SNP discovery in high-throughput resequenced microarray-enriched human genomic loci

Alena A. Antipova, Tanya D. Sokolsky, Christopher R. Clouser, Eileen T. Dimalanta, Cynthia L. Hendrickson, Cisilaya Kosnopo, Clarence C. Lee, Swati S. Ranade, Lei Zhang, Alan P. Blanchard, Kevin J. McKernan, ABI, 500 Cummings Center, Beverly, MA, USA, 01915

ABSTRACT

Identifying genetic variants and mutations that underlie human diseases and complex traits is an essential step in development of robust, cost-effective tools for routine resequencing of regions of interest in the human genome. Here we demonstrate that coupling Applied Biosystems SOLiD System to microarray-based resequencing of selected genomic regions provides an efficient and robust method for high-throughput resequencing and single nucleotide polymorphism (SNP) discovery in human protein-coding exons.

INTRODUCTION

Recent advances in high-throughput sequencing technologies have made it possible to develop a number of accurate and sensitive methods for polymorphism discovery in the whole human genome. Even with currently available parallel sequencing methods, it remains a trade-off between the power to detect localized variation in thousands of patients versus the power to detect all variation throughout the genome in a few individuals. One way of addressing this issue is to focus resequencing efforts on smaller genomic regions, selected as a result of prior investigations, such as previous linkage studies.

The enrichment approach described above has been previously used to generate up to 7-fold median resequencing coverage of 0.05-80 Mb genomic regions in a single sequencing run [1, 2, 3, 4]. However, it is not obvious from these studies if the enriched sample maintained an accurate representation of polymorphism profiles after resequencing, and could be used as a surrogate for polymorphism-detection. Here we demonstrate that SOLiD resequencing of microarray-enriched 4.3 Mb regions in the human genome provides a sensitive, accurate, and cost-efficient tool for detecting human polymorphisms, and investigate if any possible bias in polymorphism detection can result from this procedure.

MATERIALS AND METHODS

A custom oligonucleotide array (Agilent Technologies, 24k format) was designed with 80 probes targeting 4.3 Mb of human protein-coding exons. A portion (30%) of a fragment library from human DNA (NA18507) was hybridized to the array, while the remainder was reserved for control. Hybridized (enriched) fragments were eluted from the array, concentrated by precipitation, and then both control and hybridized sample were sequenced with the SOLiD System Sequencing. The 35 base pair sequencing reads were aligned against the target sequence with up to 3 mismatches. SNP detection was performed on the alignment, and identified SNPs were verified by comparison with HapMap release 23a for NA18507.

The empirical enrichment was calculated as follows: (Percent of sequence reads uniquely matching target regions in the enriched sample)/ (Percent of sequence reads uniquely matching target regions in the control sample).

The theoretical enrichment for the array-hybridized sample was calculated as [1]: (Percent of sequence reads uniquely matching target regions in the enriched sample)/100.

RESULTS

Figure 1. Microarray-based enrichment and resequencing of selected genomic loci

A random fragment library is obtained by ligation of generic adapters (highlighted in magenta) to sheared genomic fragments, and is divided into two samples. The control is set aside, while the other sample is labeled with a custom-designed biotin probe. The Probe attaches via a short adaptor to the target sequence, and is captured via streptavidin-coated magnetic beads. The beads are cleaned and eluted, and the fragment library is sequenced. Reads are aligned against the target sequence covered by at least one tag, and 87% of covered bases having accurate empirical evaluation of the enrichment efficiency of our protocol. Our results demonstrate that SOLiD resequencing of microarray-enriched genomic regions provides an accurate representation of polymorphism profile, as evidenced by the 99% accuracy of HapMap SNPs for the enriched sample. The errors in the SNP discovery primarily result from either low or missing base coverage, when there is not enough sequencing reads to form a consensus on the genotype of the SNP, or from genetic duplications, highly homologous to the probe. Overall, our results demonstrate that SOLiD resequencing of microarray-enriched genomic regions provides a powerful tool for genetic analysis and will expedite the search for genes contributing to understanding complex diseases and diseases in which somatic mutations play a role, such as atherosclerosis and cancer.

REFERENCES


ACKNOWLEDGMENTS

We thank members of the High Throughput Discovery department at Applied Biosystems and, in particular, Stephen McLaughlin and Jonathan Manning, for assistance.