BeadArray™ Technology: Enabling an Accurate, Cost-Effective Approach to High-Throughput Genotyping

Arnold Oliphant, David L. Barker, John R. Stuelpnagel, and Mark S. Chee
Illumina, San Diego, CA, USA

BioTechniques 32:S56-S61 (June 2002)

ABSTRACT

The Human Genome Project has opened the door to personalized medicine, provided that human genetic diversity can be analyzed in a high-throughput and cost-effective way. Illumina has developed a genotyping system that combines very high throughput and accuracy with low cost per SNP analysis. The system uses our BeadArray™ platform, a high level of multiplexing, and modular, scalable automation to meet the requirements for cost-effective, genome-wide linkage disequilibrium studies. As implemented in a high-throughput genotyping service facility at Illumina, the system has a current capacity of one million SNP assays per day and is easily expandable. Each SNP call is associated with a quality score that correlates with accuracy.

INTRODUCTION

Importance of SNPs in Human Genetics: The Promise and the Problem

The analysis of single nucleotide polymorphisms (SNPs) in the human genome offers the key to understanding the significance of human genetic variation (6,12,13). The promise of SNP analysis begins with its potential to reveal a detailed history of human genes and leads to the association of genes with disease and to the ability to predict the effectiveness and safety of medications for each individual (9). This promise is so compelling that the National Institutes of Health and the Wellcome Trust, the primary funding agencies of the public Human Genome Project, have announced a program to evaluate a large fraction of the more than three million SNPs in the human genome and sort them into haplotypes, the groups of neighboring SNPs that tend to be inherited together. This haplotype map will provide a framework for discovering associations between genes and disease and for using SNP analysis to support the development of personalized medicine. Applera Corporation, the primary private contributor to the human genome sequence, has also announced a program to develop a set of SNPs for use in genome-wide association studies.

Realizing this vision will require a SNP genotyping system that combines high throughput and accuracy with low cost per SNP analysis, a combination that has so far proved to be elusive (7,14). We believe that the solution to this problem is to combine a miniaturized assay platform, a high level of multiplexing, and flexible automation.

We achieve miniaturization by using a fiber-optic bundle as a substrate for the highest-density microarray available today. Ninety-six of these arrays are held together in a matrix (Array of Arrays™ matrix) that matches the spacing of a standard 96-well microplate. To efficiently use the high capacity of the platform, the SNP assays developed at Illumina have been designed for a high level of multiplexing. Currently, Illumina's service facility routinely multiplexes 288 SNP assays in each well of a microplate. The service facility is based on a proprietary laboratory information management system (LIMS) that uses informatic integration rather than mechanical integration to achieve low cost and flexibility. This design allows the use of standard modular hardware components and bar-coded microplates and ensures accurate sample tracking and data handling. The LIMS also automates oligonucleotide design and pooling for multiplexed assays, monitors internal assay controls carried out on each sample, and calculates a quality score for each assay result.

BEADARRAY PLATFORM

The randomly ordered BeadArray technology was invented by Professor David Walt and colleagues at Tufts University (10,15) and developed at Illumina as a platform for SNP genotyping and other high-throughput assays (http://www.illumina.com/prod_snp.htm). Each array is assembled on an optical imaging fiber bundle consisting of about 50 000 individual fibers fused together into a hexagonally packed matrix. The ends of the bundle are polished, and one end is etched to produce a well in each fiber. This process takes advantage of the intrinsic structure of the optical fibers in the bundle (Figure 1). After polishing and etching, a fiber bundle can hold up to 50 000 beads, each ap-

proximately 3 μm in diameter and spaced approximately 5 μm apart. This highly miniaturized array is about 1.4 mm across and has a packing density of 40 000 array elements per square millimeter—approximately 400 times the information density of a typical spotted microarray with 100- μm spacing. Each derivatized bead has several hundred thousand copies of a particular oligonucleotide covalently attached. Bead libraries are prepared by automated conjugation of oligonucleotides to silica beads, followed by quantitative pooling together of the individual bead types. The preparation of a bead library and assembly into an array are illustrated in Figure 2. The beads are stably associated with the wells under standard hybridization conditions.

For SNP genotyping, we use a library containing 1500 different bead types whose oligonucleotides have been selected to

not cross-hybridize with each other or with sequences in the human genome. When assembled on fiber bundles, this bead library forms a universal array of capture elements designed to hybridize to complementary sequences in multiplexed SNP assays. The use of a universal array simplifies manufacturing and reduces costs. Since there are 1500 bead types and approximately 50 000 beads, each bead type is present at an average redundancy of about 30-fold. Consequently, each SNP genotype is the result of data averaged from multiple beads, which increases precision and greatly reduces the possibility of error.

Array Matrix Format

To increase throughput, the arrays are formatted into a matrix that matches the wells of standard microplates (Sentry™ Array Matrix). Currently, a 96-array matrix is used in the SNP genotyping facility. However, the diameter of the bundles is small enough to permit a 384-well or even a 1536-well spacing. Therefore, the number of samples processed by one array matrix can be increased up to 16-fold with no fundamental change to the technology.

Because the arrays are assembled randomly from a library of beads, each bead must be located and identified post-assembly. This is accomplished during manufacturing by a decoding process, which comprises a series of hybridizations using fluorescently labeled oligonucleotides that are complementary to the capture probes on the beads. This process scales exponentially, is very efficient, and incorporates quantitative error checking. It also provides a quality control analysis of every array element. The decoding process will be described in detail in a separate publication (M. Chee et al., manuscript in preparation).

SNP Assays

The BeadArray platform is compatible with a variety of different SNP assays. Essentially, an assay associates a fluorescent label with an oligonucleotide sequence that is complementary to an address on a bead. This association must be made with high allele- and locus-specificity and should be amenable to high levels of multiplexing. Robust genotyping has been demonstrated

Figure 1. Structure and properties of an individual optical fiber. Each fiber has a light-conducting inner core that is surrounded by a cladding of different refractive index. The core can be chemically etched at a different rate from its surrounding cladding. By treating the polished end of an optical fiber bundle with acid, an array of microwells is generated. The geometry and dimensions of the array are determined by the physical specifications of the optical fiber and are chosen so that one bead can fit in each well in the array. Once a labeled target nucleic acid is hybridized to beads in the array, a fluorescent signal can be generated by making use of the optical properties of the fiber. An excitation beam is guided to the bead through the fiber bundle, and emitted fluorescence is guided back up the fiber, allowing the array to be imaged at the opposite end of the optical fiber bundle.

Figure 2. Assembly of a randomly ordered fiber optic array. (A) A collection of bead types, each with a distinct oligonucleotide capture probe, is pooled. An etched fiber optic bundle is exposed to the bead pool, allowing individual beads to assemble into the microwells at the bundle's end. (B) Scanning electron micrograph of an assembled array containing 3- μm diameter silica beads. (Courtesy of Todd Dickinson.)

PRODUCT & TECHNOLOGY REPORT

with versions of single-base extension (1-3,11), the oligonucleotide ligation assay (4,5,8), and allele-specific extension. Each assay is developed for automation. All steps are conducted in standard microplates and each operation is compatible with robotic procedures. An OLA-based genotyping system is being developed in collaboration with Applied Biosystems.

Automation and Informatics

The high-throughput genotyping system employs informatic integration rather than mechanical integration. Each sample and each reagent is given a unique identifier. Before each step, these unique identifiers are matched by real-time interaction with a LIMS database to identify components and confirm that they are correctly queued for the next procedure. This real-time interaction ensures accurate processing of samples, a fundamental requirement of a high-quality commercial genotyping system. Our informatic-centric system design has several other advantages. By avoiding custom mechanical integration, the production facility can use standard laboratory instruments such as thermocyclers, hybridization ovens, imagers, and pipetting robots. Therefore, the system can be scaled up simply by adding more components. The use of standard components linked informatically and aligned in parallel also contributes to robustness. Because the components are standardized, failure occurs less often, and when component failure does occur, the component redundancy does not result in failure of the system. The current operation has a capacity of one million genotypes per day and is easy to scale.

Information Management

The custom LIMS also provides file and data management. As array images are collected in two colors, corresponding to the two possible alleles present at each SNP locus, intensity values are extracted and data files are created. Intensity data from beads of the same type are averaged, and a SNP call is automatically made. The genotyping software also generates a quality score, called a GenCall™ score, based on quantitative measurements of homozygote and heterozygote cluster separation. The GenCall score correlates with the accuracy of the genotyping call. Throughout these data and file transformations, the LIMS tracks the unique identifiers of each file and controls the transformations to ensure data integrity. Genotyping calls and quality scores are exportable from the LIMS database to other databases or analysis programs via standard formats.

Genotyping Accuracy

Determining the accuracy of genotyping calls is not a straightforward exercise and deserves some comment. We offer the following definitions and procedures for evaluating genotyping accuracy. Each has its advantages and disadvantages.

1. *Concordance* refers to the comparison of two sets of genotyping results obtained by different methods. Consequently, the resolution of discordant results between the two methods requires analysis by a third technology, such as DNA sequencing. These studies are expensive, especially since genotyping should have accuracy above 99%, requiring a large number of samples and SNPs for meaningful statistics. In addition, studies of this

Figure 3. (A) Correlation of assay quality, as estimated by GenCall score, with cumulative accuracy of SNP calls. The data are progressively excluded by GenCall score, beginning at a GenCall score of zero where all of the data are evaluated. A total of 355 DNAs were assayed on both strands at 288 SNP loci (408 960 total allele calls). Accuracy is measured by the correlation of alleles determined on the two DNA strands. (B) The same 408 960 allele calls from (A) are plotted in order of GenCall score. For this study, one can select a call rate of 99% to obtain 99.7% accuracy, whereas a call rate of 90% yields 99.8% accuracy. (C) Variation of accuracy sorted as a function of DNA sample. Accuracy, as measured by strand correlation, was determined for each of 355 DNA samples. Each sample was assayed at 288 SNP loci on both strands.

PRODUCT & TECHNOLOGY REPORT

Table 1. Reproducibility and Strand Correlation from Various Studies

Study	No. of Loci	No. of Alleles	Call Rate	Reproducibility
Customer A	556	173 472	99.6%	99.83%
Customer B	313	565 278	99.9%	99.59%

Study	No. of Loci	No. of Alleles	Call Rate	Strand Correlation
Study 1	288	408 960	99.1%	99.71%
Study 2	162	585 144	97.9%	99.35%

Note: These percentages were calculated using an allele-based scoring system. This method allows us to differentiate homozygote-heterozygote discordance from homozygote-homozygote discordance.

type may not predict accuracy in further studies since the loci and DNA may vary.

2. *Reproducibility* simply refers to the reproduction of genotyping calls on replicate DNA samples. When using reproducibility as a surrogate for accuracy in a customer study, one assumes that the DNA used for the reproducibility study is equivalent to the DNA submitted by a customer. Of course, one can reproducibly generate an incorrect genotyping call. However, in efforts to rapidly develop new loci, it is one of several efficient measures of success. In most of our studies, we reserve several wells of each plate for replicate DNAs.

3. *Strand correlation* refers to the comparison of genotype calls for each SNP on both DNA strands. The assumptions for this method of calculating accuracy are that, for any particular

locus, the genotype on the top strand matches the genotype on the bottom strand, and that the ability to develop a SNP assay for a particular locus is strand-independent. We believe that the second assumption is not always true; instead, if a robust assay is developed on one strand, it is more likely that the assay will develop well on the other strand. As a result, strand correlation may slightly overestimate accuracy. Nevertheless, we have found strand correlation to be an effective way to monitor accuracy in studies to test and refine system performance.

Table 1 presents reproducibility data from two customer data sets and studies of strand correlation with two sets of loci designed to test the accuracy of the Sentrix-based genotyping system.

Call Rate

One final variable affecting accuracy is call rate. By lowering the call rate and excluding low-quality results, accuracy can be increased. The genotyping software developed at Illumina calls nearly all the genotypes and provides a quality score associated with that call. It is then left up to the analyzer of the data to determine call rate, depending on his or her need to balance completeness of data versus accuracy of data. For instance, one can lower the call rate and analyze fewer data but have a higher assurance of accuracy. This balancing is critical for linkage studies, where a 1% error rate may double the number of samples required to achieve a given power of analysis. To take advantage of this flexibility, it is necessary to have a quality score for each SNP determination that is known to correlate with accuracy. GenCall scores have been shown to correlate with accuracy as measured by strand correlation (Figure 3A). Figure 3B shows how a desired accuracy can be selected by ordering genotype calls by their GenCall score. In comparing accuracy across platforms or studies, it is important to consider the call rate in association with accuracy.

A critical factor that affects accuracy is, of course, the quality of the DNA sample used in the assay. This is shown in Figure 3C, where accuracy, as measured by strand correlation, is plotted for each of 355 DNA samples. In this study a large portion of the inaccurate calls can be assigned to relatively few DNAs.

CONCLUSION

We have briefly described Illumina's BeadArray technology and its application to SNP genotyping as practiced in our genotyping service facility. The system remains in development, as we continue to explore the limits of multiplexing and other contributors to cost reduction while maintaining high accuracy. We recently demonstrated 1152 simultaneous SNP assays per microplate well with a high success rate (Figure 4). Other applications are being developed for the BeadArray platform, including mRNA and protein profiling (16).

ACKNOWLEDGMENTS

The technology development summarized here would not have been possible without the efforts of many dedicated individuals. Our thanks to all those at Illumina who made contributions in automation, chemistry, engineering, informatics, manu-

Figure 4. Cluster diagram showing assays of a single SNP locus multiplexed together with 1151 other loci genotyped on the BeadArray platform. Results from 96 different DNA samples are shown as the fluorescence intensity in individual beads. Clusters primarily labeled with one fluor represent homozygotes; the cluster labeled with two fluors indicates heterozygotes.

facturing, molecular biology, high-throughput operations, and process development. This work was supported in part by grants from the National Institutes of Health, in particular Grant No. R44 HG02003 from the National Human Genome Research Institute to MSC.

REFERENCES

1. Cai, H., P.S. White, D. Torney, A. Deshpande, Z. Wang, R.A. Keller, B. Marrone, and J.P. Nolan. 2000. Flow cytometry-based minisequencing: a new platform for high-throughput single-nucleotide polymorphism scoring. *Genomics* 66:135-143.
2. Chen, J., M.A. Iannone, M.S. Li, J.D. Taylor, P. Rivers, A.J. Nelsen, K.A. Slentz-Kesler, A. Roses, and M.P. Weiner. 2000. A microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension. *Genome Res.* 10:549-557.
3. Fan, J.B., X. Chen, M.K. Halushka, A. Berno, X. Huang, T. Ryder, R.J. Lipshutz, D.J. Lockhart, and A. Chakravarti. 2000. Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Res.* 10:853-860.
4. Gerry, N.P., N.E. Witowski, J. Day, R.P. Hammer, G. Barany, and F. Barany. 1999. Universal DNA microarray method for multiplex detection of low abundance point mutations. *J. Mol. Biol.* 292:251-262.
5. Iannone, M.A., J.D. Taylor, J. Chen, M.S. Li, P. Rivers, K.A. Slentz-Kesler, and M.P. Weiner. 2000. Multiplexed single nucleotide polymorphism genotyping by oligonucleotide ligation and flow cytometry. *Cytometry* 39:131-140.
6. Kruglyak, L. and D.A. Nickerson. 2001. Variation is the spice of life. *Nat. Genet.* 27:234-236.
7. Kwok, P.Y. 2001. Methods for genotyping single nucleotide polymorphisms. *Annu. Rev. Genomics Hum. Genet.* 2:235-258.
8. Landegren, U., R. Kaiser, J. Sanders, and L. Hood. 1988. A ligase-mediated gene detection technique. *Science* 241:1077-1080.
9. McCarthy, J.J. and R. Hilfiker. 2000. The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nat. Biotechnol.* 18:505-508.
10. Michael, K.L., L.C. Taylor, S.L. Schultz, and D.R. Walt. 1998. Randomly ordered addressable high-density optical sensor arrays. *Anal. Chem.* 70:1242-1248.
11. Pastinen, T., A. Kurg, A. Metspalu, L. Peltonen, and A.C. Syvanen. 1997. Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Res.* 7:606-614.
12. Reich, D.E., M. Cargill, S. Bolk, J. Ireland, P.C. Sabeti, D.J. Richter, T. Lavery, R. Kouyoumjian, et al. 2001. Linkage disequilibrium in the human genome. *Nature* 411:199-204.
13. Sachidanandam, R., D. Weissman, S.C. Schmidt, J.M. Kakol, L.D. Stein, G. Marth, S. Sherry, J.C. Mullikin, et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928-933.
14. Syvanen, A.C. 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.* 2:930-942.
15. Walt, D.R. 2000. Techview: molecular biology. Bead-based fiber-optic arrays. *Science* 287:451-452.
16. Yeakley, J.M., J.B. Fan, D. Doucet, L. Luo, E. Wickham, Z. Ye, M.S. Chee, and X.D. Fu. 2002. Profiling alternative splicing on fiber-optic arrays. *Nat. Biotechnol.* 20:353-358.

Address correspondence to:

Dr. Arnold Oliphant
 Illumina
 9885 Towne Centre Drive
 San Diego, CA 92121-1975, USA
 e-mail: aoliphant@illumina.com