Three Co-expression Pattern Types across Microbial Transcriptional Networks of Plankton in Two Oceanic Waters

Ruby Sharma
Department of Computer Science
New Mexico State University
Las Cruces, New Mexico, USA
ruby49@nmsu.edu

Xuye Luo
Department of Computer Science
New Mexico State University
Las Cruces, New Mexico, USA
luoxuye@nmsu.edu

Sajal Kumar
Department of Computer Science
New Mexico State University
Las Cruces, New Mexico, USA
sajal49@nmsu.edu

Mingzhou Song*
Department of Computer Science
Molecular Biology Graduate Program
New Mexico State University
Las Cruces, New Mexico, USA
joemsong@cs.nmsu.edu

ABSTRACT

Patterns of two molecules across biological systems are often labeled as conserved or differential. We argue that this classification is insufficient. Here, we introduce three types of relationships across systems. Upon stimuli, a type-0 pattern arises from conserved circuitry with active conserved trajectory; a type-1 pattern is conserved circuitry with active differential trajectory; a type-2 pattern is rewired circuitry with active trajectory. We present a 1st-order marginal change test, prove its optimality, and establish its asymptotic chi-squared distribution under the null hypothesis of identical marginals across conditions. The test outperformed other methods in detecting 1st-order difference in simulation studies. We also introduce a zeroth-order strength test to assess association of two variables across systems. We compared gene co-expression networks of planktonic microbial communities in cold California coastal water against the warm water of North Pacific Subtropical Gyre. The frequency of type-1 patterns is much higher than those of type-2 and type-0 patterns, revealing that the microbial communities are mostly conserved in molecular circuitry but responded differentially to ocean habitats. Type-1 and 2 patterns are enriched with genes known to respond to environmental changes or stress; type-0 patterns involve genes having essential function such as photosynthesis and general transcription. Our work provides a deep understanding to effects of the environment on gene regulation in microbial communities. The method is generally applicable to other biological systems. All tests are provided in the R package 'DiffXTables' at https://cran.r-project.org/package=DiffXTables. Other source code and lists of significant gene patterns are available at https://www.cs.nmsu.edu/~joemsong/ACM-BCB-2020/Plankton

CCS CONCEPTS

• **Applied computing** → **Biological networks**; *Recognition of genes and regulatory elements*; *Computational transcriptomics*; *Computational genomics.*

KEYWORDS

First-order differential pattern, Co-expression network, Metatranscriptome, Planktonic microbial community, Oceanic ecosystem

ACM Reference Format:
Ruby Sharma, Xuye Luo, Sajal Kumar, and Mingzhou Song. 2020. Three Co-expression Pattern Types across Microbial Transcriptional Networks of Plankton in Two Oceanic Waters. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '20), September 21–24, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3388440.3412485

1 INTRODUCTION

Comparative studies of molecular biological systems have been a key contributory factor in advancing life science [14, 19]. Two dynamical systems can differ in trajectory, circuitry, or both. The trajectory of a system is a collection of states over time. Change in trajectory is directly observable; change in circuitry often has to be inferred. To study such differences, changes in a single molecule have been studied using differential gene-expression (DGE) analysis to link molecular basis to phenotypic variations [3, 5]. Specifically, DGE analysis detects mean difference in gene expression dynamics. Changes in gene interactions have been dominated by differential correlation methods to reveal changed circuitry restricted to linear or monotonic dynamic patterns across conditions [26, 28, 38]; the Sharma-Song test [37] can detect changed non-monotonic interaction patterns across conditions.

There is a gap that these methods by themselves do not address. What part of a gene network is involved in responding to environmental changes but is not rewired in circuitry? To answer this question, we classify changes to a bivariate relationship into

*Corresponding author

BCB '20, September 21–24, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7964-9/20/09…$15.00
https://doi.org/10.1145/3388440.3412485

three types as summarized in Table 1. A pair of random variables is zeroth-order if their trajectory does not form a random pattern in some condition. A pair is first-order differential if either variable differs in marginal distribution across conditions. A pair is second-order differential if the deviation of their joint distribution from the product of marginals is unequal across conditions. We categorize type-1 patterns as those that are differential in the first-order but show no change in the second order. Type-2 patterns must be second-order differential. Type-0 patterns are conserved in the first and second orders. Additionally, all three types must also be zeroth order (active in some condition); otherwise, a pair is type-null.

By DGE analysis, one cannot fully answer the first-order differential question because DGE is insensitive to difference in distribution when means are equal. To overcome this deficiency, we introduce a model-free statistical method named marginal change test that can determine first-order or marginal distribution change in a pair of random variables X and Y across conditions. We define its test statistic by summing up two chi-squared statistics from each variable across conditions. Under the null hypothesis that the marginals for both X and Y are conserved across condition, we prove that the test statistic asymptotically follows a chi-squared distribution. Additionally, we show that the test statistic is minimized to zero if and only if X has the same empirical marginal distribution across conditions and so does Y. In our simulation studies, the test demonstrates considerable advantage in detecting first-order differential components in patterns over other methods.

Microbial community covers a large fraction of the aquatic environment and supports the sustainability and functioning of marine ecosystems [7, 24, 43]. Environmental factors like light, temperature, oxygen concentration, and nutrient availability can have a major impact on plankton diversity [27, 29]. Most studies on oceanic habitats [6, 15, 16] are limited to comparing microscopic abundance of two ecosystems. To learn how microbial gene regulation may have adapted to oceanic conditions, we characterized the three pattern types on gene co-expression of planktonic microbial communities in California coastal (CC) and North Pacific Subtropical Gyre (NPSG) oceanic ecosystems, which differ substantially in water temperature and nutrient availability. We utilized metatranscriptomic data that were collected to measure expression of homologous genes in microbial clades over a span of five days in two previous studies [6, 30].

Among common 4,015 gene pairs actively co-expressed in at least one ecosystem, our analysis revealed about 62% gene-gene co-expression pairs as type 1, 22% as type 2, and 16% as type 0. Type-1 and 2 patterns are enriched with genes known to respond to environmental changes or stress; type-0 patterns involve genes required for essential function such as photosynthesis and general transcription. Type-1 patterns are over-represented in SAR116 and SAR86 clades, whereas SAR11 and SAR406 clades are prominent with type-0 and type-2 patterns in comparison to other clades. SAR116 and SAR86 clades are known to be affected by environmental factors like temperature and abundance of nutrients. SAR11 and SAR406 clades hint at adaptability as they can sustain in oxygen minimum zones. Taken together, our findings suggest that the two contrasting oceanic ecosystems have a high impact on their planktonic microbial inhabitants which have coped with the environment using mostly the same gene regulatory circuitry with a modest level of gene network rewiring.

2 METHODS

2.1 A First-order Marginal Change Test

To examine whether the marginal distribution of any variable in a pattern has changed across K conditions, we design a first-order marginal change test to examine K contingency tables for either row or column marginal differences. The null hypothesis is that row and column variables are independent in each contingency table and all K tables are observed independently of each other.

Let X and Y be discrete random variables of r and s levels, representing row and column variables, respectively. Let $p_k(X)$ be the marginal distribution of X and $p_k(Y)$ be the marginal distribution of Y in contingency table k. Let $p_k(X, Y)$ be the joint distribution of X and Y in contingency table k.

We consider the null hypothesis where the K tables are *first-order conserved*, which is defined by

$$p_1(X) = \cdots = p_K(X) \quad \text{and} \quad p_1(Y) = \cdots = p_K(Y) \tag{1}$$

In the alternative hypothesis, the K tables are *first-order differential* if for some pair of tables k and m, the following holds:

$$p_k(X) \neq p_m(X) \quad \text{or} \quad p_k(Y) \neq p_m(Y) \tag{2}$$

Let n_{ijk} be the observed counts of $X = i$ and $Y = j$ under condition k. We create an $K \times r$ contingency table C_X, where row k is the counts of X from levels 1 to r in condition k, written as

$$C_X[k, i] = \sum_{j=1}^{s} n_{ijk} \tag{3}$$

and we define a chi-squared statistic based on the Pearson's chi-squared test [32]:

$$\chi^2_{(K-1)(r-1)}(C_X) = \sum_{k=1}^{K} \sum_{i=1}^{r} \frac{(C_X[k, i] - \bar{C}_X[k, i])^2}{\bar{C}_X[k, i]} \tag{4}$$

where the expected count of $X = i$ in condition k is

$$\bar{C}_X[k, i] = \frac{\sum_{i'=1}^{r} C_X[k, i'] \cdot \sum_{k'=1}^{K} C_X[k', i]}{\sum_{k'=1}^{K} \sum_{i'=1}^{r} C_X[k', i']} \tag{5}$$

We also create an $K \times s$ contingency table C_Y, where row k is the counts of Y from levels 1 to s in condition k, given by

$$C_Y[k, j] = \sum_{i=1}^{r} n_{ijk} \tag{6}$$

which also gives rise to a chi-squared statistic

$$\chi^2_{(K-1)(s-1)}(C_Y) = \sum_{k=1}^{K} \sum_{j=1}^{s} \frac{(C_Y[k, j] - \bar{C}_Y[k, j])^2}{\bar{C}_Y[k, j]} \tag{7}$$

where the expected count of $Y = j$ in condition k is

$$\bar{C}_Y[k, j] = \frac{\sum_{j'=1}^{s} C_Y[k, j'] \cdot \sum_{k'=1}^{K} C_Y[k', j]}{\sum_{k'=1}^{K} \sum_{j'=1}^{r} C_Y[k', j']} \tag{8}$$

Table 1: Classification of relationships between two or more dynamical systems involving two variables. Each variable has the same physical meaning across systems. For example, each system can be a cell that contains the same pair of genes.

	Zeroth-order	First-order	Second-order	Interpretation		
Type	association	difference	difference	Trajectory	Circuitry	Explanation
0	Present	Absent	Absent	Same	Same	Conserved mechanism responding to same input
1	Present	Present	Absent	Changed	Same	Conserved mechanism responding to changed input
2	Present	Any	Present	Changed	Rewired	Different mechanisms responding to input
Null	Absent	—	—	—	—	No active mechanism involved

Now, we define the statistic of the first-order marginal change test by

$$\chi^2_\nu = \chi^2_{(K-1)(r-1)}(C_X) + \chi^2_{(K-1)(s-1)}(C_Y) \tag{9}$$

where the degrees of freedom (d.f.) ν is

$$\nu = (K-1)(r+s-2) \tag{10}$$

Theorem 2.1 (Null distribution). *The first-order marginal change test statistic χ^2_ν is asymptotically chi-squared distributed under the null hypothesis that X and Y are statistically independent, X is identically and independently distributed (i.i.d.) across conditions, and so does Y.*

Proof. As the counts observed for X in each condition are i.i.d. under the null hypothesis, the row and column variables of C_X are independent. It thus follows that $\chi^2(C_X)$ is asymptotically chi-squared distributed with $(K-1)(r-1)$ d.f. under the null hypothesis. We can similarly derive that $\chi^2(C_Y)$ is asymptotically chi-squared distributed with $(K-1)(s-1)$ d.f. under the null hypothesis.

As the sum of chi-squared distributed variables are also chi-squared with the summed d.f. of each variable, $\chi^2_\nu = \chi^2(C_X) + \chi^2(C_Y)$ must also be chi-squared distributed with

$$\nu = (K-1)(r-1) + (K-1)(s-1) = (K-1)(r+s-2) \tag{11}$$

d.f. under the null hypothesis. □

Using the upper tail probability of the chi-squared null distribution, we can compute a P-value, the statistical significance of observing some 1st-order differences in K conditions if the marginals are all identical.

Next, we show that the test statistic χ^2_ν is optimal in the sense that it is minimized to zero if and only if the observed marginal probabilities of both row and column variables are identical across the K tables, which corresponds to the state of the null hypothesis.

Theorem 2.2 (Optimality). *The first-order marginal change test statistic χ^2_ν is minimized to zero if and only if X has the same observed marginal probability in each condition, and so does Y.*

Proof. First we prove the sufficient condition that equal observed marginal probabilities of X and Y across conditions lead to $\chi^2_\nu = 0$. Let $p_1, \ldots, p_r$ be the equal observed marginal probabilities of X. Let the sample size for each table be $n_1, \ldots, n_K$. Then we have $C_X[k,i] = n_k p_i$, which gives rise to

$$\bar{C}_X[k,i] = \frac{\sum_{i'=1}^r C_X[k,i'] \cdot \sum_{k'=1}^K C_X[k',i]}{\sum_{k'=1}^K \sum_{i'=1}^r C_X[k',i']} \tag{12}$$

$$= \frac{\sum_{i'=1}^r n_k p_{i'} \cdot \sum_{k'=1}^K n_{k'} p_i}{\sum_{k'=1}^K \sum_{i'=1}^r n_{k'} p_{i'}} \tag{13}$$

$$= \frac{n_k \cdot (n_1 + \cdots + n_K) p_i}{n_1 + \cdots + n_K} \tag{14}$$

$$= n_k p_i = C_X[k,i] \tag{15}$$

Plugging $\bar{C}_X[k,i]$ into Eq. (4), we immediately have $\chi^2_{(K-1)(r-1)}(C_X) = 0$. Similarly, given equal observed marginal probabilities of Y, it is true that $\chi^2_{(K-1)(s-1)}(C_Y) = 0$. This implies that

$$\chi^2_\nu = \chi^2_{(K-1)(r-1)}(C_X) + \chi^2_{(K-1)(s-1)}(C_Y) = 0 \tag{16}$$

Thus, we have proven the sufficient condition for $\chi^2_\nu = 0$.

Second, we prove the necessary condition that $\chi^2_\nu = 0$ implies equal observed marginal probabilities of X and Y across K conditions. As χ^2_ν is defined by sum of squares, each squared terms must be zero if $\chi^2_\nu = 0$. This suggests that $\bar{C}_X[k,i] = C_X[k,i]$ and $\bar{C}_Y[k,j] = C_Y[k,j]$. By Eq. (5), we have

$$C_X[k,i] = \frac{\sum_{i'=1}^r C_X[k,i'] \cdot \sum_{k'=1}^K C_X[k',i]}{\sum_{k'=1}^K \sum_{i'=1}^r C_X[k',i']} \tag{17}$$

$$= \frac{n_k \cdot \sum_{k'=1}^K C_X[k',i]}{n_1 + \cdots + n_K} \tag{18}$$

Then the marginal probability of $X = i$ in table k is

$$\frac{C_X[k,i]}{n_k} = \frac{\sum_{k'=1}^K C_X[k',i]}{n_1 + \cdots + n_K} \tag{19}$$

As the right-hand side is constant with respect to k, all the observed marginal probabilities of X are identical across the K tables. Similarly, we can show that Y has the same observed marginal probabilities across the K tables. Thus, we have also proven the necessary condition for $\chi^2_\nu = 0$. □

2.2 Detecting Type 0, 1, and 2 Patterns

These three pattern types are convenient because their interpretations (Table 1) are intuitive about physical systems. We now translate them into testable hypotheses: type-0 patterns are conserved in the joint distributions $p_1(X,Y) = \cdots = p_K(X,Y)$; type-1 patterns differ in 1st-order (marginal) distribution but have no 2nd-order difference; type-2 patterns must have 2nd-order difference.

Table 2 summarizes how type 0, 1, and 2 patterns are detected across K contingency tables. We use the strength chi-squared test [37] to ensure at least one table represents a strong association. Then we use the 1st-order marginal change test to detect 1st-order difference and the Sharma-Song test [37] to detect the 2nd-order difference. The threshold is applied on adjusted P-values when applicable.

2.3 Normalizing Metatranscriptomic Data

To study the three pattern types in biological networks, we examined the planktonic microbial metatranscriptomic data previously collected in two oceanic water ecosystems: the cold water of California coastal (CC) region and the warm water of North Pacific Subtropical Gyre (NPSG). The NPSG dataset of 30 samples was first published [30]. The CC dataset of 35 samples was later collected and studied together with the NPSG data in [6], where we obtained the normalized counts of both data sets. The data were sampled over a time span of five days at every four hours. The CC data gave rise to 8844 gene clusters and NPSG data had 7276 gene clusters. For our analysis, we selected the 2958 common gene clusters based on cluster names, KEGG ID, and KO ID. We examined the combined data using principal component analysis (PCA) in Section 4. We detected outliers using the Mahalanobis distance from each sample to the center in the subspace of the first five PC dimensions [12], accounting for 95% of variance in the data. We found seven samples—six from CC and one from NPSG—that were significantly ($P \leq 0.05$) distant. After removing these seven samples, we have 29 CC and 29 NPSG samples remained. CC and NPSG were separately log scaled after normalizing them to their respective aggregated sample median and adding one, given by

$$\hat{E}_{ij} = \ln\left(\frac{E_{ij} \cdot \tilde{E}}{E_{\cdot j} \cdot F_j} + 1\right)$$

where E_{ij} is the expression of gene cluster i in sample j, $E_{\cdot j}$ is the sample aggregate for sample j, $\tilde{E}$ is the median aggregated across all samples and F_j is the normalizing factor returned by `calcNormFactors` function in 'edgeR' using the weighted trimmed mean of M-values [35].

3 BENCHMARKING THE FIRST-ORDER MARGINAL CHANGE TEST

We performed a simulation study to evaluate the performance of 1st order marginal change test in comparison with two other methods. A Z-score based method on scaled differential Spearman's correlation from DGCA [28] and the chi-squared heterogeneity test [42]. We conducted the evaluation using receiver operating characteristic (ROC) curves in detecting 1st-order components from observed patterns across two conditions. We define a full-order pattern to contain both 1st and 2nd-order differences across conditions. The four setups include detecting 1st- from 2nd-order patterns, 1st-order from conserved, full-order from 2nd-order, and full-order differential patterns from conserved patterns.

We simulated 1st-order, 2nd-order, full-order differential patterns and also conserved patterns. The algorithm for simulating the patterns are available publicly [36, 37]. We simulated 500 pairs of tables ($K = 2$) with dimension size of 3 to 5 for each table type and sample size ranging from 100 to 300. The noise levels of 0, 0.1, 0.2, 0.3, 0.4 and 0.5 are incorporated into the simulated tables.

The performance of all three tested methods is reported in Figure 1. Figure 1a gives the accuracy of our method versus other

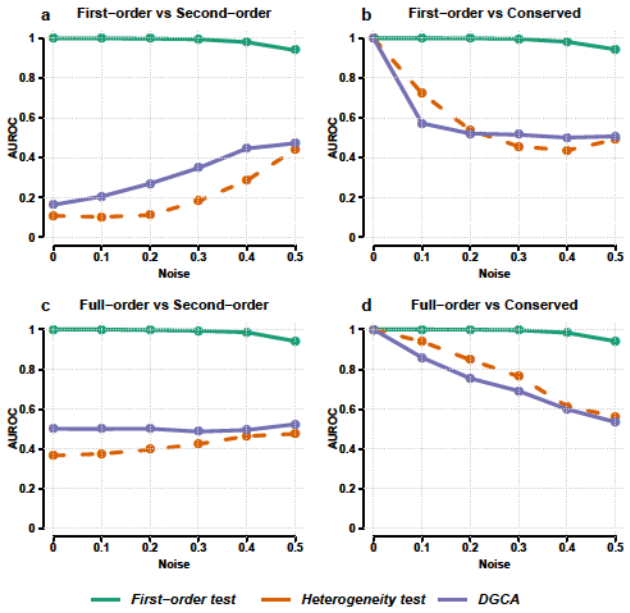

Figure 1: Benchmarking the 1st-order marginal change test and two other methods in detecting 1st-order components in patterns. We measure the performance by areas under the ROC curve over increasing noise levels. (a) 1st-order versus 2nd-order patterns. (b) 1st-order versus conserved patterns. (c) Full-order versus 2nd-order patterns. (d) Full-order versus conserved patterns.

methods in detecting 1st- from 2nd-order differential patterns. Figure 1b shows the difference in accuracy between our method and other methods in telling 1st-order patterns from conserved patterns. Figure 1c depicts the higher accuracy of our method than other methods in distinguishing full-order differential patterns from 2nd-order differential patterns. Figure 1d reveals that our method has an advantage over other methods in recognizing full-order differential patterns over conserved patterns.

4 MICROBIAL GENE CO-EXPRESSION PATTERNS ACROSS TWO ECOSYSTEMS

To understand how planktonic microbial gene networks may have evolved across oceanic waters, we study type 0, 1, and 2 gene co-expression patterns of planktonic microbial communities between CC and NPSG ecosystems. The input metatranscriptomic data of CC and NPSG include normalized data of 58 samples after removing seven outliers from the original 65 samples. Figure 2 shows the first three principal components before and after outlier removal.

We followed the protocol in Table 2 to detect three types of gene co-expression patterns across CC and NPSG ecosystems. For each

Table 2: A protocol for testing for type 0, 1, and 2 patterns across two ore more systems of two variables using three statistical tests.

Type	Zeroth-order association Strength test	First-order differential Marginal change test	Second-order differential Sharma-Song test
0	P-value ≤ 0.10	P-value > 0.05	P-value > 0.05
1	P-value ≤ 0.10	P-value ≤ 0.05	P-value > 0.05
2	P-value ≤ 0.10	Any	P-value ≤ 0.05
Null	P-value > 0.10	P-value > 0.05	P-value > 0.05

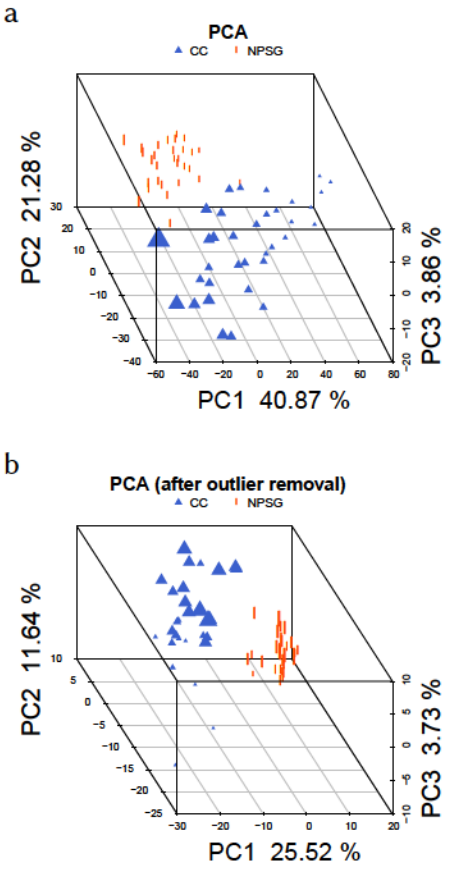

Figure 2: First three principal components of California coastal and North Pacific Subtropical Gyre samples. (a) PCA on the original 65 samples. (b) PCA on the remaining 58 samples after outlier removal.

gene pair, we removed samples where both have zero expression values. Then we discretized the continuous gene expression value using an optimal univariate clustering algorithm [39, 41].

We constructed two co-expression networks for CC and NPSG, respectively, by evaluating $\binom{2958}{2}$ gene pairs using a strength chi-squared test [32, 37] where the sum of chi-squared statistics with the summed d.f. of both CC and NPSG is used to select unique gene pairs with strong association in at least one water. We obtained 4,015 unique significant patterns at Benjamini-Hochberg [9] adjusted P-value ≤ 0.1.

We applied the 1st-order marginal change test and Sharma-Song test [37] on discretized values of the 4,015 co-expressed gene pairs across the CC and NPSG samples. Both tests return P-values as level of statistical significance, further adjusted by the Benjamini-Hochberg [9] method to account for multiple testing effects. At the significance level of 0.05, we detected 2468 type-1, 891 type-2, and 656 type-0 patterns. Table 3 summarizes the patterns detected for genes in five microbial clades. Type-1 patterns are most abundant (61.47%), aligning with the conserved gene networks of microbial communities across CC and NPSG as previously reported [6]. However, there are rewired gene pairs (22.19%) that were not reported previously; they may be fundamental for microbial communities to adapt to changed oceanic habitats.

Table 3: Numbers and percentages of type 0, 1, and 2 gene co-expression patterns in five heterotrophic bacterial clades of planktonic microbial communities between CC and NPSG oceanic ecosystems.

Clade	Total	Type 0	Type 1	Type 2
SAR116	2230	6.68 %	69.06%	24.26%
SAR86	1037	13.98%	66.25%	19.77%
SAR11	3409	21.30%	57.38%	21.33%
Roseobacter	920	16.74%	61.41%	21.85%
SAR406	434	31.80%	43.32%	24.88%
Total unique	4015	656	2468	891
Percentage		16.34%	61.47%	22.19%

We extracted five hub genes with highest degrees in the network of the 4,015 patterns. We sampled 40 patterns for each hub gene to create a subnetwork (Figure 3) of 200 interactions, where each hub gene has different proportions of type 0, 1, and 2 patterns.

From Table 3, the SAR116 clade is most abundant in type-1 patterns followed by type-2 patterns, suggesting genes have responded to temperature and nutrient differences in the two waters. The dddP gene of SAR116 is involved in the dimethylsulfoniopropionate (DMSP) cycle in ocean, regulated by temperature, UV radiation, carbon, and sulfur demands [11]. These observations highlight genes that cope with changed environmental inputs using mostly the same mechanism and some rewired mechanisms.

The SAR86 clade has a relatively high percentage of type-1 patterns with a low percentage for type-0 and type-2 patterns among

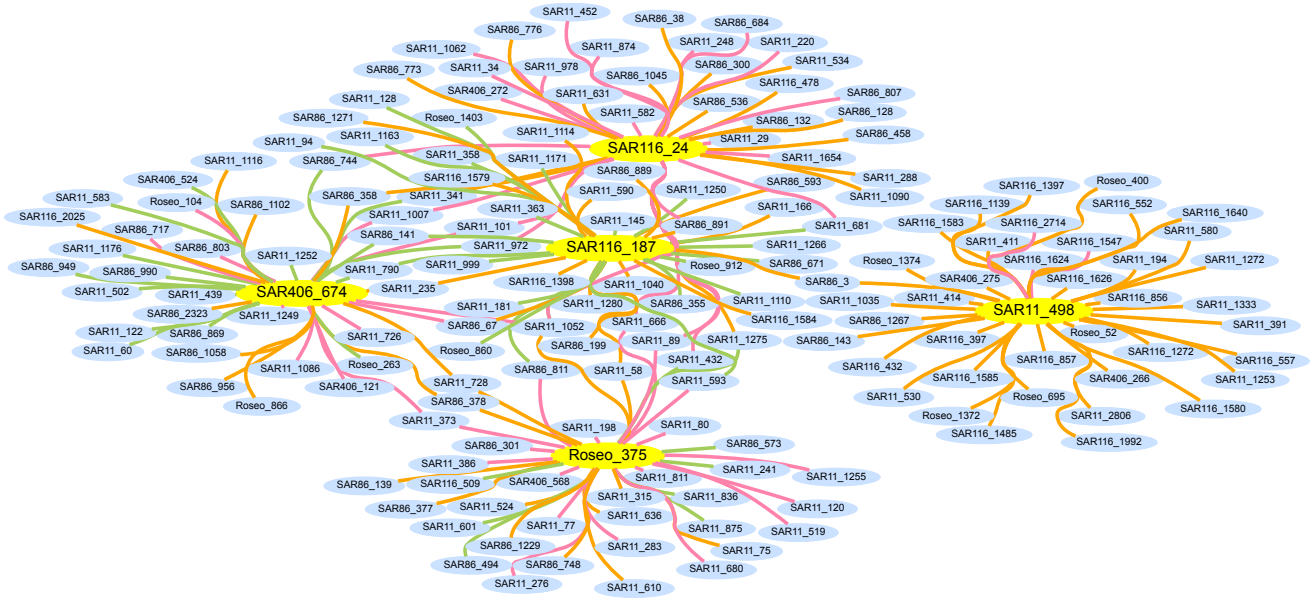

Figure 3: A subnetwork containing all three types of pattern selected for five hub genes and 200 patterns in California coast and North Pacific Subtropical Gyre. Hub genes are highlighted within yellow nodes. Type 0, 1, 2 patterns are represented by green, orange, and red edges.

all the clades. SAR86 is dominant in the coastal water [31] subject to temperature change [29].

The SAR11 clade has the highest number (3409) of active gene pairs among the clades, suggesting the overall importance of this clade to oceanic ecosystems for a prolonged desire of sustainability and endurance towards drastic environmental changes.

The SAR406 clade has the lowest number of active gene pairs among the clades. It has highest percentages of type-0 and type-2 patterns and the lowest of type-1 patterns. Like SAR11, genes involved in SAR406 have mechanisms conserved in the two waters yet exhibit rewired mechanisms probably for sustainability in oxygen minimum zones.

Both SAR11 and SAR406 are abundant in oxygen minimum zones [10]. Some operational taxonomic units (OTUs) of SAR11 and SAR406 increase with deoxygenation and others decrease [2, 8]. SAR406 can grow in diverse environmental conditions and is genetically linked to the *Fibrobacter* spp. and green sulfur bacteria in the Atlantic and Pacific oceans [18]. The SAR11 clade is the most abundantly found in marine ecosystems. SAR11 can survive in oligotrophic regions like NPSG with limiting conditions [21]. SAR11 is involved in the pathway of ocean nitrogen loss by contributing towards NO_3^- respiration in anoxic zones [40]. To adapt for life in a nutrient deficient ecosystem, a change in the molecular circuitry of SAR11 and SAR406 from cold nutrient-rich water to warm nutrient-deficient conditions is justifiable.

The Roseobacter clade has minimal presence in active gene-gene patterns after SAR406. They are observed mostly as type-1 patterns followed by type-2 patterns. Our analysis reveals Roseobacter responded to the drastic oceanic change with mostly conserved circuitry and modest network rewiring.

4.1 Type-1 Gene Co-expression Patterns and Network: Conserved Mechanisms Responding to Changed Ocean Habitats

Four significant type-1 patterns are shown in Figure 4a–d. Gene SAR86_118 (K03782) is involved in biosynthesis of secondary metabolites, phenylpropanoid biosynthesis, and tryptophan metabolism. It is positively associated with SAR11_498 (K03798), functioning as peptidases, inhibitors, chaperones, or folding catalysts. SAR11_1474 (K02030) is involved in signaling and cellular processes as transporters. It is positively co-expressed with SAR11_301 with lower expression in NPSG water than CC water. SAR86_429 (K00605) is involved in biosynthesis of secondary metabolites, glycine, serine and threonine metabolism, and carbon metabolism. It is negatively co-expressed with SAR116_535 (K03687) implicated in genetic information processing, and chaperones and folding catalysts, with high expression in NPSG water. SAR116_501 (K00963), involved in starch and sucrose metabolism, galactose metabolism, and pentose and glucuronate interconversions, is positively co-expressed with SAR11_391 with a same non-monotonic trend and high expression in NPSG water. These type-1 interactions have same mechanism between the two waters, but there exists a marginal difference suggesting an adaptive response to oceanic input without changing molecular circuitry.

Figure 4e is a network representing the top 200 type-1 interactions. Hub gene SAR86_118 represents the accumulation of katG genes which encode catalase-peroxidases, a unique bifunctional

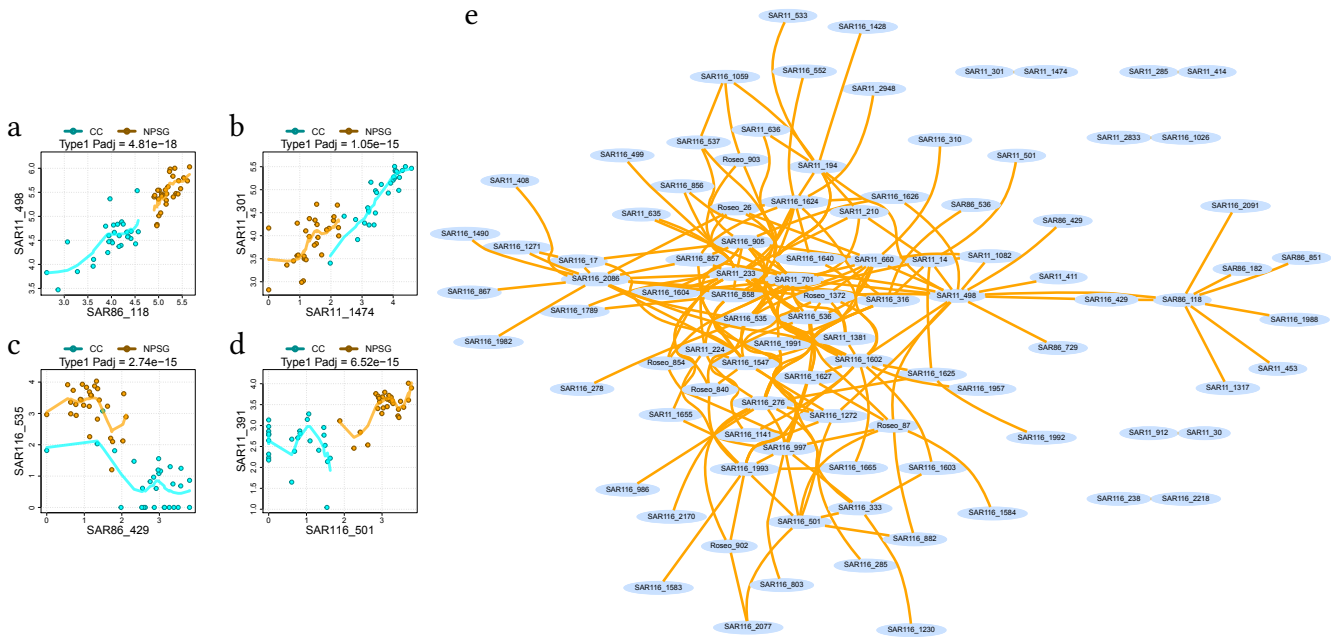

Figure 4: Type-1 patterns and networks differential in the 1st-order but not 2nd-order between California coast and North Pacific Subtropical Gyre. (a–d) Four type-1 gene co-expression patterns. The horizontal and vertical axes are the gene-expressions. A loess curve is fitted to show dynamics. (e) A type-1 network formed by top 200 type-1 patterns (edges) between CC and NPSG. Each node represents a cluster of homologous genes from microbial species in a common clade.

enzyme that expresses in soil bacteria, playing an important role in coping with the environmental stress or oxidative stress [17]. Another hub gene SAR11_233 is molecular chaperone IbpA also known as heat shock protein. The function of IbpA gene is to protect or suppress the inactivation of enzymes from the stress of heat and oxidants [22, 23]. SAR116_535, a molecular chaperone GrpE, is a heat shock protein responsible for bacterial growth and viability against all temperatures [4]. Some highly active genes are responsible for the protection of ortholog organisms against the external environmental factors. CC and NPSG are two oceanic ecosystems differing in temperature, oxygen level, nutrient availability [21]. It is conceivable that a type-1 pattern reacts to changed environmental conditions across the two ecosystems without having to rewire the molecular circuitry.

4.2 Type-2 Gene Co-expression Patterns and Network: Rewired Mechanisms

Four significant type-2 patterns are shown in Figure 5a–d, illustrating dramatic co-expression differences due to rewiring. SAR11_56 (K00128), involved in microbial metabolism in diverse environments, valine, leucine and isoleucine degradation, fatty acid degradation, and lysine degradation. It co-expresses with SAR86_378 (K00382) involved in microbial metabolism in diverse environments / Valine, leucine and isoleucine degradation, and lysine degradation, with an increasing trend in CC and a non-monotonic trend in NPSG. Roseo_1 (K03116) is involved in environmental information processing, bacterial secretion system, folding, sorting and degradation, genetic information processing, and membrane transport. It co-expresses with SAR86_437 (K03561) involved in protein families: signaling and cellular processes, transporters, biopolymer transport protein ExbB, with a non-monotonic pattern in CC while showing an independent trend in NPSG. SAR86_182 (K00789) participates in biosynthesis of secondary metabolites, biosynthesis of amino acids, cysteine and methionine metabolism. It co-expresses with SAR11_361 (K04754) as structural proteins and phospholipid-binding lipoprotein MlaA, with a negative relationship in NPSG whereas an increasing trend in CC. SAR11_155 (K00161), involved in microbial metabolism in diverse environments and HIF-1 signaling pathway, co-expresses with SAR116_282 (K00795) involved in metabolism of terpenoids and polyketides, and terpenoid backbone biosynthesis, observing an opposite non-monotonic trend in both CC and NPSG. These patterns show the rewiring of interactions between two genes which can result from ecosystem adaptability spanned over a long time.

Figure 5e is a network comprising of top 200 type-2 patterns. There are several hubs of high degrees. One hub gene SAR11_56 interacting with many genes and another gene SAR11_31, is involved in the pathway of microbial metabolism in diverse environments. The pathway is the accumulation of several metabolic pathways like nitrogen, sulfur, methane, amino acid and energy pathways [20]. SAR11_377 gene of DNA gyrase subunit B is known for DNA replication, repair and recombination [1]. It has been used as a genetic marker for exploring bacterial diversity [13, 33]. GyrB, a housekeeping gene responsible for diversity between the strains of fresh and deep water, suggests the adaptation of bacterial ecotypes in deep

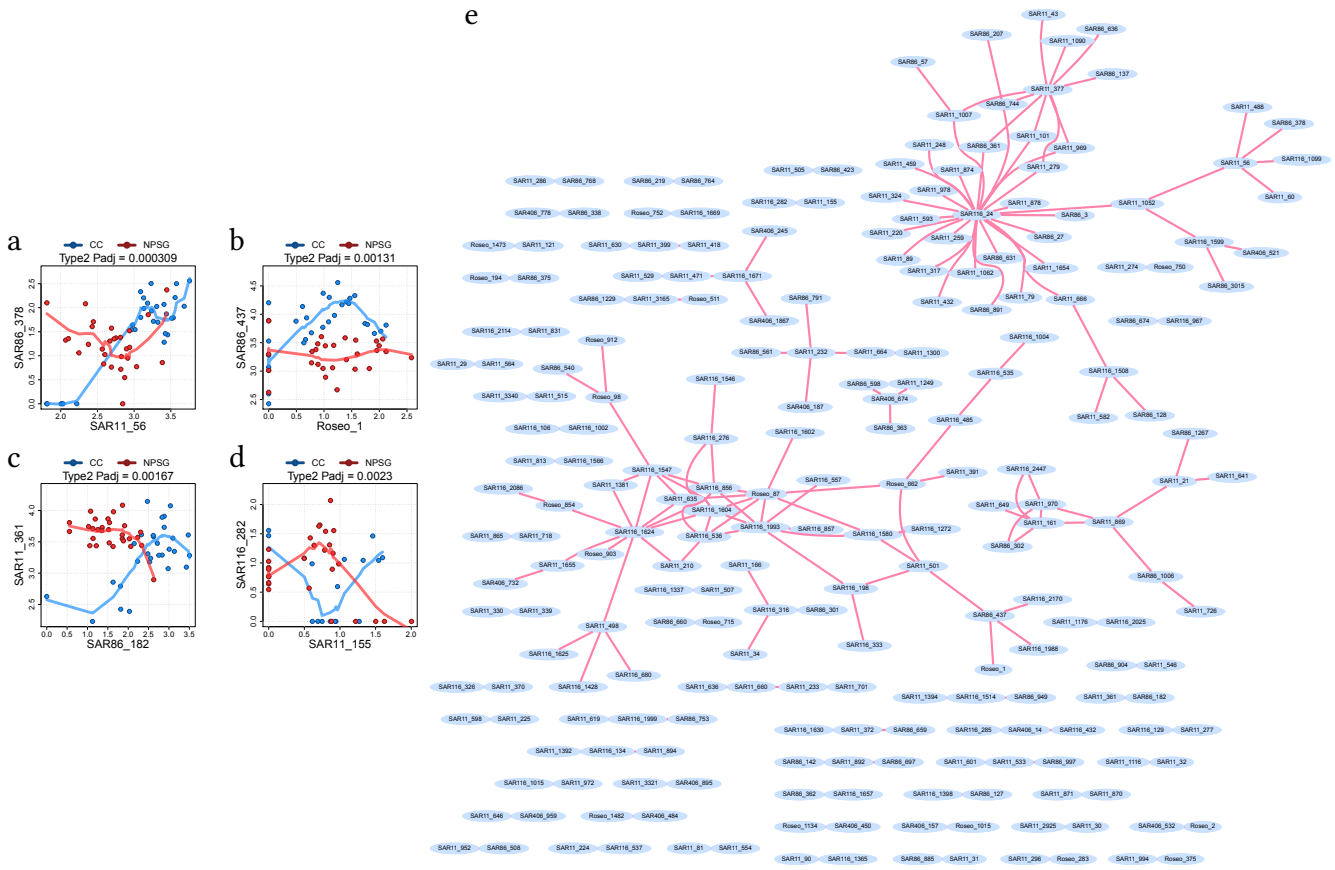

Figure 5: Type-2 patterns and networks differential in 2nd-order between California coast and North Pacific Subtropical Gyre. (a–d) Four type-2 gene co-expression patterns. The horizontal and vertical axes are the gene-expressions. Fitted loess curves show dynamics. (e) A type-2 network formed by top 200 type-2 patterns differential between CC and NPSG. The edges are statistically significant type 2 patterns involving the two nodes of each edge.

sea [25]. The SAR11 clade is the most abundant in surface water and also exists in oxygen minimum zones [40]. SAR11 lineages, having adapted to the harsh environment of oxygen minimum zones, are involved in the pathway of ocean nitrogen loss by contributing towards NO_3^- respiration in anoxic zones [40]. SAR11_1253 gene of aprA (adenylylsulfate reductase, subunit A) is involved in dissimilatory sulphate reduction [20]. Bacteria carrying aprA and aprB genes involved in dissimilatory sulphate reduction are enriched in oxygen minimum zones in the Bay of Bengal [34]. A number of genes detected as a part of significant type-2 patterns are involved in some vital metabolic pathways like nitrogen and sulfur metabolism. These metabolic pathways are key factor for sustainability of microbiome in oxygen minimum zones like NPSG. The function of these type-2 gene patterns suggests that it is possible genetic rewiring may have occurred to enhance adaptability of the microbiome in the two contrasting oceanic ecosystems.

4.3 Type-0 Gene Co-expression Patterns and Network: Conserved Mechanisms and Dynamics Unrelated to Environment

Four type-0 patterns are illustrated in Figure 6a–d, where gene co-expression patterns are dynamic but almost identical in trajectory between CC and NPSG waters.

SAR11_644 (K02879) is involved in translation and genetic information processing. It co-expresses with SAR11_639 (K06207), involved in signaling and cellular process, with an increasing trend in both ecosystems.

SAR406_1867 (K02636), involved in photosynthesis and metabolic pathways, co-expresses with SAR406_294 (K03043), involved in RNA polymerase, genetic information processing, and transcription, has a positive trend in both water systems. SAR116_1914 (K02035), involved in quorum sensing and cellular processes, has positive relation with SAR11_920. SAR11_545 (K02109), involved in photosynthesis, metabolic pathways, and oxidative phosphorylation, negatively co-expresses with SAR11_186 in both water systems. Figure 6e is a network of top 200 type-0 patterns, as ranked by increasing P-values of the zeroth-order strength test. From the

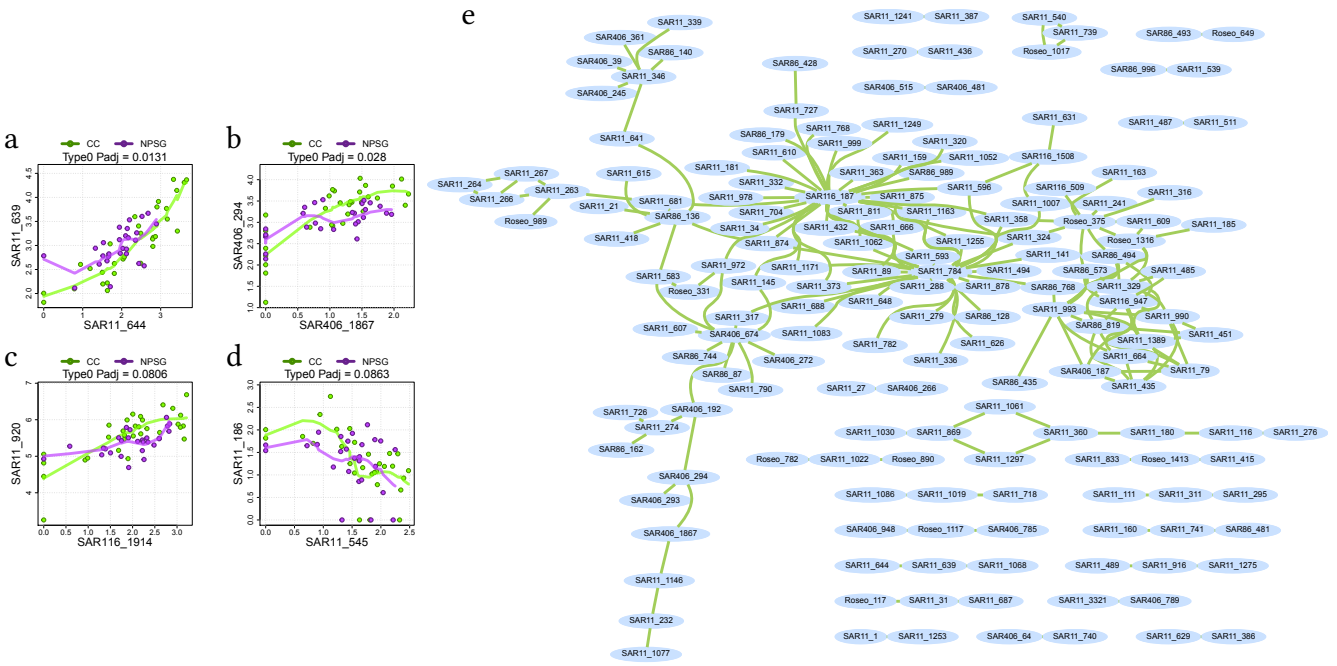

Figure 6: Type-0 gene co-expression patterns and networks conserved between California coast and North Pacific Subtropical Gyre. (a–d) Four pairs of type-0 gene patterns. The horizontal and vertical axes are the gene-expressions. Fitted loess curves show the dynamics in each ecosystem. (e) A type-0 network of top 200 patterns in conserved between CC and NPSG. The edges are type-0 patterns involving the two nodes of each edge. In a type-0 pattern, the pair of genes involved must be significantly co-expressed in at least one ecosystem, but the patterns do not differ in 1st- or 2nd-order significantly.

described function of the involved genes, these active but environmentally unchanged co-expression patterns reflect the most fundamental life processes such as photosynthesis and general gene regulation, conserved in both dynamics and molecular mechanisms in most biological systems.

5 DISCUSSION

DeLong and colleagues reported conserved transcriptional networks underlying planktonic microbial communities between California coastal and North Pacific Subtropical Gyre water ecosystems [6]. We further evaluated the gene co-expression patterns for evidence to support 1st- and 2nd-order differential networks across the two ecosystems. Our study resulted in about 62% type-1 patterns, 22% type-2 patterns and 16% type-0 patterns. Our results comply with the previous study in the aspect of conservancy in the microbial behavior as more than 70% has same interaction patterns.

Type 1 gene co-expression patterns exhibited some diversity of trajectory. Most co-expression patterns differ in mean expression of the involved genes and are thus detectable by typical differential gene expression analysis methods. However, the presented 1st-order marginal change test can also detect change in marginal distribution of involved genes which may have similar mean values. Therefore, our method is sensitive to difference in both distribution and mean, more general in recognizing changed marginal activities.

6 CONCLUSIONS

We have demonstrated the first-order marginal change test and how it is integrated with a second-order test to reveal three types of conserved and differential patterns across conditions. Applying these methods on metatranscriptomic data of planktonic microbial communities from different oceanic ecosystems, we revealed conserved and rewired microbial gene networks. The uncovered type-1 differential network suggests that planktonic microbial communities can respond to different oceanic habitats using the same underlying molecular mechanisms. Previous work focused mostly on type-0 conserved and type-2 differential networks. We emphasize the importance of type-1 differential network which can identify active network components that respond to contrasting environmental input using the same mechanism across conditions to generate different output. Therefore, the marginal change test gives new insight into molecular mechanisms beyond type 0 and 2. The three network types together comprehensively delineate how microbial communities gene networks may have evolved to respond to changed environments, also applicable to the analysis of other biological networks.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1661331, and also by Specialty

Crop Research Initiative Grant No. 2016-51181-25408/Project Accession No. 1009971 from the USDA National Institute of Food and Agriculture.

REFERENCES

[1] 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* 45, D1 (2017), D158–D169.

[2] Elke Allers, Jody J Wright, Kishori M Konwar, Charles G Howes, Erica Beneze, Steven J Hallam, and Matthew B Sullivan. 2013. Diversity and population structure of Marine Group A bacteria in the Northeast subarctic Pacific Ocean. *The ISME Journal* 7, 2 (2013), 256–268.

[3] Simon Anders and Wolfgang Huber. 2010. Differential expression analysis for sequence count data. *Nature Precedings* (2010), 1–1.

[4] DEBBIE Ang and COSTA Georgopoulos. 1989. The heat-shock-regulated grpE gene of Escherichia coli is required for bacterial growth at all temperatures but is dispensable in certain mutant backgrounds. *Journal of Bacteriology* 171, 5 (1989), 2748–2755.

[5] Alemu Takele Assefa, Katrijn De Paepe, Celine Everaert, Pieter Mestdagh, Olivier Thas, and Jo Vandesompele. 2018. Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA-sequencing data. *Genome Biology* 19, 1 (2018), 96.

[6] Frank O Aylward, John M Eppley, Jason M Smith, Francisco P Chavez, Christopher A Scholin, and Edward F DeLong. 2015. Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proceedings of the National Academy of Sciences* 112, 17 (2015), 5443–5448.

[7] Farooq Azam and Francesca Malfatti. 2007. Microbial structuring of marine ecosystems. *Nature Reviews Microbiology* 5, 10 (2007), 782–791.

[8] J Michael Beman and Molly T Carolan. 2013. Deoxygenation alters bacterial diversity and community composition in the ocean's largest oxygen minimum zone. *Nature Communications* 4, 1 (2013), 1–11.

[9] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300.

[10] Anthony D Bertagnolli and Frank J Stewart. 2018. Microbial niches in marine oxygen minimum zones. *Nature Reviews Microbiology* 16, 12 (2018), 723–729.

[11] Dong Han Choi, Ki-Tae Park, Sung Min An, Kitack Lee, Jang-Cheon Cho, Jung-Hyun Lee, Dongseon Kim, Dongchull Jeon, and Jae Hoon Noh. 2015. Pyrosequencing revealed SAR116 clade as dominant dddP-containing bacteria in oligotrophic NW Pacific Ocean. *PloS One* 10, 1 (2015).

[12] Kevin R. Coombes. 2019. *ClassDiscovery: Classes and Methods for "Class Discovery" with Microarrays or Proteomics.* https://CRAN.R-project.org/package=ClassDiscovery R package version 3.3.12.

[13] Catherine Dauga. 2002. Evolution of the gyrB gene and the molecular phylogeny of Enterobacteriaceae: a model molecule for molecular systematic studies. *International Journal of Systematic and Evolutionary Microbiology* 52, 2 (2002), 531–547.

[14] Alberto de la Fuente. 2010. From 'differential expression'to 'differential networking'–identification of dysfunctional regulatory networks in diseases. *Trends in Genetics* 26, 7 (2010), 326–333.

[15] Edward F DeLong and David M Karl. 2005. Genomic perspectives in microbial oceanography. *Nature* 437, 7057 (2005), 336–342.

[16] SJ Giovannoni. 2000. Evolution, diversity and molecular ecology of marine prokaryotes. *Microbial Ecology of the Oceans* (2000), 47–84.

[17] Jana Godočíková, Marcel Zámockỳ, Mária Bučková, Christian Obinger, and Bystrík Polek. 2010. Molecular diversity of katG genes in the soil bacteria Comamonas. *Archives of Microbiology* 192, 3 (2010), 175–184.

[18] DA Gordon and SJ Giovannoni. 1996. Detection of stratified microbial populations related to Chlorobium and Fibrobacter species in the Atlantic and Pacific oceans. *Appl. Environ. Microbiol.* 62, 4 (1996), 1171–1177.

[19] Hao He, Shaolong Cao, Ji-Gang Zhang, Hui Shen, Yu-Ping Wang, and Hong-Wen Deng. 2019. A statistical test for differential network analysis based on inference of Gaussian graphical model. *Sci Rep* 9 (Jul 2019), 10863. https://doi.org/10.1038/s41598-019-47362-7

[20] Minoru Kanehisa and Susumu Goto. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28, 1 (2000), 27–30.

[21] David M Karl and Matthew J Church. 2017. Ecosystem structure and dynamics in the North Pacific Subtropical Gyre: new views of an old ocean. *Ecosystems* 20, 3 (2017), 433–457.

[22] Masanobu Kitagawa, Mizuho Miyakawa, Yoshinobu Matsumura, and Tetsuaki Tsuchido. 2002. Escherichia coli small heat shock proteins, IbpA and IbpB, protect enzymes from inactivation by heat and oxidants. *European Journal of Biochemistry* 269, 12 (2002), 2907–2917.

[23] Dorota Kuczynska-Wisnik, Sabina Kçdzierska, Ewelina Matuszewska, Peter Lund, Alina Taylor, Barbara Lipinska, and Ewa Laskowska. 2002. The Escherichia coli small heat-shock proteins IbpA and IbpB prevent the aggregation of endogenous proteins denatured in vivo during extreme heat shock. *Microbiology* 148, 6 (2002), 1757–1765.

[24] Elizabeth B Kujawinski, Krista Longnecker, Katie L Barott, Ralf JM Weber, and Melissa C Kido Soule. 2016. Microbial community structure affects marine dissolved organic matter composition. *Frontiers in Marine Science* 3 (2016), 45.

[25] Arantxa López-López, Sergio G Bartual, Lucas Stal, Olga Onyshchenko, and Francisco Rodríguez-Valera. 2005. Genetic analysis of housekeeping genes reveals a deep-sea ecotype of Alteromonas macleodii in the Mediterranean Sea. *Environmental Microbiology* 7, 5 (2005), 649–659.

[26] Haisu Ma, Eric E Schadt, Lee M Kaplan, and Hongyu Zhao. 2011. COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. *Bioinformatics* 27, 9 (2011), 1290–1298.

[27] Isabelle Mary, DG Cummings, Isabelle C Biegala, PH Burkill, SD Archer, and MV Zubkov. 2006. Seasonal dynamics of bacterioplankton community structure at a coastal station in the western English Channel. *Aquatic Microbial Ecology* 42, 2 (2006), 119–126.

[28] Andrew T McKenzie, Igor Katsyv, Won-Min Song, Minghui Wang, and Bin Zhang. 2016. DGCA: a comprehensive R package for differential gene correlation analysis. *BMC Syst Biol* 10, 1 (2016), 106.

[29] Robert M Morris, Kevin L Vergin, Jang-Cheon Cho, Michael S Rappé, Craig A Carlson, and Stephen J Giovannoni. 2005. Temporal and spatial response of bacterioplankton lineages to annual convective overturn at the Bermuda Atlantic Time-series Study site. *Limnology and Oceanography* 50, 5 (2005), 1687–1696.

[30] Elizabeth A Ottesen, Curtis R Young, Scott M Gifford, John M Eppley, Roman Marin, Stephan C Schuster, Christopher A Scholin, and Edward F DeLong. 2014. Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science* 345, 6193 (2014), 207–212.

[31] Silvia Pajares, Francisco Varona-Cordero, and David Uriel Hernández-Becerril. 2020. Spatial distribution patterns of Bacterioplankton in the oxygen minimum zone of the Tropical Mexican Pacific. *Microb Ecol* (2020).

[32] Karl Pearson. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *London Edinburgh Dublin Philos Magazine J Sci* 50, 302 (1900), 157–175.

[33] Karolien Peeters and Anne Willems. 2011. The gyrB gene is a useful phylogenetic marker for exploring the diversity of Flavobacterium strains isolated from terrestrial and aquatic habitats in Antarctica. *FEMS Microbiology Letters* 321, 2 (2011), 130–140.

[34] Shriram N Rajpathak, Roumik Banerjee, Pawan G Mishra, Asmita M Khedkar, Yugandhara M Patil, Suraj R Joshi, and Deepti D Deobagkar. 2018. An exploration of microbial and associated functional diversity in the OMZ and non-OMZ areas in the Bay of Bengal. *Journal of Biosciences* 43, 4 (2018), 635–648.

[35] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 1 (2010), 139–140.

[36] Ruby Sharma, Sajal Kumar, Hua Zhong, and Mingzhou Song. 2017. Simulating noisy, nonparametric, and multivariate discrete patterns. *The R Journal* 9, 2 (2017), 366–377.

[37] Ruby Sharma and Mingzhou Song. 2020. *DiffXTables: Pattern Heterogeneity via Distributional Differences Across Contingency Tables.* https://CRAN.R-project.org/package=DiffXTables R package version 0.1.0.

[38] Charlotte Siska, Russell Bowler, and Katerina Kechris. 2016. The discordant method: a novel approach for differential correlation. *Bioinformatics* 32, 5 (2016), 690–696.

[39] Mingzhou Song and Hua Zhong. 2020. Efficient weighted univariate clustering maps outstanding dysregulated genomic zones in human cancers. *Bioinformatics* (Jul 2020). https://doi.org/10.1093/bioinformatics/btaa613 [Published online ahead of print, 2020 Jul 3].

[40] Despina Tsementzi, Jieying Wu, Samuel Deutsch, Sangeeta Nath, Luis Rodríguez-R, Andrew Burns, Piyush Ranjan, Neha Sarode, Rex Malmstrom, Cory Padilla, Benjamin Stone, Laura Bristow, Morten Larsen, Jennifer Glass, Bo Thamdrup, Tanja Woyke, Konstantinos Konstantinidis, and Frank Stewart. 2016. SAR11 bacteria linked to ocean anoxia and nitrogen loss. *Nature* 536 (08 2016). https://doi.org/10.1038/nature19068

[41] Haizhou Wang and Mingzhou Song. 2011. Ckmeans.1d.dp: Optimal *k*-means clustering in one dimension by dynamic programming. *The R Journal* 3, 2 (2011), 29–33. https://doi.org/10.32614/RJ-2011-015

[42] Jerrold H. Zar. 2009. *Biostatistical Analysis* (5th ed.). Prentice Hall, New Jersey.

[43] Lucie Zinger, Angelique Gobet, and Thomas Pommier. 2012. Two decades of describing the unseen majority of aquatic microbial diversity. *Molecular Ecology* 21, 8 (2012), 1878–1896.