Highly Parallel SNP Genotyping

J.-B. FAN,* A. OLIPHANT,* R. SHEN,* B.G. KERMANI,* F. GARCIA,* K.L. GUNDERSON,* M. HANSEN,* F. STEEMERS,* S.L. BUTLER,*‡ P. DELOUKAS,† L. GALVER,* S. HUNT,† C. MCBRIDE,* M. BIBIKOVA,* T. RUBANO,* J. CHEN,* E. WICKHAM,* D. DOUCET,* W. CHANG,* D. CAMPBELL,* B. ZHANG,*¶ S. KRUGLYAK,* D. BENTLEY,† J. HAAS,*§ P. RIGAULT,* L. ZHOU,* J. STUELPNAGEL,* AND M.S. CHEE*

**Illumina, Inc., San Diego, California 92121; †The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom*

The genetic factors underlying common disease are largely unknown. Discovery of disease-causing genes will transform our knowledge of the genetic contribution to human disease, lead to new genetic screens, and underpin research into new cures and improved lifestyles. The sequencing of the human genome has catalyzed efforts to search for disease genes by the strategy of associating sequence variants with measurable phenotypes. In particular, the Human Genome Project and follow-on efforts to characterize genetic variation have resulted in the discovery of millions of single-nucleotide polymorphisms (SNPs) (Patil et al. 2001; Sachidanandam et al. 2001; Reich et al. 2003). This represents a significant fraction of common genetic variation in the human genome and creates an unprecedented opportunity to associate genes with phenotypes via large-scale SNP genotyping studies.

To make use of this information, efficient and accurate SNP genotyping technologies are needed. However, most methods were designed to analyze only one or a few SNPs per assay, and are costly to scale up (Kwok 2001; Syvanen 2001). To help enable genome-wide association studies and other large-scale genetic analysis projects, we have developed an integrated SNP genotyping system that combines a highly multiplexed assay with an accurate readout technology based on random arrays of DNA-coated beads (Michael et al. 1998; Oliphant et al. 2002; Gunderson et al. 2004). Our aim was to reduce costs and increase productivity by ~2 orders of magnitude. We chose a multiplexed approach because it is more easily scalable and is intrinsically cost-efficient (Wang et al. 1998). Although existing multiplexed approaches lacked the combination of accuracy, robustness, scalability, and cost-effectiveness needed for truly large-scale endeavors (Wang et al. 1998; Ohnishi et al. 2001; Patil et al. 2001; Dawson et al. 2002; Gabriel et al. 2002), we hypothesized that some of these limitations could be overcome by designing an assay specifically for multiplexing.

To increase throughput and decrease costs by ~2 orders of magnitude, it was necessary to eliminate bottlenecks throughout the genotyping process. It was also desirable to minimize sources of variability and human error in order to ensure data quality and reproducibility. We therefore took a systems-level view to technology design, development, and integration. Although the focus of this paper is on a novel, highly multiplexed genotyping assay, the GoldenGate™ assay, four other key technologies that were developed in parallel, as part of the complete BeadLab system (Oliphant et al. 2002), are briefly described below.

BEADARRAY™ PLATFORM

We developed an array technology based on random assembly of beads in micro-wells located at the end of an optical fiber bundle (Michael et al. 1998). This technology has advantages over conventional microarrays and is particularly suited to the needs of high-throughput genotyping (Oliphant et al. 2002; Gunderson et al. 2004). Arrays currently in use have up to 50,000 beads, each ~3 microns in diameter. The beads are distributed among 1,520 bead types, each bead type representing a different oligonucleotide probe sequence. This gives, on average, ~30 copies of each bead type, with the result that a genotype call is based on the average of many replicates. The inherent redundancy increases robustness and genotyping accuracy.

We took advantage of the fact that the arrays have a small footprint to design an *array matrix*, comprising 96 arrays arranged in an 8 x 12 matrix that matches the well spacing of a standard microtiter plate (Fig. 1). With this format, samples can be processed in standard microtiter plates, using standard laboratory equipment. The array

Figure 1. The Sentrix™ array matrix.

Present addresses: ‡Pfizer Global R&D, La Jolla Laboratories, 10777 Science Center Drive, San Diego, California 92121; ¶Activx Biosciences, 11025 N. Torrey Pines Road, La Jolla, California 92037; §13438 Russet Leaf Lane, San Diego, California 92129.

matrix is then mated to the microtiter plate, allowing 96 hybridizations to be carried out simultaneously. At the current multiplex level of 1,152, a total of ~110,000 genotypes can be obtained from each matrix of 96 arrays.

BEADARRAY READER

Scanners for conventional microarrays typically have imaging spot sizes in the range of 3–5 microns, insufficient to resolve the ~5-micron-spaced features on the randomly assembled optical fiber-based arrays. We therefore developed a compact confocal-type imaging system with ~0.8-μm resolution and two-laser illumination (532 and 635 nm; Barker et al. 2003). This scanner is able to image a 96-array matrix in both color channels in about 1.5 hours, which allows a throughput of ~8–10 array matrices, corresponding to ~1 million genotypes, per scanner per day.

AUTOMATION AND A LABORATORY INFORMATION MANAGEMENT SYSTEM

Automation of a process can be achieved by building a custom instrument that performs multiple functions without human intervention. However, mechanical integration tends to be inflexible: Even minor changes in the process might require a costly redesign of the instrument. An alternative strategy is to keep the system modular, and loosely but accurately coupled through a laboratory information management system (LIMS), designed hand-in-hand with automation (Oliphant et al. 2002). This flexible design philosophy is well-suited to molecular biology assays, which can be both complex and rapidly evolving.

In collaboration with Wildtype Informatics, we developed a LIMS that tracks objects as they are processed through the laboratory. Physical objects that contain samples and reagents, such as microtiter plates and array matrices, are bar-coded. The LIMS supervises each step where information is associated with a new object. For example, when samples are transferred from one plate to another, the robot application requests permission from LIMS to perform the process for the specified plate bar codes. Should LIMS approve the transaction, the robot proceeds with the process and sample transfer. After the process and sample transfers are complete, LIMS is automatically updated with the bar code information. This fail-safe approach, called positive sample tracking, eliminates common sources of human error, such as mislabeling and plate mix-ups.

OLIGATOR® DNA SYNTHESIZER

SNP assays require one or more oligonucleotides (the GoldenGate assay requires $3n + 3$ oligonucleotides for n SNPs), which are most efficiently generated by de novo chemical synthesis. Anticipating a need to create millions of SNP assays, we developed a high-throughput, low-cost centrifugal oligonucleotide synthesizer (Lebl et al. 2001). This LIMS-integrated automated instrument is able to produce hundreds to thousands of oligonucleotides per day. With this technology, we are able to develop SNP genotyping assays on a genome-wide scale, rapidly and cost-effectively.

DESIGN OF A HIGHLY MULTIPLEXED SNP GENOTYPING ASSAY

The GoldenGate assay was developed specifically for multiplexing to high levels while retaining the flexibility to choose any SNPs of interest to assay. There are a number of key design elements. In particular, the assay performs allelic discrimination directly on genomic DNA (gDNA), generates a synthetic allele-specific PCR template afterward, then performs PCR on the artificial template. In contrast, conventional SNP genotyping assays typically use PCR to amplify a SNP of interest. Allelic discrimination is then carried out on the PCR product. By reversing the conventional order, we require only three universal primers for PCR, and eliminate primer sequence-related differences in amplification rates between SNPs. We also attach the gDNA to a solid support prior to the start of the assay proper. After attachment, assay oligonucleotides targeted to specific SNPs of interest are annealed to the gDNA (Fig. 2). This attachment step improves assay specificity by allowing unbound and nonspecifically hybridized oligonucleotides to be removed by stringency washes. Correctly hybridized oligonucleotides remain on the solid phase.

Two allele-specific oligonucleotides (ASOs) and one locus-specific oligonucleotide (LSO) are designed for each SNP (Fig. 2). Each ASO consists of a 3′ portion that hybridizes to gDNA at the SNP locus, with the 3′ base complementary to one of two SNP alleles, and a 5′ portion that incorporates a universal PCR primer sequence (P1 or P2, each associated with a different allele). The LSOs consist of three parts: At the 5′ end is a SNP locus-specific sequence; in the middle is an address sequence, complementary to one of 1,520 capture sequences on the array; and at the 3′ end is a universal PCR priming site (P3′). Currently, a typical multiplex pool is designed to assay 1,152 SNPs, and thus contains 2,304 ASOs and 1,152 LSOs. The additional capacity of the array provides some room for expansion of the multiplex pool.

After the annealing and washing steps, an allele-specific primer extension step is carried out. This employs DNA polymerase to extend ASOs if their 3′ base is complementary to their cognate SNP in the gDNA template (Pastinen et al. 2000). Allele-specific extension is followed by ligation of the extended ASOs to their corresponding LSOs, to create PCR templates. Requiring the joining of two fragments to create a PCR template provides an additional level of genomic specificity. Any residual incorrectly hybridized ASOs and LSOs are unlikely to be adjacent, and therefore should not be able to ligate.

Next, the primers P1, P2, and P3 are added. P1 and P2 are fluorescently labeled, each with a different dye. For the SNP illustrated in Figure 2, where A and G represent the two alleles, the expected products are P1-A-P3 in the case of an AA homozygote, P2-G-P3 in the case of a GG

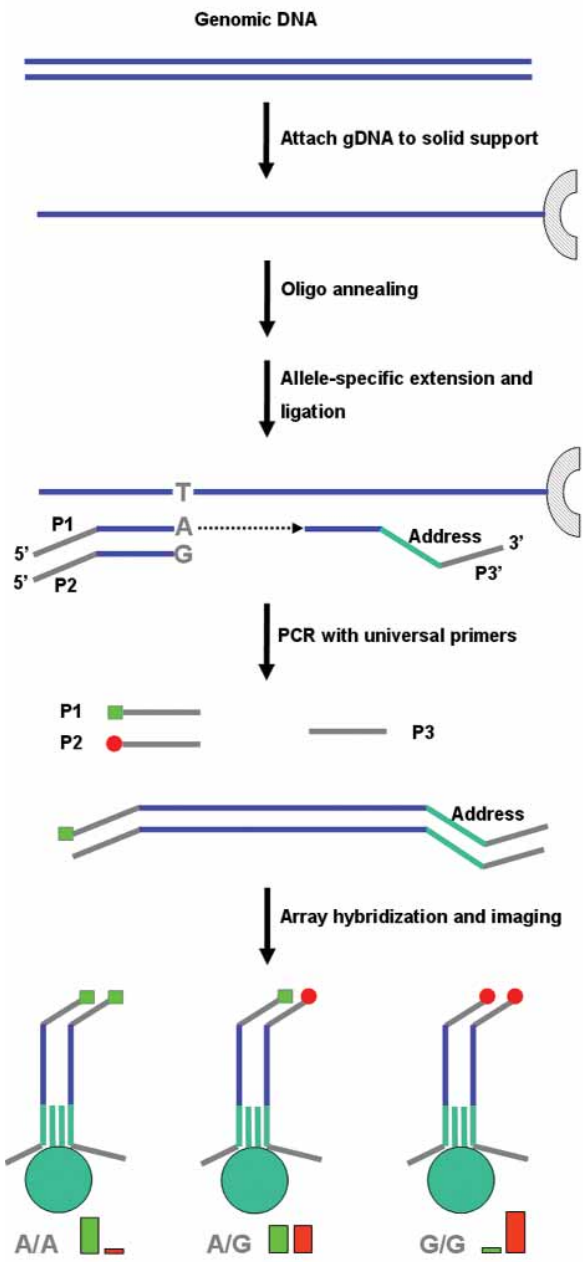

Figure 2. The GoldenGate SNP genotyping assay scheme. See Appendix for detailed procedures.

homozygote, or an equimolar mixture of P1-A-P3 and P2-G-P3 in the case of an AG heterozygote. Because P1 is associated with the A allele and P2 with the G allele, the ratio of the two primer-specific fluorescent signals identifies the genotype as AA, AG, or GG.

Each SNP is assigned a different address sequence, which is contained within the LSO. Each of these addresses is complementary to a unique capture sequence represented by one of the bead types in the array. Therefore, the products of the 1,152 assays hybridize to different bead types in the array, allowing all 1,152 genotypes to be read out simultaneously. This universal address system, consisting of artificial sequences that are not SNP specific, allows any set of SNPs to be read out on a common, standard array (Gerry et al. 1999; Cai et al. 2000; Chen et al. 2000; Fan et al. 2000; Iannone et al. 2000). This provides flexibility and reduces array manufacturing costs. Custom sets of assays can be made on demand, simply by building the address sequences into the SNP-specific assay oligonucleotides. The use of universal PCR primers to associate a fluorescent dye with each SNP allele also saves on costs. Because only three primers, two labeled and one unlabeled, are needed regardless of the number of SNPs to be assayed, the primer costs are negligible, as they are amortized over large numbers of assays.

The GoldenGate assay uses ~40 bp surrounding the SNP. Either strand can be chosen for the assay, but we use design rules to designate a preferred strand. The ASOs are designed to have a T_m of 60ºC (57–62ºC), whereas the LSO has a T_m of 57ºC (54–60ºC). It is possible to design a similar assay that omits the allele-specific polymerase extension step. However, we have found that allelic discrimination using polymerase, followed by ligation, increases the signal-to-noise ratio (data not shown) (Abravaya et al. 1995). Another advantage is that a variable gap between the ASOs and LSO (typically 1–20 bases) provides flexibility to position the LSO to avoid unfavorable sequences.

GENOTYPING RESULTS

To date, we have developed well in excess of 100,000 SNP assays. Representative data are shown in Figure 3. Figure 3A shows a fluorescence image from a single array in a 96-array matrix. The image is a false-color composite of the Cy3 (green) and Cy5 (red) images collected in separate channels. A portion of the image is expanded in Figure 3B to show individual beads. Red and green beads are indicative of homozygous genotypes. Yellow indicates a heterozygous genotype, resulting from the presence of both Cy3 and Cy5 on the same bead (Fig. 2). In the next stage of data processing, a trimmed mean intensity is calculated for each bead type, for both Cy3 and Cy5 (on average, there are ~30 beads of each of the 1,520 bead types in an array). Figure 3C shows trimmed mean intensities for 96 DNA samples genotyped on one SNP. The SNP is one of 1,152 assayed in a multiplex pool. As shown, the 96 DNA samples cluster into three groups, showing that all three genotypes are represented in the sample set.

Actual genotype calls are made after transforming the intensity values into modified polar coordinates (Fig. 4). By taking into account the intensity distribution of beads, averaging, and rejecting outliers, measurement precision is improved (Fig. 4A vs. 4B and 4C vs. 4D). It is also shown in Figure 4 that occasional beads are outliers, and would, on their own, give inaccurate genotypes. Even though there are relatively few such beads, they could have a detrimental impact given the requirements for low error rates in large-scale genotyping studies. However, the redundancy in the system ensures a minimum of 5 beads of each type, greatly reducing the chance of an incorrect call.

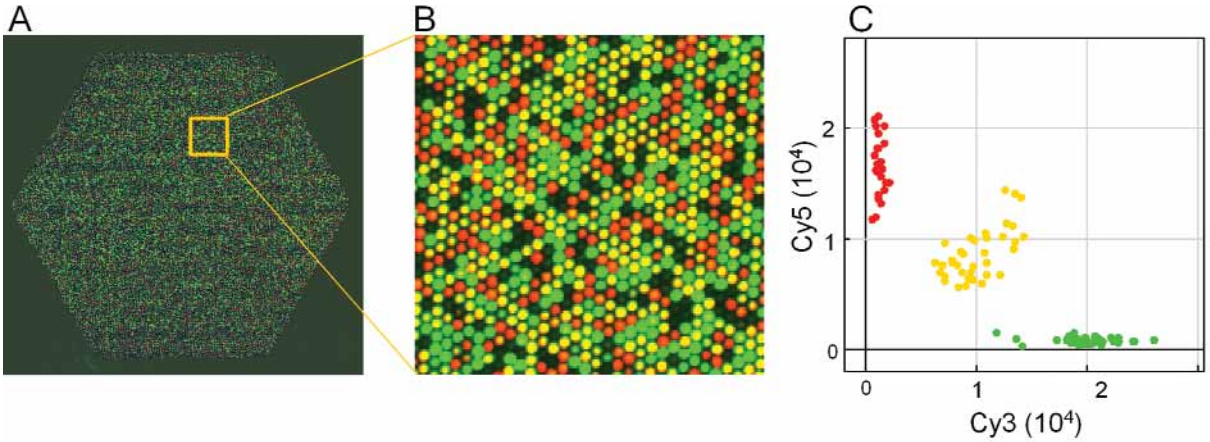

Figure 3. Views of genotyping data. (*A*) Fluorescence hybridization image of an ~1.4-mm-diameter optical fiber bundle containing 49,777 fibers in a monolithic, hexagonally packed array. (*B*) A portion of the hybridization image magnified to show individual beads. (*C*) Genotype calls for a single SNP on 96 DNA samples.

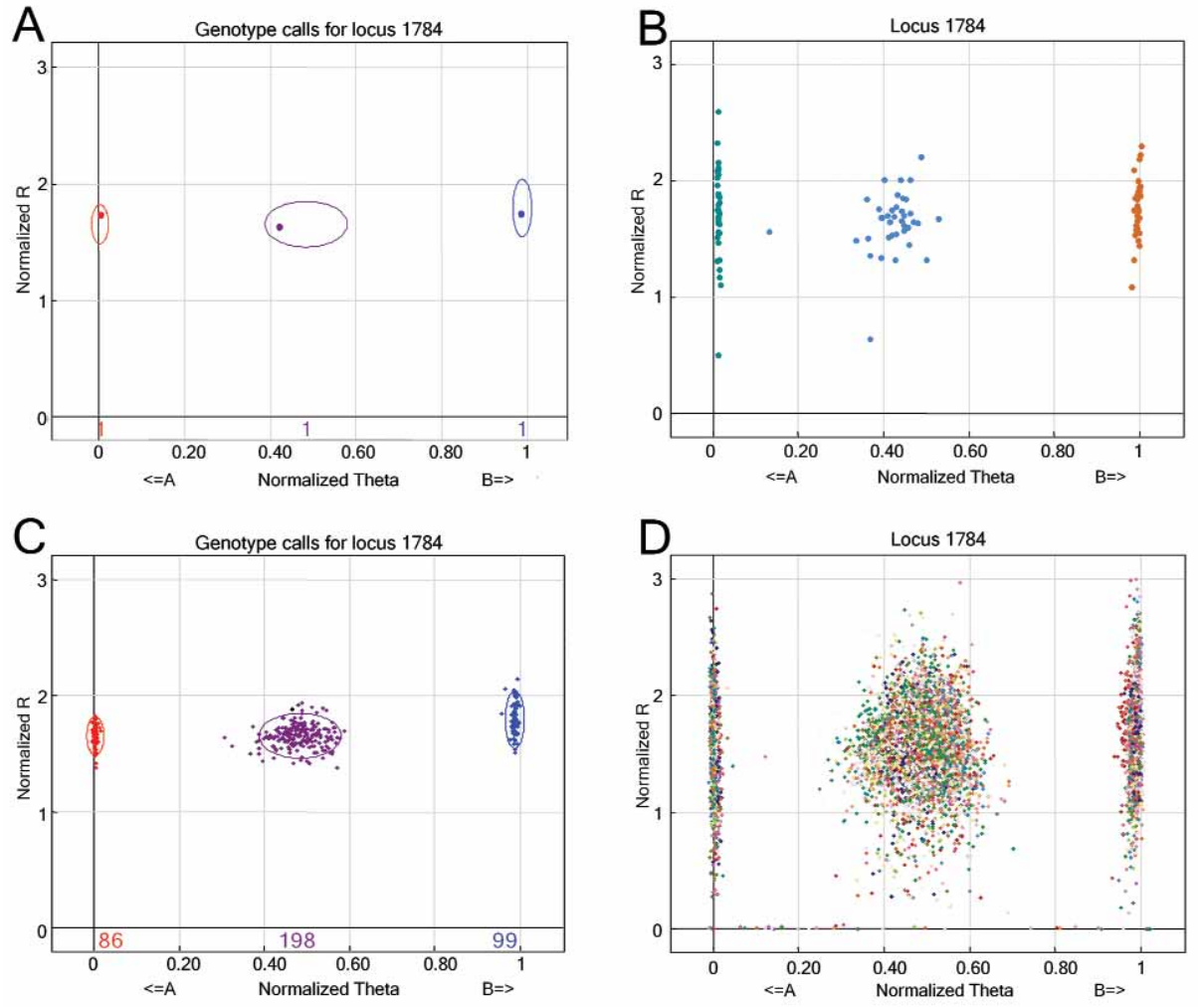

Figure 4. Genotyping data for one SNP in a 1,152 multiplex pool. Genotyping plots were created by graphing normalized Intensity (R = normalized X + normalized Y) vs. Theta = $^2/\pi\ Tan^{-1}(Cy5/Cy3)$. (*A*) Bead type data for three DNAs representing three different genotypes. Each data point represents a trimmed mean intensity derived from a population of beads. (*B*) Data from individual beads corresponding to the three data points shown in *A*. (*C*) Same as *A*, for 383 of 384 DNAs (one DNA sample failed to yield data for this SNP). The number of DNAs in each genotype cluster is shown above the *x* axis. (*D*) Same as *B*, showing individual bead data for all 383 DNAs.

ASSAY CONVERSION RATES

The fraction of SNPs that can be converted to working assays influences cost and the probability that any particular SNP can be genotyped. Currently, for most genetic studies, cost is of greater importance, and we maximize the assay conversion rate by using bioinformatic screens to rank SNPs according to likelihood of success. Sequences flanking the targeted SNP are evaluated on sequence composition and the presence of duplicated or more highly repetitive sequences, palindromes, and neighboring polymorphisms. The algorithm generates a quantitative score that reflects the likelihood of successfully developing an assay. Therefore, assay conversion rates depend on the quality of the set of input SNPs (e.g., some sets contain a higher fraction of sequencing errors and rare polymorphisms), whether or not we apply a bioinformatic screen, and the rigor of the screen. The quality of the oligonucleotides used in the assay is also an important factor. Given these variables, we have obtained assay conversion rates ranging from <50% to >97%. From a random sampling of dbSNP (http://www.ncbi.nlm.nih.gov/SNP/), attempting assays on both strands, we obtained an assay conversion rate of 81% (87% when excluding sequences that could not be assayed by other methods). This compares favorably to other technologies (Ohnishi et al. 2001; Gabriel et al. 2002; Hardenbol et al. 2003; Phillips et al. 2003).

Recently, many high-quality "double-hit" SNPs have been deposited in dbSNP (Reich et al. 2003; J. Mullikin, pers. comm.). These have each allele supported by two independent sequence reads, and therefore are more likely to be genuine SNPs and to have a relatively high minor allele frequency. To estimate the upper bound of our ability to convert candidate SNPs into assays, we selected 17,280 SNPs (corresponding to 15 x 1,152 multiplexes) that were predicted to have the highest likelihood of success from a collection of ~124,000 double hit SNPs (~14%). By choosing double-hit SNPs for this experiment, we effectively avoided the confounding variable of SNP quality (i.e., sequencing errors and rare SNPs). We developed assays for 10 x 1,152 SNPs on both strands, and assays for a further 5 x 1,152 SNPs on one strand only. This allowed us to compare success rates between the two levels of coverage. We obtained assay conversion rates of 96–99% when assaying both strands and using the best strand, and 94–96% when assaying a single strand (Table 1). Because these assays were carried out on the highest quality SNPs, we expect success rates to be more typically in the range of 90% when developing assays for double-hit SNPs on one strand, our standard approach.

Most of our assays have been developed at 1,152-plex, but we recently increased multiplexing levels to 1,536-plex. We have not yet determined the limits of multiplexing for this assay, but we have achieved excellent accuracy and call rates at multiplex levels of 1,152 to 1,536.

ACCURACY

There are several ways to estimate accuracy, including reproducibility, strand correlation, concordance with other genotyping methods, and consistency of Mendelian inheritance. Each measure has strengths and weaknesses (Oliphant et al. 2002). Here we report on an analysis of 5,704 SNPs from human chromosome 20. In a study of a 10-Mb region of Chromosome 20 (Contig NT_011362.7; 3,726,000 - 13,824,000 bp), 11,328 SNPs were selected for assay development. All assays were developed on both strands at a multiplex level of 1,152 and used to genotype 384 samples, including 100 unrelated African-Americans, 191 Caucasians (95 individuals from 12 three-generation Utah CEPH families and 96 UK Caucasians), 32 Japanese and 10 Chinese DNAs and controls,

Table 1. Assay Development and Genotyping Results for 17,280 SNPs

Bundle	Multiplex	Successful assays	Successful DNAs	Conversion rate (%)	Called genotypes	Possible genotypes	Call rate (%)
DS_1	1,152	1,136	95	99	107,882	107,920	99.96
DS_2	1,152	1,129	95	98	107,232	107,255	99.98
DS_3	1,152	1,132	95	98	107,515	107,540	99.98
DS_4	1,152	1,134	95	98	107,704	107,730	99.98
DS_5	1,152	1,121	95	97	106,464	106,495	99.97
DS_6	1,152	1,112	95	97	105,619	105,640	99.98
DS_7	1,152	1,112	95	97	105,575	105,640	99.94
DS_8	1,152	1,120	95	97	106,376	106,400	99.98
DS_9	1,152	1,114	95	97	105,794	105,830	99.97
DS_10	1,152	1,107	95	96	105,139	105,165	99.98
SS_1	1,152	1,101	95	96	104,540	104,595	99.95
SS_2	1,152	1,107	95	96	105,111	105,165	99.95
SS_3	1,152	1,097	95	95	104,189	104,215	99.96
SS_4	1,152	1,086	95	94	103,132	103,170	99.96
SS_5	1,152	1,084	95	94	102,947	102,980	99.97
Total	**17,280**	**16,692**	**95**	**97**	**1,585,219**	**1,585,740**	**99.97**

Each bundle corresponds to a SNP set assayed as a multiplex pool. All the assays are read out on a single array per sample (96 samples per 96-array matrix). Each sample plate contained 95 DNAs and a negative control. The first ten SNP sets, designated DS, were assayed on both strands. The remaining five SNP sets, designated SS, were assayed on one strand only. The DS and SS sets are sorted in the table by assay conversion rate.

including 32 duplicated DNAs. A total of 5,704 SNPs with a minor allele frequency of ≥4% in the combined set of DNA samples was selected for further analysis. Figure 4 shows genotyping data for one of the SNPs in this study. A linkage disequilibrium analysis of these data will be published elsewhere (P. Deloukas).

We used duplicate genotypes (from assays on both strands and DNA duplicates) and inheritance (CEPH family panel only) to identify discrepant genotypes. After removing 5 DNAs with poor results, the GenCall confidence score (provided with each genotype; see Analysis section below) was used as a threshold. The genotypes retained above the cutoff had a concordance rate of >99.7%. In addition, 566 of the 5,704 SNPs were also genotyped using the Homogeneous Mass Extend assay and MALDI-TOF mass spectrometry (CEPH panel only; P. Deloukas and colleagues, Wellcome Trust Sanger Institute). The concordance between the two different methods was 99.68%, based on a total of 27,901 genotypes.

We have also estimated accuracy from the sum of reproducibility and heritability errors for the data sets shown in Table 1. The accuracy of each of the 15 data sets shown in the table was ~99.9%. These results are consistent with a number of other studies (A. Oliphant, unpubl.) that have estimated the accuracy of genotypes after applying quality cutoffs to be in the range of 99.7–99.9%.

CALL RATE

There is a tradeoff between accuracy and call rate, which is also an important genotyping performance metric. We define the call rate as the fraction of genotype calls that are made as a fraction of possible calls, excluding unsuccessful assays. In the Chromosome 20 study described above, the overall accuracy of ~99.7% was achieved with an average call rate of 91.7%. We have since achieved even higher call rates while maintaining high accuracy. The more recent data set shown in Table 1 had a total of 1,585,219 genotypes called of a possible 1,585,740, for a call rate of 99.97%.

THE IMPORTANCE OF DNA QUALITY AND QUANTITY

We have analyzed the effects of key assay variables on data quality and found that gDNA concentration and purity are the most important variables in routine operation. Figure 5 shows the reproducibility of the GoldenGate assay as a function of the amount of input gDNA. A major advantage of a highly multiplexed assay is that relatively little DNA is consumed per genotype, assuming that many SNPs are assayed per sample. We routinely use 200 ng of gDNA for SNP assays multiplexed at 1,152-plex. At this level of multiplexing, DNA consumption is ~0.2 ng per genotype. Furthermore, we have shown that the assay works well with amplified gDNA, allowing large-scale genotyping from only ~ 10 ng of gDNA (D.L. Barker et al., in prep.).

ASSAY CONTROLS

Internal controls are used to monitor key steps in the procedure. These include gDNA/oligo annealing, PCR, array hybridization, and imaging. For example, assay specificity is checked by assaying nonpolymorphic sites in the genome with an ASO pair, of which one is a perfect match and the other a 3′ end mismatch. To illustrate, a site containing a G base might be assayed with an ASO containing a 3′C and a mismatch ASO containing a 3′T. The ratio of signals from the two ASOs is a measure of specificity. Similarly, imbalances in the amplification from P1 and P2 can be detected by assaying a nonpolymorphic site in the genome with two ASOs that are identical except that one incorporates P1 and the other P2. A

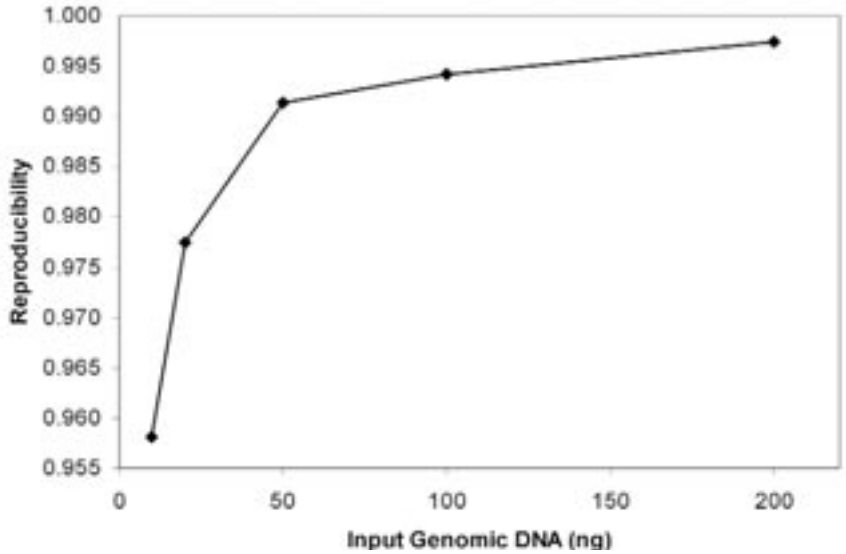

Figure 5. Relationship between the amount of input genomic DNA and genotyping reproducibility.

double-labeled control is used in array hybridization to check the optical balance of the Cy3 and Cy5 detection channels.

ANALYSIS OF LARGE GENOTYPING DATA SETS

To cope with the large amount of data generated, we developed automated methods to extract and analyze data, and derived quantitative measures of data quality. To call genotypes, we developed GenCall, a software program that interprets sample data using a model based on reference data. GenCall is used in conjunction with GenTrain software, which applies a custom clustering algorithm to a reference data set to obtain a set of locus-specific variables for each SNP. This information is provided as input to GenCall.

GenCall also calculates a quality score for each genotype called, which has been shown to correlate with the accuracy of the genotyping call (Fig. 6) (Oliphant et al. 2002). GenCall scores are in the range of 0 to 1, with 1 indicating the highest probability of the score being accurate. The score reflects the degree of separation between homozygote and heterozygote clusters and the placement of the individual call within a cluster, which can be considered key measures of signal-to-noise in the assay data. Besides being important for quality control of complex lab processes, the ability to evaluate objectively the quality of large data sets also enables process improvement to be carried out in a systematic way.

As shown in Figure 6, lower GenCall scores reflect less correlation between strands. Although the relationship is not linear, it is nevertheless useful in helping to establish a threshold for ensuring data quality. Currently, the relationship between GenCall score and accuracy can only be interpreted within a given study. Given the relatively large sizes of studies being undertaken, and the completeness of the data obtained, this is a minor limitation. Nevertheless, we are working to establish a more uniform relationship between GenCall score and accuracy.

ASSAY PANELS

Assay panels need to be readily available before any SNP genotyping system can be broadly useful for human genetic studies. We have developed a human linkage mapping panel, comprising ~4,600 highly informative SNPs, and a fine mapping panel comprising ~40,000 SNPs. We can also easily make custom panels on demand, with a high assay development success rate. The content for these panels can be chosen from the millions of available SNPs.

We are also involved in the International HapMap Project, which aims to create a haplotype map of the human genome and to make this information freely available in the public domain. This will enable genome-wide genetic association studies, potentially revolutionizing the search for the genetic basis of common diseases (http://hapmap.cshl.org). The SNPs are being genotyped in a set of samples representing African, Asian, and Caucasian populations. The data will be used to define haplotype patterns that are common in each population, and to identify a specific set of SNPs ("tag SNPs") that will be maximally informative for future genome-wide association studies. These genome-wide association studies will investigate the role of common variants in common diseases. Illumina is developing the haplotype map for Chromosomes 8q, 9, 18q, 22, and X, covering 15.5% of the genome, and the Wellcome Trust Sanger Institute is analyzing Chromosomes 1, 6, 10, 13, and 20, covering 24% of the genome. Other leading genome centers are also using the GoldenGate assay and the BeadLab system to develop HapMap assays for other regions of the genome, so far totaling an additional ~20%.

GENOTYPING RNA

Inherited variation in allelic mRNA abundance provides a means of studying the genetic basis of gene regulation and may enable associations to be made between genetic variation and disease (Yan et al. 2002; Lo et al. 2003). Studies of this type would benefit from the ability to assay efficiently large sets of SNPs occurring in coding regions (cSNPs). In preliminary studies of the GoldenGate assay for allele-specific quantitative mRNA profiling, 32 cSNPs were genotyped on matched pairs of DNA and RNA samples isolated from an ovarian tissue (Fig. 7). The RNA samples were first converted to cDNA, then genotyped using the same procedures used for gDNA. Of the 32 cSNPs, two scored as heterozygous in DNA, but

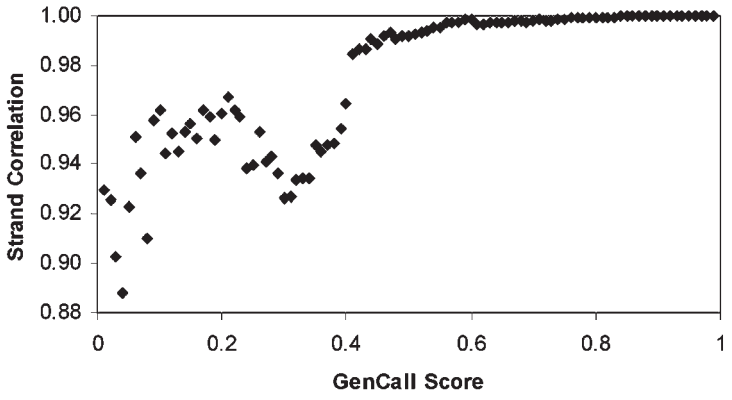

Figure 6. GenCall score predicts accuracy. The correlation between genotypes attempted on both strands of a SNP is a proxy for accuracy. A set of 2,916,654 genotype calls was analyzed. Only 99 data points representing the tail of the distribution are shown in this plot.

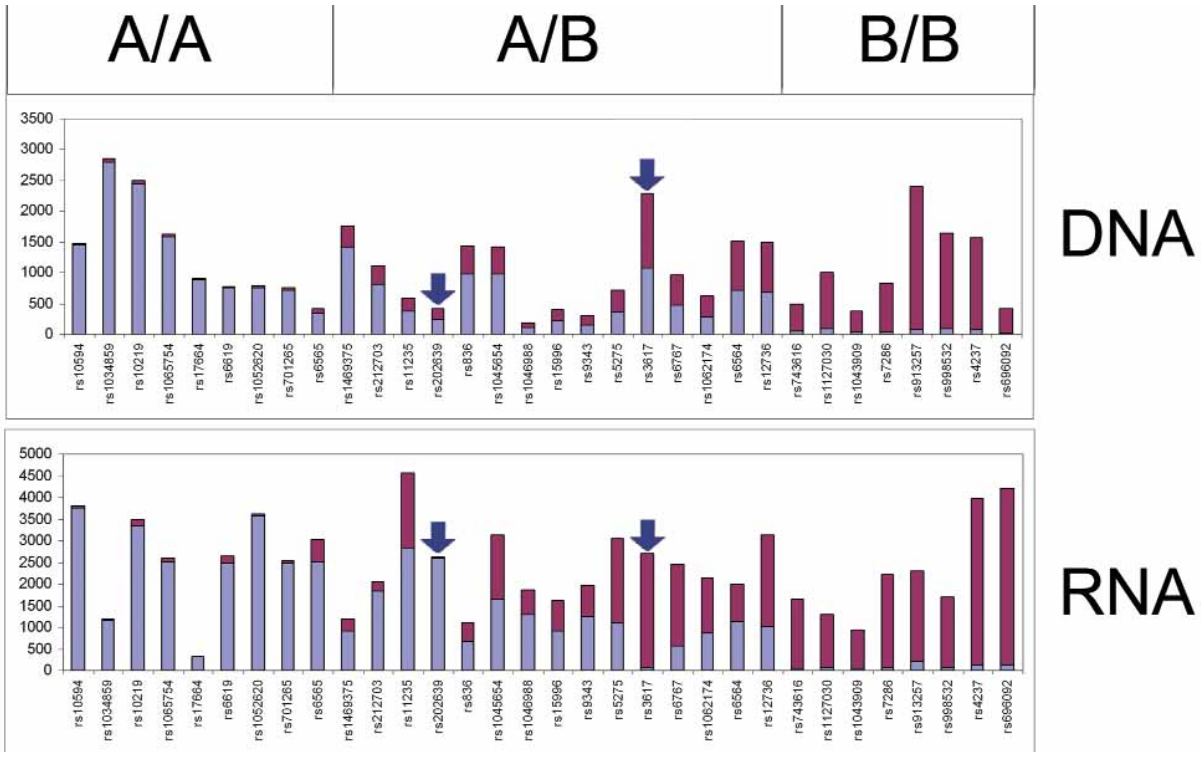

Figure 7. Allele-specific expression monitoring. Both genomic DNA and total RNA (converted to cDNA) from an ovarian tissue sample were genotyped for 32 SNPs in gene coding regions. Arrows indicate SNPs, called as heterozygous in gDNA, which show a clear allelic imbalance in mRNA abundance.

showed a strong allele-specific bias in transcript abundance. In addition, assays for 1,152 cSNPs from 450 human cancer-related genes were monitored in pairs of DNA/RNA samples isolated from different human tissues. Of the cSNPs (genes) that were expressed at detectable levels in the RNA and heterozygous in the corresponding genomic DNA, 20% showed differences in allelic abundance at 95% confidence; in many cases, only one allele from a heterozygous genomic locus was detected (J.-B. Fan, unpubl.). These preliminary studies show the feasibility of using the GoldenGate assay to genotype cDNA derived from mRNA populations.

CONCLUSIONS

We have developed a flexible, accurate, and scalable genotyping system, and have achieved high accuracy together with high call rates. Conventional wisdom was that single-plex assays would be more accurate than highly multiplexed assays, and that it would be difficult to optimize assays in a multiplex format. In fact, our accuracy and call rates are similar to those obtained from single-plex genotyping systems, but at over 1,000 times the assay sequence complexity. These results demonstrate the specificity of the GoldenGate assay format, as well as the reproducibility and accuracy of the BeadArray platform and the BeadLab genotyping system as a whole.

We previously published a ligation-based RNA assay for analysis of mRNA splice variants (Yeakley et al. 2002). In the course of this work, others have also published genotyping assays that are conceptually similar (Schouten et al. 2002; Hardenbol et al. 2003), and others have also achieved high multiplexing levels using direct hybridization-based methods, albeit without the flexibility to assay any SNPs of interest (Kennedy et al. 2003). The assay format used by Schouten et al. (2002), which is read out using gel electrophoresis, is used for the analysis of DNA copy number variation, including the detection of deletions in gDNA.

The GoldenGate assay is versatile and can be adapted to a variety of other applications, such as methylation profiling (J.-B. Fan, unpubl.). In addition, since we use a universal address scheme, different address sequences can be assigned to the same SNP locus to interrogate the same SNP in different samples. This strategy can be quite useful for studies that involve many samples but relatively few SNPs, as in some plant and animal breeding applications. We tested this scheme with an experimental design in which a set of 96 SNPs was associated with 10 discrete sets of 96 address sequences and used to genotype 10 samples in parallel, with readout on one array. We obtained exactly the same genotyping results using this pooling scheme as with our standard approach.

In conclusion, we believe that the GoldenGate assay format is an exemplar for a new class of highly multiplexed assays that utilize parallel readout systems. It represents a significant departure from single and low-multiplex assays and is well suited for large-scale analysis of complex biological systems. We expect that large-scale genotyping studies using this approach will help eluci-

date the genetic basis of complex diseases, and we are also optimistic that many new applications will spring from the general approach we have developed.

ACKNOWLEDGMENTS

The authors thank Steven Barnard, Diping Che, Todd Dickinson, Michael Graige, Robert Kain, Michal Lebl, and Chanfeng Zhao for their contributions to the development of the BeadArray and Oligator technologies that provided the foundation for this work. We also thank Steffen Oeser for assistance with graphics. The work described here was supported in part by grants R44 HG-02003, R43 CA-81952, and HG-002753 from the National Institutes of Health to M.S.C.

APPENDIX

GoldenGate™ Assay Procedures

Details may vary depending on the specifics of the genotyping system used. All robotic processes were performed on a Tecan Genesis Workstation 150 (Tecan). Up-to-date protocols are supplied with genotyping systems from Illumina, Inc.

Immobilization of genomic DNA to streptavidin-coated magnetic beads. Genomic DNA (20 μl at 100 ng/μl) was mixed with 5 μl of photobiotin (0.2 μg/μl, Vector Laboratories) and 15 μl of mineral oil, and incubated at 95ºC for 30 minutes. Trizma base (25 μl of 0.1 M) was added, followed by two extractions with 75 μl of Sec-butanol to remove unreacted photobiotin. The extracted gDNA (20 μl) was mixed with 34 μl of Paramagnetic Particle A Reagent (MPA; Illumina) and incubated at room temperature for 90 minutes. The immobilized gDNA was washed twice with DNA wash buffer (WD1) (Illumina) and resuspended at 10 ng/μl in WD1. In each subsequent reaction, 200 ng (10 μl) of DNA was used.

Annealing of assay oligonucleotides to genomic DNA. Annealing reagent (MA1; Illumina; 30 μl) and SNP-specific oligonucleotides (10 μl containing 25 nM of each oligonucleotide) were combined with immobilized DNA (10 μl) to a final volume of 50 μl. LSOs were synthesized with a 5′ phosphate to enable ligation. Annealing was carried out by ramping temperature from 70ºC to 30ºC over ~8 hours, then holding at 30ºC until the next processing step.

Assay oligo extension and ligation. After annealing, excess and mishybridized oligonucleotides were washed away, and 37 μl of master mix for extension (MME; Illumina) was added to the beads. Extension was carried out at room temperature for 15 minutes. After washing, 37 μl of master mix for ligation (MML; Illumina) was added to the extension products, and incubated for 20 minutes at 57ºC to allow the extended upstream oligo to ligate to the downstream oligo.

PCR amplification. After extension and ligation, the beads were washed with universal buffer 1 (UB1; Illumina), resuspended in 35 μl of elution buffer (IP1; Illumina) and heated at 95ºC for one minute to release the ligated products. The supernatant was then used in a 60-μl PCR. PCR reactions were thermocycled as follows: 10 seconds at 25ºC; 34 cycles of (35 seconds at 95ºC, 35 seconds at 56ºC, 2 minutes at 72ºC); 10 minutes at 72ºC; and cooled to 4ºC for 5 minutes. The three universal PCR primers (P1, P2, and P3) are labeled with Cy3, Cy5, and biotin, respectively.

PCR product preparation. Double-stranded PCR products were immobilized onto paramagnetic particles by adding 20 μl of Paramagnetic Particle B Reagent (MPB; Illumina) to each 60-μl PCR, and incubated at room temperature for a minimum of 60 minutes. The bound PCR products were washed with universal buffer 2 (UB2; Illumina), and denatured by adding 30 μl of 0.1 N NaOH. After 1 minute at room temperature, 25 μl of the released ssDNAs was neutralized with 25 μl of hybridization reagent (MH1; Illumina) and hybridized to arrays.

Array hybridization and imaging. Arrays were hydrated in UB2 for 3 minutes at room temperature, and then preconditioned in 0.1 N NaOH for 30 seconds. Arrays were returned to the UB2 reagent for at least 1 minute to neutralize the NaOH. The pretreated arrays were exposed to the labeled ssDNA samples described above. Hybridization was conducted under a temperature gradient program from 60ºC to 45ºC over ~12 hours. The hybridization was held at 45ºC until the array was processed. After hybridization, the arrays were first rinsed twice in UB2 and once with IS1 (IS1; Illumina) at room temperature with mild agitation, and then imaged at a resolution of 0.8 microns using a BeadArray Reader (Illumina; Barker et al. 2003). PMT settings were optimized for dynamic range, channel balance, and signal-to-noise ratio. Cy3 and Cy5 dyes were excited by lasers emitting at 532 nm and 635 nm, respectively.

Genotyping with RNA samples. A 20-μl reverse transcription reaction containing a reaction mix (MMC; Illumina) and total RNA (up to 1 μg), was incubated at room temperature for 10 minutes and then at 42ºC for 1 hour. After cDNA synthesis, the remainder of the assay was identical to the GoldenGate assay described above.

REFERENCES

Abravaya K., Carrino J.J., Muldoon S., and Lee H.H. 1995. Detection of point mutations with a modified ligase chain reaction (Gap-LCR). *Nucleic Acids Res.* **23:** 675.

Barker D.L., Therault G., Che D., Dickinson T., Shen R., and Kain R. 2003. Self-assembled random arrays: High-performance imaging and genomics applications on a high-density microarray platform. *Proc. SPIE* **4966:** 1.

Cai H., White P.S., Torney D., Deshpande A., Wang Z., Keller R.A., Marrone B., and Nolan J.P. 2000. Flow cytometry-based minisequencing: A new platform for high-throughput single-nucleotide polymorphism scoring. *Genomics* **66:** 135.

Chen J., Iannone M.A., Li M.S., Taylor J.D., Rivers P., Nelsen A.J., Slentz-Kesler K.A., Roses A., and Weiner M.P. 2000. A

microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension. *Genome Res.* **10:** 549.

Dawson E., Abecasis G.R., Bumpstead S., Chen Y., Hunt S., Beare D.M., Pabial J., Dibling T., Tinsley E., Kirby S., Carter D., Papaspyridonos M., Livingstone S., Ganske R., Lohmussaar E., Zernant J., Tonisson N., Remm M., Magi R., Puurand T., Vilo J., Kurg A., Rice K., Deloukas P., and Mott R., et al. 2002. A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418:** 544.

Fan J.B., Chen X., Halushka M.K., Berno A., Huang X., Ryder T., Lipshutz R.J., Lockhart D.J., and Chakravarti A. 2000. Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Res.* **10:** 853.

Gabriel S.B., Schaffner S.F., Nguyen H., Moore J.M., Roy J., Blumenstiel B., Higgins J., DeFelice M., Lochner A., Faggart M., Liu-Cordero S.N., Rotimi C., Adeyemo A., Cooper R., Ward R., Lander E.S., Daly M.J., and Altshuler D. 2002. The structure of haplotype blocks in the human genome. *Science* **296:** 2225.

Gerry N.P., Witowski N.E., Day J., Hammer R.P., Barany G., and Barany F. 1999. Universal DNA microarray method for multiplex detection of low abundance point mutations. *J. Mol. Biol.* **292:** 251.

Gunderson K.L., Kruglyak S., Graige M.S., Garcia F., Kermani B., Zhao C., Che D., Dickinson T., Wickham E., Bierle J., Doucet D., Milewski M., Yang R., Siegmund C., Haas J., Zhou L., Oliphant A., Fan J.-B., Barnard S., and Chee M.S. 2004. Decoding randomly ordered DNA arrays. *Genome Res.* (in press).

Hardenbol P., Baner J., Jain M., Nilsson M., Namsaraev E.A., Karlin-Neumann G.A., Fakhrai-Rad H., Ronaghi M., Willis T.D., Landegren U., and Davis R. W. 2003. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **21:** 673.

Iannone M.A., Taylor J.D., Chen J., Li M.S., Rivers P., Slentz-Kesler K.A., and Weiner M.P. 2000. Multiplexed single nucleotide polymorphism genotyping by oligonucleotide ligation and flow cytometry. *Cytometry* **39:** 131.

Kennedy G.C., Matsuzaki H., Dong S., Liu W.M., Huang J., Liu G., Su X., Cao M., Chen W., Zhang J., Liu W., Yang G., Di X., Ryder T., He Z., Surti U., Phillips M.S., Boyce-Jacino M.T., Fodor S.P., and Jones K.W. 2003. Large-scale genotyping of complex DNA. *Nat. Biotechnol.* **21:** 1233.

Kwok P.Y. 2001. Methods for genotyping single nucleotide polymorphisms. *Annu. Rev. Genomics Hum. Genet.* **2:** 235.

Lebl M., Burger C., Ellman B., Fambro S., Hachmann J., Heiner D., Ibrahim G., Jones A., Kim S., Nibbe M., Pires J., Santos C., Touhy S., Mudra P., Pokorny V., Poncar P., and Zenisek K. 2001. Fully automated parallel oligonucleotide synthesizer. *Collect. Czech. Chem. Commun.* **66:** 1299.

Lo H.S., Wang Z., Hu Y., Yang H.H., Gere S., Buetow K.H., and Lee M. P. 2003. Allelic variation in gene expression is common in the human genome. *Genome Res.* **13:** 1855.

Michael K.L., Taylor L.C., Schultz S.L., and Walt D.R. 1998. Randomly ordered addressable high-density optical sensor arrays. *Anal. Chem.* **70:** 1242.

Ohnishi Y., Tanaka T., Ozaki K., Yamada R., Suzuki H., and Nakamura Y. 2001. A high-throughput SNP typing system for genome-wide association studies. *J. Hum. Genet.* **46:** 471.

Oliphant A., Barker D.L., Stuelpnagel J.R., and Chee M.S. 2002. BeadArray technology: Enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* **32:** S56.

Pastinen T., Raitio M., Lindroos K., Tainola P., Peltonen L., and Syvanen A.C. 2000. A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. *Genome Res.* **10:** 1031.

Patil N., Berno A.J., Hinds D.A., Barrett W.A., Doshi J.M., Hacker C.R., Kautzer C.R., Lee D.H., Marjoribanks C., McDonough D.P., Nguyen B.T., Norris M.C., Sheehan J.B., Shen N., Stern D., Stokowski R.P., Thomas D.J., Trulson M.O., Vyas K.R., Frazer K.A., Fodor S.P., and Cox D.R. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294:** 1719.

Phillips M.S., Lawrence R., Sachidanandam R., Morris A.P., Balding D.J., Donaldson M.A., Studebaker J.F., Ankener W.M., Alfisi S.V., Kuo F.S., Camisa A.L., Pazorov V., Scott K.E., Carey B.J., Faith J., Katari G., Bhatti H.A., Cyr J.M., Derohannessian V., Elosua C., Forman A.M., Grecco N.M., Hock C.R., Kuebler J.M., and Lathrop J.A., et al. 2003. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat. Genet.* **33:** 382.

Reich D.E., Gabriel S.B. and Altshuler D. 2003. Quality and completeness of SNP databases. *Nat. Genet.* **33:** 457.

Sachidanandam R., Weissman D., Schmidt S.C., Kakol J.M., Stein L.D., Marth G., Sherry S., Mullikin J.C., Mortimore B.J., Willey D.L., Hunt S.E., Cole C.G., Coggill P.C., Rice C.M., Ning Z., Rogers J., Bentley D.R., Kwok P.Y., Mardis E.R., Yeh R.T., Schultz B., Cook L., Davenport R., Dante M., and Fulton L., et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928.

Schouten J.P., McElgunn C.J., Waaijer R., Zwijnenburg D., Diepvens F., and Pals G. 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.* **30:** e57.

Syvanen A.C. 2001. Accessing genetic variation: Genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.* **2:** 930.

Wang D.G., Fan J.-B., Siao C.-J., Berno A., Young P., Sapolsky R., Ghandour G., Perkins N., Winchester E., Spencer J., Kruglyak L., Stein L., Hsie L., Topaloglou T., Hubbell E., Robinson E., Mittmann M., Morris M.S., Shen N., Kilburn D., Rioux J., Nusbaum C., Rozen S., Hudson T.J., and Lipshutz R., et al. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280:** 1077.

Yan H., Yuan W., Velculescu V.E., Vogelstein B., and Kinzler K. W. 2002. Allelic variation in human gene expression. *Science* **297:** 1143.

Yeakley J.M., Fan J.B., Doucet D., Luo L., Wickham E., Ye Z., Chee M.S., and Fu X.D. 2002. Profiling alternative splicing on fiber-optic arrays. *Nat. Biotechnol.* **20:** 353.