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

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

MIPs are a scalable technology previously applied to ~10,000-plex single-nucleotide polymporphism genotyping[2]. Its adaptation to massively parallel exon capture had demonstrated extensive multiplexing, straightforward scalability, high specificity and low DNA input requirements[1]. However, two crippling deficiencies had been encountered. First, only ~10,000 of 55,000 targeted exons had been detectably amplified, with highly variable representation. Second, heterozygous alleles had not been equally sampled, substantially impairing variant detection. Resolution of these deficiencies was a prerequisite for MIP-based exon capture to be broadly useful.

We modified the original protocol to enhance capture efficiency (**Supplementary Methods** online). We amplified 55,000 array-derived 100-mers and converted them to single-stranded 70-mer MIPs as previously described[1], and again applied to 55,000-plex exon capture to genomic DNA (HapMap NA12248). Key changes included lengthening hybridization and gap-fill incubation durations and increasing MIP and ligase concentrations. We linearly concatenated after-capture PCR amplicons and processed them to create a standard shotgun sequencing library[3].

These modifications yielded a remarkable improvement in capture efficiency and allelic sampling while maintaining high specificity. Of 18 million uniquely mapped[4] reads obtained by Illumina sequencing, 99% overlapped with one of the 55,000 targets. We detectably captured 91% of targets (50,080 of 55,000 targets), compared to 18% previously described[1]. Representational uniformity improved markedly as well (**Fig. 1a**). Allelic sampling of heterozygous variants was dramatically improved as compared to previous data (**Fig. 1b,c**), now matching our expectation of a distribution converging to a ratio of 0.5 with increasing coverage. We assessed variant calling accuracy (and confirmed a subset of new variants by Sanger sequencing; **Supplementary Note 1** online).

Shotgun library construction remains a key bottleneck for scaling second-generation sequencing[5] to thousands of samples, as mechanical fragmentation and gel-based size-selection are challenging to automate. With longer read lengths (for example, Illumina Genome Analyzer 2; 76 base pairs (bp) or more) we realized direct sequencing of MIP-derived amplicons would enable targeted resequencing without shotgun library construction (**Fig. 2**).

To test this, we subjected genomic DNA from 16 individuals to targeted capture with 13,000 MIPs targeting 13,000 exons (**Supplementary Data 1** online; subset of 55,000 exons with greater design constraints). We made minor additional protocol changes (for example, reducing input to 750 nanograms), but the primary change was the introduction of direct resequencing. After capture, we used two multitemplate PCRs (per individual) to appended

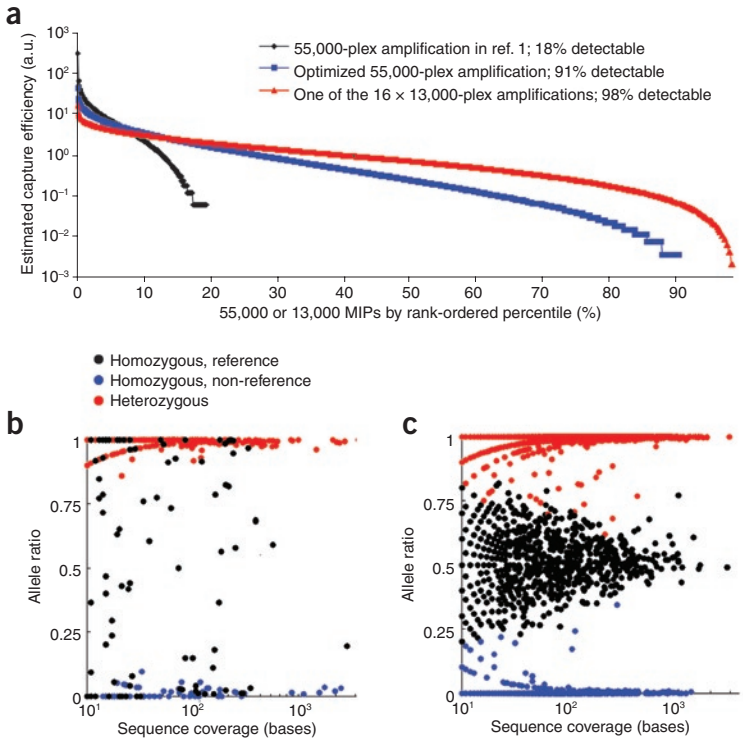

Figure 1 | Optimization yields reduced representational and allelic bias. (**a**) Capture efficiency across 55,000 or 13,000 targeted exons, estimated by the relative depth of sequence coverage of each target. (**b,c**) Sequencing-based allele ratios at positions of common variation plotted for previously described[1] (**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 an individual. The allele ratio is the frequency with which the reference allele was observed in sequence data. Genotypes are indicated according to HapMap data.

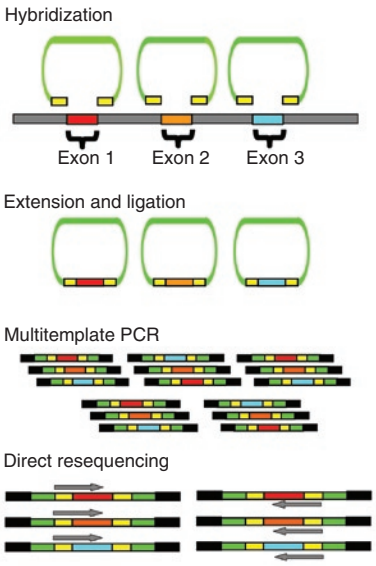

Figure 2 | Schematic of MIP-based exon capture and direct resequencing. Each MIP is an oligonucleotide that includes a common linker (green) flanked by target-specific sequences (yellow) that hybridize immediately upstream and downstream of its target. Addition of polymerase and ligase results in copying of targets and conversion to a circular form. Illumina adaptors (black) are appended by a multitemplate, inverse PCR reaction, resulting in amplicons that can be directly sequenced. All 76-bp sequencing reads (arrows) include 20 bases of targeting arm sequence and an additional 56 bases of within-target sequence.

Illumina adaptors in either orientation. We pooled amplicons from each individual's genome in a 1:1 ratio and directly sequenced them with 76-bp single-end reads in one lane of the sequencer (**Supplementary Fig. 1** online).

Capture was highly consistent across the 16 individuals' genomes (**Supplementary Table 1** online). On average, we collected 8.4 million quality-filtered reads per individual, of which 90.4% we confidently mapped[4]. The proportion of mapped reads was higher with direct sequencing than shotgun sequencing (90.4% versus 56.8%), primarily because shotgun libraries were 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). We improved representational uniformity with ~98% of all targets captured per reaction, ~58% within a tenfold and ~88% within a 100-fold range (**Fig. 1a**), approaching the performance of array-based capture[6]. 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 bidirectional sequencing was 112 bp. Because the lengths of the 13,000 targets were 100–191 bp, we focused our analysis on accessible coding bases within targets (~1.4 Mb of ~1.7 Mb), with the expectation that direct sequencing of longer capture products will be possible as read lengths increase.

Variant calling was highly reproducible across the 16 individuals' genomes (**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). Although we captured ~98% of targets in each sample, only 75% of accessible target bases (~1.0 of ~1.4 Mb) were covered sufficiently for genotype calling. We estimate that 2×, 3× and 4× increased sequence depth would increase the call rate from 75% to 85%, 89% and 91%, respectively, that is, with diminishing returns. Achieving greater completeness will likely require either additional protocol optimization or independent targeting of undercovered 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. Fifty-four percent of accessible target bases (~0.8 Mb of ~1.4 Mb) were sufficiently covered for variant calling in all 16 samples we examined.

We compared called variants to data in the dbSNP database (**Supplementary Data 2** online; 593 variants per individual genome on average). The average proportion of annotated variants was higher for Yoruba (87%; n = 10) than European (95%; n = 4) and Asian (94%, n = 2) individuals, consistent with greater diversity and poorer historical ascertainment in African individuals. Across 16 individuals' genomes, we identified 779 new variants. In contrast with previously annotated variants, most new 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 (S.B.N., E.H.T., C.L., M. Bamshad, D.A.N. and J.S., manuscript in preparation) validated 99% of all variants (n = 3,703) and 94% of new variants (n = 303) called in both datasets, consistent with a low false discovery rate.

Our results established 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: (i) concurrent targeting of at least 50,000 exons per reaction; (ii) submicrogram input DNA requirements; (iii) a single synthesis of array-derived MIP precursors that can potentially support thousands of capture reactions; (iv) small solution-based capture reactions that are highly scalable with no requirement for shotgun library construction at any stage.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

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

Emily H Turner, Choli Lee, Sarah B Ng, Deborah A Nickerson & Jay Shendure

Department of Genome Sciences, University of Washington, Seattle, Washington, USA.
e-mail: shendure@u.washington.edu

PUBLISHED ONLINE 6 APRIL 2009; DOI:10.1038/NMETH.F.248

1. Porreca, G.J. *et al. Nat. Methods* **4**, 931–936 (2007).
2. Hardenbol, P. *et al. Genome Res.* **15**, 269–275 (2005).
3. Bentley, D.R. *et al. Nature* **456**, 53–59 (2008).
4. Li, H., Ruan, J. & Durbin, R. *Genome Res.* **18**, 1851–1858 (2008).
5. Shendure, J. & Ji, H. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
6. Albert, T.J. *et al. Nat. Methods* **4**, 903–905 (2007).