Bayesian inference of community average gene copy numbers and its application for the metagenomic characterization of bacterioplankton community types in the Sargasso Sea

Bánk Beszteri[1,2], Stephan Frickenhaus[2], Stephen J. Giovannoni[1]

[1] Department of Microbiology, Oregon State University, Corvallis, OR, USA.
[2] Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany

Average genome sizes in comparative metagenomics

Comparative metagenomics builds upon shotgun sequencing of community DNA and *in silico* annotation of sequencing reads to orthologous groups to identify groups of protein coding genes occurring at different abundances among groups of samples (e.g., control vs. treatment), or ones showing trends with environmental parameters. One of the methodological issues arising by such analyses is that **apparent abundances** of fragments of protein coding genes **depend on the average genome sizes of communities** (see Figs 1 & 2). This has been ignored in published comparative metagenomic studies, although Raes et al. (2007) drew attention to its potential relevance, and Beszteri et al. (2010) explored in more depth the effect and possibilities to correct for it. Here we introduce a novel approach to incorporating average genome size effects into comparative metagenomics, on examples from a metagenomic study of the Sargasso Sea.

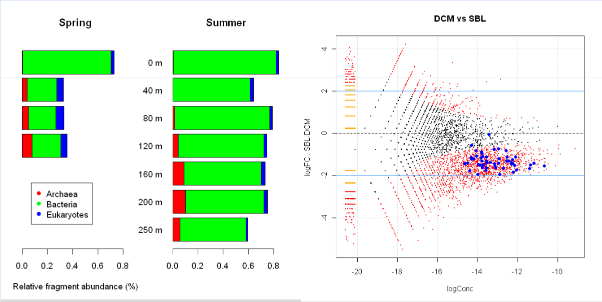

Figure 1. Illustration of the effect of average genome size differences upon gene relative abundances (left) and upon statistical inferences of differential presence (right). Left: relative abundances of fragments of 35 “universal single copy” markers (based on Raes et al. 2007a) across two Sargasso Sea depth profiles (see Fig. 2). Right: log fold change vs. log concentration of genes in a comparison of summer deep chlorophyll maximum (DCM) vs. spring bloom (SBL) from the same metagenomic samples. Red dots represent genes inferred to be present at significantly differential abundances between the two community types (black: not significantly different); blue dots mark 35 “universal single copy” genes. When ignoring the effect of average genome sizes, most “universal single copy” genes are inferred to be present at highly significantly different abundances between both community types.

Simultaenous estimation of average genome sizes and gene abundances

In Beszteri et al. (2010), we presented the first explicit statistical model of shotgun metagenomic sequencing. We suggested that apparent relative abundances of orthologous genes in metagenomic samples are proportional to the product of sampling yield (number of reads sequenced) and the “concentration” of individual genes in the sample, which is, in turn, expected to be proportional to their length, and inversely proportional to the average genome size of the community:

$$\lambda_{m,s} = \frac{l_m}{G_s} C_{m,s} N_s$$

where *m* denotes marker (gene), *s* sample, *lambda* the expected count of fragments of a gene, *l* the length of a gene, *G* the average genome size of a community, *C* the average copy number of the gene in the community concerned, and *N* the number of reads sequenced.

We model gene counts either using the Poisson (following Kristiansson et al. 2009) or the negative binomial distribution with an overdispersion parameter *omega* common to all genes (following Robinson & Smyth 2007). For ANOVA type comparisons among groups *g*, we have

$$\log(\lambda_{m,s}) = \log\frac{1}{G_s} + \log(C_{m,g}) + \log l_m N_s$$

and $Y_{m,s} \approx Pois(\lambda_{m,s})$ or $Y_{m,s} \approx NB(\lambda_{m,s}, \omega)$

Parameter identifiability requires further constraints. We fixed the average copy number (*C*) of universal single copy genes at unity, effectively normalizing gene abundances to their baseline.

We use a Bayesian framework for model fitting as this avoids difficulties of error propagation arising when sequentially estimating *G* and *C* using maximum likelihood. We implemented an adaptive Markov Chain Monte Carlo (MCMC) algorithm after Rosenthal (2007) to effectively fit these models to real-life metagenomic data sets (close to 100,000 parameters). Availability of the full posterior distribution of *G* and *C* enables flexible statistical inference and data exploration.

Seasonality in the Sargasso Sea

The **Bermuda Atlantic Time-series Study (BATS)** site hosts one of the largest long term multidisciplinary oceanographic observatory programs. One of the aims of this program is to clarify the composition of microbial communities occurring in the water column, its changes driven by environmental factors and its activities in the biogeochemical cycling of elements. We recently undertook a comparative metagenomic study of **seasonality** in this system.

Seasonality is one of the well-documented, constant features of the BATS site. Cold winds in the winter lead to deeper mixing and transport of nutrients from the mesopelagic zone towards the surface, fuelling **spring phytoplankton blooms**; whereas thermal stratification and surface nutrient depletion result in highly **oligotrophic summer** conditions.

Recent bacterial SSU fingerprinting by Treusch et al. (2009) revealed the presence of four distinct **community types** in this system: a spring bloom community; and three communities during **summer** stratification, associating with the **surface**, **deep chlorophyll maximum (DCM)**, or the **upper mesopelagic (UMP)**. On this poster, we further subdivide the **spring** samples into **surface** (**SSF**) vs. all other samples, the latter representing the **spring bloom** (**SBL**; see Figure 2).

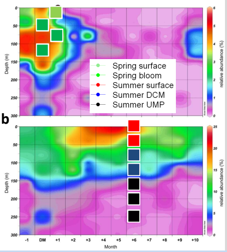

Figure 2. Mean relative abundances of SSU amplicons from Prasionphyte plastids (a; illustrating the spring phytoplankton bloom), and from the bacterial group SAR116 (b; illustrating summer stratification), after Treusch et al. (2009). The X axis represents time in months, relative to the spring bloom.

Our sampling strategy was designed to capture typical representative communities of the different community types. We sampled a **spring bloom** at **four depths** (0-120 m, Figure 2a), and a **summer** stratified water column at **seven** (0-250 m , Figure 2b) for total DNA sequencing on the 454 FLX platform.

The results presented here are based on a best-BLAST-hit annotation using the STRING database (v8.1, http://string.embl.de).

Results for the Sargasso Sea data

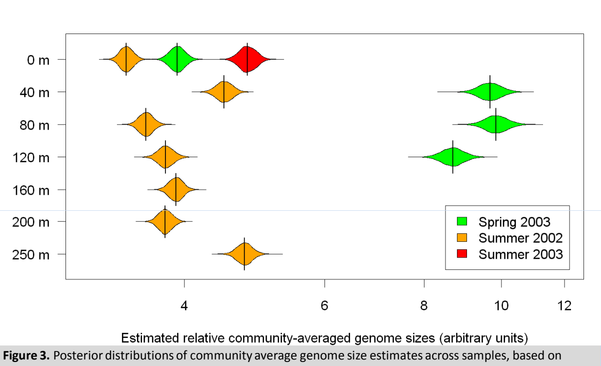

Figure 3. Posterior distributions of community average genome size estimates across samples, based on relative abundances of 35 single copy markers.

Fitting the negative binomial model to our Sargasso Sea data demonstrates the large increase in community average genome size during the spring bloom (Fig. 3). This is in line with available information on the higher abundance of photosynthetic pico-eukaryotes (mainly Prasinophytes) in these communities. Fig. 4 illustrates the posterior distribution of average copy numbers of selected genes related to inorganic nitrogen acquisition.

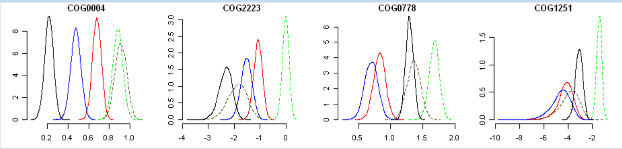

Figure 4. Estimated average copy numbers over all taxa for COG0004 (ammonia permease), COG2223 (nitrate / nitrite transporter), COG0778 (nitrate reductase) and COG1251 (nitrite reductase). Curves represent posterior densities for the five community types on a logarithmic scale (i.e, 0 means one copy per cell on average). Color code as in Fig. 2.

References

Beszteri, Temperton, Frickenhaus, Giovannoni (***2010***). Average genome size: a potential source of bias in comparative metagenomics. **ISME J.** 4: 1075.
Kristiansson, Hugenholtz, Dalevi (***2009***). ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. **Bioinformatics** 25: 2737.
Raes, Foerstner, Bork (***2007***). Get the most out of your metagenome: computational analysis of environmental sequence data. **Curr. Opin. Microbiol.** 10: 490.
Raes, Korbel, Lercher, Bork (***2007a***). Prediction of effective genome sizes in metagenomic samples. **Genome Biol.** 8: R10.
Robinson, Smyth (***2007***). Moderated statistical tests for assessing differences in tag abundances. **Bioinformatics** 23: 2881.
Rosenthal (***2007***). AMCMC: an R interface for adaptive MCMC. **Comp. Stats Data Anal.** 51: 5467.
Treusch, Vergin, Finlay, Donatz, Burton, Carlson, Giovannoni (***2009***). Seasonality and vertical structure of microbial communities in an ocean gyre. **ISME J.** 3: 1148.

Acknowledgement

This work was funded by the Gordon and Betty Moore Foundation. Thanks for the Scientific Computing of the AWI for support to participate on ECCB10, and to Alexander Treusch for Fig. 2.