

Perspectives

Inside the Development of the Human1M DNA Analysis BeadChip

A former Illumina® employee remembers the development of a new DNA Analysis BeadChip.

by Sarah Shaw Murray, Ph.D. Director of Genetics, Scripps Genomic Medicine

The genetic dissection of complex disease traits is an extremely challenging endeavor. The hallmark of a complex disease trait is that the underlying causes are the combination of several genetic susceptibility factors, in addition to environmental influences. It has been known for many years that the association study (i.e., comparison between many cases and controls) is the epidemiological study design best suited for finding the relatively small gene effects associated with complex disease traits. Unfortunately, because of the high marker density required for association studies, association studies have been limited

“In the past three months, there has been a constant stream of scientific publications...of specific genetic risk factors for complex disease traits”

in scope to studying candidate genes or regions of the genome. Because genome-wide association studies have not been practical, researchers (including myself) have conducted genome-wide linkage studies that often yielded equivocal results. These linkage studies have looked at the co-segregation of polymorphic markers

and disease traits in families, often yielding very large genomic regions that potentially harbor the complex disease trait. Although not all linkage studies have been disappointing, the highly successful studies have been few and far between. Years of hit-or-miss linkage studies are in stark contrast to the current break-neck pace of highly significant findings being published almost weekly. In the past three months, there has been a constant stream of scientific publications in *Science*, *Nature*, *Nature Genetics*, and other high-profile journals reporting discoveries of specific genetic risk factors for complex disease traits. This explosion has occurred in part due to the International HapMap Project, and in part because of new technology that enables whole genome association studies to detect these relatively small gene effects. As the technology continues to improve and more content is added to whole genome genotyping products, discoveries will continue to follow for even more complex diseases and smaller gene effects.

It was a very rewarding experience for me to be a part of the International HapMap Project and then to have the opportunity to use the HapMap data to build whole genome genotyping tools for the genetics community. The Illumina HumanHap300, HumanHap550, and HumanHap650Y BeadChips were all developed by choosing tag single nucleotide polymorphisms



Dr. Sarah Shaw Murray was a staff geneticist at Illumina from January 2003 through March 2007. She had an important role in developing Illumina's genotyping product portfolio, most recently designing the content for the whole-genome genotyping products, including the Human1M BeadChip.

Dr. Murray now serves as Director of Genetics at the newly formed Scripps Genomic Medicine program. SGM is a collaboration between Scripps Health, the dominant health care provider in the San Diego region, and The Scripps Research Institute (TSRI), the largest non-private biomedical research facility in the world. SGM's goals range from gene discovery of complex traits to gene-based clinical trials

(SNPs) from the International HapMap Project. Tag SNPs are proxies for groups of highly correlated SNPs. Information can be captured for the entire group of correlated SNPs by genotyping only one representative SNP, the tag SNP. By choosing tag SNPs, we were able to maximize efficiency, coverage, and power of carefully selected SNPs on an amazingly robust genotyping platform. The majority of HapMap common variation is either directly genotyped or “tagged” by a proxy SNP extremely well by SNPs on HumanHap300, HumanHap550 and HumanHap650Y BeadChips. In addition, under various disease models, these whole genome genotyping products provide high power to detect relatively small gene effects associated with complex disease traits with typical whole genome association study sample sizes.

When I worked at Illumina, my colleagues and I wanted to see if there were areas where we could expand coverage of the genome to improve power even more for whole genome association studies. The HapMap has done an extraordinary job of providing the research community with validated SNPs and knowledge of haplotype structure and diversity across several populations. However, we noticed some genes were not covered well by HapMap SNPs—either by not having any SNPs, or having a low density of SNPs across the gene region. We chose to focus on adding more SNPs to gene regions because the majority of all known variants associated with both simple Mendelian and complex disease traits have been either near (i.e., within 10kb) or within coding regions. Because genes and evolutionarily conserved regions (ECRs) can be so important to disease phenotypes, we targeted a high density of SNPs across all RefSeq transcript regions and in putative ECRs of the genome. We also added nearly 25,000 non-synonymous SNPs (nsSNPs), SNPs that change the amino acid sequence of the resulting protein. Non-synonymous SNPs are an important class of high-value SNPs because nsSNPs directly affect protein structure and/or function, and are therefore expected to have a major impact on phenotype.

Illumina’s HumanHap products had focused on assaying SNPs, an extremely important class of variation to study as it encompasses a major class of variation in the genome. However, we wanted to expand the class of variation to include more known and to-be-discovered copy number variants (CNVs) in the genome. Instead of just creating an expanded whole genome genotyping tool, we envisioned creating a comprehensive whole genome DNA analysis tool. CNVs are segments of DNA, ranging in size from kilobases to megabases, that include deletions, insertions, duplications, or other complex patterns. The existence of CNVs in the genome has been known for years. However, investigating the role of CNVs and their involvement with disease on a genome-wide level, with high resolution, has not occurred until very recently. CNVs are an important class of variation because they can influence gene expression, disrupt genes, and alter gene dosage. Genomic profiling for chromosomal aberrations such as amplifications and deletions has been a crucial element of cancer biology, as well as the study of congenital disorders. Genomic profiling for CNVs associated with all types of complex disease traits is now emerging, as demonstrated by the recent discovery of *de novo* CNVs associated with autism¹. Because CNV is such an important class of variation to interrogate, we targeted both public (described by the Toronto Database of Genomic Variants) and novel (created in collaboration with deCODE genetics) CNV regions of the genome with SNPs and non-polymorphic probes. We also employed a “picket fence” approach of evenly spacing SNPs across the genome to enable the discovery of new CNV regions.

Even though we focused content primarily in genes and CNV regions, we did add other classes of content to the Human1M BeadChip. With HapMap release 21 available, we chose additional tag SNPs in all HapMap populations to fill the small number of bins that were not covered. We created a higher density of SNPs in approximately 200 genes involved with drug absorption, distribution, metabolism,

“As I have now ventured back into the world of research myself, I am thrilled to continue to be a part of this exciting time in human genetics.”



and excretion (ADME), since they are an important class of genes for pharmacogenetic studies. We also targeted genes in the major histocompatibility complex (MHC), a region of the genome implicated in a large number of auto-immune and inflammatory diseases. The MHC region is gene-dense, with a high proportion of genes involved in the immune system, including the human leukocyte antigen (HLA) membrane glycoproteins that mediate T-lymphocyte signaling.

I would never have guessed that in such a short amount of time, scientists would be able to genotype more than a million SNPs so easily and affordably. The comprehensive Human1M BeadChip will be a great tool enabling researchers to discover and dissect those elusive complex disease traits. I look forward to continuing to read about—and hopefully contribute to—all the amazing discoveries that are being made with the existing and future whole genome genotyping products. As I have now ventured back into the world of research myself, I am thrilled to continue to be a part of this exciting time in human genetics. At Scripps Genomic Medicine, my colleagues and I will be using Illumina genotyping products, including the Human1M, for various common risk allele discovery projects. For rare risk alleles, we will be re-sequencing candidate regions of the genome in many cases and controls and we are developing methods to analyze these data. Since Scripps Genomic Medicine's main goal is bringing these discoveries back to the clinic, we will also be conducting gene-based clinical trials and large prospective studies aimed at determining risk profiles based on combinations of specific risk alleles.

REFERENCES

- (1) Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C et al. (2007) Strong association of de novo copy number mutations with autism. *Science* 316: 445-449.

ADDITIONAL INFORMATION

To learn more about the entire portfolio of Illumina DNA Analysis products, visit our website at www.illumina.com under Products & Services.

We are committed to providing you with the content you want as a member of the Illumina community. Please email us with comments and suggestions for topics at icomunity@illumina.com.

FOR RESEARCH USE ONLY

© 2007 Illumina, Inc. All rights reserved.

Illumina, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, and CSPro are registered trademarks or trademarks of Illumina. All other brands and names contained herein are the property of their respective owners.
Pub. No. 370-2007-018 26Jul07

