Picking up Where HapMap Left Off: The Next Generation of GWAS

Genetic analysis tools that will fuel a new era of discovery.

Introduction

The human genetics community is facing a quantum leap in the extent of known variation across diverse human populations. Using this information, the next generation of genetic analysis tools will enable the exploration of new hypotheses and usher in a new era of discovery. With proven intelligent maker selection and the ability to capture extensive genomic coverage, next-generation genotyping arrays will deliver the power needed to fuel new discoveries and uncover a greater understanding of how genetic variation contributes to human health and disease.

The International HapMap Project

Over the better part of the last decade, the International HapMap project revolutionized our understanding of genetic variation across the human genome. A collaboration that included researchers from institutions around the world, the project sought to describe patterns of genetic similarities and differences across four representative populations from Asia, Europe, and Africa. With this information, it was expected that researchers could identify genetic associations to human traits and diseases that may ultimately lead to a better understanding of disease mechanisms and new therapeutic interventions.

Through the HapMap project, researchers were able to map human haplotype blocks, chromosomal regions that contain a group of linked alleles which are transmitted together. Alleles within a haplotype block are correlated to one another, so knowing the identity of one allele allows a researcher to simultaneously identify all of the alleles within that region. The strength of the correlation between these alleles is defined by linkage disequilibrium (LD), which is commonly measured by the term r^2. At perfect LD, r^2 is equal to 1, meaning that two alleles can serve as exact proxies for one another. For Illumina genotyping microarrays that arose from the HapMap data, the phenomenon of LD proved to be a great advantage. Rather than having to genotype every allele across the genome (which is unfeasible), only a fraction from each haplotype block were needed to know the identity of the entire block with a high degree of certainty.

Armed with these new whole-genome microarrays, researchers began genome-wide association studies (GWAS), with the goal of identifying causative alleles for a trait or disease by evaluating the genotypes of a large number of DNA samples. The use of GWAS to uncover these variants has proven immensely successful, identifying over 1,900 variants in 427 publications[1] in a few short years.

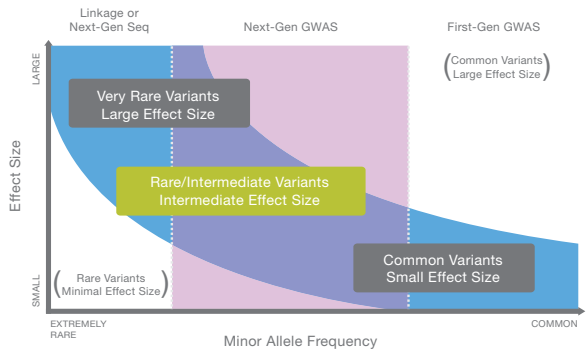
The fundamental premise behind these GWAS was that common diseases, such as diabetes or high blood pressure, were caused by common genetic variants, alleles that have high frequencies within the afflicted population (generally > 5% minor allele frequency, or MAF). By and large this proved to be true, but these common variants taken individually or in combination proved to confer only a small increase in disease risk and explained only a small proportion of the heritability—the extent that genetic variation contributes to phenotypic variance in a population[2]. The "missing heritability" not explained by common variation will be a large focus in the next generation of GWAS.

Microarrays used during the first generation of GWAS were limited by the information available in the HapMap universal reference database, which by the end of the project contained ~3.5 million variants, targeting for MAFs > 5%. As we know now, these tools only captured a small portion of the entire spectrum of human genetic variation. To increase the power of GWAS, the research community needed to begin looking at those variants that occur less commonly (down to 1% MAF).

The 1000 Genomes Project

To find these rarer alleles within human populations, researchers needed the high resolution offered by next-generation sequencing. Detecting variants that occur at such low frequencies requires a considerable number of samples to be sequenced to obtain the necessary statistical power. Even considering such an effort in 2002 when the HapMap project started would have been unthinkable. At that time the human genome had just been sequenced for the first time, an endeavor that took roughly 13 years and billions of dollars to complete[3]. Fast forward to 2010 and a complete human genome can now be sequenced in just over one week at a small fraction of the cost (less than $10,000 at

Figure 1: The Spectrum of Genetic Variation



While every phenotypically relevant variant can be plotted somewhere along these two axes, the region defined by the blue swath is most relevant for the next generation of GWAS. Until recently, genetic analysis tools only provided access to the very far left blue region (via traditional linkage mapping and next-generation sequencing) and the far right blue region (via first-generation GWAS tools). With the availability of 1kGP data, researchers now have the ability to expand their studies into the largely unexplored middle blue region, defined by rare/intermediate frequency variants of intermediate effect size.

30x coverage using Illumina's HiSeq™2000 sequencing platform). This rapid evolution in DNA sequencing has made the next generation of GWAS possible.

By leveraging these advances in next-generation sequencing technology, the 1000 Genomes Project (1kGP) will vastly extend the catalog of human variation started by HapMap. Like the HapMap project, the 1kGP involves the collective efforts of research institutions around the globe who will deliver human genome sequences for about 2,500 individuals[4]. At such a high level of sampling, the sequence data from the 1kGP will provide GWAS researchers superior power to detect variation across the genome.

The implications of this information are compelling. Where HapMap-derived genetic analysis tools focused on common variants with small effect sizes, the next generation of GWAS tools will enable researchers to examine variants that occur with rarer frequency and have intermediate to large effect sizes. These are the variants that may explain the missing heritability, the ones that occur less commonly but make more substantial contributions to the disease or trait phenotype (Figure 1).

The Next Generation of Whole-Genome Microarrays

Results from the 1kGP are powering a new generation of Illumina BeadChips that provide an unprecedented level of genomic coverage, a key metric for any whole-genome array. Genomic coverage indicates the percent of variation captured on the array at an LD of $r^2 > 0.8$. Prior to the 1kGP, coverage statistics for whole-genome arrays were based on the catalog of variants known from the HapMap project. In light of the more comprehensive data set produced from 1kGP, the reference point for coverage statistics must be adjusted.

Illumina whole-genome arrays have historically maintained an industry-leading coverage specification of at least 80%. To accomplish this level of coverage, Illumina scientists carefully selected the most highly correlated markers from each haplotype block across the genome. Considering that the 1kGP is drastically increasing the amount of known human variants, the number of markers on arrays needed to capture that variation will need to increase dramatically as well. Illumina estimates that approximately 5 million markers will be needed to maintain > 80% genomic coverage of the genome. To meet that need, the Infinium Omni5 BeadChip, containing roughly 5 million markers, will be released by Illumina in late 2010 when the complete data set from the 1kGP is expected. This array will capture human variants at minor allele frequencies down to 1%.

While the Omni5 won't be available until later this year, content from the first phase of the 1kGP has already made its way onto Illumina BeadChips. As announced at the 2009 ASHG conference, Illumina is providing a product roadmap that allows customers to begin next-generation GWAS now using any of the current Omni microarrays (Omni1, OmniExpress, or Omni2.5) and build up to the full 5 million variants through supplemental arrays that will be released as new data become available. Of particular interest among the current Omni arrays is the Omni2.5. It is the first product to be optimized around 1kGP data, providing ~2.5 million markers that capture variants down to MAF 2.5 % and deliver the highest genomic coverage rates across diverse populations (CEU: 82%, CHB/JPT: 83%, and YRI: 74%). Preliminary estimates show that it offers up to 50% greater coverage than any previous commercially available whole-genome array.

With this next generation of microarrays, researchers are now empowered to explore many new hypotheses. Coverage of variants with low minor allele frequencies will help reveal the role of rare variation on trait and disease phenotypes. Dense spacing of markers across the genome will provide superior support for copy number variation studies. Even more comprehensive coverage of common variants (> 5% MAF) will enable researchers to more definitively assess genetic associations with traits and disease.

The Marriage of Arrays and Sequencing

For the future of genomic research, investigators now have two powerful technologies at their hands—whole-genome microarrays and next-generation sequencing. The beauty of these tools is that they complementary and offer researchers a number of options for intelligent experimental design. Few technologies have risen to prominence as fast as next-generation sequencing, a transcending analytical tool that has allowed researchers to publish ground-breaking studies at an unprecedented rate. While the value of this technology goes unquestioned for genomic discovery, microarrays are still the optimal choice for validation and association studies where sample size and throughput are critically important. For known genomic content (such as that from the 1kGP), microarrays can deliver higher-quality genotyping information for a large number of samples quickly and economically. Conversely, sequencing is capable of providing data that arrays cannot, such as for un-assayable regions of the genome or for very rare variants that arrays cannot capture.

The joint application of these technologies offers a powerful approach to uncover important data for new biological discoveries. For example, whole-genome arrays can be used to identify regions of the genome that harbor trait- or disease-related variants. Using that information, additional variation in the region can be identified through targeted re-sequencing efforts. These variants can in turn be deployed on targeted custom arrays that allow researchers to perform validation studies on a large number of samples. The possibilities are limitless and are only bound by the researchers' creativity and vision.

A new era of discovery has begun. The next generation of GWAS will be driven by revolutionary genetic analysis tools that provide researchers unprecedented access to the human genome, enabling a range of new hypotheses that will catalyze the next wave of discovery.

References

1. http://genome.gov/gwastudies
2. Manolio T, Collins F, Cox N, Goldstein D, Hindorff L, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753
3. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
4. http://1000genomes.org

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

© 2010 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, GoldenGate Indexing, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, CSPro, GenomeStudio, Genetic Energy, HiSeq, and HiScan are registered trademarks or trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners.
Pub. No. 070-2010-015 Current as of 07 June 2010

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA