J. Math. Biol. (2012) 64:697–711
DOI 10.1007/s00285-011-0428-2

Mathematical Biology

A non-negative matrix factorization framework for identifying modular patterns in metagenomic profile data

Xingpeng Jiang · Joshua S. Weitz · Jonathan Dushoff

Received: 4 October 2010 / Revised: 18 March 2011 / Published online: 1 June 2011

Abstract Metagenomic studies sequence DNA directly from environmental samples to explore the structure and function of complex microbial and viral communities. Individual, short pieces of sequenced DNA ("reads") are classified into (putative) taxonomic or metabolic groups which are analyzed for patterns across samples. Analysis of such read matrices is at the core of using metagenomic data to make inferences about ecosystem structure and function. Non-negative matrix factorization (NMF) is a numerical technique for approximating high-dimensional data points as positive linear combinations of positive components. It is thus well suited to interpretation of observed samples as combinations of different components. We develop, test and apply an NMF-based framework to analyze metagenomic read matrices. In particular, we introduce a method for choosing NMF degree in the presence of overlap, and apply spectral-reordering techniques to NMF-based similarity matrices to aid visualization. We show that our method can robustly identify the appropriate degree and disentangle overlapping contributions using synthetic data sets. We then examine and discuss the NMF decomposition of a metabolic profile matrix extracted from 39 publicly available metagenomic samples, and identify canonical sample types, including one associated with coral ecosystems, one associated with highly saline ecosystems and others. We

Electronic supplementary material The online version of this article (doi:10.1007/s00285-011-0428-2) contains supplementary material, which is available to authorized users.

X. Jiang · J. Dushoff
Department of Biology, McMaster University, Hamilton, Ontario, Canada

J. S. Weitz
School of Biology and School of Physics, Georgia Institute of Technology, Atlanta, GA, USA

J. Dushoff (✉)
M. G. DeGroote Institute for Infectious Disease Research, McMaster University,
Hamilton, Ontario, Canada
e-mail: dushoff@mcmaster.ca

also identify specific associations between pathways and canonical environments, and explore how alternative choices of decompositions facilitate analysis of read matrices at a finer scale.

Keywords Non-negative matrix factorization · Overlapping clusters · Metagenomics · Metabolic profile · Spectral reordering · Microbial ecology

Mathematics Subject Classification (2010) 15A23 (Factorization of matrices) · 92D40 (Ecology)

1 Introduction

Metagenomic studies sequence DNA directly from environmental samples. Direct sequencing allows for characterization of microbial communities without the need to culture individual species (Turnbaugh and Gordon 2008). Metagenomic studies are rapidly expanding (e.g., Tyson et al. 2004; Tringe et al. 2005; Gill et al. 2006; Warnecke et al. 2007; Sogin et al. 2006; Rusch et al. 2007), with the expectation that they will provide deep insights into the function and evolution of microbial ecosystems (Hemme et al. 2010), including insights about the importance of humans' resident microbial communities for health and disease (Peterson et al. 2009).

Metagenomic studies classify "reads"—individual, short pieces of sequenced DNA—into (putative) taxonomic or metabolic groups which are then analyzed for patterns across samples. Processing (Mavromatis et al. 2007; Richter et al. 2008; Quince et al. 2009; Morgan et al. 2010) and classifying (McHardy et al. 2007; Huson et al. 2007; Kislyuk et al. 2009; Kelley and Salzberg 2010) these reads is a complex process, which we will not discuss here. In particular, metabolic profile matrices (MPMs), which describe how reads identified with particular metabolic pathways are distributed across samples, are used to probe the metabolic function of a community even if the majority of the organisms cannot be isolated and cultured (Handelsman 2004; Dinsdale et al. 2008; Gianoulis et al. 2009; Parks and Beiko 2010).

As the availability of experimental data rapidly increases, and computational methods for classification improve, an important challenge is to develop statistical methods to analyze such profiles. Analysis typically relies on a combination of: dimensional-reduction techniques, including principle component analysis (PCA) (Turnbaugh et al. 2009; Willner et al. 2009), multidimensional scaling (Willner et al. 2009) discriminant analysis (Dinsdale et al. 2008), and canonical correlation analysis (Gianoulis et al. 2009), to look for simple patterns and aid visualization; and clustering and classification techniques including hierarchical clustering (Turnbaugh et al. 2009; Willner et al. 2009), which put samples (or pathways) into groups.

Non-negative matrix factorization (NMF) provides an exciting alternative to traditional dimensional-reduction methods (Lee and Seung 1999). In NMF, samples are represented by non-negative combinations of canonical components. The structure found by NMF methods is thus often very different from, and more intuitive to interpret than, that of more traditional eigenvector-based methods, such as PCA. By constructing samples as positive combinations, NMF also has the potential to "disentangle" canonical components which often overlap to create particular community samples.

$$\begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1s} \\ X_{21} & X_{22} & \cdots & X_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{ps} \end{pmatrix} \approx \begin{pmatrix} W_{11} & \cdots & W_{1k} \\ W_{21} & \cdots & W_{2k} \\ \vdots & \ddots & \vdots \\ W_{p1} & \cdots & W_{pk} \end{pmatrix} \begin{pmatrix} H_{11} & H_{12} & \cdots & H_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ H_{k1} & H_{k2} & \cdots & H_{ks} \end{pmatrix}$$

Fig. 1 The NMF method approximates the matrix X, whose p rows are metabolic pathways, and s columns are samples, as the product of $p \times k$ and $k \times s$ non-negative matrices, for an appropriately chosen "degree" k, usually relatively small

The price of this more biologically intuitive decomposition is that the factorization is approximate, and components depend on the dimension of the decomposition, requiring greater care in interpretation.

NMF approximates a data matrix X with non-negative entries as the product of two non-negative matrices W and H (see Fig. 1). In a metagenomic example, X typically has p rows corresponding to metabolic pathways, and s columns corresponding to environmental samples; the entries would then represent the amount of evidence for a certain kind of pathway in a certain sample [often a number of DNA reads (Gianoulis et al. 2009)]. The matrix W is $p \times k$ whereas the matrix H is $k \times s$. Hence, each column of W has one entry for each of the p metabolic pathways; we can thus think of W as a collection of k "canonical samples", where k is the "degree" of the factorization. In this interpretation the s columns of H give each of the s environmental samples as linear combinations of these canonical samples. In the dual interpretation, the k rows of H are "canonical pathways" and the p rows of W give the observed pathways as linear combinations of them.

The modern approach to NMF (Lee and Seung 1999) has been used in a wide range of data mining and pattern recognition applications (Lee and Seung 1999; Montano et al. 2006) in the last decade, as well as in large-scale biological data analysis (Kim and Tidor 2003; Devarajan 2008; Brunet et al. 2004; Kim and Park 2007). Advantages of this method include its ability to simultaneously cluster the columns and rows of a data matrix (bi-clustering) (Saez et al. 2006; Montano et al. 2006), and unsupervised discovery of hierarchical structures (Brunet et al. 2004). There is also evidence that NMF can reveal overlapping structures in data (Brunet et al. 2004; Zhang et al. 2007).

Here we develop a NMF-based framework for extracting biologically relevant patterns from MPMs. We begin with a brief review of the mathematical theory underlying NMF. Next, we describe a novel means of choosing the "degree" of decomposition in NMF, using a robust model selection method that accounts for the possibility of overlapping structures in the data matrix. We also propose an ordering method to visualize the overlapping and hierarchical structure embedded in MPMs. We demonstrate the validity of our method on synthetic and empirical data. We show that our method can robustly identify the correct degree within synthetically generated data matrices with overlapping groups of canonical samples.

We then apply our model to a set of 39 metagenomic samples comprising >4,000,000 sequence reads. In doing so we hope to shed some light on Simon Levin's recent call to consider the information found in metagenomic studies in order to quantify how information "is distributed over the biota, and why specific genes are associated with particular regions of the ecosystem" (Levin 2006). We find that our method identifies canonical components corresponding to predominantly hyper-saline environments, coral environments, fish environments and others. We identify metabolic pathways with a high degree of specificity to canonical samples and discuss the biological interpretation of our findings in light of previous efforts to categorize what is similar and what is distinct about metagenomes across a broad range of habitats.

2 Datasets and methods

2.1 Datasets

We consider sequence data from metagenomic projects available on the metagenomics RAST server (MG-RAST) (Meyer et al. 2008), by Dinsdale et al. (2008) as well as an additional four hyper-saline samples from Desnues et al. (2008). We utilized the most detailed pathway breakdown available (of three provided by MG-RAST), using default parameters. In doing so, every read was assigned to a single (putative) metabolic pathway when possible. We downloaded all available data from Dinsdale et al. (2008) as of March 2010, but excluded two subterranean samples, as well as one anomalous marine sample from an area with very high nutrient levels. The resulting 39 samples are from seven biomes: hyper-saline (13), coral (7), marine (7), freshwater (4), fish (4), terrestrial animal (2), and "microbialites" (benthic microbial communities, 2). Of these, 24 samples are from aquatic environments, providing a gradient across salinity levels.

We arrange our data as a matrix X, whose p rows are pathways and s columns are samples. For our metagenomic dataset, we eliminated pathways with fewer than 0.01% of the total reads, leading to a read matrix with $p = 558$ rows and $s = 39$ columns. Each column of the matrix was normalized by dividing by the sum of the column; thus each entry is the *proportion* of reads from a given sample that correspond to a particular pathway (Gianoulis et al. 2009).

We used the statistical environment R (R Development Core Team downloaded 2010) to generate simulated data and to do our analyses. These analyses also made use of the NMF package in R (Gaujoux and Seoighe 2010).

2.2 Non-negative matrix factorization

Given the $p \times s$ matrix X, we want to find matrices W and H, (with dimension $p \times k$ and $k \times s$, respectively, where k is the *degree* of our factorization) so that $WH \approx X$. We do this by minimizing an objective function under the constraint that W and H must be non-negative. The objective function we use is the KL divergence (Lee and Seung 1999), which is frequently used in gene expression analyses (Brunet et al. 2004).

The NMF package (Gaujoux and Seoighe 2010) picks random starting values for W and H and then updates iteratively to find a *local* minimum of KL divergence. We repeated this process for 100 different starting points for each value of k, and used the results to evaluate the quality and stability of the factorization at this degree. To go forward, we used the factorization which minimized the objective function.

2.3 Model selection

NMF is not a hierarchical method; each component depends on the choice of degree, k, and thus this choice should be made with care. There are two basic approaches to

$$\begin{pmatrix} \bar{H}_{11} & \cdots & \bar{H}_{k1} \\ \bar{H}_{12} & \cdots & \bar{H}_{k2} \\ \vdots & \ddots & \vdots \\ \bar{H}_{1s} & \cdots & \bar{H}_{ks} \end{pmatrix} \begin{pmatrix} \bar{H}_{11} & \bar{H}_{12} & \cdots & \bar{H}_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{H}_{k1} & \bar{H}_{k2} & \cdots & \bar{H}_{ks} \end{pmatrix} = \begin{pmatrix} 1 & S_{12} & \cdots & S_{1s} \\ S_{21} & 1 & \cdots & S_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ S_{s1} & S_{s2} & \cdots & 1 \end{pmatrix}$$

Fig. 2 The symmetric matrix $S = \bar{H}^T \bar{H}$ is an $s \times s$ matrix showing the similarity of different samples in the projection defined by an NMF decomposition. The concordance index C reflects the stability of this matrix across different realizations of the decomposition approximation, and is used to select good values of the decomposition degree k

making this choice: we can evaluate either the *quality* of the factorization (ie., how similar is WH to X?) or its *stability* (how similar are the factorizations $W_j H_j$ that emerge from different realizations of the iterative solution process to each other?).

We are not aware of any work which has used quality to select k for NMF. In general, we would expect quality to be an increasing function of k, so a “good” value of k would not be a local maximum for quality, but instead a point where the response of quality to k changes from being steep to shallow (ie., a “good” value of k provides a substantially better approximation than nearby smaller values, but only a slightly worse approximation than nearby larger values).

Conventional degree-selection methods are based on using each factorization $W_j H_j$ to *categorize* samples into k groups, corresponding to the k canonical samples. A degree choice is stable if it repeatably produces similar categorizations, as measured by the cophenetic correlation coefficient (CPCC) (Brunet et al. 2004), or by the dispersion coefficient (Kim and Park 2007).

We are not primarily interested in NMF as a categorization tool, nor do we necessarily expect stable categorizations, given that canonical samples may “overlap”—in other words, our observed samples may be intermediate between two or more canonical samples. We therefore introduce a method for evaluating stability based on a sample similarity matrix which depends on the decomposition $W_j H_j$, but does not depend on categorizing samples into canonical groups.

We construct our similarity matrix S from the sample-projection matrix H (Fig. 2). We normalize H to $\bar{H}$ by dividing each column by its Euclidean norm. Then $S = \bar{H}^T \bar{H}$ is a symmetric similarity matrix, with ones down the diagonal, and each entry showing the similarity of two samples in the projection given by our NMF decomposition. We then define our “concordance index” $C = 1 - D$, where D is the mean squared difference between off-diagonal entries of S_j obtained from different realizations of the decomposition (Fig. 2).

2.4 Visualization and biclustering using NMF

Biclustering algorithms are used in gene-expression studies to identify patterns of interactions between types of genes and types of samples (Kluger et al. 2003; Madeira and Oliveira 2004); NMF techniques have also been used for this purpose (Saez et al. 2006). Here we suggest a new technique for biclustering, based on the sample similarity matrix S and its “dual”, the pathway similarity matrix $P = \hat{W}\hat{W}^T$. Here the $\hat{W}$ represents the row-normalization of W by dividing each row by the row norm.

We treat these two symmetric, positive, similarity matrices as adjacency matrices of a weighted graph, and applied spectral reordering after applying an "affinity" transformation (Maetschke et al. 2010). Choosing the scale r of the affinity transformation is a complex problem (Zelnik-Manor and Perona 2004; Alzate and Suykens 2010)—we used the value of r that minimized the Laplacian distance criterion for the untransformed matrix. We used these orderings of samples and pathways to reorder our dataset and provide visualizations of clustering structures. The complexity of our algorithms (for both computer time and storage space) is linear in the product $m \times k$, where $m = p \times s$ is the size of our read matrix.

It is also interesting to investigate the similarity of *canonical* pathways and samples. We calculate $k \times k$ canonical similarity matrices, analogously to the similarity matrices above, as $\hat{H}\hat{H}^T$ and $\bar{W}^T\bar{W}$, respectively. Spectral reordering can also be applied to these matrices.

2.5 Specificity

Because NMF can provide canonical representations for environmental samples and pathways simultaneously, it is possible to study the specificity and overlapping pattern of pathways among different canonical components. Each row W_i of W corresponds to a set of k weights describing the profile of a pathway over sample sites as a linear combination of canonical pathways. We define the specificity of a pathway as the sparseness of row vector W_i (Montano et al. 2006): $\sigma(W_i) = \frac{\sqrt{k}-\sum|W_{ij}|/\sqrt{\sum {W_{ij}}^2}}{\sqrt{k}-1}$. The specificity is 1 if a pathway profile can be represented using a single canonical pathway, and 0 if all k canonical pathways are equally represented. We define sample specificity in a similar manner (using columns of H).

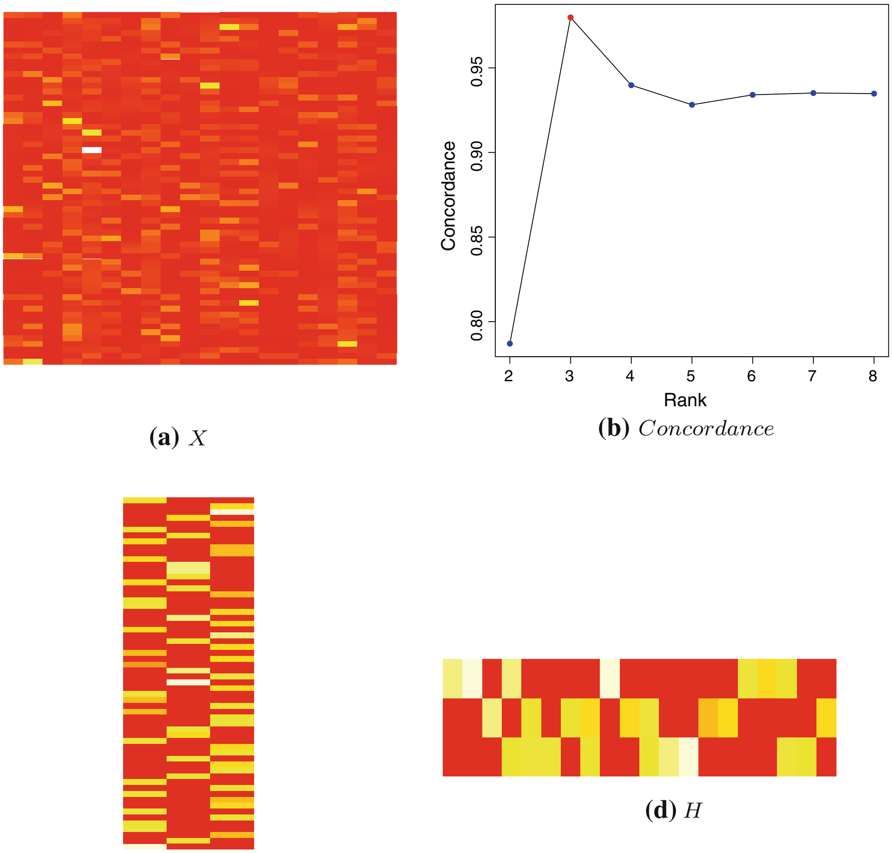
3 Results

3.1 NMF extracts structure from simulated data with overlapping patterns

To facilitate the understanding of our method, Fig. S1 gives a flowchart of the NMF framework that used in this paper. We illustrate an explicit factorization in Fig. 3.

We start with a synthetic metabolic profile (Fig. 3a) in which we embedded a stochastic realization of an overlapping pattern (with 3 "modules" consisting of disjoint pathways, but overlapping samples; see Sect. 2). The concordance plot (Fig. 3b) correctly identifies $k = 3$ as a good degree for the NMF decomposition. The best factorization found for this degree is also shown (Fig. 3c, d). To evaluate the reliability of the method, we applied this method (and two other methods, for comparison) to 100 such random matrices. Our concordance method identified the desired degree of $k = 3$ in 98/100 cases, compared to 81/100 for the dispersion method and 44/100 for CPCC. We also tested synthetic profiles with desired degrees of $k = 4$ and $k = 5$, with similar results (see Figs. S2, S3).

The re-ordering method is illustrated in Fig. 4. We generated the pathway similarity matrix $P = \hat{W}\hat{W}^T$ and the sample similarity matrix $S = \bar{H}^T\bar{H}$. Recall that $\hat{W}$ and $\bar{H}$



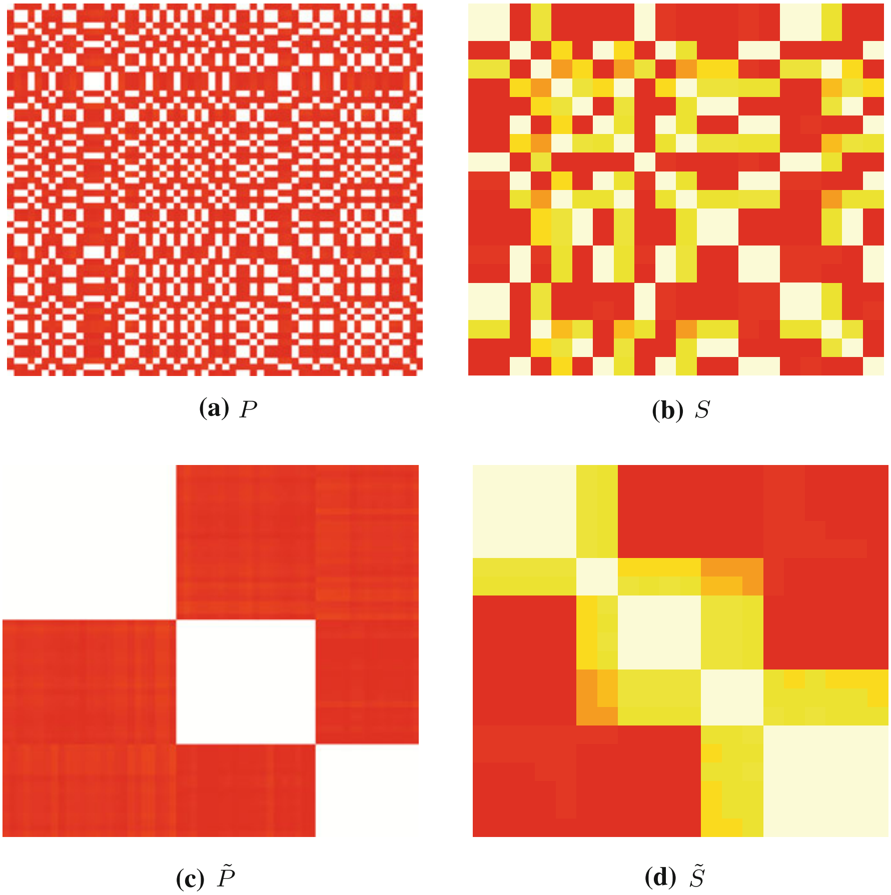
(a) X **(b)** $Concordance$

(c) W **(d)** H

Fig. 3 NMF and model selection. **a** A simulated profile matrix with overlapping modules, and pathways and samples in a random order. **b** The concordance plot shows that $k = 3$ is a stable degree for NMF. We factorized the matrix into three canonical samples (**c**) and associated weightings (**d**). Dually, these could be seen as **c** associated weightings for **d** three canonical pathways (see text). Matrix values are coded by "heat colors" (*red* is low, *yellow* intermediate, and *white* high) in all figures of this paper (color figure online)

are normalized versions of the original matrices (see Sect. 2). We then apply spectral reordering to put as much weight as possible near the diagonal (note that both ordered and original similarity matrices have high values (ones) *on* the diagonal). This calculation gives us an ordering for rows (pathways) and another one for columns (samples), which we can apply going forward. The reordered similarity matrices show clearly the structure we embedded in our data: three modules that don't overlap in pathways ($\tilde{P}$, Fig. 4c), but do overlap in some of the samples ($\tilde{S}$, Fig. 4d). We use the tilde to indicate a matrix which has been reordered using the spectral reorderings from the pathway and sample similarity matrices.

In Fig. 5, we show the re-ordered "canonical" matrices $\tilde{W}$ (Fig. 5a) and $\tilde{H}$ (Fig. 5b). We also show a reordering of the original data matrix $\tilde{X}$ (Fig. 5c), and compare it with the product $\tilde{W}\tilde{H}$ (Fig. 5d). The product shows the data as "filtered" through an NMF


(a) P

(b) S

(c) $\tilde{P}$

(d) $\tilde{S}$

Fig. 4 Reordering based on similarity matrices. We apply spectral reordering to the **a** pathway similarity matrix and the **b** sample similarity matrix to find orderings which put a lot of similarity "weight" near the diagonal. The orderings are applied to the similarity matrices in **c** and **d**; we also apply them below

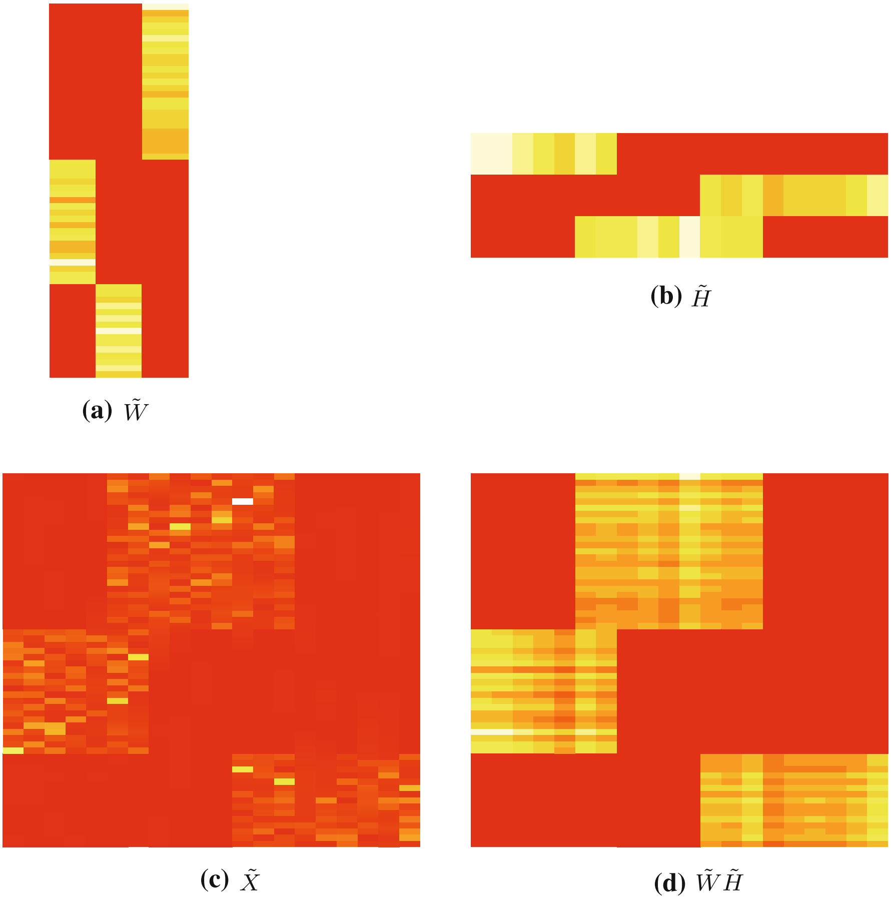
decomposition of the chosen degree (in this case 3). Both filtered and unfiltered versions of the reordered read matrix show clearly the structure that we embedded in a randomized fashion.

In contrast, a classic PCA analysis (Fig. S4) does not identify the overlapping structure of the fake dataset. This is due to the fact that unlike NMF, PCA does not rely on positive combinations of positive components, and thus would not be expected to disentangle overlapping modules (Lee and Seung 1999).

3.2 NMF analysis of metagenomic profile data

3.2.1 Canonical representation of samples and pathways

Here we consider the structure of an MPM comprised of 39 metagenomic samples and 558 pathways. The concordance plot for this MPM shows relatively complex



(a) $\tilde{W}$

(b) $\tilde{H}$

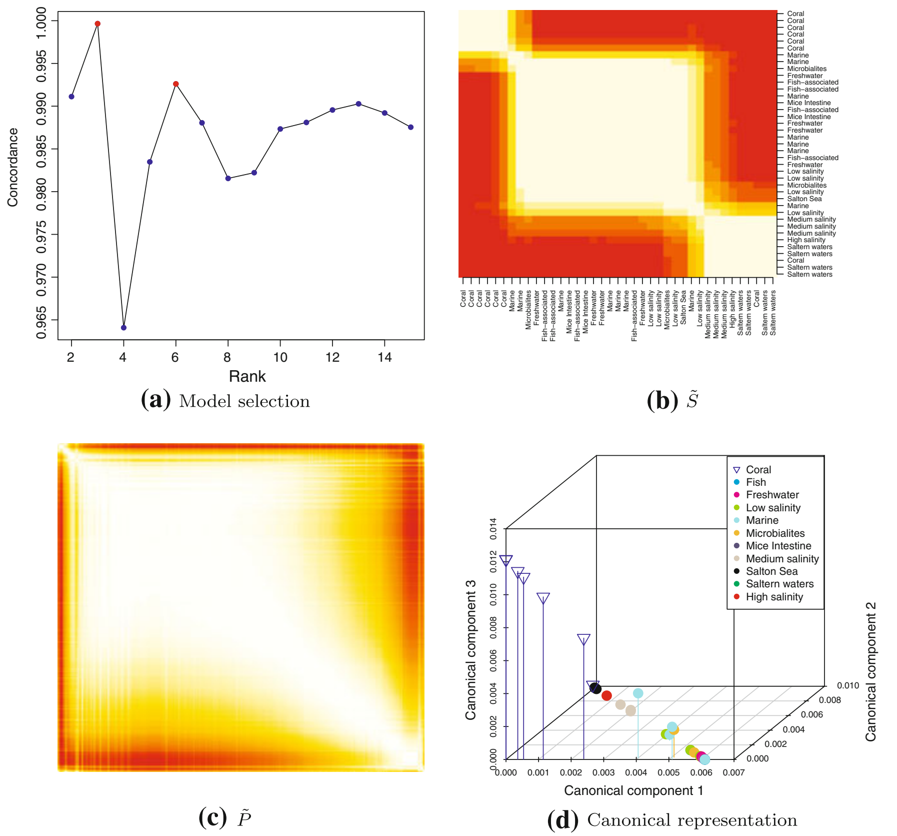
(c) $\tilde{X}$

(d) $\tilde{W}\tilde{H}$

Fig. 5 The canonical pathways (**a**) and samples (**b**), re-ordered using the ordering derived from their respective similarity matrices. We apply both of these orderings to the original MPM (**c**), and to the the "filtered" MPM, given by the product WH (**d**)

structure, with clear peaks at $k = 3$ and $k = 6$ (see Fig. 6a). We first examine the factorization for $k = 3$. The sample similarity matrix $\tilde{S}$ shows three clear clusters of samples with overlap (Fig. 6b). The pathway similarity matrix $\tilde{P}$ is more complex, with a small cluster in the upper left overlapping with the main cluster in the center, and a less-compact cluster in the lower right (Fig. 6c).

We can visualize the matrix H as projections of our samples into a space of "canonical samples" (Fig. 6d)—each of the s columns of H gives a sample approximately as a linear combination of canonical samples (the k columns of W). Many samples lie close to a single axis, while most of the others lie on a plane between two axes (representing overlapping of two types). Canonical component 1 is strongly associated with freshwater and animal-associated samples, but many other samples also lie close to component 1. Component 2 is associated with high salinity, and Component 3 is



(a) Model selection

(b) $\tilde{S}$

(c) $\tilde{P}$

(d) Canonical representation

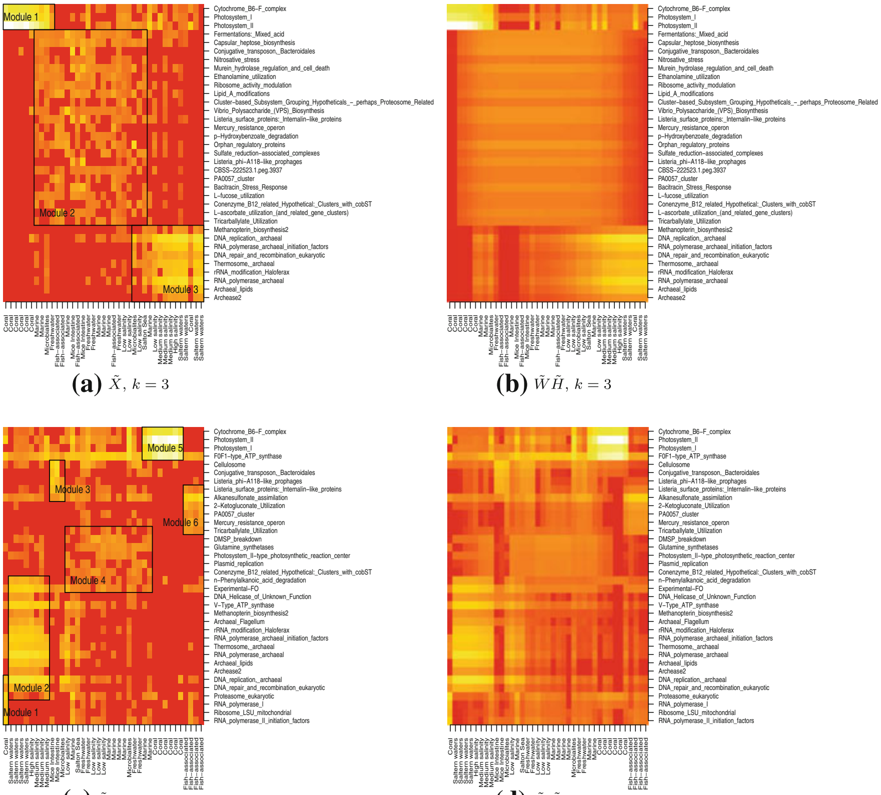
Fig. 6 MPM analysis. **a** The concordance plot shows that $k = 3$ and $k = 6$ are relatively good scales. **b** The sample similarity matrix shows well-defined clusters with overlap. **c** The pathway similarity matrix shows one main cluster, and two less-defined clusters. **d** Samples projected into a space of canonical samples, using the matrix H

associated with corals, but we see a lot of overlap (and one anomalous coral sample that lies close to axis 2).

3.2.2 Identifying adaptive and common pathways of basis environment

Figure S5 shows the MPM, reordered based on the similarity matrices P and S calculated using degree-3 NMF. Many of the pathways are overlapping, and modules are thus hard to pick out. Biological interpretation may be facilitated by separate examination of pathways with high specificity (to look for environment-specific modules) and of pathways with low specificity (to look for pathways that span across different environments). Pathway specificity, based on the three canonical pathways from our NMF decomposition, is shown along the right side.

Figure 7a and b shows the subset of the matrix in Figure S5 corresponding to pathways with high specificity. For clarity of exposition, we have outlined three relatively



(a) $\tilde{X}$, $k = 3$

(b) $\tilde{W}\tilde{H}$, $k = 3$

(c) $\tilde{X}$, $k = 6$

(d) $\tilde{W}\tilde{H}$, $k = 6$

Fig. 7 Modules in the MPM. **a** The MPM reordered using a spectral reordering derived from a degree-3 NMF (specificity $\sigma > 0.99$). **b** The same matrix, but "filtered" as well as reordered. **c** and **d** show analogous matrices, but based on a degree-6 factorization ($\sigma > 0.9$). It is worth noting that **a** and **c** show matrix entries from the original read matrix; only the order (and filtering) is different

distinct modules in Fig. 7a; these are not picked out by the method, we selected them qualitatively. Modules 1 and 3 are small, while module 2 is larger and more diffuse. Module 1 shows an association between coral samples and photosystem-related pathways, unsurprising given that corals are high-light environments where microbial photosynthesis is common. Module 3 shows an association between high-salinity samples and archaeal pathways, consistent with prior work showing archaea associated with high-salinity environments (Hollister et al. 2010). Module 2 shows a larger group of specific pathways, but these are less specific than those associated with the two smaller modules—many of them overlap the samples associated with these modules.

We also see some anomalies. In particular, one sample labelled as coral clusters along the high-salinity axis, and away from other corals. This result warrants further study. The straightforward visualization offered by NMF may help in many cases to identify anomalies like this one, and separate them from overlap samples, which may be misclassified by clustering algorithms without representing true anomalies.

At this scale, it seems that NMF has “pulled out” the coral-photosystem and saline-archaea modules, and left the rest of the reads in a less sharply defined group. This observation underlines the problems and opportunities associated with choosing a degree for NMF. While our synthetic matrix had a single, best scale, the real MPM matrix can be profitably analyzed at various scales.

The next natural scale to view this particular matrix is $k = 6$; at this scale (Fig. 7c, d) the NMF resolves three smaller modules that roughly correspond to the diffuse module 2 seen at $k = 3$, while the smaller modules stay roughly the same (although the anomalous coral sample splits off from the salinity module). Module 4 remains rather large and diffuse, including aquatic samples from freshwater, low-salinity and marine environments. Modules 3 and 6 break out the mouse- and fish-associated samples, respectively. It is interesting that the fish module has pathways that overlap both the mouse and the aquatic modules (despite the fact that the pathways shown here were chosen for specificity). Also interesting is the presence of a “mercury-resistance operon” in the fish module. We also find overlapping between samples in different modules; the sample similarity matrix corresponding to this ordering is shown in Fig. S7.

It is also interesting to examine the pathways with low specificity in Fig. S5. These pathways, found relatively evenly across environments, are shown in Fig. S6. There are a large number of these common pathways, including a variety of basic biological processes such as transport of iron and zinc, poly-hydroxybutyrate metabolism, alanine biosynthesis and many more. The grouping of metabolic pathways into relatively conserved and variable groups will assist in identification of novel organization of microbial metabolism at the sample level.

4 Discussion

We introduced a computational framework based on NMF for exploring patterns contained in read matrices from metagenomic sequencing projects. We showed that our scheme (based largely on existing NMF techniques, but with some innovations) has the ability to: identify appropriate degrees for NMF decomposition; decompose read matrices into “canonical samples” which can be combined to approximate the observed samples, and used to visualize relationships between these samples (with a dual interpretation using canonical classifications); generate similarity matrices which illuminate clustering structure of samples or classifications at a given scale; and re-order the rows and columns of the original matrix in a way that aids visualization of structure. The non-negative nature of the representation facilitates biological interpretation of the results and thus provides a useful complement to other dimension-reduction methods, such as PCA (Turnbaugh et al. 2009) and discriminant analysis (Dinsdale et al. 2008). The NMF framework also provides a direct visualization of overlapping structures, and thus provides a valuable alternative to clustering methods.

The approach outlined here offers a convenient way of extracting structure from metagenomic read matrices, which could be incorporated into metagenomic analysis pipelines, and also applied to NMF methods more broadly. We developed and applied a model-selection method based on the stability of similarity matrices, rather than of classifications, and showed that this approach is suitable for identifying overlapping structures. We also showed that spectral reordering based on NMF provides a convenient method to visualize such overlapping structures present in read matrices. Furthermore, we showed that NMF can be applied at different scales to detect different hierarchical levels of structure.

We applied our NMF scheme to the analysis of 39 publicly available metagenomic data sets, and showed that it provides a convenient method to identify structure in these sets. Our analysis found simple patterns consistent with known biology (e.g., coral ecosystems are distinguished by photosynthetic pathways; high-salinity environments by archaeal pathways), and other patterns perhaps worthy of follow-up (e.g., the association of fish with a group of pathways including a mercury-resistance operon). We also demonstrated the method's ability to identify metabolic pathways which span habitats in the sample set. The method can also help to identify sample outliers worthy of follow-up study, e.g., the finding of a coral ecosystem whose metabolic pathway profile was more similar to that of a high-saline environment.

Unlike PCA components, NMF components are not orthogonal. We thus investigate similarity patterns between our canonical pathways (Fig. S8) and canonical samples (Fig. S9). We find that the overlap patterns are similar, but the canonical pathways (Fig. S8.a) have less overlap (smaller off-diagonal similarity values) than canonical samples (Fig. S8.b). There are interesting overlapping patterns, for instance, the fifth canonical sample is broadly represented across pathways, and has relatively high overlap with other canonical samples. An analogous pattern is seen for the fifth canonical pathway, although the signature is hard to see in the similarity matrix, because the off-diagonal similarity values are low.

Whereas NMF is widely utilized in other fields (Brunet et al. 2004; Montano et al. 2006; Devarajan 2008), to our knowledge this is the first application of NMF to metagenomic data. Although we focused on metagenomic data sets in this paper, both the NMF approach in general, and the framework developed here, have potential applications to a variety of different kinds of biological profiles, including transcriptional profiles and protein expression profiles. In order to facilitate future use and method development, we make all R software used in the current analysis freely available (and modifiable) at http://lalashan.mcmaster.ca/theobio/projects/index.php/NMF.

Acknowledgments The authors thank Simon Levin for inspiration and guidance over the years, and specifically for inspiring us to investigate the ecological interpretation of metagenomic data. We thank Ben Mann for his encouragement. We thank J. Eisen, M. Fortuna, A. Kislyuk, M. Langille and R. Neches for helpful discussions. This work was supported by the Defense Advanced Projects Research Agency under grants HR0011-05-1-0057 and HR0011-09-1-0055. Joshua S. Weitz, Ph.D. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

References

Alzate C, Suykens JA (2010) Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA. IEEE Trans Pattern Anal Mach Intell 32:335–347

Brunet JP, Tamayo P, Golub TR, Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci USA 101:4164–4169

Desnues C, Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B, Ruan Y, Hall D, Angly FE, Edwards RA, Li L, Thurber RV, Reid RP, Siefert J, Souza V, Valentine DL, Swan BK, Breitbart M, Rohwer F (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 452:340–343

Devarajan K (2008) Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. PLoS Comput Biol 4:e100029

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F (2008) Functional metagenomic profiling of nine biomes. Nature 452:629–632

Gaujoux R, Seoighe C (2010) A flexible R package for nonnegative matrix factorization. BMC Bioinform 11:367

Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, Letunic I, Yamada T, Paccanaro A, Jensen LJ, Snyder M, Bork P, Gerstein MB (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. Proc Natl Acad Sci USA 106:1374–1379

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE (2006) Metagenomic Analysis of the Human Distal Gut Microbiome. Science 312(5778):1355–1359. http://10.1126/science.1124234

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68(4):669–685. http://10.1128/MMBR.68.4.669-685.2004

Hemme CL, Deng Y, Gentry TJ, Fields MW, Wu L, Barua S, Barry K, Tringe SG, Watson DB, He Z, Hazen TC, Tiedje JM, Rubin EM, Zhou J (2010) Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. ISME J 4:660–672

Hollister EB, Engledow AS, Hammett AJ, Provin TL, Wilkinson HH, Gentry TJ (2010) Shifts in microbial community structure along an ecological gradient of hypersaline soils and sediments. ISME J 4:829–838

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17(3):377–386. doi:10.1101/gr.5969107

Kelley DR, Salzberg SL (2010) Clustering metagenomic sequences with interpolated Markov models. BMC Bioinform 11. doi:10.1186/1471-2105-11-544

Kim H, Park H (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics 23:1495–1502

Kim PM, Tidor B (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res 13:1706–1718

Kislyuk A, Bhatnagar S, Dushoff J, Weitz JS (2009) Unsupervised statistical clustering of environmental shotgun sequences. BMC Bioinform 10. doi:10.1186/1471-2105-10-316

Kluger Y, Basri R, Chang JT, Gerstein M (2003) Spectral biclustering of microarray data: coclustering genes and conditions. Genome Res 13:703–716

Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791

Levin SA (2006) Fundamental questions in biology. PLoS Biol 4:e300

Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinform 1:24–45

Maetschke SR, Kassahn KS, Dunn JA, Han SP, Curley EZ, Stacey KJ, Ragan MA (2010) A visual framework for sequence analysis using n-grams and spectral rearrangement. Bioinformatics 26:737–744

Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy ACC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, Lapidus A, Grigoriev I, Richardson P, Hugenholtz P, Kyrpides NCC (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods 4:495–500. http://10.1038/nmeth1043

McHardy AC, Garcia Martin H, Tsirigos A, Hugenholtz P, Rigoutsos I (2007) Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods 4(1):63–72. doi:10.1038/NMETH976

Meyer F, Paarmann D, Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinform 9:386

Montano A, Saez P, Chagoyen M, Tirado F, Carazo JM, Marqui RD (2006) bioNMF: a versatile tool for non-negative matrix factorization in biology. BMC Bioinform 7:366

Montano A, Carazo JM, Kochi K, Lehmann D, Marqui RD (2006) Nonsmooth nonnegative matrix factorization (nsNMF). IEEE Trans Pattern Anal Mach Intell 28:403–415

Morgan JL, Darling AE, Eisen JA (2010) Metagenomic sequencing of an in vitro-simulated microbial community. PLoS One 5:e10209

Parks DH, Beiko RG (2010) Identifying biologically relevant differences between metagenomic communities. Bioinformatics 26:715–721

Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, Baker CC, Di F, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C, Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Reed P, Zakhari S, Read J, Watson B, Guyer M (2009) The NIH human microbiome project. Genome Res 19:2317–2323

Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Methods 6:639–641

R Development Core Team (2010) R Project for Statistical Computing. http://www.r-project.org/

Richter DC, Ott F, Auch AF, Schmid R, Huson DH (2008) MetaSimA Sequencing Simulator for Genomics and Metagenomics. PLoS One 3(10):e3373+. http://10.1371/journal.pone.0003373

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcon LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter CJ (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biol 5(3):e77+. http://10.1371/journal.pbio.0050077

Saez P, Marqui RD, Tirado F, Carazo JM, Montano A (2006) Biclustering of gene expression data by Non-smooth Non-negative Matrix Factorization. BMC Bioinform 7:78

Sogin MLL, Morrison HGG, Huber JAA, Welch DMM, Huse SMM, Neal PRR, Arrieta JMM, Herndl GJJ (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere”. Proc Natl Acad Sci 103:12115–12120. http://10.1073/pnas.0605127103

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM (2005) Comparative Metagenomics of Microbial Communities. Science 308(5721):554–557. http://10.1126/science.1107851

Turnbaugh PJ, Gordon JI (2008) An invitation to the marriage of metagenomics and metabolomics. Cell 134:708–713

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI (2009) A core gut microbiome in obese and lean twins. Nature 457:480–484

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428(6978):37–43. http://10.1038/nature02340

Warnecke F, Luginbühl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, Mchardy AC, Djordjevic G, Aboushadi N, Sorek R, Tringe SG, Podar M, Martin HG, Kunin V, Dalevi D, Madejska J, Kirton E, Platt D, Szeto E, Salamov A, Barry K, Mikhailova N, Kyrpides NC, Matson EG, Ottesen EA, Zhang X, Hernández M, Murillo C, Acosta LG, Rigoutsos I, Tamayo G, Green BD, Chang C, Rubin EM, Mathur EJ, Robertson DE, Hugenholtz P, Leadbetter JR (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature 450(7169):560–565. http://dx.doi.org/10.1038/nature06269

Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. PLoS One 4:e7370

Willner D, Thurber RV, Rohwer F (2009) Metagenomic signatures of 86 microbial and viral metagenomes. Environ Microbiol 11:1752–1766

Zelnik-Manor L, Perona P (2004) Self-Tuning Spectral Clustering. In: Eighteenth Annual Conference on Neural Information Processing Systems, (NIPS)

Zhang S, Wang RS, Zhang XS (2007) Uncovering fuzzy community structure in complex networks. Phys Rev E Stat Nonlin Soft Matter Phys 76:046103