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 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 fraction 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...
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 Beijing; JPT: Japanese Tokyo; YRI: Yoruban
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).

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-

---

**Table 1: Reference Data Sets for Human Genetic Analysis**

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

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 4x 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.

---

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