Empowering GWAS for a New Era of Discovery

Intelligent SNP selection and data from the 1000 Genomes Project combine to enable the next generation of genome-wide association studies.

Introduction

The use of genome-wide association studies (GWAS) to identify regions of the genome most likely to harbor variants that contribute to human traits and disease has proven immensely successful, identifying over 1900 variants in 427 publications[1] in a few short years. These successes were made possible, in large part, by the presence of a universal reference data set developed through the HapMap Project (http://hapmap.ncbi.nlm.nih.gov/). This data set has helped researchers understand linkage disequilibrium (LD) across a diverse set of samples and enabled the development of new bioinformatic techniques such as genotype imputation. Illumina has used this reference data to intelligently select powerful markers that capture the majority of variation across the genome for association studies. While immensely useful, this resource only contained ~3.5M markers, targeting minor allele frequencies (MAF) greater than 5%. As a result, the genotyping tools and studies developed around this data set are only able to capture a faction of the true spectrum of genetic variation (Figure 1).

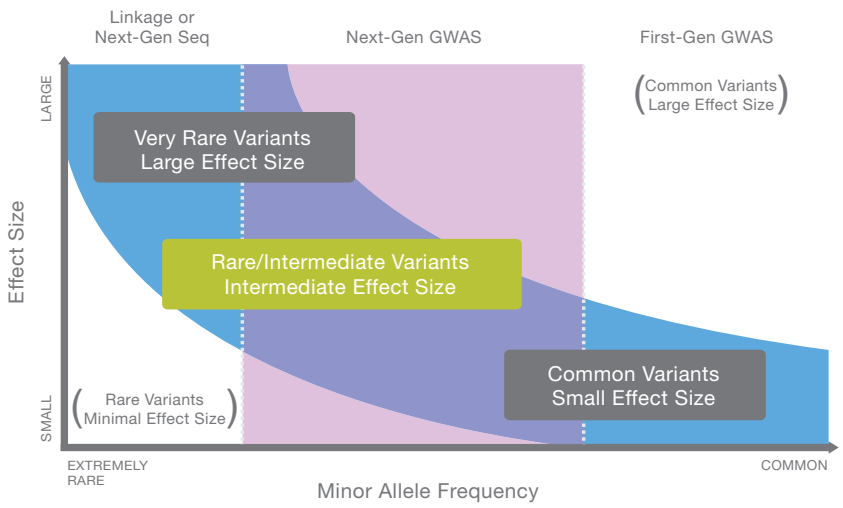
By leveraging advances in next-generation sequencing technology, the 1000 Genomes Project (1KGP) will vastly extend the catalog of human variation started by HapMap, increasing the scientific community's understanding of the full spectrum of variation in human populations. The extent of variation uncovered by the 1KGP will reset the baseline for how to assess genomic coverage (the fraction of known variants across the genome captured by the markers on the array at a certain level of LD) of all past and future human genetic studies.

Illumina's technology will continue to enable fresh discoveries as researchers harness 1KGP data to explore new hypotheses, including the role of rare and intermediate variation in human traits and disease (Figure 1). The next generation of Illumina GWAS products will incorporate proven intelligent SNP selection and high genomic coverage, along with the flexibility and data quality of the Infinium® HD Assay to provide researchers the tools to make the most of the information 1KGP will deliver.

Spectrum of Variation

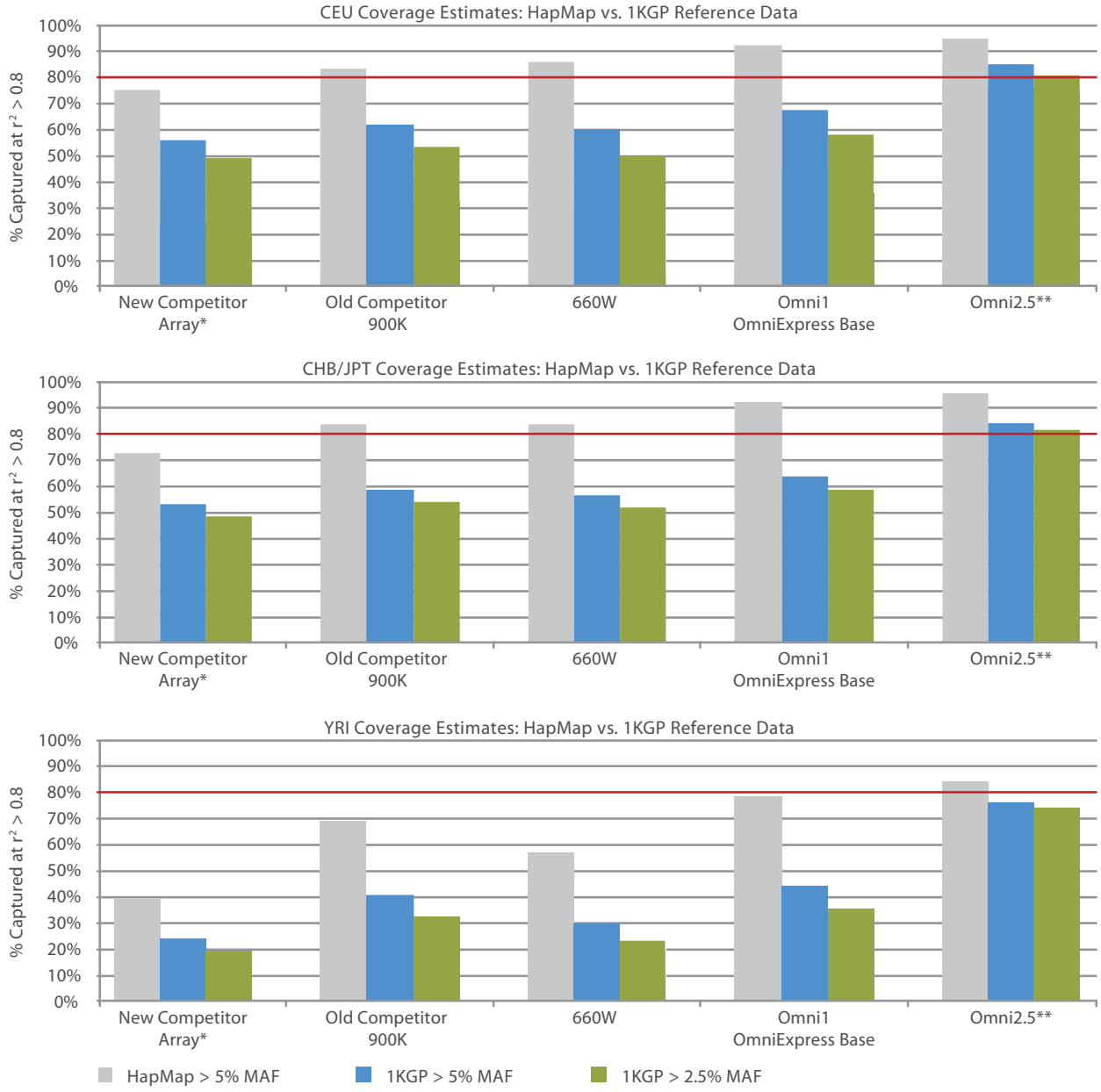
The extent of variation that researchers can expect to identify from the 1KGP is limited by the number of samples sequenced from a given population. The available 1KGP sequence data from 60 unrelated individuals from three diverse world populations (Caucasians, Yourban, and South East Asian (Chinese and Japanese)) provides sufficient power to detect ~80% of all variation down to approximately 2.5% MAF in each population. Currently, the project has identified ~17M variants (Table 1). The final data set from the 1KGP is expected to include sequence data from much larger sample size (approximately

Figure 1: Covering the Full Spectrum of 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. The variants that fall outside of this region would have already been identified through first-generation GWAS tools (upper right area), or their relevance to understanding disease is negligible (lower left area). 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 and Illumina next-generation microarrays, researchers now have the ability expand their studies into the largely unexplored middle blue region (highlighted in pink), defined by rare/intermediate frequency variants of intermediate effect size.

Figure 2: Reassessing Genomic Coverage of Commercially Available Microarrays



These graphs show the percent of variation captured for five GWAS microarrays with respect to three different reference data sets in samples from three populations—CEU, CHB/JPT, YRI[†]. The first four microarrays listed were developed around HapMap reference data, while the Omni2.5 is the first array to include broad coverage of 1KGP data. The grey bars show coverage with respect to HapMap data at a MAF cutoff of 5%. When coverage estimates are evaluated with respect to 1KGP data at 5% MAF, there is a precipitous drop (blue bars). If the reference data set is expanded to include markers down to 2.5% MAF, the coverage estimates drop even further (green bars). In light of the data coming from 1KGP, it is now apparent that Hapmap-based arrays provide little to no visibility into a large portion of the genome. For example, with respect to the CEU population, 40–50% of the genome is not captured by these arrays. In contrast, the Omni2.5 has been designed to be maximally powerful across all three reference data sets, including MAF 2.5% in 1KGP data.

* Represents product base content

** Estimated from the preliminary marker list

[†] CEU: CEPH Utah; CHB: Chinese Han Bejing; JPT: Japanese Toyko; YRI: Yoruban

Table 1: Reference Data Sets for Human Genetic Analysis

Project	Year	Approximate Cumulative SNPs	Tag SNPs Needed to Maximize Coverage	Lower Limit of Allele Frequency	% Variation Tagged ($r^2 > 0.8$)
HapMap	2003–2007	3.5M	0.6M	5%	> 90%
1KG Pilot Project	2008–2009	17M	2.5M	2.5%	~80%
1KG Main Project	2010	35M (estimated)	5.0M	1%	> 90%

The HapMap Project delivered a total catalog of approximately 3.5M variants, which were referenced for almost all human genetic studies over the past decade. In its first year, the 1KGP (the pilot project) has already expanded the spectrum of known variation almost 5-fold over HapMap. The main project is estimated to discover an additional 18M variants for a total of ~35M variants by the end of 2010. Analysis of the 1KGP pilot project data shows that approximately 2.5M tag SNPs are needed to capture ~80% of the full 17M variants identified down to 2.5% minor allele frequency (MAF). Based on these analyses, Illumina scientists project that approximately 5M tag SNPs will be needed to capture 90% of variation down to 1% MAF within a given population (assumes ~400 samples sequenced at 4× coverage within a given population).

400 Caucasian individuals, for example). Sampling at this level allows sufficient power to detect 80% of all variation down to ~0.36% MAF and 100% of all variation at 2.5% MAF*. In samples with other ancestries, the spectrum of variation identified will be similar depending on the number of individuals sequenced from that population. In aggregate, the 1KGP is expected to identify approximately 35M variants across many diverse world populations (Table 1).

Though the spectrum of variation is expanding greatly with the 1KGP, the approach to performing a GWAS remains the same. Researchers will still need to consider platform and SNP selection to maximize genomic coverage and LD, sample size, and genotyping quality to make critical decisions that will impact the downstream success of their studies.

The Power of Arrays and Tagging

Because directly genotyping every variant in the human genome is neither necessary nor cost effective, Illumina Infinium DNA Analysis BeadChips employ a proven tag SNP approach that has been successfully applied in hundreds of common variant GWAS studies[4]. The power of a tag SNP approach stems from the inherent correlation among markers that form haplotype blocks. This phenomenon allows the selection of one highly correlated marker to serve as a proxy for a number of additional highly correlated markers across the genome. Therefore, with the careful selection of tag SNPs, an array can deliver information on a much greater number of markers than those physically genotyped on the array itself. The phenomena of LD and haplotype blocks are inherent to the physical properties of DNA and recombination patterns, so as the 1KGP delivers data down to 1% MAF, the tag SNP approach will continue to be the most powerful and efficient approach to SNP selection.

The correlation between SNPs is commonly described by r^2, where a high r^2 between two SNPs indicates high correlation, making these SNPs good proxies for each other. At a maximum $r^2 = 1$, two SNPs are in perfect LD and can serve as exact proxies for each other; thus, only one SNP needs to be genotyped to know the genotype of the other with certainty. At any given r^2, different commercial genotyping products offer a range of genomic coverage levels, and therefore have different levels of power to detect association in a given sample size. Disease- and trait-specific parameters, including penetrance and effect size of risk variants, will also impact the power for an association study, but these factors are inherent to the phenotype of interest and independent from the choice of genotyping platform. Illumina DNA Analysis products offer unparalleled genomic coverage by leveraging the tag SNP approach, providing the highest average r^2 values in the industry and maximizing the likelihood of finding true associations for a given phenotype.

Prior to 1KGP, coverage estimates for all commercially available chips were commonly referenced to the HapMap project (MAF > 5%). In light of the more comprehensive data set produced from 1KGP, the reference point for coverage statistics must be adjusted. Figure 2 shows the estimated coverage of 1KGP variants down to 2.5% MAF for five different whole-genome genotyping products and three different representative populations. For reference, the coverage statistics with respect to HapMap data are also included. When the statistics for HapMap at 5% MAF are compared to 1KGP at 5% MAF, a marked decrease in coverage is observed. In light of the 1KGP data, it is apparent that microarrays developed around HapMap data provide far less coverage of the genome than previously thought. However, the next generation of microarrays, starting with Illumina's Omni2.5, will be largely derived from 1KGP data, allowing them to provide high genomic coverage across both data sets (Figure 2).

Sample Size and Genomic Coverage

GWAS typically rely on statistical analyses of the allele frequencies in the cases (individuals with the phenotype of interest) versus the controls (individuals without the phenotype of interest). If the sample size (collectively, the cases plus controls) is large enough, statistically significant associations between specific alleles and a phenotype can be identified. Though power to detect true associations is a function of many different factors, many of which are inherent to the trait of interest, researchers can ensure success of their experiment by both maximizing sample size and level of genomic coverage (Figure 3).

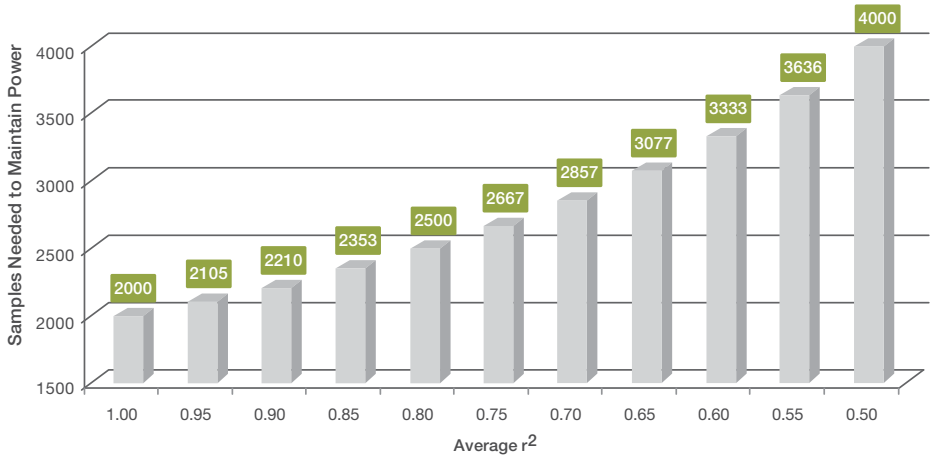
To effectively calculate the power of a GWAS to detect associations, the risk allele frequency, prevalence, and genotype relative risks must be considered in addition to sample size. These are all inherent to the phenotype and are seldom known with certainty when planning a study. If these variables are held constant, while the LD with the risk loci and sample size are varied, it is clear to see the average power delivered by commercial arrays varies dramatically depending upon the level of genomic coverage.

High-Quality Data

High-quality data is critical to the success of GWAS and rapid time to publication. Recent studies have shown that high error rates, non-random missing data, and low call rates can dramatically increase the num-

* Analysis excludes all variants only seen once in the sample (singletons), as these represent the class of variation most likely to include false positives

Figure 3: Number of Samples Needed to Maintain Power of a 2000 Sample Study



The power of a 2000 sample study was calculated under perfect LD ($r^2 = 1$). As average LD is decreased, the number of samples needed to maintain the same level of power increases to ~4000 total samples at an $r^2 = 0.5$.

ber of false positive results[3]. False positive results can create a challenge for the validation of significant SNPs and can ultimately delay identification of true regions harboring causative mutations. Furthermore, the presence of false positives can result in publication of spurious findings that cannot be replicated in other studies. Illumina products have historically shown exceptionally high data quality with regard to call rates (average > 99%), reproducibility (> 99.9%), and low sample redo rates. Even a one or two percent reduction in call rates dramatically increases the number of false positives (and decreases genomic coverage estimates), requiring time-consuming and frustrating extra analysis as well as expensive follow-up studies on erroneous associations.

The high accuracy and call rates are attributes of the powerful Infinium HD Assay and proprietary BeadArray™ technology. With 50-mer oligonucleotide probes, the Infinium HD Assay has a very high selectivity for the target DNA fragment even in very complex solutions. In addition, a separate enzymatic labeling process using a single-base extension ensures high specificity for the allele. The extension and dye incorporation occur when template DNA has hybridized to the target oligo. This dramatically reduces the background signal, which increases the signal-to-noise ratio for more accurate cluster separation and genotype calling. Illumina has developed a stable allele-calling algorithm, using high feature redundancy and two-color labeling, resulting in consistently high call rates and reproducibility between, and within, products.

CNV Analysis

Copy number variation (CNV) is a significant contributor to the genetic basis of human traits and disease. Carefully designed array content on Omni microarrays provides dense genomic coverage that minimizes large gaps, allowing researchers to evaluate CNVs across the genome. The high signal-to-noise ratios and low overall noise levels produced by the Infinium HD assay are ideal for precise copy number analysis. With Omni microarrays, researchers have a powerful tool to explore new hypotheses about the role of copy number variants.

Genotyping Controls Database

Illumina also offers the first industry-hosted genotyping controls database, iControlDB. Illumina customers can now access nearly 10,000 control samples that have been donated by researchers using Illumina's technology for SNP genotyping. The Illumina iControlDB provides investigators with an extensive set of control samples to validate their genome-wide association studies and boost the power of their studies.

Summary

The human genetics community is facing a quantum leap in the depth and extent of known variation across diverse human populations. Illumina is using this explosion of data to develop the next generation of whole-genome genotyping products, which will enable the exploration of new hypotheses and usher in a new era of discovery. With proven intelligent SNP selection and the ability to capture extensive genomic coverage with up to 5 million markers per sample, these next-generation 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.

References

1. http://www.genome.gov/gwastudies/ (accessed on 11/11/09)
2. http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html
3. http://www.illumina.com/Documents/products/whitepapers/whitepaper_gwas_array.pdf
4. http://www.illumina.com/publications/list.ilmn?subCat=10

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. 370-2010-007 Current as of 15 April 2010

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA