Published in final edited form as:

Nat Methods. 2009 May ; 6(5): 315–316. doi:10.1038/nmeth.f.248.

Massively parallel exon capture and library-free resequencing across 16 individuals

Emily H. Turner¹, Choli Lee¹, Sarah B. Ng¹, Deborah A. Nickerson¹, and Jay Shendure¹

¹Department of Genome Sciences, University of Washington, Seattle, WA 98195

To the Editor: The adoption of molecular inversion probes (MIPs) to massively parallel exon capture¹ has been limited by representational and allelic bias. We describe modifications enabling simultaneous amplification and accurate shotgun sequencing of 50,000 exons. We also prove the scalability and accuracy of direct sequencing of MIP amplicons, which circumvents all shotgun library construction, while resequencing 1 megabase of coding sequence across 16 human genomes with >99% HapMap concordance.

MIPs represent a scalable technology previously applied to ~10,000-plex SNP genotyping². Its adaptation to massively parallel exon capture demonstrated extensive multiplexing, straightforward scalability, high specificity, and low DNA input requirements¹. However, two crippling deficiencies were encountered. First, only ~10,000 of 55,000 targeted exons were detectably amplified, with highly variable representation. Second, heterozygous alleles were not equally sampled, substantially impairing variant detection. Resolution of these deficiencies was a clear prerequisite for MIP-based exon capture to be broadly useful.

We modified the original protocol to enhance capture efficiency (Supplementary Methods online). 55,000 array-derived 100-mers were amplified and converted to single-stranded 70-mer MIPs as previously described¹, and again applied to 55,000-plex exon capture on genomic DNA (HapMap NA12248). Key changes included lengthening hybridization and gap-fill incubations, and increasing MIP and ligase concentrations. Post-capture PCR amplicons were linearly concatenated and processed to a standard shotgun sequencing library³.

These modifications yielded a remarkable improvement in capture efficiency and allelic sampling while maintaining high specificity. Of 18 million uniquely mapped⁴ reads obtained by Illumina sequencing, 99% overlapped with one of the 55,000 targets. 91% of targets were detectably captured (50,080 of 55,000), compared to 18% previously described¹. Representational uniformity improved markedly as well (Fig. 1a). Allelic sampling of heterozygous variants showed a dramatic improvement as compared to previous data (Fig. 1b,c), now matching our expectation of a distribution converging to 0.5 with increasing coverage. An assessment of variant calling accuracy (including Sanger-based confirmation of a subset of novel variants) is provided in Supplementary Note 1 online.

Shotgun library construction remains a key bottleneck for scaling second-generation sequencing⁵ to thousands of samples, as mechanical fragmentation and gel-based size-selection are challenging to automate. With longer read-lengths (e.g. Illumina GA-2; 76+ base-pairs), we realized direct sequencing of MIP-derived amplicons would enable targeted resequencing without shotgun library construction (Fig. 2).

To test this, gDNA from 16 individuals were subjected to targeted capture with 13,000 MIPs targeting 13,000 exons (Supplementary Data 1 online; subset of 55,000 with greater design constraints). Minor additional protocol changes were made (e.g. reducing input to 750 nanograms), but the primary change was the introduction of direct resequencing. Following capture, two multi-template PCRs (per individual) appended Illumina adaptors in either orientation. Amplicons from each individual were pooled 1:1 and directly sequenced with 76-bp single-end reads in one lane (Supplementary Fig. 1 online).

Performance of capture was highly consistent across the 16 individuals (Supplementary Table 1 online). On average, 8.4 million quality-filtered reads were collected per individual, of which 90.4% were confidently mapped⁴. The proportion of mapped reads was higher with direct sequencing than shotgun sequencing (90.4% vs. 56.8%), primarily because shotgun libraries are contaminated by the common MIP linker. Captures were highly specific with >99% of mapped reads aligning to one of 13,000 targets (Supplementary Fig. 2 online). Representational uniformity was further improved with ~98% of all targets captured per reaction, ~58% within a 10-fold, and ~88% within a 100-fold range (Fig. 1a), approaching the performance of array-based capture⁶. The relative capture efficiencies of individual MIPs were reproducible (average pairwise rank-order correlation coefficient of 0.92).

As the useful read-length for variant calling is 56 bases (Fig. 2), the maximum target length accessible with bi-directional sequencing was 112 bp. Because the lengths of the 13,000 targets ranged from 100 to 191 bp, we focused our analysis on accessible coding bases within targets (~1.4 of ~1.7 Mb), with the expectation that direct sequencing of longer capture products will be possible as read-lengths increase.

Performance of variant calling was highly reproducible across the 16 individuals (Supplementary Table 2 online). The overall concordance to HapMap genotypes was high for both homozygous (99.8%; n = 21,346) and heterozygous genotypes (99.3%; n = 3,622). While ~98% of targets were captured in each sample, only 75% of accessible target bases (~1.0 of ~1.4 Mb) were covered sufficiently for genotype calling. We estimate that 2x, 3x, and 4x increased sequence depth would increase the call-rate from 75% to 85%, 89%, and 91% respectively, i.e. with diminishing returns. Achieving greater completeness will likely require either further protocol optimizations or independent targeting of under-covered targets. Alternatively, we estimate that performing 2, 3 or 4 capture reactions and sequencing each at equivalent depth would increase the call-rate to 90%, 92%, and 94%, respectively. 54% of accessible target bases (~0.8 Mb of ~1.4 Mb) were sufficiently covered for variant calling in all 16 samples.

Variants (Supplementary Data 2 online; 593 per individual on average) were compared to dbSNP. The average proportion of annotated variants was higher for Yoruba (87%; n = 10) than Europeans (95%; n = 4) and Asians (94%, n = 2), consistent with greater diversity and poorer historical ascertainment in Africans. Across 16 individuals, 779 novel variants were identified. In contrast with previously annotated variants, most novel variants were observed just once across 32 chromosomes (Supplementary Fig. 3 online). Comparison of variant genotypes called here to genotypes generated independently by hybrid capture and sequencing (manuscript in preparation) validated 99% of all variants (n = 3703) and 94% (n = 303) of novel variants called in both data-sets, consistent with a low false discovery rate.

Our results establish the accuracy, reproducibility, and scalability of array-derived MIPs for massively parallel exon capture and resequencing. Our simplified protocol involving direct sequencing of MIP amplicons has important advantages over alternative methodologies for targeted capture: (a) concurrent targeting of at least 50,000 exons per reaction; (b) sub-microgram input DNA requirements; (c) a single synthesis of array-derived MIP precursors

can potentially support thousands of capture reactions; (d) small solution-based capture reactions that are highly scalable with no requirement for shotgun library construction at any stage.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported in part by grants from the National Institutes of Health / National Heart Lung and Blood Institute (R01 HL094976 to D.N. and J.S.) and the National Institutes of Health / National Human Genome Research Institute (R21 HG004749 to J.S.). E.T. is supported by a training fellowship from the National Institutes of Health / National Human Genome Research Institute (T32 HG00035). S.N. is supported by the Agency for Science, Technology and Research, Singapore. For technical assistance with sequencing, we thank K. Bertucci and S. da Ponte. For assistance with genotyping data, we thank J. Smith, G. Cooper, and I. Stanway. For helpful discussions, we thank H. Ji, J. Li, K. Zhang, G. Church, J. Teer, J. Egerton, and G. Porreca. For assistance with oligo libraries, we thank E. Leproust and W. Woo.

References

1. Porreca GJ, et al. *Nat. Methods* 2007;4:931–936. [PubMed: 17934468]
2. Hardenbol P, et al. *Genome Res* 2005;15:269–275. [PubMed: 15687290]
3. Bentley DR, et al. *Nature* 2008;456:53–59. [PubMed: 18987734]
4. Li H, Ruan J, Durbin R. *Genome Res* 2008;18:1851–1858. [PubMed: 18714091]
5. Shendure J, Ji H. *Nat. Biotechnol* 2008;26:1135–1145. [PubMed: 18846087]
6. Albert TJ, et al. *Nat Methods* 2007;4(11):903–905. [PubMed: 17934467]

Figure 1. Optimizations yield reduced representational and allelic bias

(a) The uniformity of capture across 55,000 or 13,000 targeted exons is illustrated, estimated by the relative depth of sequence coverage of each target. (b,c) Sequencing-based allele ratios at positions of common variation are plotted for previously described¹ (b) and current (c) protocols. Each point represents a position where targeted resequencing data from a 55,000-plex amplification intersects with a HapMap genotype for this individual (NA12248). The allele ratio is the frequency with which the reference allele was observed in sequence data. Genotypes are indicated according to HapMap data.

Hybridization

Extension and ligation

Multi-template PCR

Direct resequencing

Figure 2. Schematic of MIP-based exon capture and direct resequencing

Thousands of MIP precursors can be cost-effectively produced by release from programmable arrays. Each MIP is an oligonucleotide that includes a common linker (green) flanked by target-specific sequences (yellow) that hybridize immediately upstream and downstream of its target. Addition of polymerase and ligase results in copying of targets and conversion to a circular form. Illumina adaptors (black) are appended by a multi-template, inverse PCR reaction, resulting in amplicons that can be directly sequenced. All 76 bp sequencing reads (grey arrows) include 20 bases of targeting arm sequence and an additional 56 bases of within-target sequence.