Computational Methods for Metagenomic Analysis

Itai Sharon
Computational Methods for Metagenomic Analysis

Research Thesis

In Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy

Itai Sharon

Submitted to the Senate of
The Technion – Israel Institute of Technology
Sivan, 5770 Haifa May 2010
This research was done under the supervision of Assoc. Prof. Oded Béjà and Assoc. Prof. Ron Pinter in the Faculty of Computer Science.

I would like to thank my family for their love and support: my dearest wife, Mayan, my beloved kids, Orr, Yahel and Peleg, and my dear parents, Nira and Dan. I appreciate all that you have done for me, and I love you all very much.

This work is dedicated to the memory of my uncle, Yair Sagi, and my grandparents, Zelda and Arie Solodukhin and Yehuda Falkovitch, Hassya and Shlomo Brat.

The Generous Financial Help of the Technion and the Joan and Erwin Jacobs Foundation Is Gratefully Acknowledged.
# Table of Contents

Abstract ........................................................................................................................................ 1  
Abbreviations and Notations ........................................................................................................ 3  
1. Introduction ................................................................................................................................ 5  
   1.1 Computational challenges in metagenomics ................................................................. 6  
   1.2 Metagenomics projects ............................................................................................... 9  
   1.3 The GOS database .................................................................................................... 10  
   1.4 *Synechococcus*, *Prochlorococcus* and their cyanophages .................................. 11  
   1.5 Organization of this work ......................................................................................... 13  
2. The Statistics of Metagenomics Functional Analysis ..................................................... 15  
   2.1 Previous work ........................................................................................................... 17  
   2.2 Functional analysis at the gene family level ................................................................ 18  
      2.2.1 Estimating the frequency of a protein family .................................................... 18  
      2.2.2 Computing confidence bounds ............................................................................ 20  
      2.2.3 Transition from genomes to metagenomes ....................................................... 21  
      2.2.4 Performing function comparison ........................................................................ 23  
   2.3 Functional analysis at the pathway level ...................................................................... 23  
      2.3.1 The Poisson model for computing gene family abundance ............................... 25  
      2.3.2 Estimating the average genome length in the DNA sample .............................. 26  
      2.3.3 Computing pathway abundance: the independent pathways model .... 26  
      2.3.4 Computing pathway abundance: the pathways intersection model .......... 27  
   2.4 Performance evaluation for the gene family model .................................................. 28  
      2.4.1 Evaluation of simulated data ............................................................................... 28  
      2.4.2 Evaluation of data derived from WGS sequencing projects ......................... 31  
      2.4.3 Evaluation of real data ..................................................................................... 33  
   2.5 Performance evaluation for the pathway models ..................................................... 33  
      2.5.1 Materials ........................................................................................................... 33  
      2.5.2 Evaluation of the different methods ..................................................................... 34  
      2.5.3 Results ............................................................................................................. 35
3. Studying Genes and Systems from Metagenomics Data ..................................41

3.1 Studying gene organizations ............................................................................. 41
3.2 Separating viral and microbial sequences .......................................................... 44
  3.2.1 Gene based classification .............................................................................. 45
  3.2.2 Classification by recruitment ....................................................................... 46
  3.2.3 Computing coverage in metagenomic datasets .............................................. 46
  3.2.4 Recruitment based sequence classification: the Northern Line Islands samples .............................................................................................................. 47
3.3 Applications: studying the tryptophan operon in the Sargasso Sea metagenome .................................................................................................................. 53
3.4 Applications: finding viral genomes with microbial genes ............................... 53
  3.4.1 Methods ......................................................................................................... 54
  3.4.2 Results ............................................................................................................ 57

4. The Discovery of Photosystem I Genes on Viral Genomes ...............................61

4.1 Results .................................................................................................................. 61
4.2 Methods ............................................................................................................... 70
  4.2.1 Collecting preliminary set of viral-PSI scaffolds in GOS ......................... 70
  4.2.2 Collecting GOS-PSI clones .......................................................................... 70
  4.2.3 The viral-PSI like set (Table 10) ................................................................... 71
  4.2.4 Recruitment against 454 databases – biomes (Table 11) ......................... 71
  4.2.5 Recruitment against 454 databases – clones (Figure 19) ......................... 71
  4.2.6 Recruitment of PSI clones against viral and microbial 454 datasets (Figure 19b) ........................................................................................................... 72
  4.2.7 Gene organization analysis (Figure 18) ......................................................... 73
  4.2.8 Estimation of PSI and PSII gene abundance in phages (Table 13 and Table 14) ............................................................................................................. 73

5. Discussion .............................................................................................................77

References ..............................................................................................................83
List of Figures

Figure 1. (a) Any read of length $L$ that begins within the $L-T$ base pairs before the target gene or somewhere on the first $g-T$ base pairs on the gene will cover at least $T$ base pairs of the gene. Overall there are $L+g-2T$ positions in which such reads may begin. (b) Each clone may cover $G$ with either its forward or backward reads. Combined with (a) this gives a total of $2(L+g-2T)$ positions.

Figure 2. (a) The independent pathways model. In this model a gene that is shared among several pathways is assumed to have a copy for each pathway in which it appears. For example: $G_5$ belongs to three pathways and thus assumed to have three copies (b) The pathway intersection model. Each gene that appears in one or more pathways is assumed to appear once. $G_5$ in this case will have a single copy, shared between $P^2$, $P^3$ and $P^4$.

Figure 3. Quality of frequency estimations as a function of (a) gene size, (b) family size, and (c) a combination of both. In all cases coverage is extremely high (94.2X) in order to demonstrate the inherent differences between the methods even when redundant data is available.

Figure 4. The influence of coverage on frequency estimation for (a) coverage=9.42X, (b) 0.942X and (c) 0.0942X.

Figure 5. (a) Reliability bounds as a function of the number of base pairs associated with protein families, $\varepsilon=0.1$ used. (b) Log-ratio error as a function of the number of base pairs associated with different families. For each range $[x \cdot 10\text{Kbps}, (x+1)\cdot 10\text{Kbps}]$ median log-ratio error for all families in the range is shown. Confidence bounds become narrow with the increase in the family's share on the genome, in line with the increased accuracy of frequency estimation.

Figure 6. Summary of data analysis for 44 single genome WGS projects. (a) Log-ratio between estimations and true frequencies, sorted by family length. For each genome, average and standard deviation of gene lengths were computed and used for binning evaluations into $x0.5$ standard deviation bins. Next the median for each bin was computed and kept with medians of the same bin from other genomes. Last the median over all medians for each bin was computed and used for generating the figure. (b) Same process as before using number of members for each family as the key. (c) Same process, using number of base pairs occupied by each family on the genome as the key.

Figure 7. Agreement between true and estimated lists of most differentially enriched pathways in selected pairs of metagenomes. For each $m \leq 100$ (X axis), the frequency of pathways that are among the $m$ most differentially enriched pathways in both the true and estimated ranked lists is plotted for (a) metagenomes M1 and M5, and (b) M3 and M4. Refer to the appendix for plots.
of all other pairs.

**Figure 8.** Ranking pathways based on their enrichment in metagenome M3. (a) The intersection between the true and predicted most abundant pathways using the various prediction methods (y-axis), for different values of m (x-axis). (b) Statistical significance (hyper-geometric p-values) for the intersection between the true and predicted highly abundant pathway sets shown in (a). Other simulated metagenomes exhibited similar behavior (see Figure 11).

**Figure 9.** Pearson correlation between predicted and true pathway abundances across the various metagenomes. (a) Correlations obtained for the entire set of 250 pathways used in the simulation, (b) correlations obtained for the set of 150 less abundant pathways.

**Figure 10.** Agreement between true and estimated lists of most differentially enriched pairs of metagenomes. Refer to the legend of Figure 7 for description. Read count (blue), independent pathways (green) and pathway intersection (red) models are compared.

**Figure 11.** Ranking pathways based on their abundances in metagenomes M1, M2, M4 and M5. Refer to the legend of Figure 8 for description. Read count (blue), normalized read-count (cyan), independent pathways (green) and pathway intersection (red) models are compared.

**Figure 12.** Gene organization graph of tryptophan genes in the Sargasso Sea metagenome. Vertices refer to genes (X is a shortcut for trpX), arrows represent to neighboring genes. Arrow size and color indicates number of occurrences. Refer to (Kagan, Sharon et al. 2008) for a full description.

**Figure 13.** Schematic description of the VirMic pipeline. Refer to text for a full description.

**Figure 14.** Cluster size distribution. Left: 12,527 out of the 13,741 clusters (91.2%) contained exactly one refseq-hit and 1,214 contained more than 1 member. Right: for each cluster size \( \geq 2 \), distribution of number of clusters of this size.

**Figure 15.** Recruitment of VirMic scaffolds (blue) and scaffolds containing 16s rRNA coding genes (red) against the Northern Line Islands viral (X axis) and microbial (Y axis) fractions, excluding the Christmas Island samples.

**Figure 16.** A maximum likelihood phylogenetic tree of psaA-deduced amino acid sequences obtained from the GOS expedition. PsaA sequences from the 27 fully sequenced and annotated Synechococcus (blue background) and Prochlorococcus (green background) genomes are shown. Sequences from the GOS expedition are shown in bold, and sequences from the original scaffolds obtained in this study are indicated in red. For clarity, the tree shows only a subset of the 583 partial PsaA sequences found in the GOS data set. The tree is
on the basis of an alignment of 94 shared amino acids.

**Figure 17.** Schematic physical maps of selected viral-suspected GOS clones (top), *Prochlorococcus* and *Synechococcus* genomes (middle) and environmental PCR products containing PSI genes (bottom). Red arrows represent ORFs with predicted viral origin, and grey arrows represent unknown ORFs. Capital letters represent the corresponding PSI genes. Gaps shown in Indian Ocean GOS clones (stations GS111 and GS117) are the result of regions that were not covered by the end-reads owing to the size of these clones (5 Kbps). Primer positions on GOS clones are indicated by triangles, and thick colored lines denote PCR products.

**Figure 18.** Distribution of neighboring genes involving at least one PSI gene. Each arrow connects neighboring genes, and its thickness represents the number of pairs found in *Synechococcus* and *Prochlorococcus* genomes (left gene-circle), microbial sequences from the GOS metagenome (middle gene-circle) and viral sequences from the GOS metagenome (right gene-circle). Note the uninterrupted clustering of PSI genes in phage genomes that contrasts the scattered arrangement of these genes in cyanobacterial genomes (in both cultures and GOS). Gene connections observed only once are not shown.

**Figure 19.** Recruitment of GOS clones carrying PSI genes with Northern Line Islands biomes. Left: Recruitment of three GOS physical clones carrying suspected viral-PSI genes by Northern Line Islands virome (green) and microbiome (red) reads. Upper panel shows recruitment at 75-100% identity while lower panel shows fold coverage by these reads. Accession numbers of the GOS reads used are presented above each clone (JCVI_READ_#). **Right:** Recruitment coverage of GOS single clones carrying PSI genes with Northern Line Islands biomes [viromes (X-axis) and microbiomes (Y-axis)]. Coverage is defined as % of GOS clone length covered by at least one recruited read.

**Figure 20.** Structural consequences of assembling the viral fusion protein PsaJF into PSI. **Left:** The structure of *T. elongatus* PSI (subunits) was illustrated by PyMOL (http://pymol.sourceforge.net/) using a PSI monomer (adopted from PDB 1jb0). PsaF (magenta), PsaJ (blue), and all the other subunits in green. **Right:** A model for the structure of the viral PsaJF fusion protein (red) substituting the original PsaF and PsaJ subunits.

**Figure 21.** Phylogenetic trees of viral GOS, *Synechococcus* & *Prochlorococcus* PsaC (left) and PsaD (right) proteins. The trees displayed are the ones having the highest likelihood of the 100 trees. Even though each tree has a different number of homologues and a different evolutionary distance between them, a pattern is evident: the cyanobacterial homologues are evolutionary closer among themselves than to any viral homologue, with the exception of viral homologue 8692 (JCVI_SCAF_1096628008692 that contains GOS clone 9984). The displayed GOS leaf names are the last digits of the corresponding accession numbers, Color code is the same as in Figure 16; *Synechococcus* sequences are
labeled with blue, *Prochlorococcus* with green and viral GOS in red. *Synechocystis* PCC6803 and *Trichodesmium erythraeum* sequences are shown for reference (in black).
List of Tables

Table 1. Information for participating genomes in the real data analysis. 32

Table 2. General information on simulated metagenomes. 34

Table 3. Genomes with highest coverage, samples 4440036.3 and 4440037.3 (Kingman). Green: microbial genomes, orange: viral genomes, blue: plasmids. 48

Table 4. Genomes with highest coverage, samples 4440040.3 and 4440039.3 (Palmyra). Green: microbial genomes, orange: viral genomes, blue: plasmids. 49

Table 5. Genomes with highest coverage, samples 4440280.3 and 4440279.3 (Tabuaeran). Green: microbial genomes, orange: viral genomes, blue: plasmids. 50

Table 6. Genomes with highest coverage, samples 4440038.3 and 4440041.3 (Christmas). Green: microbial genomes, orange: viral genomes, blue: plasmids. 51

Table 7. Number of reads carrying a segment of 16s or 23s rRNA coding genes. 52

Table 8. Largest refseq-hits clusters and their annotation. 58

Table 9. Manually validated VirMic:Microbial clusters with at least four instances. 60

Table 10. Distribution of viral-PSI clones (one or two reads sequenced from the same DNA fragment) in the GOS expedition. 64

Table 11. Number of different biomes reads recruited at different percent identity to GOS suspected viral-PSI clones. 65

Table 12. Accession numbers of Line Island 454 reads (from reads recruited on GOS scaffolds shown in Figure 19) corresponding to the junctions between the different viral PSI genes (only reads containing both end and start of the adjacent proteins are shown). Numbers in brackets represent the length of the spacer between the end and start of each adjacent protein pairs (the minus number in psaE>psaD represents overlapping ORFs). 68

Table 13. Percent cyanophages carrying PSI (psaA) or PSII (psbA) genes in the different Northern Line Islands viromes. 69

Table 14. Percent occurrence of viral PSI (psaA and psaB) or PSII (psbA and psbD) genes in cyanophages (myoviruses & podoviruses) in Northern Line Islands viromes as indicated by the presence of psaA, psaB, psbA and psbD genes respectively. Number of 454 reads recruited for each gene/virome and the estimated frequency across cyanophages (in brackets) is given. Percents were calculated in relationship to the number of 454 reads recruited with myo-cyanophage g20 gene (encoding for the phage capsid assembly protein) and g23 gene, and with podo-cyanophage DNA polymerase gene (See methods for
exact procedure).
Abstract

Metagenomics is a new field in which genetic material is extracted directly from the environment and is subsequently analyzed by a variety of biological and computational methods. Using metagenomics, it is possible to study the vast majority of microbes on earth, of which more than 99% cannot be cultured in the laboratory. Metagenomic data usually consists of many short (100-1,000 bp) DNA sequences, potentially originating from all organisms in the examined environment. Several computational challenges arise as a result, some of them are known from genomics (e.g. DNA assembly, gene prediction and functional analysis) while others are unique to metagenomics (e.g. sequence binning, in which we try to assign sequences to taxonomic bins). Many metagenomic projects have been carried out in recent years, projects that have broadened our understanding of biological processes in a way that was impossible heretofore. On-going and new projects, such as the Global Ocean Sampling (GOS) expedition, promise that the flux of discoveries will increase in the coming years.

In my PhD I chose to focus on two aspects of metagenomics analysis: (i) the statistics of functional analysis of metagenomes, and (ii) the study of genes and gene organizations from metagenomic data. The viewpoint of the first part is global: given a metagenome, we are interested in studying functional characteristics of organisms living in the examined environment which may hint us as for conditions most important in that environment. Based on the Lander-Waterman model for whole genome shotgun sequencing projects we were able to provide a statistical model that accurately estimates the expected number of sequences containing some part of a gene in a metagenome. The model was tested on both simulated and real data, and was shown to provide estimates that are in line with real values. The statistics of pathways is also discussed: in this case a different model was required that will take into account the possibility of genes that participate in more than one pathway.

The second part of this work takes a "local" view: rather than looking at microbial communities in general, we are interested in answering specific questions on specific genes or systems. This part begins with the description of our discovery of Photosystem-I (PSI) gene cassettes on viral genomes. Using metagenomic data from the Global Ocean Survey (GOS) expedition and the Northern Line Islands we were able to show that a gene cassette of eight PSI genes, potentially sufficient for coding all necessary proteins of fully functional PSI, is present on DNA sequences of viral origin. In this work we used several computational tools that were developed by me, some of them novel to this work while others were also used in other works. I will also describe a generalization to this work in which we were able to discover microbial genes on viral genomes in general, using existing and novel methods and strategies.
# Abbreviations and Notations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOS</td>
<td>Global Ocean Sampling</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>IMM</td>
<td>Interpolated Markov Model</td>
</tr>
<tr>
<td>Kbp</td>
<td>Kilo base-pairs</td>
</tr>
<tr>
<td>LSU</td>
<td>Large Subunit</td>
</tr>
<tr>
<td>Mbp</td>
<td>Mega base-pairs</td>
</tr>
<tr>
<td>NGS</td>
<td>Next Generation Sequencing</td>
</tr>
<tr>
<td>ORF</td>
<td>Open Reading Frame</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>PSI</td>
<td>Photosystem I</td>
</tr>
<tr>
<td>PSII</td>
<td>Photosystem II</td>
</tr>
<tr>
<td>SSU</td>
<td>Small Subunit</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole Genome Shotgun</td>
</tr>
</tbody>
</table>
Chapter 1.

Introduction

Metagenomics, defined as the sequencing and analysis of genomic data taken directly from the environment, is a new and rapidly developing field that makes it possible to study uncultured organisms. Several important discoveries have been made in recent years by extracting data directly from the environment, including the discovery of proteorhodopsins (Beja, Aravind et al. 2000). Unlike organism specific genome projects in which a single genome is considered, the data collected in metagenomic projects are usually highly fragmented and belong to many organisms. However, evolving computational tools help in deciphering the enormous amounts of data gathered in these projects.

There are three main approaches for the analysis of data from the environment. Functional analysis involves the expression of genes from the environment in different hosts, such as E. coli. Once genes are expressed the hosts are screened for discovering functionality of interest. This approach yielded several important findings, including the discovery of antibiotic resistant genes from environmental organisms (Riesenfeld, Goodman et al. 2004). In the sequence based approach, sequences that contain interesting genes, such as 16S rDNA, are cloned. Proteorhodopsin, for example, was discovered using this approach (Beja, Aravind et al. 2000). In the environment based approach, the entire DNA present in the environment is screened and analyzed. This approach provides a global view of an entire population, rather than the species-specific view of the other two approaches. In these projects the DNA is sequenced using the Whole Genome Shotgun (WGS) sequencing technique, in which the DNA is broken to small pieces that can be sequenced.

WGS was first used for the sequencing of single genomes (Fleischmann, Adams et al. 1995; Myers, Sutton et al. 2000; Venter, Adams et al. 2001). The data resulting from the process is composed of many short (~700 bp) sequences of DNA called reads. The first step in analyzing the data is usually the assembly of reads into longer sequences called contigs and scaffolds. The number of scaffolds, contigs and reads that are left once the assembly process is complete depends on the amount of coverage in the project, which is the average number of reads that covers a single point on a genome. The coverage depends on the amount of data collected, as well as the diversity of species in the tested environment. Abundant species may achieve good coverage that is sufficient for complete or almost complete sequencing of their genome, while the coverage of rare species is usually poor. Once the assembled sequences are available, it is possible to do further processing, such as gene finding, gene annotation, protein prediction, etc.
This chapter is organized as follows. In section 1.1 we discuss some of the major computational challenges that arise in the analysis of metagenomic data. Section 1.2 lists the most important metagenomic projects carried out so far, while section 1.3 focuses on the GOS database which is the largest metagenomic project that took place so far. This chapter finishes with a short overview of the biology of bacteria and their phages most relevant to this work, brought in section 1.4.

1.1. Computational challenges in metagenomics

Metagenomic projects pose several challenges with respect to the analysis of the data gathered. In this section some of the major computational challenges are described. For a good review on the topic refer to (Wooley, Godzik et al. 2010).

**Assembly.** Assembly is mostly relevant to Sanger sequencing data, where read length usually approaches 1,000bps. It is usually carried out using a single genome assembler such as the Jazz (Aparicio, Chapman et al. 2002) or the Celera (Myers, Sutton et al. 2000) assemblers with certain modifications (Venter, Remington et al. 2004). The use of comparative analysis with respect to template genomes may increase confidence in the final result. The transformation from the assembly of a single genome to the assembly of many genomes at once is not trivial and raises several issues. The presence of conserved regions in several different organisms is likely to harm the assembly due to the assembler's inability to differentiate closely related stretches. Such stretches are usually considered as being repetitive and are ignored. The same problem is caused by the presence of several copies of a region that belongs to an abundant genome, in particular when the region is highly polymorphic. This problem is treated manually (Venter, Remington et al. 2004).

A different type of a problem is related to the amount of DNA one has to collect in order to obtain good coverage of the target genomes. The theory for single genomes is described in (Lander and Waterman 1988), and supplies tools for estimating the amount of gaps one may expect to have in the assembly of a given genome given a certain amount of coverage. Constructing such model for multiple genomes is much more difficult and requires an estimation of the diversity of organisms in the tested environment.

Assembly of Next Generation Sequencing (NGS) Data is rarely done due to the short read length. Current assemblers for NGS data such as Velvet (Zerbino and Birney 2008) may handle de-novo assembly of microbial genomes but are not suitable for Eukaryotic genomes as well as metagenomes. Strategies and assemblers for NGS metagenomic data are therefore of great interest for the metagenomics community.

**Phylogenetic analysis.** Metagenomic data, by its nature, is composed of many short sequences from many organisms. This raises several issues related to phylogenetic analysis, both with respect to traditional and metagenomic-specific analysis. Locating an organism whose genome is known on the phylogenetic tree with respect to other known organisms is traditionally carried out using phylogenetic informative markers, primarily rRNA encoding genes. More than 1,400,000
microbial 16S Small Subunit (SSU) rRNA coding genes are available today in databases such as the Ribosomal Database Project (RDP) (Cole, Chai et al. 2005; Cole, Wang et al. 2009), which can be used for phylogenetic classification of environmental data. Other phylogenetic marker genes such as the recA and HSP70 genes are also used (Venter, Remington et al. 2004). However, most reads and scaffolds in metagenomic projects obviously do not contain phylogenetic marker genes. This raises the need for other methods that may be based on oligonucleotide frequencies(Abe, Sugawara et al. 2005) or other properties of the genome. Methods that make no use of universal genes may be useful also for phylogenetic analysis of viruses, in which no universal genes are available.

Even when reads with phylogenetic marker genes are available, the task of reconstructing a phylogenetic tree may not be trivial, due to the availability of gene fragments only. As a first step in the process, a Multiple Sequence Alignment (MSA) of all sequences is aligned. Next, the MSA is used for constructing a distance matrix that is used for constructing the phylogenetic tree. The fragmented nature of metagenomic data usually results with partial sequences of the same region, which will make it impossible to use current algorithms for MSA on such data. This problem may require a new type of algorithms for aligning and scoring multiple sequences (Chen and Pachter 2005).

**Taxonomic binning.** Sequence binning refers to the separation of sequences into taxon specific groups. A binning step may be part of the assembly process of metagenomic data (Chen and Pachter 2005), or may be used for separating the genomes of a few members in order to study the biological processes carried by each one of them(Woyke, Teeling et al. 2006). The two main approaches suggested with respect to this problem employ comparative and nucleotide frequencies techniques, with the latter being more successful. The TETRA sequence analysis software and its underlying similarity measure between sequences (Teeling, Meyerdierks et al. 2004; Teeling, Waldmann et al. 2004) use a representation of tetrancleotide frequencies for a sequence, and the Pearson correlation coefficient as a similarity measure between sequences. TETRA has been successfully used in (Woyke, Teeling et al. 2006) on a system of four organisms. The PhyloPythia classification algorithm(McHardy, Martin et al. 2007) employs an SVM for classifying sequences to known taxa, based on their nucleotide frequencies.

**Gene calling.** Several good solutions have been proposed in the past with respect to the problem of gene finding in fully sequenced genomes. Hidden Markov Model (HMM) (Borodovsky and McIninch 1993) and Interpolated Markov Model (IMM) (Salzberg, Delcher et al. 1998; Delcher, Harmon et al. 1999) based algorithms use sequence statistics for distinguishing between coding and non-coding regions. These algorithms are trained on sequences from the target genome; an initial set of genes is obtained by comparative methods. Other approaches, including searching for Open Reading Frames (ORFs) by considering start and stop codons, or aligning the genome against databases of known genes or proteins (Gish and States 1993), have been suggested as well. All these approaches may encounter difficulties when
applied to metagenomic data, difficulties that are due to the fragmented nature of the sequences in these projects. A large portion – sometimes up to 50% – of all reads in metagenomic projects cannot be assembled (Venter, Remington et al. 2004) with the rest of the sequences being usually assembled into short contigs of a few Kbps at most. Considering the length of reads (400-1,000 bps for Sanger and 454 sequencing, less than 100 for Solexa) and the fact that prokaryotic genes are usually short, lack introns and occur approximately every 1,000bps, it is probably the case that valuable information is present on most reads and contigs. Sequence-statistics based algorithms may encounter difficulties in analyzing such fragmented data due to their dependence on whole genome statistics. Comparative based methods are also expected to encounter difficulties due to the presence of many gene fractions although they seem as a more natural choice. The approach described in (Yooseph, Sutton et al. 2007) employs a clustering algorithm in which similar sequences in GOS are clustered together based on sequence similarity. The primary input for the algorithm is the pairwise sequence similarity between all sequences in the database computed using an all against all BLAST search. These similarity scores are used both for removing redundancies, and also for the construction of core sets, which contain highly conserved sequences. Next, close core sets are unified based on profile-profile comparison. Last, the profiles of the resulting sets are used for sequence recruitment using PSI-BLAST. Clusters containing sequences that are similar to annotated sequences may be assigned predicted functionality; however, approximately 25% of the clusters in GOS do not have any known homolog.

**Community level analysis.** In this case we are interested in answering questions related to microbial communities as a whole, rather than to specific species of functional systems. Community makeup is usually studied by analyzing phylogenetic marker genes, most notably the 16S SSU and 23S LSU rRNA coding genes. Species are identified based on sequence alignment against databases such as RDP (Cole, Chai et al. 2005; Cole, Wang et al. 2009) and species richness is estimated based on statistical tools such as rarefraction curves (Schloss and Handelsman 2005). In functional analysis the functional capabilities of the community as a whole are estimated based on the gene contents of the metagenome. The functional profile of the community is built by estimating the relative abundance of each gene or pathway, and then may be compared to the profiles of other environments in order to identify those functions most relevant to the environment as was done in (DeLong, Preston et al. 2006). Identification of genes and pathways is usually done by aligning all sequences in the metagenome to databases such as NCBI’s Clusters of Orthologs (COG) (Tatusov, Galperin et al. 2000; Tatusov, Fedorova et al. 2003), or the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000). Once the number of reads carrying each gene or pathway has been determined, it is possible to use these numbers in order to determine the relative abundance of each function.
1.2. Metagenomic projects

Roughly speaking, metagenomic projects carried out so far may be divided to two categories: surveys and system specific. Surveys usually attempt to study a given environment from aspects such as species diversity, gene and protein contents, and functionality. To-date such surveys were carried out on diverse environments such as the sea (Venter, Remington et al. 2004; Angly, Felts et al. 2006; DeLong, Preston et al. 2006; Rusch, Halpern et al. 2007; Yooseph, Sutton et al. 2007; Dinsdale, Edwards et al. 2008) acid mine drainage (Tyson, Chapman et al. 2004) and human distal gut (Gill, Pop et al. 2006). System specific projects usually investigate the functionality of and relationships between specific organisms, as in (Woyke, Teeling et al. 2006) and (Hallam, Putnam et al. 2004).

Venter et al. (Venter, Remington et al. 2004) collected DNA from the Sargasso Sea, an open-ocean nutrient-poor environment near Bermuda. Overall 1.66 million reads were collected with a total length of more than 1 Gbp. The data was assembled using the Celera assembler (Myers, Sutton et al. 2000), with a few modifications mainly aimed to overcome problems resulting from polymorphism in overabundant species (see the assembly paragraph in Section 1.1 above for more details). The authors used an evidence-based approach for identifying most likely coding regions, in which protein alignments against a non-redundant bacterial protein database served for identifying genes. More than 1.2 million genes were discovered that way. A few hundred species were estimated using phylogenetic markers such as rRNA genes and other proteins that are know to be encoded by exactly one gene across all known bacteria. Rhodopsin-like genes were found in a range of organisms that exceeds the previously known gammaproteobacteria.

The approach described in (DeLong, Preston et al. 2006) is different, and is mainly focused on identifying taxonomic, functional and community trends. The authors took samples from seven depths ranging from 10 to 4,000 meters at the ALOHA station near Hawaii. Each depth may be characterized in terms of environmental parameters such as light, salinity, temperature etc. Overall, 64Mbp were sequenced from all depths. The authors categorized sequences from each depth by comparing them to databases such as COG and KEGG. In addition, sample-specific genes were found by comparing the different samples and considering those genes that are present in one sample but lack from the other. As a result, it was possible to identify genes and traits that are unique to each depth, such as the presence of photosynthetic genes in samples taken from the photic zone only.

(Angly, Felts et al. 2006) characterized the prevalence of viruses in the world’s oceans. The data was extracted from four diverse oceanic regions: Sargasso Sea, the Gulf of Mexico, British Columbia coast and the Arctic Ocean. More than 180 Mbp of sequences were collected and sequenced using 454 sequencing (average read length ~100bp), which most likely prevented the assignment of many sequences to organisms and phages due to the very short sequences. The majority of phages were found to be widespread across several (or all) geographic areas. On the other hand, the assemblages of the different regions varied greatly.
Human distal gut, an environment in which some $10^{13}$-$10^{14}$ microbes live, was examined in (Gill, Pop et al. 2006). As in the cases of other metagenomic projects carried out in environments with high microbe diversity, a high portion of the reads (40% of a total of 45 Mbp) were not assembled and left as singletons. ORFs that were found (More than 50,000) were compared against the COG and KEGG databases in order to learn about functionality. The study of microbial communities in the different parts of the human body, such as the gut and mouth, is of great interest due to the important role of these communities in determining human health. Gut microbiome, for example, has been associated with various health conditions such as obesity (Turnbaugh, Ley et al. 2006), ulcers (Marshall and Warren 1984) and several types of stomach cancers (Parsonnet, Friedman et al. 1991). As a result, the amount of metagenomic projects in these environments is expected to grow.

When the examined environment is not diverse, it might be possible to get a near complete assembly of at least the most abundant species in the environment. Such is the case in the study of acid mine drainage (Tyson, Chapman et al. 2004), a harsh environment with extreme conditions including a very low pH (0.83) that is dominated by a few species. A total of 76.2 Mbp were collected and assembled using the JAZZ assembler. The fraction of assembled reads in this project was relatively high – more than 85% of the reads were assembled into scaffolds longer than 2 Kbp. The genomes of two of the species were almost completely recovered, and three other genomes were partially recovered.

Woyke et al. investigated the roles of the four co-occurring symbionts from the marine worm *O. Algarvensis*, a worm that lacks mouth, gut and nephiridia (Woyke, Teeling et al. 2006). In order to analyze the genomes of the symbionts, the authors first sequenced and assembled 20.4Mbp of DNA from the system using the JAZZ assembler. Next the scaffolds were assigned to their phylotype origin using a combinatory binning approach based on oligonucleotide frequency signatures. Each of the four resulting “bins” was scanned for ORFs that were annotated and used for explaining the role of each one of the symbionts.

The study of (Hallam, Putnam et al. 2004) tested the Anaerobic Oxidation of Methane (AOM) hypothesis, claiming that methane in anoxic marine sediments is consumed by consortia of archaea and bacteria that couple methane oxidation with sulfate reduction. Several bacteria and archaea species that were not previously cultivated are known to live in these regions. The authors enriched such cells from Eel river sediments and sequenced them using both WGS and fosmids. The resulting DNA reads were assembled and binned using %G+C content, achieving good separation for the two strains of ANME. One of the strains, ANME-1, was found to contain most of the genes required for methanogenic pathway as expected by the AOM hypothesis.

1.3. The GOS database

The Global Ocean Sampling (GOS) expedition (Venter, Remington et al. 2004; Angly, Felts et al. 2006; DeLong, Preston et al. 2006; Rusch, Halpern et al. 2007; Yooseph,
Sutton et al. 2007) is part of an effort taken by the J. Craig Venter institute to collect samples of DNA sequences from the oceans around the globe. The first leg of the project contains data from 51 samples collected over 8,000 km from northwest Atlantic Ocean through the Panama Canal and until the middle of the Pacific Ocean. The raw material that was collected consists of 7.7 million reads (6.3 Gbp), which were assembled to scaffolds as much as possible. ORFs were extracted both from the scaffolds and from unassembled reads, resulting with more than 6 million predicted proteins.

The samples were taken from diverse aquatic habitats with varying conditions, including salinity, water depth, temperatures and environment (e.g. coastal, open ocean etc.). The amount of data obtained from each sample ranges from 50,000 to 645,000 sequences. About 30% of the reads were matched to known genomes with high confidence ("recruited reads"), half of which matched three genera: pelagibacter, synechococcus and prochlorococcus. The assembly process was carried out using the Celera assembler (Myers, Sutton et al. 2000) with certain modifications (Venter, Remington et al. 2004) and resulted with 47% of the reads being assembled into scaffolds. Only 9% of the reads went into scaffolds that are longer than 10 Kbp which may be due to the low level of coverage achieved, probably due to the enormous diversity of organisms in the world’s oceans.

Prediction of proteins from the scaffolds and reads has been performed in several steps and using publicly available databases. The procedure contained the recognition of ORFs by start and stop codons, and the clustering of sequences into clusters of conserved sequences in order to increase the chances of correct prediction of ORFs. Overall, a total of 5,654,638 non-redundant proteins were predicted, of which 90.8% are likely to be of bacterial origin. A total of 297,254 protein clusters were created, of which 280,187 clusters are small (less than 20 members), 12,992 are medium (between 20 to 200 members) and 4,075 are large (contain more than 200 members). 3,995 medium and large clusters are composed of GOS-only sequences.

Recently the raw reads from the Indian Ocean samples was published, adding a total of 2.5Gbp in more than 2.4 million sequences from 22 samples.

1.4. *Synechococcus, Prochlorococcus and their cyanophages*

Marine *Synechococcus* and *Prochlorococcus* are two related genera that form a single group within the picophytoplankton clade. Despite their small evolutionary distance (>96% similarity in terms of 16S rRNA), they have ten well-defined subgroups that display wide microdiversity.

*Prochlorococcus* is the smallest known photosynthetic organism, with a diameter ranging between 0.5 to 0.7 µm. The 19 known strains of *Prochlorococcus* are ubiquitous between the 40°N to 40°S latitudinal band, occupying the upper of 100-200m layer of the oceans in these areas (Partensky, Hess et al. 1999), which presumably makes them the most abundant photosynthetic organism on earth. The
The genome size of sequenced Prochlorococcus ranges from 1.6Mbp to 2.7Mbp and the %G+C content of most strains is relatively low. The photosynthetic properties of Prochlorococcus give it a selective advantage over Synechococcus in deep water. Its absorption characteristics have maxima in the range of 430-490nm, which is better suitable than Synechococcus’ range for absorbing photons in deep water. In addition, its photosynthetic yield is higher at all depths.

Synechococcus are rod-shaped phototrophs ranging in size from 0.8 to 3 µm. They can be found in most of the world’s oceans, but are abundant mostly in nutrient-rich coastal water; in open ocean water they are usually outnumbered by Prochlorococcus. The genome lengths of fully sequenced Synechococcus strains range from 2.2 to 3.05 Mbp, and their %G+C content is high compared to Prochlorococcus. The photosynthetic genes investigated so far in Prochlorococcus are similar to genes known from other cyanobacteria except the psbA gene encoding for the D1 protein which appears in one copy in two strains, unlike its two or more copies in Synechococcus. The presence two or more copies of psbA in Synechococcus is referred to its adaptation to high and low intensity light; the transcription level of the single copy in Prochlorococcus is proportional to the light intensity (Garcia-Fernandez, Hess et al. 1998). Prochlorococcus are divided to high-light and low-light ecotypes; these two ecotypes live in the same water column at different depths. Similarity between strains of Prochlorococcus present in different depths at the same location is higher than strains located in geographical proximity (Moore, Rocap et al. 1998).

Cyanophages infecting Synechococcus and Prochlorococcus belong to three families based on morphological characteristics: myoviridae, podoviridae and siphoviridae. Most myoviruses that were discovered to-date are polyvalent (broad-range) and infect Synechococcus and low-light Prochlorococcus (Lu, Chen et al. 2001; Marston and Sallee 2003). Podoviruses, on the other hand, are usually host-specific and infect mostly high-light Prochlorococcus (Sullivan, Waterbury et al. 2003). Only a few exceptions have been reported, including host-specific podoviruses, myoviruses and siphoviruses in the nutrient rich Chesapeake bay that attack estuarine Synechococcus (Wang and Chen 2008). Reported genome sizes of myoviruses range between 150 Kbp to 250 Kbp, in comparison to 44 Kbp in podoviruses (Clokie and Mann 2006).

To date, the genomes of only six cyanophages were completely sequenced, of which two are podoviruses and four are myoviruses. Their morphology and gene content is T7-like (podoviruses) and T4-like (myoviruses) (Mann, Clokie et al. 2005; Sullivan, Lindell et al. 2006). Several genes involved in cyanobacterial photosynthetic activity, such as the psbA and the psbD were found in cyanophage genomes [Mann, 2003] (Lindell, Sullivan et al. 2004; Sullivan, Lindell et al. 2006; Wang and Chen 2008) and were shown to be transcribed and translated in the infected host(Lindell, Jaffe et al. 2005). (Sullivan, Lindell et al. 2006) examined 33 cultured cyanophages and found the psbA gene (via PCR) in the majority of the genomes, including all myoviruses and podoviruses infecting Prochlorococcus, but not in siphoviruses infecting
Prochlorococcus and Podoviruses infecting Synechococcus. Half of the genomes contained both the psbA and psbD genes, almost all of them are polyvalent.

1.5. Organization of this work

The rest of this dissertation is organized as follows. In Chapter 2 the statistics of metagenomics functional analysis with respect to both gene families and pathways is discussed. Chapter 3 focuses on methods for studying gene organizations as well as for classifying metagenomic data into viral and microbial bins. While the discussed methods are mostly heuristic, they were shown to be most effective in analyzing metagenomics data. Chapter 4 contains a description of the discovery of Photosystem I gene cassettes on the genome of viruses, while Chapter 5 contains a summary of the work and a discussion.
Chapter 2.
The Statistics of Metagenomics Functional Analysis

Metagenomic studies present us with new computational challenges that are quite different from those of classical genomics. In most cases the amount and coverage of sequence data is insufficient to ensure assembly and classification of sequences into different microbial populations, thus preventing even limited population-specific genomic and metabolic reconstruction. Consequently, a prevalent method for analyzing metagenomic datasets is to compare the relative frequencies of gene families between datasets to highlight over- and under-represented functions in a given microbial community (Tringe, von Mering et al. 2005; Rodriguez-Brito, Rohwer et al. 2006). Such comparisons require a measure of confidence in the observed differences in gene family frequencies between metagenomic datasets, which are usually based on statistical tests. Statistical tests that have been applied to metagenome dataset comparisons are based on re-sampling (Tringe, von Mering et al. 2005; Rodriguez-Brito, Rohwer et al. 2006; Warnecke, Luginbuhl et al. 2007). These methods may produce reliable results but re-sampling does not scale well computationally with increased dataset sizes.

Recall that DNA sequencing – for both metagenomic and conventional genomic studies – is done using a process known as Whole Genome Shotgun (WGS) sequencing (Sanger, Coulson et al. 1982; Fleischmann, Adams et al. 1995; Venter, Adams et al. 2001). In Sanger Sequencing the process consists of the following steps. First, the extracted DNA is sheared randomly using physical or chemical means to fragments ranging in size between a few kilo base pairs (Kbps) to a few dozen Kbps. Next, the fragments are inserted into vectors, which are used for constructing a genomic library. Last, clones from the library go through pair-end sequencing in which approximately 800bps from each side of each clone are sequenced using Sanger sequencing. Next Generation Sequencing (NGS) technologies do not require a cloning phase but rather work on the raw DNA directly. Current read lengths range between a few dozen base pairs (Illumina Solexa, ABI SOLiD) to 500 bps (454 Life Sciences), In both NGS and Sanger sequencing the outcome of the process is a set of many pair-end reads that can be used for all types of analysis including assembly, binning, and gene finding. Lander and Waterman (Lander and Waterman 1988) developed a statistical model for single-genome WGS sequencing projects that makes it possible to estimate parameters such as the expected number of contigs (continuous assembled reads) as a function of coverage depth, where coverage is the average number of reads covering each position in the genome.
Functional analysis of metagenomic data provides valuable insight into the kind of biological functions that are predominantly performed by organisms in a given environment $E$. One way of performing such an analysis is by aligning the metagenomic sequence data (also referred to as a "the metagenome") against databases such as COG (Tatusov, Galperin et al. 2000; Tatusov, Fedorova et al. 2003) or Pfam (Finn, Tate et al. 2008) and identifying the most abundant protein families in it (note that we use the terms "gene family" and "protein family" interchangeably, as implied by context). Once the frequency of each protein family in the metagenome has been estimated, it is possible to identify those families most essential to the survival of certain species in $E$ by comparing them with the frequencies observed in other metagenomes taken from similar or different environments. This process, termed function comparison, has become a common routine in metagenomic works (Tringe, von Mering et al. 2005; DeLong, Preston et al. 2006; Martin-Cuadrado, Lopez-Garcia et al. 2007; Warnecke, Luginbuhl et al. 2007) and has proven to be very useful. For example, DeLong et al. (DeLong, Preston et al. 2006) collected metagenomic samples from seven depths ranging between 10 to 4000 meters from a Pacific Ocean station near Hawaii. By performing function comparison on the seven samples using the COG database, the authors were able to identify gene families that are characteristic of certain depths, such as photosynthetic-related genes that are most abundant in shallow water. Other families had significant representation in all depths, suggesting that they represent functions required by a wide range of microbes regardless of their environment. This illustrates the importance of comparing metagenomes: high abundance of a gene family does not necessarily imply relevance to specific conditions in the environment from which it was obtained. By performing function-comparison it is possible to differentiate between environment-specific and environment-independent functions.

Frequency estimation of gene families is usually done using a process we term the read-counts approach. The frequency of a family $P$ is given by the number of reads carrying members of $P$ divided by the number of reads carrying members of any other family $P'$ (see Section 2.1 below). While being simple and straightforward, this approach "favors" families composed of longer genes and assigns them higher frequencies. In order to see why, consider two equally abundant gene families $P_1$ and $P_2$ in which members of $P_1$ are on average twice as long as members of $P_2$. Assuming that reads are sampled and sequenced uniformly across all positions in all genomes in the environment, then the above process should yield more reads containing portions of $P_1$ than reads containing portions of $P_2$ as a direct effect of the difference in lengths. In the hypothetical event of reads of length 1 (single base pair) the estimated frequency of $P_1$ should be twice the estimated frequency of $P_2$. This bias, known as the read-counts bias, is not negligible, as the length of genes may vary between a few dozen base pairs (e.g. tRNA genes) and a few thousand base pairs (e.g. the photosynthetic genes psaA and psaB and many other genes).

The scope of this chapter is functional analysis at both gene family and pathway levels. The chapter begins with a description as for the statistics of functional
analysis at the gene family level in which each family represents a single functionality based on proteins (e.g. COG, (Tatusov, Galperin et al. 2000; Tatusov, Fedorova et al. 2003)) or protein domains (e.g. Pfam (Finn, Tate et al. 2008) and TIGRfam (Haft, Selengut et al. 2003)). We present a statistical framework for estimating family frequencies that is based on the abovementioned Lander-Waterman model and explain how derived frequencies may be used for functional comparison of metagenomic datasets. In addition, we provide a novel method for assessing reliability of computed frequencies, which can be used for removing seemingly unreliable measurements. Next we describe functional analysis at the pathway level. We present two models that take different assumptions and compare their performance to the currently used read-count based method. The chapter concludes with evaluation results as for the methods presented in this chapter. As for the analysis at the gene-family level, we have tested our method on both simulated and real WGS sequencing data from projects of whole genomes; our tests indicate a substantial improvement of our method over existing ones. The methods for pathway-level analysis were tested on synthetic data and proved to significantly improve estimations produced by the read-count model.

The contribution of this chapter is 3-fold: (i) correction of the read-counts bias which is present in all methods that have been suggested to-date; (ii) for the first time a complete theoretical statistical framework with reasonable assumptions is being suggested with respect to gene-family level analysis, and (iii) the most extensive testing performed to-date, both on synthetic as well as (again, for the first time) on real data.

2.1. Previous work

Functional characterization of metagenomic data involves (i) identifying protein CoDing Sequences (CDSs) in unassembled or partially assembled metagenomic sequences using an ab initio or evidence-based gene finder, then (ii) associating these CDSs with protein families, such as COGs, Pfams, and TIGRfams, and subsequently (iii) comparing the relative abundance of protein families. Protein coding sequences are associated with protein families using BLAST against sequence databases such as COG, reverse position-specific BLAST (RPS-BLAST) against position specific scoring matrices (PSSMs, e.g. the CDD database (Marchler-Bauer, Anderson et al. 2009), or using Hidden Markov Models (HMMs, e.g. the Pfam and TIGRfam databases).

Several methods for function comparison have been proposed to date (DeLong, Preston et al. 2006; Rodriguez-Brito, Rohwer et al. 2006; Warnecke, Luginbuhl et al. 2007; Schloss and Handelsman 2008). They differ in the way they perform function comparison, but compute the frequency for a protein family \( P \) from the proportion of reads in a metagenomic sample \( M \) that are associated with \( P \) when \( M \) is compared to a function database using BLAST (Altschul, Gish et al. 1990). Observe that in such an approach, each read may be associated with multiple protein families, and hence, counted several times. We refer to this approach as the read-counts frequency computation approach; as mentioned above, this method tends to overestimate the
frequencies of longer genes. The method described in (DeLong, Preston et al. 2006) and (Tringe, von Mering et al. 2005; Rodriguez-Brito, Rohwer et al. 2006) begins with the assessment of frequencies for each family as described above, and then computes the distance between the frequencies of each protein family in the two metagenomes using simulations. The computation of the $p$-value for this distance is also done using simulations. The method is supposed to be assumption-free, but the simulation process described is – in fact – a binomial one. In addition, the generation of the random distribution is somewhat ad-hoc. Finally, the simulations may make the process computationally intensive, possibly unnecessarily since the distributions may be computed directly considering the fact that the process is binomial.

The method employed in the IMG/M system (Markowitz, Szeto et al. 2008) is also based on the computation of family-frequencies using the read-counts approach, but differs in its function comparison procedure. Given a protein family $P$, and two metagenomes $M_1$ and $M_2$ for which the respective frequencies $f_1$ and $f_2$ of association of $P$ with reads has been computed, the method first computes the distance between $f_1$ and $f_2$ and an associated $p$-value for the distance. The $p$-value computation is based on the null hypothesis that the raw counts of occurrence of $P$ among reads in both $M_1$ and $M_2$ can be approximated by a binomial distribution whose Bernoulli probability is computed as a pooled probability from the counts of occurrence of $P$ in both metagenomes. Such a computation is biased towards the frequency of $P$ in the larger metagenome, which is undesirable. Also the use of a common source distribution is likely to reduce the significance of the difference between the two metagenomes. The work described in (Schloss and Handelsman 2008) is a set of new statistical methods for comparing communities. Rather than using databases such as COG or Pfam, the authors employ clustering algorithms for putting together related proteins. This approach may extend functional analysis beyond known genes at the expense of the reliability of the results.

2.2. Functional analysis at the gene family level

In this section we propose a statistical model for computing the frequency of a protein family based on the number of reads on which it was detected in a given metagenome. This will serve as a more accurate replacement of the binomial proportion that has been widely used in the scientific literature to model protein family frequencies. The proposed statistical model derives its roots from the Lander-Waterman theory (Lander and Waterman 1988) on the statistics of WGS sequencing projects. The model is first described with respect to single genomes (Section 2.2.1) and then further developed in order to make it suitable for metagenomes (Section 2.2.3).

2.2.1. Estimating the frequency of a protein family

Consider a WGS sequencing project in which $N$ independent clones ($2N$ pair-end reads) of average read length $L$ are sequenced randomly from a genome of length $\Gamma$. Assume that each clone may start at any position in the genome with equal
probability $\alpha = \frac{N}{\Gamma}$. This description is consistent with (Lander and Waterman 1988) and may be extended for metagenomes as will be shown later in Section 2.2.3. The number of clones starting at each position is a Poisson-distributed random variable with mean $\alpha$. The probability of $j$ clones starting at a given position is:

$$f(j; \alpha) = \text{Poisson}(\alpha) = \frac{\alpha^j \cdot e^{-\alpha}}{j!} \quad (1)$$

Let $G$ be a gene with length $g$. Assume that a read must contain at least $T$ base pairs, $L>T$, of $G$ in order for $G$ to be detected. First, we estimate the number of reads containing a detectable part of $G$. Then, there are $2(L+g-2T)$ positions at which a clone might begin (see Figure 1) in order for $G$ to be detected on one of its pair-end reads.

![Diagram](image)

**Figure 1.** (a) Any read of length $L$ that begins within the $L-T$ base pairs before the target gene or somewhere on the first $g-T$ base pairs on the gene will cover at least $T$ base pairs of the gene. Overall there are $L+g-2T$ positions in which such reads may begin. (b) Each clone may cover $G$ with either its forward or backward reads. Combined with (a) this gives a total of $2(L+g-2T)$ positions.

The number of reads containing a detectable part of $G$ can be represented by a random variable $R_G$ that is the sum of $2(L+g-2T)$ independent Poisson variables with mean $\alpha$. $R_G$ is therefore Poisson distributed with mean $\lambda = 2\alpha(L+g-2T)$.

The function analysis process is based on identifying genes that belong to gene families that are captured as COGs or Pfams. Each such family contains several genes whose lengths are usually similar. Let $P$ be a protein family composed of genes whose average length is $g$. In the genome of length $\Gamma$, suppose that $\mathbb{C}_P$ genes are associated with $P$. Assuming that the occurrences of the genes associated with $P$ are independent of each other, the number of reads associated with such genes can be modeled by a random variable $R_P$, which is the sum of $\mathbb{C}_P$ independent Poisson variables, each with mean $\lambda_P$. As deduced above, $R_P$ is Poisson distributed with mean

$$\lambda_p = 2\alpha(L + g_P - 2T) \cdot \mathbb{C}_P \quad \text{(2)}$$
The numbers resulting from BLASTing the metagenome reads against the protein families database yield an empirical estimate for $R_P$, while we are interested in estimating $C_P$ for computing the frequency of $P$ from all genes in the genome of length $L$. From Equation (2) we have,

$$C_P = \frac{\lambda_P}{2\alpha \cdot (L + g_P - 2T)} \quad (3)$$

Substituting $\lambda_P$ with $R_P$, we get an estimator $\hat{C}_P$ for $C_P$ when the value of $R_P$ is known as follows:

$$\hat{C}_P = \frac{R_P}{2\alpha \cdot (L + g_P - 2T)} \quad (4)$$

In the case of metagenomes, all the required parameters for computing $C_P$ are available, except for the total length $L$ of the genome, required for the computation of $\alpha$.

Let $D$ be the set of protein families such that reads are associated with members of $D$. Then, the proportion of reads, $F_P$, associated with a protein family $P$ can be approximated by dividing $\hat{C}_P$ by the total number of genes belonging to any gene family:

$$F_P = \frac{\hat{C}_P}{\sum_{Q \in D} \hat{C}_Q} = \frac{R_P}{2\alpha \cdot (L + g_P - 2T)} \sum_{Q \in D} \frac{R_Q}{2\alpha \cdot (L + g_Q - 2T)} = \frac{R_P}{(L + g_P - 2T)} \sum_{Q \in D} \frac{R_Q}{L + g_Q - 2T} \quad (5)$$

Where $g_Q$ is the average length of genes in protein family $Q$.

Once computed, $F_P$ can be used for function analysis. As seen in Equation (5), $\alpha$ is eliminated from the expression for $F_P$ resulting in an expression comprising only known parameters. The denominator in this case covers all genes in the genome and should be accurate enough for WGS projects of reasonable size; the numerator depends on the observed number of reads for the gene family where higher read count means more accurate frequency estimation. A high read count is the result of longer genes and a higher number of occurrences of the gene family in the genome.

Next we provide a method for estimating the accuracy of the observed frequency.

### 2.2.2. Computing confidence bounds

The estimation of $F_P$ is based on the observed number of reads covering members of protein family $P$. In the event of significant inaccuracy in this estimation, later stages
such as function comparison will also be affected and will yield incorrect conclusions. Here, we provide a method for estimating a range of possible frequencies for the gene family, which may have generated the observed counts with probability higher than some user-defined threshold $\varepsilon$. For abundant families this range is expected to be narrow, while for rare families this range is going to be wide.

Specifically, we need to compute $F_p^{\text{min}}$ and $F_p^{\text{max}}$ such that $\Pr[R_p|f] \geq \varepsilon$ for every $f \in [F_p^{\text{min}}, F_p^{\text{max}}]$, and there exists no other $f'$ such that $f' \notin [F_p^{\text{min}}, F_p^{\text{max}}]$ and $\Pr[R_p|f'] \geq \varepsilon$. These upper and lower bounds on the frequencies can then be used for filtering gene families for which frequencies have ranges that are too wide.

We compute the interval in the following manner. Start with a Maximum Likelihood estimation $\hat{\lambda}_p = \text{Observed}(R_p)$ for the parameter of the Poisson distribution in Equation 3. Iteratively look for factors $C_p^{\text{min}}$ and $C_p^{\text{max}}$ such that $\Pr[R \geq R_p|C_p^{\text{min}} \cdot \hat{\lambda}_p] < \varepsilon$ and $\Pr[R \leq R_p|C_p^{\text{max}} \cdot \hat{\lambda}_p] < \varepsilon$. From Equation (2) it follows that

$$\hat{\lambda}_p \cdot \alpha = \frac{R_p}{2 \cdot (L + g_p - 2T)}$$

(6)

Therefore, multiplying $\hat{\lambda}_p$ by a constant $c$ is equivalent to multiplying $C_p$ by $c$, since $\alpha$ is not changed, and a simple multiplication of the frequencies as described here is sufficient.

### 2.2.3. Transition from genomes to metagenomes

So far we have discussed genomes, while, in fact, we are interested in analyzing metagenomes. In this section we show that Equation (5) can be used for metagenomes as well. In order to see why, consider a metagenomic sample $S$ containing $m$ different species, where species $i$ has a genome of length $\Gamma_i$ and is represented by $n_i$ members. As a first step in the WGS sequencing process all genomes in $S$ are sheared into clones; next, $N$ clones are chosen at random and undergo pair-end sequencing. Overall, there are $n_i \Gamma_i$ base pairs in $S$ associated with the genome of species $i$, and the total length of all the genomes in $S$ is $\sum_{j=1}^{m} n_j \Gamma_j$. As in the case of a single genome it is assumed that a clone may begin at each position on the genome of any organism in $S$; therefore, assuming that a total of $N$ clones undergo pair-end sequencing, the expected number of clones extracted from the genome of species $i$ is $N \cdot n_i \Gamma_i / \sum_{j=1}^{m} n_j \Gamma_j$. From Equation (1) it follows that the expected number of clones beginning at each position on the genome of species $i$ is

$$\alpha_i^S = \left( \frac{N \cdot n_i \Gamma_i}{\sum_{j=1}^{m} n_j \Gamma_j} \right) \cdot \frac{1}{\Gamma_i} = \frac{N \cdot n_i}{\sum_{j=1}^{m} n_j \Gamma_j}$$

(7)
Given species $i$ with genome of length $\Gamma_i$ and $C_P^i$ genes of average length $g$ on this genome that are associated with protein family $P$, it follows from Equation (2) that the number of reads that cover genes associated with $P$ in genome $i$ ($R_P^i$) is a Poisson random variable with mean

$$\lambda^i_P = 2\alpha^i_L(L + g_P - 2T) \cdot C_P^i$$  \hspace{1cm} (8)$$

The total number of reads covering genes associated with $P$ anywhere in the sample, $R_P^S$, can now be expressed as the sum of $m$ Poisson random variables. Using Equation (7), this is a Poisson random variable with mean

$$\lambda^S_P = \sum_{i=1}^{m} \lambda^i_P = 2(L + g_P - 2T) \cdot \sum_{i=1}^{m} \alpha^i_L \cdot C_P^i = \frac{2N(L + g_P - 2T)}{\sum_{j=1}^{m} n_j \Gamma_j} \sum_{i=1}^{m} n_i \cdot C_P^i$$  \hspace{1cm} (9)$$

Where $C_P^S$ is the total number of genes associated with $P$ in $S$. By replacing $\lambda^S_P$ with $R_P^S$, which is the observed number of reads covering genes associated with $P$ in $S$, we obtain an estimator $\hat{C}_P^S$ for $C_P^S$:

$$\hat{C}_P^S = \frac{R_P^S \cdot \sum_{j=1}^{m} n_j \Gamma_j}{2N(L + g_P - 2T)}$$  \hspace{1cm} (10)$$

Thus, the proportion of reads in $S$ that are associated with $P$ is given by

$$F_P^S = \frac{\hat{C}_P^S}{\sum_{Q \in D} \hat{C}_Q^S} = \frac{R_P^S \cdot \sum_{j=1}^{m} n_j \Gamma_j}{2N(L + g_P - 2T)} \cdot \frac{\sum_{Q \in D} R_Q^S \cdot \sum_{j=1}^{m} n_j \Gamma_j}{\sum_{Q \in D} 2N(L + g_Q - 2T)}$$  \hspace{1cm} (11)$$

For all practical purposes, Equation (11) is the same as Equation (5) since it does not contain any information related to $S$, such as the number of species and their genome lengths.

### 2.2.4. Performing function comparison

Given a reference metagenome $B$ and a metagenome of interest $A$, for a protein family $P$ we are interested in estimating the significance of the difference between
the estimated frequencies of $P$ in $B$ and $A$, $F_P^B$ and $F_P^A$, respectively. Here we propose a novel approach for assessing the significance of the difference between the two frequencies.

Recall that $F_P^B$ and $F_P^A$ are derived from the observed number of reads containing some part of $P$ in $A$ and $B$, $R_P^B$ and $R_P^A$, respectively. We are interested in computing $\text{equivalent}(R_P^A, B)$, the number of reads in $B$ that would yield the frequency $F_P^A$. Considering Equation (5) (or its metagenomic equivalent Equation (11)), this goal can be formulated by

$$F_P^A = \frac{\text{equivalent}(R_P^A, B)}{(L + g_p - 2T)} \left/ \sum_{Q \in D} \frac{R_Q^B}{(L + g_Q - 2T)} \right.$$  \hspace{1cm} (12)

This provides an explicit expression for $\text{equivalent}(R_P^A, B)$:

$$\text{equivalent}(R_P^A, B) = F_P^A \cdot (L + g_p - 2T) \cdot \sum_{Q \in D} \frac{R_Q^B}{L + g_Q - 2T}$$  \hspace{1cm} (13)

Since all components of Equation (13) are available, $\text{equivalent}(R_P^A, B)$ can be computed. Taking $R_P^B$ as the maximum likelihood estimation for $\lambda_P^B$, the parameter of the Poisson variable describing the distribution of reads covering members of $P$ in $B$, it is possible to compute the probability to observe at least (or at most) $\text{equivalent}(R_P^A, B)$ reads covering $P$ in $B$. This probability may be interpreted as a significance measure for the difference between the frequencies $F_P^B$ and $F_P^A$.

### 2.3. Functional analysis at the pathway level

Functional analysis at the pathway level is mainly used for two purposes: computation of pathway relative abundance, and pathway content comparison. Computing relative abundance of pathways within a single sample provides an overall view of the environment and was used in many studies and platforms (e.g. Overbeek, Begley et al. 2005; Edwards, Rodriguez-Brito et al. 2006; Dinsdale, Edwards et al. 2008; Markowitz, Szeto et al. 2008). Comparing pathways’ abundance between samples makes it possible to identify pathways that are enriched within one of the environments with respect to the other (DeLong, Preston et al. 2006; Edwards, Rodriguez-Brito et al. 2006). Derivatives of pathway content comparison may be used for clustering functionally similar environments using metrics over pathway abundances vectors (DeLong, Preston et al. 2006; Feingersch, Suzuki et al. 2010).

Pathway reconstruction is a related problem in which the most likely set of pathways in a genome or a metagenome is determined, without estimating their abundance. A commonly-used naive approach to this problem would be to collect all pathways with at least one representative in the data. However, this approach is
expected to yield an inflated list of pathways. Recently, a method called *MinPath* was described that attempts to deduce the minimal set of pathways required for supporting an observed set of functions (Ye and Doak 2009). The method uses Integer Programming for deciding whether a pathway is present, based on the observed functions. Note that in this case the relative abundance of the different functions is not taken into account, and no estimation of the relative abundance of the different pathways is done.

Here, we present two models for the functional analysis of metagenomes at the pathway level. Both models ignore pathway topology and treat pathways as gene sets. Based on the model described in Section 2.3.3 we deduce the *independent pathways* model that can be regarded as a natural extension of the previous work. Next we present the *pathway intersection* model that takes into account the co-occurrences of genes in more than one pathway (Section 2.3.4). We test both models on synthetic data and compare the results to the currently used read-count approach. Our tests focus on the abovementioned two common functional analysis tasks, namely sample comparison and the computation of relative abundance of pathways in the environment.

![Figure 2](image_url)

**Figure 2.** (a) The independent pathways model. In this model a gene that is shared among several pathways is assumed to have a copy for each pathway in which it appears. For example: $G_1$ belongs to three pathways and thus assumed to have three copies. (b) The pathway intersection model. Each gene that appears in one or more pathways is assumed to appear once. $G_3$ in this case will have a single copy, shared between $P^2, P^3$ and $P^4$. 
2.3.1. The Poisson model for computing gene family abundance

A metagenome \( M \) is a set of \( R \) sequence reads of length \( r \) each, extracted randomly with uniform probability for all positions across all genomes from some DNA sample of size \( L \) bps. A gene family \( G \) whose members’ average length is \( g \) represents a set of functionally similar genes, which can be defined, for example, via sequence similarity. COG, Pfam and other databases are often used as references for the identification of gene families in metagenomic data. We denote a collection of gene families by \( D_{GENE} \); the association between \( M \)’s reads and gene families is defined in terms of the read count, \( R_G \), representing the number of reads (out of \( N \)) carrying a detectable portion of \( G \)’s member. Recall from Section 2.2 that, assuming that the abundance of a gene family \( G \in D_{GENE} \) in the DNA pool is \( C_G \) (i.e. the DNA sample has \( C_G \) copies of genes that are members in \( G \)), the read count, \( R_G \), is Poisson distributed with mean \( \lambda_G \):

\[
\Pr(R_G = k) \sim \text{Poisson}(\lambda_G) = \frac{\lambda_G^k \cdot e^{-\lambda}}{k!}
\]

(14)

where

\[
\lambda_G = \frac{N}{\Gamma} (L + g - 2T)
\]

(15)

In this formula, \( N/\Gamma \) is the rate of read starts per base pair. The term \((L+g-2T)\) reflects the average number of starting positions for reads carrying a detectable portion of a single copy of \( G \) where \( g \) is the average length of \( G \)’s members, \( T \) is the minimum portion of a gene required to be present on a read in order to be associated with its family and \( L \) is the read length.

An estimator for a gene read count, \( \hat{R}_G \), can be computed using BLAST (Altschul, Gish et al. 1990) with a certain threshold. A Maximum Likelihood Estimate (MLE) \( \hat{C}_G \) for \( C_G \) can be calculated from Equation (15) and \( \hat{R}_G \):

\[
\hat{C}_G = \frac{\hat{R}_G}{N/\Gamma (L + g - 2T)}
\]

(16)

All parameters in this formula are known, except for \( \Gamma \), and hence an explicit calculation of gene family abundance is impossible. In Section 2.2 the above formula was used to compute frequency estimators for gene families, which is the relative abundance of a certain gene family out of the total abundance of all gene families in
the DNA sample pool (which eliminates the dependency on \( I \))^1. In this Section, we resolve the problem of the unknown DNA sample length \( I \) by computing the abundance of a gene family per organism in the sample, instead of the absolute abundance. This requires an estimation of the average genome length in the DNA sample, as shown next.

2.3.2. Estimating the average genome length in the DNA sample

The estimation of the average length of a genome is based on the known existence of a group of genes that are known to be present exactly once per genome in all bacterial species. Several known single-copy genes, such as bacterial \( rpoB, \) \( recA \) and \( gyrA \), were used as both phylogenetic markers (Mollet, Drancourt et al. 1997; Venter, Remington et al. 2004) as well as for the normalization of the abundance of genes in metagenomic samples (Venter, Remington et al. 2004; Howard, Henriksen et al. 2006; Rusch, Halpern et al. 2007; Yutin, Suzuki et al. 2007; Loy, Duller et al. 2009).

In the case of a single-copy gene \( SCG \), the number of copies in the entire DNA sample, \( C_{SCG} \), is equal to the number of organisms in the sample, \( S_0 \), hence it is possible to deduce an MLE for the average genome length based on Equation (16):

\[
\frac{\Gamma}{S_0} \approx \frac{\Gamma}{S_{SCG}} = \frac{N}{\hat{R}_{SCG}} (L + g_{SCG} - 2T)
\]

A more accurate estimation of the average genome length is achieved by averaging the estimated values for several single copy genes.

Utilizing the estimated average genome length, based on Equation (16), the abundance of a gene family \( G \) per organism in the DNA sample can be calculated as following:

\[
\frac{\hat{c}_G}{S_0} = \left( \frac{\Gamma}{S_0} \right) \cdot \frac{\hat{R}_G}{N \cdot (L + g_G - 2T)}
\]

2.3.3. Computing pathway abundance: the independent pathways model

In the context of the current analysis, a pathway \( P \) is defined as a set of gene families \( \{ G_1^p, \ldots, G_m^p \}, \ G_i^p \subseteq D^{GENE} \). Several repositories of pathways exist, for example KEGG and MetaCyc, and they can be used in this study. We denote a collection of pathways by \( D^{PATH} \).

---

1 Note that \( \lambda_G \) in Section 2.2 refers to the expected number of clone inserts whose two sides are sequenced. Here reads are assumed to be independent of each other, the adjustment to pair-end sequencing should be straightforward.
The independent pathways model assumes that all gene families within a certain pathway, \( P \), have the same number of occurrences (i.e. \( C_1^P = C_2^P = \cdots = C_m^P \)), and refer to this number of occurrences as the abundance of the pathway, \( C^P \). In this section, we assume that pathways’ abundances in an organism are mutually independent (Figure 2a). Analogously to the case of gene families, for each pathway \( P \in D^{PATH} \) our goal here is to compute the abundance of the pathway per organism, denoted by \( W^P \). Based on the latter assumptions, for each pathway \( P \) an estimation of its abundance per organism can be calculated by averaging the estimated abundance of the member gene families:

\[
W^P = \frac{\hat{C}^P}{S_0} = \left( \frac{\Gamma}{S_0} \cdot \frac{1}{m} \sum_{i=1}^{m} \frac{\hat{\bar{R}}_{g_{P_i}^r}}{N \cdot (L + g_{g_{P_i}^r} - 2T)} \right)
\]

Note that it is also possible to express the relative abundance of a pathway with respect to all other pathways in a sample by dividing \( W^P \) by the sum of \( W^P \) for all \( P' \in D^{PATH} \). In this case the estimation for the average genome length \( (\Gamma/S_0) \) is eliminated.

2.3.4. Computing pathway abundance: the pathways intersection model

Pathways – being a descriptive tool – are not necessarily disjoint modules, but rather they share common proteins. Ignoring the overlap in gene family content between pathways may lead the method of Section 2.3.3 to overestimate the abundance of pathways that share proteins with other pathways. Here we describe a second model that accounts for non-empty pathway intersections by jointly computing the abundance of all pathways within a collection of pathways.

The pathways intersection model assumes that a given pathway \( Y \) is either present or absent in an organism in the sample, where the presence of the pathway entails the presence of all of its member gene families in the organism. We denote by \( W^Y \) the random Boolean variable that represents the presence of a pathway \( Y \) in an organism. The probability that a gene family \( G \) is present in the genome of the organism is given by:

\[
P(G|W) = 1 - \prod_{Y \in D^{PATH} \mid g \in Y} [1 - P(W^Y = 1)]
\]

The abundance of \( G \) in the sample, \( C_{g_o} \), is deduced by multiplying this probability by the number of organisms in the sample, \( S_o \). Consequently the read count, \( R_{g_o} \), is Poisson distributed with the following mean:
\[ \lambda_G = \frac{N}{\Gamma} (L + g_G - 2T) \left[ 1 - \prod_{Y \in \text{PATH} \mid G \in Y} [1 - P(W^Y = 1)] \right] \] (21)

This can be computed for various estimates of the \( W \) variables, using the estimated average of the genome lengths in the sample (computed in Section 2.3.2).

Using the observed number of reads, \( \hat{R}_G \), we estimate \( P(W^Y = 1|\hat{R}_G) \) via a Markov Chain Monte Carlo (MCMC) posterior sampling. We assume a uniform prior for \( P(W) \), and estimate \( P(\hat{R}_G|W^Y = 1) \) using Equation (14), with \( \lambda_G \) given by Equation (21). The average of the obtained samples is used as the estimated posterior probability for the presence of each pathway in an organism.

### 2.4. Performance evaluation for the gene family model

In order to analyze the performance of our model we have used three types of data: (i) simulated data of single genomes, (ii) real data from genome sequencing projects, and (iii) real metagenomic data. Simulated data is generated according to our assumptions, and therefore provides an opportunity to evaluate the method under controlled conditions. In addition, it is possible to generate different datasets with varying parameters such as gene family distributions, coverage, and population structure. As explained above, single genomes may be regarded as a special case of metagenomes; they are simpler to analyze and generate and were therefore used for evaluation. Genome sequencing projects provide the opportunity to test our framework on real data (generated in a similar way to metagenomes) with known gene family statistics. Real metagenomes lack information regarding the underlying statistics but provide us with the opportunity to check real-world cases. Evaluation done using such data is mostly qualitative.

#### 2.4.1. Evaluation of simulated data

We have written a program that generates simulated genomes and WGS sequencing projects, including genes and their families and reads' distribution across the genome. The simulator also implements our statistical framework, as well as the read-counts based frequency estimator for protein families. In order to compare the performance of the two frequency estimators we have used the following quality measure:

\[ Q(M(P)) = \log_{10} \frac{M(P)}{F_p^{\text{true}}} \] (22)

in which \( F_p^{\text{true}} \) is the true frequency of protein family \( P \) in the (meta)genome and \( M(P) \) is the estimation. When \( M(P) = F_p^{\text{true}} \) it is the case that \( Q(M(P)) = 0 \), while \( Q(M(P)) < 0 \) when \( M(P) \) underestimates \( F_p^{\text{true}} \), and \( Q(M(P)) > 0 \) when \( M(P) \) overestimates \( F_p^{\text{true}} \).
As a first step we have tested the read-counts bias. For this purpose we have synthesized three genomes of length 10 Mbps each: (i) a genome that contains genes associated with 200 protein families of lengths ranging between a few dozens to a few thousands bps and a similar number of genes associated to each family, (ii) a genome containing genes associated with several hundreds of protein families of constant length but a different number of genes associated with each family, and (iii) a combination of (i) and (ii) – genomes with protein families of varying lengths and varying number of associated genes. Other than the length and the number of genes associated with the protein families all other parameters remained constant across all simulated genomes, including the number of reads (1M) and their lengths (942 bps) and the distance between genes (60 bps). Note that the resulting coverage is extremely high; this was done in order to test the behavior of each method when all data is available.

Figure 3. Quality of frequency estimations as a function of (a) gene size, (b) family size, and (c) a combination of both. In all cases coverage is extremely high (94.2X) in order to demonstrate the inherent differences between the methods even when redundant data is available.

Figure 3 summarizes the results of these three tests. As can be seen from Figure 3a, our framework scales with any gene length while read-counts based estimations tend to overestimate the frequencies of long genes and to underestimate the frequencies of short genes. When all genes across the genome are of exactly the same length both methods yield the same estimations, regardless of differences in the number of genes associated with each family (Figure 3b). This is not surprising considering the fact that when \( g_i = g_j \) for any two families \( i \) and \( j \), Equation (5) is reduced to simple division of read-counts with no other scaling required. As can be seen from Figure 3b, frequency estimations are less accurate for small families, which is also reasonable. Our method remains reliable also when both the number of genes associated with a family and their lengths vary (Figure 3c), while the read-counts approach produces inaccurate estimations. The accuracy of the method increases when more bps are associated with family \( P \). A large number of bps associated with family \( P \) may be the result of both long genes, or a large amount of genes associated with the family.
Next, we were interested in analyzing the influence of different coverage levels on the quality of the estimations. For this purpose we have used the setup of the genome whose behavior is described in Figure 3c and tried a different number of reads each time (resulting in different coverage levels). At high coverage levels typical to genome sequencing projects (Figure 4a) our method remains relatively accurate. As expected, the quality of the estimations decreases with lower coverage levels (Figure 4, b and c). In all cases our method outperforms the read-counts estimator and provides significantly better results (see Figure 5b).

Figure 4. The influence of coverage on frequency estimation for (a) coverage=9.42X, (b) 0.942X and (c) 0.0942X.

Figure 5. (a) Reliability bounds as a function of the number of base pairs associated with protein families, $\varepsilon=0.1$ used. (b) Log-ratio error as a function of the number of base pairs associated with different families. For each range $[x \cdot 10\text{Kbps}, (x+1) \cdot 10\text{Kbps}]$ median log-ratio error for all families in the range is shown. Confidence bounds become narrow with the increase in the family's share on the genome, in line with the increased accuracy of frequency estimation.

Figure 4a shows confidence bounds computed for different families as a function of the number of bps associated with the family. As can be clearly seen, the confidence range decreases (in terms of standard deviation units) with the more associated bps available. This property of the confidence bounds scheme is desirable, considering the fact that our method derives better frequency estimations for protein families to which many bps are associated across the genome (Figure 4b). This can also be used for filtering families with a wide confidence range.
2.4.2. Evaluation of data derived from WGS sequencing projects

Raw data of WGS sequencing projects of single genomes is available through NCBI’s Trace DB. Using data from (Sorek, Zhu et al. 2007) we were able to reconstruct read coordinates and also to identify pairs of reads that belong to the same clones. Overall we were able to reconstruct information for 44 genomes (see Table 1). COG families were assigned to these genomes by BLASTing (blastx) their genes (extracted from the IMG system (Markowitz, Szeto et al. 2008) against the COG database with an e-value threshold of 1e-50. Next the number of reads containing members of each COG family was decided and used for producing frequency estimations by both our and the read–counts approach. Evaluation of estimations was done using Equation (22) as before. Results are summarized in Figure 6 (see figure legend for description of figure generation process). As can be observed from the figure, behavior of both estimation methods fit nicely the behavior that was predicted in our simulations. While our method generated predictions that are, on average, close to the real frequencies and are not affected by the length of family members, the read-counts approach produces estimations that are biased by gene length. Both methods are not affected by the number of family members.

It is important to note that while our method produces unbiased estimations, its frequency estimations are, on average, 20-30% off the real value for all gene lengths. The estimation error is affected by the number of base pairs associated with a family on the genome (more base pairs yield better estimations). This error rate is smaller than or equal to the error of the read-counts approach for all cases.

![Figure 6](image_url)  
**Figure 6.** Summary of data analysis for 44 single genome WGS projects. (a) Log-ratio between estimations and true frequencies, sorted by family length. For each genome, average and standard deviation of gene lengths were computed and used for binning evaluations into x0.5 standard deviation bins. Next the median for each bin was computed and kept with medians of the same bin from other genomes. Last the median over all medians for each bin was computed and used for generating the figure. (b) Same process as before using number of members for each family as the key. (c) Same process, using number of base pairs occupied by each family on the genome as the key.
Table 1. Information for participating genomes in the real data analysis.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome Size</th>
<th>Coverage</th>
<th># Clones</th>
<th>Avg. Read Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acidobacterium sp. Ellin 345</td>
<td>5650368</td>
<td>11.1</td>
<td>27373</td>
<td>1035.9</td>
</tr>
<tr>
<td>Alkalilimnicolaehrichei MLHE-1</td>
<td>3275944</td>
<td>12.8</td>
<td>16913</td>
<td>1014.4</td>
</tr>
<tr>
<td>Anabaena variabilis ATCC29413</td>
<td>6365727</td>
<td>17.1</td>
<td>36406</td>
<td>934.8</td>
</tr>
<tr>
<td>Anaeromyxobacterdehalogenans 2CP-C</td>
<td>5013479</td>
<td>12.0</td>
<td>15691</td>
<td>1007.5</td>
</tr>
<tr>
<td>Bacillus anthracisstrames</td>
<td>5227293</td>
<td>20.6</td>
<td>38924</td>
<td>856.4</td>
</tr>
<tr>
<td>Baumanniacicadellinicola</td>
<td>686194</td>
<td>29.0</td>
<td>2394</td>
<td>980.0</td>
</tr>
<tr>
<td>Campylobacter jejuni rm1221</td>
<td>1777831</td>
<td>12.2</td>
<td>9316</td>
<td>881.1</td>
</tr>
<tr>
<td>Carboxydothermushydrogenoformans z-2901</td>
<td>2401520</td>
<td>14.8</td>
<td>12188</td>
<td>910.7</td>
</tr>
<tr>
<td>Chlorobium tepidum tls</td>
<td>2154946</td>
<td>13.8</td>
<td>10995</td>
<td>814.2</td>
</tr>
<tr>
<td>Colwellia psychrerythraea 34h</td>
<td>5373180</td>
<td>11.5</td>
<td>27652</td>
<td>902.2</td>
</tr>
<tr>
<td>Dechloromonas aromatica RCB</td>
<td>4501104</td>
<td>24.9</td>
<td>59247</td>
<td>850.7</td>
</tr>
<tr>
<td>Dehalococcoidesetehogenes 195</td>
<td>1469720</td>
<td>16.5</td>
<td>4861</td>
<td>907.3</td>
</tr>
<tr>
<td>Ehrlichia canis Jake</td>
<td>1315030</td>
<td>23.1</td>
<td>8229</td>
<td>968.3</td>
</tr>
<tr>
<td>Ehrlichia chaffeensis str Arkansas</td>
<td>1176248</td>
<td>14.6</td>
<td>7631</td>
<td>925.8</td>
</tr>
<tr>
<td>Frankia sp. Cl13</td>
<td>5433628</td>
<td>15.5</td>
<td>29142</td>
<td>990.6</td>
</tr>
<tr>
<td>Geobacter sulfurreducenspca</td>
<td>3814139</td>
<td>14.2</td>
<td>18579</td>
<td>916.6</td>
</tr>
<tr>
<td>Listeriamonocytogenesstr 4b f2365</td>
<td>2905187</td>
<td>17.7</td>
<td>21550</td>
<td>929.7</td>
</tr>
<tr>
<td>Methanococcus burtonii DSM6242</td>
<td>2628784</td>
<td>23.5</td>
<td>21032</td>
<td>989.5</td>
</tr>
<tr>
<td>Methanococcus capsulatus str bath</td>
<td>3304561</td>
<td>11.2</td>
<td>16517</td>
<td>884.3</td>
</tr>
<tr>
<td>Moorella thermoacetica ATCC 39073</td>
<td>2628784</td>
<td>23.5</td>
<td>21032</td>
<td>989.5</td>
</tr>
<tr>
<td>Neorickettsias sennettsustrmiyama</td>
<td>8590067</td>
<td>13.7</td>
<td>4996</td>
<td>917.8</td>
</tr>
<tr>
<td>Nitrobacter hamburgensis X14</td>
<td>4406967</td>
<td>12.5</td>
<td>20259</td>
<td>935.3</td>
</tr>
<tr>
<td>Nitrobacterwinogradskyi Nb-255</td>
<td>3402093</td>
<td>11.3</td>
<td>14703</td>
<td>958.4</td>
</tr>
<tr>
<td>Nitrosococcus oceani ATCC 19707</td>
<td>3481691</td>
<td>14.4</td>
<td>20047</td>
<td>970.0</td>
</tr>
<tr>
<td>Prochlorococcus marinus sp. NATL2A</td>
<td>1842899</td>
<td>11.0</td>
<td>10561</td>
<td>960.4</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. MIT 9312</td>
<td>1709204</td>
<td>13.6</td>
<td>9834</td>
<td>965.9</td>
</tr>
<tr>
<td>Prochlorococcus sp. CC9605 (oligotrophic)</td>
<td>2510659</td>
<td>17.0</td>
<td>16269</td>
<td>969.6</td>
</tr>
<tr>
<td>Pseudoalteromonas atlantica T6c</td>
<td>5187005</td>
<td>10.0</td>
<td>21681</td>
<td>1035.6</td>
</tr>
<tr>
<td>Psychrobacter cryoportalensis K5</td>
<td>3059876</td>
<td>12.2</td>
<td>13397</td>
<td>953.8</td>
</tr>
<tr>
<td>Rhodopseudomonas palustris B8</td>
<td>5513844</td>
<td>11.1</td>
<td>25015</td>
<td>1023.6</td>
</tr>
<tr>
<td>Rhodopseudomonas palustris B5</td>
<td>4892717</td>
<td>11.4</td>
<td>20142</td>
<td>1024.8</td>
</tr>
<tr>
<td>Rhodopseudomonas palustris HaA2</td>
<td>5331656</td>
<td>11.7</td>
<td>25012</td>
<td>959.5</td>
</tr>
<tr>
<td>Rubrobacter zylanophilus DSM 9941</td>
<td>3225748</td>
<td>14.5</td>
<td>14321</td>
<td>1193.9</td>
</tr>
<tr>
<td>Shewanella genitricans OS217</td>
<td>4545906</td>
<td>16.6</td>
<td>27125</td>
<td>967.2</td>
</tr>
<tr>
<td>Shewanella frigidimarina NCMB400</td>
<td>4845257</td>
<td>11.9</td>
<td>23999</td>
<td>978.6</td>
</tr>
<tr>
<td>Shewanella sp. MR-4</td>
<td>4706287</td>
<td>11.5</td>
<td>21464</td>
<td>1020.4</td>
</tr>
<tr>
<td>Streptococcus agalactiae a909</td>
<td>2127839</td>
<td>13.9</td>
<td>12135</td>
<td>914.9</td>
</tr>
<tr>
<td>Synechococcus sp. PPC 7942 (elongatus)</td>
<td>2695903</td>
<td>13.0</td>
<td>15863</td>
<td>891.1</td>
</tr>
<tr>
<td>Thermobifida fusca YX</td>
<td>3642249</td>
<td>31.1</td>
<td>39988</td>
<td>850.3</td>
</tr>
<tr>
<td>Thiomicrospira crunogena, XCL-2</td>
<td>2427734</td>
<td>18.5</td>
<td>18675</td>
<td>955.0</td>
</tr>
<tr>
<td>Thiomicrospira denitrificans ATCC 33889</td>
<td>2201561</td>
<td>12.0</td>
<td>9992</td>
<td>958.4</td>
</tr>
<tr>
<td>Treponemadenticola atcc 35405</td>
<td>2843201</td>
<td>15.6</td>
<td>16935</td>
<td>901.9</td>
</tr>
<tr>
<td>Trichodesmium erythraeum</td>
<td>7750108</td>
<td>17.3</td>
<td>49207</td>
<td>966.3</td>
</tr>
</tbody>
</table>
2.4.3. Evaluation of real data

While metagenomes are the goal of our method, it is impossible to validate the results generated from them. However, in order to get a sense of what the results look like we have compared two depths datasets (10 and 500 meters) from (DeLong, Preston et al. 2006). In order to assign genes to protein families we have used the COG database: each read was BLASTed against the COG database with the best hit being assigned to the read. The 10 meters dataset is composed of 7,842 reads whose average length is 954 bps, while the 500 meters dataset consists of 9,027 reads with average length of 971 bps. Overall a total of 2,426 genes from the 10 meters dataset have been assigned to COG families and 2,997 genes from the 500 meters were similarly assigned.

While family frequencies usually show correlation with the number of hits, there are many exceptions: for example, COG0187 was found 16 times in the 10 meters dataset but was assigned a frequency of 0.53%, less than the frequency of COG1024 with only 13 hits (0.59%). While COG1024's average length is 270 bps the length of COG0187 is 685 bps, which explains why it received more hits. The frequency of COG0085 with 17 hits in the 500m dataset is 0.35%, less than half the frequency of COG0316 (0.77%) with almost the same number of hits (18). Again, the difference is the result of gene lengths. Using the 500 meters dataset for the background distribution we were able to discover COG families whose frequency significantly differs between the two datasets. The results are not always observable by simply considering the read-counts; for example, the p-value assigned to the difference between the frequencies of COG0719 (ABC-type transport system involved in Fe-S cluster assembly, permease component) was quite significant (7·10^{-5}) despite a non-impressive difference in the read-counts (9 vs. 3 in the 10 and 500 meter datasets, respectively). Such cases convincingly demonstrate the need for a reliable statistical model when doing function analysis.

2.5. Performance evaluation for the pathway models

2.5.1. Materials

In order to test both our models we have generated five synthetic metagenomes based on simulated organisms, with different community complexities and metagenome sizes.

Generating organisms. We have generated two sets of organisms, KEGG10 and KEGG125 consisting of 10 and 125 synthetic species, respectively. First, the number of pathways and frequency of “dummy genes” (i.e. genes that do not belong to any pathway and that were chosen at random) were chosen either manually (KEGG10) or at random using a normal distribution with manually set parameters (KEGG125). Next, the simulated number of pathways was chosen at random from the KEGG database. Having done that, all genes from the selected pathways and the dummy genes were placed at random on the genome, using lengths as they appear in KEGG (for enzymes) or 1000 (for dummy genes). In addition to these genes, 3 single copy
genes, \textit{gyrA}, \textit{recA} and \textit{rpoB}, were also located randomly on each genome using their true lengths, averaged over instances from several bacterial genomes (2670, 1040 and 3520 bps, respectively). Note that our simulated data is based on the pathway intersection model (of Section 2.3.4), namely a single copy for every gene that appears in at least one pathway. Overall, the average genome length, number of pathways and frequency of dummy genes was 2.5Mbps, 57 and 75\% (respectively) for KEGG10 and 2.9Mbps, 81 and 68\% for the KEGG125 (The choice of parameters was made in accordance with metagenomes in the IMG/M system (Markowitz, Szeto et al. 2008)).

**Generating populations.** For each simulated population, a different organisms’ prevalence and a different population structure were used. Population complexity, which refers to the relative abundance among species, was either high (similar abundance for most species) or low (a few relatively dominant species, low abundance for the rest).

**Metagenome generation.** Number of reads per metagenome was manually set, read length ($r$) and minimum detectable gene portion ($T$) were set to 900 (typical of Sanger sequencing (Rusch, Halpern et al. 2007)) and 100 (corresponds to e-value $\approx 1e^{-100}$ in BLAST) base pairs, respectively. Number of reads per species is proportional to its DNA share in the population, defined as (genome length*frequency in the population)/(sum of (genome length*frequency in the population) over all species).

<table>
<thead>
<tr>
<th>Mtagenome</th>
<th>Organisms</th>
<th>Population complexity (% of most abundant species)</th>
<th># Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>KEGG10</td>
<td>High (10%)</td>
<td>100,000</td>
</tr>
<tr>
<td>M2</td>
<td>KEGG10</td>
<td>Low (50%)</td>
<td>100,000</td>
</tr>
<tr>
<td>M3</td>
<td>KEGG125</td>
<td>High (1.4%)</td>
<td>100,000</td>
</tr>
<tr>
<td>M4</td>
<td>KEGG125</td>
<td>Low (10.8%)</td>
<td>100,000</td>
</tr>
<tr>
<td>M5</td>
<td>KEGG125</td>
<td>Low (10.8%)</td>
<td>10,000</td>
</tr>
</tbody>
</table>

**2.5.2. Evaluation of the different methods**

**Functional comparison.** In this test, the quality of each method with respect to pathway-based functional comparison of two samples is evaluated. Given two metagenomes \textit{M} and \textit{M'} and a method for pathway abundance estimation, the frequency of each pathway in both \textit{M} and \textit{M'} is estimated and the absolute difference between the two frequencies is computed. Next, pathways are ranked based on their differential enrichment, and the intersection between the true and estimated most differentially enriched pathways is computed for every prefix size $m$ ($\leq$100).

**Pathway abundance estimations.** For each method, pathways are ranked based on their estimated frequencies. Similarly to the case of functional comparison, we use
the number of pathways that are common to both the true and estimated $m$ most abundant pathways as a measure of quality. Hyper-geometric distribution was used in order to evaluate the statistical significance of the results. In short, the probability that the intersection between the two lists of size $m$ contains exactly $k$ pathways is given by

$$\Pr(X = k) \sim \text{Hypergeometric}(k; N, m, n) = \frac{\binom{m}{k} \cdot \binom{N-m}{n-k}}{\binom{N}{n}} \tag{23}$$

where $N$ is the total number of pathways and $n=m$ is the prefix size. The significance of the observed $k$ is given by the Hyper-Geometric Tail (HGT):

$$\Pr(X \geq k) = \sum_{i=k}^{m} \frac{\binom{m}{i} \cdot \binom{N-m}{n-i}}{\binom{N}{n}} \tag{24}$$

In addition to the above, we have also used the Pearson correlation coefficient for evaluating the degree of agreement between the lists of true and estimated frequencies.

### 2.5.3. Results

To evaluate the performance of our methods in predicting pathway abundances, we generated synthetic metagenome data with various community complexities and sizes (Section 2.5.1; Table 2). Our tests focus on two of the most interesting tasks in the context of metagenomics: (i) comparing pathways’ abundance between samples and (ii) computing relative abundance of pathways within a single sample. As a baseline, we compared the performance of our methods to that of a standard read-count approach, estimating the relative abundance of each pathway $f$ as the relative abundance of its read counts out of the total number of read counts in all considered pathways:

$$freq(f) = \frac{\text{read\_count}(f)}{\sum_{f'} f' \in F \text{ read\_count}(f')} \tag{25}$$

To evaluate the performance of the various prediction methods on the task of function comparison, we compared all pairs of metagenomes and evaluated the resulting lists (Figure 7, and Figure 10). The pathway intersection model showed superior performance over the other models in 6 out of 10 scenarios (e.g. Figure 7a). The independent pathways model performed slightly better than the read-count model in most cases. The relatively low improvement in performance in this task is somewhat expected since our models aim to correct biases in the estimation of pathway abundances introduced by differences in pathway size and gene lengths,
while these biases are largely eliminated when comparing the same pathway over two samples.

Figure 7. Agreement between true and estimated lists of most differentially enriched pathways in selected pairs of metagenomes. For each \( m \leq 100 \) (X axis), the frequency of pathways that are among the \( m \) most differentially enriched pathways in both the true and estimated ranked lists is plotted for (a) metagenomes M1 and M5, and (b) M3 and M4. Refer to the appendix for plots of all other pairs.

Since the original aim of the read-count method was to address the above task of computing changes in gene set abundances across metagenomes, it is not suitable for computing the relative abundance of pathways in a sample as it does not account for differences in pathway sizes (an inherent factor for this task). Therefore, as a baseline to assessing the performance of our methods here, we implemented a fourth method, the normalized read-count, that is based on read-counts but also account for pathway sizes. The relative abundance of a pathway in this method is given by:

\[
freq(P) = \frac{\text{read}_\text{count}(P)/\text{size}(P)}{\sum_{P' \in \text{PATH}} \text{read}_\text{count}(P')/\text{size}(P')}
\]  

To evaluate the performance of the various methods in predicting the relative abundance of pathways across a single sample, we tested the agreement between the rankings of pathways based on their true abundances and predicted abundances by the various methods (Figure 8, and Figure 11). Quite expectedly, the performance of the read-count method is significantly worse than that of the other methods. A significant improvement is achieved by the normalized read-counts, taking into account pathway sizes. The independent pathway model, accounting for variation in gene lengths, achieves an additional marked improvement. The performance of the pathway intersection method is inferior to that of the independent pathways method when ranking sets of highly abundant pathways. On the other hand, the performance of the pathway intersection method is superior to all other methods when considering pathways with lower abundances.
Figure 8. Ranking pathways based on their enrichment in metagenome M3. (a) The intersection between the true and predicted $m$ most abundant pathways using the various prediction methods (y-axis), for different values of $m$ (x-axis). (b) Statistical significance (hyper-geometric p-values) for the intersection between the true and predicted highly abundant pathway sets shown in (a). Other simulated metagenomes exhibited similar behavior (see Figure 11).

To further evaluate the performance of the various methods in predicting relative abundance of pathways across a single sample, we computed the Pearson correlation coefficient between the true and predicted relative abundances (Figure 9a). Consistent with its poor performance in the ranking test, the read-counts method shows no correlation with the true pathway abundances across all metagenomes. The independent pathways and the pathway intersection methods perform better than or equal to the normalized read counts method in all cases. In particular, the pathway intersection method outperforms the other approaches when lowly abundant pathways are considered (Figure 9b), as also shown above in the ranking tests. The success of the pathway intersection method on rare (or missing) pathways may be due to the fact that it does not do multiple counting of a gene common to several pathways while the other methods do. Frequencies assigned by the other models will be higher than the true frequencies; while this also happens with abundant pathways, its influence on rare pathways is much higher. The independent pathways model does not suffer from this bias.
Figure 9. Pearson correlation between predicted and true pathway abundances across the various metagenomes. (a) Correlations obtained for the entire set of 250 pathways used in the simulation, (b) correlations obtained for the set of 150 less abundant pathways.
Figure 10. Agreement between true and estimated lists of most differentially enriched pairs of metagenomes. Refer to the legend of Figure 7 for description. Read count (blue), independent pathways (green) and pathway intersection (red) models are compared.
Figure 11. Ranking pathways based on their abundances in metagenomes M1, M2, M4 and M5. Refer to the legend of Figure 8 for description. Read count (blue), normalized read-count (cyan), independent pathways (green) and pathway intersection (red) models are compared.
Chapter 3.

Studying Genes and Systems from Metagenomics Data

Metagenomics gain increased popularity in recent years as a method for studying microbial communities and microbes that cannot be cultivated in the lab. As a result, a large amount of metagenomic data has been accumulated on public databases such as CAMERA (Seshadri, Kravitz et al. 2007), IMG/M (Markowitz, Szeto et al. 2008), MG-RAST (Meyer, Paarmann et al. 2008) and others. The wealth of metagenomic data significantly increased our understanding as for the characteristics of microbial communities in terms of community structure and functional capabilities with respect to many environments. However, metagenomic data is seldom exploited for studying genomic-level issues such as gene organizations, and gene contents of microbes and viruses that differ from cultivated ones. The use of metagenomic data for these purposes is not trivial, due to short read length available by all currently used sequencing technologies. In most cases a single read will not contain more than one or two genes, and therefore it is hard to deduce knowledge as for the organization of a certain gene in the context of other genes than its own immediate neighbors. In addition, the origin of the read cannot always be determined; in many cases it is hard or impossible to determine even the domain from which a certain read originated, or whether it belongs to a microbe or a virus.

In this chapter I describe several protocols that were used by us for studying gene organizations and the origin of reads. I refer to these as protocols since they are mainly heuristic and do not contain a significant algorithmic component. However, these protocols have proved to be powerful in several instances (Sharon, Tzahor et al. 2007; Kagan, Sharon et al. 2008; Sharon, Alperovitch et al. 2009). The utilization of these protocols for real-life cases is described in the second part of this chapter; results are described and discussed in Chapter 4.

3.1. Studying gene organizations

Here we are interested in studying conserved or abundant gene organizations from metagenomic data. Operons are one example for such organizations of interest: genes that are part of an operon are controlled by the same molecular mechanisms and may be transcribed to a single mRNA. The organization of genes in an operon may be conserved across many genomes as in the case of the tryptophan operon discussed below, or it may differ between species. The problem of determining operons and conserved gene clusters in complete genomes has been extensively studied in the past (e.g. (Fujibuchi, Ogata et al. 2000; Ermolaeva, White et al. 2001; Pavesi, Mauri et al. 2004; Westover, Buhler et al. 2005)), however the problem becomes much more complicated when metagenomic data is considered due to the
short length of the sequences. Our semi automatic protocol for studying gene organizations has been successfully used in several cases. For example, we were able to analyze the organization of the tryptophan pathway genes on microbial genomes from the Sargasso Sea and to show that their most abundant organization is probably converted into two split operons rather than the classical single operon (refer to Section 3.3 and to (Kagan, Sharon et al. 2008) for more details). Other less prevalent and non-classical organizations of the genes were also found. In another work (Sharon, Alperovitch et al. 2009) we showed that PSI genes can be found on the genomes of cyanophages, and that their organization is conserved among cyanophages from different oceans around the world: \textit{psaF} \rightarrow \textit{psaC} \rightarrow \textit{psaA} \rightarrow \textit{psaB} \rightarrow \textit{psaK} \rightarrow \textit{psaE} \rightarrow \textit{psaD}, unlike their organization on the genomes of cyanobacteria (refer to Chapter 4 for more details).

\textbf{Definitions}

- \textbf{Gene instance} – given a gene \( g \) that is defined in terms of a DNA sequence and a unique identifier, a DNA sequence is said to contain an instance of \( g \) if some portion of \( r \) is significantly similar to \( g \). Similarity is usually determined using BLAST, and its significance is set by BLAST e-value. We mark the similarity significance threshold by \textit{eval}.

- \textbf{Neighboring genes} – \( g_1 \) and \( g_2 \) are considered to be neighbors if they are located on the same DNA strand, and the distance between them does not pass some user defined distance marked by \textit{max_dist}.

- \textbf{Gene association} – given two genes \( g_1 \) and \( g_2 \) then \( g_2 \) is said to be associated with \( g_1 \) if \( g_2/\text{neighbors}(g_1) \geq \text{sig_share} \), \text{neighbors}(g_1) \) is the set of \( g_1 \)'s neighbors from either downstream or upstream directions and \textit{sig_share} is a user defined constant. Note that the relation is not symmetric, namely “\( g_2 \) is associated with \( g_1 \)” does not imply “\( g_1 \) is associated with \( g_2 \)”.
Procedures

The first two procedures, \textit{associated\_downstream} and \textit{associated\_upstream}, are used in order to determine the sets of genes associated with a gene $g$ from both the 3’ and 5’ sides, based on the data in a metagenome $M$.

\begin{center}
\textbf{Procedure} \textit{associated\_downstream}(g, M):
\begin{enumerate}
\item \textbf{Input}: a gene $g$, a metagenome $M$.
\item \textbf{Output}: the set of genes associated with $g$ in $M$ from the 3’ direction.
\end{enumerate}
\begin{enumerate}
\item Find $M_g = \{r_1^g, r_2^g, ..., r_n^g\}$, the set of all reads in $M$ on which $g$ appears.
\item Determine the gene contents of each $r \in M_g$. Identify \textit{downstream}(g), the set of $g$’s neighbors from the 3’ direction.
\item Determine and return $\text{Associated}(g, \text{downstream}(g))$, the set of genes associated with $g$ from \textit{downstream}(g).
\end{enumerate}
\end{center}

\begin{center}
\textbf{Procedure} \textit{associated\_upstream}(g, M):
\begin{enumerate}
\item \textbf{Input}: a gene $g$, a metagenome $M$.
\item \textbf{Output}: the set of genes associated with $g$ in $M$ from the 5’ side.
\end{enumerate}
\begin{enumerate}
\item Find $M_g = \{r_1^g, r_2^g, ..., r_n^g\}$, the set of all reads in $M$ on which $g$ appears.
\item Determine the gene contents of each $r \in M_g$. Identify \textit{upstream}(g), the set of $g$’s neighbors from the 5’ direction.
\item Determine and return $\text{Associated}(g, \textit{upstream}(g))$, the set of genes associated with $g$ from \textit{upstream}(g).
\end{enumerate}
\end{center}
The next procedure, `construct_graph`, constructs the graph which contains the most abundant gene organizations in which a gene g is involved.

**Procedure** `construct_graph(g, M)`:  
**Input**: a gene g, a metagenome M.  
**Output**: a directed graph G(V, E) containing organizations in which g is involved.  
**Data structures**:  
- G(V, E) – a directed graph in which V represents a set of genes and E represents association relationships between genes.  
- front – a set of genes.

1. Create G(V, E) with V={g} and E={}. Initialize front={t}.  
2. **While** front is not empty **do**  
   a. Choose f∈front and remove it from front.  
   b. `temp=associated_downstream(f, M)`  
   c. **Foreach** gene f′∈temp **do**  
      i. If f′∉V and f′∉front then front=front∪{f′}  
      ii. E=E∪{f→f′}, V=V∪{f′}  
   d. **End foreach**  
3. **End while**  
4. Initialize front={g}.  
5. **While** front is not empty **do**  
   a. Choose f∈front and remove it from front.  
   b. `temp=associated_upstream(f, M)`  
   c. **Foreach** gene f′∈temp **do**  
      i. If f′∉V and f′∉front then front=front∪{f′}  
      ii. E=E∪{f′→f}, V=V∪{f′}  
   d. **End foreach**  
6. **End while**

The graph generated by the above procedure is manually inspected for determining true organizations.

### 3.2. Separating viral and microbial sequences

Taxonomic binning is a core problem in metagenomics in which DNA reads are classified into taxonomic groups termed bins (refer to Section 1.1 for an overview on binning). To-date several binning algorithms were proposed, most of them are capable of classifying sequences into pre-defined bins based on either sequence or DNA signature similarity. Sequence similarity based algorithms attempt to align the metagenomic data to a database of fully or partially sequenced genomes (e.g. the MEGAN software, Huson, Auch et al. 2007). The approach may provide accurate results when the binned sequence has close relative(s) in the genome database, but its accuracy and sensitivity significantly drop when this is not the case (Mavromatis, Ivanova et al. 2007; Brady and Salzberg 2009). DNA signature based approaches use oligonucleotide frequencies in order to generate profile vectors for all sequences in
the metagenome. Once generated, these vectors are compared to a set of pre-defined profile vectors representing known genomes and assigned to the closest genome (e.g. the PhyloPythia algorithm, (McHardy, Martin et al. 2007)). These algorithms tend to be more accurate, but may still fail for short read lengths. Note that most existing algorithms are, in fact, classification algorithms that require a “training set” consisting of sequenced genomes. These algorithms are likely to fail for sequences with no close relatives in the training set.

Here we describe two protocols that aim to provide a solution for a subset of the binning question: given a metagenomic dataset, we are interested in identifying all sequences that are of viral origin. The first protocol is based on gene contents of the sequence, and the second is based on similarity to databases that are known to be either viral or microbial. While the goal of these protocols is much less ambitious than general purpose binning algorithms, their main advantage lies in their abilities to classify sequences with no close relatives in existing databases. This goal is achieved by exploiting additional information that we have with respect to either the way a dataset was created, or to genes that are known to be either viral or microbial specific.

3.2.1. Gene based classification

This procedure is based on the simple assumption that a sequence containing mostly viral genes is likely to be viral. It was used in (Sharon, Tzahor et al. 2007; Sharon, Alperovitch et al. 2009) and exhibited similar results to those of other, independent methods, which were used for the same purpose. While this by itself is not a proof of correctness it does suggest that the method correctly identifies viral sequences. The method is implemented as follows:

**Procedure classify_sequences(M):**

**Input:** a metagenome M.

**Output:** \( M_{viral} \), the set of all reads of likely viral origin from M.

1. Identify all ORFs on M’s reads.
2. Generate a database composed of all proteins in NCBI’s refseq_viral and refseq_microbial databases.
3. BLAST each ORF against the combined refseq_viral/refseq_microbial database. Keep the origin of each ORF’s best hit.
4. **Foreach** \( r \in M \) **do:**
   a. Compute the ratio between viral and microbial ORF’s on \( r \). Classify \( r \) as viral if this ratio is above \( viral\_ratio \).
   b. Consider the organization of genes on \( r \), and search for organizations that are likely hint on viral origin even if the read does not pass the ratio criterion.
5. **End foreach**
6. Return all reads that were classified as viral.
Gene contents based classification should take into account the possibility that a read that contains one or more viral genes is, in fact, a prophage. Both criteria do not exclude this option completely but decrease the chances of such event. In particular, in the gene organization criterion it is possible to consider organizations such as \(<\text{viral gene}>\text{--} <\text{microbial gene}>\text{--} <\text{viral gene}>\) which are not likely to be a result of a viral genomes entering the genome of bacteria. As is usually the case with classification, using stringent criteria should increase accuracy on the account of sensitivity, and vice-versa. Therefore criteria should be considered based on application.

3.2.2. Classification by recruitment

Here we utilize databases that are known to contain either viral or microbial sequences exclusively, due to measures taken during sample preparation that are supposed to leave only DNA of one origin. Note that such measures are not 100% accurate: it is usually possible to find at least some viral DNA in microbial samples and vice-versa. Therefore, prior to using a database for this type of classification it is important to test the credibility of the database.

<table>
<thead>
<tr>
<th>Procedure classify_recruitment(M):</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> a metagenome M, viral and microbial metagenomic samples.</td>
</tr>
<tr>
<td><strong>Output:</strong> (M^{\text{viral}}), the set of all reads of likely viral origin from M.</td>
</tr>
</tbody>
</table>

1. Generate two BLAST databases: a viral database, consisting of all viral input viral dataset, and a microbial database consisting of all microbial samples.
2. BLAST all reads in M against both databases.
3. For each \(r \in M\) do:
   a. Consider all alignments from the viral database that are more than \(\text{min\_length}\) bps long with percent identity that is higher than \(\text{min\_identity}\). Compute \(\text{coverage\_viral}(r)\), as described in Section 3.2.3.
   b. Repeat the process from (a) for the microbial database, and compute \(\text{coverage\_microbial}(r)\).
   c. If \((\text{coverage\_viral}(r) > 0.1 \text{ and } \text{coverage\_microbial}(r) = 0) \text{ or } (\text{coverage\_viral}(r) / \text{coverage\_microbial}(r) > 4)\) then classify \(r\) as viral.
4. End foreach
5. Return all reads that were classified as viral.

In our work we used \(\text{min\_length} = 80\) and \(\text{min\_identity} = 85\) for databases with average read length of \(~100\) bps. The criteria in step 4c can be changed, of course.

3.2.3. Computing coverage in metagenomic datasets

Traditionally (Lander and Waterman 1988), the coverage of a genome \(G\) is defined as follows:
In other words, the coverage of a genome \( G \) is the average number reads containing each position on \( G \). This formula was used for genomic projects and is also accepted when metagenomic data is analyzed. Here we propose the following alternative definition for coverage of a genome \( G \) in a metagenome \( M \):

\[
\text{coverage}^M(G) = 100 \cdot \frac{\text{# bps of } G \text{ covered by at least one read}}{\text{length}(G)}
\] (28)

We believe that Equation (28) is more appropriate for metagenomic data due to two main reasons. First, one of the known artifacts of 454 sequencing is the generation of many copies of a single read, a few hundred copies in certain cases. Indeed we have seen such cases in the Northern Line Islands datasets; such instances will result with a significantly biased outcome when Equation (27) is used, but will have no affect on the coverage computed based on Equation (28). In addition, highly conserved regions across many genomes will also result with a biased coverage when the traditional formula is used. This problem is also solved when Equation (28) is used. Our coverage formula will rise linearly with the amount of sequencing for low coverage, but will not act so when coverage becomes high. At the extreme case of \( \text{coverage}(G) = 100 \), for example, no additional increase will be observed in the coverage even if additional sequencing is done, while the value produced by Equation (28) will continue to rise linearly. This problem rarely exists when complete genomes are considered in metagenomic data because the amount of sequencing is rarely sufficient for reaching high coverage. However, this problem may be handled by: (i) compute traditional coverage in addition to the one in Equation (28), and (ii) use the Lander-Waterman model (Lander and Waterman 1988) for deducing the traditional coverage from the one given in Equation (28). We expect this coverage to be more accurate than the one computed directly due to the reasons discussed above.

### 3.2.4. Recruitment based sequence classification: the Northern Line Islands samples

In order to implement recruitment based classification we need viral and microbial samples to which we can compare our sequences. Many viral or microbial exclusive metagenomes exist, in particular those that were analyzed in (Dinsdale, Edwards et al. 2008). We have chosen to focus on those that were collected from the Northern Line Islands in the Pacific Ocean (Dinsdale, Pantos et al. 2008) due to two main reasons. First, these samples were found to contain Photosystem I (PSI) genes from viral origin in which we were interested for the work described in Chapter 4. In addition, these samples are organized in pairs, namely one viral and one microbial sample from each site that were collected at same times. Using samples that were collected from the same sites is likely to reduce biases with respect to microbial
community composition that may result when using viral and microbial samples from different sites. For example, given that we have found a certain gene in the viral fraction at some site, it may be impossible to tell whether this gene is indeed viral if we don't have a corresponding microbial fraction from the same site. The samples were collected near four islands: Kingman and Palmyra, which are small islands with little or no human activity, and Tabuaeran and Christmas, which are larger and inhabited by a few hundreds (Tabuaeran) or thousands (Christmas) of people. Human influence on the marine environment near Christmas Island is particularly significant, due to untreated sewage and other causes. Separation between viral and microbial samples at each site was achieved by taking several measures, including the usage of different filter fractions and the purification of the viral fraction using Cesium Chloride (CsCl) that is supposed to remove all free DNA and cellular microbial material. DNA was sequenced using 454 Life Sciences sequencer (average read length of ~100bps).

Despite the use of these measures some amount of viral DNA is expected to be present at the microbial fraction, and vice versa. If some problem occurred in one of the separation steps, the amount of unwanted DNA may be significant and thus prevent the usage of the samples for classifying metagenomic sequences. In order to test the credibility of these

Table 3. Genomes with highest coverage, samples 4440036.3 and 4440037.3 (Kingman). Green: microbial genomes, orange: viral genomes, blue: plasmids.

<table>
<thead>
<tr>
<th>Genome</th>
<th>Size</th>
<th>Coverage/Viral</th>
<th>Coverage/Microbial</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Candidatus Pelagibacter ubique</em> HTCC1002</td>
<td>1322994</td>
<td>1.45</td>
<td>0.37</td>
</tr>
<tr>
<td><em>Candidatus Pelagibacter ubique</em> HTCC1062</td>
<td>1308759</td>
<td>1.41</td>
<td>0.38</td>
</tr>
<tr>
<td><em>Prochlorococcus</em> phage P-SSM4</td>
<td>178249</td>
<td>1.39</td>
<td>0</td>
</tr>
<tr>
<td><em>Prochlorococcus</em> phage P-SSM2</td>
<td>252401</td>
<td>0.57</td>
<td>0</td>
</tr>
<tr>
<td><em>Prochlorococcus marinus</em> str. MIT 9301</td>
<td>1641879</td>
<td>0.21</td>
<td>1.61</td>
</tr>
<tr>
<td><em>Prochlorococcus marinus</em> str. MIT 9215</td>
<td>1738790</td>
<td>0.21</td>
<td>1.53</td>
</tr>
<tr>
<td><em>Prochlorococcus marinus</em> str. MIT 9312</td>
<td>1709204</td>
<td>0.2</td>
<td>1.44</td>
</tr>
<tr>
<td><em>Prochlorococcus marinus</em> str. AS9601</td>
<td>1669886</td>
<td>0.2</td>
<td>1.56</td>
</tr>
<tr>
<td><em>Marinobacter aquaeolei</em> VT8</td>
<td>4326849</td>
<td>0.11</td>
<td>0.18</td>
</tr>
<tr>
<td><em>Prochlorococcus marinus</em> subsp. pastoris str. MED4</td>
<td>1657990</td>
<td>0.07</td>
<td>0.38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genome</th>
<th>Size</th>
<th>Coverage/Viral</th>
<th>Coverage/Microbial</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Dinoroseobacter shibae</em> DFL 12 plasmid pDSHI03</td>
<td>126304</td>
<td>0</td>
<td>4.15</td>
</tr>
<tr>
<td><em>Dinoroseobacter shibae</em> DFL 12 plasmid pDSHI01</td>
<td>190506</td>
<td>0</td>
<td>2.59</td>
</tr>
<tr>
<td><em>Silicibacter</em> sp. TM1040 plasmid unnamed</td>
<td>130973</td>
<td>0.15</td>
<td>2.13</td>
</tr>
<tr>
<td><em>Ruegeria</em> sp. PR1b plasmid pSD25</td>
<td>148650</td>
<td>0</td>
<td>1.63</td>
</tr>
<tr>
<td><em>Prochlorococcus marinus</em> str. MIT 9301</td>
<td>1641879</td>
<td>0.21</td>
<td>1.61</td>
</tr>
<tr>
<td><em>Prochlorococcus marinus</em> str. AS9601</td>
<td>1669886</td>
<td>0.2</td>
<td>1.56</td>
</tr>
<tr>
<td><em>Prochlorococcus marinus</em> str. MIT 9215</td>
<td>1738790</td>
<td>0.21</td>
<td>1.53</td>
</tr>
<tr>
<td><em>Prochlorococcus marinus</em> str. MIT 9312</td>
<td>1709204</td>
<td>0.2</td>
<td>1.44</td>
</tr>
<tr>
<td><em>Oceanicolabatsensis</em> HTCC2597</td>
<td>4431007</td>
<td>0</td>
<td>0.46</td>
</tr>
<tr>
<td><em>Rhodobacterales bacterium</em> HTCC2654</td>
<td>4471202</td>
<td>0.01</td>
<td>0.4</td>
</tr>
</tbody>
</table>
datasets we have used two strategies: (i) recruitment of each of the eight samples against viral and microbial genomes, and (ii) search for 16s/23s rRNA coding genes.

**Sample recruitment by sequenced genomes**

refseq_viral and refseq_microbial are two datasets available from the NCBI website which contain all publically available genomes of viruses and microbes, respectively. Assuming that a sample is of the viral (microbial) fraction, we would expect viral (microbial) genomes to get more hits than microbial (viral) genomes. Note that the genomes available in our databases may not represent correctly the microbial community in the examined environment, in which case we would expect to observe relatively low coverage even for the most abundant genomes. In this case we will need to use additional methods in order to verify the origin of DNA in the sample, e.g. 16s/23s rRNA coding genes discussed below.

**Table 4.** Genomes with highest coverage, samples 4440040.3 and 4440039.3 (Palmyra). Green: microbial genomes, orange: viral genomes, blue: plasmids.

<table>
<thead>
<tr>
<th>Genome</th>
<th>Size</th>
<th>Coverage/Viral</th>
<th>Coverage/Microbial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prochlorococcus phage P-SSM4</td>
<td>178249</td>
<td>9.72</td>
<td>0.27</td>
</tr>
<tr>
<td>Prochlorococcus phage P-SSM2</td>
<td>252401</td>
<td>5.89</td>
<td>0.03</td>
</tr>
<tr>
<td>Synechococcus phage syn9</td>
<td>177300</td>
<td>1.41</td>
<td>0</td>
</tr>
<tr>
<td>Klebsiella pneumoniae subsp. pneumoniae MGH 78578 plasmid pKPN5</td>
<td>88582</td>
<td>0.62</td>
<td>0.1</td>
</tr>
<tr>
<td>Escherichia coli DU1040 plasmid NR1</td>
<td>94289</td>
<td>0.58</td>
<td>0</td>
</tr>
<tr>
<td>Synechococcus phage S-PM2</td>
<td>196280</td>
<td>0.44</td>
<td>0.04</td>
</tr>
<tr>
<td>Photobacterium damselaeepiscicida PT99-018 plasmid p99-018</td>
<td>150157</td>
<td>0.36</td>
<td>0.06</td>
</tr>
<tr>
<td>Escherichia coli plasmid pIP1206</td>
<td>168113</td>
<td>0.33</td>
<td>0</td>
</tr>
<tr>
<td>Yersinia pestis bv Orientalis IP275 plasmid pIP1202</td>
<td>182913</td>
<td>0.3</td>
<td>0.05</td>
</tr>
<tr>
<td>Salmonella enterica subsp. Entericaserовар Choleraesuis plasmid pMAK1</td>
<td>208409</td>
<td>0.26</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genome</th>
<th>Size</th>
<th>Coverage/Viral</th>
<th>Coverage/Microbial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prochlorococcus marinus str. MIT 9301</td>
<td>1641879</td>
<td>0.08</td>
<td>16.21</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. AS9601</td>
<td>1669886</td>
<td>0.13</td>
<td>16.11</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. MIT 9215</td>
<td>1738790</td>
<td>0.12</td>
<td>15.79</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. MIT 9312</td>
<td>1709204</td>
<td>0.13</td>
<td>14.81</td>
</tr>
<tr>
<td>Alteromonas macleodi 'Deep ecotype'</td>
<td>4173873</td>
<td>0</td>
<td>4.41</td>
</tr>
<tr>
<td>Paracoccus methylutens DM12 plasmid pMTH1</td>
<td>31999</td>
<td>0</td>
<td>3.86</td>
</tr>
<tr>
<td>Prochlorococcus marinus subsp. pastoris str. MED4</td>
<td>1657990</td>
<td>0.1</td>
<td>3.83</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. MIT 9515</td>
<td>1704176</td>
<td>0.14</td>
<td>3.79</td>
</tr>
<tr>
<td>Candidatus Pelagibacter ubique HTCC1062</td>
<td>1308759</td>
<td>0.02</td>
<td>1.88</td>
</tr>
<tr>
<td>Candidatus Pelagibacter ubique HTCC1002</td>
<td>1322994</td>
<td>0.02</td>
<td>1.78</td>
</tr>
</tbody>
</table>
Table 5. Genomes with highest coverage, samples 4440280.3 and 4440279.3 (Tabuaeran). Green: microbial genomes, orange: viral genomes, blue: plasmids.

<table>
<thead>
<tr>
<th>Genome</th>
<th>Size</th>
<th>Coverage/Viral</th>
<th>Coverage/Microbial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prochlorococcus phage P-SSM4</td>
<td>178249</td>
<td>6.78</td>
<td>0</td>
</tr>
<tr>
<td>Prochlorococcus phage P-SSM2</td>
<td>252401</td>
<td>4.69</td>
<td>0.09</td>
</tr>
<tr>
<td>Synechococcus phage syn9</td>
<td>177300</td>
<td>2.25</td>
<td>0</td>
</tr>
<tr>
<td>Synechococcus phage S-PM2</td>
<td>196280</td>
<td>1.01</td>
<td>0</td>
</tr>
<tr>
<td>Limnobacter sp. MED105</td>
<td>3290061</td>
<td>0.25</td>
<td>0.01</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. MIT 9312</td>
<td>1709204</td>
<td>0.16</td>
<td>0.08</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. MIT 9515</td>
<td>1704176</td>
<td>0.14</td>
<td>0.04</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. AS9601</td>
<td>1669886</td>
<td>0.13</td>
<td>0.06</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. MIT 9215</td>
<td>1738790</td>
<td>0.12</td>
<td>0.07</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. MIT 9301</td>
<td>1641879</td>
<td>0.11</td>
<td>0.07</td>
</tr>
</tbody>
</table>

4440279.3 – Tabuaeran/Microbial

<table>
<thead>
<tr>
<th>Genome</th>
<th>Size</th>
<th>Coverage/Viral</th>
<th>Coverage/Microbial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alteromonas macleodii 'Deep ecotype'</td>
<td>4173873</td>
<td>0</td>
<td>0.28</td>
</tr>
<tr>
<td>Candidatus Pelagibacter ubique HTCC1002</td>
<td>1322994</td>
<td>0.01</td>
<td>0.2</td>
</tr>
<tr>
<td>Candidatus Pelagibacter ubique HTCC1062</td>
<td>1308759</td>
<td>0.01</td>
<td>0.2</td>
</tr>
<tr>
<td>Synechococcus sp. CC9605</td>
<td>2510659</td>
<td>0.01</td>
<td>0.17</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. MIT 9312</td>
<td>1709204</td>
<td>0.16</td>
<td>0.08</td>
</tr>
<tr>
<td>Pseudoalteromonas tunicata D2</td>
<td>4938159</td>
<td>0</td>
<td>0.07</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. MIT 9301</td>
<td>1641879</td>
<td>0.11</td>
<td>0.07</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. MIT 9215</td>
<td>1738790</td>
<td>0.12</td>
<td>0.07</td>
</tr>
<tr>
<td>Synechococcus sp. RS9917</td>
<td>2579542</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td>Haemophilus influenzae R2866</td>
<td>1928482</td>
<td>0.01</td>
<td>0.06</td>
</tr>
</tbody>
</table>

We have blasted each genome in both refseq_viral and refseq_microbial against each of the eight samples and summed the results. For each genome, we have computed its coverage by each sample based on Equation (28). We have considered only reads that were aligned through at least 80bp (approximately 80% of average read length) with percent identity > 85. The results for the four pairs of datasets are summarized in Table 3 – Table 6. As can be observed, the results for the Palmyra (Table 4) and Tabuaeran (Table 5) samples are consistent with our expectations: most abundant genomes in the viral fraction are viruses or plasmids, while microbial genomes dominate the microbial samples. The coverage for the most abundant species in the microbial fraction of Tabuaeran is relatively low, which may hint on either high diversity of microbial community in Tabuaeran, or on different species than those we know in this region. The results for the Kingman samples are less definitive: the microbial fraction is dominated by plasmids and microbial genomes, while the most abundant genomes in the viral fraction are of SAR11 and cyanophages infecting Prochlorococcus (Table 3). We believe that the SAR11 DNA is present in the viral fraction due to the organism’s small dimensions: their diameter is 0.12-0.20µM, similar to the dimensions of viruses and...
less than the filter size used for the viral fraction in the Northern Line Islands (0.45µM). Another possibility is that the sequences found in the viral fraction in fact belong to an unknown phage that infects SAR11, or yet unknown SAR11 Generalized Transfer Agents (GTAs), however these cannot be proven at the current time.

Table 6. Genomes with highest coverage, samples 4440038.3 and 4440041.3 (Christmas). Green: microbial genomes, orange: viral genomes, blue: plasmids.

<table>
<thead>
<tr>
<th>Genome</th>
<th>Size</th>
<th>Coverage/Viral</th>
<th>Coverage/Microbial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidatus Pelagibacter ubique HTCC1062</td>
<td>1308759</td>
<td>19.24</td>
<td>0.08</td>
</tr>
<tr>
<td>Candidatus Pelagibacter ubique HTCC1002</td>
<td>1322994</td>
<td>19.15</td>
<td>0.08</td>
</tr>
<tr>
<td>Prochlorococcus phage P-SSM4</td>
<td>178249</td>
<td>2.27</td>
<td>0</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. MIT 9215</td>
<td>1738790</td>
<td>1.86</td>
<td>0.16</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. AS9601</td>
<td>1669886</td>
<td>1.85</td>
<td>0.13</td>
</tr>
<tr>
<td>Bacillus megaterium QM B1551 plasmid pBM400</td>
<td>53903</td>
<td>1.77</td>
<td>0.53</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. MIT 9301</td>
<td>1641879</td>
<td>1.72</td>
<td>0.14</td>
</tr>
<tr>
<td>Prochlorococcus marinus str. MIT 9312</td>
<td>1709204</td>
<td>1.68</td>
<td>0.15</td>
</tr>
<tr>
<td>Silicibacter sp. TM1040 plasmid unnamed</td>
<td>130973</td>
<td>1.6</td>
<td>0.22</td>
</tr>
<tr>
<td>Prochlorococcus phage P-SSM2</td>
<td>252401</td>
<td>1.29</td>
<td>0</td>
</tr>
</tbody>
</table>

The results for the Christmas Island samples (Table 6) hint on possible problems. By far, the most abundant genome in the viral fraction is of SAR11. This finding may be explained similarly to the finding of SAR11 in the viral fraction of Kingman samples, however, cyanobacteria seems to be similarly abundant to cyanophages, which is not supposed to happen since Prochlorococcus size is 0.5-0.8µM, larger than the viral filter fraction. In addition, Prochlorococcus seem to be much more abundant in the viral fraction than in the microbial one, which also raises questions. Coverage for even the most abundant genomes in the microbial fraction is very low; in addition, viruses are found together with microbes, some of them are clearly not expected to be found in a marine environment (e.g. “Human herpesvirus 6A”). The low coverage is probably due to both the high diversity of microbial community in the environment, as well as the lack of close genomes in refseq. The presence of non-
marine microbes and viruses may be due the conditions in the Christmas Island, e.g. untreated sewage etc.

16s (SSU) and 23s (LSU) rRNA coding genes

In order to further test the contents of each sample we searched for 16s and 23s rRNA coding genes in the eight datasets. Both genes are considered to be present on microbial genomes only, and therefore their presence may hint on the presence of microbes in a sample. The procedure we used for finding those genes was the following. First, we blasted one 16s and one 23s rRNA coding genes from SAR11 against each of the eight sample (-p blastn -F F -e 1e-5 -r 2 -q -3). Both genes are conserved, and therefore a single template of each gene is likely to capture at least most of the instances of both genes in the samples. Next, for each sample, each portion of the genes found was blasted against the nt database (same parameters as above except for e-value threshold which was set to 1e-10). For each true instance of 16s/23s rRNA coding genes we expected to find at least one such gene at its top 50 hits. The scan was done automatically with manual inspection of some of the cases. The 16s/23s gene was almost always the first gene appearing in the hit list (other hits contained mostly complete genomes), which probably mean that this strategy produced at least close numbers to the real ones. As can be observed from Table 7, the microbial samples of both Palmyra and Tabuaeran contain many 16s/23s gene copies, while the viral fractions do not contain almost any such gene. These results are consistent with the previous analysis, and increase our confidence in the credibility of these samples. As for Kingman, the microbial fraction contains 14 reads with a 16s/23s segment on them, while the viral fraction contains only 4 such reads. These results also increase our confidence in the Kingman samples, and together with the results of the previous analysis probably mean that these samples are reliable. The picture is very different as for the Christmas samples: both the viral and microbial samples contain a lot of 16s/23s genes, with the viral fraction containing much more such genes than the microbial one. These results, together with the previous analysis, raise significant questions as for the credibility of these samples. Therefore it is probably not a good idea to use these samples for viral/microbial separation.

<table>
<thead>
<tr>
<th>Site</th>
<th>Viral</th>
<th>Microbial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kingman</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>Palmyra</td>
<td>1</td>
<td>58</td>
</tr>
<tr>
<td>Tabuaeran</td>
<td>0</td>
<td>38</td>
</tr>
<tr>
<td>Christmas</td>
<td>125</td>
<td>11</td>
</tr>
</tbody>
</table>

Next we provide a concise description of two of our works, which took advantage of the abovementioned methods: the study of the tryptophan operon in the Sargasso Sea metagenome (Kagan, Sharon et al. 2008) and the study of microbial genes on viral genomes. The use of these methods will be also demonstrated in Chapter 4, describing the discovery of PSI gene cassettes in cyanophages (Sharon, Alperovitch et al. 2009).
3.3. Applications: studying the tryptophan operon in the Sargasso Sea metagenome

The tryptophan (trp) pathway and the organization of trp genes on bacterial genomes have been studied for many years. The classical organization of the genes is in a single operon (the trp operon), composed of seven genes organized as follows: \( trpE \rightarrow trpG \rightarrow trpD \rightarrow trpC \rightarrow trpF \rightarrow trpB \rightarrow trpA \). With the availability of environmental data from the Sargasso Sea we were interested in studying the organization of the trp genes on microbial genomes in the environment. For this purpose we have collected and annotated all scaffolds from the Sargasso Sea metagenome (Venter, Remington et al. 2004) that contained at least one trp gene. Trp genes proved to be quite abundant in the Sargasso Sea metagenome: at least 5% of all genes related to amino acid biosynthesis were trp genes. In order to examine the organization of trp genes on environmental genomes we have performed a gene organization analysis that is very similar to the one described in Section 3.1. Figure 12 contains the graph generated by our gene organization procedure. The two split operons \( trpE \rightarrow trpG \rightarrow trpD \rightarrow trpC \) and \( trpF \rightarrow trpB \rightarrow trpA \) can be easily identified in the graph as the likely most abundant organizations in the metagenome. Some of the genes probably appear only in the context of the tryptophan operon: \( trpD \), for example, was not found in any other context. Other genes, such as \( trpG \) and \( trpE \) also appear in other contexts (refer to (Kagan, Sharon et al. 2008) for a complete discussion).

![Gene organization graph of tryptophan genes in the Sargasso Sea metagenome. Vertices refer to genes (X is a shortcut for trpX), arrows represent to neighboring genes. Arrow size and color indicates number of occurrences. Refer to (Kagan, Sharon et al. 2008) for a full description.](image-url)
3.4. Applications: finding viral genomes with microbial genes

The discovery of PSII and PSI genes on the genomes of cyanophages (Mann, Cook et al. 2003; Lindell, Sullivan et al. 2004; Millard, Clokie et al. 2004; Lindell, Jaffe et al. 2005; Sharon, Alperovitch et al. 2009) suggests that viruses may have acquired other microbial genes and use them in order to get evolutionary advantage over other viruses. In order to test this possibility we have scanned the GOS scaffold dataset for scaffolds that are of likely viral origin that contain microbial genes previously undiscovered on viral genomes. Next we describe the process.

3.4.1. Methods

The general scheme used for finding such scaffolds is presented in Figure 13.

1. **Collecting potential viral GOS scaffolds.** As a first step, all GOS scaffolds with at least one homolog to known viral genes were collected. For this purpose all 75,079 proteins available in the NCBI refseq-viral database (April 2009) were blasted (-p blastn -F F -e 1e-20) against the GOS scaffolds dataset. Overall this stage yielded 223,687 scaffolds with at least one region that is similar, at the amino acids level, to one of the proteins in refseq-viral.

2. **Filtering scaffolds. In order to determine** whether the abovementioned regions in the scaffolds resulting from the previous stage are indeed viral, we have blasted them against a combined database of all proteins in both refseq-viral and refseq-microbial (parameters: -p blastx -F F -e 1e-10). After filtering all scaffolds whose regions got best hits from refseq-microbial we were left with 84,921 scaffolds with at least one best hit from refseq-viral.

3. **Gene finding and annotation.** Gene finding and annotation was performed using a similarity-based iterative process. At each iteration, all remaining scaffolds were blasted against a combined database of refseq-viral and refseq-microbial (parameters: -p blastx -e 1e-3 -F F). Best alignment for each scaffold was kept, and the corresponding region on the scaffold was replaced with a series of N’s before the next iteration. Sequences for which no new hit was found were removed. Overall 202,231 regions similar to 17,070 refseqproteins from the combined database were found, 4,268 of which originated from refseq-viral and 12,802 from refseq-microbial. We refer to this set of refseq proteins as refseq-hits.

4. **refseq-hits tagging.** Tagging refers to the process of assigning an origin for each refseq-hit. A protein may be either viral-exclusive (if its origin is refseq-viral and it has no microbial homolog), microbial-exclusive (coming from refseq-microbial with no viral homologs) or viral-microbial (if it can be found in both databases). For the purpose of tagging, each refseq-hit was blasted against the other database (-p blastp -e 1e-3 -F F) and tagged according to the results. Overall 1,840 refseq-hits were tagged as viral-exclusive, 5,829 were tagged as microbial-exclusive and 8,231 were viral-microbial.
5. **refseq-hits clustering.** Different refseq-hits may be homologous, thus representing proteins with the same predicted function. At this stage we aimed to cluster similar refseq-hits, based on sequence similarity. For this purpose we have used the following clustering algorithm:
identity(rh1, rh2) and conserved(rh1, rh2) are the number of identical and conserved amino acids, respectively, in the alignment of rh1 and rh2. The algorithm assumes that the graph resulting from the graph construction stage is a collection of strongly connected components. If this is not the case the algorithm will fail. Exclusion of vertices from $C_{rh}$ (step 4e) is based on a heuristic; other alternatives also exist, but we have found this one to be both fast and accurate and therefore chose to use it. The criterion is stringent, and prefers accuracy over sensitivity.

6. **refseq-hits cluster tagging.** At this stage each cluster is assigned a viral (V), microbial (M) or viral-microbial (VM) tagging based on the tagging of the cluster members. Any of the following conditions is considered to be non-conflicting:
• All members are of viral-exclusive – cluster is viral

• All members are of microbial-exclusive – cluster is microbial

• Members are of viral-microbial and possibly viral- and microbial-exclusive – cluster is viral-microbial

Clusters with both viral- and microbial-exclusive members and no viral-microbial members are considered to be conflicting clusters. In that case the cluster is split into two sub-clusters, one viral and one microbial.

7. **Filtration of non-viral scaffolds, and viral scaffolds that do not contain members of microbial clusters.** All scaffolds containing one of the following makeup of genes were considered to be of interest:

- A gene that is a member of a microbial cluster located between two genes with viral tagging

- At least one gene that belongs to a microbial-exclusive cluster and at least 4 viral-exclusive genes such that at least 20% of the genes on the scaffold are viral-exclusive.

Note that two kinds of tagging are used: in order to decide whether the scaffold is viral or microbial we use the gene’s tagging, namely the origin of the gene’s best refseq-hit. In order to decide whether a specific gene is viral or microbial we use the tagging of the gene’s refseq-hit’s cluster tagging, because we are interested only in genes that were not previously reported to be of viral origin.

8. **Manual inspection of all cases of resulting microbial clusters on viral scaffolds.** This stage included validation of the finding, exclusion of multiple copies of the same scaffold, etc.

### 3.4.2. Results

Overall, 13,741 clusters resulted from the above process of which 12,257 were singletons, namely contained exactly one refseq-hit (Figure 14). 6,851 (49.9%) of the clusters received a VM tagging, 5,182 (37.7%) received an M tagging and 1,708 (12.4%) received a V tagging. Conflicting tagging was found in only 38 (0.3%) of the clusters, supporting the credibility of the process. Table 8 summarizes the 11 largest clusters and their tagging and annotations. Annotation of these clusters was done manually, based on the annotation of most cluster members. In most cases annotation of the different cluster members was consistent, again, supporting the credibility of the process.

Overall, 5,267 scaffolds were identified as viral containing at least one gene that was not previously observed on viral genomes. After removing redundant scaffolds we were left with 3,083 scaffolds, which were further processed. Next we were interested in validating the origin of our scaffolds; this was done by recruitment of
Table 8. Largest refseq-hits clusters and their annotation.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Size</th>
<th>Tagging</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>C133</td>
<td>37</td>
<td>VM</td>
<td>PSII D1</td>
</tr>
<tr>
<td>C83</td>
<td>31</td>
<td>VM</td>
<td>ribonucleoside-diphosphateductase beta subunit (small chain)</td>
</tr>
<tr>
<td>C322</td>
<td>24</td>
<td>M</td>
<td>scaffold protein/FeS cluster assembly scaffold IscU/ NifU family protein</td>
</tr>
<tr>
<td>C310</td>
<td>21</td>
<td>M</td>
<td>ribosomal protein S6 modification protein/ alpha-L-glutamate ligase</td>
</tr>
<tr>
<td>C542</td>
<td>21</td>
<td>VM</td>
<td>ribonucleotide-diphosphateductase subunit alpha</td>
</tr>
<tr>
<td>C34</td>
<td>19</td>
<td>VM</td>
<td>Ferredoxin</td>
</tr>
<tr>
<td>C22</td>
<td>17</td>
<td>VM</td>
<td>guanosine 5´-monophosphate oxidoreductase</td>
</tr>
<tr>
<td>C47</td>
<td>17</td>
<td>VM</td>
<td>GDP-mannose 4,6-dehydratase</td>
</tr>
<tr>
<td>C45</td>
<td>16</td>
<td>VM</td>
<td>ribonucleoside-diphosphateductase, alpha subunit (large chain)</td>
</tr>
<tr>
<td>C410</td>
<td>16</td>
<td>M</td>
<td>Redoxin/ antioxidant, AhpC/Tsa family protein/ peroxidoxin/glutaredoxin family protein/ putative hydroperoxidereuctase protein</td>
</tr>
<tr>
<td>C276</td>
<td>15</td>
<td>VM</td>
<td>PSII D2</td>
</tr>
</tbody>
</table>

the scaffolds against the Northern Line Islands samples, excluding the two samples from the Christmas Island. As a control, we added all scaffolds from GOS containing some part of the 16S rRNA coding gene.

Results are presented in Figure 15. Overall, 47% of the VirMic scaffolds recruited at least 2 reads. Of those, the vast majority got significant recruitment from the viral sample and no or very little recruitment from the microbial fraction. Control scaffolds, on the other hand, recruited much more reads from the microbial samples than from the viral samples. These results are in good agreement with our process, which was based on gene-content based separation of scaffolds into viral and microbial bins. All scaffolds which recruited only microbial reads were excluded from further analysis; these included mostly scaffolds with mostly microbial and very few viral genes.

We refer to all microbial clusters that were found on viral scaffolds as VirMic:Microbial. Similarly, viral viral-microbial clusters are termed VirMic:Viral and VirMic:viral_microbial. All VirMic:Microbial clusters with at least 4 occurrences

![Figure 14. Cluster size distribution. Left: 12,527 out of the 13,741 clusters (91.2%) contained exactly one refseq-hit and 1,214 contained more than 1 member. Right: for each cluster size≥2, distribution of number of clusters of this size.](image)
on viral scaffolds went through manual inspection, which included verification of viral origin of their scaffolds, and their annotation. After this step we were left with 34 VirMic:Microbial clusters which represent the set of newly discovered microbial genes on viral genomes (Table 9). This set contains genes such as psaA (T109), psaJ (T204), psaC (T1557) and NADH dehydrogenase subunit I (T603), all of them reported in (Mann, Clokie et al. 2005; Sharon, Alperovitch et al. 2009). The gene psaJ and NADH dehydrogenase subunit I were found in other contexts than the PSI gene cassette previously reported. The psbN gene (T443), a PSII coding gene, which was not previously known to be on viral genomes, was found in eight copies in our database. The rest of the genes are currently studied, with their biological significance to be reported in the near future.

In the next chapter I describe how the methods presented in this chapter were used for discovering PSI gene cassettes on the genomes of cyanophages.
Table 9. Manually validated VirMic:Microbial clusters with at least four instances.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Size</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>T128</td>
<td>70</td>
<td>Peptidedeformylase</td>
</tr>
<tr>
<td>T1396</td>
<td>58</td>
<td>glycerol-3-phosphate cytidylyltransferase</td>
</tr>
<tr>
<td>T982</td>
<td>36</td>
<td>Exported protein</td>
</tr>
<tr>
<td>T451</td>
<td>31</td>
<td>Glycerodehydrogenase</td>
</tr>
<tr>
<td>T768</td>
<td>30</td>
<td>Antioxidant, AhpC/Tsa family protein</td>
</tr>
<tr>
<td>T1414</td>
<td>28</td>
<td>fructose-1,6-bisphosphate aldolase class I</td>
</tr>
<tr>
<td>T100</td>
<td>20</td>
<td>glycosyltransferase, group 1</td>
</tr>
<tr>
<td>T156</td>
<td>15</td>
<td>Serine hydroxymethyltransferase</td>
</tr>
<tr>
<td>T212</td>
<td>13</td>
<td>NAD(P)H-quinoneoxidoreductase subunit 4</td>
</tr>
<tr>
<td>T603</td>
<td>13</td>
<td>NADH dehydrogenase subunit I</td>
</tr>
<tr>
<td>T17</td>
<td>13</td>
<td>membrane protein</td>
</tr>
<tr>
<td>T90</td>
<td>13</td>
<td>phosphoribosylaminomimidazole-succinocarboxamidesynthase</td>
</tr>
<tr>
<td>T1338</td>
<td>12</td>
<td>Glycosyltransferase involved in cell wall biogenesis-like</td>
</tr>
<tr>
<td>T1486</td>
<td>11</td>
<td>NH(3)-dependent NAD+ synthetaseNadE</td>
</tr>
<tr>
<td>T71</td>
<td>10</td>
<td>glycine cleavage system protein P2</td>
</tr>
<tr>
<td>T885</td>
<td>10</td>
<td>iron-sulfur cluster-binding protein</td>
</tr>
<tr>
<td>T891</td>
<td>8</td>
<td>Phosphorylase</td>
</tr>
<tr>
<td>T327</td>
<td>8</td>
<td>Mannitol-1-phosphate/altronate dehydrogenase</td>
</tr>
<tr>
<td>T198</td>
<td>8</td>
<td>alkylhydroperoxiredoxuctase/ Thiol specific antioxidant/ Mal allergen</td>
</tr>
<tr>
<td>T1113</td>
<td>8</td>
<td>putative NH(3)-dependent NAD synthetase</td>
</tr>
<tr>
<td>T596</td>
<td>8</td>
<td>scaffold protein</td>
</tr>
<tr>
<td>T158</td>
<td>7</td>
<td>sugarisomerase (SIS)</td>
</tr>
<tr>
<td>T255</td>
<td>6</td>
<td>translation initiation factor IF-1</td>
</tr>
<tr>
<td>T1119</td>
<td>6</td>
<td>coenzyme PQQ biosynthesis protein A</td>
</tr>
<tr>
<td>T445</td>
<td>6</td>
<td>Adenylatekinase</td>
</tr>
<tr>
<td>T444</td>
<td>5</td>
<td>iron-sulfur cluster insertion protein ErpA</td>
</tr>
<tr>
<td>T883</td>
<td>5</td>
<td>glycine cleavage system aminomethyltransferase T</td>
</tr>
<tr>
<td>T712</td>
<td>5</td>
<td>antoxidant, AhpC/TSA family protein</td>
</tr>
<tr>
<td>T446</td>
<td>5</td>
<td>Photosystem II reaction center protein N</td>
</tr>
<tr>
<td>T361</td>
<td>5</td>
<td>S'-methylthioadenosine nucleosidase/S'-adenosylhomocysteinemucleosidase</td>
</tr>
<tr>
<td>T72</td>
<td>5</td>
<td>Ribosomal protein S21</td>
</tr>
<tr>
<td>T204</td>
<td>4</td>
<td>Photosystem I reaction center subunit IX</td>
</tr>
<tr>
<td>T1557</td>
<td>4</td>
<td>Photosystem I subunit VII</td>
</tr>
<tr>
<td>T109</td>
<td>4</td>
<td>Photosystem I P700 chlorophyll a apoprotein A1</td>
</tr>
</tbody>
</table>
Chapter 4.
The Discovery of Photosystem I Genes on Viral Genomes

Bacteriophages have the ability to manipulate the life histories and evolution of their hosts (Rohwer and Thurber 2009) and evolved many adaptation and defense mechanisms for efficient survival and multiplication. Most of these involve manipulation of the host DNA, as well as the incorporation, into the phage genomes, of bacterial genes that encode proteins with a potential to facilitate bacteriophage reproduction (Brown, Wickham et al. 2006). Recently, it was discovered that marine cyanophages (bacteriophages that infect cyanobacteria) carry photosynthetic genes, and it was suggested that these genes increase phage fitness (Mann, Cook et al. 2003; Lindell, Sullivan et al. 2004; Millard, Clokie et al. 2004; Lindell, Jaffe et al. 2005). Cyanobacterial photosynthetic membranes contain two photosystems, of which PSII mediates the transfer of electrons from water, the initial electron donor, to the plastoquinone pool whereas PSI mediates electron transfer from plastocyanin to ferredoxin, thereby generating reducing power needed for CO$_2$ fixation in the form of NADPH. While PSII is known to be sensitive to photodamage, PSI is considered to be a more stable complex. The PSII gene psbA coding for the labile D1 protein is readily detected in various cultured and environmental cyanophages infecting Prochlorococcus and Synechococcus (Mann, Cook et al. 2003; Lindell, Sullivan et al. 2004; Sullivan, Lindell et al. 2006; Sharon, Tzahor et al. 2007). In addition, other photosynthesis genes encoding the PSII D2 protein (Mann, Cook et al. 2003; Lindell, Sullivan et al. 2004), high-light inducible proteins, pigment biosynthesis proteins (Ho1, PebA, PcyA), or the photosynthetic electron transport proteins plastocyanin (PetE) and ferredoxin (PetF) were also identified in several cyanophage genomes(Lindell, Sullivan et al. 2004; Sullivan, Coleman et al. 2005).

4.1. Results

To assess the possible presence of other photosynthesis-related genes in viruses, we set up a designated search scheme for publicly available metagenomic data. Initially we searched for the cyanobacterial PSI gene, psaA. Together with PsbB, the PsbA protein forms the heterodimeric core of PSI that binds the primary electron donor P700, formed by a special chlorophyll pair (Jordan, Fromme et al. 2001; Jordan, Fromme et al. 2001). Using tBLASTx, different Synechococcus and Prochlorococcus psaA gene sequences were used as queries against the GOS expedition (Rusch, Halpern et al. 2007) dataset.
Figure 16. A maximum likelihood phylogenetic tree of \( psaA \)-deduced amino acid sequences obtained from the GOS expedition. \( psaA \) sequences from the 27 fully sequenced and annotated \textit{Synechococcus} (blue background) and \textit{Prochlorococcus} (green background) genomes are shown. Sequences from the GOS expedition are shown in bold, and sequences from the original scaffolds obtained in this study are indicated in red. For clarity, the tree shows only a subset of the 583 partial \( psaA \) sequences found in the GOS data set. The tree is on the basis of an alignment of 94 shared amino acids.
We detected 574 psaA-containing GOS scaffolds. These were further screened to identify those that were likely to originate from viruses using tBLASTx against refseq_viral, a database that contains all known viral genomes. This procedure reduced the number of suspected scaffolds to five. The PsA homologues encoded by these sequences showed only 65-75% identity to Prochlorococcus or marine Synechococcus PsA proteins. On a maximum likelihood tree, four of these proteins clustered together on a well supported branch related to Prochlorococcus PsA, whereas the fifth sequence (JCVI_SCAF_1096628008692) was retrieved near the base of the Synechococcus branch (Figure 16). As the GOS general scaffold assembly represents reads that come from different GOS sample sites or from different clones and hence are chimerical by definition, we restricted all further analysis to sequences assembled from single clone reads only. Analysis of the GOS clones containing the modified psaA genes confirmed their viral origin (likely cyanophages of the Myoviridae family) as indicated by the presence, in the vicinity of psaA, of typical viral genes, such as nrdA and B (that encode the alpha-2 and beta-2 subunits of viral ribonucleosidediphosphatereductase, respectively) or the T4-like neck gp13 protein gene (Figure 17). In addition to psaA, these clones contained clusters of PSI genes, including psaB, psaC and a unique fused version of the psaF and psaJ genes (psaJF). An analysis of the GOS datasets with other PSI peptides as baits revealed the

Figure 17. Schematic physical maps of selected viral-suspected GOS clones (top), Prochlorococcus and Synechococcus genomes (middle) and environmental PCR products containing PSI genes (bottom). Red arrows represent ORFs with predicted viral origin, and grey arrows represent unknown ORFs. Capital letters represent the corresponding PSI genes. Gaps shown in Indian Ocean GOS clones (stations GS111 and GS117) are the result of regions that were not covered by the end-reads owing to the size of these clones (5 Kbps). Primer positions on GOS clones are indicated by triangles, and thick colored lines denote PCR products.
presence of several other PSI clusters also containing \textit{psaE}, \textit{psaK} and \textit{psaD} genes (see distribution in the different GOS sites in Table 10). As in the case of \textit{PsaA}, phylogenies made with these additional PSI protein sequences showed that they were all clustered apart from the homologous proteins of \textit{Prochlorococcus} and \textit{Synechococcus}, except \textit{PsaC} and \textit{PsaD} from GOS clone 106100809984 (hereafter clone 9984; a clone used to build the previously mentioned scaffold JCVI_SCAF_1096628008692) which were retrieved closer to corresponding cyanobacterial sequences than to other viral sequences (Figure 21). Examining the \textit{Prochlorococcus} and \textit{Synechococcus} genome arrangements (middle panel in Figure 17) or gene-pairs frequency modeling showed that the organization observed in most viral clones, \textit{psaF-C-A-B-K-E-D}, differs from that observed in these cultured cyanobacterial genomes and in most other (probably cyanobacteria-derived) GOS sequences (Figure 18). The PSI genes found on clone 9984 (represented by GOS reads 1095964115098 & 1095975140994 in Figure 17) had a different order (\textit{psaD-C-A}) than on other clones, consistently with their distinct positions in phylogenetic trees (Figure 21).

\textbf{Table 10.} Distribution of viral-PSI clones (one or two reads sequenced from the same DNA fragment) in the GOS expedition.

<table>
<thead>
<tr>
<th>GOS site</th>
<th># Clones</th>
<th>Site Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS025</td>
<td>1</td>
<td>Dirty Rock, Cocos Island</td>
</tr>
<tr>
<td>GS027</td>
<td>1</td>
<td>Devil's Crown, Floreana Island</td>
</tr>
<tr>
<td>GS029</td>
<td>1</td>
<td>North James Bay, Santiago Island</td>
</tr>
<tr>
<td>GS034</td>
<td>1</td>
<td>North Seamore Island</td>
</tr>
<tr>
<td>GS037</td>
<td>8</td>
<td>Equatorial Pacific TAO Buoy</td>
</tr>
<tr>
<td>GS045</td>
<td>1</td>
<td>400 miles from French Polynesia</td>
</tr>
<tr>
<td>GS047</td>
<td>11</td>
<td>201 miles from French Polynesia</td>
</tr>
<tr>
<td>GS111</td>
<td>1</td>
<td>Indian Ocean</td>
</tr>
<tr>
<td>GS117</td>
<td>1</td>
<td>Seychelles</td>
</tr>
</tbody>
</table>

To validate the viral origin of these genes and their unique cluster organization, data obtained from the GOS project were cross-referenced with recently released 454 pyrosequencing metagenomic sequences obtained from a variety of marine and non-marine viral and microbial biomes datasets (Dinsdale, Edwards et al. 2008). This was a critical step in increasing the credibility of the results because the two approaches each introduce different biases (Harismendy, Ng et al. 2009). The various viral-suspected PSI GOS clones identified were used to recruit reads from these different datasets. Marine virome fragments were readily recruited to all of the viral GOS clones regions, whereas virome or microbiome fragments coming from other environments were scarcely recruited (Table 11), with a much lower identity (Figure 18a), further supporting a marine viral origin for the PSI clones. Overall coverage measure of viromes and microbiomes to all different GOS clones containing PSI genes (Figure 18b) clearly points to 2 distinguished populations, one from bacteria (cyanobacteria) and one from viruses (phages). Except for clone 9984, all our identified viral clones are falling in the viral population. In addition, marine virome fragments were also recruited to regions between
Table 11. Number of different biomes reads recruited at different percent identity to GOS suspected viral-PSI clones.

<table>
<thead>
<tr>
<th>Type</th>
<th>Microbial Metagenomes</th>
<th>Viral Metagenomes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>85% 90% 95%</td>
<td>85% 90% 95%</td>
</tr>
<tr>
<td>Coral</td>
<td>0 0 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>Fish</td>
<td>0 0 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>Freshwater</td>
<td>2 0 0</td>
<td>1 1 1</td>
</tr>
<tr>
<td>Hypersaline</td>
<td>1 0 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>Marine</td>
<td>1 0 0</td>
<td>207 144 91</td>
</tr>
<tr>
<td>Microbialites</td>
<td>0 0 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>Terrestrial</td>
<td>0 0 0</td>
<td>0 0 0</td>
</tr>
</tbody>
</table>

Figure 18. Distribution of neighboring genes involving at least one PSI gene. Each arrow connects neighboring genes, and its thickness represents the number of pairs found in *Synechococcus* and *Prochlorococcus* genomes (left gene-circle), microbial sequences from the GOS metagenome (middle gene-circle) and viral sequences from the GOS metagenome (right gene-circle). Note the uninterrupted clustering of PSI genes in phage genomes that contrasts the scattered arrangement of these genes in cyanobacterial genomes (in both cultures and GOS). Gene connections observed only once are not shown.
the photosynthesis genes, linking neighbor genes in the observed viral cassette (Figure 18a & Table 12), an observation that supports the gene cluster organization observed on the GOS clones.

Figure 19. Recruitment of GOS clones carrying PSI genes with Northern Line Islands biomes. Left: Recruitment of three GOS physical clones carrying suspected viral-PSI genes by Northern Line Islands virome (green) and microbiome (red) reads. Upper panel shows recruitment at 75-100% identity while lower panel shows fold coverage by these reads. Accession numbers of the GOS reads used are presented above each clone (JCVI_READ_#). Right: Recruitment coverage of GOS single clones carrying PSI genes with Northern Line Islands biomes [viromes (X-axis) and microbiomes (Y-axis)]. Coverage is defined as % of GOS clone length covered by at least one recruited read.

To validate the juxtaposition of the genes in the identified viral-PSI gene clusters, DNA from the Northern Line Islands marine virome (Dinsdale, Pantos et al. 2008) was used to perform ‘continuous’ overlapping and long PCR with primers assigned to the different genes. The results of the ‘continuous’ overlapping PCR (lower panel in Figure 17) and the amplification of a ~6.2 kb long PCR amplicon (lower panel in Figure 17) spanning the entire PSI cassette and including the viral nrdB gene, show that the different genes in the cluster nrdB-hyp-psaJF-C-A-B-K-E-D (and also a novel arrangement nrdB-psaJF-C:genbank #EU926755) are physically linked and exist as one photosynthetic cluster.

Although the data presented here are derived from environmental genomic datasets (non-continuous data), and therefore the lack of genes is not a proof of absence, it is notable that the PSI genes psaI, psaL and psaM were not found in the viral psa gene cassettes. The psaM gene is naturally absent from plants and its inactivation in cyanobacteria shows that it is mainly required for the formation of stable PSI trimers (Naithani, Hou et al. 2000). Similarly, targeted inactivation of cyanobacterial psaL produces functional PSI complexes unable to form trimers, whereas Psal is mostly required for stabilizing PsaL (Xu, Hoppe et al. 1995). Therefore, these three proteins are mainly involved in the trimer formation of cyanobacterial PSI, and their potential absence from the viral clone might indicate on the formation of a monomeric PSI complex as in plants (Kouril, van Oosterwijk et al. 2005) and not a trimeric complex as in cyanobacteria (Boekema, Dekker et al. 1987). All genetic information required to form this putative minimal, monomeric PSI is clustered onto a very small cyanophage genome fragment (~5.9 kb). To our
knowledge, gene clusters encoding all the components of a photosystem from an oxygenic phototroph have not been previously reported, and neither have there been reports on cyanobacterial PSI genes outside a cyanobacterial chromosome.

The potential structural consequences of assembling the phage proteins into the PSI complex were modeled in relation to the 2.5 Å structure of PSI from the cyanobacterium *Thermosynechococcus elongates* (Jordan, Fromme et al. 2001). We modeled the PsajF fusion protein (where the C-terminus of Psaj is fused to the N-terminus of Psaf) at the position of subunits J and F of PSI. Figure 5 shows that the viral PsajF fusion protein fits perfectly at the position of subunits J and F in the PSI structure. The only prominent change was the absence of the N-terminus of subunit F, which is responsible for the specific binding of the natural electron donor (plastocyanin) of PSI (Hippler, Drepper et al. 1997). In chloroplasts of green algae and plants, this part of subunit F is elongated, resulting in higher affinity of plastocyanin to the chloroplast PSI (Hippler, Drepper et al. 1997; Ben-Shem, Frolow et al. 2003; Nelson and Yocum 2006). While both plastocyanin and cytochrome c₆ are capable of donating electrons to PSI (Nelson and Yocum 2006) in *Chlamydomonas reinhardtii*, this site in higher plants is specific for plastocyanin (Merchant and Sawaya 2005). However, the electron donation to PSI is not at all promiscuous, and several soluble cytochromes, including the respiratory cytochrome c₆, fail to donate electrons to PSI (Kerfeld and Krogmann 1998). We hypothesize that the replacement of Psaj and Psaf with the viral PsajF fusion protein enables electron donation through additional electron carriers, including cytochromes that usually function as electron donors to cytochrome oxidase.

**Figure 20.** Structural consequences of assembling the viral fusion protein PsajF into PSI. **Left:** The structure of *T. elongatus* PSI (subunits) was illustrated by PyMOL (http://pymol.sourceforge.net/) using a PSI monomer (adopted from PDB 1jb0). Psaf (magenta), Psaj (blue), and all the other subunits in green. **Right:** A model for the structure of the viral PsajF fusion protein (red) substituting the original Psaf and Psaj subunits.

The mechanistic consequence of a less selective electron donation to PSI might be the possibility of sharing reducing power generated by the respiratory chain with the photosynthetic electron transport chain. A similar phenomenon, called chlororespiration, detected in both cyanobacteria and chloroplasts, was attributed to the
plastid terminal plastoquinoneoxidase (PTOX) (Rumeau, Peltier et al. 2007). The electron mediator in this process is plastoquinone, which shuttles between the respiratory-like chain and the chloroplast b6f complex (Rumeau, Peltier et al. 2007). After phage infection and the incorporation of the phage gene products into PSI, the function of electron mediation could be carried out by a soluble cytochrome. In addition, the phage might boost the amount of PSI in order to lead the infected cyanobacterial cells toward a cyclic photosynthesis for the generation of ATP in expense for the production of reducing power for CO₂ fixation. The PSI levels are notably low in both oceanic Synechococcus (Bailey, Melis et al. 2008) and in Prochlorococcus (Partensky, LaRoche et al. 1997), possibly as a result of adaptation to low iron levels, and it was recently proposed that a compensatory mechanism might exist, involving alternative electron flow to O₂ (Bailey, Melis et al. 2008).

The phage PSI gene fusion psaJF described here is the first example of a phage gene innovation that involves structural membrane proteins. Modification toward a new function of existing cyanobacterial proteins by their phages was recently demonstrated for the divergent phage PebA homolog (Dammeyer, Bagby et al. 2008) (renamed PebS [phycoerythrobilinsynthase]). The phage PebS single-handedly catalyzes a reaction for which uninfected host cells require two consecutive enzymes, PebA and PebB. Considering these findings and our calculations that suggest a high likelihood of gene cluster formation in phage genomes (see supplemental information), the oceanic virome could be an almost unlimited source of naturally bioengineered gene cassettes.

Table 12. Accession numbers of Line Island 454 reads (from reads recruited on GOS scaffolds shown in Figure 19) corresponding to the junctions between the different viral PSI genes (only reads containing both end and start of the adjacent proteins are shown). Numbers in brackets represent the length of the spacer between the end and start of each adjacent protein pairs (the minus number in psaE>psaD represents overlapping ORFs).

<table>
<thead>
<tr>
<th>psaJF&gt;psaC (144 nt)</th>
<th>psaC&gt;psaA (10 nt)</th>
<th>psaA&gt;psaB (3 nt)</th>
<th>psaB&gt;psaK (5 nt)</th>
<th>psaK&gt;psaE (1 nt)</th>
<th>psaE&gt;psaD (-3 nt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3402488</td>
<td>27779839</td>
<td>27876865</td>
<td>27779457</td>
<td>3799329</td>
<td>27854187</td>
</tr>
<tr>
<td>277645314</td>
<td>3810600</td>
<td>3416584</td>
<td>27801041</td>
<td>27577460</td>
<td></td>
</tr>
<tr>
<td></td>
<td>27834459</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 21. Phylogenetic trees of viral GOS, *Synechococcus* & *Prochlorococcus* PsaC (left) and PsaD (right) proteins. The trees displayed are the ones having the highest likelihood of the 100 trees. Even though each tree has a different number of homologues and a different evolutionary distance between them, a pattern is evident: the cyanobacterial homologues are evolutionary closer among themselves than to any viral homologue, with the exception of viral homologue 8692 (JCVI_SCAF_1096628008692 that contains GOS clone 9984). The displayed GOS leaf names are the last digits of the corresponding accession numbers, Color code is the same as in Figure 16; *Synechococcus* sequences are labeled with blue, *Prochlorococcus* with green and viral GOS in red. *Synechocystis* PCC6803 and *Trichodesmium erythraeum* sequences are shown for reference (in black).

Table 13. Percent cyanophages carrying PSI (*psaA*) or PSII (*psbA*) genes in the different Northern Line Islands viromes.

<table>
<thead>
<tr>
<th></th>
<th>Kingman</th>
<th>Palmyra</th>
<th>Tabuarean</th>
<th>Kiritimati</th>
</tr>
</thead>
<tbody>
<tr>
<td>% phage(^a) with PSI genes</td>
<td>6.5</td>
<td>16.1</td>
<td>54.4</td>
<td>61.6</td>
</tr>
<tr>
<td>% phage with PSII genes</td>
<td>96.0</td>
<td>58.0</td>
<td>111.2</td>
<td>99.8</td>
</tr>
</tbody>
</table>

\(^a\)Cyanophage (myovirus & podovirus) abundance estimates were based on the myovirus viral-capsid-assembly-gene (g20), the myovirus major-capsid-protein-gene (g23), and the podovirus T7-like DNA polymerase markers.
Table 14. Percent occurrence of viral PSI (psaA and psaB) or PSII (psbA and psbD) genes in cyanophages (myoviruses & podoviruses) in Northern Line Islands viromes as indicated by the presence of psaA, psaB, psbA and psbD genes respectively. Number of 454 reads recruited for each gene/virome and the estimated frequency across cyanophages (in brackets) is given. Percents were calculated in relationship to the number of 454 reads recruited with myo-cyanophage g20 gene (encoding for the phage capsid assembly protein) and g23 gene, and with podo-cyanophage DNA polymerase gene (See methods for exact procedure).

<table>
<thead>
<tr>
<th>Line Island site (Dinsdale, Pantos et al. 2008)</th>
<th>psaA</th>
<th>psaB</th>
<th>psbA</th>
<th>psbD</th>
<th>g20</th>
<th>g23</th>
<th>DNA pol.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kingman</td>
<td>1 (6.5%)</td>
<td>2 (13.5%)</td>
<td>7 (96.0%)</td>
<td>1 (14.1%)</td>
<td>10</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Palmyra</td>
<td>21 (16.1%)</td>
<td>20 (15.9%)</td>
<td>36 (58.0%)</td>
<td>11 (18.2%)</td>
<td>85</td>
<td>82</td>
<td>3</td>
</tr>
<tr>
<td>Tabuaeran</td>
<td>71 (54.4%)</td>
<td>57 (45.2%)</td>
<td>69 (111.2%)</td>
<td>45 (74.5%)</td>
<td>98</td>
<td>74</td>
<td>1</td>
</tr>
<tr>
<td>Kiritimati</td>
<td>13 (61.6%)</td>
<td>19 (93.3%)</td>
<td>10 (99.8%)</td>
<td>4 (41.0%)</td>
<td>14</td>
<td>9</td>
<td>3</td>
</tr>
</tbody>
</table>

4.2. Methods

4.2.1. Collecting preliminary set of viral-PSI scaffolds in GOS

In order to find candidate viral-PSI sequences in GOS we have used the following 2-step method: (i) identify all GOS scaffolds containing psaA genes; and (ii) identify viral genes on psaA containing scaffolds. The first step was done by blasting (tBLASTx) the database of GOS scaffolds with several cyanobacterial psaA probes (e-value threshold of 1e-20). In the second step, all psaA containing scaffolds were blasted (tBLASTx) against the refseq-viral database (again, with an e-value threshold of 1e-20). The scan revealed five scaffolds containing both psaA and viral genes. These scaffolds were annotated (BLASTx against the nr database) and found to contain both viral genes such as nrdA and nrdB, as well as PSI genes such as psaA, psaC, psaD and a fusion of psaJ and psaF.

4.2.2. Collecting GOS-PSI clones

The following steps were taken in order to identify viral and non-viral PSI clones in GOS: (i) tBLASTx searches of psaA, psaB, psaC, psaD, psaE, psaF, psaJ, psaK probes against the dataset of GOS scaffolds. This step yielded 1,167 scaffolds. (ii) Identify all reads composing the scaffolds found in the previous stage and their division into clones. Overall, 3,758 reads from 2,147 clones were found (536 single-read and 1,611 pair-end clones). (iii) “In-clone assembly” – the reads of each pair-end clone were aligned (bl2seq) and assembled; 50 N’s were added between non-overlapping reads. (iv) Annotation – an iterative procedure was used for gene discovery and annotation: at each iteration all clones were BLASTxed against nr (e-value threshold=10), first hit for each clone was saved and the clone’s segment in the alignment was replaced with N’s. For each clone, the process halted when no new hits were found. (v) All clones with no PSI hit were removed. Overall we were left with 1,585 GOS clones carrying at least one PSI gene.
4.2.3. The viral-PSI like set (Table 10)

This set contains all GOS-PSI clones whose gene organizations are similar to those observed in the preliminary viral-PSI set but differ from the organization found on sequenced microbial genomes were considered. Overall, a total of 26 clones composed of 49 reads belong to this set. Distribution of clones in this set across different GOS sites is shown in Table 10.

4.2.4. Recruitment against 454 databases – biomes (Table 11)

*Choosing viral PSI recruiters:* all reads from the viral-PSI like set were used, with only the part that contains PSI genes as well as intergenic regions being considered for the analysis.

**Datasets:** all datasets from (Dinsdale, Edwards et al. 2008) were downloaded and divided into the following biomes: coral, fish, freshwater, hypersaline, marine, microbialites and terrestrials.

**Recruitment process:** all recruiters were BLASTned against the different biomes (parameters -F F -r 2 -e 1e-3). For each % identity level (85%, 90% and 95%), all hits that were aligned to at least one of the recruiters with % identity higher than or equal to that level and alignment length ≥ 80bps were collected. The presence of several copies of the same read is one of the known artifacts of 454 sequencing: in order to prevent such instances we have BLASTned each read against all others, and left only one of each set of multiple copies of the same read (reads that share more than 80% of their length at 95% identity were considered to be identical). For each biome and % identity level, the numbers in the table is the size of the resulting set.

4.2.5. Recruitment against 454 databases – clones (Figure 19)

*Choosing clones for the figure:* the following three clones were chosen due to the fact that they cover the whole stretch of viral PSI genes with a few other viral genes on both sides:

- JCVI_TMPL_1061008141871 (forward read: JCVI_READ_1095975009715, reverse read: JCVI_READ_1095964243645, gene organization: NrdB->Hypothetical *prochlorococcus* phage gene->psaF->psaC->psaA),
- JCVI_TMPL_1061007850113 (forward read: JCVI_READ_1095904013142, reverse read: JCVI_READ_1095900032010, gene organization: psaA->psaB),

In addition to the above three clones, all clones that were used as recruiters for the generation of Figure 19b were also analyzed and showed similar patterns of
recruitment (results not shown here), with the exception of clone 9984 (discussed later).

454 datasets: a significant amount of hits from the marine biome arrived from the four viral databases of the Northern Line Islands (Dinsdale, Pantos et al. 2008). Each of these datasets has a corresponding microbial dataset that was sampled from the same location; the availability of both microbial and viral datasets from the same locations made these datasets most suitable for comparative analysis as shown in Figure 19a & b.

Process: A BLASTn search of each of the three clones (with parameters -F F -r 2 -e 1e-3) against the four viral and four microbial Northern Line Islands samples yielded several hundred hits longer than 80 bps with percent identity exceeding 85% in the viral samples (average read length in these datasets is approximately 100 bps), but only a few hits in the four corresponding microbial datasets in a few specific locations on the psaA and psaB genes. This phenomenon can be explained by the presence of small stretches that are conserved on the amino acid level across viral and microbial genomes.

4.2.6. Recruitment of PSI clones against viral and microbial 454 datasets (Figure 19b)

For a complete discussion of recruitment-based separation of viral from microbial sequences refer to Section 3.2.

The complete set of 1,585 GOS-PSI clones carrying at least one PSI gene was BALSTNned against the 4 viral and 4 microbial Line Islands datasets with the same parameters as above (alignment length ≥ 80bps, % identity ≥ 85). Coverage is traditionally defined as (number of bps in recruited reads)/(length of recruiting clone); however in this case we believe that using this using this definition may not be suitable, due to the possibility of multiple copies of the same read, or the possible presence of conserved region across both microbial and viral versions of the same gene. Therefore we used the alternative definition of coverage discussed in Paragraph 3.2.3.

Figure 19b shows the results for the coverage measure: the Y-axis represents overall coverage by microbial samples, while the X-axis represents coverage over viral samples. In addition to GOS clones we have also added psaA and psaB genes from 27 fully sequenced Synechococcus and Prochlorococcus strains. As can be observed, almost all members of the viral-PSI like set are located at the high-viral low-microbial end of the figure, while most of the rest are located at the low-viral end, together with the cyanobacterial genes. Clone 9984 recruited 6% and 0% from the viral and microbial datasets respectively, and probably represents a phage type that is not abundant in the Northern Line Islands. Its gene organization is different from all other viral-PSI clones (psaA<psaC<psaD)>hypothetical Prochlorococcus phage>T4 neck protein). Several clones that were not identified as viral due to their gene organization show
recruitment patterns similar to those of the viral-PSI clones. A close inspection of each of them shows that they all carry either \textit{psaA}, \textit{psaB} or both genes and therefore could not be identified as viral due to their gene organization.

The presence of two distinct clusters of PSI clones, one of them contain all (except one) suspected viral-PSI clones with high coverage from the viral samples while the other contains all cyanobacterial genes and most of the other GOS clones, supports our claim that the observed uniquely-organized PSI genes are indeed viral. Interestingly, microbial clones did not recruit as much reads as viral one. This could be the result of different \textit{Synechococcus} and \textit{Prochlorococcus} strains living in the Northern Line Islands than those that were cultured and seen in the GOS samples. Also, it is possible that the Northern Line Islands microbial samples are dominated by organisms other than cyanobacteria; still, the fact that most of these clones recruited more microbial than viral reads, while the viral clones recruited very few microbial reads, means that they are likely microbial.

### 4.2.7. Gene organization analysis (Figure 18)

**Obtaining viral and microbial GOS Clones:** all clones in the "viral cluster" of Figure 19b (>50% coverage by viral samples, <15% coverage by microbial samples) were considered as viral, while the rest were considered to be microbial.

**Obtaining \textit{Synechococcus} and \textit{Prochlorococcus} genomes:** 27 fully sequenced and annotated genomes were retrieved from the IMG system (Markowitz, Szeto et al. 2008) (\texttt{http://img.jgi.doe.gov/}) and used for the generation of Figure 18a.

**Constructing the organization graphs:** The following criteria were used to identify neighboring genes: (i) the two genes must be on the same clone; (ii) the genes must be on the same strand; (iii) no other gene should be located between the two genes (also on the opposite strand); and (iv) the distance between the two genes must be less than 150 bps. Once a list of neighboring genes was prepared, a transition matrix between all PSI genes+\textit{nrdA}, \textit{nrdB} and \textit{hyp} (hypothetical gene between \textit{nrdB} and \textit{psaJF}) was created and used to generate the figure.

### 4.2.8. Estimation of PSI and PSII gene abundance in phages (Table 13 and Table 14)

Concept and rationale: the goal of this analysis is to estimate the abundance of phages carrying PSI genes in each of the Northern Line Islands stations. Assuming that the genes are carried by cyanophages (as indicated by the top hits to the genes), and assuming that the PSI genes appear once on the phage genomes, it is possible to estimate their abundance by normalization to known single copy genes on the genome of cyanophages. While this process may not be very accurate, it does give a rough estimation of the abundance; as a control we have also estimated the abundance of PSII genes \textit{psbA} and \textit{psbD}.

**Genes:** \textit{psaA} and \textit{psaB} were taken as representatives for PSI. These genes have similar lengths, they are relatively long with visible differences at the DNA level.
between microbial and viral copies. \textit{psbA} and \textit{psbD} represented PSII: \textit{psbA} was shown to be present in 88\% of cyanophages in culture (Sullivan, Lindell et al. 2006), while \textit{psbD} appears in smaller frequencies (~50\%). For normalization we have used \textit{g23} and \textit{g20} (myoviruses), and DNA polymerase (podoviruses).

\textbf{Probes:} for \textit{g20} and \textit{g23} we have used copies from S-PM2, Syn9, P-SSM4 and P-SSM2. For DNA polymerase copies from Syn5 and P60 were used. Copies of \textit{psaA} and \textit{psaB} were extracted from our set of viral-PSI clones as well as from all fully sequenced \textit{Synechococcus} and \textit{Prochlorococcus} strains while \textit{psbA} and \textit{psbD} were collected from cultured cyanophages (available in genbank) and fully sequenced \textit{Synechococcus} and \textit{Prochlorococcus}.

\textbf{Recruitment process:} the following procedure was used for each viral Northern Line Island dataset and for each gene:

1. Perform a tBLASTx search for all gene probes against the dataset (BLAST parameters: -F F -e 1e-3). Collect all hits.

2. Remove all but one identical reads. Reads are considered to be identical if they share at least 85\% of their length at 95\% identity.

3. Perform a BLASTx search for all reads from the previous stage against nr. Remove all reads that did not yield the target gene as their top hit.

4. Determine origin of species:
   a. For \textit{g20}, \textit{g23} and DNA polymerase: perform a BLASTn search against refseq viral, remove all reads which did not yield cyanophages as their top cultured hits.
   b. For \textit{psbA} and \textit{psbD}: perform a BLASTn search against both refseq microbial and a dataset of all viral \textit{psbA} and \textit{psbD} used as probes, remove all reads which received better similarity for refseq microbial with respect to the probes, or received a best viral hit that is different from the target gene (e.g. \textit{psbA} instead of \textit{psbD}). Comparison between alignments was done by comparing the % identity multiplied by the alignment length of the two alignments and taking the highest one. Note that for small databases composed of specific type of sequences the BLAST statistical model may not work well due to background statistics that is very different from the one assumed by BLAST, and therefore we preferred not to rely on the BLAST e-values.
   c. For \textit{psaA} and \textit{psaB}: similarly to Figure 19b.

   In cases of alignments shorter than 50bps against one of the DNA databases, affiliation of best hit in the search against nr was used.

5. Count the remaining reads.
**Frequency estimation:** given that a gene $p$ of length $L(p)$ was found on $R_p$ reads in our database with an average read length $g$ and a minimum segment of $T$ bps from the gene required in order for the gene to be discovered and a coverage of $\alpha$ it follows from Chapter 2 (Sharon, Pati et al. 2009) that

$$\hat{C}_p = \frac{R_p}{2\alpha \cdot (L(p) + g_p - 2T)}$$  \hspace{1cm} (29)

Is an estimator of $C_p$, the number of $p$'s copies in the sample. In our case $g \approx 100$ and $T \approx 50$, therefore we get

$$\hat{C}_p = \frac{R_p}{2\alpha \cdot L(p)}$$  \hspace{1cm} (30)

Following, if $p$ is normalized by another gene $q$ of length $L(q)$ then the frequency of $p$ is given by

$$frequency(p) = \frac{\hat{C}_p}{\hat{C}_q} = \frac{R_p \cdot L(p)}{R_q \cdot L(q)}$$  \hspace{1cm} (31)

For g23 and g20 it is assumed that

$$C_{g_{20}} = C_{g_{23}} = \text{number of myoviruses in the sample}$$

And therefore we use the average of $\hat{C}_{g_{20}}$ and $\hat{C}_{g_{23}}$ in order to get an estimation of the number myoviruses. Adding the estimator for the number of podovirus DNA polymerase $\hat{C}_{DNAPol}$ we get

$$frequency(p) = \frac{\hat{C}_p}{\hat{C}_{DNAPol} + (\hat{C}_{g_{20}} + \hat{C}_{g_{23}})/2} =$$

$$= \frac{R_p}{L(p)/[L(DNAPol) + (R_{g_{20}}/L(g_{20}) + R_{g_{23}}/L(g_{23}))/2]}$$  \hspace{1cm} (32)

which is the formula used for frequency estimation for each of the photosynthetic genes.
Chapter 5.

Discussion

Metagenomics is a rapidly growing field that has become a standard technique for studying environments and microbes that cannot be cultivated. The ability to acquire community-level view of microbes can be used for understanding ecological processes that were previously not understood. Community structure and dynamics, cooperation between species, response to environmental changes and even evolutionary processes that involve the gain and loss of functions can now be studied for the first time. In addition, microbes and viruses that cannot be cultivated are now within reach; in many cases these are the dominant species in their environments and therefore their study is crucial in order to fully understand life in these environments. The flux of discoveries and insights that were acquired using metagenomics is on the rise and expected to increase in the next few years. From the personal point of view, I feel fortunate to become involved in metagenomics at this time and looking forward for the many exciting discoveries that are still ahead of us.

In this dissertation I have described my contribution to the field of metagenomics, both from the computational and the biological aspects. Chapter 2 discussed the statistics of functional analysis of metagenomes, the process in which the abundance of different functional capabilities at the studied environment is studied (also described at (Sharon, Pati et al. 2009; Bercovici, Sharon et al. 2010)). All the methods suggested to-date use the proportion between a gene’s read-count and the total number of reads for computing the relative abundance of the gene in metagenomes and, as a result, suffer from the read-counts bias. Our methods are the first to correct this problem, allowing for more reliable and credible analyses of this kind of data. At the gene-family level this is achieved by offering a comprehensive statistical framework, based on sound theoretical foundations, that – with reasonable assumptions – allows us to assess the significance of the evidence for the presence of a protein family of interest in a given genome. Moreover, we present the most extensive evaluation performed to-date of such methods, which includes – for the first time – validation on real data. Our results suggest that the proposed framework provides a good approximation to the statistics of gene families in WGS projects. Observed estimation errors may be explained by reasons that are related to the definition of gene families. For example, the assignment of a gene to a family is hardly ever to the whole length of the family but rather to some part of it. This may result from the target family carrying several domains, while the assigned gene may carry only one of them. However, when we compute the expected frequency we consider the whole family’s length, which is clearly unrealistic. The use of domain-specific databases, such as Pfam, may address this problem and improve frequency estimations.
At the pathway level we proposed two models reflecting two different assumptions regarding the sharing of genes among pathways. The two models eliminate biases resulting from variations in number of genes across pathways and also biases resulting from variation in genes’ lengths. Our methods performed much better with respect to predicting relative abundance of pathways. Each of our two methods was shown to have its own strength: the pathway intersection method outperforms the other approaches in predicting pathway abundances when focusing on lowly abundant pathways; the independent pathways method is superior in ranking pathway abundances for highly abundant pathways. Both our methods performed only slightly better than the read-count method when used for functional comparison, despite the failure of the later in the second task of predicting the absolute frequencies of the different pathways. One possible explanation for this behavior is that frequency estimation biases of specific pathways tend to be similar in both compared datasets and thus cancel each other when computing the relative abundances. For example, the relative abundance of a gene family or a pathway whose members are relatively long is likely to be overestimated by the read-count method in both samples. Such mutual compensation does not hold in the general case, suggesting that a more robust method is in place.

The pathway intersection method relies on the availability of single copy genes that are present in the vast majority of species in the studied environment. Single copy genes were used in the past as phylogenetic markers (Yoossoph, Sutton et al. 2007) and for estimating gene abundance (Rusch, Halpern et al. 2007; Yutin, Suzuki et al. 2007; Loy, Duller et al. 2009). There are several families of single copy genes that are known to be present across all known bacterial species, but these families are not present in Archaea and Eukaryotes. Hence, the pathway intersection method is more appropriate for environments in which the vast majority of sampled microbes are bacteria such as marine environments, but is likely to yield skewed frequencies when applied to environments in which either Archaeal or Eukaryotic species are abundant (such as acid mine drainage).

Functional characterization of metagenomic data such as that discussed in this dissertation depends, first and foremost, on the quality of the employed pathway annotation data. Specifically, all pathway analysis methods rely on the basic assumption that a pathway is a coherent functional module that is either entirely present or absent in an organism. However, pathways defined in databases such as KEGG and MetaCyc do not fully address this requirement, and in many cases have only a fraction of their genes actually present in many species. Future advances in pathways curation are expected to significantly improve the outcome of the presented methods.

In Chapter 3 I discussed methods for separating viral from microbial sequences and for studying gene organizations. These methods, while being heuristic, proved to be valuable and very useful in some of my works (Sharon, Tzahor et al. 2007; Kagan, Sharon et al. 2008; Sharon, Alperovitch et al. 2009). Separating viral from microbial sequences is part of the broader binning problem in which we aim to cluster
metagenomic sequences based on their taxonomic relevance. Unlike most of the currently available binning algorithms which classify sequences based on the availability of related fully sequenced genomes, we aim to classify sequences both with and without such reference genomes. I have described two different protocols for tackling this problem. The first one, which was not originally suggested by me, determines the origin of the sequence based on the gene contents of the classified sequence. The second one, which is original to this work, is based on similarity to sequences whose origin is known, as is the case in the Northern Line Islands samples (Dinsdale, Pantos et al. 2008). Results produced by both approaches were consistent despite the fact that the two methods rely on different datasets and different sequence type (peptides for the gene-contents based method vs. DNA for recruitment). While not being a proof for correctness by itself, this fact increases our confidence in the method. In addition, I have thoroughly analyzed the Northern Line Islands datasets in order to test their suitability for recruitment-based classification and found the data from the Kingman, Palmyra and Tabuaeran to be reliable while the data from the Christmas Island was found to be problematic. Currently, the number of databases that are suitable for recruitment-based classification is relatively small, despite the availability of many viral and microbial metagenomes (e.g. the data used in (Dinsdale, Edwards et al. 2008)). Using viral and microbial datasets from different locations is possible but may yield biases. For example: some microbial DNA can be found in viral samples (and vice versa) even when the stringent lab measures are taken for separating viral from microbial material. Despite this fact DNA is expected to be correctly classified when a pair of viral and microbial samples is available from each location, since much more similar sequences are expected to be found in the correct sample with respect to the wrong one. When only one sample, either viral or microbial, is available from a certain position and the target sequence recruits DNA from this sample, there is always a possibility that the recruited material is wrong, and that much more similar DNA would have been found the complementary sample (either viral or microbial).

Gene organization analysis still heavily relies on manual work. The tools developed by me automate the procedure for the most part, but in order to conduct large-scale studies of gene organizations that are conserved across several species in large datasets such as GOS we still need to develop sophisticated algorithms that will be able to accurately deduce gene organizations. The development of such algorithms is not trivial, due to the fact that each DNA read usually holds no more than two genes, and the fact that the same gene may appear in several contexts (as in the case of the trpE and trpG genes in the tryptophan pathway, (Kagan, Sharon et al. 2008)). It might be possible to separate gene contexts by using DNA signature or simply by sequence similarity. For examples: trpG and trpE genes that appear in the context of the trp operon may be identified as such assuming that the intra-trp sequence alignment produces higher alignment scores than the scores resulting from alignment against the same genes in other contexts than trp.

Chapter 4 described the main biological discovery in which I was involved, that is the discovery of Photosystem I gene cassettes on viral genomes (Sharon,
Alperovitch et al. 2009). Using the methods described in Chapter 3 I was able to prove the viral origin of the sequences and the expected organization of the genes inside the cassette. This work was done in close cooperation with Ariella Alperovitch from Prof. Béjà’s laboratory that conceived the idea for this project and validated my findings by conducting long PCR of the seven PSI and the neighboring viral \textit{nrdb} genes. PSI is the second complex involved in oxygenic photosynthesis in plants, algae and cyanobacteria. Our finding arrives a few years after the discovery of the PSII genes \textit{psbA} and \textit{psbD} on the genomes of cyanophages (Mann, Cook et al. 2003). The discovery of PSI gene cassettes on the genome of viruses is fascinating from several aspects. First, this is the first discovery of this amount of genes that belong to the same biological system on viral genomes (seven, one of them is a novel fusion of the genes \textit{psaJ} and \textit{psaF}). In principle, these genes are sufficient for encoding all proteins required for a fully functional PSI, a fact that raises many questions with respect to the role these genes play. Second, the organization of the genes we have found is unique and was never observed neither on the genomes of cyanobacteria nor on the genomes of plants. This fact raises questions as for the acquisition of the PSI genes by viruses: did it happen during a single kidnapping event, or by a series of events? At the current time neither possibility is supported by available data. The fact that the seven PSI genes were never observed in a single cassette on any genome do not support the possibility of a single kidnapping event, at least from any known genome. In the event of multiple kidnapping events, on the other hand, we would expect subsets of the PSI genes to be beneficial for the viruses as well. However, to-date, the only cases of PSI genes on viral genomes that are not in line with the PSI cassettes are the cases of clone 9984 and a few cases of the \textit{psaJ} gene that were found by us in other contexts than the PSI gene cassettes (result to be reported as part of the VirMic project, see Section 3.4). These cases are anecdotal, and require more support in order to deduce multiple kidnapping events. At the global level the finding of PSI genes on viral genomes raises questions as for the type of microbial genes acquired by viruses and the dynamics of this process. These issues will be addressed by us in the VirMic project, described briefly in Section 3.4.

Our results (Table 13 and Table 14) suggest that PSI and PSII genes may be found together on the same genomes. Recently we have found evidences that this is in fact the case, including one clone from the GOS Indian Ocean dataset that contains both the PSI \textit{psaB} gene and the PSII \textit{psbA} gene. This clone do not contain viral genes, however it achieves high recruitment from the viral fraction of the Northern Line Islands datasets comparing to low coverage from the microbial fraction. This finding raises additional questions as for the evolutionary advantages of carrying these genes, and the contexts in which the PSI genes can be found.

The future of metagenomics will be largely impacted by progression in both sequencing technology as well as computational methods for analyzing this data. The main computational challenge facing metagenomics is, without doubt, the need to handle ever increasing amounts of data resulting from NGS technologies. For example, the amount of data that can be generated in a single run of the Illumina Solexa genome analyzer more than triples the amount of DNA in the Atlantic/Pacific
Ocean leg of the GOS expedition (Rusch, Halpern et al. 2007; Yooseph, Sutton et al. 2007), published just three years ago. The throughput continues to grow rapidly, a trend that is expected to continue with the improvement in chemical processes and the introduction of new technologies such as single molecule sequencing (Braslavsky, Hebert et al. 2003; Eid, Fehr et al. 2009). Computing capabilities and computational methods do not progress at the same pace, a fact that already poses difficulties even with respect to problems that were considered to be solved such as data storage, alignment against databases, assembly and more. Short read length is another issue, which requires the adjustment and development of algorithms that will fit reads shorter than 100 bps (comparing to ~800 bps in Sanger sequencing and 500 bps in 454 pyrosequencing). This problem is expected to decrease with the improvement in chemistry that will enable longer.

The increased throughput provides us with the opportunity to gain deeper insights into microbial communities that was not previously possible. Ultra-high throughput sequencing may make it possible to sequence and assemble complete genomes directly from the environment, thus eliminating entirely the need to cultivate organisms prior to the sequencing of their genomes. This seems to be unfeasible considering today’s read lengths, but may become feasible with the introduction of improved chemistry and new technologies that may be able to generate a few Kbp and even a few tens Kbp long reads (Eid, Fehr et al. 2009). In addition, the sequencing of community transcriptomes (metatranscriptomics) and proteomes (metaproteomics) that already takes place is expected to increase and add important dimensions to the study of microbial communities. All these trends promise that metagenomics will continue to be an exciting field in the coming years, and that the flux of discoveries resulting from metagenomics studies will grow.
References


