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, Cisilya 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 requires 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 sequencing platform with microarray capture of targeted regions provides an efficient and robust method for high-coverage resequencing and single nucleotide polymorphism (SNP) discovery in human protein-coding exons.

INTRODUCTION

Recent advances in high-throughput sequencing technologies have resulted in development of a number of accurate and sensitive methods for polymorphism discovery in the whole human genome. Still, even with massively parallel sequencing technologies, there 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 provide up to 7-fold median resequencing coverage of 0.05-80 Mb genomic regions in a single sequencing run ([1], [2], [3], [4]) with an estimated average theoretical enrichment of 300-400 fold. 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 regions of the human genome provides a sensitive, accurate, and cost-efficient tool for detecting human polymorphisms, and investigate if any possible biases in polymorphism detection can result from this procedure.

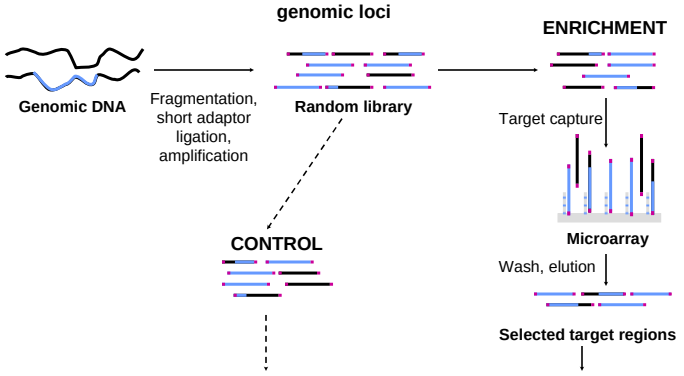
MATERIALS AND METHODS

A custom oligonuclotide array (Agilent Technologies, 244K format) was designed with 60 bp probes targeting 4.3 MB of human protein-coding exons. A portion (30 µg) 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 interrogated with 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 in [1]: (Percent of sequence reads uniquely matching target regions in the enriched sample/ Percent of sequence reads uniquely matching human genome in the enriched sample)*Maximum enrichment, where Maximum enrichment is defined as a ratio of genome length vs. target length and is equal to 722 for our 4258893 bp enrichment target.

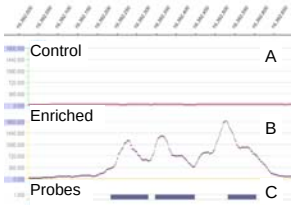
RESULTS

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



A random fragment library is obtained by ligating generic adapters (highlighted in magenta) to sheared genomic fragments, and is divided into two samples. The control is set aside, while the other sample undergoes hybridization to a custom designed high-density oligonucleotide array to enrich for the target DNA regions (highlighted in blue). The eluted captured regions are then SOLiD™-sequenced in parallel with the control sample.

Figure 2. Sequence coverage by uniquely matching sequence reads generated by the SOLiD™ System for a sample region of chromosome 22.



The X axis indicates the chromosomal position of the mapped reads, and the Y axes represents per base coverage. A. Coverage in unhybridized control sample. B. Coverage in an enriched sample. C. Placement of hybridization probes in the genome. The per base sequence coverage for the enriched sample (panel A) exceeds 340x (reaching almost 1800x), while staying under 11x in the control sample (Panel B).

Comparison with the chromosomal location of the probe targets (Panel C) demonstrates that the vast majority of sequencing reads in the enriched sample map to regions overlapping array hybridization probes. Averaging over the whole 4.3 Mb enrichment target, we obtained 138-fold per base coverage (median coverage is 59-fold) with 91% of target sequence covered by at least one tag, and 87% of covered bases having coverage within one standard deviation of the average. This level of coverage enabled highly accurate and sensitive SNP detection, with 99.5% of identified HapMap SNPs called correctly, establishing suitability of a combined enrichment/SOLiD™ -resequencing approach for polymorphism discovery.

Table 1. Accuracy of SNP discovery

HapMap SNPs	Found by SOLiD™	Correctly identified by SOLiD™	Correctly identified by HapMap	SOLiD™ accuracy	HapMap accuracy
Homozygous for reference allele	4568	4567	4552	99.98	99.65
Homozygous for alternative allele	696	695	694	99.86	99.71
Heterozygous	744	718	742	96.51	99.73
Unknown	35	31	0	88.57	0

We evaluated accuracy and sensitivity of SNP discovery by comparing our results to the HapMap database and, where necessary, resequencing by conventional Sanger sequencing. Two approaches were utilized. Firstly, we tested if our sequencing data corroborated presence and identity of known HapMap SNPs (we applied this strategy to SNPs homozygous for the reference allele, as these SNPs are indistinguishable from reference when a single individual is resequenced). Secondly, we performed *de novo* SNP detection, as described in Materials and methods, aiming to identify all heterozygous SNPs and all SNPs homozygous for the alternative allele in the target regions. Those "newly-found" SNPs were then compared to HapMap reference, and only SNPs present in HapMap were selected to evaluate the accuracy of our SNP discovery.

Table 2. Duplications overlapping chr7:150795122

Duplication coordinates	Strand	Reference in SNP position
chr7:150794730-150795233	Forward	A
chr10:46334616-46335119	Reverse	G
chr10:46816688-46817192	Reverse	G
chr22:16882878-16883370	Reverse	A

Table 3. Comparison of SOLiD, HapMap, and Sanger calls for chr7:150795122

	Reference	SOLiD	HapMap	Sanger
chr7:150795122	A	GA	GG	AA

The only HapMap SNP homozygous for the alternative allele and erroneously called heterozygous by SOLiD™ appears to overlap a 500 base pair homologous duplication (88-98 % identity) on chromosomes 7, 10, and 22 (Table 2), which probably contributed to its erroneous identification by both HapMap and SOLiD (Table 3).

Alena.Antipova@appliedbiosystems.com

Table 4. Sensitivity of *de novo* SNP finding

HapMap SNPs	Overlap probes	Non-zero coverage	3+ fold coverage	Found by SOLiD™
Homozygous for alternative allele	914	819	753	696
Heterozygous	1139	1050	975	745
Unknown	133	108	98	35

Since, at a minimum, *de novo* SOLiD™ SNP discovery required 3 sequencing reads to call a SNP, we estimated how many HapMap SNPs had 3-plus coverage in our target regions. Positions of 83% of the HapMap SNPs homozygous for the alternative allele and 86% of the heterozygous HapMap SNPs were covered by at least 3 reads. The proportion of "unknown" HapMap SNPs with 3-plus coverage was lower at 74%, suggesting that sequences surrounding "unknown" HapMap SNPs, where conventional SNP calling was not successful, might also present some challenge to our platform. Indeed, SOLiD™ found only 36% of the "unknown" HapMap SNPs with 3-plus coverage. Still, of the SNPs found, 89% were identified correctly, thus demonstrating suitability of our SNP detection approach even in the regions with complex sequences not amenable to SNP discovery by conventional methods.

Table 5. Comparison of matching statistics for the enriched and control samples

	Enriched	Control
Total	2.6706 Gb	3.9068 Gb
Reads uniquely mapping to human genome	1.0825 Gb	1.2853 Gb
Reads uniquely mapping to target regions	0.5899 Gb	0.0047 Gb

The empirically calculated enrichment, a ratio of enriched sample reads matching target vs. control sample reads, was equal 184. The theoretical enrichment, commonly used to evaluate the efficiency of the enrichment procedure [1], was predicted to be 391 (see Materials and methods). Thus the theoretical enrichment twice exceeded the empirical enrichment derived from direct comparison of the enriched and the control samples. The measure of empirical enrichment is more reflective of how much more sequence is required if whole genome shotgun sequencing were utilized as opposed to array enrichment.

CONCLUSIONS

In summary, by coupling the Applied Biosystems SOLiD™ System sequencing platform with microarray capture of targeted regions, we developed an efficient and customizable method for high-coverage resequencing and polymorphism discovery in human protein-coding exons. The theoretical enrichment was calculated to be 391-fold, which is higher or on par with previously reported data [1]. However, while commonly used, we believe this evaluation of the enrichment is misleading, as it does not take into account the sequence complexity of the target. Here we propose an alternative metric for estimating efficacy of the enrichment procedure, by comparing amplification of the target regions in the enriched sample and in the control. In this study, we report that 184 times more sequencing reads mapped to the target after the microarray selection procedure, thus providing an accurate empirical evaluation of the enrichment efficiency of our protocol. Our results demonstrate that combination of array enrichment with SOLiD™ sequencing provides an accurate representation of polymorphism profile, as evidenced by the 99.5% accuracy of SNP discovery for the enriched sample. The errors in the SNP discovery primarily result either from low per base coverage, when there is not enough sequencing reads to form a consensus on the genotype of the SNP, or from genomic duplications, highly homologous to the probes. 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 inherited common diseases and diseases in which somatic mutations play a role, such as atherosclerosis and cancer.

REFERENCES

1. Albert TJ, Molla MN *et al*: Direct selection of human genomic loci by microarray hybridization. *Nature methods* 2007, 4(11):903-905.
2. Dahl F, Stenberg J *et al*: Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proceedings of the National Academy of Sciences of the United States of America* 2007, 104(22):9387-9392.
3. Hodges E, Xuan Z *et al*: Genome-wide in situ exon capture for selective resequencing. *Nature genetics* 2007, 39(12):1522-1527.
4. Porreca GJ, Zhang K *et al*: Multiplex amplification of large sets of human exons. *Nature methods* 2007, 4(11):931-936.

ACKNOWLEDGEMENTS

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