Prediction of effective genome size in metagenomic samples

Raes, J.*, Korbel, J.O.*[1], Lercher, M.J., von Mering, C[2]. and Bork, P[†].
European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany.

* Contributed equally

[†] Correspondence to peer.bork@embl.de

[1] Present address: Molecular Biophysics & Biochemistry Department, Yale University, 266 Whitney Av., New Haven, CT.

[2] Present address: Institute of Molecular Biology, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

Submitted as a Method Article to Genome Biology

Abstract

We introduce a novel computational approach to predict effective genome size (EGS – a measure that includes multiple plasmid copies, inserted sequences and associated phages and viruses) from short sequencing reads of environmental genomics (or metagenomics) projects. We observe considerable EGS differences between environments and link this with ecological complexity as well as species composition (i.e. eukaryotes). For example, we estimate EGS in a complex, organism-dense farm soil sample at about 6.3 Mb whereas that of the bacteria therein is only 4.7 Mb; for bacteria in a nutrient-poor, organism-sparse ocean surface water sample, EGS is as low as 1.6 Mb. The method also permits evaluation of completion status and assembly bias in single-genome sequencing projects.

Background

Because of its direct link with the functional repertoire, microbial genome size is an important ecological parameter, which is believed to be closely coupled to the functional complexity and environmental niche of an organism[1-4]. For over three decades, numerous studies have provided estimates of average microbial genome size for various environments, but results vary greatly. Estimates of the average DNA content per cell (converted to megabases (Mb) for comparison) range from 1.5 to 8.0 Mb for soil and from 1.5 to 9.5 Mb for aquatic environments (see [5, 6] for an overview of estimates). However, the diversity of techniques and parameters used (e.g. sample filtering, DNA staining and cell counting) greatly hampers the interpretation and comparison of these results. All currently used methods also have several important drawbacks: For instance, they have difficulties discriminating between the different ploidy levels of cells[7-9], so any technique measuring DNA content does not necessarily measure genome size. In addition, DNA binding of stains used in the majority of these studies (e.g. DAPI) is not always specific, and important biasing factors (GC content, permeability, salinity, influence of debris etc.) have hardly been compensated for [7, 10-12]. Finally, some estimates have been obtained in studies using cultured isolates only (e.g.[8, 13]), which does not reflect the actual environmental species composition.

Because of all these difficulties, the average genome size of microorganisms living in particular environments is still uncertain, and the influence of environments on genome size remains a matter of speculation. Recently however, several studies have provided unprecedented insights into the microbial DNA content of complete ecosystems using massive random shotgun sequencing of environmental samples[14-16]. Through comparative metagenomics, various aspects of ecological complexity can now be studied[15, 17-20]. Here, we use these data to study the relationship between environment and microbial genome size.

Results and Discussion

The concept of effective genome size (EGS). An assembled, sequenced genome is a non-redundant representation of the naturally occurring amount of base pairs that a cell supports. It neither reflects actual copy number of inserted elements and plasmids nor the amount of associated phages and viruses. However, the total amount of DNA replicated per cell division is what determines the metabolic cost and

what has to be balanced against the full functional spectrum of genes available to a given organism. To estimate this latter, ecologically more meaningful measure of genome size (subsequently referred to as “Effective Genome Size (EGS)”), we have developed a novel computational approach to predict EGS directly from raw shotgun sequencing data, thereby avoiding experimental biases such as mentioned above. When applied to metagenomics data, our method measures the average EGS of organisms living in the sampled environment.

Deriving a method for EGS prediction. In brief, we use a set of marker genes that typically occur only once per genome, to extrapolate the average genome size from the density of these genes found in the total set of sequence reads. Even in complex metagenomics data, the total number of marker genes should be proportional to the number of genome equivalents (i.e. individuals) present, and the marker gene density (i.e. number of marker genes divided by the number of sequenced base pairs) should thus be inversely correlated to the average size of the genomes in the sample. This approach would also be able to normalize, unlike previous DNA measurements, for intermittent episodes of polyploidy (for example in the case of fast-growing microbes that may replicate multiple concurrent copies of their genomes); in these situations our marker genes themselves are present in multiple copies and their density does not change.

Previous studies have shown that genes involved in translation, ribosome structure and biogenesis generally show a low number of duplicates per genome and their number does not expand much with genome size[2, 21-24], and thus would constitute suitable marker genes to estimate the number of individuals in a sample, irrespective of their genome size. However, when applying orthologous group (OG) categories for identifying such genes, we still observed a slight positive correlation between genome size and the number of translation-related genes (Figure 1a). Therefore, we selected a universally occurring set of marker genes (largely overlapping with the ones used in [25]) that only very rarely occur as duplicates, such that the total marker gene count remains constant with increasing genome size (Figure 1a; Materials and methods). The selected marker genes (most of them, but not all, involved in translation) can be considered to be both essential for cellular life and very ancient; they evolve at a slow rate and are members of basal cellular processes, showing little variation across phyla. To identify the relationship between the density (count per megabase (Mb)) of the combined set of these selected markers and genome size, we calibrated our method on simulated shotgun reads from 154 completely sequenced bacterial and archaeal genomes (see Materials and

methods). Indeed, although the relationship between the true number of occurrences of each marker gene and the number of individuals (and thus the relationship between their density and average genome size) is simple, the relationship between the number of BLAST-observable instances of a combined set of marker genes in incomplete environmental shotgun data and the number of individuals is not so straightforward. In addition, the length of sequencing reads can seriously influence the likelihood of successfully detecting a marker gene using BLAST (data not shown). Therefore, we performed a three-dimensional calibration relating marker gene density x, read length L and genome size (Figure 1b; Materials and methods), resulting in the relationship

$$EGS = \frac{a + b \times L^{-c}}{x}$$ (with *a*=21.2, *b*=4230, *c*=0.733)

to predict genome size (in Mb) from the two other parameters. This formula indeed shows the inverse relationship between EGS and marker gene density, however corrected by a read-length-dependent factor following a power law. An analysis of sources and magnitudes of errors in predicted genome sizes showed that inaccuracies stem mostly from finite sequencing depth, from uncertainties associated with identification of marker gene sequences using BLAST, and from residual biological variation in genomic marker gene count (see additional file 1). The error contribution from finite depth is small when more than ~4× the average genome size was sequenced (additional file 7; this is the case in the environmental shotgun-sequenced samples currently available). On the whole, the median unsigned prediction error on the simulated shotgun data was 5.3% (standard deviation (SD) 8.7%), largely independent of genome size and read length (additional file 7).

As marker genes are equally present in all species, our method should work well on complex, mixed samples (a theoretical proof can be found in additional file 1). We further support this by performing simulations mixing species and readlengths. For mixtures, the contribution of each species should weighted by the fraction (in numbers of genomes) it takes up in the sample. E.g., for a mixture of 2 species of 4Mb (90% of genomes) and 12Mb (10% of genomes), we should get EGS=0.9*4Mb + 0.1*12Mb. Our simulations show that this is indeed this case (see Materials and methods, additional file 1 and additional file 8).

Method validation on real shotgun data and detection of sequencing artifacts. To confirm that simulated data represent a valid approximation of real sequencing reads, we measured the prediction error on publicly available sets of microbial whole-genome raw shotgun sequencing reads (Figure 2; Note that only 32 such datasets were present in NCBI's trace archives at the time of analysis - a larger training set of whole-genome sequencing reads should allow to further improve the method). The analysis of this prediction error on 'real' reads showed a systematic shift of predicted versus known genome sizes by 15.9%, most likely reflecting unequal representation of certain genomic regions in sequencing libraries ('cloning bias' due to library preparation, toxicity, restriction site biases etc). After correction for this bias (which leads to an adjustment of the values for a and b in formula 1, see Materials and methods), the median error on real shotgun data is as low as ~7.8% (standard deviation 14.4%; Figure 2; Materials and methods). The two outliers with larger errors can be linked to anomalies in the deposited reads, caused by contamination (*Wolbachia*) or collapsing of repeated insert sequences (*Dehalococcoides*) (Figure 2). The latter example illustrates that the EGS predicted by our method is not just reflecting the principal, non-redundant chromosome, but indeed considers the actual copy number of inserted elements and plasmids. The importance of this additional genomic repertoire ('mobilome'; [26]) is not to be underestimated. For example, a 20% variation in chromosome length was described for different isolates of Escherichia coli[27].

Since these cases indicate that our approach can identify assembly artifacts or incomplete cloning material, we applied the method to unfinished genome sequencing projects that have not been updated recently, and might have had a problematic project history (additional file 5). Our method seems to yield plausible results on the whole, as in the majority of cases our predictions stay within the error ranges of previous estimates based on e.g. pulse-field gel electrophoresis. In a few cases, our method reveals a larger genome size than was initially anticipated, which might explain problems in achieving sufficient sequence coverage (e.g. for the *Chloroflexus aurantiacus* genome); here our predictions can provide guidance as to the amount of sequence data needed for genome completion. This is expected to be particularly useful in sequencing projects that utilize the recently developed low-cost sequencing techniques[28, 29], which produce short reads and are thus more difficult to assemble.

Characterizing prokaryotic genome sizes as comparable subsets of samples. In complex environmental samples, the proportion of eukaryotic DNA present may have a large impact on EGS measurements. Because of their disproportionally bigger genome size, even a minor fraction of eukaryotes in the sample could inflate EGS. Thus, the relationship between organism complexity, EGS and environment becomes difficult to interpret, even though eukaryotes are a valid part of an environmental sample and a higher proportion of eukaryotes should go hand in hand with a higher degree of complexity.

In order to better understand these effects, we adapted the method to measure EGS specifically for only the bacterial or archaeal fraction of the sample (see Materials and methods). To do this, we divide the number of hits to marker genes of bacterial/archaeal origin (as determined by the best BLAST hit in the STRING database [30]) by the number of hits to any bacterial/archaeal gene. Calibration of this domain-specific marker gene density on known genomes shows that, as expected, it scales inversely linear to genome size. The error is readlength-independent and is not influenced by genome size. Median prediction error for real bacterial shotgun reads is 8.0% (standard deviation 14.6%) for the bacteria-specific measure (see additional files 1, 10 and 11). The archaea-specific measure is associated with a higher prediction error (see Materials and methods) due to the small number of genomes available for calibration and will improve when more genomes become available for this domain.

Measuring EGS of real environments using metagenomics data. Having established methods for measuring EGS, we applied them to twelve publicly available environmental sequence data sets: five communities that were sampled without particle size filtering (a soil community, an acidophilic underground biofilm (‘acid mine drainage’) and three deep sea whale carcass scavenger communities (‘whale falls’)[14, 15]), and also seven Sargasso Sea water samples of different cell size fractions (sea water was pumped through two consecutive filters: a first ‘prefilter’ to remove larger organisms and debris and a second ‘collection filter’ for sampling. Therefore, each Sargasso Sea sample should be interpreted with the corresponding organismal size range in mind [16, 31]).

Measurements of the EGS in the samples shows that the Soil sample has the largest EGS, together with two of the Sargasso Sea samples of large cell size fractions, while the other Sargasso Sea samples have very low EGS estimates (Table 1). In order to test whether these values reflect functional complexity of the micro-organisms in the sample or only reflect phylogenetic composition of the samples we

applied our method of measuring bacterial/archaeal EGS. The comparison of the latter with the general EGS measure shows that in a number of the samples analyzed here, the presence of eukaryotes indeed has an important effect on the estimated EGS. For example, the value for soil is reduced by ~25% when eukaryotes are excluded. However, the most drastic differences are seen in the Sargasso samples 5 and 6 (the two largest cell size fractions), which are reduced by more than 75 and 50%, respectively, now causing all Sargasso size fractions to become statistically indistinguishable and converge to an average bacteria-specific EGS of 1.6 Mb (sample 1 excluded – see below; Table 1; Figure 3). In addition, the fact that the samples originate from four different sampling sites in the Sargasso Sea (stations 11/13, 3, 13 and S), and were taken at different time points [16] suggests that micro-environmental differences are not influencing genome size or, alternatively, that there are very little differences in the bacterial populations of these sites. Indeed, water currents can allow for a rapid and continuous homogenization of communities and previous studies showed very little variation in GC content between samples with similar filtering treatment [32], arguing for the latter explanation.

The only outlier in these estimates is that of sample 1 (station 13/11), which has a bacterial-specific EGS of 3.4Mb, i.e. about twice as large as all other samples (Table 1). However, for this sample, contamination with *Shewanella* and *Burkholderia*, two terrestrial species, has been proposed [19, 33], which could explain the EGS differences.

The acid mine drainage dataset derived from a biofilm at pH 0.83 [14] provides a first large-scale glimpse of genomic properties of free-living extremophiles. When only considering the bacterial EGS for the acid mine drainage sample, our results show an increase of 50% compared to the overall measure (3.2Mb vs 2.1Mb; Table 1; Figure 3). This might be explained by the presence of small genome archaea (*Ferroplasma acidarmanus fer1 and Ferroplasma type II*), which (as estimated by BLAST-based phylotyping, data not shown) seem to dominate the deposited reads. Indeed, the calculated EGS of archaea in this sample (1.8Mb) is perfectly in line with the genome sizes of the two assembled species. Intriguingly, the bacterial-specific EGS in the acid mine drainage sample (3.2Mb) is more than twice as large as the average parasite/symbiont genome size[34] and is thus in conflict with the proposed theory that genome evolution patterns of free-living extremophiles are similar to those of intracellular pathogens or symbionts[35].

EGS correlates with environmental complexity. Although the Sargasso Sea samples are separated into size fractions, the convergence of all samples to a narrow common average bacterial EGS of ~1.6Mb suggests that this value gives a correct general EGS of bacteria living in this environment. When considering only the bacterial fraction, we can hence compare these samples with the other (unfiltered) samples in order to investigate the influence of environmental factors on genome size. Our results show that soil bacterial EGS is significantly larger than that of the pooled non-contaminated Sargasso samples ($p=5.5\times10^{-6}$ after correction for multiple comparisons) and marginally significantly larger than AMD ($p=0.04$) and the pooled whale falls ($p=0.053$). Although the genomes of fully sequenced soil-dwellers were already noted to have a tendency to be larger than others [1, 21], we provide here for the first time conclusive evidence for this hypothesis based on an unbiased sampling of thousands of soil bacteria. Soil is a very challenging environment, because of 1) the high organism density leading to strong competition for nutrients as well as complex communication and cooperation strategies, and 2) the highly variable living conditions (e.g. seasons, weather)[36]. Therefore, a broad functional repertoire is needed in order to survive competition and adapt to ever-changing conditions. Together with the fact that only about 50% of the predicted soil genes have a match in current protein databases, our results imply a wide variety of novel functions and processes in soil, potentially including biotechnologically relevant ones such as defense (antibiotics), biosynthesis and biodegradation.

The EGS of the Sargasso Sea samples is significantly smaller than that of all other samples. The explanation could lie in the lower organism density in Sargasso surface waters (about 3 orders of magnitude smaller than soil[16, 37]), which would allow organisms to shed the functional repertoire presumably needed for survival in densely populated, substrate-bound habitats or, alternatively, 'genome streamlining' to optimize replication under limiting nutrient resources, as was seen for *Pelagibacter ubique[34]*, a member of the SAR11 clade, and *Prochlorococcus*[38-40], which dominate oceanic surface waters. Our estimated EGS is consistent with the genome sizes of these organisms (1.3Mb for Pelagibacter and 1.6Mb for the Prochlorococcus high-light-adapted ecotype[34, 39]) and the previously reported low GC content of Sargasso sequences[32], as GC content scales with genome size[1].
As expected, AMD and whalefall EGS estimates are in between both extremes, in accordance with their densely populated substrate-bound lifestyle, but under relatively stable environmental conditions[14, 15].

Conclusions

Using a novel computational approach, we have shown how effective genome size can be directly determined from raw sequencing reads, either for single species or entire organismal communities. EGS can be reliably estimated for complex environments as a whole, but also for bacterial or archaeal subcommunities in a sample. Applying this method to diverse environmental data sets, we could establish a relationship between genome size and environment, suggesting a clear correlation between environmental complexity and the diversity of the cellular repertoire that is required to cope with various external challenges. As EGS directly reflects the functional diversity of a community, it will not only serve as a useful ecological parameter but might also play a role in the search for novel biological activities.

Furthermore, an accurate estimate of average genome sizes for different environments is paramount for other approaches to understand the totality of ecosystem composition and functioning. For example, widely-used techniques such as DNA reassociation kinetics help to understand ecosystem species composition and biodiversity as a whole[37, 41], but they require knowledge of the average genome size to translate genetic diversity into species diversity. Even environmental cell counts (used for various applications) heavily depend on the selection of a reference species with a genome size comparable to the sampled ecosystem average[42]. For the analysis of metagenomics data, the average genome size allows to calculate early on in a pipeline the amount of sequencing necessary for completion of the most dominant species[15] and is needed for the deduction of community structures from assembly data[43]. Currently, our method is limited to predicting only the average EGS, without describing the distribution of genome sizes within the sample. However, improvements in phylogenetic separation of metagenomic sequences should allow the adaptation of our method to predict genome size distributions in the future.

The applications above illustrate the importance of this parameter (together with the functional and phylogenetic characterization of samples) in the process of understanding ecosystem properties from metagenomics data. The Effective Genome Size, as predicted here, is thus applicable to a broad range of questions and techniques ranging from genomics via population genetics to ecology.

Materials and methods

Detection of marker genes. We used a set of 35 OGs that are widely conserved (present in most species), rarely occur as duplicate genes in known genomes, are not subject to horizontal gene transfer, and do not scale with genome size as marker genes (set largely overlapping with the one used in [25]; Figure 1; additional file 2). Marker gene counts were carried out using an approach similar to the one described previously[15], based on comparisons to known proteins. In brief, DNA sequence reads (or alternatively, randomly generated genome fragments) were searched against the extended database of proteins assigned to OGs in the latest STRING release (6.3)[30], using BLASTX[44], and an OG was called present when a hit matching one of its proteins occurred (with a BLAST score of at least 60 bits). Note that this procedure is largely independent on varying annotation qualities across genomes, and avoids biases owing to lower gene prediction quality on short sequences[17], such as the reads used in this study. In order to avoid potential biases introduced by any uneven phylogenetic representation within the reference set of known proteins, all BLASTX matches exceeding an overall protein identity of 50% were discarded. This latter step is needed to avoid artefacts introduced by the occasional organism in the sample that happens to be closely related to a known organism in the BLAST database. For such an organism, even marker genes contained only partially on a read can be detectable by BLAST due to their high sequence identity to known genes. In contrast, most environmental genes have low sequence identity to known genes, so fragmented marker genes often escape detection. Thus, without the above threshold, marker gene counts would be higher for well-known organism, biasing the results (note that the threshold does not select *against* well-known organisms either – their genes will still generate hits with other organisms in the database, with identities below 50%, and will thus be counted like any other environmental marker gene). Query sequences were allowed to map to several OGs, provided these overlapped by no more than 50% of the shortest assignment. BLASTX was run using the BLOSUM62 matrix, and low-complexity filtering was enabled. Marker gene density x was then defined as the number of matches to reference genes, divided by the total number of Mb surveyed.

Calibrating marker gene density with genome size, and genome size prediction. To determine the relationship between the occurrence of marker genes and genome size we used fully sequenced genomes for calibration. We simulated

the widely used whole genome shotgun (WGS) sequencing process by randomly extracting 'reads' of variable readlength from 154 previously completely sequenced bacterial and archaeal genomes (EBI Genome Reviews release 17). In total, 50 genomes were randomly chosen per readlength bin (300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200bp), and reads were sampled until 3x coverage was achieved (see additional file 3 for a list of genomes). We did not distinguish between plasmid and chromosomal DNA; i.e. each DNA fragment of a completely sequenced genome was equally likely to be considered.

We determined the occurrences of marker genes among these 'reads' as described above. On these counts, we based a three-dimensional calibration, relating known genome size to the parameters read length and marker gene density. Because the total number of marker genes per genome does not vary with genome size, we expect that genome size increases proportionally to the inverse marker gene density 1/*x* at any given read length *L*: EGS = *c*(*L*) / x, where *c*(*L*) is a read-length dependent calibration factor. The exact form of *c*(*L*) is determined not only by the read-length dependence of the probability of sequencing a portion of a marker gene, but also by the probability of identifying a read as a marker gene. Because the latter depends on sequence features of individual marker genes, it is not easily possible to specify the analytical form of *c*(*L*) *a priori*. Based on manual comparison of a variety of possible functional forms, we found that *c*(*L*) is well approximated by a power law, *c*(*L*) = a + b L^{-c}. This is indeed the best (R^2=0.97) 'simple' 3-parameter formula relating genome size to marker gene density and read length, as confirmed using the TableCurve3D v.4.0 package; a 'simple' equation is defined here as a three parameter equation consisting of a constant and two coefficients which multiply a function of x or *L*. This resulted in the prediction formula

$$EGS = \frac{a + b \times L^{-c}}{x} \quad [1]$$

We estimated the parameters of this formula with a non-linear least-squares fit, as implemented with the *nls* function in the R programming environment[45]. First, we randomly selected half of the species in the simulated data. Their complete sets of simulation results (marker gene densities *x* at specified read lengths *L*), together with the known genome sizes *z*, were used as calibration data (the remaining data was later used for detailed error estimation, see additional file 1 and additional file 7). Parameters *a, b and c* were chosen such as to minimize the weighted sum of squares $((a + b \times L^{-c})/x - z)^2/z$. This led to the parameter estimates *a*=21.2, *b*=4230, *c*=0.733.

Because Eq.(1) is linear in the inverse marker gene density x^{-1}, it can be directly applied to mixtures of genomes, which was supported by our simulations. In the case of species mixtures, the estimated mean *EGS* is the number of megabases per genome present in the sample; i.e., effective genome sizes of different species are weighted by their genome count, not by their contribution to the number of sequenced base pairs. As Eq.(1) is further approximately linear in the inverse read length L^{-1}, it can also be applied to sequence datasets with mixed read lengths (For a full discussion and simulation of EGS prediction in mixtures, see additional file 1 and additional file 8).

So far, calibration was based on simulated reads from fully sequenced genomes, ignoring potential biases introduced, e.g., by cloning. To test our predictions on actual sequencing data, we downloaded and analyzed shotgun data of completed genomes from the NCBI and Ensembl trace repositories (ftp://ftp.ncbi.nih.gov/pub/TraceDB/; ftp://ftp.ensembl.org/traces/), excluding those projects where sequencing coverage was low (<1.5×; additional file 4). In order to ensure consistency, we applied a uniform quality clipping method to all 32 datasets, rather than using the provided coordinates of each sequencing centre (phred quality cut-off score of 15, using a perl script kindly provided by Jarrod Chapman (JGI); clipped reads with less then 100 nucleotides remaining were discarded). We found that Eq.(1) overestimates genome size on average by 15.9%. This reflects a strong bias against the marker genes, likely resulting from the fact that these 35 OGs were chosen for their strongly conserved single-copy distribution across genomes, and hence introduction of additional genes into the cloning vector is often lethal. Removal of this bias by an additional scaling factor (excluding the outliers *Wolbachia* and *Dehalococcoides ethenogenes 195* as discussed in the main text) results in the new parameter values a = 18.26 and b = 3650 for Eq.(1).

Species mixture simulations. In order to generate species mixtures, we first randomly picked 60 out of the entire list of completely sequenced cellular genomes, and simulated WGS sequencing for all 60 genomes using readlength bins from 600 to 900 as described in the Materials and methods. We then generated 1000 simulated metagenomes, by repeating the following procedure 1000 times:

(1.) A random number i of species were picked, with $1<i\leq 60$.

(2.) For each of the *i* species its contributing nucleotides n_i was randomized, using the following condition regarding the total number of nucleotides in the metagenome:

[2] $$\sum_i \frac{n_i}{n_{total}} = 1$$.

(3.) A readlength was randomly picked from the available 600, 700, 800, or 900 bp.

(4.) A 'large' metagenome was generated (total sequence ≈40 *Escherichia coli* K12 sized genomes) by randomly extracting reads from each of the i contributing species, using the readlength randomly picked in step (3.)

(5.) The theoretical genome size *T* of the simulated metagenome is calculated from the actual contributions c_i and from the genome sizes s_i of the given species, using the following equations:

[3] $$T = \sum_i c_i s_i = \sum_i \frac{n_i}{s_i u} s_i = \sum_i \frac{n_i}{u}$$, with $$c_i = \frac{n_i}{s_i u}$$, and $$u = \sum_i \frac{n_i}{s_i}$$.

(6.) The effective genome size for the simulated metagenome is predicted from the randomly extracted reads as described in the main text (formula for estimating the EGS).

(7.) Errors *e* are calculated using:

[4] $$e = \frac{EGS - T}{T}$$.

Results are given in additional file 8.

Mixed read lengths simulations. We further generated datasets with mixed read lengths. Namely, we applied the following procedure 1000 times:

(1.) A species S was randomly picked from the pool of completely sequenced genomes.

(2.) A random number j of readlengths with 1<j≤4 were picked.

(3.) A 'large' metagenome was randomly generated (total sequence ≈40 *Escherichia coli* K12 sized genomes), consisting only of species *S* (with genome size *s*).

(4.) Genome size is predicted as described in the main text (formula for estimating the EGS).

(5.) Errors *e* are calculated using:

[5] $$e = \frac{EGS - s}{s}$$.

Results are given in additional file 8.

EGS estimation restricted to bacteria. For EGS estimation restricted to only the bacteria in the sample, we calculate a domain-specific marker gene density x_{bacteria}, by dividing the number of hits to marker gene OGs (n) by the number of hits to any OG (n_{total}), with the limitation that an OG mapping is only counted if the best BLAST hit of that read region to STRING is a bacterial protein. This way, only reads of bacterial origin are considered. Because marker gene density is now estimated per read rather than per base pair, this measure requires a new calibration analogous to the one described above. Again, we found Eq.(1) to be the best fitting simple formula for simulated data (R^2=0.93), with parameter estimates *a*=0.0389, *b*=0.81, and *c*=0.78 from a weighted fit to a training dataset consisting of data for half of the included genomes.

Performance was tested on the remaining data as for the general measure (See additional file 1 and additional files 10 and 11). Comparison to real reads as above revealed an average bias of 5.2%, which lead to adjusted parameter values of *a*=0.0370 and *b*=0.770. After correction, we found an unsigned median error of 8.0% (standard deviation 14.6%). An analogous procedure was also performed to estimate *EGS* restricted to archaeal genomes in the sample. However, currently only very few archaea are fully sequenced, and hence there was insufficient data for a full fit and error estimation. To allow at least an approximate analysis of the archaeal fraction in the AMD sequences (see below), we obtained a rough measure by scaling Eq.(1) with the bacterial parameter set to fit simulated data from the available fully sequenced archaea. This resulted in the parameter estimates *a*=0.045, *b*=2.91, and *c*=0.78 for archaea (R^2=0.87). Median unsigned error on the simulated data was 14.7%, and standard deviation was 15.8%.

Before comparing bacterial EGS estimates across environments, we first confirmed that there were no significant differences among the three whalefall samples and among the 6 Sargasso Sea samples (excluding sample 1), by calculating a *z*-score and *P*-value (Additional file 1; all pairwise raw P>0.05). We then pooled all whalefall samples, and separately all Sargasso Sea samples, to reduce the total number of comparisons to be made. Statistical significance of differences in EGS was then

estimated by calculating *z*-score and *P*-value in all remaining pairwise comparisons, and correcting the resulting raw *P*-values for multiple comparisons[46].

Environmental sequencing data. The same data was used as in [32], with the exception of the Sargasso Sea data, where now all samples were used. Reads were trimmed as described above. Additional file 6 gives an overview of sequence data after trimming.

Scripting, statistical analyses and parameter estimation was performed using the R environment for statistical computing[45] and perl[47].

Additional data files

Additional file 1: Supplemental Materials and methods

Additional file 2: List of orthologous group markers

Additional file 3: Randomly selected genomes and read lengths used for calibration

Additional file 4: Shotgun sequencing projects used to estimate cloning bias.

Additional file 5: Estimated genome sizes for available unfinished genomic sequencing project datasets.

Additional file 6: Data statistics for available environmental shotgun sequencing datasets (measured after quality clipping)

Additional file 7: Error distribution for EGS prediction on simulated reads.

Additional file 8: EGS predictions work well when analyzing mixtures of different species or read lengths.

Additional file 9: Error distribution for EGS prediction on real reads.

Additional file 10: Error distribution for EGS prediction on simulated reads, using the bacterial-specific version of the prediction formula.

Additional file 11: Error distribution for EGS prediction on real reads, using the bacterial-specific version of the prediction formula.

Acknowledgments

The authors would like to thank Sean Hooper, Lars Juhl Jensen and other members of the Bork group, as well as Shannon McWeeney for stimulating discussions, Susannah Green Tringe and Jarrod Chapman at the Joint Genome Institute for sharing their expertise in quality trimming of sequence reads, and Sam Pitluck (JGI) as well as the staff of the NCBI trace archive for their assistance in acquiring additional shotgun data. This work was supported by the EU 6th Framework Programme (Contract Nº: LSHG-CT-2004-503567).

Table 1: Predicted EGS on environmental samples (in Mb±SD)

Sample	EGS (complete sample)	EGS (only bacteria)
AMD	2.11 ± 0.30	3.16 ± 0.46
Soil	6.29 ± 0.91	4.74 ± 0.69
Whalefall 1 ('agzo')	3.42 ± 0.49	3.39 ± 0.49
Whalefall 2 ('ahaa')	4.50 ± 0.65	4.02 ± 0.59
Whalefall 3 ('ahai')	3.35 ± 0.48	3.24 ± 0.47
Sargasso sample 1	3.25 ± 0.47	3.39 ± 0.49
Sargasso sample 2	1.48 ± 0.21	1.46 ± 0.21
Sargasso sample 3	1.68 ± 0.24	1.57 ± 0.23
Sargasso sample 4	1.59 ± 0.23	1.50 ± 0.22
Sargasso sample 5	6.20 ± 0.89	1.71 ± 0.25
Sargasso sample 6	4.04 ± 0.58	1.94 ± 0.28
Sargasso sample 7	1.32 ± 0.19	1.35 ± 0.20

Figure legends

Figure 1: Predicting effective genome size from marker gene density

(a) Gene counts for various functional classes[48] and their relationship with genome size. While counts of genes belonging to the categories T (Signal transduction mechanisms), K (Transcription) and J (translation, ribosome structure and biogenesis) scale (to a greater or lesser extent) with genome size, the set of 35 universal, single-copy genes used in this study, does not.

(b) Calibration plane used to identify the relationship between marker gene density, read length and genome size. The calibration was based on a simulated shotgun dataset of randomly extracted 'reads' from the sequenced genomes (see Materials and methods), as insufficient raw shotgun sequence data is currently available in the trace archives to allow a robust calibration based on 'real' data. Circles represent shotgun datasets. Circle fill color indicates the goodness of fit to the plane (blue: <1SD, green: <2SD, yellow: <3SD, red: >3SD). Circle border indicates position relative to plane (blue: above, red: below).

Figure 2: **Prediction error and identification of sequencing artifacts**

Distribution of the prediction error ((predicted – known genome size) / known genome size) of 32 complete genome shotgun datasets downloaded from the NCBI's trace archive (see additional file 4 for a list). The majority of predictions have an error estimate <20%, with a median value of ~9%.

There are, however, two exceptions in which the error is significantly larger. The first is the *Wolbachia* endosymbiont of *Drosophila melanogaster*. The marker OG density in the simulated reads is considerably higher than in the real shotgun data, leading to a 70% difference in predicted genome size. After further investigation of the raw reads, we noticed that this difference was caused by an important contamination of the dataset by reads originating from the organisms host, *Drosophila*, that were filtered out during the assembly of the genome, but that are still present in the shotgun data available at the trace archive. The second exception is the genome of the PCE-dechlorination bacterium *Dehalococcoides ethenogenes*. Also here, the marker OG density in the shotgun data is lower than in the simulated dataset. Mapping of the publicly available reads to the genome sequence showed a peak of read density in a region that was identified to be an integrated element that is believed to exist in variable copy numbers in different individuals but was only included once in the published genome sequence [49].

Figure 3: Predicted effective genome sizes for environments. (a) Comparison of predicted EGS for total samples *vs.* the bacterial fraction. AMD: Acid Mine Drainage; WF; Whale fall deep sea samples.; S; Sargasso sea samples. Error bars indicate standard deviation for total (horizontal) and bacteria-specific (vertical) estimate.
(b) Overview of cell size in the different Sargasso sea samples due to filtering during sampling.

References

1. Bentley SD, Parkhill J: **Comparative genomic structure of prokaryotes.** *Annu Rev Genet* 2004, **38:**771-792.
2. van Nimwegen E: **Scaling laws in the functional content of genomes.** *Trends Genet* 2003, **19:**479-484.
3. Mira A, Ochman H, Moran NA: **Deletional bias and the evolution of bacterial genomes.** *Trends Genet* 2001, **17:**589-596.
4. Gregory TR, DeSalle R: **Comparative genomics in prokaryotes.** In *The Evolution of the Genome.* Edited by Gregory TR. San Diego: Elsevier; 2005: 585-675
5. Loferer-Krossbacher M, Witzel K-P, Psenner R: **DNA content of aquatic bacteria measured by densitometric image analysis.** *Arch Hydrobiol Spec Issues Advanc Limnol* 1999, **54:**185-198.
6. Torsvik V: **Total bacterial diversity in soil and sediment communities - a review.** *J Industr Microb* 1996, **17:**170-178.
7. Button DK, Robertson BR: **Determination of DNA content of aquatic bacteria by flow cytometry.** *Appl Environ Microbiol* 2001, **67:**1636-1645.
8. Christensen H, Bakken LR, Olsen RA: **Soil bacterial DNA and biovolume profiles measured by flow-cytometry.** *FEMS microbiology ecology* 1993, **102:**129-140.
9. Bakken LR, Olsen RA: **DNA-content of soil bacteria of different cell size.** *Soil Biol Biochem* 1989, **21:**789-793.
10. Weinbauer MG, Beckmann C, Hofle MG: **Utility of green fluorescent nucleic acid dyes and aluminum oxide membrane filters for rapid epifluorescence enumeration of soil and sediment bacteria.** *Appl Environ Microbiol* 1998, **64:**5000-5003.
11. Kepner RL, Jr., Pratt JR: **Use of fluorochromes for direct enumeration of total bacteria in environmental samples: past and present.** *Microbiol Rev* 1994, **58:**603-615.
12. Zweifel UL: **Total counts of marine bacteria include a large fraction of non-nucleoid-containing bacteria (ghosts).** *Appl Environ Microbiol* 1995, **61:**2180-2185.
13. Torsvik V, Salte K, Sorheim R, Goksoyr J: **Comparison of phenotypic diversity and DNA heterogeneity in a population of soil bacteria.** *Appl Environ Microbiol* 1990, **56:**776-781.
14. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment.** *Nature* 2004, **428:**37-43.
15. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, et al: **Comparative metagenomics of microbial communities.** *Science* 2005, **308:**554-557.
16. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, et al: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304:**66-74.
17. Tringe SG, Rubin EM: **Metagenomics: DNA sequencing of environmental samples.** *Nat rev genet* 2005, **6:**805-814.
18. Schloss PD, Handelsman J: **Metagenomics for studying unculturable microorganisms: cutting the Gordian knot.** *Genome Biol* 2005, **6:**229.
19. DeLong EF: **Microbial community genomics in the ocean.** *Nat Rev Microbiol* 2005, **3:**459-469.
20. Foerstner KU, von Mering C, Bork P: **Comparative analysis of environmental sequences: potential and challenges.** *Philos Trans R Soc Lond B Biol Sci* 2006, **361:**519-523.
21. Konstantinidis KT, Tiedje JM: **Trends between gene content and genome size in prokaryotic species with larger genomes.** *Proc Natl Acad Sci U S A* 2004, **101:**3160-3165.
22. Ranea JA, Buchan DW, Thornton JM, Orengo CA: **Evolution of protein superfamilies and bacterial genome size.** *J Mol Biol* 2004, **336:**871-887.
23. Taylor JS, Raes J: **Duplication and divergence: the evolution of new genes and old ideas.** *Annu Rev Genet* 2004, **38:**615-643.

24. Gevers D, Vandepoele K, Simillon C, Van de Peer Y: **Gene duplication and biased functional retention of paralogs in bacterial genomes.** *Trends Microbiol* 2004, **12:**148-154.
25. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life.** *Science* 2006, **311:**1283-1287.
26. Ou HY, Smith R, Lucchini S, Hinton J, Chaudhuri RR, Pallen M, Barer MR, Rajakumar K: **ArrayOme: a program for estimating the sizes of microarray-visualized bacterial genomes.** *Nucleic Acids Res* 2005, **33:**e3.
27. Bergthorsson U, Ochman H: **Distribution of chromosome length variation in natural isolates of Escherichia coli.** *Mol Biol Evol* 1998, **15:**6-16.
28. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate multiplex polony sequencing of an evolved bacterial genome.** *Science* 2005, **309:**1728-1732.
29. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376-380.
30. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P: **STRING: known and predicted protein-protein associations, integrated and transferred across organisms.** *Nucleic Acids Res* 2005, **33:**D433-437.
31. Remington KA, Heidelberg K, Venter JC: **Taking metagenomic studies in context.** *Trends Microbiol* 2005, **13:**404.
32. Foerstner KU, von Mering C, Hooper SD, Bork P: **Environments shape the nucleotide composition of genomes.** *EMBO Rep* 2005, **6:**1208-1213.
33. Falkowski PG, de Vargas C: **Genomics and evolution. Shotgun sequencing in the sea: a blast from the past?** *Science* 2004, **304:**58-60.
34. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, et al: **Genome streamlining in a cosmopolitan oceanic bacterium.** *Science* 2005, **309:**1242-1245.
35. Cases I, de Lorenzo V, Ouzounis CA: **Transcription regulation and environmental adaptation in bacteria.** *Trends Microbiol* 2003, **11:**248-253.
36. Daniel R: **The metagenomics of soil.** *Nat Rev Microbiol* 2005, **3:**470-478.
37. Torsvik V, Ovreas L, Thingstad TF: **Prokaryotic diversity--magnitude, dynamics, and controlling factors.** *Science* 2002, **296:**1064-1066.
38. Dufresne A, Garczarek L, Partensky F: **Accelerated evolution associated with genome reduction in a free-living prokaryote.** *Genome Biol* 2005, **6:**R14.
39. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, et al: **Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation.** *Nature* 2003, **424:**1042-1047.
40. Strehl B, Holtzendorff J, Partensky F, Hess WR: **A small and compact genome in the marine cyanobacterium Prochlorococcus marinus CCMP 1375: lack of an intron in the gene for tRNA(Leu)(UAA) and a single copy of the rRNA operon.** *FEMS Microbiol Lett* 1999, **181:**261-266.
41. Gans J, Wolinsky M, Dunbar J: **Computational improvements reveal great bacterial diversity and high metal toxicity in soil.** *Science* 2005, **309:**1387-1390.
42. Glavin DP, Cleaves HJ, Schubert M, Aubrey A, Bada JL: **New method for estimating bacterial cell abundances in natural samples by use of sublimation.** *Appl Environ Microbiol* 2004, **70:**5923-5928.
43. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J, Rohwer F: **PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information.** *BMC Bioinformatics* 2005, **6:**41.
44. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
45. **R: a Language and Environment for Statistical Computing** [http://www.R-project.org]

46. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing.** *Journal of the Royal Statistical Society B* 1995, **57:**289 -300.
47. **Perl** [http://www.perl.com]
48. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4:**41.
49. Seshadri R, Adrian L, Fouts DE, Eisen JA, Phillippy AM, Methe BA, Ward NL, Nelson WC, Deboy RT, Khouri HM, et al: **Genome sequence of the PCE-dechlorinating bacterium Dehalococcoides ethenogenes.** *Science* 2005, **307:**105-108.

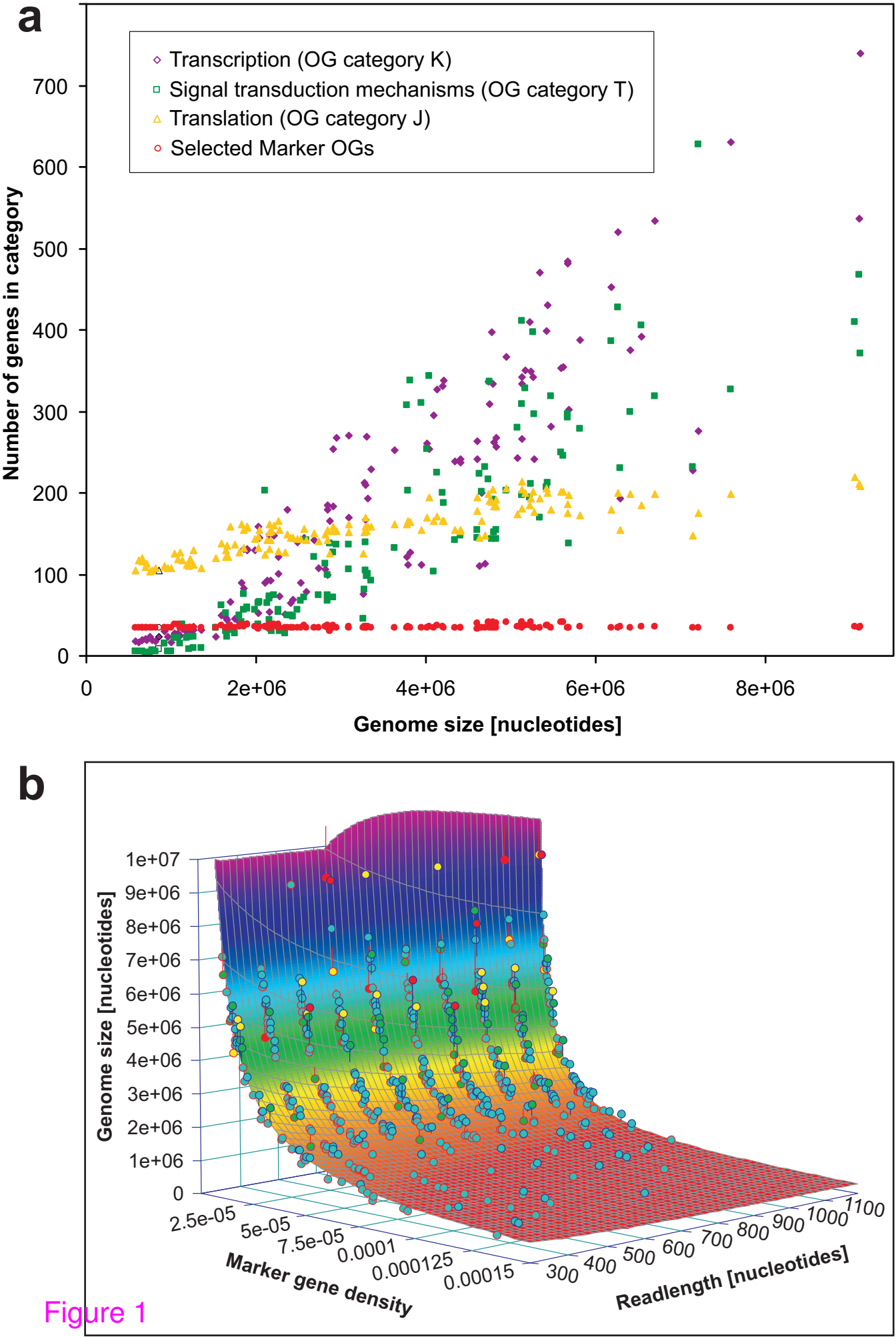
a
Transcription (OG category K)
Signal transduction mechanisms (OG category T)
Translation (OG category J)
Selected Marker OGs
Number of genes in category
0
100
200
300
400
500
600
700
0
2e+06
4e+06
6e+06
8e+06
Genome size [nucleotides]
b
Genome size [nucleotides]
1e+07
9e+06
8e+06
7e+06
6e+06
5e+06
4e+06
3e+06
2e+06
1e+06
0
2.5e-05
5e-05
7.5e-05
0.0001
0.000125
0.00015
Marker gene density
300
400
500
600
700
800
900
1000
1100
Readlength [nucleotides]

Figure 1

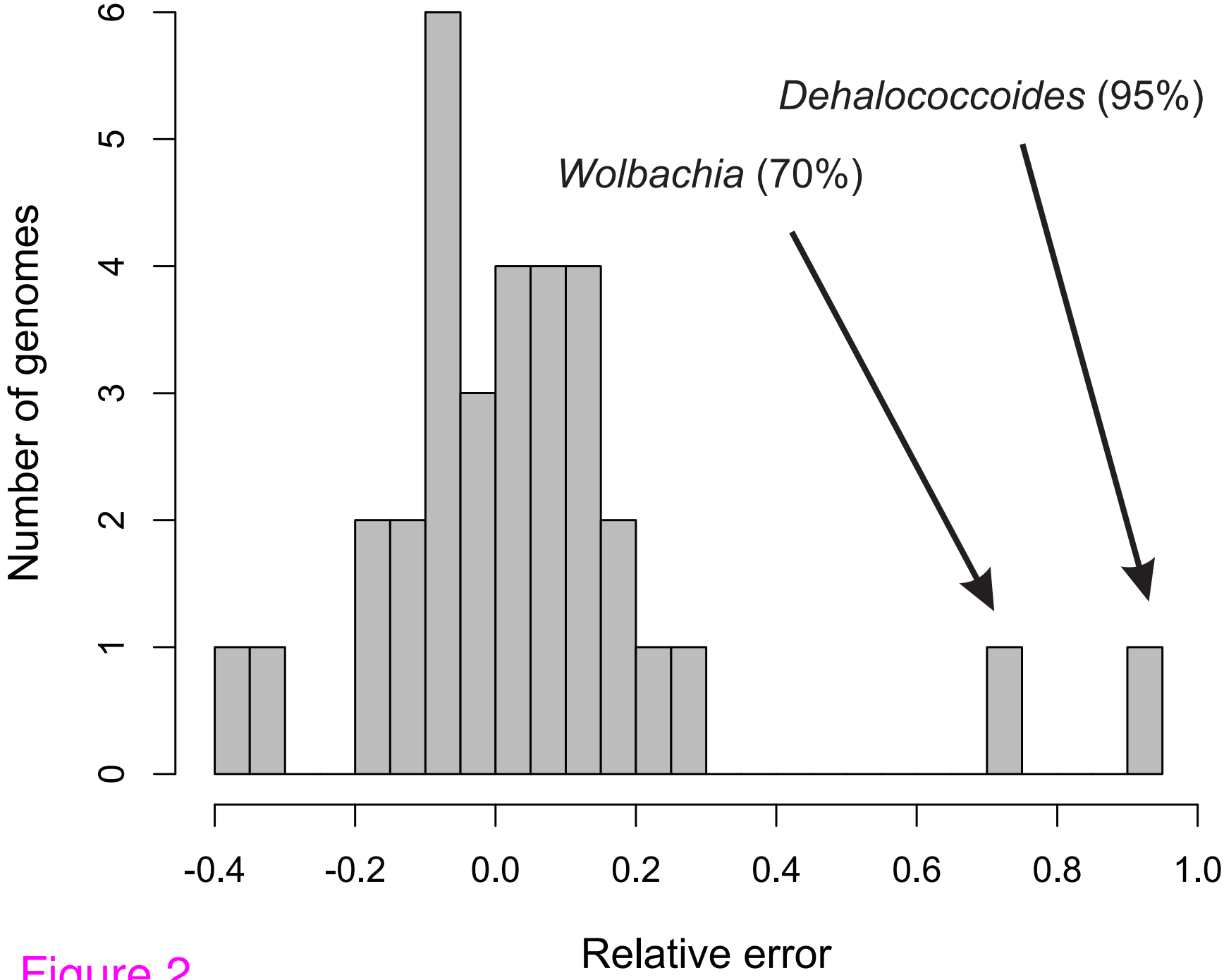
Dehalococcoides (95%)
Wolbachia (70%)
Number of genomes
0
1
2
3
4
5
6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
Relative error

Figure 2

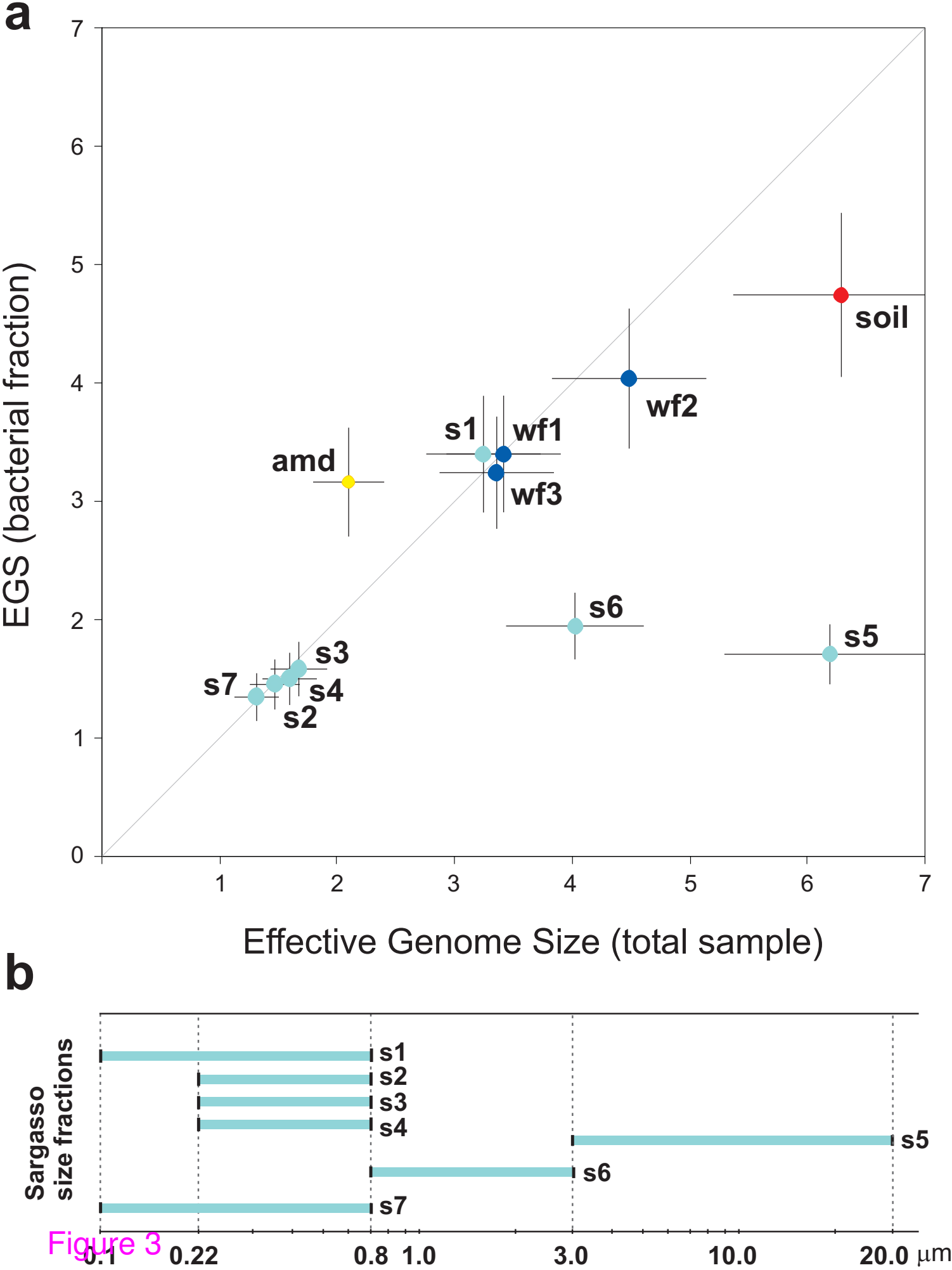
a
EGS (bacterial fraction)
Effective Genome Size (total sample)
soil
wf2
s1
wf1
wf3
amd
s6
s5
s3
s7
s4
s2
b
Sargasso size fractions
s1
s2
s3
s4
s5
s6
s7
0.1
0.22
0.8
1.0
3.0
10.0
20.0 µm

Figure 3

Additional files provided with this submission:

Additional file 1: raes_etal_genomebiol_supplementary_methods.pdf, 124K
http://genomebiology.com/imedia/8515386441244730/supp1.pdf
Additional file 2: raes_etal_GB_add_t1.pdf, 41K
http://genomebiology.com/imedia/2104562741136246/supp2.pdf
Additional file 3: raes_etal_GB_add_t2.pdf, 55K
http://genomebiology.com/imedia/3496901781136246/supp3.pdf
Additional file 4: raes_etal_GB_add_t3.pdf, 52K
http://genomebiology.com/imedia/3023147361136246/supp4.pdf
Additional file 5: raes_etal_GB_add_t4.pdf, 87K
http://genomebiology.com/imedia/3801944241136246/supp5.pdf
Additional file 6: raes_etal_GB_add_t5.pdf, 52K
http://genomebiology.com/imedia/5888620111136246/supp6.pdf
Additional file 7: raes_etal_GB_add_f1.pdf, 90K
http://genomebiology.com/imedia/1240440371113624/supp7.pdf
Additional file 8: raes_etal_GB_add_f2.pdf, 70K
http://genomebiology.com/imedia/3661808821136246/supp8.pdf
Additional file 9: raes_etal_GB_add_f3.pdf, 65K
http://genomebiology.com/imedia/1310973751113624/supp9.pdf
Additional file 10: raes_etal_GB_add_f4.pdf, 65K
http://genomebiology.com/imedia/4535539631136246/supp10.pdf
Additional file 11: raes_etal_GB_add_f5.pdf, 66K
http://genomebiology.com/imedia/1117484638113624/supp11.pdf