Direct selection of human genomic loci by microarray hybridization

Thomas J Albert[1], Michael N Molla[1], Donna M Muzny[2], Lynne Nazareth[2], David Wheeler[2], Xingzhi Song[2], Todd A Richmond[1], Chris M Middle[1], Matthew J Rodesch[1], Charles J Packard[1], George M Weinstock[2] & Richard A Gibbs[2]

We applied high-density microarrays to the enrichment of specific sequences from the human genome for high-throughput sequencing. After capture of 6,726 approximately 500-base 'exon' segments, and of 'locus-specific' regions ranging in size from 200 kb to 5 Mb, followed by sequencing on a 454 Life Sciences FLX sequencer, most sequence reads represented selection targets. These direct selection methods supersede multiplex PCR for the large-scale analysis of genomic regions.

The targeted enrichment of specific regions in complex genomes is a fundamental practice of modern molecular biology. For the last two decades, PCR, in conjunction with innumerable procedures for detection, characterization and manipulation of specific sequences, has been the dominant enrichment technology[1]. Despite this success, the vast size of mammalian genomes (~3 Gb), coupled with the wide spatial distribution of individual elements (for example, ~20,000 scattered genes) and the upper limit of reliable PCR amplification length (5–50 kb), has made it difficult to perform very large surveys in single experiments. Meanwhile, new methods for rapid and inexpensive production of dense DNA microarrays[2,3], and for highly parallel DNA sequencing[4] have been introduced. We combined these new developments to enable the efficient targeted capture and sequencing of specific DNA sequences from complex genomes to provide an alternative to PCR-based methods.

Direct genomic selection for sequence enrichment has previously been demonstrated using cDNA or bacterial artificial chromosome (BAC) clones. Two rounds of hybridization, elution and amplification of fragmented genomic DNA to a biotinylated BAC clone had recovered ~50% of sequenced clones from the enriched fraction corresponding to the targeted region[5]. The remaining sequences represent genomic repeats or other distant sequences, showing that in addition to the burden of manipulating the large BACs, the presence of repeat sequences in the selecting DNA compromized the final enrichment.

Here we report the design of custom high-density oligonucleotide microarrays (NimbleGen) to capture both dispersed short genome segments, encompassing individual gene exons, and single long segments, corresponding to entire gene loci. The 'exonic' design aimed to capture DNA representing 6,726 genomic regions (**Supplementary Table 1** online; minimum size 500 bases, ~5 Mb of total sequence) from 660 genes dispersed throughout the genome, and a second capture array series targeted areas of 200 kb, 500 kb, 1 Mb, 2 Mb and 5 Mb surrounding the human *BRCA1* gene locus. Each of these microarrays was designed with long oligonucleotide probes (>60 bases) spaced on average between 1 and 10 bases apart, depending on the locus size covered by the particular array, and excluded regions identified by 'repeat mask' computer software as well as all additional probes that were not unique in the human genome.

To test the reproducibility of the capture system, we used the 'exonic' design to capture fragments from a commonly used human cell line (Burkitt's Lymphoma cell line NA04671). First we subjected the genomic DNA to whole-genome amplification and sonication. Next we ligated linkers to the ends of the approximately 500-base fragments for subsequent PCR amplification and hybridized single-stranded products to the capture arrays. We removed non-hybridized materials by washing, eluted the retained single-stranded fragments and amplified them by ligation-mediated PCR, followed by quantitative PCR directed to eight of the target regions (**Fig. 1** and **Supplementary Methods** online). After a single round of microarray capture, an average of 378-fold enrichment was achieved for three capture replicates (because 5 Mb of total sequence was targeted for exon capture, the theoretical maximum enrichment level was 3,000,000,000/5,000,000 or 600-fold).

DNA sequencing of each of the three replicate 'exonic' capture products on the 454 FLX instrument generated 63, 115 and 93 Mb of total sequence. BLAST analysis showed that 91, 89 and 91% of

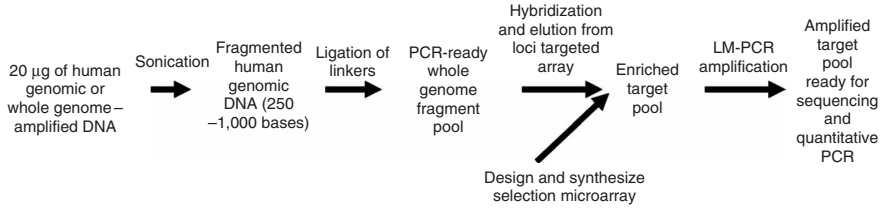

Figure 1 | Schematic of the microarray direct genomic selection scheme. LM-PCR, linker-mediated PCR.

[1]NimbleGen Systems Inc., 1 Science Court, Madison, Wisconsin 53711, USA. [2]Human Genome Sequencing Center, One Baylor Plaza, Houston, Texas 77030, USA. Correspondence should be addressed to T.J.A. (talbert@nimblegen.com) or R.A.G. (agibbs@bcm.tmc.edu).

RECEIVED 26 JULY; ACCEPTED 24 SEPTEMBER; PUBLISHED ONLINE 14 OCTOBER 2007; CORRECTED ONLINE 21 OCTOBER (DETAILS ONLINE); DOI:10.1038/NMETH1111

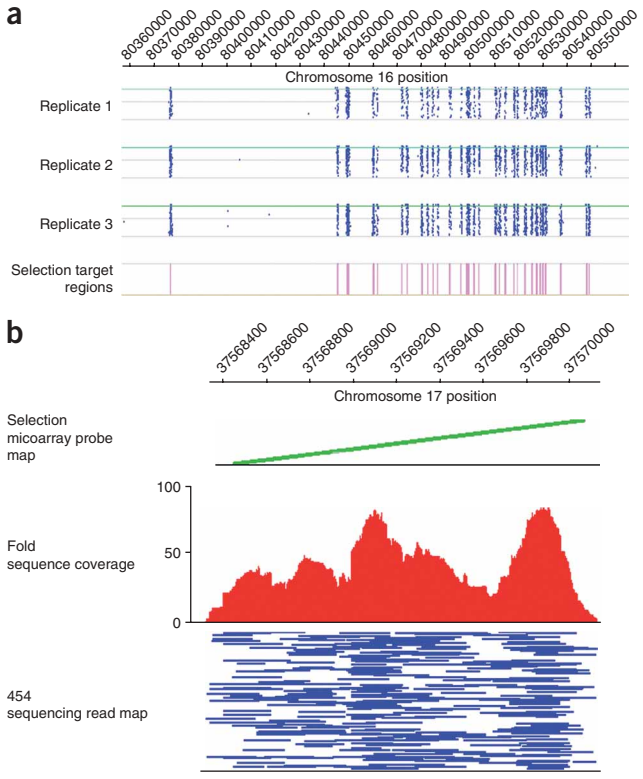

Figure 2 | Captured sequences. (**a**) Illustrative sequence read map detail of 'exonic fragments' from a segment of chromosome 16 from three microarray genomic selection replicates, indicating the reproducibility of targeted sequencing. Blue dots indicate the position of the best blast score for individual 454 sequencing reads, and pink bars indicate positions targeted by selection microarray probes. (**b**) Sequence read map detail of ~2,000 bases of chromosome 17 from a microarray selection of a 2-Mb contiguous region that contains the *BRCA1* gene. Microarray selection probe positions are indicated by green lines. Probes are spaced every 10 bases and staggered along the *y*-axis to aid visualization. Sequence coverage from one 454 run is indicated as per base (red) and as per read (blue).

reads, respectively, mapped back uniquely to the genome; 75, 65 and 77% were from targeted regions, and 96, 93 and 95% of target sequences contained at least one sequence read (**Supplementary Table 1**). This represents an average enrichment of 432 fold. We mapped sequencing reads back to the genome to generate read maps (**Fig. 2**) and calculated the per-base sequence coverage for the aggregate capture region (**Fig. 3**). The median per-base coverage for each sample was 5, 7 and 7 fold, respectively.

To determine our ability to discover variation in the human genome, we also captured and sequenced DNA from four samples derived from the human HapMap collection. We treated the samples as above, but did not subject them to whole-genome amplification before capture. The capture results were similar to those above, except that sequence coverage was consistently more uniform than before. This may reflect bias introduced during whole-genome amplification. We assembled the sequence from the four HapMap samples, identified the mutations and compared them to the HapMap single-nucleotide polymorphism data for each sample (**Supplementary Table 2** online). Between 94 and 79% of the variant positions known in the HapMap samples were identified with at least one sequence read, which was expected based on the overall sequence coverage. There was no apparent bias against the alleles that were not present on the capture array when we compared the coverage of targets that contained 0, 1 or >1 known variants (7.95, 8.48 and 8.82-fold coverage, respectively).

To analyze large contiguous genomic regions we tested the capture microarrays targeting segments from 200 kb to 5 Mb surrounding the human *BRCA1* gene with the human Burkitt's Lymphoma cell line NA04671 DNA. All capture targets performed well, with up to 140 Mb of raw sequence generated in a single sequencing machine run, generating ~18-fold coverage, from a 5-Mb capture region (**Supplementary Table 3** online). Note that the percentage of reads that map to the target sequence increased with the size of the target region.

Lists of exon target regions (**Supplementary Table 4** online) and microarray probe sequences (**Supplementary Data** online) are available.

The power of microarray-based direct selection methods for enrichment of targeted sequences is illustrated by these data. In addition to the specificity of the assay, the high yields of the downstream DNA sequencing steps are consistently superior to

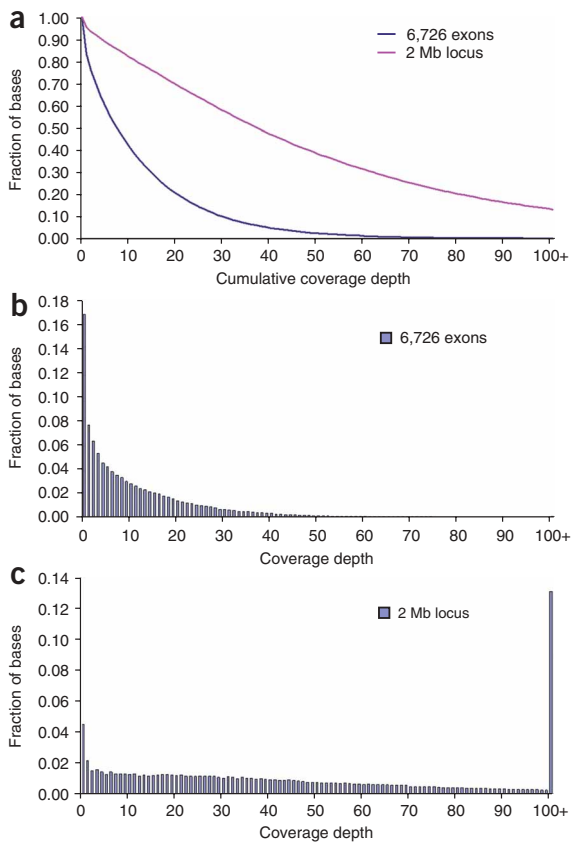

Figure 3 | Sequence coverage depth. (**a**) Fraction of bases of each aggregate target region and the corresponding cumulative depth of sequence coverage after one 454 FLX run. 'Exon' sample represents 6,726 exon-sized regions. The 2-Mb *BRCA1* region was targeted from positions 37490417 to 39490417 on human chromosome 17. Only the unique genome fraction was targeted by selection probes. (**b**) Histogram of per-base sequence coverage depth for the 'exon' example. (**c**) Histogram of per-base coverage depth for the 2-Mb locus example.

the routine average performance using non-captured DNA sources. This is attributed to the capture-enrichment process providing a useful purification of unique sequences away from repeats and other impurities that can confound the first emulsion PCR step of the 454 sequencing process.

Recently a study using 'padlock probes' demonstrated the ability to select and amplify 177 exons from the human genome in parallel followed by 454 sequencing[6]. This method provided high sequence coverage and allowed for variant discovery in the target regions, but required the serial synthesis of hundreds of long oligonucleotides and subsequent re-optimization to amplify several target loci. The limit of scalability of the method is not known. An approach using a variation of 'molecular inversion probes' is described in this issue of *Nature Methods*[7].

Another report in this issue describes a similar method to the one we describe here to select genomic loci using microarrays, which are then sequenced by subsequent microarray hybridization[8]. This technique was able to identify single-nucleotide polymorphisms from HapMap samples with >99% accuracy, and also demonstrates that the capture arrays can be reused.

The data presented here and in reference 8 use a programmable high-density array platform for the enrichment of targeted sequences. With 385,000 probes we were readily able to capture at least 5 Mb of total sequence. A method facilitating parallel analysis of all human exons is presently under development using a 2.1 million feature capture array.

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

ACKNOWLEDGMENTS

We thank M. Zwick and D. Okou for helpful conversations, and the US National Human Genome Research Institute (grant U54 HG003273) and the US National Cancer Institute (grant R21HG004553) for support for this effort.

AUTHOR CONTRIBUTIONS

T.J.A., G.M.W. and R.A.G., experimental design, analysis and interpretation; M.N.M. and T.A.R., experimental design and data analysis; D.M.M. and L.N., sequencing; D.W. and X.S., data analysis; C.M.M. and C.J.P., laboratory experimentation (arrays); M.J.R., experimental design and laboratory experimentation (arrays).

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. Mullis, K.B. & Faloona, F.A. *Methods Enzymol.* **155**, 335–350 (1987).
2. Nuwaysir, E.F. *et al. Genome Res.* **12**, 1749–1755 (2002).
3. Albert, T.J. *et al. Nucleic Acids Res.* **31**, e35 (2003).
4. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
5. Bashiardes, S. *et al. Nat. Methods* **2**, 63–69 (2005).
6. Dahl, F. *et al. Proc. Natl. Acad. Sci. USA* **104**, 9387–9392 (2007).
7. Porreca, G.J. *et al. Nat. Methods* advance online publication 14 October 2007 (doi:10.1038/nmeth1110).
8. Okou, D.T. *et al. Nat. Methods* advance online publication 14 October 2007 (doi:10.1038/nmeth1109).