J. Math. Biol. (2013) 67:1141–1161
DOI 10.1007/s00285-012-0586-x

Mathematical Biology

Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens' theorem

**Michael C. Wendl · Karthik Kota ·
George M. Weinstock · Makedonka Mitreva**

Received: 6 September 2011 / Revised: 28 August 2012 / Published online: 11 September 2012

Abstract Metagenomic project design has relied variously upon speculation, semi-empirical and ad hoc heuristic models, and elementary extensions of single-sample Lander–Waterman expectation theory, all of which are demonstrably inadequate. Here, we propose an approach based upon a generalization of Stevens' Theorem for randomly covering a domain. We extend this result to account for the presence of multiple species, from which are derived useful probabilities for fully recovering a particular target microbe of interest and for average contig length. These show improved specificities compared to older measures and recommend deeper data generation than the levels chosen by some early studies, supporting the view that poor assemblies were due at least somewhat to insufficient data. We assess predictions empirically by generating roughly 4.5 Gb of sequence from a twelve member bacterial community, comparing coverage for two particular members, *Selenomonas artemidis* and *Enterococcus faecium*, which are the least (∼3 %) and most (∼12 %) abundant species, respectively. Agreement is reasonable, with differences likely attributable to coverage biases. We show that, in some cases, bias is simple in the sense that a small reduction in read length to simulate less efficient covering brings data and theory into essentially complete accord. Finally, we describe two applications of the theory. One plots coverage probability over the relevant parameter space, constructing essentially a "metagenomic design map" to enable straightforward analysis and design of future projects. The other

M. C. Wendl · K. Kota · G. M. Weinstock · M. Mitreva
The Genome Institute, Washington University, St. Louis, MO 63108, USA

M. C. Wendl · G. M. Weinstock · M. Mitreva
Department of Genetics, Washington University, St. Louis, MO 63108, USA

M. C. Wendl (✉)
Department of Mathematics, Washington University,
4444 Forest Park Blvd., Campus Box 8501, St. Louis, MO 63108, USA
e-mail: mwendl@wustl.edu

gives an overview of the data requirements for various types of sequencing milestones, including a desired number of contact reads and contig length, for detection of a rare viral species.

Keywords DNA sequencing · Coverage · Microbiome · Metagenomics

Mathematics Subject Classification (2000) 05A10 · 60D05 · 62K05 · 92B99

1 Introduction

Microbes are both ubiquitous and singularly important to almost every aspect of life as we know it. There is no shortage of remarkable statistics that might be quoted, for example symbiont microbial cells outnumber human somatic cells by about 10 fold in most individuals, microbes represent about half the world's biomass, and most of the probably more than 10 million bacterial species remain to be discovered. Such numbers contrast starkly with our relatively limited understanding of these organisms, which stems largely from difficulties in isolating and culturing most species in a laboratory setting. However, technology has lately reached the point where comprehensive metagenomic approaches are now being used. Here, whole-genome shotgun (WGS) sequencing is applied directly to the collective DNA of a community of organisms. A number of metagenomes have already been examined in this way (Breitbart et al. 2002; Tyson et al. 2004; Venter et al. 2004; Tringe et al. 2005; Gill et al. 2006; Culley et al. 2006; Angly et al. 2006; Martín et al. 2006; Rusch et al. 2007; Schlüter et al. 2008; Qin et al. 2010; Hess et al. 2011).

Project design remains a significant issue facing metagenomic research. In particular, it is difficult to know how much sequence data should be generated for any particular community. Early projects in the Sanger-era of sequencing often made pragmatic choices based simply on speculation (Handelsman et al. 1998) or budgetary constraints (Kunin et al. 2008). Sequencing was relatively expensive, limiting the amount of data. This meant that while simple metagenomic communities could still be mostly reconstructed (Tyson et al. 2004; Culley et al. 2006), large tracts within highly complex communities would necessarily be left uncharted (Venter et al. 2004; Tringe et al. 2005).

The commonality across all sequencing scenarios is that project success depends strongly on the notion of *covering* (Wendl and Wilson 2008, 2009a,b), i.e. the process that randomly places one-dimensional DNA segments onto larger genomic DNA targets. Venter et al. (2004) summarize the coverage idiosyncrasies of metagenomic sequencing in terms of the differences in both genome size among the member species and among their relative abundances. In essence, if abundance levels are roughly uniform, any single sequencing read is more likely to have come from a large genome rather than a small one. If instead genome sizes are all similar, this read probably represents an abundant species of individuals rather than a rare one. The sampling dynamics of an actual metagenomic project are a community-specific mixture of these two phenomena and the obvious danger is one of missing the proverbial "needle in the haystack" (Kowalchuk et al. 2007). That is, data may not adequately capture a member

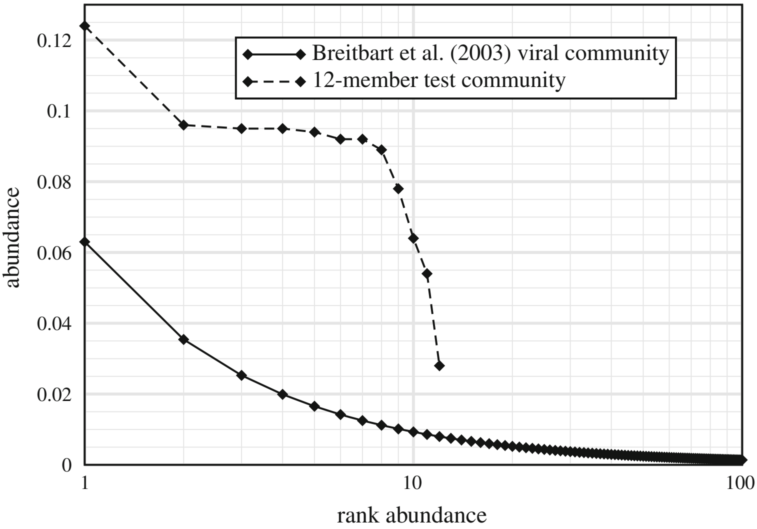

Fig. 1 Rank abundance curves are shown for the 12-member test microbial community used here for comparison and for the viral community analyzed by Breitbart et al. (2003). The latter was estimated to have around 1,200 species and to be distributed according to the power law $y = 0.063x^{-0.831}$, assuming a 50 kb average genome size

that plays some particularly vital internal role within the community and/or has some otherwise important biomedical relevance outside the community. The serendipitous discovery of the proteorhodopsins is a good example (Béjà et al. 2000).

Abundance biases are especially important in metagenomic projects because they can be quite extreme. Consider the viral community studied by Breitbart et al. (2003), which was estimated to contain around 1200 species. Its top 10 members, numbering about 0.8 % of those species, account for about 22 % of the community biomass (Fig. 1). Sequence representation will accrue rapidly for them, while their rare counterparts having abundances only on the order of 10^{-4} will be much more difficult to recover.

The economics of DNA sequencing have improved dramatically with the commercialization of so-called next-generation technologies (Harismendy et al. 2009), suggesting that comprehensive studies of some of the more complex metagenomes are now becoming feasible. It is likely that the amounts of data that will have to be generated in such projects will be larger than what is now typical. For example, the remarkable figure of 10 Tb (more than 3,000 human genomes) has been floated for a single instance of a soil metagenome (Riesenfeld et al. 2004).

With a few exceptions, the current *de facto* standard methods for making such calculations rely on an elementary extension of traditional single-genome coverage theory (Venter et al. 2004; Tringe et al. 2005; Allen and Banfield 2005; Kunin et al. 2008). Specifically, a species is taken to have a sequence redundancy of $\rho = \alpha\, R\, L/\gamma$, i.e. the respective product of its abundance, total number of project reads and read length, all divided by the species' sequence-accessible genome size. (Mathematical notation is listed in Table 1). This formula is simply the expected redundancy of data that will represent the species. Substitution into the classical coverage equation

Table 1 Mathematical notation

Variable	Meaning
α	Abundance of a species within metagenomic community
γ	Size in nucleotides of sequenceable genome
L	Average length in nucleotides of a sequence read
R	Total number of sequenced reads for a community
μ	Expected number of reads for target species: $\mu = \alpha R$
φ	Probability of a position being covered: often L/γ
ρ	Avg. number of reads spanning a position (redundancy): $\rho = \mu\varphi$
η	Steven's series limiter: the smaller of R and $\mathrm{int}(1/\varphi)$
B	Number of sequence gaps in target species (random variable)
T	Number of reads hitting target species (random variable)
λ	Contig size in target species (random variable)
C	Coverage: amount of genome covered by reads (random variable)
V	Vacancy: complement of coverage (random variable)

(Clarke and Carbon 1976) then yields the expected number of covered bases (Venter et al. 2004; Wooley et al. 2010) as $E\langle C\rangle = \gamma\,(1 - e^{-\rho})$. Lander–Waterman theory (Lander and Waterman 1988) can be further combined to obtain the expected contig length (Rusch et al. 2007) for this species as $E\langle\lambda\rangle = L\,(e^{\rho} - 1)/\rho$. Other results are similarly derived, for example for multiple-read coverage (Chen and Pachter 2005; Tringe et al. 2005).

While such formulae are attractive because of their simplicity, the salient question is whether they are sufficient for project design. Consider the calculation by Rusch et al. (2007). They predicted that 6-fold Sanger redundancy for a 10 Mb genome at 1 % abundance should give an average contig length of about $E\langle\lambda\rangle = 50$ kb. Now consider another hypothetical species in the same project whose abundance and genome size are only 0.1 % and 1 Mb, respectively, whereby ρ remains at 6-fold. The chance that a randomly-selected read represents this second species has now been reduced by a factor of 100, but the model still predicts $E\langle\lambda\rangle = 50$ kb. Rusch et al. actually reported that most of their data falling outside the dominant species remained "strikingly fragmented", with the majority not assembling at all. This scenario illustrates a subtle property of expectation-based formulae: measures such as coverage necessarily collapse onto "universal" curves that only depend upon redundancy. In a sense, expectation theory lacks the resolution to say something about specific species, as one might be able to do with a probability model. For instance, the probabilities of the two above species being fully covered are certainly different.

We briefly mention a few other results which, however more sophisticated, are still unsuited to this particular design problem. There is an appreciable body of work in the statistics literature regarding abundance estimation and these methods are readily applied to coverage-type calculations, for example as recently described by Hooper et al. (2009). They propose an expected coverage whose modeling parameters rely on fitting data to a user-chosen kernel function. Reported shortcomings include iterative tuning of parameters, limitations of kernel fidelity, and the need to discard certain portions of the data to preserve the model's integrity. Perhaps even more important is that calculations can only be made once the project is already underway, having generated enough data for parameter-fitting. The model described by Breitbart et al. (2002) has similar technical issues and does not account for variation in genome size. Alternatively, Wendl (2008) developed the density function for the project-wide number of sequence gaps, but that equation also does not adequately consider the sampling biases mentioned above. Stanhope proposed an approximation model (Stanhope 2010) based on the idealized "occupancy" concept of covering (Wendl 2006b). That approach either takes all species at uniform genome size and abundance, or requires speculative distributions for these unknowns. Finally, there are scattered rules-of-thumb (Dutilh et al. 2009; Riesenfeld et al. 2004) whose origins are not entirely clear and upon which we also comment further below (Sect. 3.1).

These observations collectively point to the need for improved theoretical tools to quantify the metagenomic sequencing process. We propose several such results here. Most are corollaries of a generalization of Stevens' theorem (Stevens 1939; Fisher 1940; Solomon 1978; Wendl and Waterston 2002), suitably extended to account for the distribution aspect of multiple species and its ensuing "abundance bias". Like all of the methods above, this work does not strictly consider effects related to particular DNA sequence or instrumentation biases, within-species variation, or choices regarding computational processing. Consequently, we view it merely as another installment within a broader research program of metagenomic sequencing theory.

2 Results

The basic premise is to develop useful and rigorous quantitative tools for designing metagenomic projects based on the community members and the level at which one desires to characterize them. The goal might range anywhere from light sampling simply to estimate community membership, to reconstructing the dominant species, to fully recovering an extremely rare member within a very complex constituency. Consequently, we will speak of the *target species* as the basis of design. Species that are more readily accessible to sequencing than the target will almost certainly be even better characterized, while the converse is true for less accessible members. This is an inherent property of all random metagenomic sequencing.

The concept of a "target species" is implicit in expectation models and enables quantitative analysis without having to first speculate closures for the invariably unknown properties of the larger metagenomic community. This aspect is enormously practical. The closure problem is necessarily present for semi-empirical models (Hooper et al. 2009; Breitbart et al. 2002), but our theory does not depend on closure estimates.

2.1 Generalization of Stevens' theorem

The problem of covering a one-dimensional domain with finite segments had been examined for some time before being solved successfully by W. L. Stevens in 1939 using a form of the well-known probability concept of inclusion–exclusion and a clever geometric observation (Stevens 1939; Fisher 1940; Solomon 1978). We generalize this result to the scenario of covering one particular domain from among a population of distinct domains. The abstraction is clearly applicable to metagenomic sequencing.

Consider a case in which R reads of length L have been processed and define the Bernoulli probability, α, as the chance that a randomly selected read represents the target species. This parameter, often understood as the "abundance", is project-dependent. Also, let φ represent the probability that this read covers a particular base position within the target species' genome. It may simply be L/γ, or it might be assigned other values to account for overlap detection (Lander and Waterman 1988) and/or the effects of bias (see below). We can now state the following salient result.

Theorem 1 (Gap Census) *If B is a random variable denoting the number of sequence gaps within the target species' composite genome, then the probability of k gaps is*

$$P(B=k) \;=\; \binom{R}{k} \sum_{\beta=k}^{\eta} \binom{R-k}{\beta-k} (-1)^{\beta-k}\, \alpha^{\beta}\, (1-\beta\varphi)^{\beta-1}\, (1-\beta\varphi\alpha)^{R-\beta}$$

for $0 < \varphi < 1$ and $0 < \alpha \leq 1$, and where η is the smaller of R and int$(1/\varphi)$. The latter quantity represents the maximum number of reads that can be placed without overlap on the target and arises from Stevens' geometric observation (Stevens 1939; Solomon 1978). Stevens' original theorem is readily shown to be a special case for $\alpha = 1$.

This theorem can be applied either directly, or in various derivative ways to obtain rigorous probabilistic quantifiers for metagenomic sequencing. We discuss two of the more useful implementations in Sect. 2.2: the probability of complete target species coverage and the probability that the average size of contiguous regions of coverage in the target exceeds some threshold. (There are other possibilities, though of lesser practical interest; Roach 1995). Finally, we give another handy formula for community sampling, not related to Theorem 1, but derivable rather from elementary considerations.

2.2 Implementations of Theorem 1 for metagenomic sequencing

As alluded to in the above discussion of expectation models, let C and λ be the respective random variables representing the number of base positions covered in the target species' composite genome and the length of a contiguously covered segment, i.e. a "contig".

Corollary 1 (Complete Coverage) *Complete coverage of the target species, $C = \gamma$, also occurs by virtue of all gaps being filled, i.e. $B = 0$. The probability of this event is*

$$P(B=0) = \sum_{\beta=0}^{\eta} \binom{R}{\beta} (-\alpha)^{\beta}\, (1-\beta\varphi)^{\beta-1}\, (1-\beta\varphi\alpha)^{R-\beta}\,.$$

This is a high standard of coverage. More relaxed conditions based on contig size are also relevant (Roach 1995; Stanhope 2010). Here, we exploit the fact that $C \to \gamma$ much more rapidly than $E\langle\lambda\rangle \to \gamma$. That is, coverage increases appreciably faster than contig size, with a large fraction of the process existing in a state of high or even nearly complete coverage, yet still having numerous small gaps (Roach et al. 1995). This phenomenon is nicely illustrated by considering the last few events of the process, where the remaining tiny gaps are closed just before attaining complete coverage. It is only here that $E\langle\lambda\rangle$ grows rapidly as $\ldots, \gamma/3, \gamma/2, \gamma$. The effect has been confirmed empirically from the earliest sequencing projects (Fleischmann et al. 1995) and holds for metagenomic projects, as well (Martín et al. 2006).

Corollary 2 (Average Contig Size) *If coverage is almost complete, the average contig length is, to a very good approximation, a function only of the target size, γ, and the (random) number of gaps, $B = k$, and is likewise itself then a random variable. The tail probability of a value at least γ/k is*

$$P\left(E\langle\lambda\rangle \geq \frac{\gamma}{k}\right) \approx \sum_{j=0}^{k} P(B = j),$$

where $k \neq 0$ and where the coverage provision can be checked in any suitable way. For example, if $E\langle V\rangle$ is the largest allowable fractional vacancy for the target, say 1 %, then a simple corollary of expectation theory, $\rho \geq \ln(1/E\langle V\rangle)$, might be used. This result can also be generalized by replacing γ with $E\langle C\rangle$ (see above), though at an obvious additional degree of approximation.

2.3 Formula for community sampling

Sequencing can also be used in a diagnostic capacity to assess what species are present in a community (Eisen 2007; Kunin et al. 2008). In the simplest case, coverage structure and contiguity are subordinated by raw counts of reads, especially if their lengths are sufficient to identify species merely by alignment against reference sequences.

Theorem 2 (Read Count) *Let T be the random variable representing the number of reads hitting the target species. Its distribution is Poissonian, $P(T = k) = \mu^k \exp(-\mu)/k!$, with a rate $\mu = \alpha R$.*

2.4 Numerical evaluation

Theorem 1 and its corollaries have a number of interesting mathematical properties, the most relevant here of which is the convergence rate. Evaluation requires summing terms that are themselves products of progressively larger and smaller numbers. Consequently, round-off error overwhelms slowly converging series unless extended precision arithmetic is employed. While such is required for much of the parameter space, standard precision can be used for Corollary 1 if the heuristic

$$\varphi \geq \frac{\ln(\alpha R/\zeta_o)}{\alpha R} \tag{1}$$

is satisfied, where ζ_o is a constant having an empirically determined value on the order of 10. The accessible range is then roughly $P(B = 0) > 10^{-3}$, which includes most scenarios of practical interest.

2.5 Parameter estimation

The formulae above can be used either parametrically or applied for specific species. In the former role, calculations will reveal the attributes of the most extreme member, i.e. its size and abundance, that could be captured for a given P-value and amount of data. In the latter, specific estimates of φ and α can be used to determine the required data for a given probability or *vice versa*. Estimates for φ are straightforward, for example one can take advantage of the fact that bacteria largely fall within $1 \leq \gamma \leq 5$ Mb if setting $\varphi = L/\gamma$, as discussed above. Conversely, α can be approximated in various ways, including 16S rRNA screening (Liles et al. 2003; Tyson et al. 2004), or methods that utilize light shotgun data, such as protein-coding markers (von Mering et al. 2007), single-copy single-marker complements to 16S, e.g. *rpoB* (Vos et al. 2011), fitting (Hooper et al. 2009), or probabilistic modeling (Xia et al. 2011).

3 Discussion

3.1 Coverage probability as a design variable

We already mentioned above some of the shortcomings of using an expectation-based quantity such as $E\langle C\rangle$ as a *measure* for the metagenomic design problem. While Stanhope (2010) is similarly critical, several additional factors support replacement with a probability-based metric, such as $P(B = 0)$ in Corollary 1.

The more obvious issues are based on the ensemble nature of expectations themselves. That is, they only characterize trials collectively and not necessarily any single one taken alone. In most instances, variances will not be terribly large compared to respective expectations. For example, the expected number of reads hitting the target species is αR (Theorem 2) with a standard deviation of $\sqrt{\alpha R}$ (Feller 1968) and the deviation in coverage $E\langle C\rangle$ is approximately $\sqrt{\gamma}\,\exp(-\rho)$ (Wendl 2006a). Consequently, this aspect is the source of *some* uncertainty, but not its main contributor.

The much more substantive concern is actually based on the sensitivity of predictions to small changes in the measure itself. Let us first be clear about the differences in what these measures mean. $E\langle C\rangle$ represents the desired percentage of bases recovered from the target species and its value is typically chosen as something approaching, but not actually equal to 100 %. (That case is mathematically undefined). Conversely, $P(B = 0)$ is the actual probability of 100 % coverage and would be picked in roughly the same context as statistical power, e.g. 90 %.

Figure 2 shows the characteristics of both measures for a 1 Mb target species using 100 bp reads. Here, the expectation results were plotted according to $\alpha R = -\ln(1 - E\langle C\rangle)/\varphi$, which follows directly from the traditional coverage equation. $E\langle C\rangle$ gives an extremely wide range of predictions for the required data. The physical

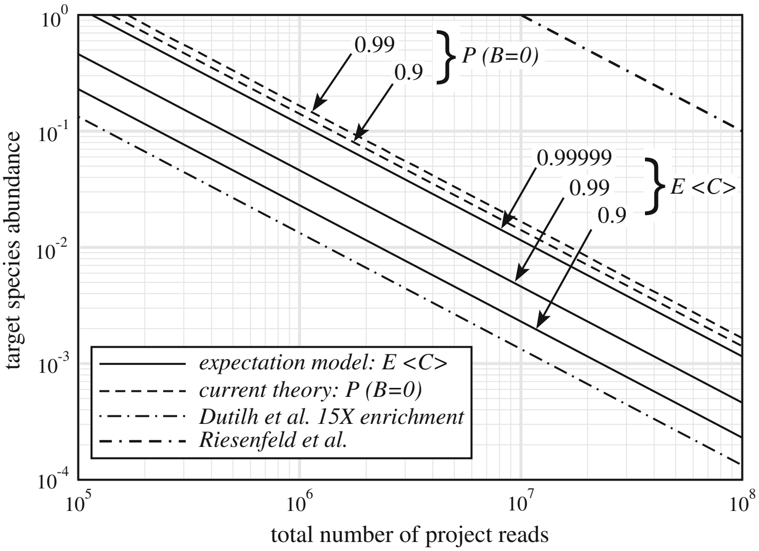

Fig. 2 Abundance versus required number of project reads for a 1 Mb target species using 100 bp reads as specified by various theories. Here, $E\langle C\rangle$ is plotted in its fractional context, i.e. as the quotient of covered bases to the genome size. The rule-of-thumb given by Dutilh et al. (2009) is plotted for an enrichment factor of 15. The rule described by Riesenfeld et al. (2004) is plotted for the metagenomic redundancy of 1,000, which leads to their sometime-quoted figure of 10 Tb for a soil metagenome

spread of curves is much greater, c.f. the distances between constant-value lines of 0.9 and 0.99 for $P(B = 0)$ and $E\langle C\rangle$, which is basically a consequence of the latter's long asymptotic tail (Wendl and Barbazuk 2005). This is further exacerbated by the somewhat subjective nature of choosing values of the measure itself. Consider that picking $E\langle C\rangle$ "close to 100 %" usually means anything between roughly 99 % (Bouck et al. 1998) and 99.996 % (Green 2001) and these bounds translate to over a two-fold difference in the required data! In short, $E\langle C\rangle$ is inherently ambiguous because of subjective thresholds chosen from within an extremely sensitive sub-domain of this function. Conversely, there is very little ambiguity in using $P(B = 0)$; it is chosen within a fairly narrow range for which the lines are very closely spaced.

A final argument, compelling more from an empirical standpoint, is that $E\langle C\rangle$ uniformly specifies fewer required data than $P(B = 0)$. Most projects that relied on the former measure reported significant assembly and contiguity problems (Venter et al. 2004; Tringe et al. 2005; Rusch et al. 2007), which seem to be at least partial by-products of having insufficient data. It is also consistent with a more general opinion that current levels of redundancy are inadequate for resolving lower-abundance organisms (Venter et al. 2004; Allen and Banfield 2005; Gill et al. 2006; Rusch et al. 2007; Nicholls 2007; Kunin et al. 2008; Schlüter et al. 2008; Wooley et al. 2010).

Figure 2 also shows two rules-of-thumb gleaned from the literature: the product of target species enrichment and redundancy should be at least 20 (Dutilh et al. 2009) and the metagenome redundancy should be around 1000 (Riesenfeld et al. 2004). The former is plotted for an enrichment factor of 15, again showing clearly insufficient data. This factor is largely arbitrary, being adjustable down to values that move the curve well past those of $E\langle C\rangle$ and $P(B=0)$. Consequently, this rule appears to be entirely too vague and unsupported to be of any practical use. The latter rule is the source of the 10 Tb soil metagenome prediction quoted above and apparently results from the mistaken presumption that the target species redundancy is the product of the species abundance and the redundancy of the metagenome as a whole. As such, it is also unsuitable for further use.

Lastly, we comment on another class of models based on contig length. The standard expectation result, quoted above (Sect. 1), is readily derivable as the ratio of coverage expectation to gap expectation, the latter obtained from Lander–Waterman theory. The formula is often avoided because it is divergent (Lander and Waterman 1988; Roach 1995), a consequence of the fact that gaps approach zero much faster than coverage approaches completion. (This can be demonstrated through simple differentiation.) More recently, Stanhope proposed a metagenomic coverage theory based on the occupancy concept (Stanhope 2010). It furnishes the probability that the largest contig exceeds some length, f. Using a property of logarithms (Beyer 1984), the main result in that paper, Eq. 1, can be written in the appreciably simpler form

$$P(\text{max contig} \geq f) = 1 - \exp\left(-\delta^f\left[(1-\delta)(2\gamma/L-3)+2\right]\right), \qquad (2)$$

where $\delta \approx 1-[1-L/(2\gamma)]^R$. This expression is not monotonically increasing, as strict addition of data requires, and violates the boundary condition of 100 % coverage in the limit of infinite R. In particular, because $\delta \to 1$, it is easy to see from Eq. 2 that $P(\text{max contig} \geq f) \to 1-\exp(-2) \approx 0.865$.

3.2 Empirical comparison for a 12-member microbial community

On the more pragmatic side, a model's ability to make worthwhile predictions can be assessed empirically. Here, we compare Corollary 1 to the data obtained from a 12-member bacterial community for which we generated roughly 4.47 Gb of sequence (about 46 million reads) from 1 lane on an Illumina GA-IIx instrument. Table 2 shows the project parameters, where "data" and "size" indicate the total amount of data generated for each genome and actual genome size, respectively. The "depth" column is their quotient, representing the average number of reads spanning each position in the genome, while the "vacant" column indicates the amount of genome remaining uncovered in the assembly. Though having only mild complexity, abundance bias is certainly evident in this population, given that the ratio of highest to lowest abundances exceeds a factor of 4. Larger values are admittedly more common, for example Breitbart et al. (2003) estimate a ratio of more than 300 (Fig. 1). However, the assemblies for such communities generally remain fragmented (Rusch et al. 2007; Hess et al. 2011) and are therefore unworkable as comparisons for the metric $P(B=0)$. The community in Table 2 is highly redundant, averaging roughly 125× data per species, with the rarest member, *S. artemidis*, still surpassing 50×. Consequently, this community is a rigorous, if preliminary test of the theory's ability to account for abundance bias.

Table 2 Sequence data for 12-member microbial community

Species (NCBI accession number)	Depth (fold)	Size (Mb)	Vacant (kb)	Data (Mb)	α
E. faecalis (AEBQ00000000)	142.5	3.00	2.26	427.4	0.096
E. coli (AJGD01000000)	62.8	4.57	2.16	287.1	0.064
F. prausnitzii (AECU00000000)	80.7	2.96	3.24	239.0	0.054
S. artemidis (AECV01000000)	56.8	2.22	2.19	126.0	0.028
E. faecalis (AEBB00000000)	139.2	2.85	1.00	396.6	0.089
E. faecalis (AEBP00000000)	115.1	3.01	3.45	346.4	0.078
E. faecalis (AEBF00000000)	148.5	2.83	1.64	420.1	0.094
E. faecalis (AEBD00000000)	147.7	2.88	1.26	425.5	0.095
E. faecalis (AEBN00000000)	132.2	3.12	1.55	412.5	0.092
E. faecalis (AEBO00000000)	131.4	3.12	3.36	409.5	0.092
E. faecalis (AEBE00000000)	131.3	3.26	2.56	426.5	0.095
E. faecium (AEBC00000000)	188.0	2.94	1.25	552.8	0.124

An important, but more subtle aspect in all empirical-theoretical comparisons is controlling for the unavoidable differences that arise as a consequence of project-specific factors, including DNA sequence and instrumentation biases (Harismendy et al. 2009) and the vagaries related to specific combinations of software packages used for processing, alignment, and assembly. In metagenomic projects, we must add inter-strain variation within species as another confounder. These factors, which we will henceforth refer to collectively as "coverage bias", tend to reduce actual performance below predictions because portions of each species' genome are inclined against locally spanning reads. While simplistic bias models have been used for posterior fitting (Port et al. 1995; Schbath 1997; Wendl et al. 2001), there is no established, general methodology for resolving this aspect of the design problem *a priori*.

Table 2 shows that the covering process for this community is indeed biased. Specifically, the amount of uncovered genome (vacancy) for each species is on the order of kilobases, despite sequence depths that often substantially exceed 100×. Empirical-theoretical comparisons for various other scenarios show that biases do not begin to manifest themselves until significant amounts of coverage have been obtained (Wendl and Barbazuk 2005; Wendl and Wilson 2008). In other words, it is not unusual that much of a genome has little to no bias and closely follows theoretical coverage predictions, some fraction is moderately biased and consequently more difficult to cover, and a small amount is extremely averse to being covered. In essence, the amount of the genome that is accessible to "routine sequencing" is somewhat smaller than the actual genome size (Thousand Genomes Project Consortium 2010; Ajay et al. 2011). This aspect can be particularly problematic for an analysis such as ours, which relies on "100 % coverage" as its metric.

We compare Corollary 1 specifically to *S. artemidis* and *E. faecium*, which are the least (∼3 %) and most (∼12 %) abundant species in this community, respectively (Fig. 3). Following the above observations, we estimate each species' accessible

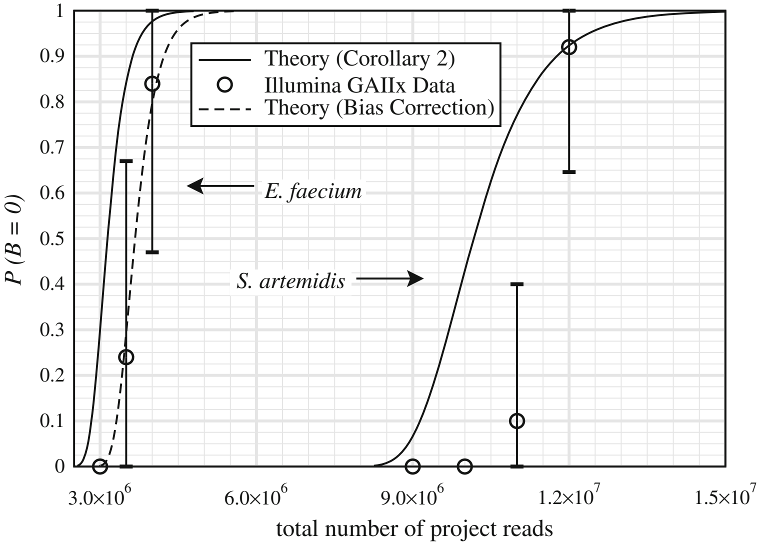

Fig. 3 Comparison of data from bacterial community in Table 2 (*circles*) to the probability of total genome coverage given by Corollary 1 (*solid curves*). Each datum represents the average of 50 random drawings from ∼46 million reads, where a corresponding indicator variable was set to 1 or 0 depending upon whether total coverage was achieved or not, respectively. *Error bars* are plotted at one standard deviation. *Dashed curve* represents simple bias correction for *E. faecium* in the form of a 3 % lowered relative read length from $\varphi = 3.046 \times 10^{-5}$ to 2.95×10^{-5}

genome size as a minimum breadth of coverage obtained after repeated sampling and assembly of a number of reads R such that $P(B = 0)$ is close to unity. For example, *E. faecium* (NCBI accession: AEBC00000000) has an actual genome size of 2,936,981 bp and $R = 4.5 \times 10^6$ implies $P(B = 0) \approx 0.997$ for this species, given an average read length of $L = 100$. We then did 50 separate assemblies of 4.5 million reads, all randomly-chosen without replacement, and found each assembly attained coverage of at least 2,933,000 bp. This figure is taken as the amount of the *E. faecium* genome that is routinely accessible to sequencing and the remaining 3,981 bases (0.14 %) are taken to be non-compliant. (Note that roughly 1.25 kb still remains uncovered, even after 188-fold redundancy for this species!) A similar calculation yields an estimate of 2,211,400 accessible bases from the total genome size of 2,215,616 of *S. artemidis*. We deemed 50 simulations per datum to be sufficient, given that the maximum coefficients of variation (quotient of standard deviation and mean) for *E. faecium* and *S. artemidis* were 0.000083 and 0.000185, respectively, for all the assemblies represented in Fig. 3.

The plots show reasonable agreement when considered in light of the bias problem. Although our elementary truncation procedure referenced above corrects somewhat for the worst factors, it unquestionably falls short. If biases are “simple”, meaning relatively benign and not distributed in complicated or extreme ways, it may be possible to further compensate by artificially lowering the read length to simulate less efficient covering. This procedure is demonstrated on *E. faecium*, where we reduced φ by about 3 %, from its actual value of 3.046×10^{-5} to a compensatory value of 2.95×10^{-5}, thereby fitting Corollary 1 almost exactly to the data. Conversely, coverage biases can

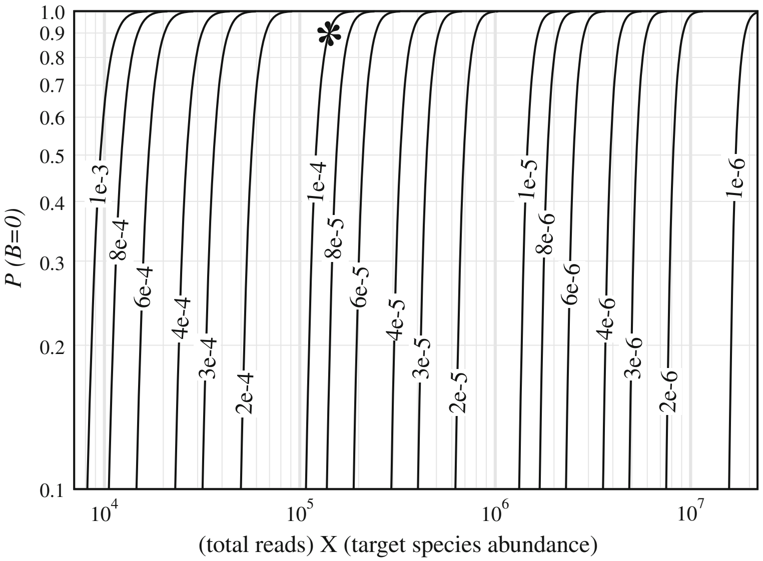

Fig. 4 The metagenomic sequencing project design map. Coverage probability is plotted as a function of the product $R \cdot \alpha$ for various values of φ. Example scenario discussed in Sect. 3.3 is denoted by the *asterisk*. *Curves* do not include any compensation for bias

also be stronger and more complicated. In such instances, simple read reduction will not help substantially, as is clear in the case of *S. artemidis*. Broadly speaking, it is difficult to characterize biases *a priori* to a degree that could be formally incorporated into a model. This remains a major unsolved problem in genomic coverage theory.

3.3 Empirical simplification and the metagenomic design map

Theorem 1 is completely general in that it describes probability as a function of all four independent variables: $P = P(B, R, \alpha, \varphi)$. Metagenomic sequencing projects impose additional *empirical* constraints on these variables such that, to a very good approximation, α and R act as a product rather than independently (implied in Fig. 2 for $B = 0$ and demonstrated in Methods), effectively reducing the problem to just three variables for gap census, $P = P(B, \alpha R, \varphi)$, and two for coverage, $P = P(\alpha R, \varphi)$. Contrast this to the functional dependence of $E\langle C\rangle$ on only a single variable, the redundancy, which lumps φ into the product $\alpha R \varphi$.

With respect to coverage, the two-variable dependence enables us to construct what is essentially a "design map" for all metagenomic projects in the form of a single plot (Fig. 4). Assuming estimates of φ and α are available for the target species, one simply picks the covering probability on the ordinate, moves horizontally from there to the intersection with the appropriate φ curve, then moves vertically down and to the corresponding αR value on the abscissa. The required number of reads for the project is then found by simply dividing this value by α. The largely vertical stature of the curves reiterates the observation that predicted data requirements are relatively insensitive to the chosen measure.

Let us illustrate the process with a brief example. Suppose our hypothetical 1 Mb target discussed above in the context of the Rusch et al. (2007) project is to be fully recovered at 90 % power using 100 bp reads. This scenario is denoted by the asterisk in Fig. 4 and corresponds to an abscissa value of roughly $\alpha R = 1.4 \times 10^5$, or a target redundancy of 14×. Given its abundance of 0.1 %, the total number of project reads is then about 140 million, or 14 Gb of total sequence data. For comparison, we cite Rusch's actual figure of about 6.4 million assembled Sanger reads (5.9 Gb of data), as well as the expectation-based prediction of 92 million next-gen 100 bp reads (9.2 Gb of data), assuming we have chosen $E\langle C\rangle = 99.99\,\%$. Note that this calculation does not include any reduction of φ to compensate for bias, as discussed above.

Let us also illustrate the compounding effect of size by now increasing the target to 10 Mb while holding all other parameters constant. Expectation theory simply multiplies everything by 10, according to the rule that the redundancy is constant if we maintain $E\langle C\rangle = 99.99\,\%$. That is, 920 million reads would now be generated. However, the probability equation accounts for the fact that it is indeed harder to cover a bigger target with constant-size reads. Instead of simply multiplying by 10 to get 1.4 billion reads, the above calculation procedure specifies 1.65 billion reads, or 16.5× target redundancy at 90 % power.

3.4 Assessing community membership

So far, we have concentrated on the special case $P(B = 0)$, i.e. full coverage, as the relevant measure, which will be useful primarily for discovery-oriented projects that rely on assembly of previously unknown species. However, there is also increasing interest in application-oriented projects that seek instead to assess community membership, the goal being to accumulate enough sequence to determine whether a known species is present or not (Eisen 2007; Kunin et al. 2008; Stanhope 2010). Because these will rely on alignment more than assembly, read hits and contig lengths are also relevant, suggesting application of Corollary 2 and Theorem 2.

Consider the example of a 50 kb target at 0.05 % abundance (Fig. 5), which is characteristic of a relatively rare virus. Assuming a read length of 100 bp, i.e. $\varphi = 0.002$, we plot a broad spectrum of representative sequencing milestones. (These are not necessarily indicative of the minimum or maximum required amounts of information to reliably indicate species presence, since those undoubtedly vary with the species. The analysis is also predicated on the existence of a well-posed viral reference sequence, which can be an issue in cases of high evolutionary rate). Calculations show that even a moderate number of project reads will very likely result in some number of reads hitting the target organism. For example, for $P(T = k) = 0.95$, the $k = 5$, 20, and 100 thresholds are reached with around $R = 18{,}000$, 55,000, and 235,000 project reads, respectively. These cases are not predicated on any coverage model and it is likely that the reads will exist almost exclusively as singletons without any real contig structure. Expectation theory concurs, for instance it suggests around 20 % coverage for the 100 read hit threshold.

Substantive contigs start to form only at higher levels of coverage. For example, average contig length reaches 2 kb (20 read lengths) at 95 % probability only after

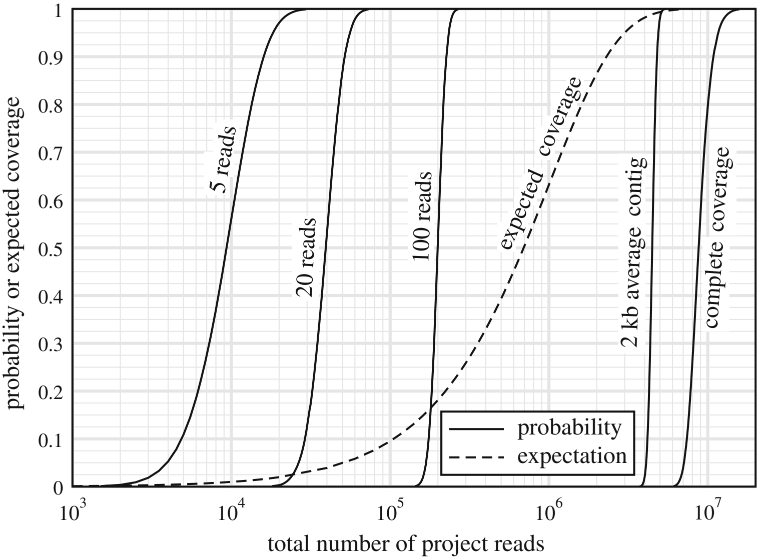

Fig. 5 Quantification of various sequencing milestones when using 100 bp reads for a 50 kb target at an abundance of 0.0005 (i.e. 0.05 %). The curve for expected coverage is computed from classical theory, while probability results are calculated from Corollaries 1 (complete coverage) and 2 (average contig size ≥2 kb, implying upper limit of 25 gaps) and Theorem 2 (number of reads hitting target, here ≥5, ≥20, or ≥100, without regard to their associated coverage structures). Curves do not include any compensation for bias

about 4.9 million project reads. Expectation theory predicts $> 99\,\%$ coverage at this point, consistent with our earlier assertions regarding loss of resolution of $E\langle C\rangle$. In other words, despite "almost" complete coverage, the actual genome is still appreciably fragmented. The apparent contradiction is simply a consequence of the fact that the rate of change of coverage in the late-stages of a project is very small. For comparison, Fig. 5 also shows the complete coverage curve, $P(B = 0)$. Here, roughly 11.7 million project reads are required at the same 95 % probability level.

The results for this target virus are readily transformed to other abundance values on the basis that α and R act asymptotically as a product. For example, this same virus at an abundance of 0.5 %, i.e. 10 times more frequent than above, would require approximately 23,500 project reads for 100 hitting reads, 490,000 for 2 kb average contig length, and 1.17 million reads for complete coverage. Note that transforming does not hold in the case of changing read length or genome size for Theorem 1 or its derivative implementations. These expressions would require new evaluation. It does hold for Theorem 2, since that result does not speak to coverage structure and is independent of read and genome length.

Finally, it is interesting to assess the maximum data required for very complex communities. For instance, the (Breitbart et al. 2003) power law estimation suggests the least abundant species in their viral community (Fig. 1) is on the order of $\alpha = 10^{-4}$, about 5 times more rare than the example just discussed. Taking a conservative value of $\varphi = 0.0008$ to account for both bias-related coverage inefficiencies and overlaps in assembling unknown species, Fig. 4 suggests $\alpha \cdot R \approx 20,000$ for high probability of complete coverage. This implies $R = 200$ million reads, or 20 GB of data for 100 bp read length. For equally rare 2 Mb bacteria take $\varphi = 2 \times 10^{-5}$, whereby $\alpha \cdot R \approx 1 \times 10^{6}$, indicating 10 billion reads and 1 Tb of data. Such communities are probably near the outer edge of the design space, suggesting an approximate upper bound for the required data.

3.5 Closing remarks

We have described a rigorous mathematical framework for the analysis and design of metagenomic sequencing projects that does not suffer from various resolution, consistency, or closure problems of earlier works. Though it does not address every outstanding issue, including those related to bias, the theory will be useful for a broad spectrum of calculations. We demonstrated several such aspects above, including use of $P(B = 0)$ as a coverage metric, empirical comparison to a bacterial community, use of the "product simplification", and community membership assessment. Numerical implementations of the mathematical results are straightforward, though we are glad to furnish our own code upon request.

Some have argued that sufficiently complex communities will necessarily remain beyond reach (DeLong 2005; Wooley et al. 2010), primarily because of limitations in sampling, while others have maintained that it is simply a matter of generating enough data (Venter et al. 2004; Tyson et al. 2004; Allen and Banfield 2005). This issue may be debatable in the philosophical sense of "proving a negative". Yet, in a practical sense, our theory furnishes quantitative conditions under which even the most complex metagenomes can be decoded and the least abundant species recovered. Developments in instrumentation continue apace, suggesting many of these communities will be within reach in the near future.

4 Methods

4.1 Proof of Theorem 1

Assume all reads are independently and identically distributed (IID) among species in the metagenome, each with a Bernoulli probability α of representing the target species. The probability that i of R reads are indeed derived from the target species is then binomial. Consequently, the probability of k gaps in the target species is obtained by further conditioning Stevens' theorem (Stevens 1939; Solomon 1978; Wendl and Waterston 2002) upon the number of resident reads $i \in \{k, k+1, \ldots, R\}$,

$$P(B = k) = \sum_{i=k}^{R} \binom{R}{i} \alpha^{i} \, (1-\alpha)^{R-i} \binom{i}{k} \sum_{j=k}^{\eta} \binom{i-k}{j-k} (-1)^{j-k} \, (1 - j\varphi)^{i-1} \, .$$

Here, $\eta = \min\Big(i, \mathrm{int}(1/\varphi)\Big)$ is the appropriate Stevens limiter. Distribute the binomial and re-order the resulting set of terms, effectively switching the inner and outer summations, to obtain

$$P(B=k)=\sum_{j=k}^{\eta}\sum_{i=j}^{R}\binom{R}{i}\binom{i}{k}\binom{i-k}{j-k}\alpha^{i}(1-\alpha)^{R-i}(-1)^{j-k}(1-j\varphi)^{i-1}.$$

The following combinatorial identity can readily be constructed

$$\binom{R}{i}\binom{i}{k}\binom{i-k}{j-k}=\binom{R}{k}\binom{R-k}{j-k}\binom{R-j}{R-i}$$

and substituting this result leads to the factored expression

$$P(B=k)=\binom{R}{k}\sum_{j=k}^{\eta}\binom{R-k}{j-k}\frac{(-1)^{j-k}f^{j}}{1-j\varphi}\sum_{i=j}^{R}\binom{R-j}{R-i}g^{R-i}f^{i-j},$$

where $f=\alpha(1-j\varphi)$ and $g=1-\alpha$. Reversing the order of terms for the inner summation and making a suitable change of variables shows that the inner summation collapses via the binomial theorem to $(f+g)^{R-j}$. Theorem 1 follows from straightforward algebra.

4.2 Proof of Theorem 2

Given the IID property of reads, the Bernoulli proposition of either hitting or missing the target species implies binomial distribution. Theorem 2 follows directly from its Poisson approximation (Feller 1968), justified by the fact that α is sufficiently close to zero and $R \gg 1$.

4.3 Derivation of numerical heuristic in Equation 1

The heuristic is based on the notion that rate of growth of successive terms in Corollary 1 is bounded to the degree that the largest one does not overwhelm standard arithmetic precision. The first term is always unity, so we focus on the second. Given $R \gg 1, \alpha < 1$, and φ generally less than 0.002, we use asymptotic approximation, finding $R\alpha\,\exp\bigl(-\alpha\varphi R\bigr) \leq \zeta_o$, where ζ_o is our empirically-chosen limiter. Straightforward algebra leads to Eq. 1.

4.4 Collapse of variables

The independent variables in Theorem 1 are governed by $B \geq 0$, $R > 0$, $0 < \varphi < 1$, and $0 < \alpha \leq 1$. However, metagenomic sequencing projects place the further empirical restrictions that R and φ are very large and small compared to 1, respectively, and furthermore that $R \gg 1/\varphi$. The last equation means that the overall number of reads in a project is far more than the minimum number required to cover just the target species. These conditions further imply $R \gg \beta$ and $R \gg B$, enabling a significant simplification of the system, wherein the number of independent variables is reduced by one.

The outer and inner combinatorial terms are well-approximated by $R^k/k!$ and $R^{\beta-k}/(\beta-k)!$, respectively, whereby $(\alpha R)^\beta$ can be factored. Asymptotic approximation also applies, such that $(1-\beta\varphi\alpha)^{R-\beta} \sim \exp(-\alpha R \cdot \beta \cdot \varphi)$. Finally, the series is always limited by $1/\varphi$ rather than R, meaning that in all places where α and R appear, they act as a product.

4.5 Sequence generation and analysis

Whole genome shotgun libraries were constructed from 1 μg of starting DNA. The DNA samples were fragmented, end repaired, A-tailed, and ligated. The ligation was size selected for 300–500 bp fragments via ampure beads and 5 μl were then amplified. The final library was quantitated via Qubit and size was verified by Agilent. A 5nM stock was then made from equal pooled volumes of each library followed by qPCR. Sequencing was performed on the Illumina GA-IIx instrument following manufacturer's instructions. We obtained 50,085,061 reads from the 12 known bacterial genomes, 25,077,278 and 25,007,783 from the two respective ends. A small fraction, 266,146 reads (0.53 %), could not be assigned to any of the 12 species (Table 2).

Analytical processing and assembly of the 12 genomes were managed with the Genome Institute automated pipeline. It initially performs a BWA-style trim (Li and Durbin 2009) to a threshold of q10 on all input instrument data. Reads trimmed to less than 35 bp were discarded. The pipeline then runs Velvet (Zerbino and Birney 2008), which cycles through the 31–35 kmer range, optimizing for the kmer which produces the longest N50 contig length. The entire data set is publicly available through the NCBI Sequence Read Archive (SRA) under the accession numbers listed in Table 2.

BWA (Li and Durbin 2009) was used to align clean paired end reads to the 12 bacterial assemblies, ultimately placing 46,079,563 reads. Up to 5 mismatches were allowed per read, corresponding roughly to minimum 95 % identity. The distribution among the 12 organisms was then assessed using an in-house program called Refcov (Todd Wylie, unpublished) based on the generated alignments. Experimental coverage was then simulated by randomly picking reads from the total pool and assessing subsequent coverage for the target organism again using Refcov. For *E. faecium* (NCBI accession: AEBC00000000), we ran 50 simulations each of 3, 3.5, 4, and 4.5 million reads and for *S. artemidis* (NCBI accession: AECV01000000), which were based on 50 selections each of 9, 10, 11, 12, and 15 million reads. These numbers were based on species abundance within the community in Table 2.

Acknowledgments The authors wish to acknowledge funding sources for this work: National Human Genome Research Institute grants HG003079 and HG004968.

References

Ajay SS, Parker SCJ, Abaan HO, Fuentes-Fajardo KV, Margulies EH (2011) Accurate and comprehensive sequencing of personal genomes. Genome Res 21(9):1498–1505

Allen EE, Banfield JF (2005) Community genomics in microbial ecology and evolution. Nat Rev Microbiol 3(6):489–498

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F (2006) The marine viromes of four oceanic regions. PLoS Biol 4(11), article no. e368

Béjà O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, Spudich EN, DeLong EF (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289(5486):1902–1906

Beyer WH (1984) CRC standard mathematical tables. CRC Press, Boca Raton

Bouck J, Miller W, Gorrell JH, Muzny D, Gibbs RA (1998) Analysis of the quality and utility of random shotgun sequencing at low redundancies. Genome Res 8(10):1074–1084

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci 99(22):14250–14255

Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F (2003) Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 185(20):6220–6223

Chen K, Pachter L (2005) Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput Biol 1(2):106–112

Clarke L, Carbon J (1976) A colony bank containing synthetic Col El hybrid plasmids representative of the entire *E. coli* genome. Cell 9(1):91–99

Culley AI, Lang AS, Suttle CA (2006) Metagenomic analysis of coastal RNA virus communities. Science 312(5781):1795–1798

DeLong EF (2005) Microbial community genomics in the ocean. Nat Rev Microbiol 3(6):459–469

Dutilh BE, Huynen MA, Strous M (2009) Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly. Bioinformatics 25(21):2878–2881

Eisen JA (2007) Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. PLoS Biol 5(3), article no. e82

Feller W (1968) An introduction to probability theory and its applications. Wiley, New York

Fisher RA (1940) On the similarity of the distributions found for the test of significance in harmonic analysis and in Stevens' problem in geometrical probability. Ann Eugen 10:14–17

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, Fitzhugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu LI, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269(5223):496–512

Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE (2006) Metagenomic analysis of the human distal gut microbiome. Science 312(5778):1355–1359

Green ED (2001) Strategies for the systematic sequencing of complex genomes. Nat Rev Genet 2(8):573–583

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5(10):R245–R249

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10, article no. R32

Hess M, Sczyrba A, Egan RWKT, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, Mackie RI, Pennacchio LA, Tringe SG, Visel A, Woyke T, Wang Z, Rubin EM (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331(6016):463–467

Hooper SD, Dalevi D, Pati A, Mavromatis K, Ivanova NN, Kyrpides NC (2009) Estimating DNA coverage and abundance in metagenomes using a gamma approximation. Bioinformatics 26(3):295–301

Kowalchuk GA, Speksnijder AGCL, Zhang K, Goodman RM, van Veen JA (2007) Finding the needles in the metagenome haystack. Microb Ecol 53(3):475–485

Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P (2008) A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev 72(4):557–578

Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2(3):231–239

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760

Liles MR, Manske BF, Bintrim SB, Handelsman J, Goodman RM (2003) A census of rRNA genes and linked genomic sequences within a soil metagenomic library. Appl Environ Microbiol 69(5):2684–2691

Martín HG, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, Yeates C, He S, Salamov AA, Szeto E, Dalin E, Putnam NH, Shapiro HJ, Pangilinan JL, Rigoutsos I, Kyrpides NC, Blackall LL, McMahon KD, Hugenholtz P (2006) Metagenomic analysis of two enhanced biological phosphorus removal EBPR sludge communities. Nat Biotechnol 24(10):1263–1269

Nicholls H (2007) Sorcerer II: the search for microbial diversity roils the waters. PLoS Biol 5(3), article no. e74

Port E, Sun F, Martin D, Waterman MS (1995) Genomic mapping by end-characterized random clones: a mathematical analysis. Genomics 26(1):84–100

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Paslier DL, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD, Wang J (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464(7285):59–65

Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. Annu Rev Genet 38:525–552

Roach JC (1995) Random subcloning. Genome Res 5(5):464–473

Roach JC, Boysen C, Wang K, Hood L (1995) Pairwise end sequencing: a unified approach to genomic mapping and sequencing. Genomics 26(2):345–353

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC (2007) The Sorcerer II global ocean sampling expedition: Northwest Atlantic through eastern tropical Pacific. PLoS Biol 5(3), article no. e77

Schbath S (1997) Coverage processes in physical mapping by anchoring random clones. J Comput Biol 4(1):61–82

Schlüter A, Bekel T, Diaz NN, Dondrup M, Eichenlaub R, Gartemann KH, Krahn I, Krause L, Krömeke H, Kruse O, Mussgnug JH, Neuweger H, Niehaus K, Pühler A, Runte KJ, Szczepanowski R, Tauch A, Tilker A, Viehöver P, Goesmann A (2008) The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. J Biotechnol 136(1–2):77–90

Solomon H (1978) Geometric probability. Society for Industrial and Applied Mathematics, Philadelphia

Stanhope SA (2010) Occupancy modeling, maximum contig size probabilities and designing metagenomic experiments. PLoS ONE 5(7), article no. e11,652

Stevens WL (1939) Solution to a geometrical problem in probability. Ann Eugen 9:315–320

Thousand Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM (2005) Comparative metagenomics of microbial communities. Science 308(5721):554–557

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428(6978):37–43

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso sea. Science 304(5667):66–74

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. Science 315(5815):1126–1130

Vos M, Quince C, Pijl AS, DeHollander M, Kowalchuk GA (2011) A comparison of rpoB and 16S rRNA as markers in pyrosequencing studies of bacterial diversity. PLoS ONE 7(2), article no. e30,600

Wendl MC (2006a) A general coverage theory for shotgun DNA sequencing. J Comput Biol 13(6):1177–1196

Wendl MC (2006b) Occupancy modeling of coverage distribution for whole genome shotgun DNA sequencing. Bull Math Biol 68(1):179–196

Wendl MC (2008) Random covering of multiple one-dimensional domains with an application to DNA sequencing. SIAM J Appl Math 68(3):890–905

Wendl MC, Barbazuk WB (2005) Extension of Lander-Waterman theory for sequencing filtered DNA libraries. BMC Bioinform 6, article no. 245

Wendl MC, Waterston RH (2002) Generalized gap model for bacterial artificial chromosome clone fingerprint mapping and shotgun sequencing. Genome Res 12(12):1943–1949

Wendl MC, Wilson RK (2008) Aspects of coverage in medical DNA sequencing. BMC Bioinform 9, article no. 239

Wendl MC, Wilson RK (2009a) Statistical aspects of discerning indel-type structural variation via DNA sequence alignment. BMC Genom 10, article no. 359

Wendl MC, Wilson RK (2009b) The theory of discovering rare variants via DNA sequencing. BMC Genom 10, article no. 485

Wendl MC, Marra MA, Hillier LW, Chinwalla AT, Wilson RK, Waterston RH (2001) Theories and applications for sequencing randomly selected clones. Genome Res 11(2):274–280

Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. PLoS Comput Biol 6(2), article no. e1000,667

Xia LC, Cram JA, Chen T, Fuhrman JA, Sun F (2011) Accurate genome relative abundance estimation based on shotgun metagenomic reads. PLoS ONE 6(12), article no. e27,992

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18(5):821–829