Available online at www.sciencedirect.com
SCIENCE DIRECT®
Mutation Research 573 (2005) 70–82
Fundamental and Molecular Mechanisms of Mutagenesis
www.elsevier.com/locate/molmut
Community address: www.elsevier.com/locate/mutres

Review

High-throughput SNP genotyping on universal bead arrays

Richard Shen*, Jian-Bing Fan, Derek Campbell, Weihua Chang, Jing Chen, Dennis Doucet, Jo Yeakley, Marina Bibikova, Eliza Wickham Garcia, Celeste McBride, Frank Steemers, Francisco Garcia, Bahram G. Kermani, Kevin Gunderson, Arnold Oliphant

Illumina Inc., 9885 Towne Centre Drive, San Diego, CA 92121, USA

Received 7 July 2004; accepted 21 July 2004
Available online 11 February 2005

Abstract

We have developed a flexible, accurate and highly multiplexed SNP genotyping assay for high-throughput genetic analysis of large populations on a bead array platform. The novel genotyping system combines high assay conversion rate and data quality with >1500 multiplexing, and Array of Arrays™ formats. Genotyping assay oligos corresponding to specific SNP sequences are each linked to a unique sequence (address) that can hybridize to its complementary strand on universal arrays. The arrays are made of beads located in microwells of optical fiber bundles (Sentrix® Array Matrix) or silicon slides (Sentrix BeadChip). The optical fiber bundles are further organized into a matrix that matches a 96-well microtiter plate. The arrays on the silicon slides are multi-channel pipette compatible for loading multiple samples onto a single silicon slide. These formats allow many samples to be processed in parallel. This genotyping system enables investigators to generate approximately 300,000 genotypes per day with minimal equipment requirements and greater than 1.6 million genotypes per day in a robotics-assisted process. With a streamlined and comprehensive assay, this system brings a new level of flexibility, throughput, and affordability to genetic research.

Keywords: SNP; Genotyping; Microarray; BeadArray

Contents

* Corresponding author. Tel.: +1 858 202 4633; fax: +1 858 202 4680.
E-mail address: rshen@illumina.com (R. Shen).

0027-5107/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.mrfmmm.2004.07.022

1. Introduction

With the complete sequencing of the human genome [1,2], large numbers of DNA sequence variants, mainly single nucleotide polymorphisms (SNPs) are revealed. As of June 2, 2004, 9,856,125 SNPs have been found in the human genome and deposited to public databases (NCBI dbSNP Build 121). The analysis of SNPs in the human genome may offer the key to understand genetic differences between individuals and disease states, and eventually improve medical treatments by allowing the prediction of genetically related disease risk and drug response. To meet these goals, major international collaborative efforts have been made to carry out large scale genomic studies that require the determination of hundreds of thousands of genotypes performed in many individuals [3,4].

Large-scale genetic epidemiological studies can be done with a genome-wide approach or with a candidate gene approach [5,6]. The scale of SNP genotyping needed for such studies is several orders of magnitude greater than what has been required for conventional family-based linkage mapping. Realizing the vision of comprehensive genome-wide genetic association studies will require a SNP genotyping system that combines very high throughput and accuracy with very low cost per SNP analysis, i.e., to be able to economically genotype large number of SNPs in large sample sets [7,8].

The BeadArray technology described in this study provides a solution to these problems by combining a miniaturized array platform with a high level of assay multiplexing and scalable automation. The system uses a high-density BeadArray technology in combination with an allele-specific extension, adapter ligation and amplification assay protocol that achieves high multiplexing in a fully integrated production environment. In turn, the high level of multiplexing simultaneously lowers the genotyping cost while increasing the throughput through judicious use of materials and automation.

The multiplexed assay detects up to 1536 SNPs in a single DNA sample. Genotyping 96 samples at once, using a single Sentrix Array Matrix, allows an individual researcher to determine up to 150,000 genotype calls simultaneously. Genotyping throughput may be increased proportionately by processing more Sentrix Array Matrices.

2. BeadArray technology and Sentrix Array formats

At the heart of Illumina genotyping products lies a fundamentally different way of building arrays: the random self-assembly of beads into patterned microwell substrates [9,10]. Illumina has used technological advances in both the fiber-optics and microelectronics industries to build substrates containing tens of thousands to many millions of wells across their surfaces. Quantitatively pooled bead libraries are then self-assembled into the etched microwell substrates, resulting in the highest density array platform currently available. To meet the broad range of researchers' needs, Illumina has developed two different Array of Arrays formats, the Sentrix Array Matrix and the Sentrix BeadChip.

The Array Matrix uses fiber-optic bundles containing nearly 50,000 individual light-conducting strands chemically etched to create a 3-μm well at the end of each strand [11]. Bundles are grouped together into a 96-array configuration matching the well spacing of standard microtiter plates (Fig. 1a). This unique format allows users to simultaneously conduct experiments on 96 arrays simultaneously. The diameter of the bundles

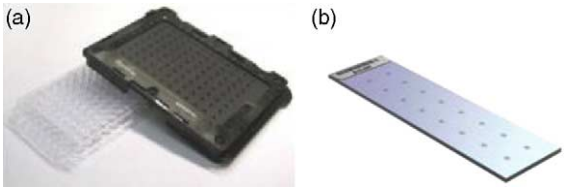

Fig. 1. Two different array of arrays formats: the Sentrix Array Matrix and the Sentrix BeadChip.

is small enough to accommodate a 384-well or 1536-well spacing. Moreover, the platform can be readily incorporated into automated routines using standard robotic equipment, leading to reduced error and human labor requirements.

For users with more moderate throughput requirements, Illumina has introduced the novel BeadChip format (Fig. 1b). This slide-sized platform allows processing of 16 samples at a time, with the same feature-to-feature spacing as standard multi-channel pipettes.

On the Sentrix Array Matrix bundle, up to 1624 unique bead types containing different probe sequences are represented in each array, with an average 30-fold redundancy of each bead type. Independent of the array format, each bead in every array contains hundreds of thousands of covalently attached oligonucleotide probes. After bead assembly into the wells of the array a hybridization-based procedure is used to decode the array, determining which bead type resides in each well. The decoding process is described in detail in a separate publication [12]. This final process validates the performance of each bead type and provides a level of quality control unmatched in the microarray industry. A set of universal bead types is used for the Illumina GoldenGate^{TM} genotyping assay [13]. Each bead type has a unique oligonucleotide sequence, corresponding to the complementary DNA sequences (address sequences) designed into the genotyping assay oligonucleotides. These address sequences hybridize to the universal bead type probes in the last hybridization step of the genotyping assay.

3. GoldenGate genotyping assay

The GoldenGate genotyping assay protocol, illustrated in Fig. 2, allows for a high degree of loci multiplexing in a single reaction through highly specific extension and amplification steps. One of the most important features of the assay is that it genotypes directly on the genomic DNA and does not require prior PCR amplification of the genotyping target.

The DNA sample used in this assay is activated for binding to paramagnetic particles. This activation step is a highly robust process and requires a minimum input of DNA (250 ng at 50 ng/μl). Depending upon the multiplex level, this equates to as low as 0.16 ng of DNA per SNP genotype call. Assay oligonucleotides, hybridization buffer, and paramagnetic particles are then combined with the activated DNA in the assay hybridization step. Three assay oligonucleotides are designed for each SNP locus. Two oligos are specific to each allele of the SNP site, called the allele-specific oligos (ASOs). A third oligo that hybridizes between 1 and 20 bases downstream from the ASO site is the locus-specific oligo (LSO). The 1–20 bp spacing between SNP and LSO allows probe design flexibility to avoid "bad" sequences flanking the SNP or neighboring SNPs. All three oligonucleotide sequences contain regions of genomic complementarity and universal PCR primer sites; the LSO also contains a unique address sequence complementary to a particular bead type. Up to 1536 SNPs may be interrogated simultaneously in this manner. During the hybridization process, the assay oligonucleotides hybridize to the genomic DNA sample bound to paramagnetic particles. Because hybridization occurs prior to any amplification steps, no amplification bias can be introduced into the assay.

Following hybridization, several wash steps are performed, significantly reducing the noise level in the genotyping assay, as properly hybridized assay oligonucleotides are retained and mishybridized and excess assay oligonucleotides are washed away. After oligo hybridization, a polymerase with high specificity for 3′ match is added and only extends the ASO(s) that perfectly match the target sequence at the SNP sites. As the polymerase used has no strand displacement or exonuclease activity, the polymerase fills the gap between the ASO and LSO. When it reaches the LSO the polymerase simply drops off the genomic DNA. High locus specificity is achieved by the requirement that both the ASO and LSO oligos need to hybridize to the same target site. A DNA ligase seals the nick between the extended sequence of the ASO and the LSO to form PCR templates that can be amplified with universal PCR primers. Typically, over 1536 loci on

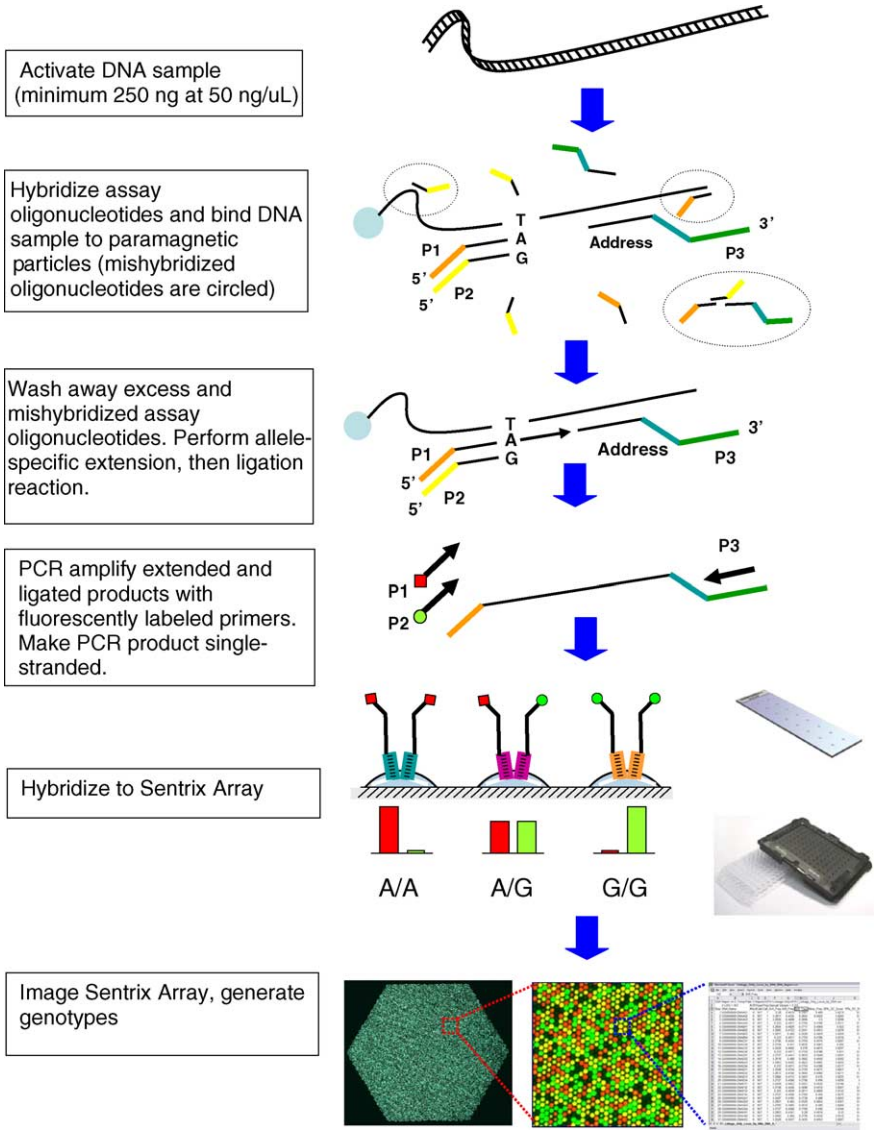

Fig. 2. Illustration of the GoldenGate genotyping assay process.

the genomic DNA are being interrogated simultaneously, and as such a population of over 4600 distinct oligos are present in the reaction tube. Extension of the appropriate ASO and ligation of the extended product to the LSO joins information about the genotype present at the SNP site to the address sequence on LSO. These joined, full-length products provide a template for PCR using only three universal PCR primers P1, P2, and P3.

Universal PCR primers P1 and P2 are labeled with Cy3 and Cy5 dyes, respectively. After thermal cycling and downstream-processing the single-stranded, dye-labeled DNA products (the GoldenGate assay products) are hybridized to their complement bead type through their unique address sequences. Hybridization of the GoldenGate assay products onto the Sentrix Arrays allows for the readout of the highly multiplexed SNP genotyping assay. After the hybridization, the

BeadArray Reader [11] is used to analyze fluorescence signal on the Array Matrix or BeadChip.

Illumina's BeadArray Reader simultaneously scans Sentrix Arrays at two different wavelengths with submicron resolution. A scan of 96 hybridized samples in a Sentrix Array Matrix, representing data acquisition from over 4.5 million discrete beads, takes approximately 90 min. This represents the highest throughput at the highest resolution of any scanning platform used for microarray-based genetic analysis applications.

The BeadArray Reader offers broad dynamic range, high sensitivity, and a low limit of detection. The BeadArray Reader automatically scans an array and uses information from a bead map file (unique to each array) to extract intensity information at 550 and 630 nm wavelengths for each bead type.

The GoldenGate genotyping platform offers flexibility and scalability. The ability to use either the Sentrix BeadChip or the Sentrix Array Matrix allows the user to perform as few as 6000 genotype calls (384 loci multiplex on a 16-sample BeadChip), to as many as 300,000 genotype calls (1536 loci multiplex on two 96-sample Array Matrices) in a single day with minimal equipment requirements. Automating the GoldenGate assay with a few liquid handling robots and a laboratory information management system allows throughput of over 1.6 million genotype calls per day.

4. Oligonucleotide design and synthesis

Quality oligonucleotides are necessary to meet the performance specifications of the GoldenGate assay. Illumina's Oligator® DNA synthesis technology enables parallel synthesis of many plates of oligonucleotides, providing the throughput and quality necessary to support our genotyping products.

Proper design of the assay oligonucleotides is one factor that significantly contributes to the success of the genotyping assay. Our proprietary OligoDesigner software is optimized for the design of GoldenGate assay probes. It evaluates sequences flanking the targeted SNP such as repeated sequences across the genome; palindromic sequences; GC and AT content; neighboring polymorphisms. The allele-specific sequence is designed at T_m of 60 °C (57–62 °C), while the locus-specific probe is designed at T_m of 57 °C (54–60 °C).

Illumina has experience with the design of many hundreds of thousands of GoldenGate genotyping assays and have incorporated this experience into the OligoDesigner software. The customer supplies SNPs along with flanking DNA sequences to Illumina, then these SNPs are entered into the OligoDesigner software, and a file containing the oligo designs and scoring information about the likelihood of success for that design is returned to the user. The user then has the option of selecting which SNP designs to order or can repeat the process with additional SNPs.

5. Contamination control safeguards

Two aspects of contamination control are designed into the genotyping assay: the first detects contamination; the other assists in removal of the contamination source should it occur. For contamination detection, one of four PCR contamination detection controls is added to each tube of assay oligonucleotide pools. When a single assay oligo pool is run, it is expected that only a single contamination control type should have high signal. Should two or more contamination control types have high signal, then significant contamination may have occurred.

The genotyping assay uses dUTP instead of dTTP in the PCR amplification, so that PCR carry-over contamination can be rendered inactive by the use of uracil DNA glycosylase (UDG) [14]. UDG cleaves the uracil base from the phosphodiester backbone of uracil-containing DNA, but has no effect on thymine-containing DNA. True PCR templates generated through the genotyping assay will contain only thymine and no uracil nucleotides.

6. Automatic genotype scoring

The automatic calling of genotypes is performed by genotype calling software (GenCall) genotyping software, using a Bayesian model. GenCall also assigns a confidence score, bound between 0 and 1 to each genotype call. In order to make the calls and assign the corresponding scores, GenCall software creates a locus-specific variables file (LSV). These variables are extracted by a proprietary custom-designed clustering algorithm on a population of DNA's. The advantage of

the custom-designed clustering algorithm over the classical clustering algorithms, e.g., hierarchical clustering or *K*-means, is the incorporation of genotyping-specific heuristics into the energy function of the optimization. This expected improvement is on the basis of Bayesian modeling, which states that when domain-specific knowledge (a.k.a., prior information) are available, one could enhance the outcomes by incorporating them into the model. In other words, the integration of likelihood (i.e., the fit of the data to the model) and prior probability (the domain knowledge) to form posterior probability contains more information than just the likelihood.

The expected clusters are A/A, A/B, and B/B. Often times, due to small minor-allele frequencies, one homozygote cluster or one homozygote and the heterozygote clusters are not present. In such cases, the clustering algorithm in GenCall computes the location and the form of the missing clusters using an artificial neural network.

A quality score, the GenCall score, is calculated for each SNP call, reflecting the degree of separation between homozygote and heterozygote clusters for that SNP and the placement of the individual call within a cluster. To make a genotype call, the software looks at many factors but one of the first is the distribution of beads of the same type and in this way outliers are rejected to ensure genotyping accuracy. The GenCall score is composed of various sub-scores, of which the most important one is the clustering score. This score is a locus-specific score, and is computed by a fuzzy

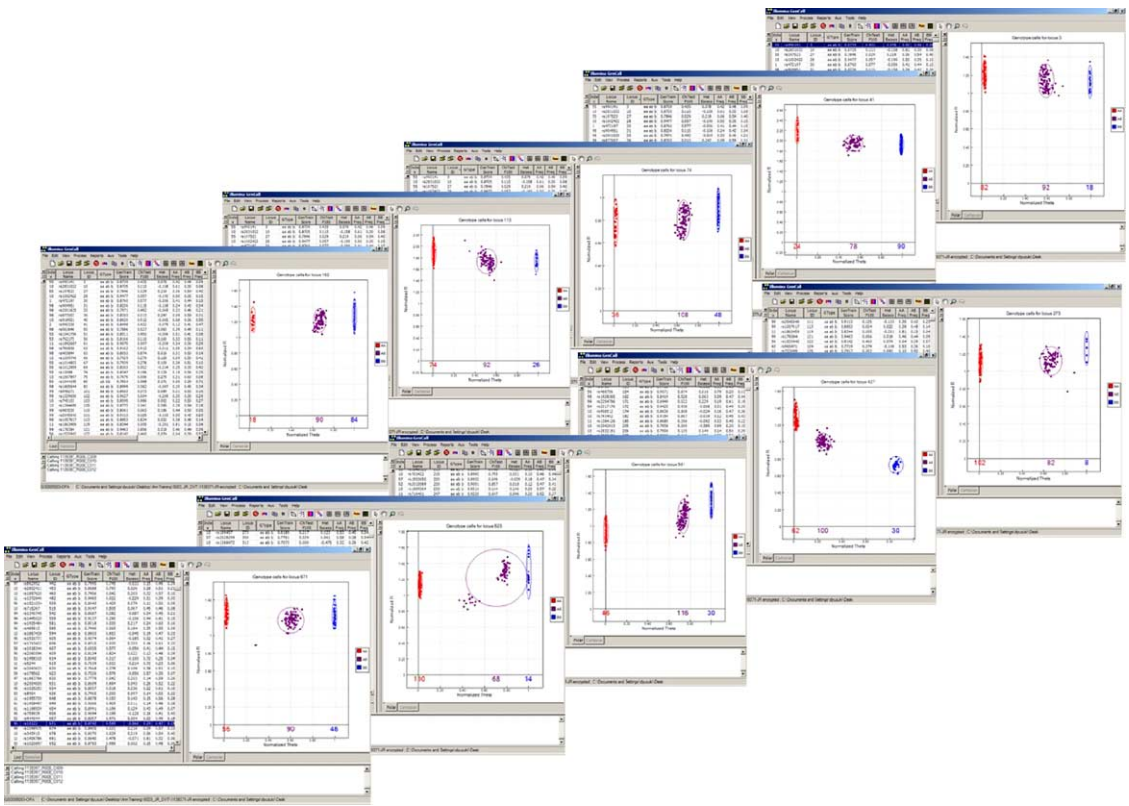

Fig. 3. GenCall software-produced plots of 10 randomly selected loci of 192 samples in polar coordinate representation. Ninety-six DNA samples were run in duplicate (a total of 192 DNAs) across 1263 loci. Each image is a graph of a single locus with 192 “dots” representing individual DNA samples. The *y*-axis is normalized intensity (sum of intensities of the two channels) and the *x*-axis is the “theta” value. Theta values near 0 (left side of graph) are homozygotes for allele “A” and theta values near 1 (right side of graph) are homozygotes for allele “B”. The GenCall software has also automatically grouped the 192 DNAs for each locus into the two homozygote clusters (red and blue) and the heterozygote cluster (purple).

logic inference system. It varies from 0.0 to 1.0, and correlates with accuracy of the genotype call. GenCall scores have been shown to correlate with the accuracy of the genotyping call [13,15].

Figs. 3 and 4 are screen capture images from our genotype calling software. They are unbiased representations of the quality of the genotype clusters generated from the Illumina genotyping assay on a linkage set oligonucleotide pool (GS0005003-OPA). Separation between homozygote clusters is excellent and variation within a cluster is small. In fact, because there is such high precision within a cluster, linked polymorphisms are also captured by this assay. The second graph from the left on the bottom row shows potentially five clusters. Individuals falling into the lower purple cluster (“heterozygote”) and the lower blue cluster (“homozygote”) have a linked polymorphism under one of the ASOs. The allele frequency of the linked polymorphism is sufficiently low that no individuals in this panel of 96 individuals are homozygote for the linked polymorphism (hence no lower red cluster). This linked polymorphism has destabilized one of the ASOs and lowered the resultant assay intensity. This “extra” cluster phenomenon is analogous to the split peaks sometimes seen in STR analysis when a single base change in one of the repeats alters the normal migration pattern.

7. Genotyping performance

Various internal assay controls are used to assess the GoldenGate assay and array hybridization at various experimental steps, including gDNA/oligo annealing, PCR, array hybridisation, and imaging. For example, G/T or G/G mismatches; amplification balance of the

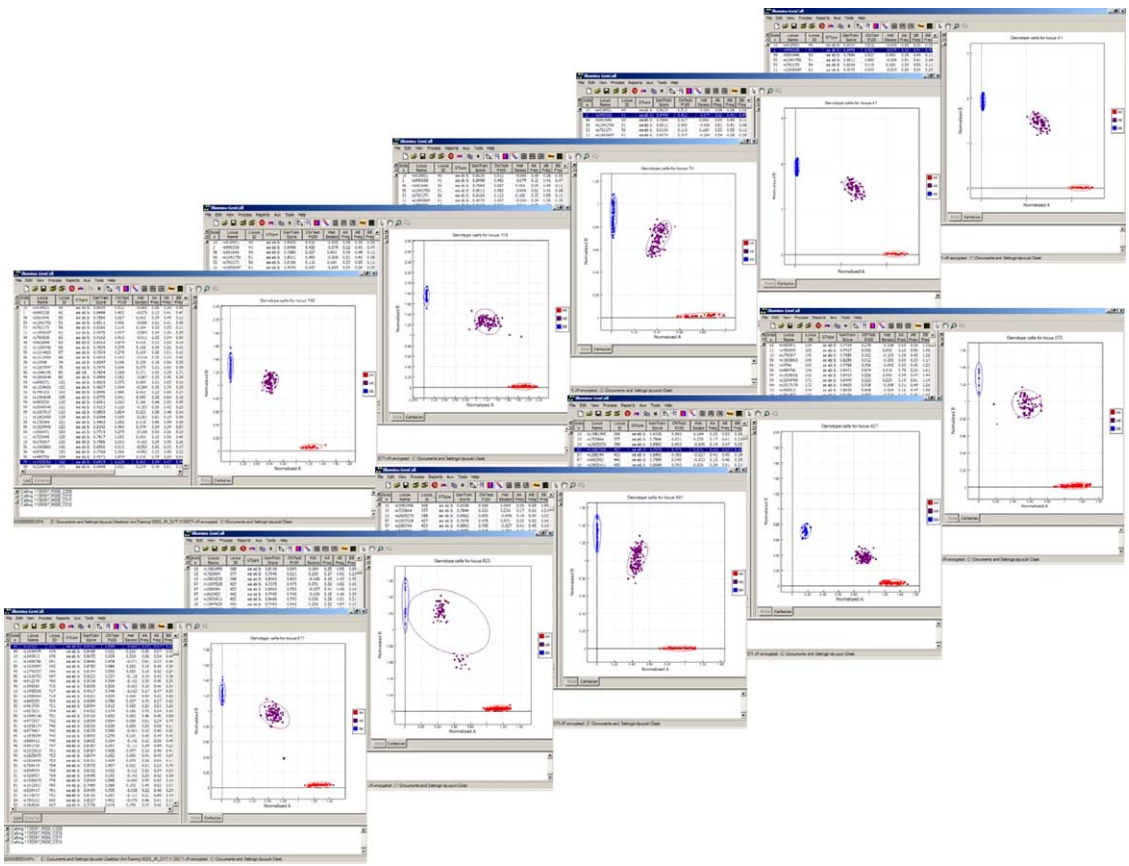

Fig. 4. GenCall software-produced plots of 10 randomly selected loci of 192 samples in Cartesian coordinate representation.

Table 1
Genotyping performance

	Reproducibility (%)	Heritability (%)	Call rate (%)	Assay development success rate (%)
Oligo pool 1: 1536 loci multiplex				
96 DNAs Run 1	100	100	99.96	93.68
96 DNAs Run 2	100	100	99.97	93.95
Oligo pool 2: 640 loci multiplex				
96 DNAs Run 1	100	99.99	100	93.91
96 DNAs Run 2	99.95	99.99	100	93.28

two types of universal PCR primers incorporated into the allele-specific oligos; hybridization controls; double label control to estimate optical balance of Cy3/Cy5 channels.

To estimate the impact of multiplexing level on genotyping performance, two oligo pools at two different multiplex levels (1536 and 640 loci) were run through the genotyping assay. These runs were performed in duplicate on 95 DNAs and a water control (not included in this analysis). The reproducibility was calculated from five duplicate samples among the 95 DNAs for all successfully developed assays in these two pools. The heritability was calculated from 35 trios among the 95 DNAs and is a measure determining whether the genotype calls follow Mendelian inheritance. The call rate is the fraction of genotypes called from among the total possible genotype calls for successfully genotyped DNAs. The assay development success rate is the fraction of assays actually functional from the total number of loci designed into the oligo pool.

As can be seen in Table 1, the reproducibility, heritability, and call rate are near 100%. The assay development success rate is at the low 90% level for these two particular oligo pools. Individual assay development success rate may vary and is highly dependent upon the SNPs selected and the informatics tools used to design the assay oligonucleotides. An international consortium of leading investigators (including Illumina scientists) has been formed to determine a haplotype map of the human genome using single nucleotide polymorphisms (HapMap project) [3]. The goal of the human HapMap project is to identify "tag" SNPs that may be used for linkage disequilibrium studies of complex genetic diseases. Illumina's technology platform is being used to genotype over 60% of the SNP loci in the HapMap project. To date, 200,000 SNP assays have been developed using the BeadArray technology.

The GoldenGate genotyping assay has been in use in our service operations group for 2 years, and has proven to be a robust process that generates high quality data. Table 2 shows the assay success rate, genotype call rate, and DNA sample success rate from our service operations. A total of 28,850 DNA samples were run from various customers; 96.46% of them were successfully genotyped and high quality data returned to the customer. The genotyping assay is robust to variation in DNA quantity and quality. However, the genotyping assay is significantly compromised when less than 20% of the required input DNA is provided; a primary reason for the failure to successfully generate genotype calls was insufficient DNA quantity.

A total of 293,505 SNP loci were designed and oligonucleotides synthesized. 80.2% of the attempted assays were successful. The assay success rate is highly dependent upon the SNP loci selected. In many projects, the list of SNPs was fixed and assays had to be designed and synthesized, even though SNP loci may have scored poorly on our OligoDesigner software. Individual experience with assay success rate may vary

Table 2
Assay success rate, genotype call rate, and DNA sample success rate

	Total attempted	Total successful	Assays successfully developed (%)
Number of DNA samples	28850	27828	96.46
Number of SNP loci	293505	235391	80.20
Genotypes produced (DNA samples × SNP loci)	54111293	53969348	99.74

depending upon the choice of SNP loci. The genotypes produced or genotype call rate is a measure of the completeness of the generated data set. It is the number of delivered genotypes on successfully genotyped DNAs and successfully developed SNP assays. Illumina's highly multiplexed genotyping assay delivers a high call rate of 99.74%. In a blind study, conducted by independent third party, 26,850 genotypes were compared with Mass Spec data, and a 99.7% concordance was observed [13].

To measure the accuracy of the genotyping system in allele frequency estimation, a series of linear dilutions were performed between DNA samples of two individuals. The first concentration contained 100% Individual 1 (and 0% Individual 2), and the eighth concentration contained 0% Individual 1 (and 100% Individual 2). For each bead type (i.e., one SNP locus), Cy3 and Cy5 intensities (representing A and B alleles, respectively) were collected and extracted, using Illumina's imaging and image analysis system. The ratios of Cy3/(Cy3 + Cy5) and Cy5/(Cy3 + Cy5) were used as proxies for A-allele and B-allele frequencies, respectively. For each concentration, 12 repeated experiments were performed; each experiment was performed on a single fiber-bundle at 1170 multiplex. Of the total 1170 loci typed, 501 loci were found to have different genotypes between the two individuals. This subset of loci was used to estimate the accuracy of allele-frequency computations. For each locus, at each concentration, the sufficient statistics were computed. Based on these statistics, it was determined that on average, with 95% confidence, our technology can detect 8% difference in allele frequencies with a single experiment. This number can be significantly improved, if one further selects a subset of high-performing loci or measures the samples multiple times.

8. Adaptation of the methylation profiling assay to the SNP genotyping platform

Illumina is adapting the SNP genotyping system to high throughput DNA methylation detection, based on "SNP" genotyping of bisulfite-converted genomic DNA. In this assay, non-methylated cytosines (C) are converted to uracil (U) when treated with bisulfite, while methylated cytosines remain unchanged. Hybridization behavior of uracil is similar to that of thymine (T). The detection of the methylation status of a particular cytosine can thus be carried out using a genotyping assay for a C/T polymorphism. Methylation status of the interrogated locus can be determined by calculation of the ratio of the fluorescent signals from the "C" (methylated) and "T" (unmethylated) alleles. In a previous study [unpublished data], we successfully demonstrated the feasibility of DNA methylation profiling on fiber optic arrays and developed a controlled system for monitoring assay performance. We have established a standard process for probe design and data analysis, a standard bisulfite conversion protocol, and a set of internal controls and reference samples for assay development and calibration. Current assay sensitivity and specificity is shown to be sufficient to detect changes in methylation status at more than 100 different sites simultaneously, in 1 μg of human genomic DNA. A minimum of three levels of methylation can be distinguished at any single site: fully methylated, semi-methylated, and unmethylated. We are performing experiments to measure >1000 CpG sites simultaneously.

9. Adaptation of the gene expression profiling assay to the SNP genotyping platform

Illumina has developed a flexible, sensitive, accurate, and cost-effective gene expression profiling assay, dubbed the DASL™ assay (for cDNA-mediated annealing, selection, extension, and ligation) based on the same universal address array used for SNP genotyping [16]. In this assay, three oligos are designed to target a specific mRNA sequence in a way similar to the SNP assay design. If a sample contains a cDNA target (the RNA is converted to cDNA by random priming), the corresponding query oligos will bind to the cDNA, and become extended and ligated enzymatically. The ligated products are then amplified and fluorescently labeled during PCR, and finally detected by binding to address sequences on the array. The hybridization intensity is proportional to the original mRNA abundance in the sample and used to represent the expression level of the targeted transcripts.

The DASL assay multiplexes to 1536 sequence targets, i.e., 512 genes at three probes per gene. DASL is characterized by a dynamic range of 2.5–3 logs, using 100 ng total RNA, and a 1.3-fold difference

measurement precision. This results in a mid-density expression profiling system with high sample throughput, high sensitivity and fills a gap between two existing RNA profiling technologies: quantitative RT-PCR (high sample throughput/low gene content) and high-density oligo or cDNA microarray technology (high gene content/low sample throughput).

Due to its high sequence specificity, the DASL assay can be used for splice variant detection [17] as well as differential allele-specific expression monitoring (i.e., quantitative measuring the abundance of each allele of target gene transcripts) [13]. We have developed genotyping assays for 1152 cSNPs derived from 380 cancer related genes. Each assay target is chosen from the sense strand and within one exon, thus, they can be used to interrogate both genomic DNA and RNA. With this assay design, we are able to obtain not only genotype information at the genomic DNA level, but potentially differential allele-specific expression information as well. It is known that levels of gene transcripts originating from paternal and maternal chromosomes may differ, thus constituting the basis of differential allele expression. Inheritance of allelic expression levels may provide an important link between individual genetic variation and the origin of disease. In addition, comparison of allele expression profiles may facilitate the identification of dominant susceptibility alleles in case/control studies where the frequency of heterozygotes is higher in cases versus controls [18,19].

10. Discussion

One key aspect of the GoldenGate genotyping assay design is the incorporation of an address sequence, so that the assay products can be read out on a universal array [20–23]. The probes on the array are random, artificial sequences that are not SNP-specific. Any set of SNPs can be analyzed simply by building the address sequences into the SNP-specific assay oligonucleotides. The approach offers substantial flexibility. The universal probe set represented on the array comprises 1624 different sequences selected to not cross hybridize with each other or with sequences in the human genome. The use of a universal array greatly simplifies the manufacturing process and reduces costs.

Since we use a universal address sequence tagging approach, different address sequences can be assigned to the same SNP locus and used to interrogate the same SNP in different samples (pooled for hybridization). In this way, a single array could be used to analyze the same loci from different individuals. This pooling scheme can be quite useful for studies, which require genotyping large number of samples with relatively small SNPs (e.g., <100 SNPs). We have tested this scheme with an experimental design in which a set of 96 SNPs were 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 compared to that obtained with standard single-sample-single-array approach.

The GoldenGate assay can tolerate certain degree of DNA degradation. We have completed highly successful studies using DNA samples isolated from formalin-fixed, paraffin-embedded tissues (unpublished data) and whole genome amplified samples [24]. Formalin-fixed archival tissues represent an invaluable resource for genetic analysis, as they are the most widely available materials for studies of human diseases. The ability to perform genetic analysis in archived tissues, for which clinical follow-up is already available, will greatly facilitate research in correlating genetic profiles with clinical parameters, and eventually in developing biomarkers for therapeutic decision making.

Most recently, Illumina has developed an array-based whole genome genotyping (WGG) assay approach, which comprises a novel combination of well proven technologies (Gunderson et al., in preparation). It utilizes direct hybridization capture of processed genomic DNA to sequence-specific capture probes. After target capture, the SNP is genotyped by performing an array-based allele-specific primer extension reaction. After extension with labeled nucleotides and signal amplification the array is read out on a BeadArray Reader. Genotyping quality was demonstrated by assaying a set of HapMap SNPs, and the call rate and accuracy are similar to that of GoldenGate assay. This technology enables simultaneously genotyping hundreds of thousands of loci from a single sample without the need for PCR or ligation steps, significantly reducing labor and potential sample-handling errors, and the cost per data point. The number of SNPs that can be assayed from

one sample is limited only by the numbers of features on the array with unconstrained locus selection. Thus, it allows large-scale interrogation of variation in the human genome at many levels of resolution, accelerating the ability of researchers to cost-effectively unlock the genetic basis of disease. The new assay fulfills the growing need for fixed-content genotyping and complements the GoldenGate custom genotyping assay.

11. Summary

Our genotyping platform combines a highly efficient genotyping assay with high quality Sentrix Arrays to deliver unprecedented quality and throughput at a reasonable cost. The Illumina genotyping assay has demonstrated exceptional performance by measurements of: reproducibility, call rate, allelelic heritability, and assay development success rate. The scalability of the platform allows the user to perform small pilot studies or large scale SNP genotyping association studies under the same system for complex human genetic disease studies, pharmacogenomic applications, as well as rapid development of molecular diagnostics.

Acknowledgements

The authors thank Marc Laurent, Dale Yuzuki, Steve Barnard, Diping Che, Todd Rubano, Chanfeng Zhao, and Lixin Zhou for their dedicated efforts and contributions to the development of the SNP genotyping assay, bead chemistry, imaging systems, and image analysis tools, bioinformatics, and process automation. We thank Mark Chee and David Barker for their support and guidance for this work.

References

[1] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W.J. Kent, P. Kitts, E.V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J.V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, A. Patrinos, M.J. Morgan, J. Szustakowki, P. de Jong, J.J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y.J. Chen, Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.

[2] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, J.D. Gocayne, P. Amanatides, R.M. Ballew, D.H. Huson, J.R. Wortman, Q. Zhang, C.D. Kodira, X.H. Zheng, L. Chen, M. Skupski, G. Subramanian, P.D. Thomas, J. Zhang, G.L. Gabor Miklos, C. Nelson, S. Broder, A.G. Clark, J. Nadeau, V.A. McKusick, N. Zinder, A.J. Levine, R.J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J.

Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A.E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T.J. Heiman, M.E. Higgins, R.R. Ji, Z. Ke, K.A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G.V. Merkulov, N. Milshina, H.M. Moore, A.K. Naik, V.A. Narayan, B. Neelam, D. Nusskern, D.B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M.L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N.N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J.F. Abril, R. Guigo, M.J. Campbell, K.V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, X. Zhu, The sequence of the human genome, Science 291 (2001) 1304–1351.

[3] Project The International HapMap Project, Nature 426 (2003) 789–796.

[4] F.S. Collins, E.D. Green, A.E. Guttmacher, M.S. Guyer, A vision for the future of genomics research, Nature 422 (2003) 835–847.

[5] N. Risch, K. Merikangas, The future of genetic studies of complex human diseases, Science 273 (1996) 1516–1517.

[6] D. Botstein, N. Risch, Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease, Nat. Genet. 33 (Suppl.) (2003) 228–237.

[7] A.C. Syvanen, Accessing genetic variation: genotyping single nucleotide polymorphisms, Nat. Rev. Genet. 2 (2001) 930–942.

[8] P.Y. Kwok, Methods for genotyping single nucleotide polymorphisms, Annu. Rev. Genomics Hum. Genet. 2 (2001) 235–258.

[9] K.L. Michael, L.C. Taylor, S.L. Schultz, D.R. Walt, Randomly ordered addressable high-density optical sensor arrays, Anal. Chem. 70 (1998) 1242–1248.

[10] D.R. Walt, Techview: molecular biology. Bead-based fiber-optic arrays, Science 287 (2000) 451–452.

[11] D.L. Barker, D. Theriault, D. Che, T. Dickinson, R. Shen, R. Kain, Self-assembled random arrays: high-performance imaging and genomics applications on a high-density microarray platform, Proc. SPIE 4966 (2003) 1–11.

[12] K.L. Gunderson, S. Kruglyak, M.S. Graige, F. Garcia, B.G. Kermani, C. Zhao, D. Che, T. Dickinson, E. Wickham, J. Bierle, D. Doucet, M. Milewski, R. Yang, C. Siegmund, J. Haas, L. Zhou, A. Oliphant, J.B. Fan, S. Barnard, M.S. Chee, Decoding randomly ordered DNA arrays, Genome Res. 14 (2004) 870–877.

[13] J.B. Fan, A. Oliphant, R. Shen, B. Kermani, F. Garcia, K. Gunderson, M. Hansen, F. Steemers, S.L. Butler, P. Deloukas, L. Galver, S. Hunt, C. McBride, M. Bibikova, J. Chen, E. Wickham, D. Doucet, W. Chang, D. Campbell, B. Zhang, S. Kruglyak, D. Bentley, J. Haas, P. Rigault, L. Zhou, J. Stuelpnagel, M.S. Chee, Highly parallel SNP genotyping, in: Cold Spring Harbor Symposia on Quantitative Biology, vol. 68, 2003, pp. 69–78.

[14] M.C. Longo, M.S. Berninger, J.L. Hartley, Use of uracil DNA glycosylase to control carry-over contamination in polymerase chain reactions, Gene 93 (1990) 125–128.

[15] A. Oliphant, D.L. Barker, J.R. Stuelpnagel, M.S. Chee, BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping, Biotechniques 56–58 (Suppl.) (2002) 51–60.

[16] J.B. Fan, J.M. Yeakley, M. Bibikova, E. Chudin, E. Wickham, J. Chen, D. Doucet, P. Rigault, B. Zhang, R. Shen, C. McBride, H.R. Li, X.D. Fu, A. Oliphant, D.L. Barker, M.S. Chee, A versatile assay for high-throughput gene expression profiling on universal array matrices, Genome Res. 14 (2004) 878–885.

[17] J.M. Yeakley, J.B. Fan, D. Doucet, L. Luo, E. Wickham, Z. Ye, M.S. Chee, X.D. Fu, Profiling alternative splicing on fiber-optic arrays, Nat. Biotechnol. 20 (2002) 353–358.

[18] H. Yan, W. Yuan, V.E. Velculescu, B. Vogelstein, K.W. Kinzler, Allelic variation in human gene expression, Science 297 (2002) 1143.

[19] R. Kaplan, K. Luettich, A. Heguy, N.R. Hackett, B.G. Harvey, R.G. Crystal, Monoallelic up-regulation of the imprinted H19 gene in airway epithelium of phenotypically normal cigarette smokers, Cancer Res. 63 (2003) 1475–1482.

[20] N.P. Gerry, N.E. Witowski, J. Day, R.P. Hammer, G. Barany, F. Barany, Universal DNA microarray method for multiplex detection of low abundance point mutations, J. Mol. Biol. 292 (1999) 251–262.

[21] M.A. Iannone, J.D. Taylor, J. Chen, M.S. Li, P. Rivers, K.A. Slentz-Kesler, M.P. Weiner, Multiplexed single nucleotide polymorphism genotyping by oligonucleotide ligation and flow cytometry, Cytometry 39 (2000) 131–140.

[22] J. Chen, M.A. Iannone, M.S. Li, J.D. Taylor, P. Rivers, A.J. Nelsen, K.A. Slentz-Kesler, A. Roses, M.P. Weiner, A microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension, Genome Res. 10 (2000) 549–557.
[23] J.B. Fan, X. Chen, M.K. Halushka, A. Berno, X. Huang, T. Ryder, R.J. Lipshutz, D.J. Lockhart, A. Chakravarti, Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays, Genome Res. 10 (2000) 853–860.
[24] D.L. Barker, M.S. Hansen, A.F. Faruqi, D. Giannola, O.R. Irsula, R.S. Lasken, M. Latterich, V. Makarov, A. Oliphant, J.H. Pinter, R. Shen, I. Sleptsova, W. Ziehler, E. Lai, Two methods of whole-genome amplification enable accurate genotyping across a 2320-SNP linkage panel, Genome Res. 14 (2004) 901–907.