Applications of Next-Generation Sequencing in Genetic Epidemiology

Francisco M. De La Vega*, Jon Sorenson*, Fiona Hyland*, Wonkuk Kim¶, Stephen J. Finch¶, and Derek Gordon†. Applied Biosystems, 850 Lincoln Centre Dr., *Foster City, CA 94404, ¶Stony Brook University, Stony Brook NY; †Rutgers University, Piscataway, NJ, USA.

ABSTRACT

Next-generation sequencing (NGS) platforms produce 10-100's of millions of short reads (25-50bp) in a single run. This allows to use this platforms for high-throughput resequencing of DNA samples to discover and genotype variants simultaneously. We sought to understand the relationships between number and length of reads and per-base sequencing error to the overall genotyping accuracy in NGS through simulations and experimental results from the Applied Biosystems SOLiD™ system. We also studied the power to detect genetic association in situations where the disease variant is initially not typed, but can be discovered and typed by following-up by sequencing leading candidate regions in cases and controls. We discover that while significant power loss may occur when the disease variant is absent from the screening phase (as is more typical), power can be recovered by full SNP ascertainment by resequencing of cases and controls, presaging a significant role of NGS in the identification of disease-causing genes.

INTRODUCTION

An important application for NGS is the resequencing of targeted regions for the identification of causal disease susceptibility alleles. The ultimate goal would be to genotype by sequencing full disease association cohorts for identification of disease susceptibility alleles. Here, a shotgun sequencing of template DNA is performed, where reads are derived from clonal fragments and genotypes are derived by counting. The requirements in terms of coverage an accuracy for such workflows with short-read technologies is still not well understood. A critical question in association studies is the power to detect genetic association with a fixed sample size. Recent work has focused on two-stage designs, where subsets of individuals and markers are typed each in screening and replication stages. Here, we examined situations where the disease variant is initially not typed, but can be discovered and typed by NGS of the leading candidate regions in the cohort. Properly deploy of NGS platforms such as SOLiD in disease association studies of complex traits requires a better understanding of the aforementioned issues.

MATERIALS AND METHODS

We developed a model to simulate digital sequencing with pooled samples, in the presence of error. Given a series of basecalls at a position in the genome, we modeled the likelihood that these basecalls represent a homozygous AA, heterozygous AB, or homozygous BB genotype through Bayesian hypothesis testing as follows:

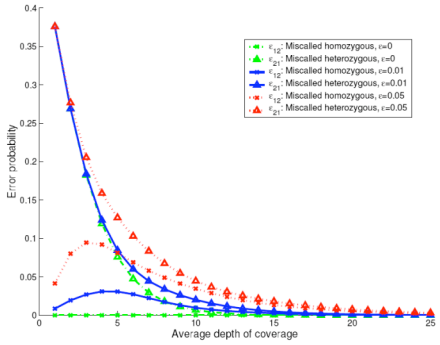
$$P(H_i \mid x_1 \ldots x_n) = \frac{P(x_1 \ldots x_n \mid H_i)P(H_i)}{\sum_j P(x_1 \ldots x_n \mid H_j)P(H_j)}$$

We automated the decision making between hypotheses and averaged across the depth of coverage distribution in a shotgun sequencing experiment. Error is expressed completely as a function of basecalling error rate and average depth of coverage.

To calculate power under a two-stage scenario, we extend the method of Skol et al. (Nat Genet 38: 209-213, 2006) by considering situations where the disease variant is not typed but is in linkage disequilibrium with a SNP in the discovery panel, and can be typed in stage 2 due to full resequencing in cases and controls of the lead regions identified in the first phase. The latter was previously unfeasible due to the cost of Sanger sequencing, but now is becoming possible due to the advent of NGS methods. We investigated what factors most significantly determine power using a factorial design with two settings for nine parameters and performing a backward stepwise regression to determine a "best-fitting" model for these data.

RESULTS

Figure 1. Genotyping error as function of read depth of coverage



As shown in Fig. 1, on our simulation results highlight the importance of the NGS error rate, and indicates that coverage of 20-40X is needed to reduce heterozygous misclassification errors to acceptable rates, whereas homozygote calling requires less coverage (10-15X). Since the shotgun process creates a Poisson distribution of coverage, some missing data still exists, suggesting a role for statistical imputation methods based on population genetics. The SOLiD System currently exhibits an average accuracy >99.94% with 35-bp reads.

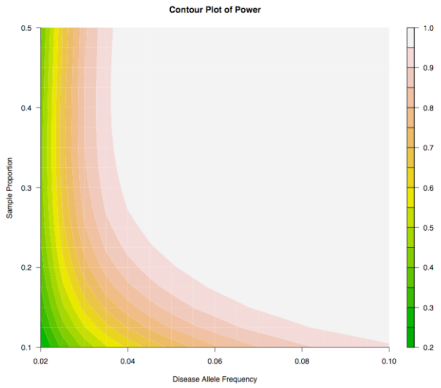
Table 1. Power of two-stage association study design with and without follow-up resequencing in second stage

	Estimated Power	
	GWA Stage	Replication
Functional SNP typed in GWA and replication/fine mapping	0.978	0.878
SNP in LD in both stages (indirect association)	0.654	0.470
SNP in LD GWA; Functional SNP typed in 2nd stage	0.654	0.843

1250/1250 cases/control; Significance 0.05; Multiplicative model; GRR 1.89; DAF 0.05; SNP MAF 0.075; D' 0.9

A major reason for considering two stage is the reduction in genotyping cost. However, these designs often assumed that the disease variant is contained in the stage 1 panel Often, this is not true as due to technical limitations of high density SNP chips. Table 1 shows that in this case power loss is significant, but can be mostly recovered by full resequencing of lead regions in the second stage.

Figure 2. Contour plot of joint analysis power as function of disease allele frequency and sampling proportions



The variable most significantly affecting power is relative risk, followed by the interaction between relative risk and sampling proportion. We found that typing additional markers at stage 2 due to full SNP ascertainment does not significantly affect power. We computed power for joint analysis as a function of nine parameters computing 2^9 = 512 power values. We then applied a stepwise regression analysis to ideintify factor relevance. Parameters used: Relative risk ration=1.8; sample size (c/c) = 3000/3000 ; r^2 = 0.9; additional markers in stage II=10; prevalence = 0.01 ; Δp=0.01 ; π_{ind}= 0.05

CONCLUSIONS

Since NGS platforms typically produce short reads (25-50bp), coverage needs to increase to 15-20X to reduce heterozygous misclassification errors to acceptable rates, whereas homozygote calling requires less coverage (10-15X). Error rate significantly influences coverage requirements highlighting its importance in "genometyping". Due to shotgun coverage some missing data will persist, suggesting a role for statistical imputation methods. Our simulation results are confirmed overall by empirical data obtained by resequencing long-range PCR amplicons from ENCODE regions performed on a HapMap Yoruba sample where shotgun Sanger data was generated and can be used as reference (data not shown - see poster 2134 by Hyland et al. in this meeting).

We discover that while significant power loss may occur when the disease variant is absent from the screening phase, power can be recovered by full SNP ascertainment by resequencing. These results suggest that NGS could become a valuable tool in the identification of susceptibility genes for complex disease.