The Power of Intelligent SNP Selection

The Infinium Assay provides the freedom to design the most powerful genotyping panels. Targeting the most informative SNP loci supports the most efficient study designs. This power provides the fastest path to discovery and publication.

Introduction

Every single nucleotide polymorphism (SNP) marker on Illumina Infinium® DNA Analysis BeadChips has been selected by Illumina scientists based on the high information content and genomic coverage it provides. Employing rationally selected SNPs rather than randomly chosen SNPs greatly improves the power to detect association and drastically reduces the number of samples required for successful genome-wide association studies (GWAS). This technical note explores this impact of array power on study efficiency.

Illumina's strategy has proved successful at optimizing several parameters critical to successful GWAS. Metrics for genomic coverage, array efficiency, genic coverage, call rate, and call accuracy were compared between different array platforms in a white paper[1] and in a publication by University of Michigan researchers[2].

SNPs and Genome-Wide Association Studies

It is estimated that more than 10 million SNPs exist in the human genome3. SNPs are known to contribute to population diversity and phenotypic differences between individuals, and cause predispositions to diseases. Genome-wide association studies hold the promise of identifying SNPs associated with a certain phenotype of interest as well as those that can serve as diagnostic markers[4–8]. Genes in the vicinity of SNPs that appear to cause the phenotype could be qualified as new drug targets.

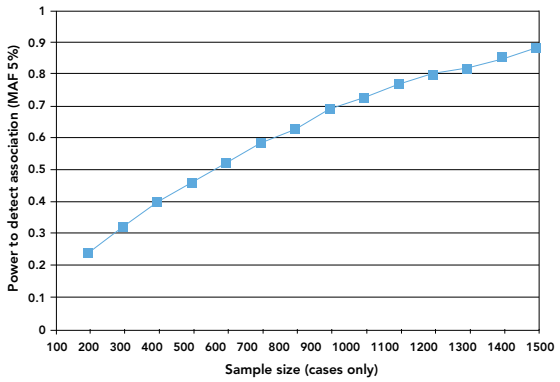
To successfully identify candidate SNPs using genome-wide association analysis, the researcher needs to consider sample size, multiple testing correction, SNP selection (to maximize genomic coverage and linkage disequilibrium [LD]), and genotyping quality.

Sample Size and Multiple Testing Correction

Genome-wide association studies rely on statistical analyses of the allele frequencies in the cases (individuals with the disease phenotype) versus the controls (individuals without the disease phenotype). If the sample size is large enough, a statistically significant association between a specific allele and a phenotype can be determined. The sample size, and therefore the power to detect the association, is largely dependent on the disease frequency and the odds ratio of the phenotype in the population selected. The odds ratio is defined as the odds of an experimental patient suffering an adverse event relative to a control patient.

The large sample numbers inherent to genome-wide association studies require researchers to statistically correct for false positives (type I errors) that can occur by chance with multiple sample experiments. Usually a Bonferroni correction is used, but less conservative corrections such as the false discovery rate (FDR) are also employed. Knowing the phenotype frequency and odds ratio and calculating a statistical correction for false positives allows the researcher to determine the required sample size to detect a statistically significant association.

Figure 1: Sample Sizes Required at Different Powers of Detecting Association



The power to detect association determines the sample size required for cases. This example is based upon a disease with an odds ratio of 1.3 and a minor allele frequency (MAF) of 5%. It assumes that the disease SNP is measured directly.

Comprehensive Genomic Coverage with Tag SNPS

Because studying every SNP in the human genome is not cost-effective with current microarray technology, Illumina Infinium DNA Analysis BeadChips employ a subset of SNPs called tag SNPs that can be used as proxies for all common SNPs (minor allele frequency ≥ 5%) in the genome. Using tag SNPs for a genome-wide association study allows the investigator to maximize information content and minimize sample size without losing the power to detect genetic association.

More than 2.2 million common SNPs (minor allele frequency ≥ 5%) in four ethnic groups (Caucasians in Utah, Han Chinese in Beijing, Japanese in Tokyo, and Yoruba in Ibadan, Nigeria) were genotyped as part of the International HapMap Project. From this data set, Illumina scientists select the SNPs with the optimal minor allele frequency and the best genomic coverage to serve as tag SNPs for each population.

The selection process is dramatically enhanced by the Infinium Assay, which allows unrestricted access to SNPs throughout the genome and enables even coverage of tag SNPs across the genome. This tag SNP approach represents the foundation for content selection for Illumina genotyping products and the resulting industry-leading level

of genomic coverage. In this context, genomic coverage is defined as the number of SNPs that are in LD with a reference SNP set. Illumina scientists use as the reference SNP set all common SNPs typed by the HapMap Project.

A high r^2 between two SNPs indicates high correlation, making these SNPs good proxies for each other. At a maximum r^2 of 1, two SNPs are in perfect LD and can serve as pure proxies; thus, only one SNP needs to be genotyped to know the genotype of the other. At any given r^2, different genotyping products have a certain genomic coverage, and therefore a certain power to detect association at a given sample size and odds ratio of the disease. Illumina DNA Analysis products offer unparalleled genomic coverage using tag SNPs with the highest average r^2 values in the industry.

Illumina scientists have calculated the genomic coverage provided by the Infinium DNA Analysis BeadChip product line as well as the coverage offered by microarrays relying on randomly selected SNPs to interrogate genetic variation (Figure 2).

Comparing the Genomic Coverage of the Infinium

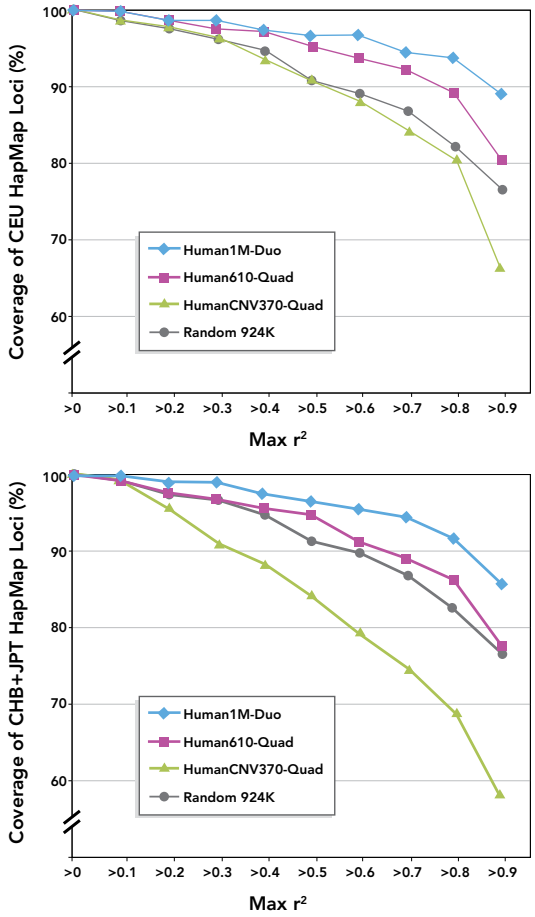
Human1M-Duo, Human610-Quad, and HumanCNV370-Quad BeadChips, all of which use tag SNPs, to the genomic coverage of 924,000 randomly selected SNPs, it is clear that Illumina products offer unmatched genomic coverage at greater power for genome-wide association studies (Figure 2).

Tag SNPs and Smaller Sample Size

Since current genome-wide association technologies rely on genotyping SNPs near a disease locus, the power to detect association depends on the linkage disequilibrium of the genotyped markers with the adjacent disease-causing SNP. Any reduction in power can be overcome by increasing the sample size. However, using a tag SNP approach optimizes the power at any sample size, reducing cost and time. This adjustment is based on a simple calculation which takes the expected linkage disequilibrium between the marker and the disease locus into account. The required sample size, determined by the power calculation and illustrated in Figure 1, must be multiplied by $1 / r^2$ to derive the actual number of samples that need to be genotyped to achieve the same power as when the disease locus is directly assayed.

For a desired genomic coverage of 95%, Table 1 shows the necessary increase in sample size to maintain the power to detect association within the top 95% of tagged SNPs for a Caucasian population. To maintain power within the bottom 5% requires even larger sample sizes. This calculation indicates that ~600,000 Illumina-designed tag SNPs require significantly fewer samples than 924,000 random SNPs. The Human1M-Duo BeadChip, with over one million tag SNPs, requires half the sample size compared to 924,000 randomly selected SNPs to maintain power in the top 95% of the tagged SNPs, demonstrating that Illumina's tag SNP approach provides more power for detecting genetic associations throughout all possible SNPs.

Figure 2: Genomic Coverage Using Tag SNPS Versus Random SNPS in Caucasian and Asian Populations



Max r^2 indicates the minimum LD with a tag SNP when calculating the fractions. Corresponding genomic coverage is shown for the Illumina Infinium HumanCNV370-Quad, Human610-Quad, and Human1M-Duo BeadChips versus a microarray designed with 924,000 randomly selected SNPs. Note that these genomic coverage values are averaged over the chromosome and may vary based on the specific genomic location.

High-Quality Data

Recent studies have also shown that low genotype data quality (i.e., high error rates), non-random missing data, and low call rates can increase the number of false positive results. False positive results can make the validation of significant SNPs challenging and can ultimately delay the identification of real causative mutations or cause the publication of false findings that cannot be replicated in other studies. Illumina products have historically shown exceptionally high data quality with regard to call rates (> 99%), reproducibility (> 99.9%), and low redo rates. Even a one or two percent reduction in call rates dramatically increases the number of false positives, requiring time-consuming and expensive follow-up studies on erroneous associations.

The high accuracy and call rates are attributes of the powerful Infinium Assay and proprietary BeadArrayTM technology. With 50-mer oligonucleotide probes, the Infinium Assay has a very high selectivity for

Table 1: Comparison of Power of Detection and Sample Size using Tag SNPS Versus Random SNPS

HumanCNV370-Quad	Human610-Quad	Human1M-Duo	Random 924K	Comments
2,606	2,606	2,606	2,606	Required sample size (cases and controls) for 80% power at an odds ratio risk 1.3 (MAF 5%)
0.35	0.55	0.8	0.4	r^2 at which 95% of SNPs are tagged*
912	1,433	2,085	1,042	Effective sample size (r^2 × sample size)
43%	59%	70%	48%	Power at effective sample size
2.86	1.82	1.25	2.50	Sample multiplier (1 / r^2)
7,446	**4,738**	**3,258**	**6,515**	**Total samples needed to maintain the same power across all SNPs (sample multiplier × sample size)**

*based on the coverage listed in Figure 2 for CEU HapMap samples

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, reducing the background signal dramatically and therefore increasing the signal-to-noise ratio for more accurate cluster separation and genotype calling. Illumina has developed a stable allele-calling algorithm, utilizing high feature redundancy and two-color labeling, resulting in consistently high call rates and reproducibility between and within products.

Genotyping Controls Database

Illumina also offers the first industry-hosted genotyping controls database, iControlDB. Combined with the database of Genotype and Phenotype (dbGaP), the scientific community 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, allowing researchers to significantly reduce the time and cost of their genotyping studies.

Summary

Illumina's tag SNP approach in tandem with the powerful Infinium Assay delivers the high-quality genotyping data, information content, and genomic coverage necessary for identifying susceptibility loci in human disease. In addition, using intelligently selected tag SNPs drastically reduces the number of samples needed for confident detection of candidate regions in any genome-wide association study, reducing the overall cost per study. Illumina whole-genome genotyping solutions provide industry-leading levels of accuracy, flexibility, and affordability.

References

1. http://www.illumina.com/downloads/GWASArrayWhitePaper.pdf
2. Li M, Li C, Guan W (2008) Evaluation of Coverage Variation of SNP Chips for Genome-Wide Association Studies. Eur J Hum Genet 16: 635-643. Available for free download from http://www.illumina.com/GWASArray.
3. Kruglyak L and Nickerson DA. (2001) Variation is the spice of life. Nat Genet 27(3):234-236.
4. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, et al. (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 39(5): 596-604.
5. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445(7130): 881-885.
6. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, et al. (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 39(5): 645-649.
7. Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, et al. (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 39(5): 631-637.
8. Schymick JC, Scholz SW, Fung HC, Britton A, Arepalli S, et al. (2007) Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol 6(4): 322-328.

Additional Information

To learn more about Illumina Whole-Genome Genotyping Products and Illumina iControlDB, please visit www.illumina.com.

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. 370-2007-016 Current as of 28 July 2010

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA