AEM
Journals.ASM.org

Strain/Species-Specific Probe Design for Microbial Identification Microarrays

Qichao Tu,[a] Zhili He,[a] Ye Deng,[a] Jizhong Zhou[a,b,c]

Institute for Environmental Genomics, University of Oklahoma, Norman, Oklahoma, USA[a]; Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA[b]; State Key Joint Laboratory of Environmental Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, China[c]

Specific identification of microorganisms in the environment is important but challenging, especially at the species/strain level. Here, we have developed a novel *k*-mer-based approach to select strain/species-specific probes for microbial identification with diagnostic microarrays. Application of this approach to human microbiome genomes showed that multiple (≥10 probes per strain) strain-specific 50-mer oligonucleotide probes could be designed for 2,012 of 3,421 bacterial strains of the human microbiome, and species-specific probes could be designed for most of the other strains. The method can also be used to select strain/species-specific probes for sequenced genomes in any environments, such as soil and water.

Understanding the diversity, composition, structure, activity, interaction, and dynamics of microorganisms, especially pathogens in the environment, is crucial to reveal host/environment-microbe interactions. However, accurate identification of microorganisms in a complex environment is challenging, especially at the strain/species level, due to the limited number of reference genomes and low resolutions of currently used methods. For example, the most commonly used 16S rRNA sequencing can identify only microorganisms at the genus level but not at the species/strain levels. Recent advances in sequencing technology have greatly facilitated genome sequencing for microorganisms at low cost, generating thousands of draft or finished microbial genomes so far. For example, the human microbiome project (1) has released more than 828 draft or finished reference genomes from different human body sites, with a total of 2,467 funded for sequencing. Although huge sequence data sets are accumulating, transforming them into useful information and knowledge remains challenging. Microarrays are one of the technologies that can be used for highly parallel detection of complex microbial communities in many environments (2, 3). So far, a variety of microarrays, such as GeoChip and PhyloChip, have been developed and widely used for functional and phylogenetic profiling of microbial communities from different habitats (4–6). However, those types of microarrays largely target conserved 16S rRNA or key functional genes, which are not suitable for identifying microbial species/strains. This is especially true for the human microbiome, which is characterized by low diversity at the class level or above but extremely high diversity at the genus, species, and/or strain levels (7). Therefore, it is necessary to develop strain/species-level identification microarrays for microbial community studies, which is largely challenged by the selection of strain/species-specific probes from huge sequence databases. However, it is almost impossible to design strain/species-specific probes using currently available software or approaches (8–11), especially with an exponential increase of genome sequences in the future. Using the human microbiome as an example, here we have developed a *k*-mer-based approach to quickly and comprehensively select 50-mer strain/species-specific probes for sequenced microbial strains/species, which can be used to construct microarrays for strain/species-level identification of microorganisms in complex microbial communities, potentially leading to rapid and accurate clinical diagnosis of environmental problems.

The major challenge in strain/species-specific probe design is to ensure probe specificity. We use three criteria for selecting candidate-specific probes, including maximum continuous stretch length, sequence similarity, and free energy of designed probes with their nontarget sequences (12, 13). Considering the above-described issues, a flowchart of strain/species-specific probe design (SSPD) was constructed for selecting optimal strain/species-specific probes (Fig. 1), including five major steps, as follows.

(i) Data preparation. We downloaded both GenBank format contigs and assembled draft/finished genome sequences for 5,417 bacterial species released by NCBI GenBank and HMP DACC databases (www.hmpdacc.org). To avoid cross-hybridization with contaminated human DNA in the sample, human genome sequences were also downloaded and included for specificity checking.

(ii) Generation of 50-mer DNA fragments for each genome. The first challenge was to identify strain-specific regions in each reference genome; however, to our knowledge, no comparative genomic tools or probe design software is able to perform such tasks. Even for CommOligo (10), which was specifically developed to design probes for highly homologous genes, genomes, and metagenomes, this is still a difficult problem. To solve this issue, we took advantages of *k*-mer-based approaches in analysis of next-generation sequencing data, which is used mainly by *de novo* assemblers like Velvet, which assembles a deluge of short reads into contigs for microbial communities or individual genomes (14). Here, we first split each reference genome into 50-mer fragments without ambiguity nucleotide (Ns and other consensus nucleotides). Thus, for a genome size of L, the number of 50-mer fragments is up to $L - 50$.

Received 7 April 2013 Accepted 2 June 2013
Published ahead of print 7 June 2013
Address correspondence to Jizhong Zhou, jzhou@ou.edu, or Zhili He, zhili.he@ou.edu.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/AEM.01124-13.

doi:10.1128/AEM.01124-13

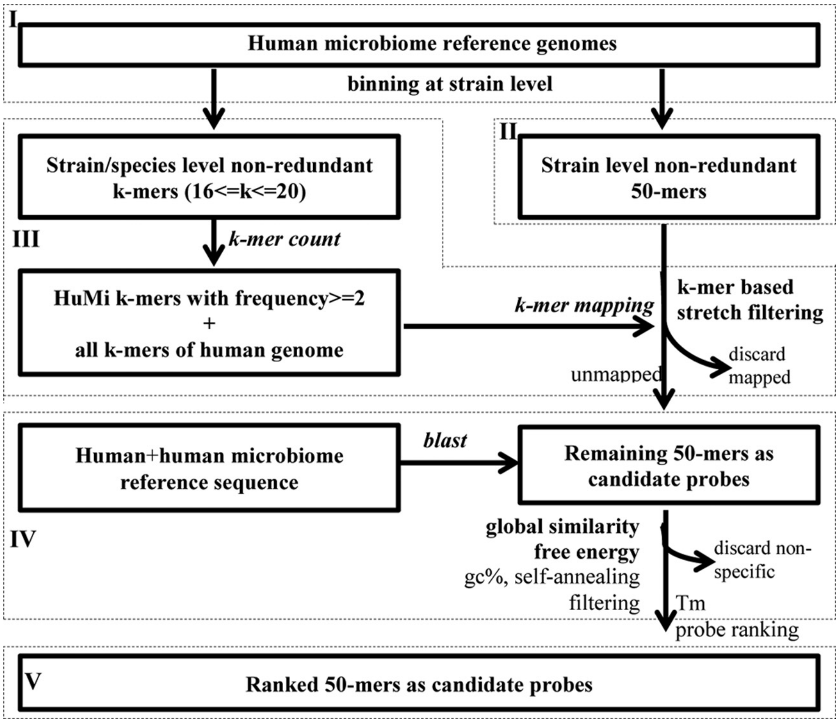
I
Human microbiome reference genomes
binning at strain level
Strain/species level non-redundant k-mers (16<=k<=20)
II
Strain level non-redundant 50-mers
III
k-mer count
HuMi k-mers with frequency>=2 + all k-mers of human genome
k-mer mapping
k-mer based stretch filtering
discard mapped
unmapped
Human+human microbiome reference sequence
blast
Remaining 50-mers as candidate probes
IV
global similarity
free energy
gc%, self-annealing
filtering
discard non-specific
Tm
probe ranking
V
Ranked 50-mers as candidate probes

FIG 1 Flowchart of SSPD, including five major steps: (i) data preparation; (ii) probe-mer generation; (iii) *k*-mer-based continuous stretch filtering; (iv) further specificity check; and (v) probe ranking and output.

(iii) *k*-mer-based continuous stretch filtering. An appropriate *k*-mer size for continuous filtering is important to warrant the probe specificity. Based on our experimentally validated probe design criteria (12, 13), a 50-mer probe with ≤20-base continuous stretches with nontarget sequences should not have significant cross-hybridization, so *k* values of ≥16 and ≤20 were tested in this study. Also, probes with shorter stretch lengths with nontarget sequences have less chance to cause cross-hybridization. Our experience with GeoChip 3.0 development (15) showed that the majority of nonspecific probe candidates (~90%) were discarded because they could not meet the criteria: <20-base stretches with nontarget sequences. Thus, this step was first performed to remove nonspecific 50-mers. In order to quickly filter nonspecific 50-bp probes sharing *k*-mer ($16 \leq k \leq 20$)-length stretches with other strains/species, strain/species-level nonredundant *k*-mers were generated for all human microbiomes and human genomes. *k*-mers that occurred in two or more bacterial strains/species were extracted and combined with all *k*-mers of human genomes as a database for stretch filtering. A *k*-mer table was then built by the Meryl program adopted from the *k*-mer package (16). All *k*-mers in the *k*-mer table were mapped to the 50-mers for each genome generated in the 2nd step by the mapMers program (16). Mapped 50-mers were discarded since they shared the same *k*-mers with other strains.

(iv) Further specificity check. The remaining 50-mers for each genome were then searched against all human microbiome and human genomes for further global sequence identity and free energy filtering using BLAST to search the closest nontarget sequences and recalculate global sequence identities. GC content (~20% to 80%), self-annealing (≤8 bp), and melting temperature (T_m) properties were also calculated and applied for probe filtering.

(v) Ranking. Then, 50-mer candidate probes were ranked based on continuous stretch, sequence similarity, and free energy between the probes and their nontarget sequences using a similar probe quality score function described in CommOligo (10). Finally, for each genome, probes passing all filtering criteria were output in a text-format file with detailed information, such as start/end positions in the genome, maximum global similarity, maximum continuous stretch, minimum free energy to nontarget sequences, GC content, T_m, targeted gene or intergenic region, gene annotation, and probe quality scores.

As a result, *k*-mer-based stretch filtering from 16-mer to 20-mer showed that more than 8 million 50-mer probe candidates remained for these bacterial genomes at the 18-base stretch cutoff (Fig. 2A), indicating that most currently sequenced genomes could be differentiated at the 18-mer size. Specifically, 1,325 bacterial strains could have ≥10 probes designed for each strain when the 18-base continuous stretch was used, and the number increased to 2,012 when 20-base continuous stretches were used, among which at least 1,447 strains could have more than 100 probes per strain designed (Fig. 2B; see also Table S1 in the supplemental material). Of the 2,404,311 strain-specific probe candidates designed with 18-base to 20-base continuous stretches, 1,001,855 (41.67%) were located in genes, 342,470 in the intergenic region (14.24%), 106,611 (4.43%) between a gene and an intergenic region, and the rest (953,375; 39.65%) were from unannotated genomes for which gene/intergenic information is not yet available (Fig. 2C; see also Table S1). Considering the size of genes and intergenic regions in a bacterial genome (~4.9:1), it could be concluded that a similar rate of intergenic regions contains strain-specific elements, indicating that intergenic regions in bacterial genomes could be as important as genes in designing strain-specific probes for microbial identification microarrays.

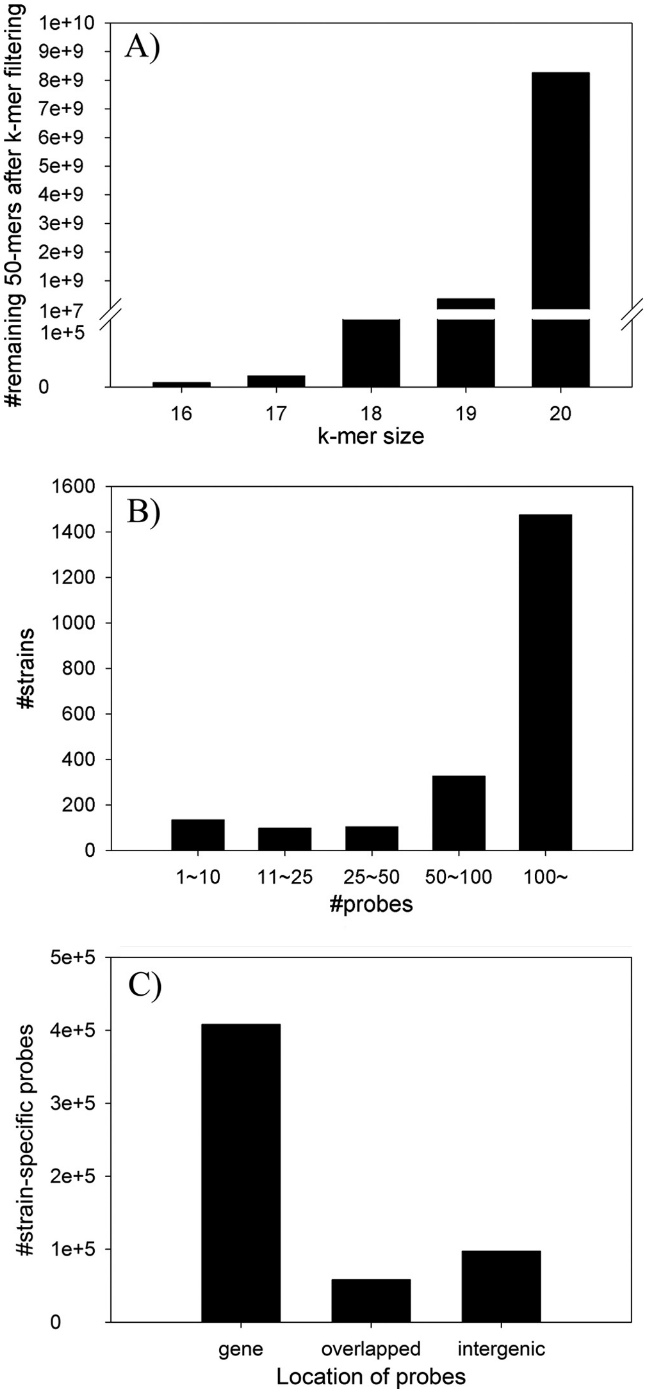

FIG 2 Summary of strain-specific probes designed for human microbiome strains. (A) Number of remaining 50-mers after *k*-mer filtering using *k*-mer sizes from 16 to 20; (B) distribution of strain numbers with a different number of probes designed per strain; (C) distribution of strain-specific probes at different regions in the genome, including gene, intergenic region, and the overlap between gene and intergenic regions. Locations of probes were summarized based on completed genomes.

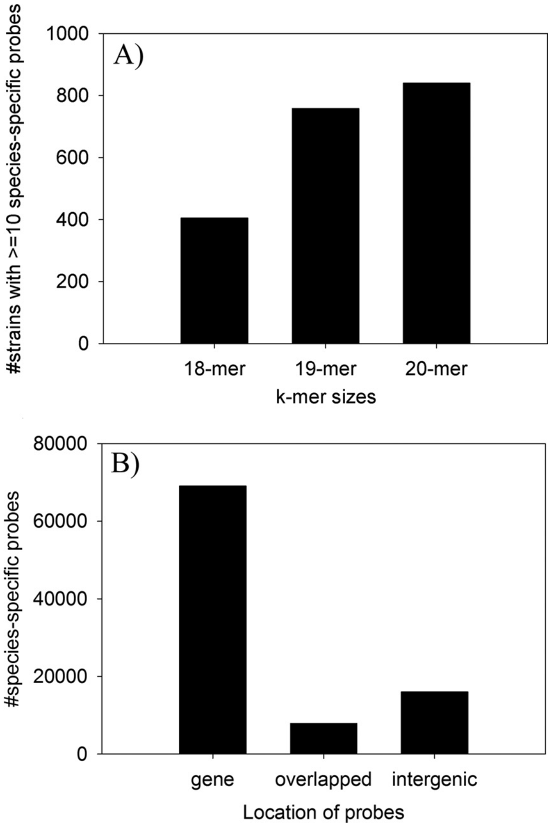

FIG 3 Summary of species-specific probes designed for human microbiome strains without strain-specific probes. (A) Number of strains with more than 10 species-specific probes per strain designed at different *k*-mer sizes; (B) distribution of species-specific probes at different regions in the genome, including gene, intergenic, and the overlap between gene and intergenic regions. Locations of probes were summarized based on completed genomes.

For most bacterial genomes, a shorter *k*-mer should be used for the continuous stretch cutoff since the number of 50-mer candidate probes will dramatically increase when a longer *k*-mer for stretch filtering is used, resulting in a much longer computational time for probe checking. For microarray fabrication, we recommend that multiple (e.g., ~10 to 50) strain-specific probes should be used for each strain to ensure reliable and accurate microbial identification as well as sufficient statistical power for data analysis.

It is noticed that no strain-specific probes could be designed for 1,208 bacterial strains. Further investigation of these strains showed such a problem was due mainly to the existence of extremely closely related strains, such as strains named *Propionibacterium acnes* HLxxxx (where xxxx is the strain number), which were very similar strains isolated from the same study (17). In order to design probes for these strains, species-specific probes that target multiple strains in the same species but not strains in other species were introduced and designed, enabling species-level identification of microorganisms without strain-specific probes. As a result, 405 and 840 of the 1,208 bacterial strains could have more than 10 species-specific probes per strain designed when an 18-mer continuous stretch and a 20-mer continuous stretch were used, respectively (Fig. 3A; see also Table S2 in the supplemental material). Again, a rate similar to the intergenic region/gene ratio in bacterial genomes was measured for the probes (38,641 versus 142,585) (Fig. 3B; see also Table S2), suggesting the specificity of intergenic regions also exists at the species level. To cover all strains/species, longer probe length and/or more relaxed continuous stretch filtering thresholds could be used for the remaining strains without strain/species-specific probes. For example, only 6 of 23 O157:H7 *Escherichia coli* strains could have ≥10 probes/strain designed when a ≤20-bp continuous stretch was used, but the number increased to 17 with a ≤21-bp continuous stretch length.

It should be noted that some potential pitfalls may remain in applying this approach to design strain/species-specific probes. Generally, three concerns are identified: (i) selection of strain-specific probes for very closely related strains in the same species, (ii) cross-hybridization with not-yet-sequenced genomes in the environment, and (iii) genome variations among individuals of the same strain/species. The first concern may result in no/few specific probes designed for a portion of sequenced strains, affecting the microarray coverage. To address this issue, two solutions are suggested. One is to slightly relax the probe design criteria, such as longer probe length, longer stretch length, and/or other parameters (e.g., sequence identity, free energy). In this case, more probes could be chosen for each genome/strain. The other is to design species-specific probes for each particular species with multiple closely related strains. In this case, the coverage should be increased, although it may lose the ability to detect those microorganisms at the strain level. The second concern may lead to nonspecific identification. Although the human microbiome has a high portion of genomes sequenced, most genomes in the environment, especially in soil, are not sequenced yet, so it is impossible to include unsequenced genome information for unique *k*-mer identification, which may result in cross-hybridization with nontarget sequences in the environment. To solve this problem, we will update our databases to the current status so that newly sequenced genomes can be included for more specific probe design. Also, we will use experimental data to statistically define a reasonable cutoff (e.g., 5% of probes) for potential positives during data analysis. It is expected that such approaches should minimize or eliminate potential cross-hybridizations. The third concern may cause misidentification of microorganisms at the species or strain level in a complex environment. The current approach is based on the available genome sequences and their taxonomical information in the database, and genome variations among individuals of the same strain or species are not specially considered. Since the information about variations within a species or strain is largely unavailable due to the limited number of genome sequences, one current possible solution is to randomly select more probes (e.g., ~50 to 100) for each genome so that the effect of some probes from highly variable regions could be reduced during data analysis, during which we may eliminate such probes using a cutoff (e.g., 5%) for false positives and/or false negatives. However, this issue may be minimized by avoiding some highly variable regions when more and more individual genomes of the same strain/species are available.

In summary, we have developed a novel approach for strain/species-specific probe design by taking advantage of *k*-mer-based strategies. Although the approach here was initially evaluated to design 50-mer strain/species-specific probes for the human microbiome, it can also be applied to design strain/species-specific probes for microbial communities in any environments with user-specified criteria as long as there is a good coverage of reference genomes. For example, the ongoing soil TerraGenome project (18) is expected to generate many reference genomes, and this developed tool could be used to design strain/species-specific probes to construct microarrays for identifying microbial strains in soil.

A Web-based portal to select strain/species-specific probes for 5,417 bacterial genomes is available at http://ieg.ou.edu/SSPD.

ACKNOWLEDGMENTS

This work was supported by Ecosystems and Networks Integrated with Genes and Molecular Assemblies (ENIGMA) under contract no. DE-AC02-05CH11231 and by the Oklahoma Center for the Advancement of Science and Technology (OCAST) through Oklahoma Applied Research Support (OARS) project AR11-035.

REFERENCES

1. **Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI.** 2007. The human microbiome project. Nature **449:**804–810.
2. **Roh SW, Abell GCJ, Kim K-H, Nam Y-D, Bae J-W.** 2010. Comparing microarrays and next-generation sequencing technologies for microbial ecology research. Trends Biotechnol. **28:**291–299.
3. **Stralis-Pavese N, Abell GCJ, Sessitsch A, Bodrossy L.** 2011. Analysis of methanotroph community composition using a *pmoA*-based microbial diagnostic microarray. Nat. Protoc. **6:**609–624.
4. **Brodie EL, DeSantis TZ, Joyner DC, Baek SM, Larsen JT, Andersen GL, Hazen TC, Richardson PM, Herman DJ, Tokunaga TK, Wan JM, Firestone MK.** 2006. Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. Appl. Environ. Microbiol. **72:**6288–6298.
5. **He Z, Deng Y, Zhou J.** 2012. Development of functional gene microarrays for microbial community analysis. Curr. Opin. Biotechnol. **23:**49–55.
6. **He Z, Van Nostrand JD, Zhou J.** 2012. Applications of functional gene microarrays for profiling microbial communities. Curr. Opin. Biotechnol. 23:460–466.
7. **Ley RE.** 2010. Obesity and the human microbiome. Curr. Opin. Gastrenterol. **26:**5–11.
8. **Dugat-Bony E, Missaoui M, Peyretaillade E, Biderre-Petit C, Bouzid O, Gouinaud C, Hill D, Peyret P.** 2011. HiSpOD: probe design for functional DNA microarrays. Bioinformatics **27:**641–648.
9. **Lemoine S, Combes F, Le Crom S.** 2009. An evaluation of custom microarray applications: the oligonucleotide design challenge. Nucleic Acids Res. **37:**1726–1739.
10. **Li X, He Z, Zhou J.** 2005. Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. Nucleic Acids Res. **33:**6114–6123.
11. **Parisot N, Denonfoux J, Dugat-Bony E, Peyret P, Peyretaillade E.** 2012. KASpOD—a Web service for highly specific and explorative oligonucleotide design. Bioinformatics **28:**3161–3162.
12. **He Z, Wu L, Li X, Fields MW, Zhou J.** 2005. Empirical establishment of oligonucleotide probe design criteria. Appl. Environ. Microbiol. **71:**3753–3760.
13. **Liebich J, Schadt CW, Chong SC, He Z, Rhee S-K, Zhou J.** 2006. Improvement of oligonucleotide probe design criteria for functional gene microarrays in environmental applications. Appl. Environ. Microbiol. **72:** 1688–1691.
14. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18:**821–829.
15. **He Z, Deng Y, Van Nostrand JD, Tu Q, Xu M, Hemme CL, Li X, Wu L, Gentry TJ, Yin Y, Liebich J, Hazen TC, Zhou J.** 2010. GeoChip 3.0 as a high-throughput tool for analyzing microbial community composition, structure and functional activity. ISME J. **4:**1167–1179.
16. **Marçais G, Kingsford C.** 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics **27:**764–770.
17. **McDowell A, Barnard E, Nagy I, Gao A, Tomida S, Li H, Eady A, Cove J, Nord CE, Patrick S.** 2012. An expanded multilocus sequence typing scheme for propionibacterium acnes: investigation of 'pathogenic,' 'commensal' and antibiotic resistant strains. PLoS One **7:**30. doi:10.1371/journal.pone.0041480.
18. **Vogel TM, Simonet P, Jansson JK, Hirsch PR, Tiedje JM, van Elsas JD, Bailey MJ, Nalin R, Philippot L.** 2009. TerraGenome: a consortium for the sequencing of a soil metagenome. Nat. Rev. Microbiol. **7:**252.