q-mer analysis: a generalized method for analyzing RNA-Seq data.

Tatsuma Shoji [1,2,*], Sato Yoshiharu [2]

*: Corresponding Author

1. qmer LLC. Inc., 2-1-20-706, Honcho, Chuo-ku, Chiba, 260-0012, Japan
2. DNA Chip Research Inc. 1-15-1 Kaigan, Suzue Baydium 5F, Minato-ku, Tokyo 105-0022, Japan

E-mail

Shoji Tatsuma: tatsumashoji@bioinforest.com

Sato Yoshiharu: yo-sato@dna-chip.co.jp

Abstract

Background. RNA-Seq data are usually summarized just by counting the number of reads aligned to each gene. However, this count-based method ignores the alignment information: where each read was mapped in the gene. This is very problematic because the alignment information retains the information regarding the second or third structure

of mRNA, or protein-mRNA interaction including ribosomal distribution over mRNAs, which potentially reflects the sample conditions that we need.

Method. To include the alignment information, we introduced "q-mer analysis", a generalized method for analyzing RNA-Seq data. The q-mer analysis summarizes the RNA-Seq data using the "q-mer vector": the ratio of 4^q kinds of q-length oligomer in the alignment data. Note that, by increasing the q value, q-mer analysis can produce the vector with a higher dimension than the one from the count-based method. This "dimensionality increment" is the key point to describe the sample conditions more accurately than the count-based method does.

Results. We first determined the appropriate q value in *H. sapiens* by investigating how sparse each q-mer vector is using the publically available RNA-Seq dataset taken from 17 healthy or 19 cocaine-addicted cases (GEO accession number: GSE99349) as an example. As a result, q-mer vectors with $q \leqq 9$ had no zero elements, indicating at least q=9 is required to summarize the RNA-Seq data in *H. sapiens*. The $4^9 = 262{,}144$ is approximately 10 times larger than the number of genes in *H. sapiens* (20,022 genes). Next, to examine the ability of q-mer analysis to describe the sample conditions especially with $q > 9$, we prepared the "q-mer matrix" by aligning the 36 q-mer vectors. Then, we decomposed the matrix using the principal component analysis (PCA) and

(PC1, PC2) of each case was plotted onto a two-dimensional plane. As a result, the q-mer matrix with q ≧ 13 with PCA were successfully linearly distinguished cocaine-addicted cases from the healthy control cases on the plane while the matrix derived from the count-based method failed to separate, indicating the q-mer analysis describes the sample conditions more accurately than the count-based method does.

Conclusion. We introduced q-mer analysis, a new method for analyzing RNA-Seq data, and demonstrated q-mer analysis can describe the sample conditions more accurately than the count-based methods does. The combination of RNA-Seq experiment and q-mer analysis will be the promising strong tool when describing the samples conditions or diseases in the future.

Keywords. RNA-Seq, q-mer

Introduction

RNA-Seq is now commonly used to describe sample conditions by counting the reads aligned to each gene [1]. However, this count-based method ignores the alignment information. In other words, the information regarding where each read was mapped in the gene is completely lost. This is very problematic because the number of reads aligned to the transcriptome at a unique region is indicative of the stability of the

region, which retains the information about the second or third structure of mRNA or protein-mRNA interaction including the ribosomal distribution over mRNAs, which potentially reflects the sample conditions that we need [2,3]. Here, we provide the new method to summarize the RNA-Seq data without losing the alignment information.

Methods

Example RNA-Seq dataset

The example RNA-Seq dataset in this study were retrieved from GEO (GEO accession number: GSE99349) [4]. This dataset contains 36 fastq files: 17 are from healthy control cases, 19 are from cocaine-addicted cases. We selected this dataset because we aim to show the q-mer analysis is useful especially in psychiatric diseases in which transcriptome analysis tends to be complicated [4] and because the RNA-Seq libraries in the study were constructed not by the poly-A method. The Poly-A method only catches the alignment information near the 3' end, which is not appropriate for our purpose.

Count-based matrix

The 36 fastq files were preprocessed by clipping Illumina adapter sequences using Trimmomatic v.0.39 [5] and aligned to the human genome sequence (GenBank assembly accession: GCA_000001405.28) using HISAT2 v.2.2.1[6]. Then the read counts for each gene were calculated by featureCounts v.2.0.1 [7] using the human gene feature file (INSDC Assembly: GCA_000001405.28, Database version: 103.38) as a reference. Finally, the resulting 36 count data were aligned to get the count-based matrix. The size of the matrix was 36 x 20,022.

qmer vector and qmer matrix

The number of the frequency for each 4^q kinds of q-length oligomer in the sam file produced by HISAT2 was counted and then normalized by the number of the frequency of the oligomer in the genome. Note that we skipped the line in the sam file if the read was not uniquely mapped and if the read contained the character "N". The resulting 4^q dimension vector was defined as the q-mer vector. To get the q-mer matrix, we just aligned 36 q-mer vectors. The size of the matrix was 36 x 4^q.

Decomposition of the matrix

To decompose the matrix and plot 36 cases onto a two-dimensional plane, we used principal component analysis (PCA) implemented with "scikit-learn (0.24.1)" [8]. Briefly, we first selected 10 columns from the matrix. The 10 columns had the top 10 highest correlation coefficients against the y vector, where the element is 0 if the corresponding case is the healthy control case otherwise 1. Then we decomposed the matrix whose size is 36 x 10 using PCA. Finally, the (PC1, PC2) for each case was plotted onto a two-dimensional plane. To reproduce our result, follow the jupyter notebook at https://github.com/tatsumashoji/qmer.

Results

q-mer analysis of the RNA-Seq data.

The simplest way to approximate the RNA-Seq data is the ratio of A, T, G, and C. The second to the simplest way to approximate the RNA-Seq data is the ratio of the AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG, GC, CA, CT, CG, and CC. In this way, we can approximate the RNA-Seq data with the ratio of 4^q kinds of q-length oligomer (Figure 1). If q is large enough, this approximation includes the count-based RNA-Seq data. Hereinafter, we introduce this approximation as "q-mer transformation"

and the resulting 4^q dimension vector and matrix as "q-mer vector" and "q-mer matrix" respectively.

q=9 is required to express the RNA-Seq data in *H. sapiens*.

To investigate what q value is appropriate to express the RNA-Seq data in *H. sapiens*, we calculated the q-mer vector with the q value of 1 to 14 respectively using the RNA-Seq data from *H. Sapiens* as an example. The example data is retrieved from the NCBI database with GEO accession number GSE99349 [4]. The rationale for picking the value of 14 is that the total length of all mRNAs in *H. sapiens* (= 589,150,963) is larger than 4^{14} but less than 4^{15}, which mean the q-mer vector with the q value of more than 14 is sparse. Table 1 summarizes the number of the zero elements observed in q-mer vectors in the example data. Surprisingly, the zero elements are not observed at all in q-mer vectors with the q value of less than 10, indicating q=9 is at least required to express *H. sapiens* RNA-Seq data.

The ability of q-mer analysis to describe the sample conditions.

The dimension of q-mer vector with the q value of 9 is 4^9 = 262,144. This value is much more than the number of genes in *H. sapiens* (= 20,022). This led us to

speculate q-mer analysis could express the sample condition more accurately than the count-based method does. To investigate the ability of q-mer analysis to describe the sample conditions, we reused the publically available 36 RNA-Seq data (GEO accession number: GSE99349, same as the above). This RNA-Seq experiment was performed on neuronal nuclei isolated from the post-mortem dorsolateral prefrontal cortex of 19 cocaine-addicted and 17 healthy control cases. We first summarized the 36 RNA-Seq data with the count-based method and got a gene expression matrix whose size is 36 x 20,022. We then decomposed the matrix using principal component analysis (PCA) and plotted (PC1, PC2) of each case (Figure 2(A)). The resulting plot failed to clearly distinguish the cocaine-addicted cases from the healthy cases. On the other hand, when we summarized the 36 RNA-Seq data using the q-mer analysis with the q value of 1 to 14 and decompose the resulting q-mer matrix with PCA, the cocaine-addicted cases and healthy control cases were linearly subgrouped with the q value of 13 or 14 (Figure 2 (N), (O)). These results show the q-mer analysis's ability to describe the sample conditions.

Discussion

In this study, we suggested the new method "q-mer analysis" for analyzing the RNA-Seq data to include the alignment information. This approach produces the vectors with much higher dimension than the count-based method does and this “dimensionality increment” was demonstrated to be the key point to successfully describe the sample conditions in that non-supervised subgrouping was successful when analyzing the RNA-Seq data with q-mer analysis with the large q values (Figure 2(N)(O)) while the count-based method failed to subgroup (Figure 2(A)).

The interpretability of the q-mer analysis is an important issue. In other words, what kind of biological natures did the q-mer analysis capture to discriminate the cocaine-addicted cases from the healthy control cases? The q-mer analysis can easily answer this question by examining what kind of q-length oligomers contributed to PC1 or PC2. For example, the contribution rate of PC1 in Figure 2(O) was 0.70 and the major contributors for PC1 were the “CAAAAAGCGTAAAT” and “AAGTATCCCAAGGG”. These two 14-length oligomers were observed only in the healthy control group. Thus, we can say the q-mer analysis defined the cocaine-addicted cases as the low content of these two 14-length oligomers in the transcriptome.

For the detailed biological and structural mechanisms among sample conditions, the q-mer analysis cannot provide a clear answer. However, the q-mer analysis can

provide some hints for further investigation. For instance, the two oligomers above mentioned were observed in XYLT1 which encodes xylosyltransferase 1, and in LINK02250 which is one of the lincRNAs respectively. This is consistent with the report [4] that alterations in lincRNAs were observed between healthy control and cocaine-addicted cases. Although they characterized the cocaine-addicted cases based on the genes or the group of genes in the report [4], the specific region in the gene might more accurately characterize the cases. We suggest that there might be a limit for the gene-level characterization to understand the complicated sample conditions such as psychiatric diseases. The q-mer analysis can tell us there might be complicated mechanisms that require the region-level explanation. This helps us choose what kind of experiments would be appropriate for investigating the detailed mechanisms.

The dimension required to express the transcriptome was estimated at least 4^9 = 262,144 (Table 1). Further, to catch the difference among the sample conditions accurately, the required dimension may be 4^{14} = 268,435,456 or larger. Thus, the dimension of RNA-Seq data potentially has approximately 10,000 times the number of the genes in *H. sapiens*. However, some reports say that the affective dimension of the transcriptome should be far less than the total number of genes it contains [9,10,11]. This contradiction may be because they only investigated in the ideal situation. In the

real situation or the situations that we have not investigated yet, especially psychiatric diseases in *H. sapiens*, transcriptome would have a large number of dimensions.

The nature of the high dimensionality of the RNA-Seq data might be the strong point of the RNA-Seq experiment. Recently, we describe the sample conditions by combining the multi-omics data to get more explanatory variables. However, this approach is very expensive and hard to combine. Just combining the RNA-Seq experiment with the q-mer analysis may be enough to describe the sample conditions because the number of the dimension of the RNA-Seq is much more than the total dimension derived from the multi-omics approach. The q-mer analysis will be the new standard instead of the omics-related research in the future.

Conclusions

We introduced q-mer analysis, a generalized method for analyzing RNA-Seq data, and demonstrated the RNA-Seq data are much more informative than what is expected to be by the count-based method. At least, we need 4^9 dimensions to express the RNA-Seq data. Further, the additional information introduced by the q-mer analysis is very important in that q-mer vectors with q values larger than 13 enable us to catch the difference among samples, which the count-based methods could not catch. The

combination of RNA-Seq experiment and q-mer analysis is the promising strong method when giving the biological definition to sample conditions or psychiatric diseases.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Author contributions

ST suggested q-mer analysis and showed its usefulness in giving the biological definition to the sample conditions especially psychiatric diseases in humans by analyzing the publically available RNA-Seq data taken from cocaine-addicted or healthy control cases. Also, ST was responsible for writing manuscripts and making Figures and Tables. SY checked the programs that ST wrote, reviewed the manuscript, and supported the research grant.

Acknowledgments

Not applicable.

References

1. Stark, Rory, Marta Grzelak, and James Hadfield. "RNA sequencing: the teenage years." Nature Reviews Genetics 20.11 (2019): 631-656.

2. Shoji, Tatsuma, et al. "Ribosome profiling in Streptococcus pneumoniae reveals the role of methylation of 23S rRNA nucleotide G748 on ribosome stalling." J Genet Genomic Sci 6 (2021): 024.

3. Shoji, Tatsuma. "Methylation of 23S rRNA G748 and the ribosomal protein L22 Lys-94 are critical factors for maintaining the association between ribosome stalling and proteome composition in Streptococcus pneumoniae." J Genet Genomic Sci 6 (2021): 026.

4. Ribeiro, Efrain A., et al. "Gene network dysregulation in dorsolateral prefrontal cortex neurons of humans with cocaine use disorder." Scientific reports 7.1 (2017): 1-10.

5. Bolger, Anthony, and F. Giorgi. “Trimmomatic: a flexible read trimming tool for illumina NGS data.” Bioinformatics 30.15 (2014): 2114–2120.

6. Kim, Daehwan, et al. “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.” Nature biotechnology 37.8 (2019): 907–915.

7. Liao, Yang, Gordon K. Smyth, and Wei Shi. "The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote." Nucleic acids research 41.10 (2013): e108-e108.

8. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825-30.

9. Heimberg, Graham, et al. "Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing." Cell systems 2.4 (2016): 239-250.

10. Biswas, Surojit, et al. "Tradict enables accurate prediction of eukaryotic transcriptional states from 100 marker genes." Nature communications 8.1 (2017): 1-10.

11. Kobayashi-Kirschvink, Koseki J., et al. "Linear regression links transcriptomic data and cellular Raman spectra." Cell systems 7.1 (2018): 104-117.

Figure Legends

Fig. 1 The schematic image for the count-based matrix and the q-mer matrices. The traditional count-based summary of the RNA-Seq data is shown on the left. q-mer matrices are shown on the right. The number of the explanatory variables is shown above each table. The q value is shown at the right of the table.

Fig. 2 The ability of the matrices to distinguish the healthy control cases from the cocaine-addicted cases.

The distribution of the healthy control cases (closed black circle) and the cocaine-addicted cases (closed red triangle) on the PC1 PC2 plane based on the matrix from (A) the count-based method and (B)-(O) the q-mer analysis. For the panels of (B)-(O), the q value was shown above the panel. In the panel (N) and (O), the borderline that linearly separates the cocaine-addicted cases from the healthy control cases is inserted.

Tables

Table 1 The sparsity of q-mer vectors in *H. sapiens*.

The mean value and S.D. of the number of zero elements in 36 q-mer vectors are shown in the column of "# of zero elements". The mean value and S.D. of the ratio of the zero elements against the size of vector (i.e. 4^q) are shown in the column of "% of zero elements".

q	4^q	# of zero elements	% of zero elements
1	4	0.0 ± 0.0	0.0 ± 0.0

2	16	0.0 ± 0.0	0.0 ± 0.0
3	64	0.0 ± 0.0	0.0 ± 0.0
4	256	0.0 ± 0.0	0.0 ± 0.0
5	1024	0.0 ± 0.0	0.0 ± 0.0
6	4096	0.0 ± 0.0	0.0 ± 0.0
7	16384	0.0 ± 0.0	0.0 ± 0.0
8	65536	0.0 ± 0.0	0.0 ± 0.0
9	262144	0.0 ± 0.0	0.0 ± 0.0
10	1048576	0.14 ± 0.42	0.0 ± 0.0
11	4194304	1944.53 ± 1489.84	0.05 ± 0.04
12	16777216	299921.78 ± 123628.22	1.79 ± 0.74
13	67108864	8527168.92 ± 2068364.84	12.71 ± 3.08
14	268435456	99835336.33 ± 12089958.3	37.19 ± 4.5

Figure 1

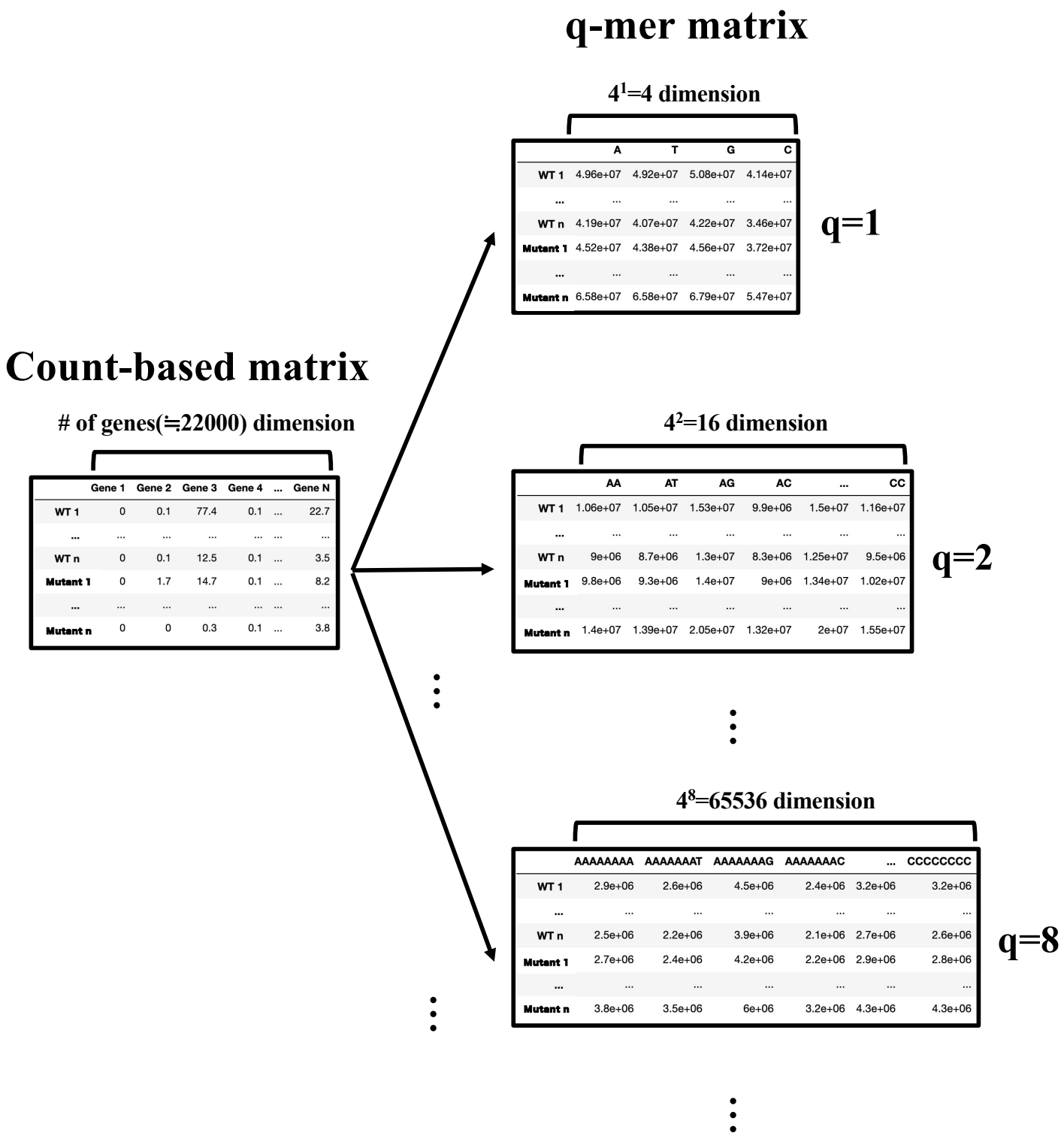
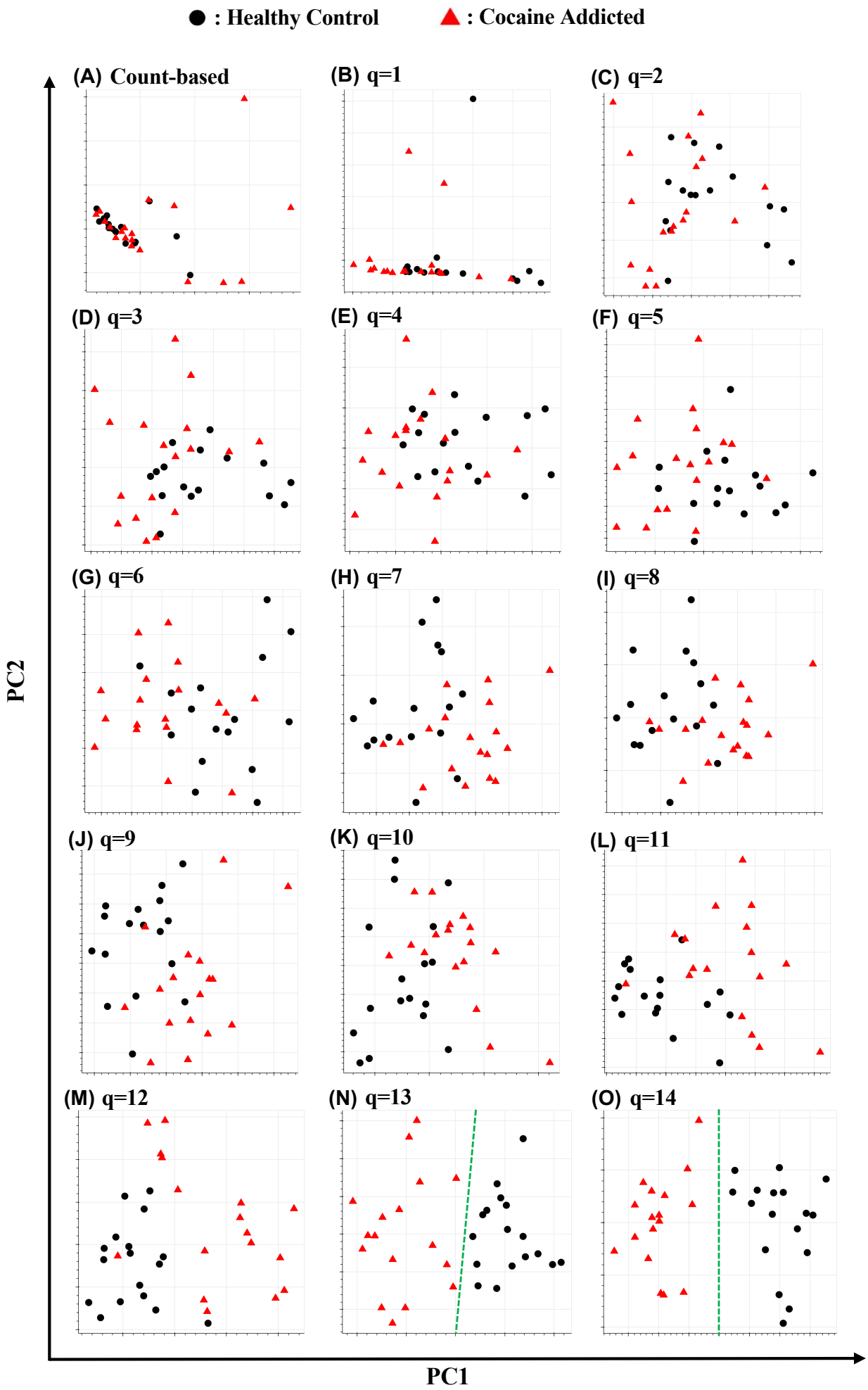

q-mer matrix
4^1=4 dimension
A T G C
WT 1 4.96e+07 4.92e+07 5.08e+07 4.14e+07
... ...
WT n 4.19e+07 4.07e+07 4.22e+07 3.46e+07
Mutant 1 4.52e+07 4.38e+07 4.56e+07 3.72e+07
... ...
Mutant n 6.58e+07 6.58e+07 6.79e+07 5.47e+07
q=1
Count-based matrix
# of genes(≒22000) dimension
Gene 1 Gene 2 Gene 3 Gene 4 ... Gene N
WT 1 0 0.1 77.4 0.1 ... 22.7
... ...
WT n 0 0.1 12.5 0.1 ... 3.5
Mutant 1 0 1.7 14.7 0.1 ... 8.2
... ...
Mutant n 0 0 0.3 0.1 ... 3.8
4^2=16 dimension
AA AT AG AC ... CC
WT 1 1.06e+07 1.05e+07 1.53e+07 9.9e+06 1.5e+07 1.16e+07
... ...
WT n 9e+06 8.7e+06 1.3e+07 8.3e+06 1.25e+07 9.5e+06
Mutant 1 9.8e+06 9.3e+06 1.4e+07 9e+06 1.34e+07 1.02e+07
... ...
Mutant n 1.4e+07 1.39e+07 2.05e+07 1.32e+07 2e+07 1.55e+07
q=2
4^8=65536 dimension
AAAAAAAA AAAAAAAT AAAAAAAG AAAAAAAC ... CCCCCCCC
WT 1 2.9e+06 2.6e+06 4.5e+06 2.4e+06 3.2e+06 3.2e+06
... ...
WT n 2.5e+06 2.2e+06 3.9e+06 2.1e+06 2.7e+06 2.6e+06
Mutant 1 2.7e+06 2.4e+06 4.2e+06 2.2e+06 2.9e+06 2.8e+06
... ...
Mutant n 3.8e+06 3.5e+06 6e+06 3.2e+06 4.3e+06 4.3e+06
q=8

Figure 2


● : Healthy Control
▲ : Cocaine Addicted
(A) Count-based
(B) q=1
(C) q=2
(D) q=3
(E) q=4
(F) q=5
(G) q=6
(H) q=7
(I) q=8
(J) q=9
(K) q=10
(L) q=11
(M) q=12
(N) q=13
(O) q=14
PC2
PC1