Open Access Bioinformatics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

METHODOLOGY

Microarray oligonucleotide probe designer: a Web service

This article was published in the following Dove Press journal:
Open Access Bioinformatics
1 November 2010
Number of times this article has been viewed

Viren C Patel[1*]
Kajari Mondal[1*]
Amol Carl Shetty[1*]
Vanessa L Horner[1]
Jirair K Bedoyan[2]
Donna Martin[2,3]
Tamara Caspary[1]
David J Cutler[1]
Michael E Zwick[1]

[1]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA; [2]Department of Pediatrics, [3]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA; *These authors contributed equally to this work.

Abstract: Methods of genomic selection that combine high-density oligonucleotide microarrays with next-generation DNA sequencing allow investigators to characterize genomic variation in selected portions of complex eukaryotic genomes. However, choosing the specific oligonucleotides to be used can pose a major technical challenge. To address this issue, we have developed a software package called MOPeD (microarray oligonucleotide probe designer), which automates the process of designing genomic selection microarrays. This Web-based software allows individual investigators to design custom genomic selection microarrays optimized for synthesis with Roche NimbleGen's maskless photolithography. Design parameters include uniqueness of the probe sequences, melting temperature, hairpin formation, and the presence of single-nucleotide polymorphisms. We generated probe databases for the human, mouse, and rhesus macaque genomes and conducted experimental validation of MOPeD-designed microarrays in human samples by sequencing the human X chromosome exome, where relevant sequence metrics indicated superior performance relative to a microarray designed by the Roche NimbleGen proprietary algorithm. We also performed validation in the mouse to identify known mutations contained within a 487-kb region from mouse chromosome 16, the mouse chromosome 16 exome (1.7 Mb), and the mouse chromosome 12 exome (3.3 Mb). Our results suggest that the open source MOPeD software package and Web site (http://moped.genetics.emory.edu/) will make a valuable resource for investigators in their sequence-based studies of complex eukaryotic genomes.

Keywords: genomic selection, oligonucleotide, microarray, next-generation sequencing, software

Introduction

Next-generation sequencing platforms enable individual investigators to harness enormous raw sequencing power at a dramatically lower cost per sequenced base than traditional Sanger sequencing.[1,2] Although sequencing complete eukaryotic genomes can still be prohibitively expensive for many types of studies, the recent development and validation of methods of isolating target DNA from complex eukaryotic genomes offer a way forward for many investigators[3–10] (see review by Mamanova et al[11]). These methods have been used recently to perform targeted next-generation sequencing of human exomes to identify causative variants underlying monogenic disorders.[12,13] Similarly, it appears that targeted next-generation sequencing to reveal mutations induced in forward genetic screens of model organisms is bound to be used in a progressively increasing extent to identify causative mutations. Ultimately, given a reference genome sequence, these improved methods of target DNA isolation combined

Correspondence: Michael E Zwick
Department of Human Genetics, Emory University School of Medicine, Whitehead Biomedical Research Building, Suite 301, Atlanta, GA 30322, USA
Tel +1 404 727 9924
Fax +1 404 727 3949
Email mzwick@emory.edu

with next-generation sequencing platforms will allow a more complete and comprehensive ascertainment of DNA sequence variation.

Nevertheless, to fully realize this experimental paradigm, an investigator must obtain a specialized, often custom-designed set of reagents. The development of maskless array synthesis allows the custom design and production of high-density oligonucleotide microarrays.[14] The central challenge is then the selection of specific oligonucleotides to be placed on a genomic selection array.[4,9] Although there are a number of algorithms for designing tiling arrays for genome-wide transcriptome or ChIP-Seq experiments[15–18] (see review by Lemoine et al[19]), we still lack easily accessible open source tools for building genomic selection arrays.

To address this issue, we have developed a software package named MOPeD (microarray oligonucleotide probe designer), which automates the process of designing genomic selection microarrays. This Web-based software allows individual investigators to easily design custom genome capture arrays that have been optimized for maskless array synthesis by Roche NimbleGen (Madison, WI). Here, we experimentally validated the performance of MOPeD-designed genomic selection microarrays by sequencing the human X chromosome exome, a mouse chromosome 16 genomic region, and the mouse chromosome 12 and 16 exomes. Our data show that MOPeD can provide investigators a valuable resource for their sequence-based studies of complex eukaryotic genomes.

Materials and methods

The MOPeD software package is implemented in two parts. The first part involves creation of a probe database for a specific reference genome. Operations in this first part are required once for a specific genome. The second part involves obtaining user parameters, querying the previously created probe database, and selecting optimal probes for target regions. This process may be repeated for the design of different microarrays. MOPeD was developed in C and Perl and is licensed under GPL 3.0. The source code is available in the MOPeD Web site (http://moped.genetics.emory.edu/) and via SourceForge (http://moped.sourceforge.net).

Construction of the MOPeD probe database

Creation of the probe database for a specific reference genome is implemented in two steps (Figure 1). The first step involves creation of a database that contains the count of every *k*-mer ($k = 10–15$) in the given genome. The second step involves computation of attributes for both forward and reverse probes of size 55–65 bp. UCSC reference assemblies for human (hg18), mouse (mm9), and rhesus macaque (rheMac2), along with their respective dbSNP tracks, were used for the current implementation of MOPeD.

Construction of *k*-mer database

We constructed a database containing the count of all *k*-mers in a given genome, where *k* ranges from 10 to 15. Each *k*-mer is given an index from 0 to $4^k - 1$ according to its alphabetical position. In this scheme, a 10-mer consisting of all As would have index 0, and a 10-mer consisting of all Ts would have index $4^{10} - 1$. A 15-mer consisting of all As would also have index 0; however, a 15-mer consisting of all Ts would have index $4^{15} - 1$. Distinct files were used for each *k*. This facilitated searching and locating the count of any particular *k*-mer. This database was used to compute a weighted score that estimates the uniqueness of probes.

Computation of probe attributes

The final probe database contained all possible probes of size *n* ($n = 55–65$) from the genome of interest, excluding any probes that contained N (unknown base). The database stored four attributes of each probe: the potential to form hairpin structures, a weighted uniqueness score, the number and positions of single-nucleotide polymorphisms (SNPs), and the Roche NimbleGen synthesis cycle length.

Hairpin

Each probe was tested for the potential to form a hairpin structure by computing the cumulative melting temperature (T_m) of Watson–Crick pairings in the preloop and postloop segments for varying sizes of preloop, postloop, and loop segments. If the cumulative T_m exceeded a predefined T_m limit (eg, annealing temperature), the probe was considered a candidate for hairpin formation and noted as such. The T_m of the probes was calculated for oligonucleotides bound to a surface using the model and parameters described.[20] T_m limit was 40°C.

Uniqueness score

A weighted uniqueness score for each forward and reverse probe was computed. The weighting scheme gave proportionally more weight to larger *k*-mers ($k = 10–15$), because larger *k*-mers are more unique in the genome. For each probe of size n (n = 55–65), all possible *k*-mers present in the probe are extracted and their counts obtained from the

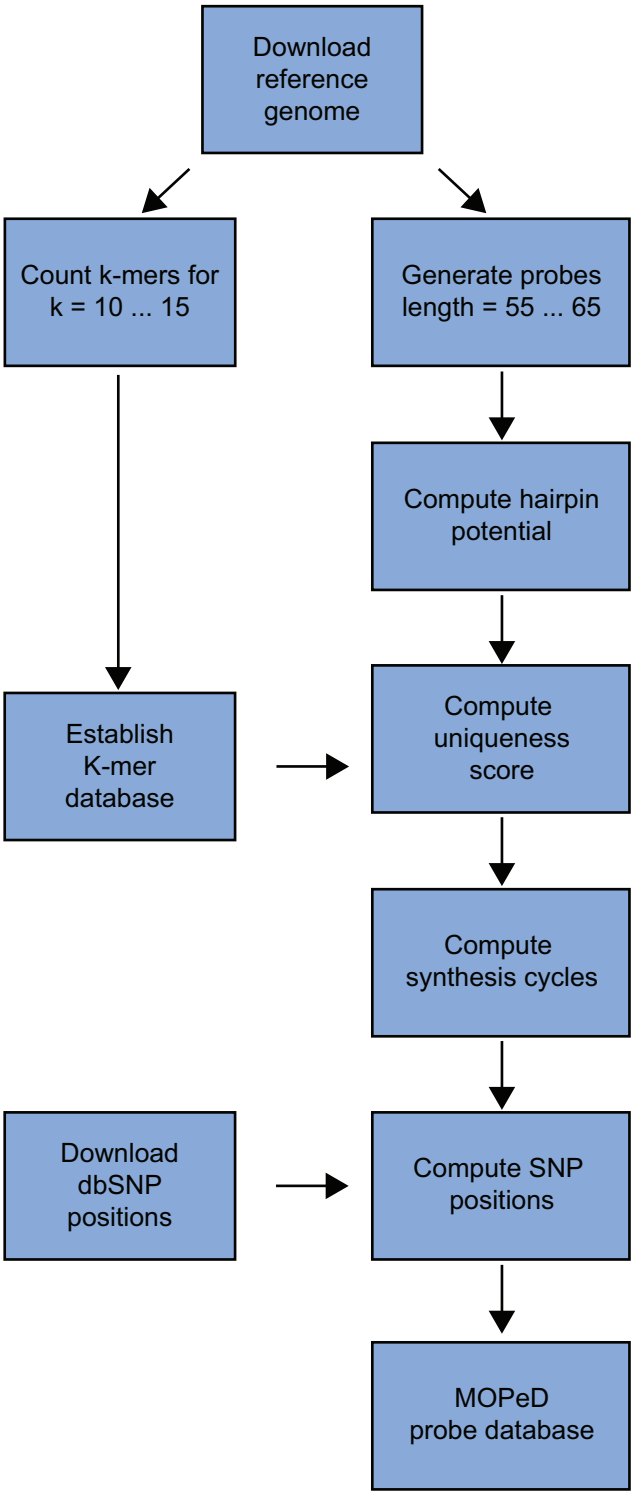

Figure 1 Steps required to generate the MOPeD probe database.

k-mer database. The counts were summed and divided by the number of *k*-mers to provide an average score. The score was further adjusted to account for the larger counts associated with smaller *k*-mers. In this scheme, lower score values indicate higher specificity of the probes.

Synthesis cycle length

The Roche NimbleGen synthesis cycle length was computed for each probe and added to the final probe database. Roche NimbleGen cycle length was computed using their published algorithm.[21] Synthesis cycles computations and limits for other manufacturers may be easily incorporated into the software.

SNP variation

The probes were analyzed for the presence/absence of SNPs, and their positions on the probe were noted. SNPs were determined using UCSC SNPs track for hg18 (dbSNP build 130) and mm9 (dbSNP build 128). Probes with SNPs have been implicated in lower performance in array comparative genomic applications.[22]

Design of microarray-based genomic selection microarrays

MOPeD design of microarrays requires user input; the software selects optimal probes and outputs a text file that can be transmitted to the manufacturer of a microarray (Figure 2). User inputs include the following: the genome of interest; minimum and maximum values for probe size, coverage, and T_m; number of chip features; upper bounds on the number of synthesis cycles; and number of SNPs on a probe. Also specifiable is the priority of probe filtering by various parameters. Optionally, a BED file containing regions that should be biased for additional probe coverage may be specified. Finally, a BED file containing target regions from the genome of interest is required.

The format of the BED file submitted by the user is then verified. Duplicate regions are removed, while overlapping regions are merged. Preliminary probe allotment is then computed for all regions taking into account the user-specified parameters, as well as the characteristics of the genomic region under consideration such as size and GC content.

Dynamic allocation of probes

Previous studies have shown that the performance of oligonucleotide probes can vary as a function of sequence content and context.[22,23] To improve performance, MOPeD uses genomic variation information to aid in the selection and dynamic allocation of probes to targeted genomic regions. Two variables, targeted fragment size and GC content, can alter the performance of a genomic selection array. To achieve more uniform sequence capture, we employ a set of linear models to guide the dynamic allocation of probes (Figure 3). Fragments with high GC content (GC_{max}) have maximum

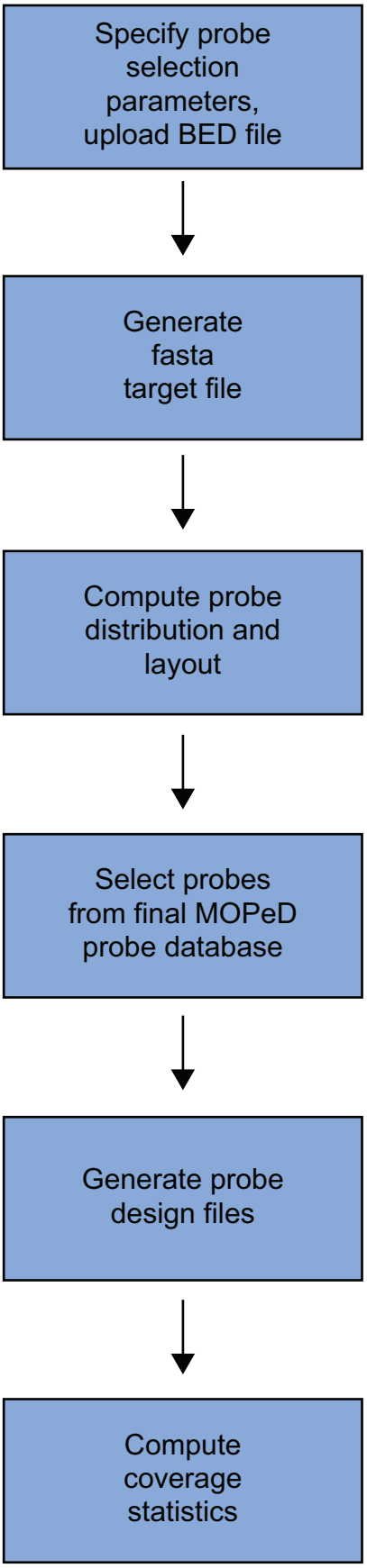

Figure 2 Steps required to generate a probe design file with MOPeD.

coverage (C_{max}) and correspondingly smaller shift (S_{min}). Similarly, large fragments (L_{max}) have smaller coverage (C_{min}) and correspondingly larger shift (S_{max}). Our protocol attempts to ensure that every base in the region of interest has at least the minimum coverage (C_{min}) of probes.

Selection of probes

The final step of the protocol involves selecting probes for each fragment in the region of interest based on parameters such as probe length, T_m, uniqueness score, hairpin potential, Roche NimbleGen cycle length, and SNPs. The first part involves selection of the best probes to ensure maximum coverage of the target region according to the algorithm outlined below. For each fragment:

1. Query probe database for all probes that tile over the fragment
2. Evaluate probes to meet user-specified parameters for hairpin, length, T_m, synthesis cycle length, and SNPs; create viable probe set (VPS)
3. If (VPS is not empty)
    a. Set V(X) = 0 for every base X in the fragment
    b. Loop until V(X) ≠ 0 for every base X in the fragment
        i. Set B(X,s) to first base X where V(X) = = 0
        ii. Set B(X,e) to last base X where V(X) = = 0 and V(B(X,s) .. B(X,e)) = = 0
        iii. Set M = (B(X,s) + B(X,e))/2
        iv. Query VPS for all probes that tile over M; create probe set MPS for position M; each probe P_i has uniqueness score U_i
        v. If (MPS is empty)
            1. Mark V(M) = 'N'
        Else
            2. Select probe P_i with lowest uniqueness score U_i
            3. Set Ps = start coordinate of P_i
            4. Set Pe = stop coordinate of P_i
            5. Mark V(Ps .. Pe) = 1
        End If
    End loop
End If

The final part involves replication of the tiled probes to satisfy the fragment coverage allotment computed beforehand.

MOPeD design files and coverage statistics

The output consists of a text file in FASTA format with all of the unique probes selected for the regions specified in the

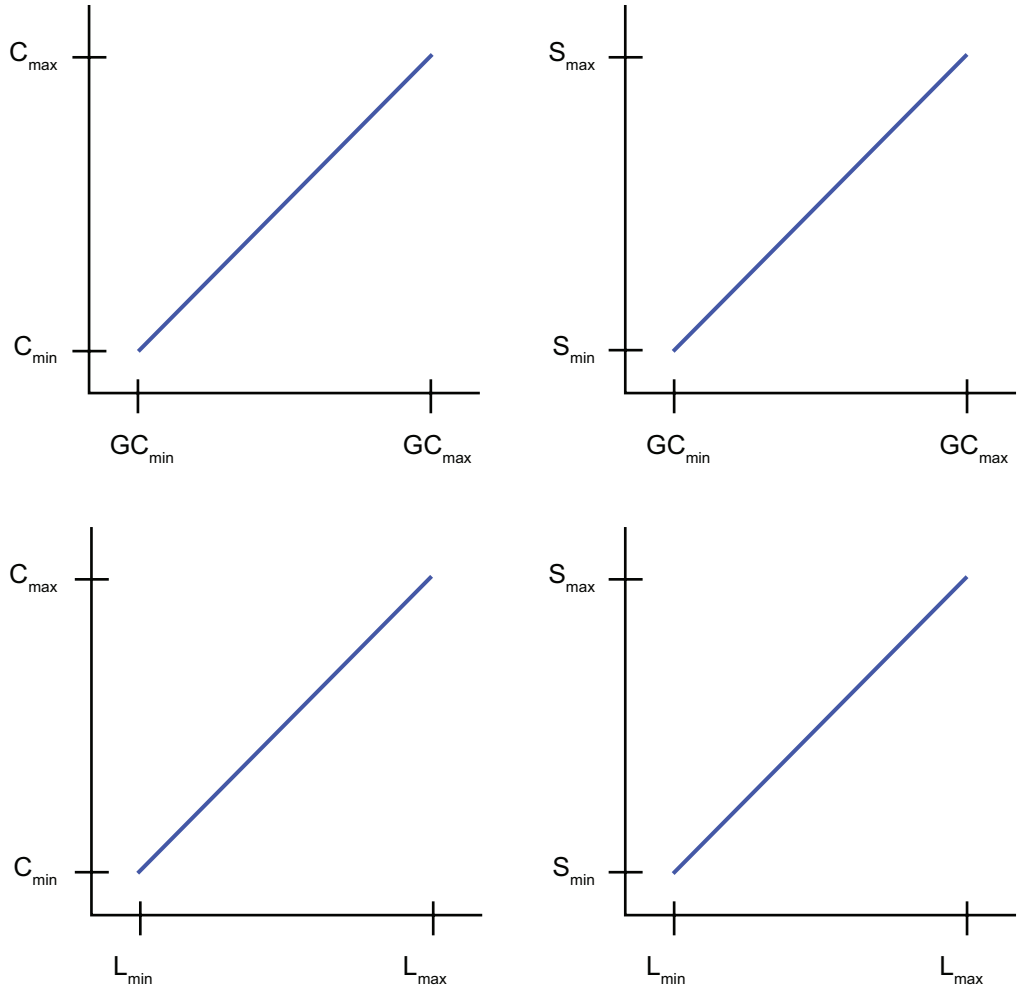

Figure 3 Linear models MOPeD uses to select and dynamically allocate probes when generating a probe design file.

user-supplied BED file. A text file with the complete probe list (385,000 or 2.1 million oligonucleotides) in a format suitable for providing to Roche NimbleGen is generated. Also provided are a summary of the design statistics and the distribution of the selected probes across the user-defined criteria, along with probe coverage analysis for individual fragments in the targeted region. For each fragment, a BED file that can be uploaded to the UCSC Genome Browser is supplied. These files show the overlay of unique probes in the target region.

MOPeD design parameters

Four different microarray-based genomic selection (MGS) arrays were designed and experimentally validated. For the human X chromosome exome microarray, the target region was preprocessed to remove fragments smaller than 25 bases and repeat regions greater than 25 bases. The MOPeD design was generated using the following criteria: probe size ranged from 55 to 65; the probe T_m range was 65°C–75°C; number of SNPs per probe was limited to 2; and the synthesis cycles limit was 192. The selected probes were further filtered to remove probes with more than 33% repeat content.

The mouse chromosome 16 487-kb microarray and the chromosome 16 and 12 exome microarray designs were generated using MOPeD with the following parameters: probe size ranged from 55 to 65; number of SNPs per probe was limited to 2; and the synthesis cycle limit was 192.

Validation of MOPeD using MGS

Experiments were carried out as outlined by Okou et al[4,9] with the following changes to the MGS protocol. Instead of 20–25 µg of fragmented DNA, 5 µg of fragmented DNA was used while repairing the ends of the DNA library. After purification of the adaptor-ligated product, the samples were run on Invitrogen 2% SizeSelec™ gels (catalog # G6610-02; Invitrogen, Carlsbad, CA). A 300-bp band was selected and placed in a plastic tube. The entire 300-bp size-selected DNA was then amplified using the following primers: 5′-AAT

GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T-3′ and 5′-CAA GCA GAA GAC GGC ATA CGA GCT CTT CCG ATC T-3′ and high-fidelity polymerase. This precapture PCR product was purified, and 1 μL of the purified product was run on a Bioanalyzer DNA 7500 chip (Agilent, Santa Clara, CA) for DNA quantitation and also for ensuring that most of the DNA fragments fell between 250 and 350 bp. To 1 μg of the precaptured PCR-purified sample, a 100-fold amount (in μg) of Human Cot-1 DNA® (Invitrogen) was added. The samples were dried down to a pellet in a Speed-Vac at medium heat (75°C). To each pellet, 2.8 μL of water and 1 μL each of two hybridization-enhancing oligos (5′-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T-3′ and 5′-CAA GCA GAA GAC GGC ATA CGA GCT CTT CCG ATC T-3′) were added. To this, we added 8 μL of 2X SC hybridization buffer (Roche NimbleGen) and 3.2 μL of SC Hybridization Component A (Roche NimbleGen). The sample pellet was then gently resuspended, and hybridization on a 385K chip was done following Roche NimbleGen's SeqCap User's Guide version 3.2. After hybridization, arrays were eluted following the protocol mentioned in Roche NimbleGen's SeqCap User's Guide version 3.2. Each eluted sample was split into 10 tubes, and postcapture PCR was done using the following primers: 5′-AAT GAT ACG GCG ACC ACC GAG A-3′ and 5′-CAA GCA GAA GAC GGC ATA CGA G-3′.

After PCR, the products were pooled from 10 tubes and were purified using the Qiagen QIAquick PCR Purification Kit (Qiagen, Germantown, MD). We then analyzed 1 μL of the purified products on a Bioanalyzer DNA 7500 chip (Agilent). More than 1 μg of DNA was obtained for each sample. PCR products were then subjected to quantitative PCR using a KAPA Library Quant Kit (catalog # 4852; Kapa Biosystems, Cambridge, MA). Based on the qPCR quantification, each sample was diluted to 10 nM using water. The samples were then denatured using NaOH, and 120 μL of 8 pM of each sample was loaded onto each lane of the flow-cell on the Illumina Cluster Station (Illumina, San Diego, CA). Following cluster amplification, the flow-cell was transferred to the Illumina Genome Analyzer (Illumina). A 76-cycle stepwise sequencing-by-synthesis using four-color nucleotides was performed according to the manufacturer's instructions (Illumina).

DNA samples analyzed

Human DNA samples used included a HapMap sample, NA18503 obtained from the Coriell Cell Repositories (Coriell, Camden, NJ). Whole genomic DNA was isolated from blood samples obtained from two additional male anonymous samples, M1 and M2, to be used in the X chromosome exome experiments. Consent was obtained and the study was approved by the University of Michigan Institutional Review Board. Mouse DNA was isolated from the liver of heterozygous carrier females. Approximately pea-sized fragments of liver were homogenized in 10 mL of DNA extraction buffer (10 mM tris pH 8.0, 0.1 M EDTA pH 8.0, 0.5% SDS, 20 μg/mL RNAse A). After homogenization, the samples were incubated at 37°C for 1 h to degrade RNA. Proteinase K was added at a concentration of 100 μg/mL, and samples were incubated at 50°C overnight. DNA was extracted three times using an equal volume of phenol equilibrated with 0.5 M tris pH 8.0. After the final extraction, an equal volume of chloroform was added to remove traces of phenol. DNA was precipitated with 0.2 volumes of 3 M sodium acetate and 2 volumes of ethanol. After precipitation, DNA was washed once with 70% ethanol and dissolved in 100–200 μL of water. The MGS protocol was then carried out as described previously.

Results

We performed three distinct targeted sequence capture experiments to validate MOPeD. Sequences targeted for genomic selection and sequencing were derived from the human and mouse genomes. These experiments exemplify potential applications of MOPeD and next-generation sequencing.

Targeted sequencing of the human X chromosome exome

We first used MOPeD to design a MGS array capable of capturing the human X chromosome exome. Targeted sequences included all coding and noncoding (3′ and 5′ untranslated regions) exons. The total reference sequence, consisting of 7429 fragments with a total size of 2,477,787 bases, was used to design capture microarrays using MOPeD and Roche NimbleGen's proprietary algorithm (Figure 4). MOPeD successfully selected oligonucleotide probes for 95.1% (7061) of the targeted fragments. As a comparison, the Roche NimbleGen design selected probes for 6% (436) fewer fragments, or 89.1% (6625) of the targeted fragments.

Comparing the coverage of the two methods revealed that there was a significant number of exons where only one algorithm successfully chose target probes (Figure 4). A total of 301 exons were covered only in the Roche NimbleGen design. These were not found in the MOPeD design because of high T_m (239), sequence repeats (33), and small fragment size (29). Relaxing the T_m parameter or increasing the size of the region

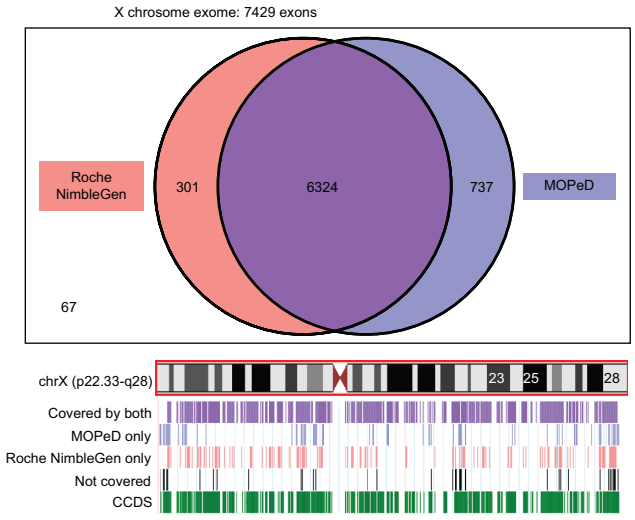

Figure 4 Comparison of MOPeD and Roche NimbleGen microarray designs for the human X chromosome exome.

searched would likely allow MOPeD to successfully design probes for these exons. The 737 exons covered only in the MOPeD design were clustered in regions of the genome (telomere, pericentromeric) expected to contain higher levels of repetitive sequences. These data suggest that the MOPeD algorithm is better at finding unique probes in regions composed of repetitive sequences.

The two designs were then empirically evaluated using the identical experimental protocol. The MOPeD-designed microarray mapped ~12% more reads uniquely to the reference target sequence and 13% fewer reads outside of the target region (Table 1). The MOPeD-designed microarray also had fewer exons with zero coverage (Supplemental Figure 1). The Roche NimbleGen-designed microarray had slightly fewer (1.5%) reads that failed to map uniquely to a single location in the target sequence. To assess data accuracy, genotype calls at 1679 known X chromosome HapMap sites in sample NA18503 showed comparable rates of data completion (97%) and accuracy (98.8%) for both designs. To assess repeatability, we performed MGS with the MOPeD-designed microarray two additional times with non-HapMap samples and obtained comparable results (Supplemental Table 1 and Supplemental Figure 2). Thus, performance differences between the MOPeD- and Roche NimbleGen-designed microarrays are repeatable.

Design and validation of a mouse chromosome 16 microarray

To assess whether MOPeD can speed the identification of mutations in the mouse, we first asked whether it could design microarrays that could be combined with next-generation sequencing to identify single base pair changes, such as those induced by the alkylating chemical, *N*-ethyl-*N*-nitrosourea (ENU). We focused on the ENU-induced mouse mutant *Hnn,* which was identified in a forward genetic screen as a recessive mutation that disrupts normal embryogenesis.[24] The *Hnn* mutation was induced on a C57/BL6 background and mapped using a C3H/HeJ backcross to mouse chromosome 16. The region targeted for genomic selection, and sequencing consisted of unique coding and noncoding DNA contained within 729 fragments with a total size of 487,615 bases. The MOPeD design successfully selected oligonucleotide probes for all 729 fragments. MGS and next-generation sequencing were performed on a DNA sample from a mouse heterozygous for the known mutation (Table 2). Only two fragments out of 729 had a median depth of zero after mapping (Supplemental Figure 3). After mapping the reads, the causative mutation (a T-to-G mutation) in a splice donor site at position 62830567 (mm9 assembly) was successfully identified as a heterozygote with a total coverage of 480.

Design and validation of a mouse chromosome 16 exome microarray

Mapping a newly induced mutation to a specific chromosome in the mouse can be accomplished inexpensively and rapidly with any number of SNP genotyping arrays. The major bottleneck and cost arise from the need to reduce the size of the region containing the mutation to make it amenable to sequencing. An alternative strategy would be to simply sequence the entire exome of a mouse chromosome suspected to harbor a mutation that results in a visible phenotype when homozygous. To evaluate how MOPeD could make this strategy feasible, we designed a genomic selection microarray targeting the chromosome 16 exome. The targeted

Table 1 Results of targeted sequencing of human X chromosome exome

Sample ID	NA18503	NA18503
Design algorithm	Roche NimbleGen	MOPeD
Size of target reference sequence (bp)	2,477,787	2,477,787
Total number of reads	6,072,205	11,006,867
Median depth (bp)	107	184
Proportion of reads map to target	0.436	0.551
Proportion of reads that fail to map uniquely to target	0.002	0.019
Proportion of reads mapping outside target region	0.562	0.431

sequence consisted of 4280 unique fragments with a total size of 1,712,120 base pairs. MOPeD was able to successfully design oligonucleotide probes for all chromosome 16 fragments. We then performed genomic selection and targeted sequencing using DNA from a mouse heterozygous for the *Hnn* mutation (Table 2). Again, we successfully identified the *Hnn* mutation (a T-to-G mutation at position 62830567, mm9 assembly) as a heterozygote with a total coverage of 498. None of the chromosome 16 fragments had a median depth of zero (Supplemental Figure 3).

Design and validation of a mouse chromosome 12 exome microarray

To further validate MOPeD, we designed a mouse chromosome 12 exome microarray to identify an induced mutation. The targeted sequence consisted of 6200 unique fragments with a total size of 3,345,769 base pairs. The MOPeD design successfully selected oligonucleotide probes for all fragments. Genomic selection and Illumina sequencing were then performed (Table 2), and a putative mutation was identified as a heterozygote with a total sequence depth of 108. Subsequent Sanger sequencing confirmed the variant, and complementation testing demonstrated that it was in fact the causative mutation (data not shown). For the chromosome 12 exome microarray, only 45 out of 6200 fragments had a median sequence depth of zero (Supplemental Figure 3).

Discussion

Methods of direct genomic selection, especially when combined with next-generation sequencing platforms, offer a number of significant advantages over traditional PCR-based methods of target DNA preparation.[3–10,25,26] Our software package, MOPeD, enables individual investigators to use a fully open source set of software tools to optimize the design of high-density oligonucleotide microarrays for genomic selection. When integrated with maskless synthesis commercially available from Roche NimbleGen, MOPeD can be especially useful for experiments requiring custom designs, or for those instances when only a limited number of samples need to be characterized.

MOPeD offers a number of advantages over the standard Roche NimbleGen design algorithm. First, MOPeD-designed arrays are able to capture a larger proportion of a targeted reference sequence, and at the same time, have more reads map to the targeted sequence than the equivalent Roche NimbleGen microarray. Second, because the MOPeD software is fully open source and freely available to the scientific community, the methods used are thoroughly described and are available to be improved upon by the larger scientific community. Furthermore, synthesis cycle computations and limits for other manufacturers could be easily incorporated into the software. Third, MOPeD allows the user to know the complete sequence of all oligonucleotide probes. This information is not made available to users of Roche NimbleGen-designed microarrays. Finally, the approach we employ is general, thereby enabling analysis of genomes beyond the human and the mouse. Presently, the MOPeD Web site (http://moped.genetics.emory.edu/) also includes a rhesus macaque probe database, and we intend to support additional reference genomes in the future.

We believe there are a number of potential future directions MOPeD could help pursue. The current implementation uses a dynamic probe allocation scheme that uses linear models to guide probe selection. The software and performance of the genomic selection microarray might be further improved with the development of nonlinear models to help guide probe distribution. Recently, methods of genomic selection that use oligonucleotides in solution are becoming more prevalent and offer some advantages. Regardless of the specific experimental protocol used, the fundamental technical challenge lies in designing oligonucleotides that can

Table 2 Results of targeted sequencing of mouse chromosomes 16 and 12

Sample ID	Mouse chromosome 16 region	Mouse chromosome 16 exome	Mouse chromosome 12 exome
Design algorithm	MOPeD	MOPeD	MOPeD
Size of target reference sequence (bp)	487,615	1,712,120	3,345,769
Total number of reads	11,219,282	15,444,662	12,933,835
Median depth (bp)	331	435	119
Proportion of reads map to target	0.444	0.723	0.577
Proportion of reads that fail to map uniquely to target	0.035	0.011	0.050
Proportion of reads mapping outside target region	0.521	0.266	0.373

uniquely and successfully bind targets from a given genome, and MOPeD offers a fully open method that can be used to address this important technical challenge.

Conclusion

Here, we describe an open source software package named MOPeD that efficiently designs high-density oligonucleotide genomic selection microarrays. At present, individual investigators can access the MOPeD Web site and design oligonucleotide microarrays for the human, mouse, and rhesus macaque genomes (http://moped.genetics.emory.edu/). Experimental validation of four different MOPeD-designed microarrays shows improved performance on a number of standard metrics when compared with the proprietary Roche NimbleGen design algorithm.

Acknowledgments

The ELLIPSE Emory High-Performance Computing Cluster was used for this project. This work was supported in part by the National Institutes of Health/National Institute of Mental Health Gift Fund (grant number MH076439) to MEZ, the Simons Foundation Autism Research Initiative (MEZ), and the PHD Grant (UL1 RR025008, KL2 RR025009, or TL1 RR025010) from the Clinical and Translational Science Award Program, National Institutes of Health, National Center for Research Resources.

Disclosure

The authors report no financial interest or conflicts of interest in this work.

References

1. Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: methods and goals. *Nat Rev Genet*. 2004;5(5):335–344.
2. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008;26(10):1135–1145.
3. Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett M. Direct genomic selection. *Nat Methods*. 2005;2(1):63–69.
4. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. Microarray-based genomic selection for high-throughput resequencing. *Nat Methods*. 2007;4(11):907–909.
5. Porreca GJ, Zhang K, Li JB, et al. Multiplex amplification of large sets of human exons. *Nat Methods*. 2007;4(11):931–936.
6. Albert TJ, Molla MN, Muzny DM, et al. Direct selection of human genomic loci by microarray hybridization. *Nat Methods*. 2007;4(11):903–905.
7. Hodges E, Xuan Z, Balija V, et al. Genome-wide in situ exon capture for selective resequencing. *Nat Genet*. 2007;39(12):1522–1527.
8. Krishnakumar S, Zheng J, Wilhelmy J, Faham M, Mindrinos M, Davis R. A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proc Natl Acad Sci U S A*. 2008;105(27):9296–9301.
9. Okou DT, Locke AE, Steinberg KM, et al. Combining microarray-based genomic selection (MGS) with the Illumina Genome Analyzer platform to sequence diploid target regions. *Ann Hum Genet*. 2009;73(Pt 5):502–513.
10. Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol*. 2009;27(2):182–189.
11. Mamanova L, Coffey AJ, Scott CE, et al. Target-enrichment strategies for next-generation sequencing. *Nat Methods*. 2010;7(2):111–118.
12. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*. 2009;461(7261):272–276.
13. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*. 2010;42(1):30–35.
14. Singh-Gasson S, Green RD, Yue Y, et al. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat Biotechnol*. 1999;17(10):974–978.
15. Graf S, Nielsen FG, Kurtz S, et al. Optimized design and assessment of whole genome tiling arrays. *Bioinformatics*. 2007;23(13):i195–i204.
16. Lipson D, Yakhini Z, Aumann Y. Optimization of probe coverage for high-resolution oligonucleotide aCGH. *Bioinformatics*. 2007;23(2):e77–e83.
17. Schliep A, Krause R. Efficient algorithms for the computational design of optimal tiling arrays. *IEEE/ACM Trans Comput Biol Bioinform*. 2008;5(4):557–567.
18. Hovik H, Chen T. Dynamic probe selection for studying microbial transcriptome with high-density genomic tiling microarrays. *BMC Bioinformatics*. 2010;11:82.
19. Lemoine S, Combes F, Le Crom S. An evaluation of custom microarray applications: the oligonucleotide design challenge. *Nucleic Acids Res*. 2009;37(6):1726–1739.
20. Vainrub A, Pettitt BM. Theoretical aspects of genomic variation screening using DNA microarrays. *Biopolymers*. 2004;73(5):614–620.
21. Roche Nimblegen Systems I. Technical note: Roche Nimblegen probe design fundamentals. Part No. TN-ARAY0100. 2007.
22. Mulle JG, Patel VC, Warren ST, Hegde MR, Cutler DJ, Zwick ME. Empirical evaluation of oligonucleotide probe selection for DNA microarrays. *PLoS One*. 2010;5(3):e9921.
23. Cutler DJ, Zwick ME, Carrasquillo MM, et al. High-throughput variation detection and genotyping using microarrays. *Genome Res*. 2001;11(11):1913–1925.
24. Caspary T, Larkins CE, Anderson KV. The graded response to Sonic Hedgehog depends on cilia architecture. *Dev Cell*. 2007;12(5):767–778.
25. Dahl F, Stenberg J, Fredriksson S, et al. Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc Natl Acad Sci U S A*. 2007;104(22):9387–9392.
26. Bau S, Schracke N, Kränzle M, et al. Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays. *Anal Bioanal Chem*. 2009;393(1):171–175.

Supplementary material

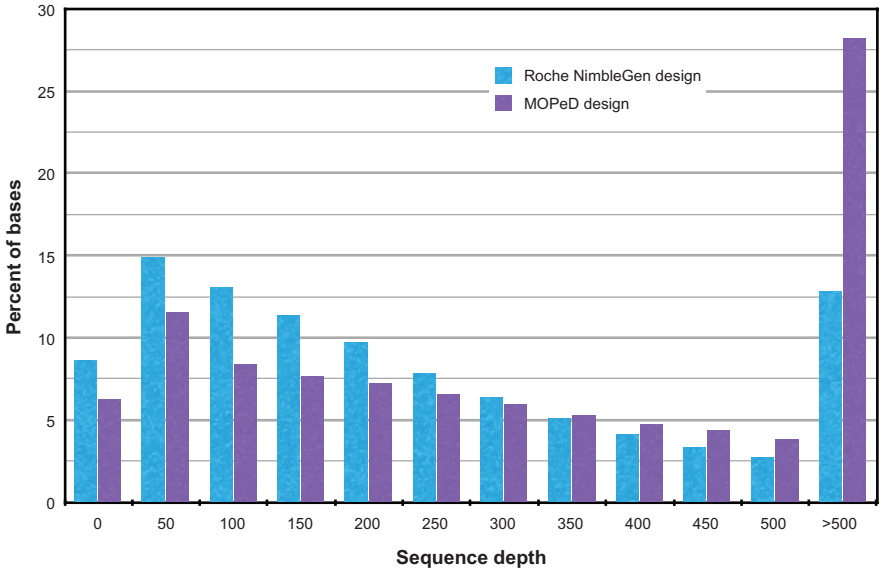

Figure S1 Sequence depth obtained for two different human X chromosome exome microarrays. Data shown compare the performance of a MOPeD-designed MGS microarray and a Roche NimbleGen-designed microarray.

Table S1 Results of targeted sequencing of human X chromosome exomes in samples M1 and M2

Sample ID	M1	M2
Design algorithm	MOPeD	MOPeD
Median depth	324	202
Total number of reads	14,024,708	9,461,956
Proportion of reads map to target	0.553	0.474
Proportion of reads that fail to map uniquely to target	0.017	0.025
Proportion of reads mapping outside target region	0.430	0.501

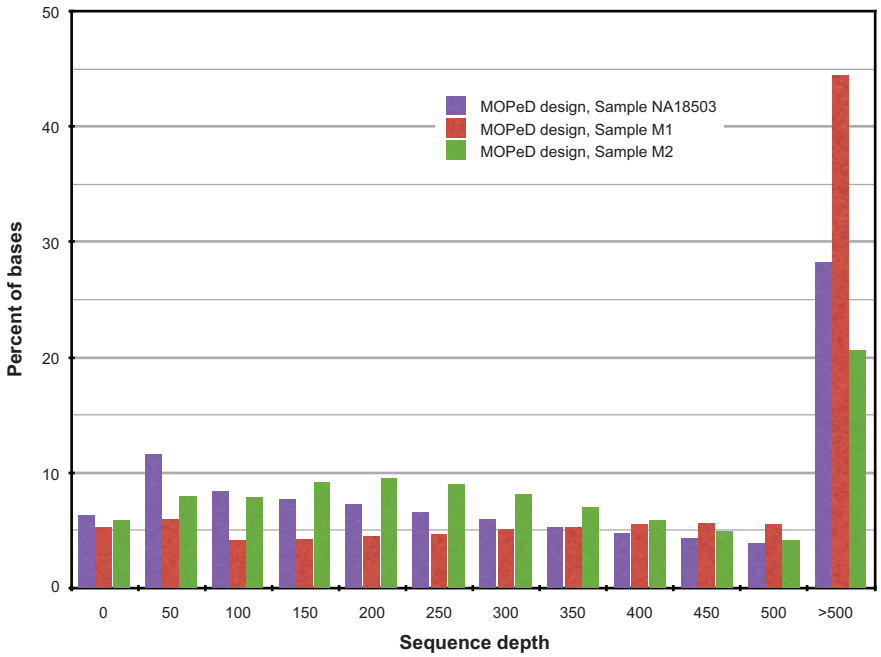

Figure S2 Sequence depth obtained in three different human X chromosome exome sequencing experiments. Each experiment used a single MOPeD-designed MGS microarray. The identical design was used for each sample.

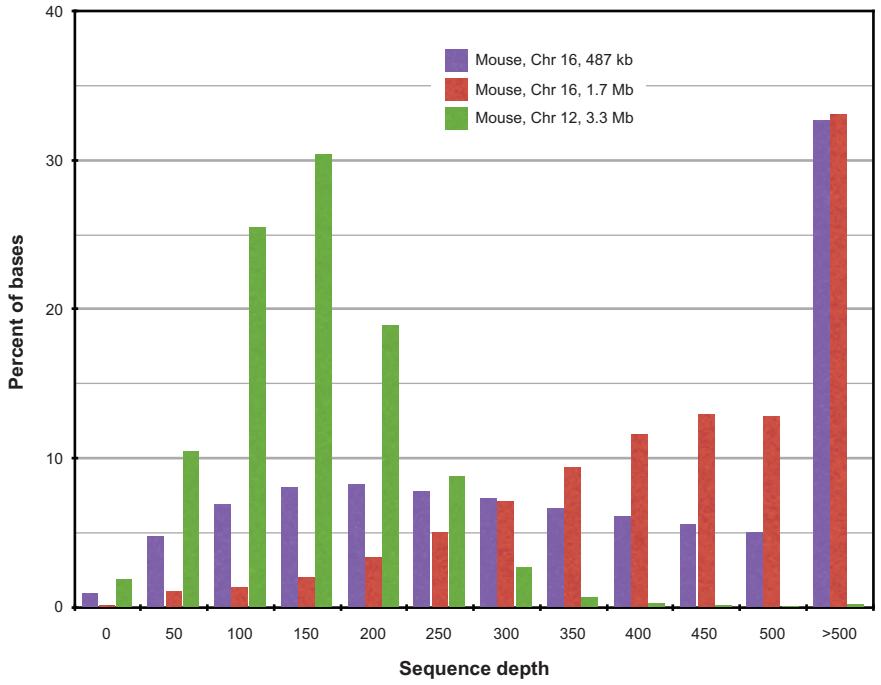

Figure S3 Sequence depth obtained in the three different mouse MGS and sequencing experiments. Each experiment used a different MOPeD-designed MGS microarray.