BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 15 2011, pages 2134–2140
doi:10.1093/bioinformatics/btr326

Systems biology

Advance Access publication June 7, 2011

Mathematical modeling of 16S ribosomal DNA amplification reveals optimal conditions for the interrogation of complex microbial communities with phylogenetic microarrays

Oleg Paliy[1,*] and Brent D. Foy[2]

[1]Department of Biochemistry and Molecular Biology and [2]Department of Physics, Wright State University, Dayton, OH 45435, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Many current studies of complex microbial communities rely on the isolation of community genomic DNA, amplification of 16S ribosomal RNA genes (rDNA) and subsequent examination of community structure through interrogation of the amplified 16S rDNA pool by high-throughput sequencing, phylogenetic microarrays or quantitative PCR.

Results: Here we describe the development of a mathematical model aimed to simulate multitemplate amplification of 16S ribosomal DNA sample and subsequent detection of these amplified 16S rDNA species by phylogenetic microarray. Using parameters estimated from the experimental results obtained in the analysis of intestinal microbial communities with Microbiota Array, we show that both species detection and the accuracy of species abundance estimates depended heavily on the number of PCR cycles used to amplify 16S rDNA. Both parameters initially improved with each additional PCR cycle and reached optimum between 15 and 20 cycles of amplification. The use of more than 20 cycles of PCR amplification and/or more than 50 ng of starting genomic DNA template was, however, detrimental to both the fraction of detected community members and the accuracy of abundance estimates. Overall, the outcomes of the model simulations matched well available experimental data. Our simulations also showed that species detection and the accuracy of abundance measurements correlated positively with the higher sample-wide PCR amplification rate, lower template-to-template PCR bias and lower number of species in the interrogated community. The developed model can be easily modified to simulate other multitemplate DNA mixtures as well as other microarray designs and PCR amplification protocols.

Contact: oleg.paliy@wright.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on March 2, 2011; revised on May 2, 2011; accepted on May 22, 2011

1 INTRODUCTION

Owing to the development and refinement of novel DNA and RNA interrogation technologies, there is a surge of studies in the current literature exploring the populational structure and function of various complex microbial communities (Brodie *et al.*, 2007; Gao *et al.*, 2007; Huber *et al.*, 2007). The ability to isolate and subsequently examine total community DNA and RNA without any need to culture individual microbial species and cells allows analysis of systems that would otherwise be difficult to profile and examine including microbiota of human intestine and other epithelial surfaces and the microbes of soils and ocean waters (Eckburg *et al.*, 2005; Gao *et al.*, 2007; Huber *et al.*, 2007; Kent and Triplett, 2002).

Gene coding for the small ribosomal subunit RNA molecule (16S rRNA in prokaryotes and 18S rRNA in eukaryotes) has been used in the vast majority of such studies due to its ubiquitous presence in all organisms and because of the conservation of its nucleotide sequence (Cannone *et al.*, 2002). In a typical experimental design to profile microbial community structure, total genomic DNA (gDNA) isolated from a sample of interest is subjected to rounds of 16S rRNA gene (rDNA) specific amplification in polymerase chain reaction (PCR) using two 'universal' primers complementary to the beginning and the end of prokaryotic 16S rRNA molecule (Frank *et al.*, 2008). The amplified DNA is then interrogated by a detection method of choice such as DNA sequencing or microarray analysis. Because on average 16S rRNA gene constitutes only $\sim$0.25% of the total genomic DNA (see below), selective 16S rDNA amplification is crucial to increase the sensitivity of detection (Paliy *et al.*, 2009) and to obtain good measures of bacterial presence and relative abundance in the community samples. However, the optimal use of such PCR amplification in relation to microarray and DNA sequencing detection have not been yet fully explored.

A number of studies have been published, though, examining the thermodynamic behavior of DNA molecules during DNA amplification, and the biases that can be observed during many rounds of PCR amplification (Kanagawa, 2003; Kurata *et al.*, 2004; Polz and Cavanaugh, 1998). Because most microbial communities consist of a large number of different microbial species with varied 16S rRNA gene sequences, any PCR amplification of community DNA is multitemplate. PCR amplification of such gDNA has been shown to introduce a deviation of the post-amplification fractions from the initial ratios of DNA molecules (termed PCR bias) due to unequal amplification of different DNA molecules during PCR (Polz and Cavanaugh, 1998). Several mechanisms of this effect have been described that include (i) unequal denaturation of templates based on GC content of DNA sequences; (ii) higher binding efficiency of GC-rich variants of degenerate amplification primers to the template at the same annealing temperature; and (iii) competitive re-annealing

*To whom correspondence should be addressed.

of abundant templates at later amplification cycles discriminating against major templates (Kurata *et al.*, 2004; Polz and Cavanaugh, 1998; Sachse, 2004; Sipos, *et al.*, 2007). In addition, the stochastic nature of the PCR amplification process during the first few cycles of amplification can lead to a reaction-specific skew in relative distribution of amplified templates producing so-called PCR drift (Polz and Cavanaugh, 1998; Wagner *et al.*, 1994). Finally, genomic DNA isolated from environmental sources or human bodily fluids such as fecal matter is known to often contain PCR inhibitors that are able to significantly impede DNA amplification sample-wide (Lantz *et al.*, 2000). All these mechanisms result in a skew of the product-to-template ratios for most DNA species in complex samples.

Several studies have provided theoretical framework aimed to describe the individual components of the PCR or to simulate a particular PCR bias. Initial efforts focused on the modeling of individual biochemical reactions including template–enzyme kinetics (Schnell and Mendoza, 1997), DNA renaturation (Wetmur and Davidson, 1968) and DNA re-annealing (Suzuki and Giovannoni, 1996). Two recent studies described detailed simulations of kinetic processes of PCR based on the calculations of mass preservation and equilibrium coefficients (Lee *et al.*, 2006; Mehra and Hu, 2005). Both studies focused on the behavior of single-template amplification process, and in addition provided considerations of biochemical constants for each equilibrium reaction.

In this work, we extend these previous PCR simulations with a quantitative model of multitemplate PCR amplification of complex 16S ribosomal DNA mixture derived from the simulated genome pool of 400 different microbial species. The multitemplate PCR amplification is computationally assessed through the detection of amplified 16S ribosomal DNA by phylogenetic microarray.

2 METHODS

2.1 General model description

A typical experimental procedure of microbial community analysis consists of selective amplification of 16S rDNA from the total genomic DNA, applying the pool of amplified and total gDNA onto a microarray, and then measuring hybridization of this mixture to different microarray probes. Detection of bacterial species of intestinal microbiota with Microbiota Array (Paliy *et al.*, 2009) was used as a model of this method. In the *in silico* simulation of the interrogation process, developed in Matlab (version 7.3, The Mathworks, Inc.), the starting gDNA sample was subjected to successive rounds of PCR amplification producing a mixed sample containing original gDNA and amplified 16S rDNA molecules. After each PCR amplification round, detection of 16S rDNA species in the mixture was assessed by the microarray detection algorithm.

2.2 Model of the multitemplate PCR

The process of PCR amplification was simulated for each bacterial species in the mixture individually, and the accumulation of amplified 16S rDNA was modeled for each species as:

$$P_N = P_0 \times \prod_{i=1}^{N} (1 + \mathrm{AE}_i) \quad (1)$$

where P_N is the total amplified 16S rDNA after N cycles of PCR amplification, P_0 is the starting amount of 16S rDNA and AE_i is amplification efficiency at cycle i for each species 16S rDNA (Schnell and Mendoza, 1997).

Table 1. Definition of parameter values used in simulation experiments

Parameter	Description	Value
N_{SP}	Number of species in sample	400[a]
16Sfrac	Fraction of 16S rDNA per genome	0.0025
$\mathrm{AR}_{\mathrm{MAX}}$	Maximum amplification rate, $1 + \mathrm{AE}_{\mathrm{MAX}}$	1.75
$\mathrm{VOL}_{\mathrm{PCR}}$	PCR reaction volume	50 μl
$\mathrm{CONC}_{\mathrm{NUC}}$	Concentration of total nucleotides	8.0×10^{-4} M [b]
$\mathrm{CONC}_{\mathrm{ENZ}}$	Concentration of DNA polymerase	2.7×10^{-9} M [b]
$\mathrm{CONC}_{\mathrm{PR}}$	Concentration of each primer	2.0×10^{-7} M [b]
$K_{\mathrm{EQ}}^{\mathrm{TT}}$	Template–template self-annealing	1.0×10^{10} M^{-1} [c]
$K_{\mathrm{EQ}}^{\mathrm{TP}}$	Template–primer annealing	1.1×10^{8} M^{-1} [c]
$K_{\mathrm{EQ}}^{\mathrm{ET}}$	Template–primer-enzyme annealing	1.0×10^{8} M^{-1} [c]
$K_{\mathrm{EQ}}^{\mathrm{ETN}}$	Template–primer–enzyme-nucleotide annealing	1.8×10^{6} M^{-1} [c]
DET_{AVG}	Average detection limit of the microarray	5.0 pg[a]
LOAD	Amount of sample loaded onto microarray	1.5 μg[a]

[a]Based on data from (Rigsbee *et al.*, 2011) and (Paliy *et al.*, 2009).
[b]Based on available mixture composition of standard Taq PCR kits.
[c]Based on parameter estimation from (Mehra and Hu, 2005).

The overall species-specific PCR amplification rate $\mathrm{AR}_i = 1 + \mathrm{AE}_i$ at cycle i was defined as:

$$\mathrm{AR}_i = 1 + (\mathrm{AE}_{\mathrm{MAX}} - K_{\mathrm{INH}} \times g\mathrm{DNA}) \times \left(\frac{\mathrm{CONC}_{\mathrm{EQ}}^{\mathrm{TEM.PR.ENZ.NUCL}}}{\mathrm{CONC}_{\mathrm{IN}}^{\mathrm{TEM}}} \right) \quad (2)$$

where $\mathrm{AE}_{\mathrm{MAX}}$ is maximum amplification efficiency, $K_{\mathrm{INH}} \times g$DNA term defines amplification reaction inhibition due to the presence of PCR inhibitors, $\mathrm{CONC}_{\mathrm{IN}}^{\mathrm{TEM}}$ is the total concentration of single-stranded template at the melting step and $\mathrm{CONC}_{\mathrm{EQ}}^{\mathrm{TEM.PR.ENZ.NUCL}}$ is the concentration of the template–primer–enzyme–nucleotide complex (growing complementary nucleotide chain) at equilibrium. The last term in the equation models the adjustment of amplification efficiency based on biochemical reaction equilibria and takes into account template self-annealing (Mehra and Hu, 2005), template-specific PCR bias due to variability of the 16S rDNA nucleotide sequences complementary to universal primers among different species (Polz and Cavanaugh, 1998) and sample-wide reduction in amplification efficiency due to enzyme and nucleotide resource limitation (Lee *et al.*, 2006; Mehra and Hu, 2005; Sachse, 2004). A detailed description of the mathematical model is available in Supplementary Material.

2.3 Simulation of microarray detection

Microarray analysis of species 16S rDNA in the amplified sample was simulated by comparing the amount of each species rDNA versus the microarray detection limit (see below). If the species-specific 16S rDNA amount was above detection limit, the corresponding species was treated as detected, and its gDNA abundance was equal to its 16S rDNA amount in the amplified mixture. Otherwise, the species was treated as undetected, and its measured gDNA abundance was assumed to be 0.

2.4 Parameter estimation

The following modeling parameters and coefficients were estimated (Table 1):

(1) Yield of PCR amplification as a function of the amount of starting material: PCR amplification yield was estimated from semi-quantitative PCR tests that utilized between 50 and 500 ng of gDNA template and between 15 and 30 cycles of amplification (Supplementary Fig. S1). The yield rate (defined as the ratio of the final amplified amount divided by the starting DNA amount)

diminished as the starting DNA amount increased. The yield rates were used to empirically estimate K_{INH} (Table 1).

(2) Sample-wide rate of PCR amplification: quantitative PCR tests, utilizing small amounts of genomic DNA to reduce possible inhibitory effect(s), were carried out as described previously (Paliy *et al.*, 2009) to estimate AE_{MAX}. The average amplification rate was 1.75 for the exponential phase of PCR giving AE_{MAX} of 0.75, a value comparable to other reports (Acinas *et al.*, 2005; Frank *et al.*, 2008).

(3) Distribution of PCR biases due to primer annealing variability: based on experimental data (Rigsbee *et al.*, 2011), the PCR bias was modeled as lognormal distribution of the template–primer complex formation constant. The SD of the natural logarithm was set to 1.74 in order to produce a distribution of species-specific amplification rates matching experimental median per-cycle bias of 1.59% (Rigsbee *et al.*, 2011).

(4) Microbiota array detection and experimental conditions: the minimum detection limit of the microarray was estimated from the microarray validation experiments to be between 2.5 and 10 pg of 16S rDNA (Paliy *et al.*, 2009). The detection limit was modeled as a normal distribution function with mean of 5 pg, SD of 2 pg and minimum of 1 pg. The number of bacterial species in the interrogated sample was set to 400 (Rigsbee *et al.*, 2011). The distribution of starting amounts of each bacterial species in the mixture was modeled by the square of an exponential function and varied over four orders of magnitude. This produced a distribution of species DNA abundances heavily skewed toward low abundance members to match experimental data (Paliy *et al.*, 2009; Rigsbee *et al.*, 2011). The model assumed that 1500 ng of combined pool of amplified 16S rDNA and initial gDNA was added to each microarray (Paliy *et al.*, 2009). If necessary, multiple simulated PCRs were combined to obtain this 1500 ng of total material.

(5) 16S gDNA fraction: fraction of the 16S rDNA per bacterial genome was estimated by multiplying the average 16S rDNA gene copy number (6.0) by average 16S rDNA gene length (1500 bp) and dividing the product by average prokaryotic genome size (3.6 Mb)—this produced a 0.0025 fraction. NCBI microbial genomes and rrnDB databases were used to obtain these averages.

2.5 Measured outcomes of the simulated processes

Model simulations were focused on the analysis of how species detection and accuracy of community composition measurements by phylogenetic microarray depended on the amount of starting gDNA material for PCR amplification and the number of PCR amplification cycles the samples are subjected to (Bonnet *et al.*, 2002; Rigsbee *et al.*, 2011).

The model was run with all parameters defined as shown in Table 1, while varying the amount of starting material and the number of 16S rDNA-specific cycles of PCR amplification. One hundred repeated simulations were carried out for each different amount of starting material and different number of PCR amplification cycles, and the simulation results were averaged among repeats.

Two outputs of the model were used to assess the simulation: (i) the fraction of bacterial species detected by the microarray and the (ii) accuracy of quantitative measurement of bacterial relative abundance. Additional model outcomes such as fraction of nucleotides and primers used, total reaction yield and cycle-to-cycle amplification rate changes were also tracked. The accuracy of species abundance measurements was defined as:

$$\sum_{i=1}^{N} \frac{\left|\mathrm{SP.FRAC}_i^{\mathrm{TRUE}} - \mathrm{SP.FRAC}_i^{\mathrm{EXP}}\right|}{\mathrm{SP.FRAC}_i^{\mathrm{TRUE}}} \Big/ N_{\mathrm{SPECIES}}$$

where $\mathrm{SP.FRAC}_i^{\mathrm{TRUE}}$ is the true fraction of 16S rDNA represented by species i in the initial sample, $\mathrm{SP.FRAC}_i^{\mathrm{EXP}}$ is the fraction of 16S rDNA represented by species i estimated by microarray interrogation of PCR-amplified sample and N_{SPECIES} is the total number of species in the sample. $\mathrm{SP.FRAC}_i^{\mathrm{EXP}}$ for all undetected species was set to 0. In words, this accuracy measure is the average of the absolute fractional error for each species fraction estimate. A lower value represents greater accuracy, with a zero value being perfect accuracy.

2.6 Model limitations

Several parameters known to influence the behavior of standard PCRs were omitted from the model for simplicity because they were not found to be modified often in phylogenetic studies examined. These included Mg^{2+} concentration, varying annealing temperature and temperature ramp rate. Because DNA polymerases currently used for PCRs are very thermostable, denaturation of the enzyme during PCR was considered to be insignificant. Since our model was focused on the determination of amplification efficiency of each template at each PCR cycle, it did not include detailed simulation of the biochemistry of melting and extension reactions that were considered in detail previously (Lee *et al.*, 2006; Mehra and Hu, 2005). Nevertheless, the behavior of our PCR simulation was similar to those reported elsewhere (Lee *et al.*, 2006; Mehra and Hu, 2005). Finally, our algorithm did not model complexities of microarray hybridization behavior and bias; rather it assumed a perfect measurement of each species abundance as long as it was above the microarray detection limit.

3 RESULTS

3.1 Fraction of bacterial detection

Figure 1A and B displays detection fraction of microbial species in the sample as a factor of the number of PCR amplification cycles and the amount of starting material. As expected, bacterial detection increased with an initial increase in the number of PCR amplification cycles. The curves of species detection fraction were similar when starting material was kept <50 ng. After initial increase, the detection fraction started to plateau out around 15th PCR cycle and reached maximum level around cycle 20 (91–93% of bacterial species detected for samples with <50 ng starting gDNA). The detection slowly decreased with further PCR amplification to $\sim$90%.

The detection fraction curve was different for samples with higher starting amounts of gDNA. When increasing initial gDNA template from 100 ng to 1000 ng, the rate of detection became progressively lower. At 15 cycles of PCR amplification, 86, 78, 71 and 63% of all species in the sample were detected for samples containing 100, 300, 600 and 1000 ng of starting genomic DNA, respectively. The lower rate of detection is explained in part by an increased amount of PCR inhibitors in large gDNA samples reducing the rate of template amplification. The initial sample-wide amplification rate dropped from 1.75 for reactions with 1–50 ng starting material to 1.50 for samples with 1000 ng of starting gDNA (Supplementary Fig. S2). Elevated template self-annealing in large samples also contributed to a lower amplification rate. In addition, because of higher initial concentration of DNA templates, PCR reached resource limitation stage faster for larger gDNA samples. A drop in the amplification rate occurred at cycle 20 for PCR with 1 ng starting gDNA versus cycle 10 for reactions with 1000 ng starting gDNA (Supplementary Fig. S2). This lowered the 16S rDNA enrichment factor of the amplified samples (Supplementary Fig. S3). Whereas 16S rDNA constituted 86% of total sample after 15 cycles of PCR amplification for sample containing 1 ng of starting gDNA (99% after 20 cycles), this fraction was 73% for 100 ng starting gDNA sample (83% after 20 cycles, 93% after 40 cycles) and only 22% for 1000 ng starting gDNA sample (30% after 20 cycles, 51% after 40 cycles). The initial

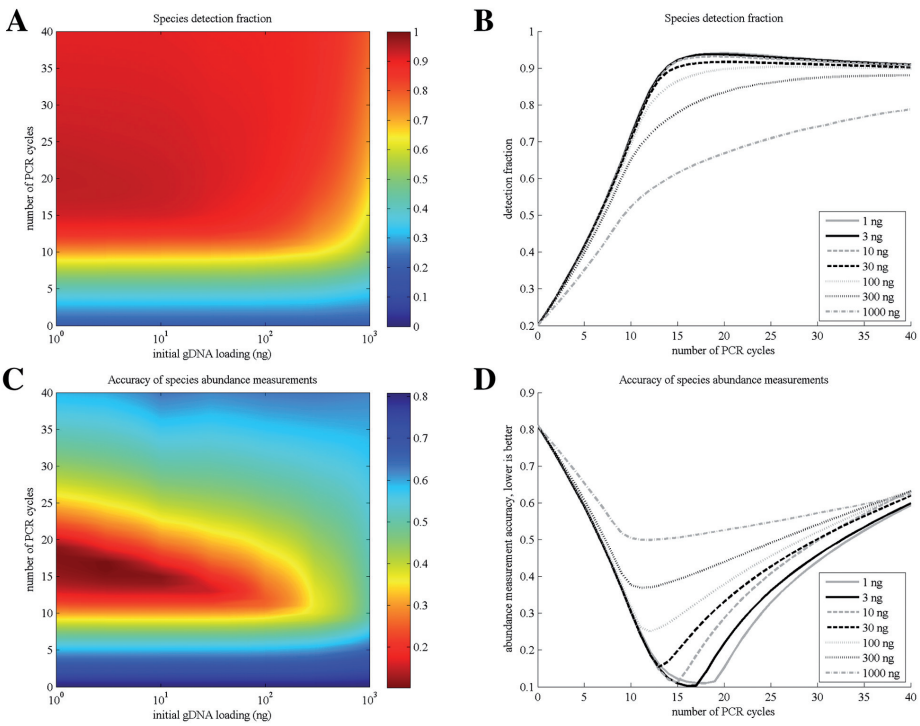

Fig. 1. Species detection (**A** and **B**) and accuracy of abundance measurements (**C** and **D**) as a function of the amount of starting material and the number of PCR amplification cycles. (A and C) The *X*-axis shows the initial amount of gDNA in a simulated sample ($\log_{10}$ scale), and the *Y*-axis shows the number of PCR amplification cycles. The species detection (A) and abundance measurement accuracy (C) are represented by a color gradient as shown in the legend. (B and D) The *X*-axis shows the number of PCR amplification cycles, whereas either the detection fraction (B) or the abundance measurement accuracy (D) are plotted on the *Y*-axis. Drawn lines correspond to specific initial gDNA sample amounts as shown in the legend. Note that in (C) and (D), lower value corresponds to better accuracy.

fraction of 16S rDNA in genomic DNA samples was 0.25%. The relative fractions of detection at 15, 20, 25 and 30 cycles of PCR amplification observed in our simulations were in good accordance with our previous experimental analyses of human fecal samples using Microbiota Array (Rigsbee *et al.*, 2011).

Carrying out PCR amplification past 20 cycles actually led to a slight reduction of species detection in samples with small starting amount of gDNA. This was connected to the existence of the simulated PCR bias among different species, since species detection did not drop when PCR bias was turned off (Fig. 2C). Our simulation results are consistent with previous observations that the use of large number of PCR amplification cycles to enrich genomic DNA samples with 16S rDNA results in reduced estimates of microbial community diversity (Bonnet *et al.*, 2002).

3.2 Accuracy of 16S rDNA abundance measurements

Figure 1C and D displays the accuracy of microarray estimation of species relative abundance in a sample. Independent of the amount of starting material used, the accuracy of community composition estimates improved with an initial increase in the number of PCR amplification cycles a sample was subjected to. The optimum accuracy was inversely proportional to the starting gDNA amount: it was similar for samples with 1–10 ng of starting gDNA, but was progressively worse for samples with higher starting gDNA amounts. The best accuracy (minimum value on Figure 1C and D graphs) was observed at cycle number 17, 15, 12 and 11 for samples with 1, 10, 100 and 1000 ng of starting gDNA, respectively. In all cases, the accuracy then started to decline sharply as the number of PCR amplification cycles continued to grow.

Several different effects influenced the observed accuracy of 16S rDNA abundance measurements. Since many bacterial species are not detected in samples subjected to a low number of PCR amplification cycles, this leads to a poor concordance of relative abundance values between initial (true) and measured sample compositions. However, because each additional PCR amplification cycle skews the original distribution of species abundances due to unequal amplification efficiency of 16S rDNA of different species, use of many PCR amplification cycles leads to a significant overall deviation of measured signals from the initial species distribution (Fig. 2D). In addition, template self-annealing introduced another amplification deviation in samples with higher starting gDNA that influenced the position of optimum accuracy (see Supplementary Fig. S4). Thus, even though a progressively higher fraction of community members is detected until at least the 20th PCR amplification cycle, accumulation of PCR amplification biases leads to optimum accuracy being reached at an earlier cycle.

3.3 Sensitivity analysis of estimated model parameters

To test the dependence of model predictions on values chosen for key parameters, sensitivity analyses were carried out where each parameter was systematically permuted while keeping all other

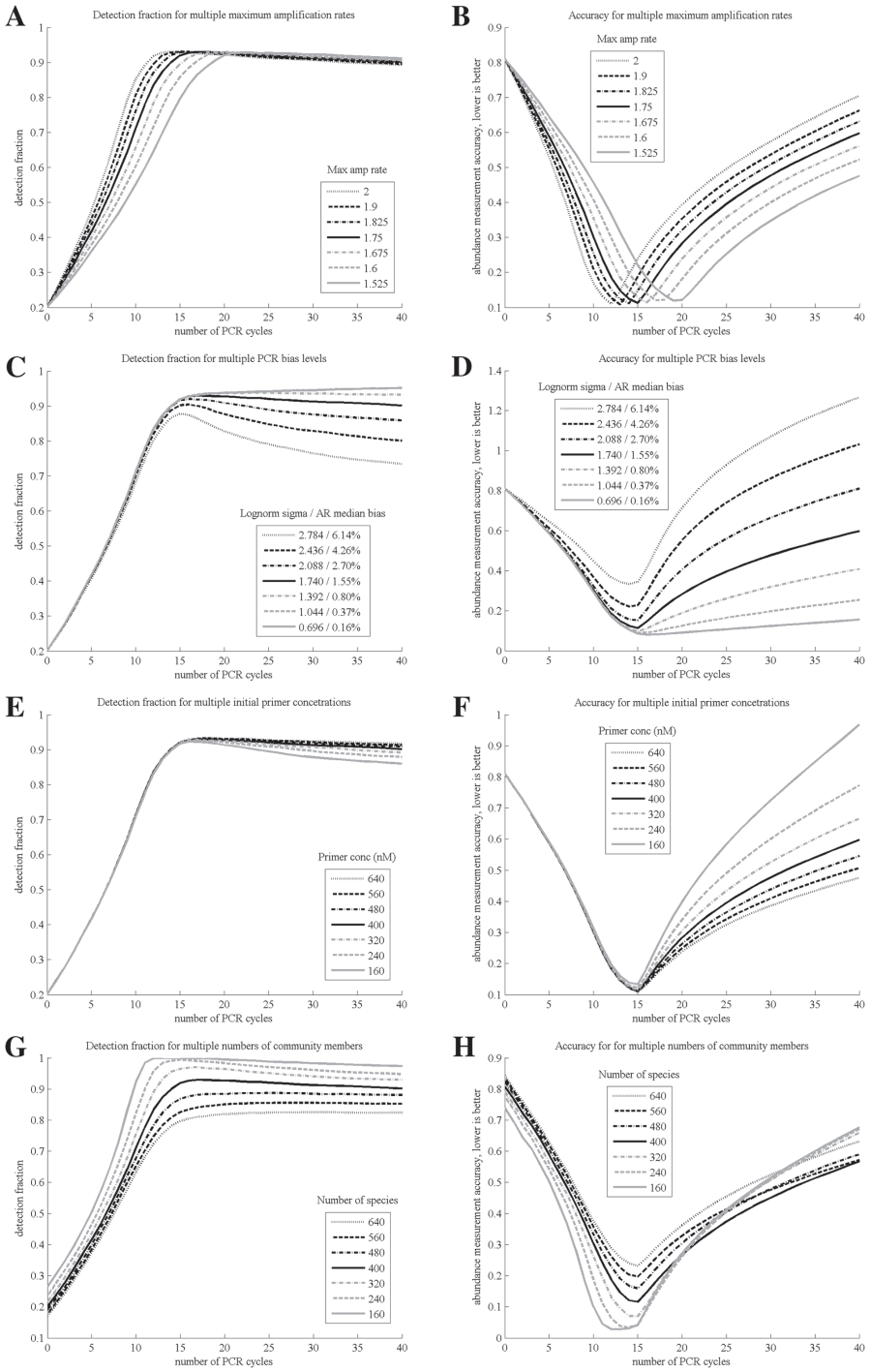

Fig. 2. Sensitivity analysis results. Figure displays the results of key parameter permutations for maximum amplification rate (**A** and **B**), PCR bias level (**C** and **D**), initial combined primer concentration (**E** and **F**) and the number of community members (**G** and **H**). Each parameter was varied in the range of ±60% of the established parameter value (±30% for maximum amplification rate). Left-hand charts display detection fraction of community members; right-hand charts show accuracy of abundance measurements. On all panels, solid black line corresponds to the default parameter value.

values unchanged. In all cases, the starting gDNA amount was set to 10 ng.

Maximum amplification rate: the rate of gDNA amplification can be influenced by the amount of inhibitors in the sample, by the complementarity of amplification primers used and by specific PCR conditions employed. Figure 2A and B shows how different maximum PCR amplification rates affect species detection and accuracy of abundance estimates. As expected, the slope of detection

fraction was generally proportional to the PCR amplification rate. The 90% detection point was reached at PCR cycle 12, 14, 17 and 21 for reactions with amplification rates of 2.00, 1.75, 1.60 and 1.45, respectively. The highest detection fraction was the same for all simulations and thus was independent from the starting amplification rate. The optimum accuracy was generally observed for the PCR cycle that corresponded to detection fraction reaching 90% level.

Template PCR amplification bias: due to the differences in the primer annealing and in sequence replication rates, different 16S rDNA species display slightly different rates of PCR amplification, which in a complex sample can lead to deviations of the post-PCR template ratios from the original sample composition. A number of approaches to experimentally affect PCR bias level are known and can be used (Kurata *et al.*, 2004; Sipos *et al.*, 2007). For this sensitivity analysis, the SD of the template–primer equilibrium K_{EQ}^{TP} constant was systematically permuted to determine the influence of PCR bias level on both the detection and accuracy performance of microarray measurements (Fig. 2C and D). Maximum detection fraction was inversely proportional to the bias level. With a 0.15% median per-cycle amplification rate bias, 95% of species were detected in our simulations; however, this fraction dropped to 87% for amplifications with 5.9% median per-cycle bias (Fig. 2C).

An even larger negative effect of PCR amplification bias was observed for the accuracy of abundance measurements (Fig. 2D). This was expected since each PCR cycle introduces deviations into relative species abundances. With PCR bias close to 0, accuracy stayed at the same level after reaching an optimum.

Concentration of amplification primers: Figure 2E and F shows simulation results of varying the initial concentration of amplification primers. Because in standard PCR primers are always in abundance compared with the amount of template (2.6×10^{-12} M of 16S rDNA template in 50 ng of total genomic DNA versus 2.0×10^{-7} M of each primer), primer concentrations are often assumed to be invariant during PCR. Somewhat unexpectedly, both of the main assessed parameters showed primer concentration dependence. Though the optimum accuracy and the highest detection fraction were generally not affected, the decline in both parameters with higher number of amplification cycles was inversely proportional to the overall primer concentration in the reaction. Increasing each primer concentration beyond 350–400 nM had negligible additional effect. Our finding can be explained by the fact that because phylogenetically conserved primers with several degenerate positions are often used in 16S rDNA amplification studies (Paliy *et al.*, 2009), the actual effective concentrations of each forward and reverse primer for each specific 16S rDNA template are likely significantly lower. This effect was included in our model. A lower effective primer concentration will lead to a lower overall template amplification rate through shifting the reaction equilibrium. This reduction will be particularly noticeable when concentration of the templates is comparable with primer amounts at the later stages of PCR amplification.

Number of species in the community sample: based on the previous analysis of intestinal microbial communities (Rigsbee *et al.*, 2011), our simulations were run with 400 species members. However, community diversity is expected to be vastly different for diverse ecological environments and microbial communities. Figure 2G and H shows detection and accuracy results for simulations where community complexity was varied between 160 and 640 members. As expected, overall fraction of species detection decreased with an increase in the number of community members, since the same amount of genomic DNA was distributed among larger number of members. We detected all members of the 160-member community but only ~84% of the 640-member community. A similar relationship was also observed for accuracy estimates—microarray results were generally more accurate for a low-member community due to the higher fraction of detected members.

4 DISCUSSION AND CONCLUSIONS

In this work, we have developed the first detailed model of PCR amplification and microarray detection for the quantitative analysis of microbial community structure. Model performance matched well available experimental data including total number of detected species, total PCR reaction yield and species detection differences for 15, 20, 25 and 30 cycles of PCR amplification [see Supplementary Table S1 and (Rigsbee *et al.*, 2011)]. Model simulations led to the following conclusions:

(1) Species detection showed a positive relationship with the number of PCR amplification cycles and lower template-to-template amplification bias. On the other hand, it was inversely proportional to the number of species comprising the interrogated community. A detection fraction plateau was reached in each case due to a significant and quick drop in the PCR amplification rate once the template concentration became high enough to affect chemical reaction equilibria.

(2) Accuracy of 16S rDNA abundance measurements had a non-monotonic relationship with most tested parameters. The accuracy initially improved with each successive PCR amplification cycle due to a significant increase in the fraction of detected species. However, once reaching an optimum, the accuracy started to decline due to the rising PCR amplification bias with each additional amplification cycle. Overall, PCR amplification bias had the most profound effect on the detection and accuracy estimates among all parameters tested in our model.

(3) A caution has to be exercised when comparing species composition and abundance among microbial communities with significantly different numbers of community members. With the same experimental protocol used, a community with many members will have a lower detection fraction than a community with fewer species. Thus, the actual differences among such communities would be underestimated.

(4) Not surprisingly, our experiments indicated that the total yield of PCR amplification correlated positively with the amount of starting material as well as with the number of amplification cycles employed. The yield neared a plateau in each case (Supplementary Fig. S5) which coincided with a drop in the PCR amplification rate (Supplementary Fig. S2). Previous studies also showed that efficiency of DNA duplex denaturation is decreased at higher concentrations, which leads to poor PCR performance when too much sample DNA is added to the mixture (Sachse, 2004). While this indicates that carrying out many PCRs with smaller amounts of starting material will produce the best overall yield from the same amount of available gDNA, the use of many PCRs carries significant financial and labor costs that also must

be considered. The same amount of material (1500 ng) was 'loaded' onto the microarray in our simulations; therefore, for reactions with small starting amounts and low number of PCR amplification cycles, the simulation assumed that many PCRs were combined to achieve the needed load. This is not likely to be practical for most experiments: assuming a 50% loss of PCR products during reaction purification, 27 separate PCR amplification reactions would have to be combined for samples with 10 ng of starting gDNA subjected to 15 cycles of PCR amplification (see Supplementary Fig. S6).

Overall, when assuming maximum sample-wide PCR amplification rate of 1.75, the optimal conditions for PCR amplification of complex 16S rDNA samples included a combination of smaller amount of starting gDNA (< 50 ng) and moderate number of PCR amplification cycles (15–20 cycles). The use of a higher number of PCR amplification cycles should be avoided (Kanagawa, 2003). The particular optimal values for starting gDNA amount and the number of amplification cycles will be affected by the specific microbial community under investigation and the details of the experimental protocol. The developed model is well suited to provide these optimal condition estimates for different experimental setups. Our model predicts that combination of 50 ng starting material and 18–20 cycles of PCR amplification will require pooling of between 5 and 8 independent PCRs (assuming 1500 ng microarray load), which is of reasonable cost and labor-wise. An additional advantage of pooling multiple independent PCR amplifications of the same sample is the mitigation of potential measurement biases associated with PCR drift (Polz and Cavanaugh, 1998; Wagner *et al.*, 1994). Our experimental PCR amplification and purification tests [data not shown and Paliy *et al.* (2009); Rigsbee *et al.* (2011)] are consistent with these predictions.

In summary, this work develops a mathematical model which can assist with establishing the optimal experimental conditions for the selective amplification and microarray-based measurements of community 16S rDNA. The developed model can be easily modified to simulate the interrogation of other microbial communities as well as other microarray designs or PCR amplification protocols. For example, to simulate different microbial communities the number of species and their expected abundance distribution in the sample can be modified (Fig. 2G). On the other hand, microarray detection limit and the number of interrogating probesets can also be changed to simulate a different microarray design. Furthermore, the model can be further expanded to include additional parameters influencing microarray detection such as cross-hybridization behavior among different probes and 16S rDNA fragments (Rigsbee *et al.*, 2011).

Funding: National Institutes of Health grant (AT003423) to O.P.

Conflict of Interest: none declared.

REFERENCES

Acinas,S.G. *et al.* (2005) PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl. Environ. Microbiol.*, **71**, 8966–8969.

Bonnet,R. *et al.* (2002) Differences in rDNA libraries of faecal bacteria derived from 10- and 25-cycle PCRs. *Int. J. Syst. Evol. Microbiol.*, **52**, 757–763.

Brodie,E.L. *et al.* (2007) Urban aerosols harbor diverse and dynamic bacterial populations. *Proc. Natl Acad. Sci. USA*, **104**, 299–304.

Cannone,J.J. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC bioinformatics*, **3**, 2.

Eckburg,P.B. *et al.* (2005) Diversity of the human intestinal microbial flora. *Science*, **308**, 1635–1638.

Frank,J.A. *et al.* (2008) Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Appl. Environ. Microbiol.*, **74**, 2461–2470.

Gao,Z. *et al.* (2007) Molecular analysis of human forearm superficial skin bacterial biota. *Proc. Natl Acad. Sci. USA*, **104**, 2927–2932.

Huber,J.A. *et al.* (2007) Microbial population structures in the deep marine biosphere. *Science*, **318**, 97–100.

Kanagawa,T. (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J. Biosci. Bioeng.*, **96**, 317–323.

Kent,A.D. and Triplett,E.W. (2002) Microbial communities and their interactions in soil and rhizosphere ecosystems. *Annu. Rev. Microbiol.*, **56**, 211–236.

Kurata,S. *et al.* (2004) Reevaluation and reduction of a PCR bias caused by reannealing of templates. *Appl. Environ. Microbiol.*, **70**, 7545–7549.

Lantz,P.G. *et al.* (2000) Biotechnical use of polymerase chain reaction for microbiological analysis of biological samples. *Biotechnol. Annu. Rev.*, **5**, 87–130.

Lee,J.Y. *et al.* (2006) Simulation and real-time monitoring of polymerase chain reaction for its higher efficiency. *Biochem. Eng. J.*, **29**, 109–118.

Mehra,S. and Hu,W.S. (2005) A kinetic model of quantitative real-time polymerase chain reaction. *Biotechnol. Bioeng.*, **91**, 848–860.

Paliy,O. *et al.* (2009) High-throughput quantitative analysis of the human intestinal microbiota with a phylogenetic microarray. *Appl. Environ. Microbiol.*, **75**, 3572–3579.

Polz,M.F. and Cavanaugh,C.M. (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.*, **64**, 3724–3730.

Rigsbee,L. *et al.* (2011) Optimizing the analysis of human intestinal microbiota with phylogenetic microarray. *FEMS Microbiol. Ecol.*, **75**, 332–342.

Sachse,K. (2004) Specificity and performance of PCR detection assays for microbial pathogens. *Mol. Biotechnol.*, **26**, 61–80.

Schnell,S. and Mendoza,C. (1997) Theoretical description of the polymerase chain reaction. *J. Theor. Biol.*, **188**, 313–318.

Sipos,R. *et al.* (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol. Ecol.*, **60**, 341–350.

Suzuki,M.T. and Giovannoni,S.J. (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.*, **62**, 625–630.

Wagner,A. *et al.* (1994) Surveys of gene families using polymerase chain-reaction - PCR selection and PCR drift. *Syst. Biol.*, **43**, 250–261.

Wetmur,J.G. and Davidson,N. (1968) Kinetics of renaturation of DNA. *J. Mol. Biol.*, **31**, 349–370.