Microarray-based genomic selection for high-throughput resequencing

David T Okou[1], Karyn Meltz Steinberg[1,2],
Christina Middle[3], David J Cutler[1],
Thomas J Albert[3] & Michael E Zwick[1,2]

We developed a general method, microarray-based genomic selection (MGS), capable of selecting and enriching targeted sequences from complex eukaryotic genomes without the repeat blocking steps necessary for bacterial artificial chromosome (BAC)-based genomic selection. We demonstrate that large human genomic regions, on the order of hundreds of kilobases, can be enriched and resequenced with resequencing arrays. MGS, when combined with a next-generation resequencing technology, can enable large-scale resequencing in single-investigator laboratories.

Technological innovation in DNA sequencing offers the promise of a more comprehensive, cost-effective and systematic ascertainment of genetic variation[1–5]. A major bottleneck, however, is in isolating the target DNA to be sequenced. Complex eukaryotic genomes, such as the human genome, are too large to explore without complexity reduction using methods that directly amplify specific sequences. Current approaches for target DNA isolation include short PCR[6,7], long PCR[1,5], fosmid library construction and selection[8], transformation-associated recombination (TAR) cloning[9,10], selector technology[11] and direct genomic selection with BACs[12]. PCR using primer pairs complementary to specific genomic regions of interest is still the most common method of sample preparation, but it is difficult to scale to large genomic regions, is labor-intensive and, when primers are multiplexed, is subject to failure or artifacts. Random clone-based methods offer the advantage of obtaining complete haplotypes but are relatively expensive to scale.

Direct genomic selection, using BAC clones as hybridization 'hooks', has previously demonstrated the ability to isolate specific genomic regions without requiring specific amplification[12], but its adoption has been limited. Because BAC clones consist of many highly repetitive sequences, several protocol steps are required to maximize the enrichment of these types of sequences. Furthermore, because a single BAC is the unit of selection, isolating discontinuous unique sequence regions from across the genome would require multiple BACs. Finally, the existing protocol depends upon the presence of restriction sites adjacent to the targeted regions of interest that produce sticky ends for the ligation of generic adaptors. This limits coverage in regions lacking these restriction sites. Although random shearing followed by repair has been mentioned as a possible alternative approach, it has not been demonstrated[12].

To address these challenges, we developed MGS, a method capable of isolating user-defined unique genomic sequences from complex eukaryotic genomes. The MGS protocol consists of five main steps: (i) physical shearing of genomic DNA to create random fragments with an average size of 300 bp, (ii) end repair of the fragments, which includes adding 3′-adenine overhangs, followed by ligation to unique adaptors with complementary thymine overhangs, (iii) fragment hybridization and capture using a custom high-density oligonucleotide microarray consisting of complementary sequences identified from a reference genome sequence, (iv) elution of fragments bound to the probes, and (v) amplification of selected fragments through one round of PCR using the adaptors as a single set of primers (**Fig. 1**). The complete protocol is outlined in **Supplementary Methods** online.

To test MGS, we captured and resequenced two X chromosome–linked genomic regions (**Fig. 2**). In an initial experiment we examined a 50-kb region that included coding and noncoding sequences surrounding the fragile X mental retardation 1 gene (*FMR1*). In a second, larger-scale experiment we isolated and resequenced 304 kb of unique coding and noncoding sequences contained within a 1.7-Mb genomic region that includes *FMR1*, *FMR1NB* and *AFF2*. Each custom MGS array consisted of ~385,000 long oligonucleotide capture probes covering the regions of interest; the arrays were manufactured by NimbleGen Systems, Inc. Capture probe sequences included both the forward and reverse strands manufactured on a standard commercially available microarray to our specifications (**Supplementary Data 1** and **2** online). For the 50-kb region, there were four pairs of probes for every targeted base, and the 304-kb region had one pair of probes for every 1.5 targeted bases. The capture oligonucleotides were 50–93 bp and were designed to achieve optimal isothermal hybridization across the microarray.

We processed 20 μg of whole genome amplified genomic DNA for each sample using the MGS protocol. Upon eluting the selected target from the capture MGS chip, we obtained yields of 700 ng to 1.2 μg. We split the eluted sample into 5–10 PCRs, each of which was carried out using high fidelity *Taq* polymerase at an optimal concentration of 3 ng/μl of PCR template. We could reuse the MGS capture chips at least once with no apparent contamination or effect on data quality (data not shown).

To assess MGS, we first sought to resequence a 50-kb genomic region containing the *FMR1* locus in cell lines derived from two patients with known *FMR1* mutations: patient sample Tr91

[1]Department of Human Genetics, Emory University School of Medicine, 615 Michael Street, Suite 301, Atlanta, Georgia 30322, USA. [2]Program in Population Biology, Ecology and Evolution, Emory University, 1510 Clifton Road, Atlanta, Georgia 30322, USA. [3]NimbleGen Systems, Inc., 1 Science Court, Madison, Wisconsin 53711, USA. Correspondence should be addressed to M.E.Z. (mzwick@genetics.emory.edu).

RECEIVED 27 JULY; ACCEPTED 20 SEPTEMBER; PUBLISHED ONLINE 14 OCTOBER 2007; DOI:10.1038/NMETH1109

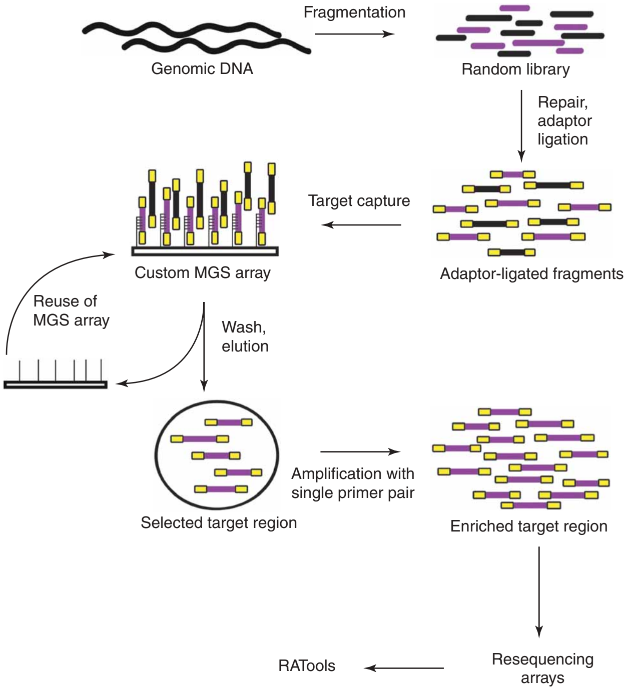

Figure 1 | Microarray-based genomic selection and resequencing of complex genomes. Sheared genomic fragments are repaired and ligated to generic adaptors. Hybridization to a custom designed high-density oligonucleotide microarray allows the capture of the target DNA regions. The selected target is eluted and amplified using a one-step PCR and a single primer pair per template. Amplified targets were resequenced with resequencing arrays and analyzed with RATools.

contains a disease-causing point mutation (A to T) at position 146825745 on the X chromosome, and sample DM316 harbors a large deletion in *FMR1* (refs. 13,14). We designed a custom NimbleGen 50-kb resequencing array that covered the targeted regions, containing both coding and noncoding sequences in the vicinity of *FMR1* (**Fig. 2**), and resequenced both samples in triplicate using resequencing array (**Supplementary Data 3** online). Analysis of the Tr91 sequence identified the expected A to T point mutation when compared to the human genome reference sequence in all three replicates. Six additional variants were detected in Tr91, 5 of which were successfully validated by independent sequencing (Agencourt Bioscience; see **Supplementary Methods** and **Supplementary Table 1** online). As we expected, each of the three DM316 replicates exhibited an absence of hybridization on the resequencing array in the regions corresponding to the known deleted sequences (**Supplementary Fig. 1** online).

To evaluate MGS on a larger genomic region, we selected a total of 304 kb from 10 individual genomes represented by two populations of different ancestry: a European descent population ($n = 5$) selected from the Centre d'Etude du Polymorphisme Humain (CEPH) panel and an African descent population ($n = 5$) selected from the Hapmap (Coriell Cell Repository numbers are available in **Supplementary Methods**). We replicated MGS twice for each of the ten samples. Using quantitative PCR, we estimated that MGS enriched targeted sequences ~1,000-fold (**Supplementary Fig. 2** online).

Our resequencing results provide three lines of evidence demonstrating the efficacy of our MGS protocol (**Supplementary Data 4** online). First, our total base-calling rate over all 20 replicates (10 samples, each processed twice) was 99.1% (6,528,393 called out of 6,585,832 total). This very high level of coverage implies that our MGS protocol efficiently enriches for the variety of sequences contained in the genomic regions we targeted. Second, for each sample we counted the number of bases called identically and differently between both replicates. The reproducibility of resequencing array base calls was 99.98%. Third, for each sample, to assess accuracy of base calls, we compared our resequencing array base calls with genotype calls generated by the HapMap project (http://www.hapmap.org). We initially observed 39 discrepancies between resequencing array and HapMap genotype calls. To identify the nature of the discrepancy, we resequenced each of them independently via conventional ABI chemistry (Agencourt Bioscience). The resulting sequence data showed that 27 of the discrepancies agreed with our resequencing array call, and 12 agreed with the HapMap genotype call. Hence, more than two-thirds of the discrepancies we observed arose because of errors in HapMap genotyping. Our final accuracy at segregating sites was thus 99.81%.

The MGS protocol we describe uses routine enzymatic reactions and protocols that increase efficiency while minimizing risk of contamination and artifacts. The capture arrays are standard high-density long oligonucleotide arrays and are commercially available. The user can design the array to select multiple unique sequence fragments located throughout the genome for resequencing or to comprehensively resequence genomic regions without the repeat blocking step necessary for BAC genomic selection. We are continuing to pursue the trade-off between probe density and sequence coverage and to increase the level of enrichment. Our current MGS microarrays contain ~385,000 capture probes. Presently available state-of-the-art microarrays contain 2.1 million features, and arrays with 4.5 million probes will be available in the near future. We believe that obtaining high coverage from genomic regions on the scale of megabases will soon be feasible.

Owing to the quality and comprehensive coverage of the data obtained, we believe that MGS, in addition to other general methods of multiplex amplification[11,15] or sample enrichment[16], will significantly contribute to a future where single-investigator laboratories, with limited infrastructure and relatively few

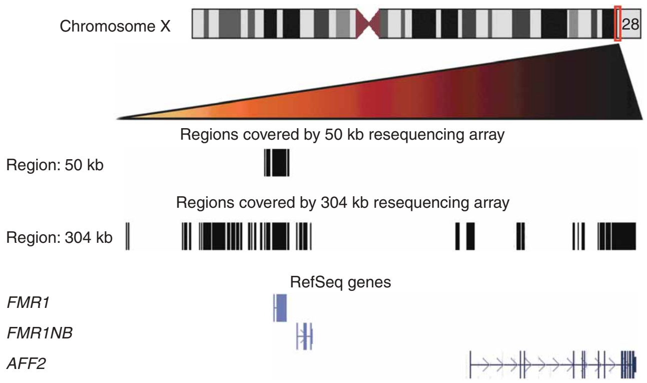
Chromosome X
28
Regions covered by 50 kb resequencing array
Region: 50 kb
Regions covered by 304 kb resequencing array
Region: 304 kb
RefSeq genes
FMR1
FMR1NB
AFF2

Figure 2 | Genomic regions (50 kb, 304 kb) resequenced in the two MGS validation experiments. Targeted sequences included both coding and unique noncoding genome sequences adjacent to Xq28. The shaded triangle represents an expanded view of regions resequenced.

personnel, will be able to generate sequences at levels comparable to a conventional genome sequencing center. The ability of MGS to select multiple targets permits a comprehensive large-scale resequencing of user-defined genomic regions that provide potentially important clues to the pathogenesis of complex diseases[7], or to find human genetic variation and functional sequences in coding as well as noncoding regions[11]. Our method is useful for candidate-gene studies that have been limited by sequencing capabilities and offers the opportunity to select hundreds of genes in known pathways for resequencing. MGS would be useful for studying other eukaryotic model systems (that is, mouse, zebrafish, fruit fly) to speed the sequencing of regions known to contain induced mutations. Finally, although we chose to use resequencing arrays, our approach is quite general. With continuing improvements in levels of enrichment, it should be possible to incorporate MGS into existing sample-preparation pipelines for instruments from Solexa[17] or 454 (refs. 2,16), enabling even greater throughput at lower costs in the near future.

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

ACKNOWLEDGMENTS
Funding for this work was provided by the US National Institutes of Health (NIH) National Institute of Mental Health R01 MH076439-03 (MEZ) and by the Gift Fund, NIH.

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. Cutler, D.J. *et al. Genome Res.* **11**, 1913–1925 (2001).
2. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
3. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. *Nat. Rev. Genet.* **5**, 335–344 (2004).
4. Shendure, J. *et al. Science* **309**, 1728–1732 (2005).
5. Zwick, M.E. *et al. Genome Biol.* **6**, R10 (2005).
6. Hinds, D.A. *et al. Science* **307**, 1072–1079 (2005).
7. Sjoblom, T. *et al. Science* **314**, 268–274 (2006).
8. Raymond, C.K. *et al. Genomics* **86**, 759–766 (2005).
9. Raymond, C.K., Sims, E.H. & Olson, M.V. *Genome Res.* **12**, 190–197 (2002).
10. Kouprina, N., Noskov, V.N. & Larionov, V. *Methods Mol. Biol.* **349**, 85–101 (2006).
11. Dahl, F. *et al. Proc. Natl. Acad. Sci. USA* **104**, 9387–9392 (2007).
12. Bashiardes, S. *et al. Nat. Methods* **2**, 63–69 (2005).
13. De Boulle, K. *et al. Nat. Genet.* **3**, 31–35 (1993).
14. Gu, Y., Lugenbeel, K.A., Vockley, J.G., Grody, W.W. & Nelson, D.L. *Hum. Mol. Genet.* **3**, 1705–1706 (1994).
15. Porreca, G.J. *et al. Nat. Methods*, advance online publication 14 October 2007 (doi:10.1038/nmeth1110).
16. Albert, T.J. *et al. Nat. Methods*, advance online publication 14 October 2007 (doi:10.1038/nmeth1111).
17. Bentley, D.R. *Curr. Opin. Genet. Dev.* **16**, 545–552 (2006).