Multiplex amplification of large sets of human exons

Gregory J Porreca[1,8], Kun Zhang[1,7,8], Jin Billy Li[1], Bin Xie[2], Derek Austin[2], Sara L Vassallo[1], Emily M LeProust[3], Bill J Peck[3], Christopher J Emig[4], Fredrik Dahl[5,7], Yuan Gao[2,6], George M Church[1,8] & Jay Shendure[1,6,8]

A new generation of technologies is poised to reduce DNA sequencing costs by several orders of magnitude. But our ability to fully leverage the power of these technologies is crippled by the absence of suitable 'front-end' methods for isolating complex subsets of a mammalian genome at a scale that matches the throughput at which these platforms will routinely operate. We show that targeting oligonucleotides released from programmable microarrays can be used to capture and amplify ~10,000 human exons in a single multiplex reaction. Additionally, we show integration of this protocol with ultra-high-throughput sequencing for targeted variation discovery. Although the multiplex capture reaction is highly specific, we found that nonuniform capture is a key issue that will need to be resolved by additional optimization. We anticipate that highly multiplexed methods for targeted amplification will enable the comprehensive resequencing of human exons at a fraction of the cost of whole-genome resequencing.

Recently several DNA sequencing platforms have been demonstrated, which decrease cost and increase throughput by massively parallel interrogation of arrayed polymerase colonies[1–4]. Although conventional technology continues to account for the overwhelming majority of DNA sequencing, the remarkable cost-effectiveness of these new platforms makes it unlikely that this will be the case in several years. As the canonical genome sequences of all major model organisms, as well as of our own species, are nearly complete, much of the enthusiasm for how to apply these new technologies is directed at the discovery of somatic mutations and germline variation. As costs still remain too high to support routine resequencing of complete human genomes, these efforts will likely focus initially on the resequencing of specific subsets of the genome in multiple individuals.

Conventionally, targeted variation discovery is achieved by Sanger sequencing of PCR amplicons[5,6]. PCR provides the 'front end,' permitting amplification of discrete regions that can be sequenced by an individual read, for example, 1 kb. As new technologies displace the Sanger method, is PCR adequate to continue in this role? To illustrate the key difficulty with an example, 1 Gb of sequencing would permit 20-fold coverage of 50,000 1-kb regions. But this would still require 50,000 PCRs. Our ability to fully leverage the power of next-generation sequencing technologies is markedly crippled by the lack of corresponding 'front-end' targeting technologies, analogous to PCR, that are matched to the scale at which these new sequencing technologies will operate.

An ideal targeting method serving this role would enable the single-reaction, multiplex capture of arbitrary subsets of a complex genome (for example, 100,000 exons). Despite decades of effort, PCR multiplexes poorly[7,8]. Other methods, however, are more compatible with multiplexing. For example, 'molecular inversion probes' can interrogate more than 10,000 single nucleotide polymorphisms (SNPs) in a single reaction[9,10]. Recently one report described the multiplex capture of 425 targeted restriction fragments by selective circularization[11] and another described a modified version of multiplex PCR that allowed simultaneous amplification of 170 exons[12]. A hybridization-based capture protocol has shown 10,000-fold enrichment of sequences derived from bacterial artificial chromosome–sized genomic regions[13]. Finally, in this issue of *Nature Methods*, two reports present a selection method that facilitates enrichment by hybridization of target sequences to oligonucleotide microarrays[14,15].

In evaluating multiplex targeting methods, key performance parameters to consider include multiplexity, specificity and uniformity. Multiplexity refers to the number of independent capture reactions performed simultaneously in a single reaction. Specificity is measured as the fraction of captured nucleic acids that derive from targeted regions. Uniformity is defined as the relative abundances of targeted sequences after selective capture. Ideally, a multiplex targeting method will perform adequately by all three measures. An additional concern is cost; targeted capture necessarily requires one or more oligos to specify each target, which is potentially very expensive at high degrees of multiplexing. We therefore explored DNA synthesis on programmable microarrays, previously used for long DNA assembly[16], as an opportunity to

[1]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [2]Center for the Study of Biological Complexity, Virginia Commonwealth University, 1000 W. Cary St. Richmond, Virginia 23284, USA. [3]Genomics Solution Unit, Agilent Technologies Inc., 5301 Stevens Creek Blvd., Santa Clara, California 95051, USA. [4]Codon Devices Inc., One Kendall Square, Building 300, Third Floor, Cambridge, Massachusetts 02139, USA. [5]Stanford Genome Technology Center, Clark Center W300, 318 Campus Drive, Stanford, California 94305, USA. [6]Department of Computer Science, Virginia Commonwealth University, 601 West Main Street, Richmond, Virginia 23284, USA. [7]Present addresses: Department of Bioengineering, University of California at San Diego, 9500 Gilman Dr., La Jolla, California 92093, USA (K.Z.), Complete Genomics Inc., 2071 Stierlin Court, Suite 100, Mountain View, California 94043, USA (F.D.), and Department of Genome Sciences, University of Washington, 1705 NE Pacific St., Seattle, Washington 98195, USA (J.S.). [8]These authors contributed equally to this work. Correspondence should be addressed to J.S. (shendure@u.washington.edu) and G.M.C. (http://arep.med.harvard.edu/gmc/email.html).

RECEIVED 7 AUGUST; ACCEPTED 21 SEPTEMBER; PUBLISHED ONLINE 14 OCTOBER 2007; DOI:10.1038/NMETH1110

substantially reduce the upfront costs of synthesizing complex mixtures of targeting oligos (**Supplementary Table 1** online).

Here we describe a new strategy for the targeted amplification of nucleic acids, showing its utility by capturing ~10,000 human exons in a single multiplex reaction. We evaluate the advantages and limitations of the method, in terms of multiplexity, specificity and uniformity, as well as its potential utility for targeted variation discovery.

RESULTS

Multiplex exon capture

In this method 100-mer oligos are synthesized and released from a programmable microarray. This complex pool is PCR amplified, then restriction digested to release a single-stranded 70-mer 'capture probe' mixture (**Fig. 1a**). Individual probes consist of a universal 30 nucleotide motif flanked by unique 20 nt segments ('targeting arms'). Each linked pair of targeting arms is designed to hybridize immediately upstream and downstream of a specific genomic target, for example, an exon. The capture event itself, a modification of the 'molecular inversion probe' strategy developed for multiplex genotyping[10], is achieved by polymerase-driven extension from the 3′ end of the capture probe to copy the target, followed by ligation to the 5′ end to complete the circle (**Fig. 1b**). Subsequent steps enrich and amplify these circles (**Fig. 1c**) or generate products amenable to shotgun sequencing library production (**Fig. 1d**).

Design of targeting oligos

Our initial experiments established this method as successful for 480-plex exon capture (**Supplementary Fig. 1** online). To test this method further, we designed 55,000 oligos, each intended to capture a single human exon. Specifically, we targeted a subset of the Consensus CoDing Sequence (CCDS) curation of well-annotated protein-coding regions (http://www.ncbi.nlm.nih.gov/CCDS). We selected exon targets within a defined size range (60–191 bp), excluding targets where either targeting arm overlapped with repetitive sequence, or had high (>70%) or low (<30%) (G+C) content. The 55,000 targeted sequences totaled 6.7 Mb (13% of CCDS protein-coding sequences).

Performance of capture reaction

The 55,000 targeting oligos were generated by parallel release of 100-mers synthesized on a programmable microarray, followed by amplification and enzymatic conversion to a single-stranded 70 nt form (**Fig. 1a**). We performed subsequent steps (**Fig. 1b,c**) in duplicate on genomic DNA derived from a single anonymous individual from the HapMap CEPH/UTAH population (GM12248). The size distribution of captured material in duplicate reactions (**Fig. 2**) was highly consistent with our expectation (60–191 bp targets plus 40 bp targeting arms plus 40 bp common primers yields 140–271 bp amplicons).

Evaluation of specificity

To verify that captured material represented intentionally targeted exons, we gel-purified and sub-cloned amplicons within the relevant size range. We obtained 356 Sanger sequencing reads (~50% from each duplicate). Of 329 reads that aligned to the human genome, 98% ($n = 322$) most strongly aligned to one of the 55,000 targets, indicating that the capture reaction is highly specific. For 2% ($n = 7$), this was not the case. Five of these exceptions involved capture of a highly paralogous sequence (93–98% identical to one of the 55,000 targets over the captured region; 90–100% identical in the targeting arm regions). Two exceptions involved capture of a clearly 'off-target' genomic sequence, that is, no significant alignment with any of the 55,000 targets.

Twenty-seven reads had no significant alignment to the human genome. A review of these chromatograms found that nearly all (26 of 27) were not high-quality traces and primarily appear to represent mixing of multiple clones. Alternative possibilities include artifacts consequent to the capture protocol or the capture of contaminating, nonhuman DNA. Notably, none of these 27 had significant alignments to sequences deposited in GenBank.

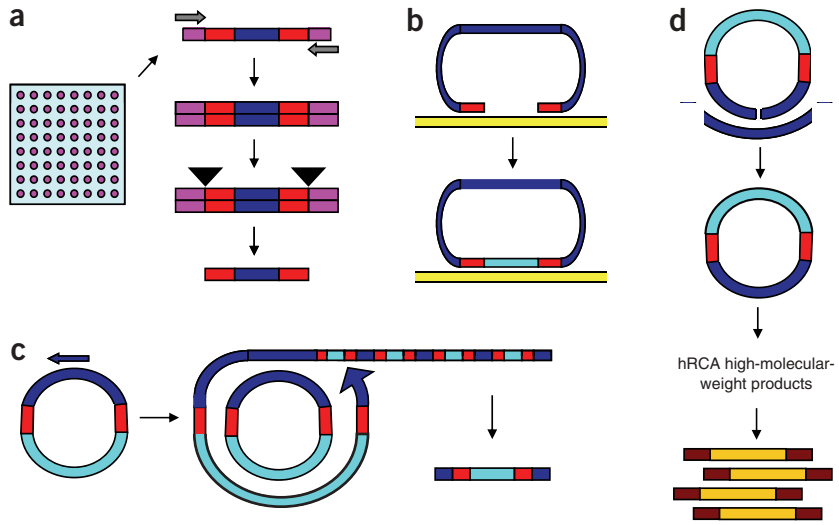

Figure 1 | Schematic of multiplex exon capture. (**a**) A library of 100-mer targeting oligo precursors is synthesized in parallel on a programmable microarray. Each oligo consists of common 15 nt flanking sequences (purple), unique 20 nt targeting arms (red) and a common 30 nt linker (dark blue). The oligos are released and PCR amplified with a single set of primers directed at the common flanking sequences. Double digestion with nicking restriction endonucleases (black triangles) releases a library of single-stranded 70-mers. (**b**) The unique targeting arms (red) of individual targeting oligos are designed to hybridize immediately upstream and downstream of each exon of interest. Hybridization to genomic DNA (yellow) is followed by an enzymatic gap-filling and ligation step, such that a copy of the sequence of interest is incorporated into a circle (light blue). The figure is not drawn to scale; the length of the gap-fill ranged from 60 to 191 bp. (**c**) Enrichment and amplification of exon-capture circles includes exonuclease digestion of linear material (not shown), linear rolling circle amplification and PCR amplification. Primers for the latter two steps are directed at the common linker sequence (dark blue). (**d**) To generate a shotgun sequencing library, amplicons are recircularized via their common linker (dark blue) and subjected to hyperbranched RCA with phi29 polymerase and random hexamers. High-molecular-weight products of the hRCA reaction are randomly sheared and appended with universal adaptors to generate the shotgun sequencing library.

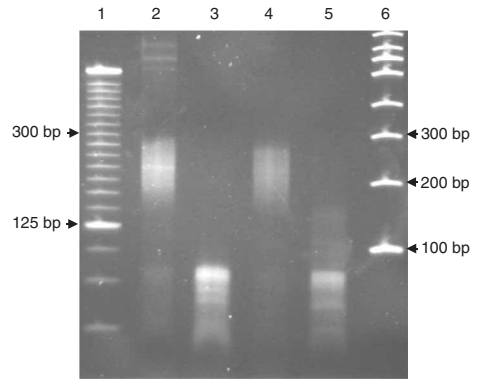

Figure 2 | Specificity of multiplex exon capture. The products of duplicate multiplex exon capture reactions are shown on a 6% nondenaturing polyacrylamide gel. Lane 1, 25 bp ladder (Invitrogen); lane 2, exon capture reaction, replicate 1; lane 3, negative control (no genomic DNA); lane 4, exon capture reaction, replicate 2; lane 5, negative control (no genomic DNA); lane 6, 100 bp ladder (NEB). Controls containing genomic DNA but no capture probe resulted in no substantial amplification as measured by real-time PCR (data not shown).

Evaluation of multiplexity and uniformity

Other key parameters to evaluate include multiplexity (how many targets were simultaneously captured within each reaction?) and uniformity (how uniform are their abundances relative to one another? how many of the 55,000 targets fail to be captured?). To address these questions, we performed high-throughput 'end sequencing' on a Illumina Genome Analyzer (Solexa). Specifically, we estimated the abundances of targets by sequencing ~35 bp from one end of amplicons and then counting the frequency with which each target appeared. For the duplicate reactions, 1,109,756 and 1,584,331 sequencing reads were obtained that could be confidently mapped to one of the 55,000 targets. Plots of rank-ordered abundances for observed species are shown in **Figure 3a**. We observed 15,380 of the 55,000 potential targets (28%) one or more times in one of the duplicates (10,660 and 10,587 for the reactions independently). Of targets that we observed within each reaction, abundances varied over 2–3 logs, with ~75% of observed targets falling within a 100-fold range.

Given that we observed ~10,000 of the 55,000 targets in each reaction, the number of targets observed in both reactions (n = 5,867) is significantly greater than expected by chance (chi-squared, $P < 1.0 \times 10^{-250}$). This suggests a systematic contribution to the abundance distribution. Consistent with this, the abundances of targets observed in both reactions are modestly correlated (**Fig. 3b**; Pearson correlation coefficient = 0.54). But we also observed that many targets were abundant in one replicate but absent in the other, suggesting a mixed picture in which both systematic and random factors govern the distribution, relative efficiency and reproducibility with which individual targets are captured.

Another concern is the high 'dropout' rate, that is, in each replicate, ~80% of the 55,000 intended targets were not observed by deep sequencing. This contrasts with the 480-plex experiment (**Supplementary Fig. 1**), in which we observed no dropout. To evaluate whether these missing species are present at a very low abundance, or are simply not captured, we performed PCR of sequences internal to 72 targets (a random subset of the 55,000 targets) and 24 negative controls (that is, exons not targeted), using the products of capture reactions as templates. In the duplicate reactions, 25 and 23 of the 72 targets (35% and 32%) amplified (and 1 and 2 of the 24 negative controls), suggesting that although some targets unobserved by deep end-sequencing are present at a very low abundance, many are either absent or are present at an extremely low frequency (undetectable by both direct PCR and end-sequencing of more than 1 million tags).

Potential sources of nonuniformity include both the original pool of array-generated oligos as well as biases intrinsic to the targeting protocol itself. To assess the quality of the targeting oligos, we quantified two subsets of targeting oligos, representative of exons observed in both replicates and exons observed in neither, by individual real-time PCR (**Supplementary Fig. 2** online). We found no significant difference between these two categories, suggesting that bias in the targeting oligo pool is not a primary source of nonuniformity.

We also evaluated whether target-specific characteristics such as (G+C) content and target length contributed to dropout (**Supplementary Fig. 3** online). Intermediate (G+C) content, of either targeting arm or the targeted sequence itself, is associated with a lower rate of failed capture. These results are consistent with an analogous analysis of the 480-plex experiment (**Supplementary Fig. 1c**). Shorter targets had a higher rate of failed capture than longer targets, suggesting that gap-fills longer than 191 bp are possible.

Shotgun sequencing of captured exons

To assess this method in the context of an integrated variation discovery pipeline, we converted captured amplicons generated

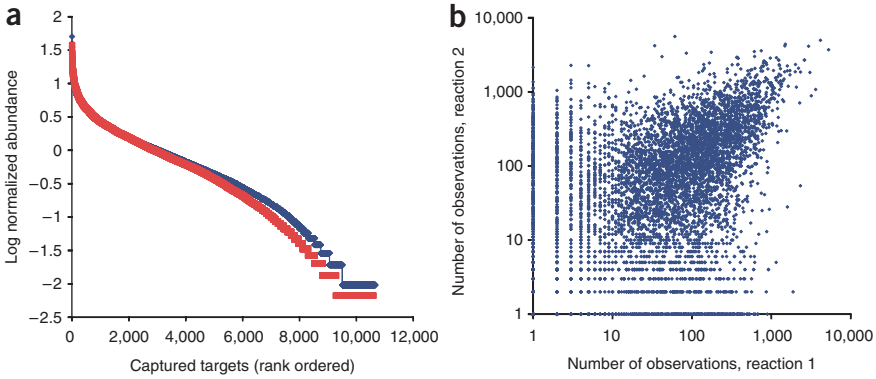

Figure 3 | Quantification of uniformity by deep end-sequencing of captured amplicons. (**a**) Amplicons resulting from the replicate multiplex capture reactions were end-sequenced to a high depth on an Illumina Genome Analyzer (>1 million alignable reads per reaction). Counts were normalized relative to the mean abundance for each reaction. The logs (base 10) of the estimated relative abundances were calculated, sorted and plotted for reaction 1 (blue) and reaction 2 (red). In each reaction, more than 10,000 targets are observed one or more times, but their abundances vary over several orders of magnitude. (**b**) For capture targets observed one or more times in duplicate reactions (n = 5,867), the number of observations of each target is plotted for each reaction. Pearson correlation coefficient = 0.54.

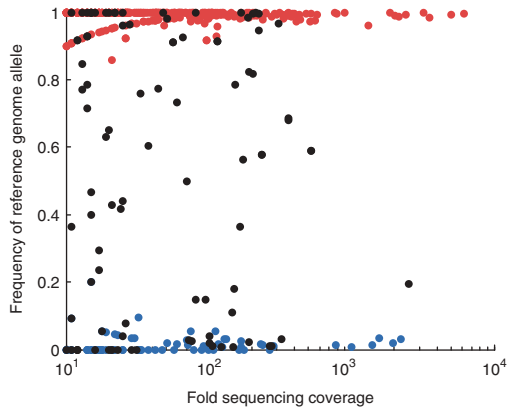

Figure 4 | Validation of exon resequencing data. A shotgun sequencing library generated from the products of multiplex exon capture was sequenced on an Illumina Genome Analyzer to obtain ~125 Mb of sequence that could be confidently aligned to targeted sequences. There are 478 positions within these data (plotted points in the figure) for which we have >10 times coverage and for which HapMap genotyping data are available for this individual at validated SNP positions. The fold coverage for a given position in the resequencing data is plotted versus the frequency in the resequencing data of the allele corresponding to the reference genome sequence. Colors correspond to genotypes in the HapMap data. Red, homozygous, reference genome allele; black, heterozygous; blue, homozygous, nonreference genome allele.

above into a shotgun sequencing library (**Fig. 1d**). With an Illumina Genome Analyzer, we obtained 125 Mb of sequence that could be confidently aligned to targeted exon sequences. There were ~500 kb of genomic positions for which we obtained tenfold or greater coverage. To evaluate the accuracy of these resequencing data, we focused on 478 validated HapMap SNPs, genotyped in the individual from whose DNA the sequencing library was generated (GM12248), that overlapped with positions for which we had generated 10 times or greater coverage (**Fig. 4**). All evaluated positions at which the genotype is homozygous were dominated by the correct base-call ($n = 388$). The data on heterozygous positions ($n = 90$) are more ambiguous. For example, if we were to call positions with allele frequencies between 0.2 and 0.8 as heterozygous, then only 25 of 90 heterozygous positions (31%) would be called correctly. An alternative perspective is that because the goal of variation discovery is often to interrogate for rare variants, the key task is to identify positions that deviate from the reference genome with high sensitivity and specificity. If we call positions with observed allele frequencies of <0.8 as likely to be mutated relative to the reference genome, we can estimate a sensitivity of 63% for detecting rare mutations while maintaining high specificity (sensitivity, 57 of 90 heterozygous positions had a reference allele frequency of less than 0.8; specificity, 0 of 313 homozygous positions had a reference allele frequency of less than 0.8).

DISCUSSION

We developed a method for highly multiplex amplification of arbitrary sets of short sequences from a complex genome. Notable attributes of our approach include: (i) unlike conventional PCR, the capture reaction is compatible with extensive multiplexing, with ~10,000 targets captured in individual reactions; (ii) capture is highly specific, with ~98% of amplicons corresponding to targets; (iii) the method allows for precise specification of target boundaries; and (iv) the use of targeting oligos derived from programmable microarrays greatly reduces the upfront costs of oligo synthesis.

Comparing our approach with the 'Selector' technology[11,17], a key difference is in the constraints on target definition, as the boundaries of targets captured by the Selector method are dependent on the natural distribution of restriction enzyme recognition sites. In contrast, the constraints on this method are analogous to those that govern PCR primer design, for example, avoiding targeting arms with excessively high or low (G+C) content. This may translate into greater ease of use and flexibility. In terms of performance, we note that current implementations of both methods demonstrate high specificity but poor uniformity. Additionally, although we have demonstrated a substantially higher number of concurrent amplifications, pools of array-generated oligos can likely also be used to push the Selector approach (or other methods) to higher multiplexities.

Although this method has great potential, improving the uniformity with which individual targets are captured and amplified remains crucial to fully realize its potential. Steps that may contribute to nonuniformity include: (i) microarray-based synthesis and processing of the targeting oligo library; (ii) stochastic events or systematic, sequence-dependent biases during the capture reaction itself; and (iii) linear rolling circle amplification (RCA) and exponential PCR amplification steps. Thus far, our analysis suggests that biases intrinsic to the capture protocol have a primary role (**Supplementary Figs. 1c**, **2** and **3**).

It is both concerning and interesting that the resequencing data show skewing of base-calls at heterozygous positions away from the expected ratio of 0.5 (**Fig. 4**). This skewing might be expected if stochastic events (for example, a finite number of capture events during the multiplex capture reaction) are a primary source of bias. Given that the concentration of individual targeting oligos in the capture reaction is very low (~0.36 pM per species), it is likely that their hybridization to genomic DNA is inefficient. This problem might therefore be mitigated simply by increasing the concentration of reactants. If either the targeting oligos or genomic DNA can be defined as limiting, one could potentially hybridize with an excess of the complement, that is, to saturation, such that small differences in capture efficiency would be 'normalized' away.

Although this method is generally useful for the multiplex amplification of large sets of short sequences from a complex genome, we chose to focus on human exons. The successful amplification of ~10,000 exons in a single reaction raises the prospect that the full set of annotated protein-coding sequences can be captured and amplified in one or a few reactions. To the extent that nonuniform amplification of targets is systematic and therefore reproducible, the uniformity of a given reaction can be empirically optimized. One could synthesize and evaluate sets of oligos that comprehensively target the full set of annotated protein-coding sequences, for example, twenty 10,000-plex reactions. After characterizing the efficiency of individual capture oligos in each reaction, a second iteration could be carried out, in which capture oligos performing similarly to one another are grouped together within the same oligo mixture (thereby maximizing uniformity within any single reaction). We also note that if a broadly useful oligo targeting mix can

be validated, column-based synthesis followed by pooling at empirically informed concentrations could deliver an even lower cost per genome, provided that the large upfront cost of synthesis could be amortized over a very large number of samples (**Supplementary Table 1**).

If comprehensive sequencing of all exons were feasible for ~$1,000 per genome, what studies would be enabled? First, an emerging approach to study the role of rare, nonsynonymous variants involves direct sequencing of candidate genes in cohorts of patients at the extremes of the phenotypic distribution[18–20]. The natural extension of this paradigm is to move beyond candidate genes to the full complement of protein-coding sequences[21]. Second, the recurrent identification of nonsynonymous, functional somatic mutations in tumors is frequently the means by which a gene is implicated in oncogenesis[22]. Recent studies conducting extensive sequencing of exons in tumor DNA suggest that we identified only a fraction of cancer-related genes[6,23]. Given the scale at which tumor DNA sequencing is being proposed, that is, 'The Cancer Genome Atlas'[24], an early investment in developing multiplex exon targeting methods will likely give rise to an enormous savings of research dollars.

METHODS

Design of 55,000 targeting oligos. Targets were defined as contiguous protein-coding sequences in the human genome (US National Center for Biotechnology Information (NCBI), CCDS 27 February 2007 update for hg36), extended by 2 bp in both the 5′ and 3′ directions (see **Supplementary Methods** online for a description of the structure of array-synthesized 100-mers). Sequences of all 55,000 targeting oligos and their targets are available in **Supplementary Data 1** online. For the 480-plex experiment (**Supplementary Fig. 1**), sequences are in **Supplementary Data 2** online.

Processing of targeting oligo precursors. We synthesized 55,000 oligos (100-mers), released them from a programmable microarray and shipped them lyophilized (Agilent Technologies), at an estimated yield of 0.3 fmol/species. PCR amplification was performed in 200 µl with 2.5 nM oligos (total), 200 µM dNTPs, 400 nM eMIP_CA1_F primer, 400 nM eMIP_CA_1_R primer, 0.8× SybrGreen, 40 units Pfu polymerase in 1× Pfu buffer (Stratagene), at 95 °C for 5 min, eight cycles of 95 °C for 30 s, 55 °C for 2 min, 72 °C for 8 min, and finally 72 °C for 10 min. We column-purified the amplicons and eluted them in 85 µl dH_20.

Release of single-stranded 70 nt targeting oligos. Flanking sequences of 100-mers contained recognition sites for nicking restriction endonucleases at their junctions with the targeting arms. Digestions were as follows: 85 µl column-purified PCR amplicons, 10 µl 10× NEB Buffer 2 (NEB) and 5 µl *Nt.AlwI* (10 U/µl; NEB) were mixed and incubated at 37 °C for 1 h, 80 °C for 20 min. We added 5 µl *Nb.BsrDI* (10 U/µl, NEB) and then incubated at 65 °C for 1 h. We column-purified the reaction, eluted DNA in 100 µl dH_20 and concentrated it in a vacuum centrifuge to 20 µl. We separated this sample on a 6% denaturing acrylamide gel, and recovered DNA from a band corresponding to expected single-stranded 70 nt species, purified it and eluted the DNA in 20 µl dH_20. Gel-based estimation of the total concentration of targeting oligos was 125 nM.

Multiplex capture of targeted sequences. Genomic DNA was derived from a lymphoblastoid cell line from an anonymous individual (CEPH, GM12248). We hybridized targeting oligos to genomic DNA in 20 µl 1× Ampligase buffer (Epicentre), with 1.5 µg genomic DNA and 20 nM targeting oligos (total concentration; ~0.36 pM for each species), incubating the reactions at 20 °C for 4 min, 95 °C for 5 min and 60 °C for 1.5 h. Then we added 1 µl gap-filling mix (200 µM dNTPs, 2 units of *Taq* Stoffel Fragment (Applied Biosystems) and 0.25 units Ampligase in 1× Ampligase buffer), and incubated the reaction at 60 °C for 15 min and 37 °C for 1 min. To degrade linear species, we added 4 µl of exonuclease mix (containing 40 units of exonuclease I and 200 units of exonuclease III; NEB), and incubated the reaction at 37 °C for 15 min and then at 95 °C for 2 min.

Amplification of capture circles. We carried out linear RCA as follows: 5 µl capture reaction and 0.5 µl primer-dNTP mix (RCA_2_RA at 5 µM, dNTPs at 5 mM each) were mixed and incubated at 94 °C for 3 min and 60 °C for 3 min. We added 0.5 µl BST polymerase (8 U/µl; NEB) and incubated the reaction at 60 °C for 1 h and 85 °C for 2 min. We amplified this material by PCR in 50 µl reactions with 2 µl template (linear RCA reaction), 200 µM dNTPs, 200 nM CP2-FA primer, 200 nM CP2-RA primer, 0.8× SybrGreen and 2.75 units Amplitaq Gold in 1× Amplitaq Gold PCR buffer with 1.5 mM $MgCl_2$ (Applied Biosystems) at 95 °C for 10 min, 23 cycles of 95 °C for 30 s, 55 °C for 2 min, 72 °C for 8 min, and finally 72 °C for 10 min. We separated the resulting amplicons on a 6% nondenaturing polyacrylamide gel (**Fig. 2**). We recovered amplicons corresponding to the expected size range (140–271 bp), purified them and resuspended the products in 20 µl TE (pH 8.0).

Evaluation for specificity and uniformity. We subjected 10% of the recovered amplicons to an additional eight cycles of PCR (as above) and subcloned them with the TOPO TA cloning kit (Invitrogen). Products of colony PCR were submitted to a core facility for Sanger sequencing. Alignment of sequencing reads to the human genome was performed with NCBI Blast. We prepared a library for 'end sequencing' by appending Solexa adaptor sequences during eight cycles of PCR of gel-purified amplicons. We used a custom sequencing primer such that the ~35 bp reads began at the first base of the variable targeting arm. Sequencing was performed on the Illumina Genome Analyzer according to manufacturer's instructions. For the duplicate reactions, we obtained 2,247,111 and 2,547,479 sequencing reads, of which 1,109,756 and 1,584,331 could be confidently mapped back to one of the 55,000 targets.

Additional methods. Oligo sequences, and methods related to the shotgun sequencing of capture products and the analysis of Solexa sequencing data are available in **Supplementary Methods**.

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

ACKNOWLEDGMENTS

This work was supported by a Center for Excellence in Genome Sciences grant from the National Human Genome Research Institute, and a SPARC grant from the Broad Institute of Massachusetts Institute of Technology and Harvard University. We are grateful to G. Buck, M. Davis, N. Sheth, C. Childress, Jr. and J. Noble (Center for High Performance Computing and Center for the Study of Biological Complexity, Virginia Commonwealth University) for setting up the Illumina Genome Analyzer

analysis pipeline. We thank H. Ji, S. Fredriksson, A. Gnirke, E. Lander, D. Jaffe and C. Nusbaum for discussions.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* **5**, 335–344 (2004).
2. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
3. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
4. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
5. Stephens, M., Sloan, J.S., Robertson, P.D., Scheet, P. & Nickerson, D.A. Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat. Genet.* **38**, 375–381 (2006).
6. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
7. Edwards, M.C. & Gibbs, R.A. Multiplex PCR: advantages, development, and applications. *PCR Methods Appl.* **3**, S65–S75 (1994).
8. Markoulatos, P., Siafakas, N. & Moncany, M. Multiplex polymerase chain reaction: a practical approach. *J. Clin. Lab. Anal.* **16**, 47–51 (2002).
9. Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **21**, 673–678 (2003).
10. Hardenbol, P. *et al.* Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Res.* **15**, 269–275 (2005).
11. Dahl, F. *et al.* Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc. Natl. Acad. Sci. USA* **104**, 9387–9392 (2007).
12. Fredriksson, S. *et al.* Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. *Nucleic Acids Res.* **35**, e47 (2007).
13. Bashiardes, S. *et al.* Direct genomic selection. *Nat. Methods* **2**, 63–69 (2005).
14. Okou, D.T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods,* advance online publication 14 October 2007 (doi:10.1038/nmeth1109).
15. Albert, T.J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat. Methods,* advance online publication 14 October 2007 (doi:10.1038/nmeth1111).
16. Tian, J. *et al.* Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* **432**, 1050–1054 (2004).
17. Dahl, F., Gullberg, M., Stenberg, J., Landegren, U. & Nilsson, M. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res.* **33**, e71 (2005).
18. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* **37**, 161–165 (2005).
19. Farooqi, I.S. *et al.* Clinical and molecular genetic spectrum of congenital deficiency of the leptin receptor. *N. Engl. J. Med.* **356**, 237–247 (2007).
20. Romeo, S. *et al.* Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.* **39**, 513–516 (2007).
21. Topol, E.J. & Frazer, K.A. The resequencing imperative. *Nat. Genet.* **39**, 439–440 (2007).
22. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
23. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
24. Collins, F.S. & Barker, A.D. Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart anew course across the complex landscape of human malignancies. *Sci. Am.* **296**, 50–57 (2007).