is located at 11.2 Mb on chromosome 2. Bulk segregant studies using Affymetrix arrays placed the *ERECTA* mutation within a 12 cM interval (Borevitz *et al.*, 2003). With the indel array, one can approximate the location of the *erecta* mutation with a similar resolution, sufficient for reliable first-pass mapping.

Disparate ecotypes. To investigate the utility of the array for genotyping lines derived from crosses other than Col × L*er*, we used PCR to establish that a randomly selected set of 20 indel markers are polymorphic among 18 accessions of Arabidopsis. At many loci, each state of the indel allele is shared among several accessions (Table S7), suggesting that Col/L*er* indel polymorphisms are useful for genetic mapping with a wide variety of Arabidopsis accessions. Therefore, we performed CGH experiments for Col-0 versus the accessions Kas-2, Nd-1, Tsu-1 and Cvi-0, which have been used as parents of RIL sets. We identified between 66 and 91 markers with an accession/Col signal ratio of less than 0.5 (Table S8). We also showed that 82 out of 215 markers are polymorphic between Bay-0 and Shahdara accessions and 161 out of 191 markers are polymorphic between L*er* and Cvi-0 (Table S8). Some markers failed to hybridize, presumably due to the deletion allele of the indel occurring in both accessions, or other sequence divergence. Note that the current array design is optimized for Col × L*er*;

Ratio bin	p(Col)	p(L*er*)	% Markers
<0.1	0.011	0.989	23
0.1–0.2	0.009	0.991	12
0.2–0.3	0.031	0.969	5
0.3–0.4	0.086	0.914	3
0.4–0.5	0.306	0.694	2
0.5–0.6	0.696	0.304	2
0.6–0.7	0.901	0.099	3
0.7–0.8	0.964	0.036	5
0.8–0.9	0.962	0.038	9
0.9–1.0	0.983	0.017	11
1.0–1.1	0.976	0.024	10
1.1–1.2	0.995	0.005	6
1.2–1.3	0.992	0.008	3
>1.3	0.986	0.014	6

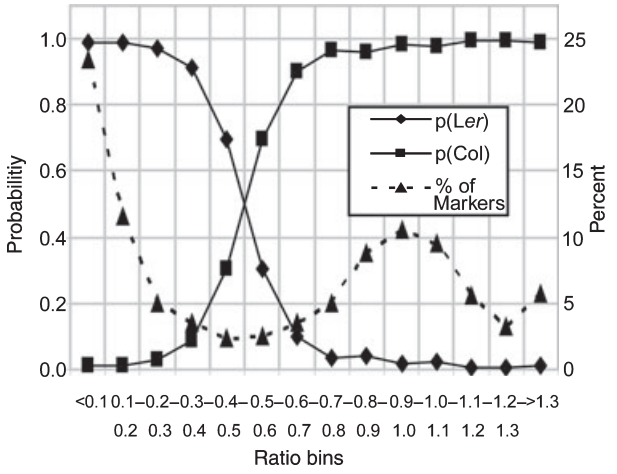

Figure 5. Error rate analysis.
The probability of a particular signal ratio corresponding to a particular marker genotype was determined by comparing indel array ratios from four RILs (CS1982, CS1990, CS1992 and CS1994; Lister and Dean, 1993) with high-resolution SNP hybridization maps. Top, the probability that a marker of known genotype falls into an array ratio bin and the percentage of markers in each bin. This probability is used as the false call error rate for determining the genotype of an unknown marker displaying such a ratio. Bottom, graphical representation.

further markers polymorphic between other accessions can be identified from the current L*er* sequence (Jander *et al.*, 2002) and on-going large-scale resequencing efforts (Perlegen Sciences Inc., http://www.perlegen.com/index.htm?newsroom/pr/2004/2004_10_05_Perlegen_Planck_Salk_Arabidopsis_Press_Release.html).

To confirm that these potential polymorphisms are genuine, we performed PCR for 8–15 randomly selected markers per accession. No contradictions were observed between the PCR and array data for Kas-2 × Col, Nd-1 × Col, Tsu-1 × Col and Bay-0 × Shahdara datasets. Thus, whereas Illumina SNP-based genotyping offers more markers at a comparable price for diverse RIL populations containing several hundred lines, our data demonstrate that the indel array is useful and cost-efficient for mapping small populations from crosses of diverse Arabidopsis accessions (Table S1).

Identifying heterozygous blocks. We further demonstrate the ability of the indel array to identify segments of heterozygosity (Figure 8). One Col × L*er* CSS (chromosome substitution strain) and two STAIRS lines (stepped aligned inbred recombinant strains; Koumproglou *et al.*, 2002) were backcrossed to Col, and the F_1 progeny were genotyped with the indel array. Blocks of heterozygosity are expected to result in intermediate signal ratios; indeed, we observed this result for heterozygous segments in all three crosses. Whereas single markers could not reliably be scored as heterozygous, a sliding-window analysis yielded signal ratios for heterozygous segments significantly different from Col. As heterozygous ratios are far more similar to those from Col segments, they can be easily distinguished from typical L*er* ratios.

Discussion

We have developed a technique employing a spotted microarray platform to genotype Arabidopsis, with 277 array elements representing 240 unique indel markers polymorphic between Col and L*er*. Offering high-throughput genotyping using commonly available equipment at low cost, the indel array is applicable for both large and small sample sizes, thereby accommodating routine mapping requirements of Arabidopsis research. The high quality of indel array data eliminates the need for replicate experiments, and data analysis is streamlined. We supply PCR primer sequences for all markers on the array, which can be used by the Arabidopsis community when genome-wide mapping is not desired.

A primary application of the array will be the mapping of segregating recessive mutations in F_2 populations of diverse Arabidopsis accessions using bulk segregant analysis. In Arabidopsis, such mutations have been mapped previously using an Affymetrix platform (Borevitz *et al.*, 2003; Hazen *et al.*, 2005a,b). By mapping the well-known developmental mutation *erecta* (Figure 7), we showed that the indel array offers similar precision without replication and at a fraction of the cost (Table S1). Numerous mutants may result from large forward mutagenesis screens, many of which may correspond to the same or previously identified loci. Complementation analysis is traditionally used to exclude previously mapped and multiply sampled loci (Weigel and Glazebrook, 2002, p. 48). The labor required for comprehensive complementation analysis increases exponentially with both the number of identified mutants and the number of known complementation groups. Given the low cost and high throughput of indel array genotyping, bulk segregant analysis of F_2 populations can be used instead of complementation analysis.

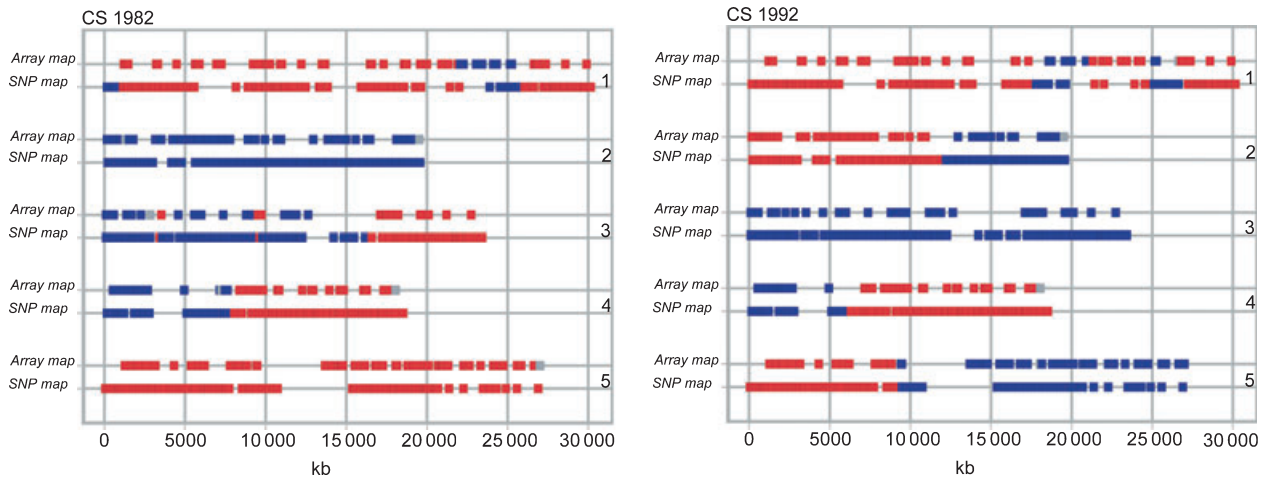

Figure 6. Comparative array accuracy.
Representative indel array data for two Col × L*er* RILs are compared to SNP hybridization data. The map for CS1982 is derived from raw data in Figure 4. Mapping data for each chromosome are shown above the corresponding SNP data. Markers of the Col and L*er* genotypes are colored red and blue, respectively. Missing marker data or markers whose genotype could not be determined from the indel array are represented in gray.

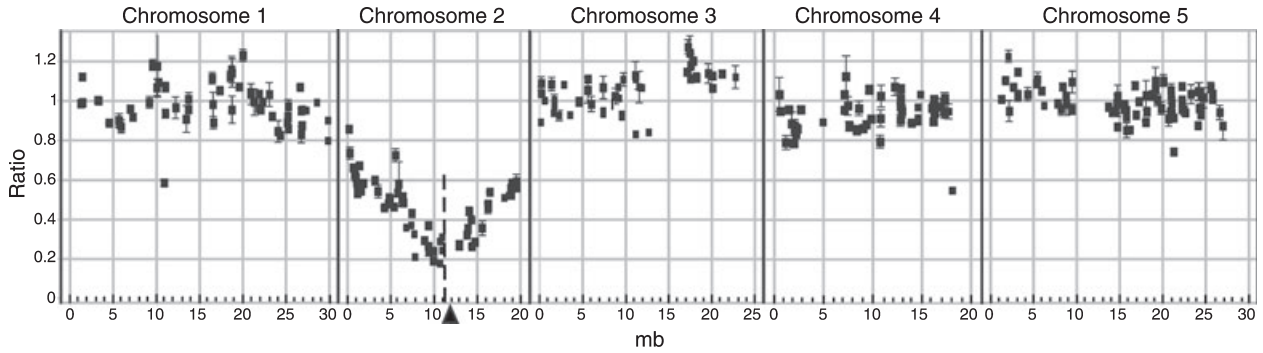

Figure 7. Utility of indel array for bulk segregant analysis.
An F_2 population segregating for *erecta* was divided into phenotypically wild-type and mutant pools. DNA from the pools was hybridized competitively. The raw signal ratio is plotted along all Arabidopsis chromosomes. Whereas the ratios on chromosomes 1, 3, 4 and 5 are consistently near equality, the increasingly low ratios on chromosome 2 demonstrate preferential transmission of the L*er* genotype to the mutant pool. The lowest ratios cluster around the actual location of *ERECTA* (filled triangle).

Another application of the array will be in genotyping RIL populations, especially those containing fewer than 96 lines and those based on either Col or L*er*. Such populations are of considerable interest for researchers investigating the interaction of specific genetic perturbations with natural variation. Most Arabidopsis mutants are isolated in either the Col or the L*er* background. We demonstrate that RIL maps generated by the indel array are highly accurate. Moreover, we provide PCR markers for all array elements, thereby facilitating rapid and simple characterization of near-isogenic lines (NILs) as required for QTL confirmation and fine-mapping.

Using the indel array, we estimate that a skilled researcher can genotype 96 lines per week. Genotyping each line with 277 markers costs under US$50. This calculation includes the list price of significant reagents required for a single hybridization in a 16-well array (DNA probe clean-up, glass slide, DNA-labeling kit and fluorophore dyes). Due to the minimal requirements for experimental replication and labor, our array method is significantly cheaper than PCR and SNP-based Affymetrix or NimbleGen technology for any sample size. The costs of commercial genotyping services using Illumina technology increase exponentially with decreasing sample size, whereas the costs of our genotyping method scale linearly with the number of mapped lines. For mapping a large sample size of Col × L*er* lines, our approach offers a similar number of markers as Illumina SNP-based genotyping at a comparable price, while offering significant cost savings for small populations. Taken together, the indel array offers an affordable complementary method to Illumina SNP-based genotyping. A detailed breakdown of costs and a comparison to other currently available genotyping methods is given in Table S1.

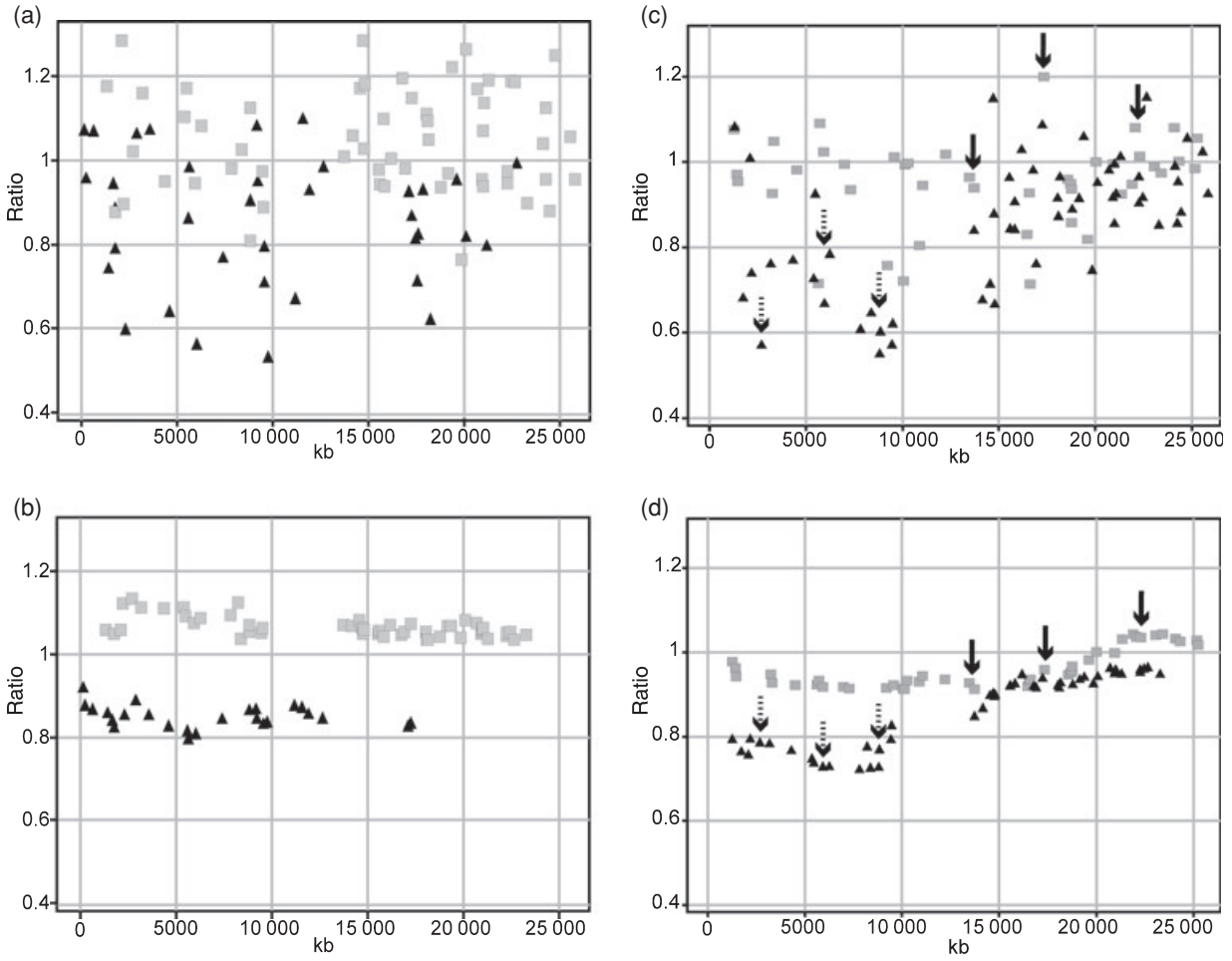

Figure 8. Identifying heterozygous blocks using the indel array.
(a, b) Ratio data of Col × CSS line CS9434 F_1 line versus Col hybridization. Black triangles represent ratios for chromosome 3, which is heterozygous for Col and L*er*. Gray squares represent ratios for a representative homozygous Col chromosome (chromosome 5).
(a) Raw ratio data.
(b) Raw ratio data using sliding-window averages of 10 marker ratios.
(c, d) Ratio data of a representative F_1 line (Col × STAIRS CS9470) versus Col hybridization. Black triangles represent ratios for the partial Col–L*er* heterozygous chromosome 5. Regions of heterozygosity and homozygosity were confirmed by PCR (dashed arrows, heterozygous Col × L*er* loci; solid arrows, Col homozygous loci). Gray squares represent ratios for a representative homozygous Col chromosome (chromosome 1).
(c) Raw ratio data.
(d) Raw ratio data using sliding-window averages of 10 marker ratios.
Regions of heterozygosity are readily distinguishable from Col homozygous regions. Note that distinguishing Col × L*er* heterozygous regions from L*er* homozygous regions is far easier than differentiating from Col homozygous regions due to the significantly greater separation of ratios (see Figures 4 and 5).

Further comparison of the available Col and L*er* genomic sequences (Arabidopsis Genome Initiative, 2000; Jander *et al.*, 2002), along with the ongoing large-scale genotyping efforts for many diverse Arabidopsis accessions (Perlegen Sciences Inc., http://www.perlegen.com/index.htm?newsroom/pr/2004/2004_10_05_Perlegen_Planck_Salk_Arabidopsis_Press_Release.html), will provide additional indel markers that can be easily included as microarray elements and will increase the versatility of the indel array for diverse accessions. Indel arrays may prove particularly powerful for mapping in organisms with more complex genomes than Arabidopsis, such as barley, where DNA hybridization cannot currently be used for SFP detection (Cui *et al.*, 2005). Although RNA hybridization data can aid SFP detection in barley and Arabidopsis, experimental preparation time and cost are substantially increased, and the resulting map is biased toward gene-rich regions (Cui *et al.*, 2005; West *et al.*, 2006). Further, it is already possible to construct indel arrays for other systems, as many studies have reported an abundance of indels between strains of model organisms such as *Caenorhabditis elegans* and *Drosophila melanogaster* (Britten *et al.*, 2003), as well as important crop species such as rice (Shen *et al.*, 2004). Indeed, a similar approach to genotyping is

being developed for detecting indel polymorphisms in rice mapping populations (D. Galbraith, University of Arizona, personal communication). With the decreasing costs of large-scale sequencing, such as those provided by the 454 Life Sciences and Solexa platforms (Bentley, 2006; Kaller *et al.*, 2007; Margulies *et al.*, 2005), partial genome sequences of many non-model organisms will identify polymorphic elements. These elements will be usable for both indel arrays and SNP-based methods, allowing similar approaches to genotyping as described here for Arabidopsis.

Experimental procedures

All documentation of reagents, protocols and analysis scripts can also be obtained from the Queitsch lab web-page: http://www.sysbio.harvard.edu/csb/queitschlab/array. Printed microarray slides (AL and MPX) are available at cost (cqueitsch@cgr.harvard.edu).

Oligonucleotide design

Indel sizes and their 20 bp flanking sequences were obtained from the Monsanto *Arabidopsis* polymorphism and L*er* sequence collection available on TAIR (sequence release 3; http://www.arabidopsis.org/Cereon/). We chose indels longer than 25 bp to maximize differential hybridization. Flanking sequences were localized in the Arabidopsis genome through local alignments using BLAST (Altschul *et al.*, 1997). Insertion positions in the Col and L*er* genomes were determined using the raw genomic sequence reads from TAIR (sequence release 6.0, downloaded 19 July 2006). Flanking sequences that did not return perfect BLAST hits or for which the location of the 5′ and 3′ flanking sequences did not correspond to indel size were discarded. We then extracted 40 bp of sequence on both sides of the center of the insertion to derive a fragment of 80 bp. These fragments were then locally aligned using BLAST in both genomes, and only those that returned a perfect hit in the accession with the insertion and no perfect hit in the accession with the deletion were considered for further steps. Based on these 80 bp sequences, eleven 70 bp oligos were designed for each indel marker by sliding the start position by one nucleotide in the 80 bp window. The best 70 bp oligonucleotide for each indel marker was chosen based on its GC content (percentage of cytosine and guanine), which was held as close as possible to 50%. Another criterion for oligonucleotide choice was the lack of homopolymeric runs longer than 12 bp (calculated as the longest run of adenine and/or thymine and cytosine and/or guanine).

For microarray printing, C6–5′amino modified 70-mer salt-free oligonucleotides were synthesized by Operon Biotechnologies (http://www.operon.com) at a 0.2 μmol scale. Oligonucleotide DNA was diluted to 60 μM in 3× SSC/0.01% SDS for printing.

PCR testing of indel markers

PCR primers flanking the predicted indel sequences were designed using Primer3 software (Rozen and Skaletsky, 2000). Primer sequences and expected amplicon sizes for Col and L*er* template DNA are given in Table S2. PCR cycle conditions were: 96°C for 5 min, then 40 cycles of 94°C for 30 sec, 60°C for 30 sec and 72°C for 2 min, followed by 72°C for 10 min. PCR products were run on a 2.5–3% agarose gel. PCR reagents and DNA extraction method were as described previously (Weigel and Glazebrook, 2002, p. 168).

Sentrix® bead array SNP genotyping

A previously described protocol (Kliebenstein *et al.*, 2007) was used for fiber optic detection of 516 Arabidopsis SNP genotyping reactions on glass beads self-assembled into wells on fiber optic bundles arrayed in a 96-well format [Sentrix® Array Matrix (SAM), Illumina, http://www.illumina.com]. This set of 516 SNPs was used to genotype four Col × L*er* RILs to estimate the indel array error rate (Table S6). SNP maps generated for an additional 56 Col × L*er* RILs using this method are provided in Table S9.

Slide surface

Aldehyde slide surface chemistry was chosen as yielding the best results. For genotyping RILs, the 16-array/slide format Schott Nexterion slides (MPX) (www.schott.com), with double-sided adhesive superstructures, were used. Schott Nexterion slides (AL) were used for bulk segregant mapping of the *erecta* mutation in F_2 lines.

Printing and processing of indel oligonucleotide array slides

Slides were printed using a Genemachines® Omnigrid™ machine (Genemachines® software version 4.2.0.2, http://www.genomicsolutions.com). Indel oligonucleotides were printed from a 60 μM DNA, 3× SSC, 0.01% SDS solution. Printed indel oligonucleotides were immobilized to the glass surface upon dehydration by a covalent bond via the Schiff base. Slides were blocked to quench remaining active aldehyde groups using 0.05 M $NaBH_4$, and washed with NH_4OH (28–30% NH_3) to increase the binding efficiency of DNA hybridization. A detailed description of this protocol is given in Appendix S1.

Lines used for heterozygote block identification

Three lines were used for array-based heterozygote block identification: CSS line CS9434 and STAIRS lines CS9431 and CS9470 (Koumproglou *et al.*, 2002). These were backcrossed to Col wild-type plants. DNA from the resulting F_1 plants was hybridized with Col wild-type DNA. Representative data from two lines are given in Figure 8.

DNA preparation

One gram (fresh weight) of plant tissue was frozen in liquid nitrogen and ground to a fine powder. DNA was extracted using the Plant DNAeasy MaxiPrep kit (Qiagen; http://www.qiagen.com/), substituting the supplied RNase with 1× RNase ONE ribonuclease (Promega; http://www.promega.com/; 50 U RNase/1 g fresh weight tissue). Extracted DNA was washed using Microcon YM30 centrifugal filters (Millipore, http://www.millipore.com) and sonicated (550 sonic dismembrator, Fisher Scientific, http://www.fishersci.com), yielding fragments averaging 100–200 bp. DNA was concentrated to 600 ng μl^{-1} using Microcon YM30 filters. A detailed protocol is given in Appendix S2.

DNA labeling, hybridization and washing

Direct labeling of sonicated DNA was performed using the BioPrime® DNA labeling kit (Invitrogen; http://www.invitrogen.com/), but biotinylated dNTP were substituted with non-biotinylated dNTP (Invitrogen). Cy-3 and Cy-5 fluorophores (Amersham; http://www5.amershambiosciences.com/) were used in labeling reactions. Comparative hybridization was performed with 6 μg (600 ng μl^{-1}) of each labeled probe DNA to indel array oligonucleotides. For the AL and MPX slides, respectively, 37 and 50 μl of hybridization solution (6 μg DNA, 4.6× SSC, 14 μg yeast tRNA, 0.001% SDS, 0.001 M DTT) were used per hybridization. Hybridization was carried out at 65°C for 16 h in a water bath or a rotating hybridization oven for AL and MPX slide types, respectively. Following this, slides were washed to remove non-specific hybridization (2× SSC, 0.001% SDS, 0.001 M DTT for 5 min, 0.1× SSC, 0.001% SDS, 0.001 M DTT for 5 min, 0.1× SSC, 0.001 M DTT for 2 min, 0.01× SSC, 0.001 M DTT for 2 min). Washes were performed in an ozone-reduced room with ozone levels below 20 ppb (parts per billion) to minimize Cy-5 bleaching. Details are provided in Appendix S3.

Data acquisition and analysis

Slides were scanned using an Axon 4000B scanner (http://www.moleculardevices.com). Image files were analyzed using GenePix Pro 5.1 software (http://www.moleculardevices.com), which calculates the median signal intensity minus background noise of each fluorophore channel for each marker. For genotyping, we calculated preliminary signal ratios by dividing the signal of Cy-3-labeled DNA by that of Cy-5-labeled Col DNA. These ratios were normalized with respect to positive control markers on the array (see Table S5 for marker type). A 0.5 ratio cut-off was used to infer polymorphic status in the Col-0 × Kas-2, Col-0 × Nd-1, Col-0 × Tsu-1 and Col-0 × Cvi-0 CGH experiments. For Bay-0 × Sha and L*er*-0 × Cvi-0 hybridizations, ratios of <0.5 or >2 were indicative of polymorphic marker status. In the *ERECTA* mapping experiment, we divided the signal from the *erecta* mutant phenotype DNA pool (labeled with Cy-3) by that of the *ERECTA* wild-type DNA pool (labeled with Cy-5), therefore low array ratio values indicate regions enriched for the L*er* genomic composition.

To automate analysis of GenePix output data, a Visual Basic macro module was developed. Markers flagged by the user as having poor binding in an experiment or previously verified as being sub-optimal for genotyping purposes are removed. Indel markers with combined Cy-3 and Cy-5 median signals <75 fluorescence units above background are also deleted. The macro calculates the ratio of median dye signals minus background for each indel marker. A normalization factor is calculated by averaging ratios of the positive controls. Positive control markers that have a ratio of <0.3 or >3.0 are excluded in the normalization factor calculation. Indel marker ratios are subsequently normalized by dividing by the normalization factor, giving ratio values for each marker. Ratio values for technical replicate markers and markers of different sequence representing the same indel polymorphism are averaged. A script to implement this analysis is given in Appendix S4.

To estimate the false call rate at various normalized ratio values, we used a training set of four Col × L*er* RILs (CS1982, CS1990, CS1992 and CS1994; Lister and Dean, 1993) and compared indel array ratios to high-resolution SNP hybridization maps. The known genotypic state of each marker was placed into the experimentally determined array ratio bin, with a bin size of 0.1. The percentage of markers with a particular genotypic state in an array bin is applied to hybridizations of unknown genotype to estimate the confidence of each marker call. Non-informative marker calls may be improved by Bayesian posterior probability analysis. Briefly, consider three consecutive loci A, B and C with possible genotypic states *x* and *y*. By Bayes' theorem:

$$P(\mathrm{B}=x|\mathrm{A}=x,\mathrm{C}=x) = \frac{P(\mathrm{A}=x,\mathrm{C}=x|\mathrm{B}=x)P(\mathrm{B}=x)}{P(\mathrm{A}=x,\mathrm{C}=x|\mathrm{B}=x)P(\mathrm{B}=x)+P(\mathrm{A}=x,\mathrm{C}=x|\mathrm{B}=y)P(\mathrm{B}=y)} \quad (1)$$

where $P(\mathrm{B} = x)$ is the prior probability of genotype *x* and locus B from experimental training data, $P(\mathrm{B} = y) = 1 - P(\mathrm{B} = x)$, and $P(\mathrm{A} = x,\mathrm{C} = x|\mathrm{B} = x)$ is estimated from genetic map distances. The final posterior probability for genotype *x* at locus B may be calculated by multiplying equation (1) by the prior probabilities of the assigned genotypes of A and C from the training data, and summing over all four combinations of genotypic states of A and C:

$$P(\mathrm{B}=x) = \sum_{m=x}^{y}\sum_{n=x}^{y} P(\mathrm{B}=x|\mathrm{A}=m,\mathrm{C}=n)P(\mathrm{A}=m)P(\mathrm{C}=n) \quad (2)$$

A script to implement this analysis is given in Appendix S5.

Acknowledgements

This work was supported by National Science Foundation Arabidopsis 2010 grant number 0313473. The Monsanto Company is gratefully acknowledged for providing access to Arabidopsis marker data. Todd A. Sangster is an HHMI pre-doctoral fellow. We would like to thank Paul Grosu and Dr Mike Slack for help with indel marker design and Christian Daly for technical assistance in printing slides. We thank Dr Susan Lindquist (Whitehead Institute for Biomedical Research, Cambridge, MA, USA) for discussions and support. We would also like to thank the two anonymous reviewers for helpful comments and suggestions.

Supplementary material

The following supplementary material is available for this article online:

Table S1. Cost comparison of indel array technology.
Table S2. Indel array markers with PCR marker details.
Table S3. Polymorphic PCR markers not present on the indel array.
Table S4. Indel marker quality testing.
Table S5. Characteristics of markers present on the indel array.
Table S6. Indel array genotyping data and comparison to SNP maps.
Table S7. PCR testing of marker segregation in 17 Arabidopsis accessions.
Table S8. Polymorphic indel marker ratio data in accession hybridization experiments.
Table S9. Sentrix® bead array SNP genotyping of 60 Columbia/Landsberg *erecta* recombinant inbred lines.
Appendix S1. Printing, immobilization of indel oligonucleotides and blocking of slides.
Appendix S2. Pre-hybridization DNA preparation.
Appendix S3. Hybridization and washing protocols for (a) the AL slide type and (b) the MPX slide type.
Appendix S4. Visual Basic script for marker filtering and normalization.
Appendix S5. Script to implement Bayesian probability analysis.
This material is available as part of the online article from http://www.blackwell-synergy.com

References

Albert, T.J., Dailidiene, D., Dailide, G., Norton, J.E., Kalia, A., Richmond, T.A., Molla, M., Singh, J., Green, R.D. and Berg, D.E. (2005) Mutation discovery in bacterial genomes: metronidazole resistance in *Helicobacter pylori*. *Nat. Methods*, **2**, 951–953.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **406**, 796–815.

Bell, C.J. and Ecker, J.R. (1994) Assignment of 30 microsatellite loci to the linkage map of Arabidopsis. *Genomics*, **19**, 137–144.

Bentley, D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* **16**, 545–552.

Borevitz, J. (2006) Genotyping and mapping with high-density oligonucleotide arrays. *Methods Mol. Biol.* **323**, 137–145.

Borevitz, J.O., Liang, D., Plouffe, D., Chang, H.-S., Zhu, T., Weigel, D., Berry, C.C., Winzeler, E. and Chory, J. (2003) Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* **13**, 513–523.

Britten, R.J., Rowen, L., Williams, J. and Cameron, R.A. (2003) Majority of divergence between closely related DNA samples is due to indels. *Proc. Natl Acad. Sci. USA*, **100**, 4661–4665.

Copenhaver, G.P., Nickel, K., Kuromori, T. *et al.* (1999) Genetic definition and sequence analysis of Arabidopsis centromeres. *Science*, **286**, 2468–2474.

Cui, X., Xu, J., Asghar, R., Condamine, P., Svensson, J.T., Wanamaker, S., Stein, N., Roose, M. and Close, T.J. (2005) Detecting single-feature polymorphisms using oligonucleotide arrays and robustified projection pursuit. *Bioinformatics*, **21**, 3852–3858.

Erickson, D.L., Fenster, C.B., Stenoien, H.K. and Price, D. (2004) Quantitative trait locus analyses and the study of evolutionary process. *Mol. Ecol.* **13**, 2505–2522.

Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G. and Chee, M.S. (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.* **37**, 549–554.

Gunderson, K.L., Steemers, F.J., Ren, H. *et al.* (2006) Whole-genome genotyping. *Methods Enzymol.* **410**, 359–376.

Hazen, S.P., Borevitz, J.O., Harmon, F.G., Pruneda-Paz, J.L., Schultz, T.F., Yanovsky, M.J., Liljegren, S.J., Ecker, J.R. and Kay, S.A. (2005a) Rapid array mapping of circadian clock and developmental mutations in Arabidopsis. *Plant Physiol.*, **138**, 990–997.

Hazen, S.P., Schultz, T.F., Pruneda-Paz, J.L., Borevitz, J.O., Ecker, J.R. and Kay, S.A. (2005b) *LUX ARRHYTHMO* encodes a Myb domain protein essential for circadian rhythms. *Proc. Natl Acad. Sci. USA*, **102**, 10387–10392.

Jander, G. (2006) Gene identification and cloning by molecular marker mapping. *Methods Mol. Biol.* **323**, 115–126.

Jander, G., Norris, S.R., Rounsley, S.D., Bush, D.F., Levin, I.M. and Last, R.L. (2002) Arabidopsis map-based cloning in the post-genome era. *Plant Physiol.* **129**, 440–450.

Kaller, M., Lundeberg, J. and Ahmadian, A. (2007) Arrayed identification of DNA signatures. *Expert Rev. Mol. Diagn.* **7**, 65–76.

Kane, M.D., Jatkoe, T.A., Stumpf, C.R., Lu, J., Thomas, J.D. and Madore, S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.* **28**, 4552–4557.

Kliebenstein, D.J., D'Auria, J.C., Behere, A.S., Kim, J., Gunderson, K.L., Breen, J.N., Gershenzon, J., Last, R.L. and Jander, G.(2007) Characterization of seed-specific benzoyloxyglucosinolate mutations in *Arabidopsis thaliana*. *Plant J.* (in press).

Koornneef, M., Alonso-Blanco, C. and Vreugdenhil, D. (2004) Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annu. Rev. Plant Biol.* **55**, 141–172.

Koumproglou, R., Wilkes, T.M., Townson, P., Wang, X.Y., Beynon, J., Pooni, H.S., Newbury, H.J. and Kearsey, M.J. (2002) STAIRS: a new genetic resource for functional genomic studies of Arabidopsis. *Plant J.*, **31**, 355–364.

Lister, C. and Dean, C. (1993) Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. *Plant J.* **4**, 745–750.

Margulies, M., Egholm, M., Altman, W.E. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Ponce, M.R., Robles, P., Lozano, F.M., Brotons, M.A. and Micol, J.L. (2006) Low-resolution mapping of untagged mutations. *Methods Mol. Biol.* **323**, 105–113.

Rozen, S. and Skaletsky, H.J.(2000) Primer3 on the WWW for general users and for biologist programmers. In *Bioinformatics Methods and Protocols: Methods in Molecular Biology* (Misener, S. and Krawetz, S.A., eds) Totowa, NJ: Humana Press, pp. 365–386.

Selzer, R.R., Richmond, T.A., Pofahl, N.J., Green, R.D., Eis, P.S., Nair, P., Brothman, A.R. and Stallings, R.L.(2005) Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosomes Cancer*, **44**, 305–319.

Shen, Y.J., Jiang, H., Jin, J.P. *et al.*(2004) Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. *Plant Physiol.* **135**, 1198–1205.

Storm, N., Darnhofer-Patel, B., van den Boom, D. and Rodi, C.P. (2003) MALDI-TOF mass spectrometry-based SNP genotyping. *Methods Mol. Biol.* **212**, 241–262.

Törjék, O., Berger, D., Meyer, R.C., Mussig, C., Schmid, K.J., Rosleff Sorensen, T., Weisshaar, B., Mitchell-Olds, T. and Altmann, T. (2003) Establishment of a high-efficiency SNP-based framework marker set for Arabidopsis. *Plant J.* **36**, 122–140.

Weigel, D. and Glazebrook, J. (2002) *Arabidopsis: A Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

West, M.A., van Leeuwen, H., Kozik, A., Kliebenstein, D.J., Doerge, R.W., St Clair, D.A. and Michelmore, R.W. (2006) High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis. *Genome Res.* **16**, 787–795.

