ASSOCIATION STUDY

Application Note

Solutions for Each Step of an Association Study

Overview
Affymetrix offers solutions for each step of a typical whole-genome association study—from initial discovery, to replication and fine mapping, sequence analysis, and functional testing of genes. Researchers worldwide are now using these tools to identify the genetic causes of a wide range of common diseases such as autism, obesity, cancer, asthma, and diabetes.

WHOLE-GENOME SCAN
Scan whole-genome hypothesis free
Products:
• GeneChip® Mapping 500K Array Set

FINE MAPPING/REPLICATION
Reduce to candidate region/gene; repeat in another population
Products:
• Custom Targeted Genotyping

SEQUENCE ANALYSIS
Identify candidate nucleotide variants
Products:
• GeneChip CustomSeq® Arrays

FUNCTIONAL TESTS
Understand the function
Products:
• Gene Expression
• Exon Expression
• Gene Regulation
• Full Transcriptome

WHOLE-GENOME SCAN
In this initial step, an unbiased subset of all SNPs are genotyped on a microarray to determine which are statistically associated with a given phenotype—a predisposition to disease, drug response, diagnosis, or prognostic outcome.

SEQUENCE ANALYSIS
In this third step, genomic regions are chosen for detailed sequence analysis to identify all SNPs in these regions within a particular population. Regions are chosen based on proximity to statistically significant SNPs and, often, functional annotations. Those SNPs can then be analyzed for statistical association with the phenotype and researchers can ultimately pinpoint one or more causal variants.

FINE MAPPING/REPLICATION
In this second step, additional SNPs in the regions identified during the initial whole genome scan are analyzed for statistical association with the phenotype. SNPs are re-genotyped in additional sample populations, or in the same population, to replicate the initial discoveries.

FUNCTIONAL TESTING
In this final step, researchers look to verify and explore the association between the causal variant that has been discovered and the phenotypic effects.
Whole-Genome Scan

Study Design and Genetic Power

LOWER COSTS. MORE SAMPLES. HIGHER GENETIC POWER.

The number of samples genotyped in your study is the single most important factor in powering your whole-genome scan. The *GeneChip® Human Mapping 500K Array Set* reduces genotyping costs to $250 per sample,

which allows you to dramatically increase the number of samples you can run in a single study. More samples translate into higher genetic power to detect true associations (Figures 1 and 2).

1,000,000 SNPS IN 2007

Use a platform that will grow with your future needs. Affymetrix introduce a 1 million-SNP product by the first quarter of 2007. Additionally, by the end of 2006, Affymetrix will begin offering a single-array genotyping product for whole-genome association studies to replace the Mapping 500K Array Set. This new product, developed in collaboration with the Broad Institute of MIT and Harvard, will allow researchers to increase throughput, devote more resources toward performing larger experiments, and analyze more copy number variation in their studies.

CONTROLS DATABASE

Access 6 billion control genotypes to supplement your data—without spending a penny more. Over 10 investigators worldwide have agreed to make their 500K genotyping data—more than 13,000 samples from multiple ethnic backgrounds—accessible to other scientists.

Analysis has shown that doubling the number of control genotypes can result in a 20 percent increase in genetic power (Figure 3). Also, by accessing these freely available data, you may be able to run fewer controls.

Researchers using the controls database will be able to select age- and sex-matched controls from ethnic backgrounds similar to their cases, and add them to their current study.

COPY NUMBER AND SNP VARIATION

Analyze SNP and copy number variation in a single assay. A growing number of geneticists are studying germ line copy number variation in their association or linkage studies. The Mapping 500K Array Set allows you to identify disease-causing copy number variation on the same arrays, with no additional cost. Genotype information, in addition to quantitative copy number information, allows researchers to distinguish between copy-neutral events, such as Uniparental Disomy (UPD).

More resolution provides more power to detect chromosomal aberrations and better define boundaries and break points. The Mapping 500K Array Set provides a mean resolution of 5.8kb and a median marker distance of 2.5kb, giving scientists the highest physical coverage of the genome for their study.

Affymetrix offers a growing family of software solutions for copy number analysis, including standardized third-party software providers:

Figure 1: This chart illustrates the influence of sample size on genetic power for varying percentages of genome coverage.

Figure 2: This example power calculation illustrates how more samples impact genetic power.
The Affymetrix® Copy Number Analysis Tool (CNAT) and the Affymetrix Integrated Genome Browser (IGB) are available for download at www.affymetrix.com. CNAT supports copy number and Loss of Heterozygosity (LOH) analysis of Mapping Array data, including 10K, 50K, and 100K. Look for expanded support for the 500K in late 2006.

Partek® Genomics Solution™ is the first commercial software application available for copy number analysis in support of Mapping 10K, 100K, and 500K arrays.

CNAG, from Dr. Seishi Ogawa at the University of Tokyo, and dChip SNP, from Dr. Cheng Li at Harvard University and the Dana Farber Cancer Institute, are both academic software tools available for copy number and LOH analysis of Affymetrix Mapping Arrays, including 10K, 100K, and 500K.

Genotyping Software, Downstream Analysis, and Sample Tracking

Affymetrix uses an open informatics platform and partners with a community of over 25 software development partners, to provide flexible and innovative genotyping software solutions for researchers in different laboratory environments, with different software development resources.

GENECHIP® OPERATING SYSTEM (GCOS)

Affymetrix GeneChip® Operating Software (GCOS) uses a database structure for storage and management of samples, sample attributes, associated arrays, and the resulting data. GCOS is an out-of-the-box software solution for primary data generation. Using a standard Microsoft Windows XP workstation, GCOS can associate sample attributes with the resulting files. These data are stored in a lightweight, freely available database.

INPUT DATA AND SAMPLE ATTRIBUTES QUICKLY AND EASILY INTO GCOS

Today, Affymetrix supports a number of methods to push data into GCOS through a variety of input data types—from simple CSV or XML files, to ties to external databases through embedded libraries. Once the data reside within GCOS, sample attributes are natively associated with the resulting files, and samples run on multiple arrays are properly associated. Researchers can use GCOS as the master data repository, or use existing methods to pull data out of GCOS into another data repository. To learn more about integration options, please contact the Affymetrix Developers’ Network Mailbox: devnet@affymetrix.com.

GENECHIP GENOTYPING ANALYSIS SOFTWARE (GTYPE)

Affymetrix GeneChip® Genotyping Analysis Software (GTYPE) is part of the GeneChip Mapping Array System, specifically designed to give highly accurate, automated SNP allele calls for GeneChip Mapping Arrays. In addition, GTYPE was developed to enable workflows for a variety of SNP-related applications, such as whole-genome association, linkage, and copy number studies.

GTYPE provides high-quality genotyping, quality control reporting, and streamlined data analysis for projects with thousands of samples. GTYPE integrates with the BRLMM genotyping algorithm; BRLMM performs multiple-chip analysis, fitting probe effects to increase precision on signal estimates for the two alleles of each SNP. This is followed by a Bayesian classification approach to make genotype calls.

Learn more about BRLMM, the genotype calling algorithm, at www.affymetrix.com/support/technical/product_updates/brlmm_algorithm.affx

Watch a video with Greg Marcus, Affymetrix Sr. Product Manager, discussing the algorithm improvements implemented in BRLMM at www.affymetrix.com/userForum/news/collaborations/marcus.affx

Look at the GTYPE Data Analysis Workflow at www.affymetrix.com/support/technical/other/gtype_e_workflow_diagram.pdf

GENECHIP-COMPATIBLE™ SOFTWARE

Over 25 companies have built software that is integrated with the GeneChip system. If you tend to “buy” rather than “build” software, Affymetrix has already taken the first step to help you qualify your software vendors. Affymetrix’ GeneChip-compatible™ Application Program features commercial software vendors that have integrated their software products to work cleanly with the Affymetrix platform. The GeneChip-compatible Software Solutions Catalog (www.affymetrix.com/products/software/compatible/snp.affx) provides a comprehensive list of software packages both at the level of SNP analysis as well...
as laboratory management. Simply look for the GeneChip-compatible logo on any software package that you buy.

GeneChip-compatible providers for SNP Analysis:

- **Golden Helix HelixTree® Genetics Analysis Software** View a recent GeneChip-compatible™ webcast illustrating a complete analysis workflow of data from the Mapping 500K Array Set at www.affymetrix.com/corporate/events/processEventRegistrationArchive.jsp?event=71
- **Progeny Software’s Progeny Lab** View a recent webcast illustrating how Progeny Lab can aid in data management from the GeneChip Human Mapping 500K Array Set. www.affymetrix.com/corporate/events/processEventRegistrationArchive.jsp?event=68
- **Sapios Sciences® Exemplar Genotyping Analysis Suite**
- **Biocomputing Platforms’ BC/GENE JMP® Genetics from SAS**

### INTEGRATE EXPRESSION AND GENOTYPING DATA
The Integrated Genome Browser (IGB, pronounced “ig-bee”) from Affymetrix allows researchers to visualize genomic data and annotations from multiple data sources. IGB enables researchers to integrate genotyping and expression data into a single view of the genome, on a single platform.

### AFFYMETRIX GENECHIP® COMMAND CONSOLE™
Affymetrix has announced the upcoming release of the next generation of instrument control software—Affymetrix GeneChip® Command Console™ (AGCC), which will co-exist with GCOS. Command Console software will leverage the latest software development technologies, including web-based user interfaces, file indexing, and integration via web services. Licenses to Command Console software will be freely available to all current Affymetrix customers. Learn more at: www.affymetrix.com/userForum/news/newProducts/GC_CommandConsole.uf

### The Platform

**PERFORM YOUR RESEARCH ON A PROVEN, STANDARDIZED PLATFORM**
The Mapping 500K Array Set is being used by over 200 investigators worldwide. There have been nearly 120 scientific publications in peer-reviewed journals, across multiple applications based on Affymetrix genotyping technology—including ground-breaking association studies in sudden cardiac death (Arking, et al.), obesity (Christman, et al.) and macular degeneration (Klein et al).

Read over 120 Peer-reviewed Publications using the Mapping Assay at www.affymetrix.com/support/technical/other/mapping_publications_august_2006.pdf

### SUDDEN CARDIAC DEATH

**OBESITY**

Watch an interview with Johns Hopkins School of Medicine’s, Aravinda Chakravarti whose team has identified a genetic variant associated with electocardiographic QT interval, a trait associated with increased risk of sudden cardiac death at www.microarraybulletin.com/community/article.php?p=241

Read an interview with Harvard University School of Public Health’s Christoph Lange on algorithm development that enabled this study at www.microarraybulletin.com/community/article.php?p=206

### THE TARGETED GENOTYPING ASSAY

Affymetrix Custom Targeted Genotyping allows you to genotype more of the SNPs you want. The conversion rate specification for the assay (the percentage of working SNPs that make it into the final assay panel from what was submitted) is greater than 80 percent, and typically greater than 90 percent. This conversion rate is considerably higher than other custom SNP genotyping assays on the market. Guarantees from other companies about conversion rates are based on the number of SNPs that have already passed a bioinformatic screening step.

---

**Fine Mapping/Replication**

### 50,000 CUSTOM SNPS AT YOUR FINGERTIPS

Until recently, researchers have been pursuing just a handful of SNPs in the replication or fine mapping phase of an association study because there was no high-throughput, affordable technology that allowed them to genotype a customized set of SNPs. By applying a robust replication technology, the initial discovery scan can be maximally informative.

With Affymetrix Custom Targeted Genotyping Panels, you no longer need to let technology limit your study. Using this system, which is based on Molecular Inversion Probe Technology, you can genotype up to 50,000 SNPs of your choice, enabling large-scale fine mapping and replication studies (Figures 4, 5, and 6).

**DIAGNOSTICS**

Plan a diagnostic path for your study, and your business. Affymetrix has the first and only microarray instrument to be cleared by the FDA for in vitro diagnostics (IVD) use, providing a standardized platform for nucleic acid diagnostics. Additionally, Affymetrix will soon be the only microarray manufacturer to offer a CLIA-approved laboratory for development of clinical assays and samples. Under the Powered by Affymetrix™ program, companies can license technology from Affymetrix to develop innovative microarray products.
QUALITY CHECKS EVERY STEP OF THE WAY
Every Affymetrix Custom Targeted Genotyping oligo panel undergoes functional quality checks before it ever leaves our facility. Additionally, only Affymetrix Custom Targeted Genotyping allows you to see raw data from all four possible genotype calls. This gives researchers a built-in quality check for each SNP—you will know if a SNP call has failed if there is too much background signal. With allele-specific probes, a researcher will not know that is the case.

GET A STREAMLINED WORKFLOW
There is one protocol for all custom assay panels—from 3,000 to 50,000 SNPs; the assay leverages routine laboratory techniques, materials, and equipment.

CUSTOM ASSAY SPECIFICATIONS
- Conversion Rate: >80 percent
- Data Completeness: ≥97.0 percent
- Repeatability: ≥99.3 percent
- Accuracy: ≥99.3

The Custom Genotyping Platform, Services, and Throughput

THE PLATFORM
Buy a standardized platform that’s proven in the field. The Affymetrix Custom Targeted Genotyping Assay has already been used to discover genetic associations in common diseases such as juvenile diabetes.

JUVENILE DIABETES
Watch an interview with Dr. John Todd, from the JDRF/Wellcome Trust Diabetes Laboratory at www.microarraybulletin.com/community/article.php?p=203

BOVINE TRAITS
Watch an interview with Baylor College of Medicine’s Richard Gibbs at www.microarraybulletin.com/community/article.php?p=213

GENOTYPING SERVICES
Researchers can perform targeted genotyping on their existing Affymetrix system with a simple upgrade, or have a service provider do the work. Genotyping services offer researchers a way to complete their projects quickly—simply send the samples in and receive data back. Affymetrix can provide this service, or researchers can work with one of our expert service providers—they have years of experience, extensive scientific expertise, and tight process controls. Affymetrix works with scientists to determine whether an in-house system or a service provider will best meet their needs.

THROUGHPUT
The Affymetrix Targeted Genotyping Assay genotypes up to 50,000 SNPs in 48 samples per day (including controls) and 192 samples per week. That is 2.4 million genotypes/day (Figure 7). Throughputs can be even higher by using additional equipment.

Figure 4: Custom Targeted Genotyping combines custom SNP assay panels with Universal Tag Arrays.

<table>
<thead>
<tr>
<th>Custom SNP Assays</th>
<th>Universal Tag Array</th>
</tr>
</thead>
<tbody>
<tr>
<td>You can genotype 3,000 to 50,000 SNPs on the same system and the same type of assay.</td>
<td>Custom assays don’t require custom arrays. Universal Tag Arrays contain novel, bioinformatically designed tag sequences that result in minimal potential for cross-hybridization and do not require redesign or recreation of custom GeneChip arrays.</td>
</tr>
<tr>
<td>Custom 3K SNP Kit</td>
<td>Universal 3K Array</td>
</tr>
<tr>
<td>Custom 5K SNP Kit</td>
<td>Universal 5K Array</td>
</tr>
<tr>
<td>Custom 10K SNP Kit</td>
<td>Universal 10K Array</td>
</tr>
<tr>
<td>Custom 20K SNP Kit</td>
<td>Universal 25K Array</td>
</tr>
<tr>
<td>Custom 50K SNP Kit*</td>
<td>Universal 70K Array</td>
</tr>
</tbody>
</table>

*50K services only

Figure 5: Affymetrix Targeted Genotyping Custom Panel Order Process.

GET HIGHLY ACCURATE GENOTYPE CALLS

GTGS implements an enhanced genotype calling algorithm that improves the robustness and quality of genotyping. The expectation-maximization (E-M) algorithm handles the enormous number of data points (10^10) that need to be analyzed in a typical targeted genotyping experiment. It requires no manual intervention at the detailed level of a single call, sample, or marker. The algorithm performs data normalization followed by the application of a clustering algorithm for genotype calling. Optional marker filters are then applied to the clustered data, and per-
Genomic DNA  
- 46 S  
- 20K A  
Per Day:  
- 46 S  
~ 1 Million A  

**Figure 6: The Custom Targeted Genotyping System is based on the Molecular Inversion Probe Assay.**

MAKE QUALITY CONTROL CHECKS MORE EFFICIENT  
Quality control (QC) is streamlined with GTGS because the software automatically flags samples or markers that do not pass quality control criteria. GTGS quality checks on three levels: for each call individually (signals from all four nucleotides are measured), at the sample level (is the overall data quality from a single array sufficient?), and at the marker level (is the data quality for a marker across multiple samples sufficient?). GTGS provides a visualization of each QC step, such as trend charts to look at QC metrics vs. time, and sample location, as well as a visual inspection of each array at each nucleotide channel.

TRACK YOUR SAMPLES FROM START TO FINISH  
The complete custom genotyping assay is divided into two major steps: pre-amplification and post-amplification. Dedicated laboratory space and computer stations are recommended for both steps. The GTGS software allows you to track your samples from sample plate to GeneChip array between these laboratory areas.

EXPORT YOUR DATA IN ONE OF MULTIPLE TEXT-BASED FORMATS  
Genotypes are exported as a single file, collating the data across all of your samples belonging to a single-batch genotyping operation. The tab-delimited output provides the flexibility to reformat as desired.

Resequencing

The Affymetrix CustomSeq® Resequencing Program  
The Affymetrix CustomSeq® Resequencing Program enables researchers to design custom arrays for rapid identification of all SNPs—common, rare, or non-synonymous—in a specific genomic region. Complete sequence analysis ensures the highest probability of finding causative SNPs, since many causal variants are rare or heterogeneous in the population. CustomSeq Arrays provide a high-throughput method of SNP detection by enabling the analysis of large contiguous chromosomal regions, or multiple candidate genes, on a single array.

SEQUENCE MORE DNA FOR LESS THAN A TENTH OF A PENNY PER BASE

CustomSeq Arrays enable the analysis of up to 300,000 bases of double-stranded sequence (600,000 bases total) on a single array, providing the most efficient and cost-effective method for interrogating large amounts of sequence in a single experiment.

GET HIGH-QUALITY SEQUENCING DATA IN A SINGLE PASS

Affymetrix CustomSeq Arrays call greater than 90 percent of bases at greater than 99.9 percent accuracy and 99.9 percent reproducibility.

PERFORMANCE ON CEPH SAMPLES

Affymetrix has characterized the performance of the 300kb resequencing arrays across multiple designs, representing homozygous and heterozygous model systems. To demonstrate performance for a typical human region, resequencing array data and dideoxy sequence data were collected from 16 diploid CEPH (Centre d’etude du polymorphisme humaine) individuals across 115kb of non-repetitive sequence on Chromosome 4—1.84 mega base pairs (Mb) in total. While the performance of every custom-designed array depends on the specific genomic sequence, the overall call rate in this study was found to be 96.56 percent; overall accuracy was 99.95 percent.

CUSTOMSEQ ARRAY THROUGHPUT

CustomSeq Arrays give you data faster, so you can find your SNPs faster. CustomSeq arrays deliver completed sequence in 48 hours with minimal alignment, curation, or hand editing.

Put your time and money into your experiments, not your reagents. By leveraging long-range PCR, the number of amplification reactions required can be dramatically reduced, decreasing the cost and time associated with PCR.

Do your resequencing, whole-genome scanning, and targeted genotyping on one reliable, integrated Affymetrix system.

CUSTOM RESEQUENCING PUBLICATIONS

Take advantage of a proven, cutting-edge technology. The publications below highlight the advantages and system performance in two candidate gene studies.


EXPERIMENT SUMMARY:

- Sequenced 155 exons from 35 patients (29,214 bp) from genes associated with autosomal recessive Retinitis Pigmentosa (arRP), a condition leading to irreversible blindness or severe visual impairment, and affecting one in 3500 individuals worldwide
- Retinitis Pigmentosa shows broad genetic heterogeneity with at least 32 genes known to be associated with various forms (autosomal dominant, autosomal recessive, and X-linked)

BASE Calling PERFORMANCE

- Average Call Rates=97.60 percent (individual arrays ranged from 96.0 percent to 98.5 percent)
- Accuracy >99 percent
- Reproducibility >99 percent

SNP DETECTION

- 506 sequence changes identified
- Accurately detected 382 previously reported SNPs and identified 113 novel SNPs
- Accurately detected five previously reported mutations and identified seven novel rare mutations


EXPERIMENT SUMMARY:

- Sequenced 164 exons (23,966 bp) from genes associated with cancer in 20 lung tumor samples with matched normal controls
- Dideoxy sequencing was performed on a subset of exons in order to evaluate the performance of the arrays

COMPARISON TO DIDEOXY SEQUENCING

- Call rate=97.53 percent
- Overall accuracy=99.99 percent
- Only four SNP call errors reported
- Three homozygous SNPs called heterozygous and one heterozygous SNP called reference
Functional Testing Through Expression Analysis

Finding a causal variant or gene is just the beginning—understanding how that affects the entire biological system is the critical next step to ultimately developing a therapeutic or diagnostic. A new generation of expression microarrays allows you to examine all aspects of the genome—exons, introns, and other non-coding regulatory regions—to decipher pathways and functional effects.

GeneChip® Exon Arrays enable you to identify expression profiles affected by a disease-causing variant or gene, including altered splicing patterns. GeneChip® Tiling Arrays for ChIP-on-chip analysis allow you to identify regions of DNA with which your specific protein interacts—at the highest resolution, and with the most complete coverage of the human genome. Affymetrix 3'IVT Expression Arrays are the industry standard and have been used in thousands of studies.

Functional Tests