fortune Journals

Research Article

JOURNAL OF BIOINFORMATICS AND SYSTEMS BIOLOGY

ISSN: 2688-5107

Q-Mer Analysis: A Generalized Method for Analyzing RNA-Seq Data

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

Abstract

Background: RNA-Seq data are usually summarized by counting the number of transcript reads aligned to each gene. However, count-based methods do not take alignment information, where and how each read was mapped in the gene, into account. This information is essential to characterize samples accurately. In this study, we developed a method to summarize RNA-Seq data without losing alignment information.

Results: To include alignment information, we introduce "q-mer analysis," which summarizes RNA-Seq data with 4q kinds of q-length oligomers. Using publicly available RNA-Seq datasets, we demonstrate that at least q ≧ 9 is required for capturing alignment information in Homo sapiens. It should be noted that 49 = 262,144 is approximately 10 times larger than the number of genes in H. sapiens (20,022 genes). Furthermore, principal component analysis showed that q-mer analysis with q = 14 linearly distinguished samples from controls, while a count-based method failed. These results indicate that alignment information is essential to characterize transcriptomics samples.

Conclusions: In conclusion, we introduce q-mer analysis to include alignment information in RNA-Seq analysis and demonstrate the superiority of q-mer analysis over count-based methods in that q-mer analysis can distinguish case samples from controls. Combining RNA-Seq research with q-mer analysis could be useful for identifying distinguishing transcriptomic features that could provide hypotheses for disease mechanisms.

Keywords: RNA-Seq; count-based method; q-mer analysis; alignment information; q-mer vector; dimensionality increment.

Background

RNA-Seq is commonly used in molecular biology [1] since its development more than a decade ago [2–6]. Many studies have characterized samples at the transcriptional level using RNA-Seq or at the post-transcriptional level by coupling the appropriate biochemical assay with RNA-Seq [7–9]. The number of publications containing RNA-Seq data is approximately 36,000 in 2021 (PubMed). Furthermore, many types of software have been developed for analyzing RNA-Seq data [10–12]. RNA-Seq is now an indispensable technology.

In terms of RNA-Seq data analysis, the core task of mapping and counting reads is common to many kinds of software. After the mapping process, RNA-Seq data is summarized by counting the reads aligned to each exon, gene, or transcript [11] and is reported in the form of gene expression tables [13–16]. However, count-based methods do not consider alignment information.

Affiliation:

[1]qmer LLC. Inc. 2-1-20, 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

***Corresponding author:**
Tatsuma Shoji qmer LLC. Inc. 2-1-20, Honcho, Chuo-ku, Chiba, 260-0012, Japan

Email: tatsumashoji@bioinforest.com

Citation: Tatsuma Shoji and Yoshiharu Sato. Q-Mer Analysis: A Generalized Method for Analyzing RNA-Seq Data. Journal of Bioinformatics and Systems Biology. 6 (2023): 60-73.

Received: February 04, 2023
Accepted: February 11, 2023
Published: March 13, 2023

Shoji T and Sato Y., J Bioinform Syst Biol 2023
DOI:10.26502/jbsb.5107049

In other words, count-based methods discard information regarding where and how the aligner mapped each read in the exon, gene, or transcript. If the difference between samples and controls is well characterized at the transcriptional level, and differential gene expression analysis is sufficient for the data analysis, count-based methods work well [17]. There is a valid statistical theory that explains the fluctuations of the read counts [18–21], and many software products test whether the observed difference between samples and controls is significant [19–23]. However, if the aim of the study involves post-transcriptional regulation or mutations in the genome, the loss of alignment information is problematic for two reasons. First, count-based methods ignore mismatches in alignment information, which are ”expressed mutations,” and thus are critical features of samples. Second, count-based methods do not take the number of reads aligned to the transcriptome at a unique region into account, which retains post-transcriptional information: the secondary or tertiary structure of mRNA or protein–mRNA interactions, including the ribosomal distribution across mRNAs [24], which potentially reflects the sample characteristics of interest [25]. Therefore, RNA-Seq data should be summarized in a more informative way by taking alignment information into account to accurately describe samples.

Although several studies have described the usefulness of alignment information and used it to detect mRNA isoforms [17-26], applications for taking alignment information into account are limited. Currently, no studies focus on summarizing alignment information in RNA-Seq data.

In this study, we provide a method to describe samples more accurately by including alignment information. First, we introduce a new method called “q-mer analysis” that summarizes RNA-Seq data without losing alignment information using a “q-mer vector,” which is the occurrence rate of 4q kinds of q-length oligomers in RNA-Seq data. We then determine the appropriate q value and assess the ability of q-mer analysis to describe samples using two publicly available RNA-Seq datasets from Homo sapiens. Further investigation shows the superiority of q-mer analysis to count-based methods.

Results

q-mer Analysis of RNA-Seq Data

The simplest way to approximate RNA-Seq data is using the ratio of A, T, G, and C nucleotides in the alignment data. The second simplest way to approximate the RNA-Seq data is using the ratio of the 42 kinds of oligomers, namely, AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG, GC, CA, CT, CG, and CC, in the alignment data. In this way, we can approximate the RNA-Seq data and express alignment information at the same time using the occurrence rate of 4q kinds of q-length oligomers (Figure 1). If q is large enough, this approximation includes count-based RNA-Seq data. Hereafter, we introduce this approximation as “q-mer analysis” and the resulting 4q dimension vector and matrix as the “q-mer vector” and “q-mer matrix,” respectively.

q ≧ 9 is Required to Express Alignment Information in H. sapiens

To investigate the appropriate q value to express alignment information of RNA-Seq data in H. sapiens, we calculated the q-mer vector with a q value of 1 to 14 for two publicly available RNA-Seq datasets from H. sapiens (GEO accession numbers GSE99349 [27] and GSE106589 [28]). The first RNA-Seq experiment (GSE99349) was performed on neuronal nuclei isolated from the post-mortem dorsolateral prefrontal cortex of 19 cocaine-addicted and 17 healthy control cases. The second RNA-Seq experiment (GSE106589) was performed on 18 human-induced pluripotent stem cell-derived neural progenitor cells (hiPSCs-NPCs) derived from 14 individuals with childhood-onset schizophrenia (COS) and 20 hiPSCs-NPCs from 12 unrelated healthy controls. Regarding the selection of the RNA-Seq datasets, see the selection criteria described in Section 4. The rationale for picking a range of q values from 1 to 14 is that the total length of all mRNAs in H. sapiens (589,150,963) is larger than 414 but less than 415, which means the q-mer vectors with q values greater than 14 are sparse.

Table 1 summarizes the number of zero elements observed in the q-mer vectors in the example data. Surprisingly, zero elements were not observed in q-mer vectors with q values less than 10, indicating that q = 9 is at least required to express alignment information from RNA-Seq data in H. sapiens. Note that 49 (262,144) is approximately 10 times the number of genes in H. sapiens, suggesting that the dimensions of gene expression tables are too low to accurately summarize RNA-Seq data.

Sample Description with q-mer Analysis

The high dimensionality of RNA-Seq data in H. sapiens may result from the inclusion of alignment information. This very large number led us to speculate that q-mer analysis could describe the samples more accurately than could count-based methods. To investigate the ability of q-mer analysis to describe these samples, we applied principal component analysis (PCA) to the two RNA-Seq datasets described above after preprocessing (Figure 2).

Briefly, we first summarized these RNA-Seq datasets using a count-based method and obtained gene expression matrices with the sizes 36 × 20,022 or 38 × 20,022 for the first and second studies, respectively (Figure 2, top-left table). Then, after selecting 10 genes that showed the top 10 magnitudes of the correlation coefficient against the y vector (Figure 2, center), we decomposed the resulting matrix using PCA and plotted principal component (PC) 1 and PC2 for

Shoji T and Sato Y., J Bioinform Syst Biol 2023
DOI:10.26502/jbsb.5107049

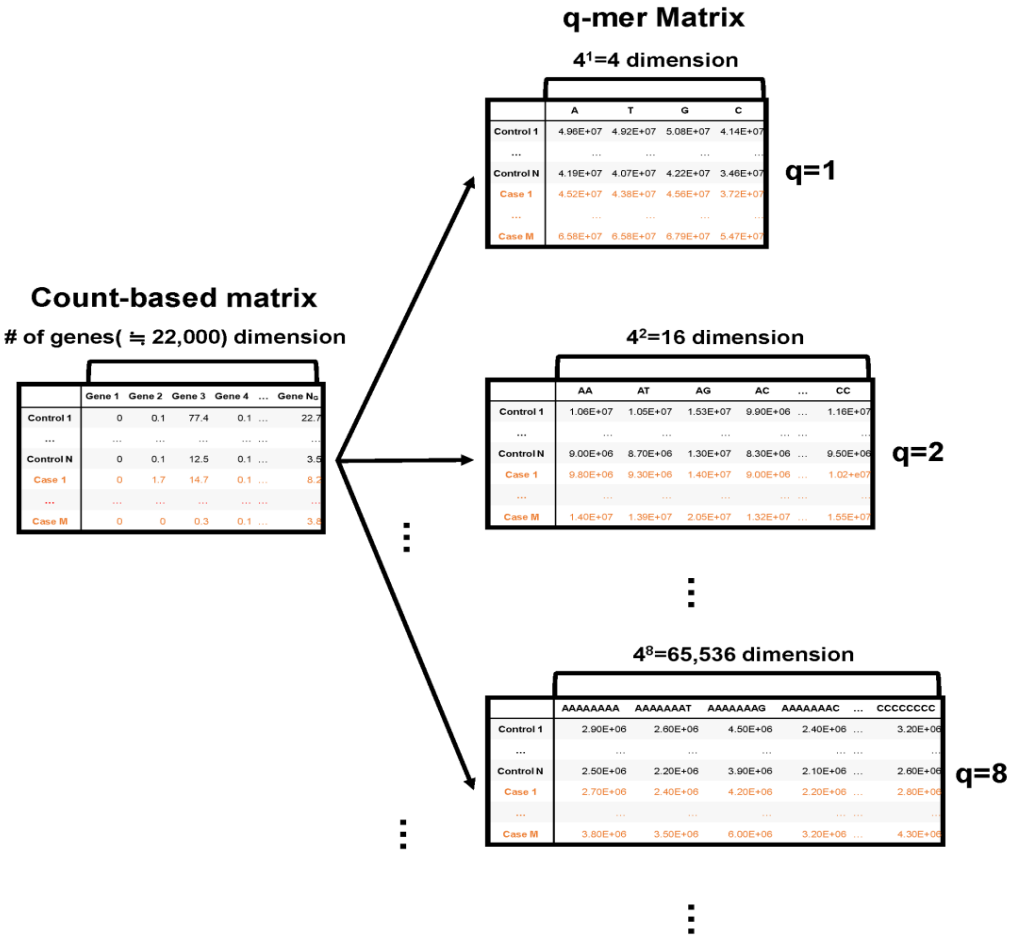


Figure 1: Schematic image of the count-based matrix and the q-mer matrices. The count-based summary of RNA-Seq data is shown on the left, while q-mer matrices are shown on the right. The number of the explanatory variables is shown above each table. The q value is shown to the right of each table.

Table 1: The sparsity of q-mer vectors in H. sapiens. The mean and S.D. of the number or percentage of zero elements in the q-mer vectors are shown in the "Number" or "Percentage" columns, respectively.

q	4^q	GSE99349		GSE106589	
		Number	Percentage	Number	Percentage
1	4	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
2	16	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
3	64	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
4	256	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
5	1,024	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
6	4,096	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
7	16,384	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
8	65,536	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
9	2,62,144	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
10	10,48,576	0.14 ± 0.42	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
11	41,94,304	1,944.53 ± 1,489.84	0.05 ± 0.04	400.11 ± 195.84	0.01 ± 0.00
12	1,67,77,216	299,921.78 ± 123,628.22	1.79 ± 0.74	137,672.45 ± 33,426.43	0.82 ± 0.20
13	6,71,08,864	8,527,168.92 ± 2,068,364.84	12.71 ± 3.08	5,563,000.03 ± 713,139.60	8.29 ± 1.06
14	26,84,35,456	99,835,336.33 ± 12,089,958.30	37.19 ± 4.50	81,236,506.71 ± 5,119,176.48	30.26 ± 1.91

Shoji T and Sato Y., J Bioinform Syst Biol 2023
DOI:10.26502/jbsb.5107049

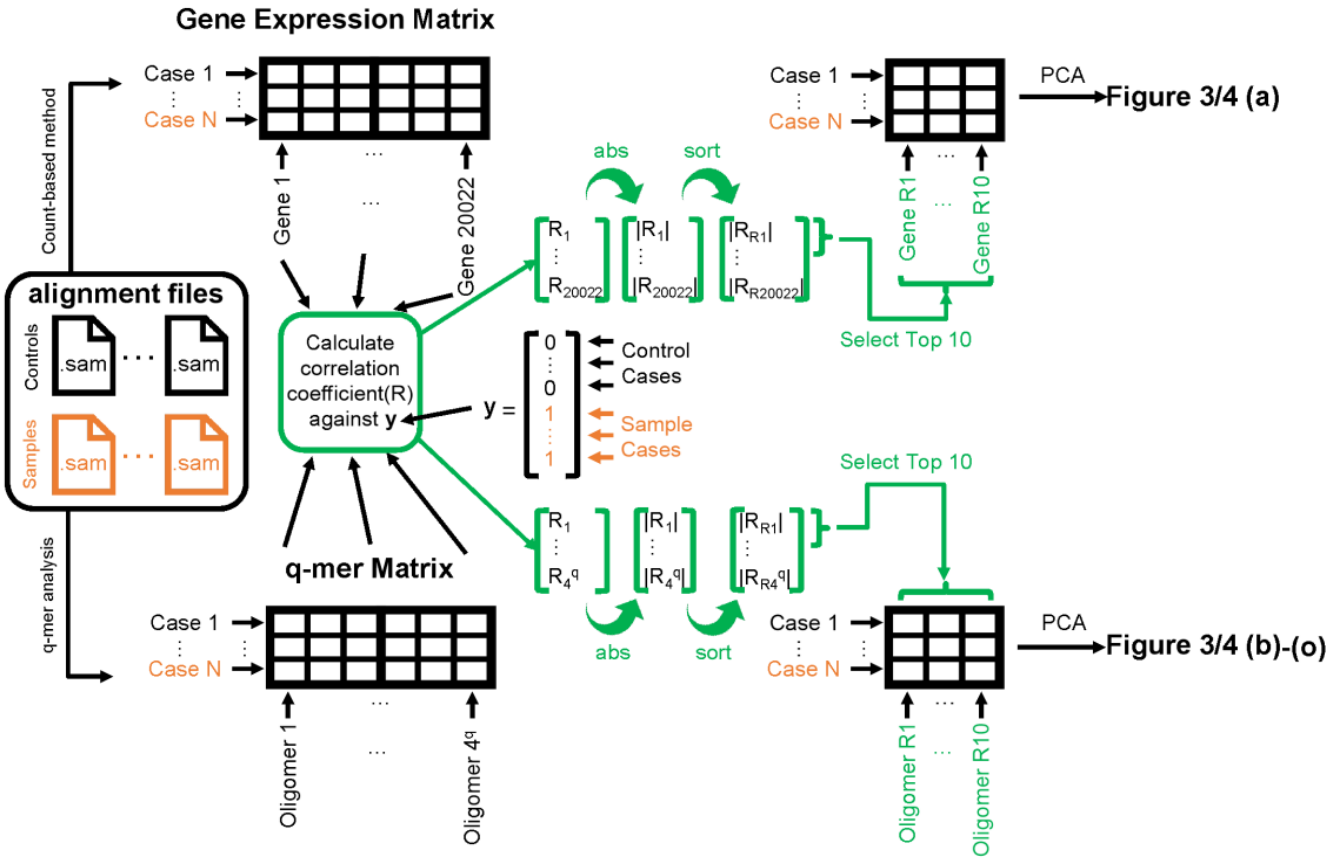


Figure 2: Schematic image of steps to apply principal component analysis (PCA) and draw figures 3, 4. The left panel shows the alignment files (black for healthy control cases and orange for cocaine-addicted or childhood-onset schizophrenia cases). The exact entity of the alignment file is a .sam file. These alignment files are converted to gene expression matrices (top-left table) by the count-based method or q-mer matrices (bottom-left table) by q-mer analysis. The correlation coefficient "R" is calculated against the y vector for each column in the table (center). The 10 columns on the top-right or bottom-right table are selected based on the size of the absolute value of "R" (indicated in green). Figures 3 and 4 were drawn based on the results of the PCA of the top-right table or bottom-left table.

each sample (Figure 2, top-right table). The resulting plot was unable to distinguish the cocaine-addicted cases or the COS cases from the respective healthy controls (Figure 3(a) and Figure 4(a), respectively). On the other hand, when we summarized the RNA-Seq datasets using q-mer analysis with q values of 1 to 14 (Figure 2, bottom, Figure 3(b-o), Figure 4(b-o)), the cocaine-addicted cases and the COS cases were linearly separated from the respective healthy control cases with q values of 12–14 and 14, respectively (Figure 3(m–o) and Figure 4(o)). Note that applying PCA to the gene expression matrix directly also failed to separate the cases from the controls (See Figure S1, Additional File 1). These results support the idea that q-mer analysis is able to describe the samples more accurately than other methods.

Interpretability of q-mer Analysis

The interpretability of q-mer analysis is an important issue to confirm its utility. In other words, we next sought to investigate which biological features q-mer analysis captures to distinguish cocaine-addicted or COS cases from their respective healthy control cases.

Regarding the study of cocaine addiction, the contribution of PC1 in Figure 3(o) was 0.91 (Figure 5(a)), and all of the oligomers contributed to PC1 to the same extent (Figure 5(b)). These oligomers were only expressed either in the controls or in the cocaine-addicted cases and as part of one or two genes (Figure 5(c)). Interestingly, some of the identified genes are expressed specifically in the brain and are reported to be involved in neurological function. When examining the alignment for the region where the "ACTCGACCAAAAAT" oligomer was observed, for example, the shape of the alignment was different between the cocaine-addicted cases and the healthy controls (Figure 5(d)), indicating a difference in the post-transcriptional regulation between the two groups. The same feature was true for other oligomers (See Figure S2, Additional File 2). Note that the false discovery rates (FDRs) calculated based on the count-based method suggested that none of the genes in Figure 5(c) were significantly differentially expressed between the two groups. Furthermore, PCA of the count-based matrices only with the genes in Figure 5(c) failed to separate the cocaine-addicted

Shoji T and Sato Y., J Bioinform Syst Biol 2023
DOI:10.26502/jbsb.5107049

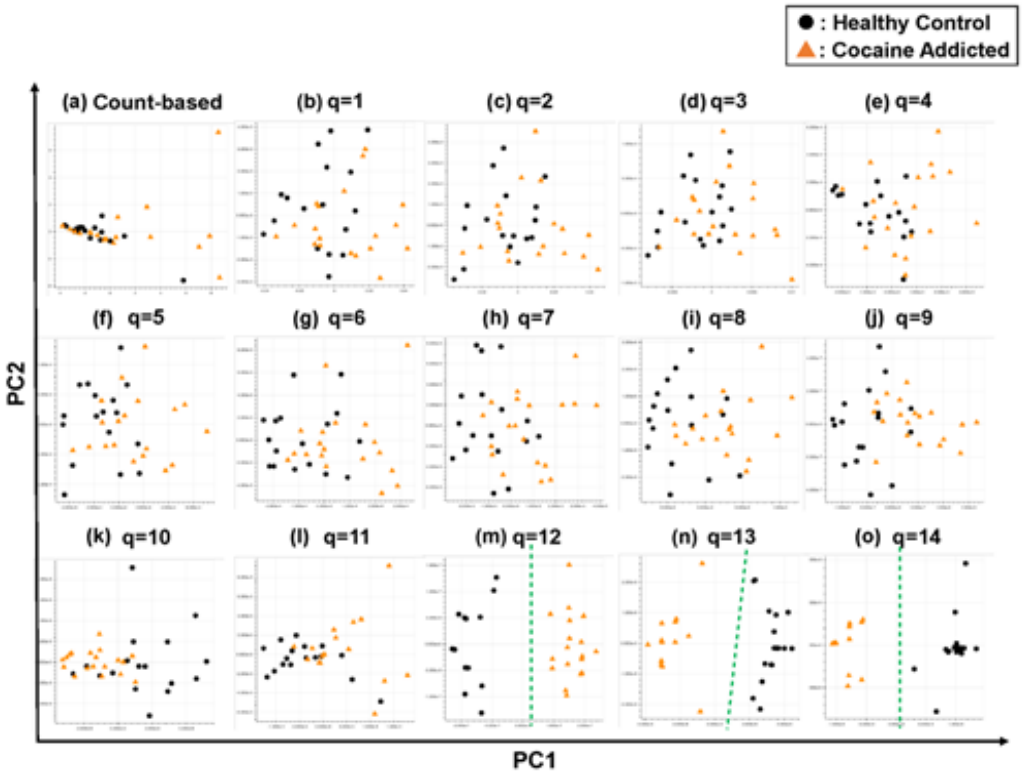


Figure 3: The ability of the matrices to distinguish healthy controls from the cocaine-addicted cases. The distribution of the healthy control cases (closed black circles) and cocaine-addicted samples (closed red triangles) on the PC1/PC2 plane were plotted based on the matrix from (a) the count-based method and (b–o) q-mer analysis. For panels (b)–(o), the q value is shown above the panel. In panels (m)–(o), a green dotted line separates the cocaine-addicted cases from the healthy controls.

cases from the healthy controls (See Figure S3(a), Additional File 3), implying that it is essential to be able to discriminate between specific gene regions. These results suggest that q-mer analysis focuses on specific regions of each gene and is able to capture these differences in the shape of the alignments to discriminate the cocaine-addicted cases from healthy controls in this RNA-Seq dataset.

For the study of COS, the contribution rate of PC1 in Figure 4(o) was 0.90 (Figure 6(a)), and nine oligomers contributed to PC1 to the same extent (Figure 6(b)). These nine oligomers were only expressed in the healthy controls (Figure 6(c)), six of which were observed in the same region of the TTBK2 gene (Figure 6(c)–(d)). However, there was not a clear difference in the shape of the alignment. Instead, we detected a mutation in the region. The same feature was confirmed with the other oligomers (See Figure S4, Additional File 4). The FDRs calculated based on the count-based method suggested that none of the genes in Figure 6(c) were significantly differentially expressed between the two groups. Furthermore, PCA of the count-based matrices only with the genes in Figure 6(c) failed to separate the COS cases from the healthy controls (See Figure S3(b), Additional File 3). These results support the idea that q-mer analysis in this RNA-Seq dataset detected mismatch information that the count-based method does not include to separate the cases from the controls.

Methods

Criteria for Choosing Example RNA-Seq Datasets

Example RNA-Seq datasets were selected based on five criteria: (1) the dataset was related to neurological or psychiatric disorders, (2) the raw sequence data were available, (3) the RNA-Seq libraries were not constructed using the poly-A method, (4) more than 20 RNA-Seq samples were available, and (5) the RNA-Seq data were from H. sapiens.

Example RNA-Seq Datasets

Two RNA-Seq datasets were chosen based on the criteria mentioned above: GEO accession numbers GSE99349 [27] and GSE106589 [28]. The first dataset, GSE99349, contained 36 fastq files: 17 from healthy controls and 19 from cocaine-addicted cases. All 36 fastq files were used for q-mer analysis. The second dataset, GSE106589, contained 94 fastq files: 20

Shoji T and Sato Y., J Bioinform Syst Biol 2023
DOI:10.26502/jbsb.5107049

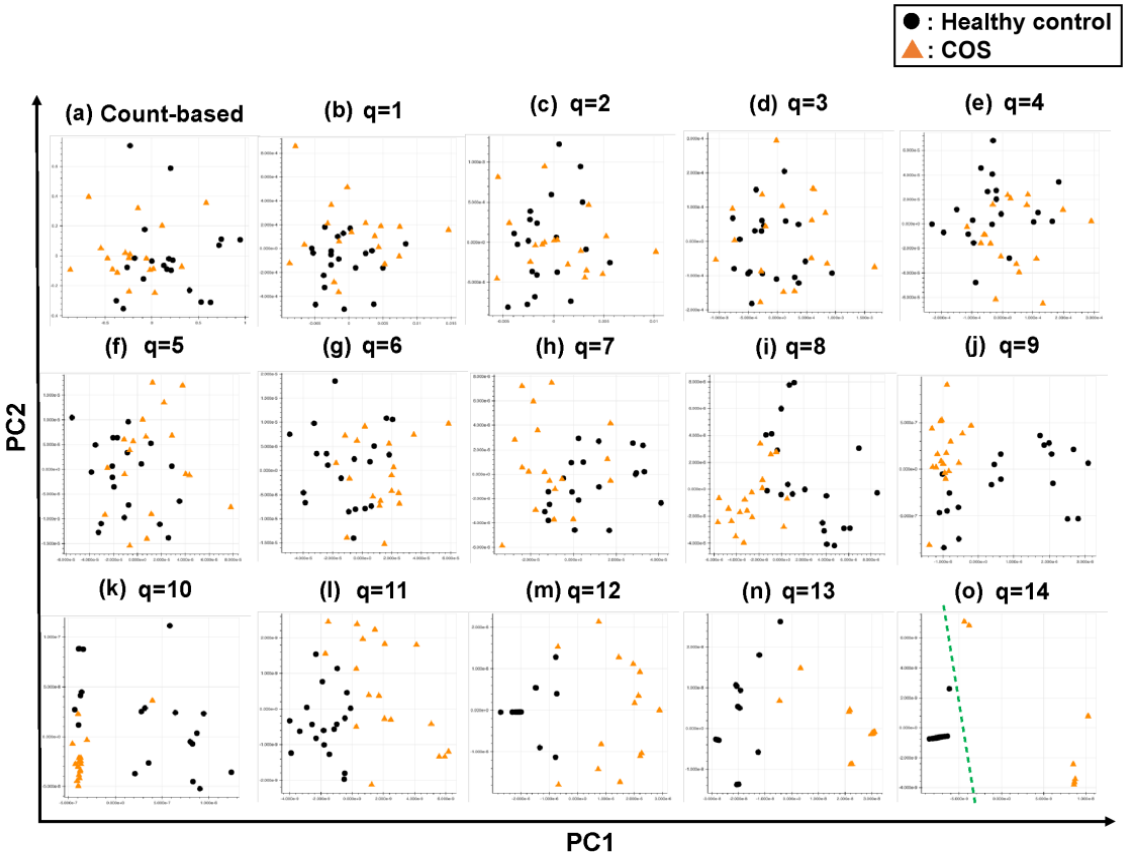


Figure 4: The ability of the matrices to distinguish healthy control cases from the COS cases. The distribution of the healthy control cases (closed black circles) and COS cases (closed red triangles) on the PC1/PC2 plane were plotted based on the matrix from (a) the count-based method and (b)–(o) the q-mer analysis. For panels (b)–(o), the q value is shown above the panel. In panel (o), the green dotted line separates the cocaine-addicted cases from the healthy control cases.

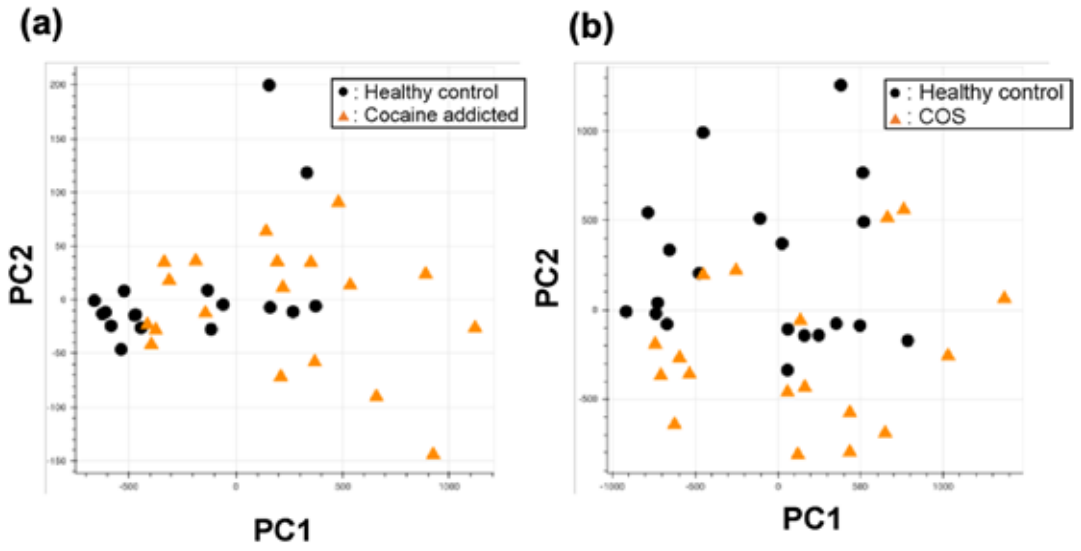


Figure S1: PCA of the gene expression table without gene selection for the study of cocaine addiction (a) and childhood-onset schizophrenia (COS) (b).

Shoji T and Sato Y., J Bioinform Syst Biol 2023
DOI:10.26502/jbsb.5107049

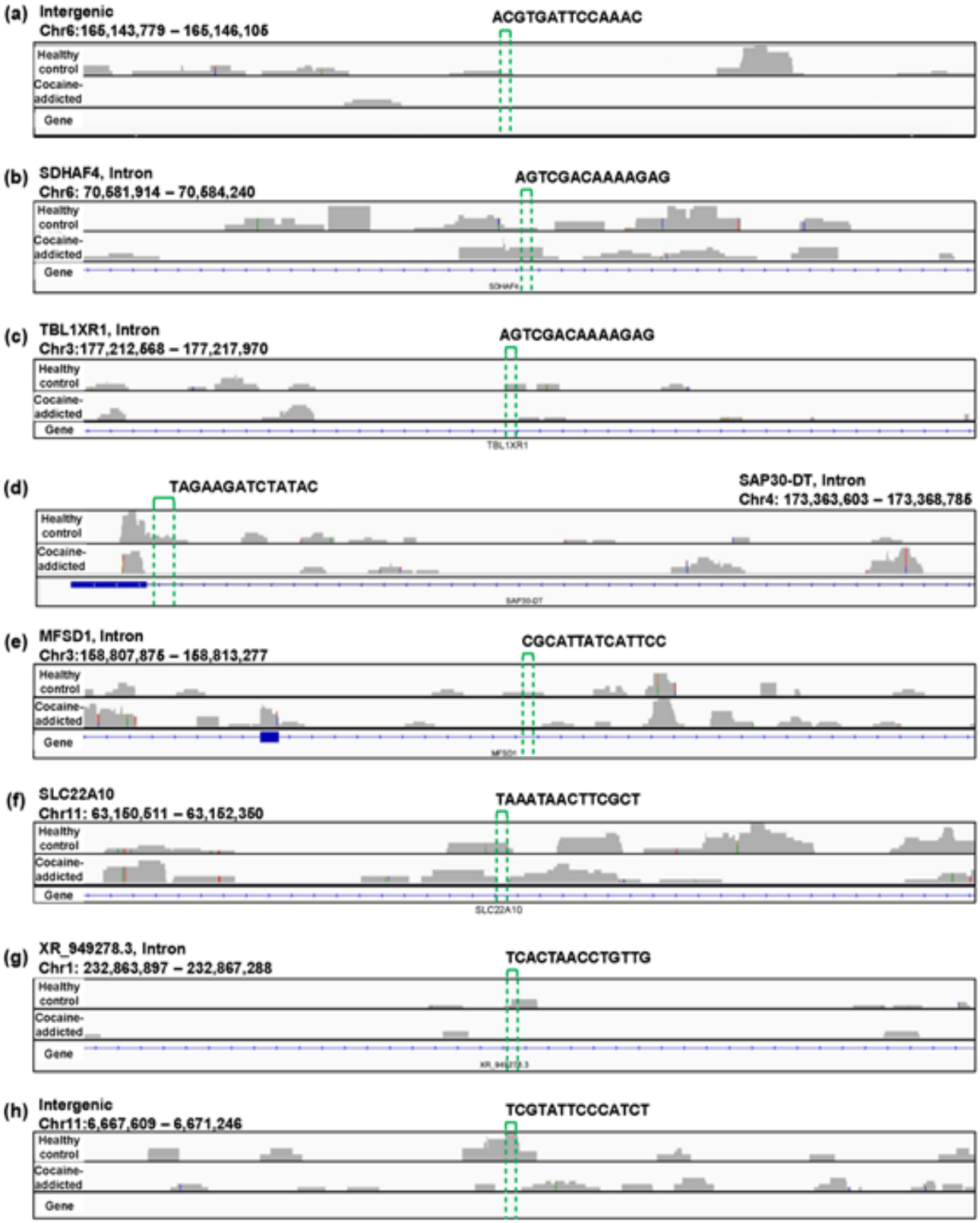


Figure S2: Example alignments near the region where the oligomers in Figure 5(c) are mapped. The positions of the oligomers are indicated by green dotted lines.

Shoji T and Sato Y., J Bioinform Syst Biol 2023
DOI:10.26502/jbsb.5107049

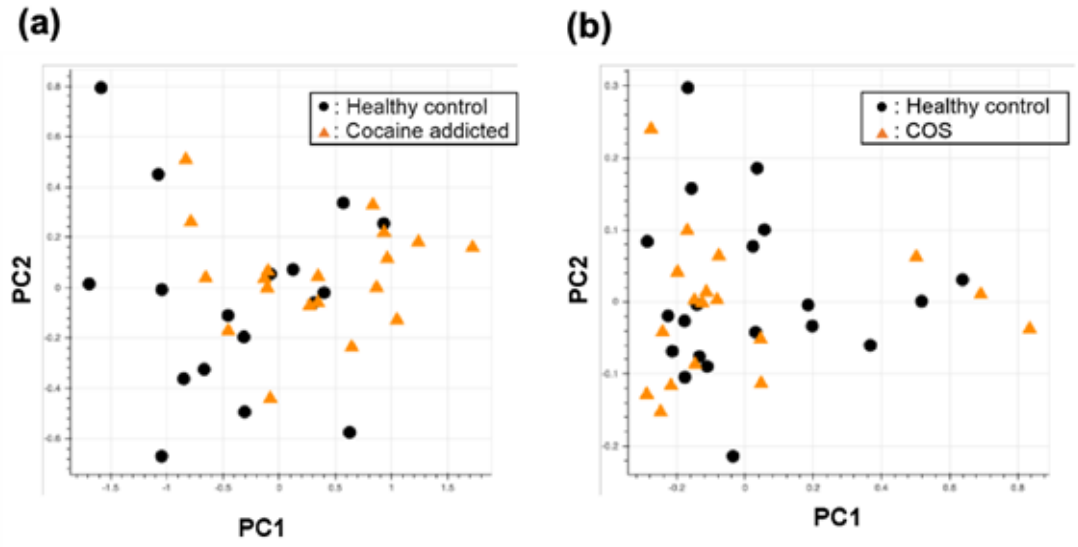


Figure S3: PCA of the gene expression table with the genes in Figure 5(c) and in Figure 6(c).

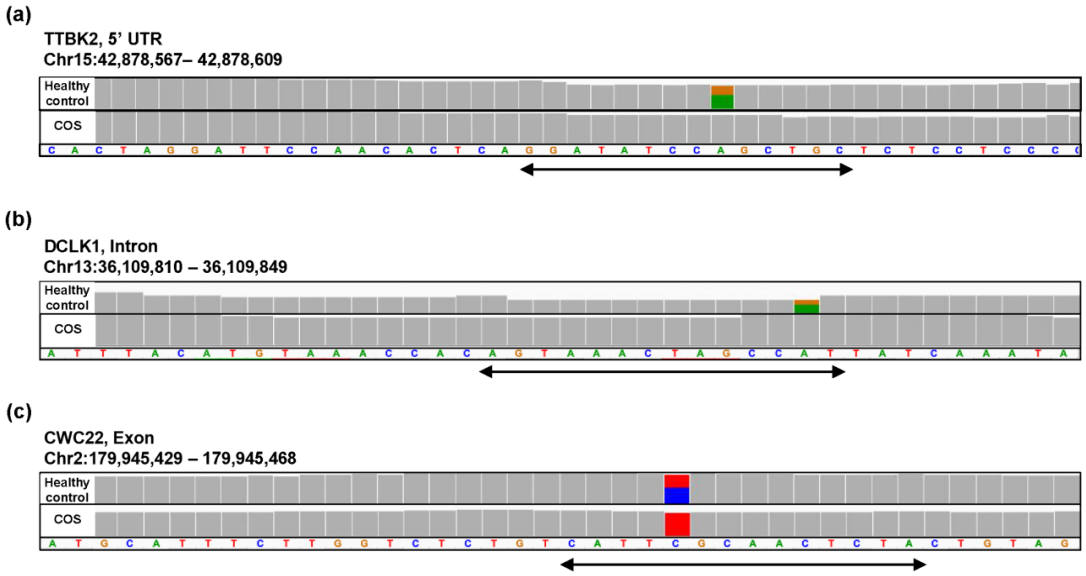


Figure S4: Example alignments near the region where the oligomers in Figure 6(c) are mapped. The position of the oligomers is indicated by the arrow.

from hiPSCs-NPCs derived from the healthy control group, 20 from total neurons derived from the same control group, 18 from hiPSCs-NPCs derived from 14 individuals with COS, 18 from total neurons derived from the same COS group, and 18 were excluded [28]. Out of 94 files, only the 38 hiPSCs-NPC fastq files were used in order to focus our analyses on a single cell type [28].

Count-based Matrices and Differential Gene Expression Analysis

The 36 or 38 fastq files from the cocaine addiction study or the COS study, respectively, were preprocessed by clipping Illumina adapter sequences using Trimmomatic v.0.39 [59] and were aligned to the human genome sequence (GenBank assembly accession: GCA_000001405.28) using HISAT2 v.2.2.1 [55]. Then, the read counts for each gene were calculated by featureCounts v.2.0.1 [60] using the human gene feature file (INSDC Assembly: GCA_000001405.28, Database version: 103.38) as a reference. Finally, the resulting 36 and 38 count data files were aligned with each other to obtain the count-based matrices. The sizes of the matrices were 36 × 20,022 and 38 × 20,022, respectively.

Shoji T and Sato Y., J Bioinform Syst Biol 2023
DOI:10.26502/jbsb.5107049

Differential gene expression was performed to obtain the FDR as shown in Figures 5(c) and 6(c), and the edgeR library [61] was applied to the count-based matrices.

q-mer Vectors and q-mer Matrices

First, 4q-dimension vectors were produced by counting the frequency of each 4q kind of q-length oligomer in each .sam file. Next, the count of each oligomer was normalized by the frequency of the oligomer in the transcriptome. Then, the data were further normalized so that the sum of the elements in each vector equaled 1. Finally, the resulting 4q dimension vectors were defined as the q-mer vectors. To obtain the q-mer matrices, the 36 or 38 q-mer vectors were aligned with each other. The sizes of the matrices were 36 × 4q and 38 × 4q. Reads that were not uniquely mapped or those that contained the character "N" were skipped. The code for producing q-mer vectors from .sam files is publicly available through GitHub: https://github.com/tatsumashoji/qmer.

Decomposition of Matrices

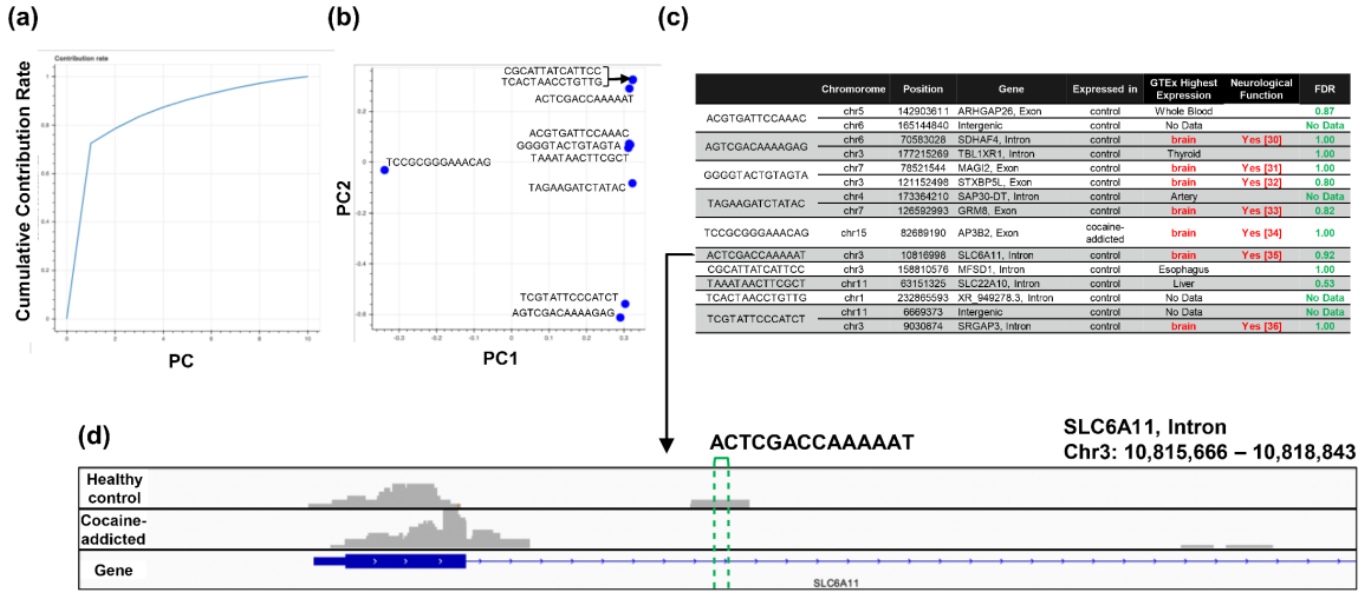
To decompose the matrices and plot the 36 or 38 cases onto a two-dimensional plane, PCA was implemented with "scikit-learn (0.24.1)" [62]. Briefly, 10 columns were selected from each matrix. These 10 columns had the top 10 highest correlation coefficients against the y vector, where the element is 0 if the corresponding case is the healthy control and 1 if otherwise. Then, the matrix with a size of 36 × 10 was decomposed using PCA. Finally, PC1 and PC2 for each case were plotted onto a two-dimensional plane. The library "scikit-learn (0.24.1)" was used for the detailed analysis of the PCA plots shown in Figure 5(a, b) and Figure 6(a, b). To reproduce our results, we have provided all code in a Jupyter notebook at https://github.com/tatsumashoji/qmer.

Identification of the Genome Position of the Oligomers

Not all of the cases in the cocaine-addicted group, the COS group, or the healthy control group expressed each oligomer. The genome position for each oligomer as shown in Figure 5(c) and Figure 6(c) was defined if the oligomer was expressed at the same genome position among more than four cases.

Discussion

In this study, we describe a new method, q-mer analysis, that includes alignment information in RNA-Seq data analysis. q-mer analysis focuses not on the expression of whole genes but on oligomers and produces vectors with higher dimensionality than do count-based methods to summarize alignment information. This "dimensionality increment" was critical for characterizing samples so that non-supervised subgrouping was successful when analyzing the RNA-Seq datasets with large q values (Figure 3(m)–(o), Figure 4(o)), while the count-based method failed to distinguish between subgroups (Figure 3(a), Figure 4(a); See Figure S1, Additional File 1).



	Chromorome	Position	Gene	Expressed in	GTEx Highest Expression	Neurological Function	FDR
ACGTGATTCCAAAC	chr5	142903611	ARHGAP26, Exon	control	Whole Blood		0.87
	chr6	165144840	Intergenic	control	No Data		No Data
AGTCGACAAAAGAG	chr6	70583028	SDHAF4, Intron	control	brain	Yes [30]	1.00
	chr3	177215269	TBL1XR1, Intron	control	Thyroid		1.00
GGGGTACTGTAGTA	chr7	78521544	MAGI2, Exon	control	brain	Yes [31]	1.00
	chr3	121152498	STXBP5L, Exon	control	brain	Yes [32]	0.80
TAGAAGATCTATAC	chr4	173364210	SAP30-DT, Intron	control	Artery		No Data
	chr7	126592993	GRM8, Exon	control	brain	Yes [33]	0.82
TCCGCGGGAAACAG	chr15	82689190	AP3B2, Exon	cocaine-addicted	brain	Yes [34]	1.00
ACTCGACCAAAAAT	chr3	10816998	SLC6A11, Intron	control	brain	Yes [35]	0.92
CGCATTATCATTCC	chr3	158810576	MFSD1, Intron	control	Esophagus		1.00
TAAATAACTTCGCT	chr11	63151325	SLC22A10, Intron	control	Liver		0.53
TCACTAACCTGTTG	chr1	232865593	XR_949278.3, Intron	control	No Data		No Data
TCGTATTCCCATCT	chr11	6669373	Intergenic	control	No Data		No Data
	chr3	9030874	SRGAP3, Intron	control	brain	Yes [36]	1.00

Figure 5: Detailed results of the q-mer analysis in the study of cocaine addiction. The cumulative contribution rate of PC1 to PC10 in the PCA of the q-mer matrix with q = 14 is shown in (a). The contribution rate of each oligomer for PC1 and PC2 is shown in (b). The position in the genome, the gene symbols for each oligomer described in the .sam files, and the false discovery rate (FDR) calculated by differential gene expression analysis based on the count-based matrix are shown in (c). Example alignments near the region "TAGAAGATCTATAC" are shown in (d). The position of the oligomer is indicated by green dotted lines.

Shoji T and Sato Y., J Bioinform Syst Biol 2023
DOI:10.26502/jbsb.5107049

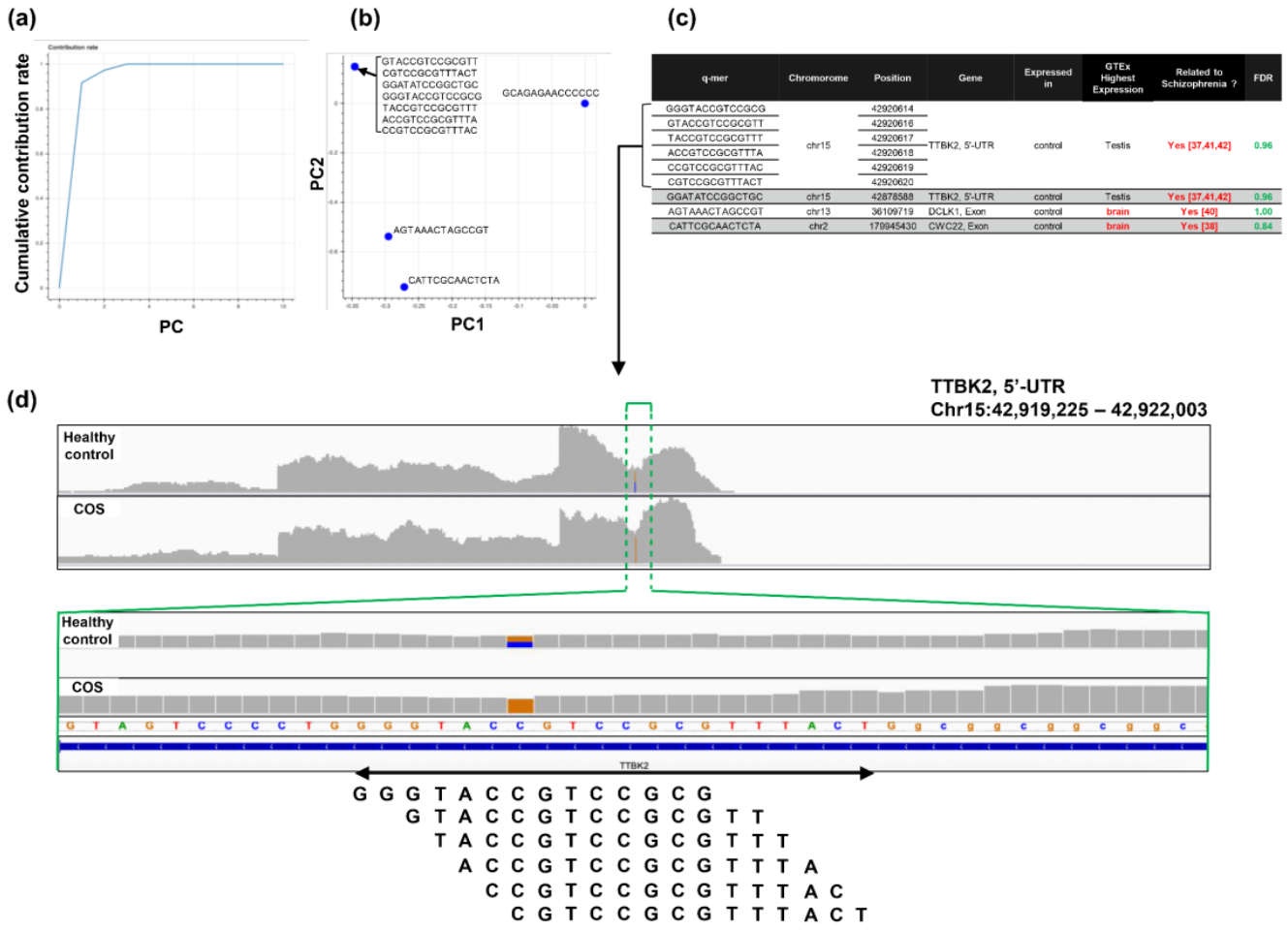


Figure 6: Detailed results of q-mer analysis in the COS study. The cumulative contribution rate of PC1 to PC10 in the PCA of the q-mer matrix with q = 14 is shown in (a). The contribution rate of each oligomer for PC1 and PC2 is shown in (b). The position in the genome and the gene symbols for each oligomer described in the .sam files and the FDR calculated by differential gene expression analysis based on the count-based matrix are shown in (c). Example alignments near the region where the six oligomers map are shown in (d). The position of the oligomers is indicated in green dotted lines.

The ability of q-mer analysis to accurately characterize transcriptomics samples could be helpful when identifying the biological mechanisms underlying disorders. Currently, the diagnosis of most psychiatric or neurological diseases, such as major depressive disorder, bipolar disorder, autism spectrum disorder, and attention-deficit hyperactivity disorder, depends on behavioral and symptomatic characterization. However, q-mer analysis could help to define these disorders using q-mer vectors, potentially revealing unique characteristics that could be used for diagnosis.

q-mer analysis cannot provide a clear description of the detailed biological mechanisms that explain the differences between samples and controls; however, q-mer analysis can offer hypotheses for further investigation by capturing differences in post-transcriptional regulation or expressed mutations. Through q-mer analysis, we detected differentially expressed candidate genes in the study of cocaine addiction (Figure 5(c)). These genes are not highlighted in the original study because of the high FDR. Since q-mer analysis indicated that the post-transcriptional regulation of these genes were different between cases and controls, subsequent studies should quantify the protein levels of the genes in Figure 5(c). Interestingly, most of the genes in Figure 5(c) were specifically expressed in the brain [29] and were functionally related to neurogenesis [30–36]. In the COS study, we identified mutations as discriminators between samples from COS cases and controls. Surprisingly, all of these mutations are novel [37–40]. Although many small nucleaotide polymorphisms related to COS have been reported [37–40], the mutations we identified are worth additional follow-up studies because the genes carrying these mutation were expressed and are related to COS [37-38] [40–42].

Importantly, since q-mer analysis does not require any additional experiments other than the RNA-Seq dataset,

Shoji T and Sato Y., J Bioinform Syst Biol 2023
DOI:10.26502/jbsb.5107049

scientists can re-mine existing RNA-Seq data and may find new results. For example, studies could immediately apply q-mer analysis to publically available RNA-Seq data [43–48]. However, the aligner, such as bowtie [49], bowtie2 [50], BWA [51-52], STAR [53], and HISAT2 [54-55], that is used must be considered; use of different aligner programs from study to study may produce different results. If the alignment data are different, then q-mer analysis results may differ as well.

Furthermore, RNA-Seq data derived from libraries constructed using the poly-A method should be avoided because this method only captures the 3′-end of each mRNA transcript. Thus, this method does not capture alignment information from the 5′-end and is thus not appropriate for q-mer analysis. In addition, currently, we do not have a valid statistical method to quantify the impact of q-mer analysis. In this study, we selected the top 10 oligomers that showed the highest correlation coefficients and observed linear separation of case and control samples using PCA. However, ideally, significant oligomers should be identified statistically. Investigations of the probability distribution of q-mer results remains to be addressed in future studies.

The dimension required to express alignment information of RNA-Seq data was estimated at least 49 (262,144; Table 1). Furthermore, to identify differences among the case samples and controls accurately, the required dimension may be 414 (268,435,456) or larger. Thus, the dimensionality of RNA-Seq data potentially has approximately 10,000 times the number of genes than are found 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 [56–58]. This contradiction may be because those studies only investigated ideal situations. In actual cases or conditions that have not yet been examined, such as in neurological or psychiatric disorders in H. sapiens, the transcriptome could have a large number of dimensions owing to the complexity of the brain and the etiology of these diseases.

The high dimensional nature of RNA-Seq data may be its advantage. Recently, scientists have attempted to describe sample conditions by combining multi-omics data to obtain additional explanatory variables. However, this approach is costly and hard to comprehend. Combining RNA-Seq experiments with q-mer analysis may be sufficient to describe samples because the dimensionality of RNA-Seq is much higher than that derived from multi-omics approaches. In the future, q-mer analysis may be the new standard rather than omics-related methods.

We suggest that there might be a limit for gene-level characterization in understanding complicated samples such as those derived from neurological and psychiatric disorders because networks and pathways often share similarities despite different disorders in each study [27]. The number of parameters (i.e., the number of genes) is probably not sufficient to explain these disorders. In contrast, q-mer analysis provides many parameters by focusing on oligomers and can separate samples if the number of samples and the q value are large enough. Further, q-mer analysis can identify candidate genes and biological mechanisms underlying differences between samples. Therefore, we propose using q-mer analysis to increase dimensionality, to identify novel mechanistic hypotheses based on differences in oligomers between different conditions, and to study underlying biology. In conclusion, differential transcriptomics based on q-mer analysis can provide novel data for clinical studies, diagnosis, prognosis, and identification of new genetic markers for diseases.

Conclusion

We introduced q-mer analysis, a generalized method for analyzing RNA-Seq data that includes alignment information and demonstrated that this alignment information was essential to characterize the samples appropriately. Aspects that q-mer analysis can more correctly represent than count-based approaches could be helpful for studies regarding gene expression. In the future, combining RNA-Seq with q-mer analysis could help identify biologically relevant features in transcriptomics data.

List of abbreviations

hiPSCs-NPCs: human-induced pluripotent stem cell-derived neural progenitor cells

COS: childhood-onset schizophrenia

PCA: principal component analysis

PC: principal component

FDRs: false discovery rates

Declarations

Ethics approval and consent to participate

The studies providing the example RNA-Seq in this study [27, 28] were conducted according to the guidelines and regulations established at Icahn School of Medicine at Mount Sinai for the use of human samples and were approved by the Institutional Review Board of Icahn School of Medicine at Mount Sinai. Informed consent was obtained from all subjects involved in the study by the authors of those studies [27, 28].

Consent for publication

Not applicable

Availability of data and materials

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE99349/GSE99349and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106589/GSE106589.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no external funding.

Authors' contributions

ST contributed to the conceptualization, developed the method and software, analyzed and interpreted the data, and was a major contributor in writing the manuscript. SY validated the manuscript and supervised the project. All authors read and approved the final manuscript.

Acknowledgments

Not applicable

References

1. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet 20 (2019): 631-56.
2. Emrich SJ, Barbazuk WB, Li L, et al. Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res 17(2007): 69-73.
1. Lister R, O'Malley RC, Tonti-Filippini J, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133(2008): 523-36.
2. Mortazavi A, Williams BA, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5(2008): 621-8.
3. Holt RA, Jones SJ. The new paradigm of flow cell sequencing. Genome Res 18(2008): 839-46.
4. Nagalakshmi U, Wang Z, Waern K, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320(2008): 1344-9.
5. Ingolia NT, Ghaemmaghami S, Newman JR, et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324(2009): 218-23.
6. Licatalosi DD, Mele A, Fak JJ, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 456 (2008): 464-9.
7. Yeo GW, Coufal NG, Liang TY, et al. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat Struct Mol Biol 16(2009): 130-7.
8. Koch CM, Chiu SF, Akbarpour M, et al. A beginner's guide to analysis of RNA sequencing data. Am J Respir Cell Mol Biol 59(2018): 145-57.
9. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. Genome Biol 17(2016): 13.
10. Kukurba KR, Montgomery SB. RNA sequencing and analysis. Cold Spring Harb Protoc 2015 (2015): 951-69.
11. Zhang Z, Hernandez K, Savage J, et al. Uniform genomic data analysis in the NCI genomic data commons. Nat Commun 12(2021): 1226.
12. Thul PJ, Lindskog C. The human protein atlas: a spatial map of the human proteome. Protein Sci 27(2018): 233-44.
13. Uhlen M, Karlsson MJ, Zhong W, et al. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. Science 366(2019): 9198.
14. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. GT. Science 369(2020): 1318-30.
15. Chhangawala S, Rudy G, Mason CE, et al. The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome Biol 16(2015): 131.
16. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11(2010): R25.
17. Xiao Z, Zou Q, Liu Y, et al. Genome-wide assessment of differential translations with ribosome profiling da-ta. Nat Commun 7(2016): 11194.
18. Risso D, Perraudeau F, Gribkova S, et al. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun 9(2018): 284.
19. Vargo AHS, Gilbert AC. A rank-based marker selection method for high throughput scRNA-seq data. BMC Bioinform-ICs 21(2020): 477.
20. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(2010): 139-40.
21. Varet H, Brillet-Guéguen L, Coppée JY, et al. SARTools: a DESeq2- and EdgeR-based R pipeline for comprehensive differential analysis of RNA-Seq data. PLOS ONE 11(2016): e0157022.

22. Shoji T, Takaya A, Kusuya Y, et al. Ribosome profiling in Streptococcus pneumoniae reveals the role of methylation of 23S rRNA nucleotide G748 on ribosome stalling. J Genet Genom Sci 6(2021): 024.

23. Shoji T. 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 Genom Sci 6(2021): 026.

24. Wenzel MA, Müller B, Pettitt J. SLIDR and SLOPPR: flexible identification of spliced leader trans-splicing and prediction of eukaryotic operons from RNA-Seq data. BMC Bioinformatics 22(2021): q140.

25. Ribeiro EA, Scarpa JR, Garamszegi SP, et al. Gene network dysregulation in dorsolateral prefrontal cortex neurons of humans with cocaine use disorder. Sci Rep 7(2017): 1-10.

26. Hoffman GE, Hartley BJ, Flaherty E, et al. Transcriptional signatures of schizophrenia in hiPSC-derived NPCs and neurons are concordant with post-mortem adult brains. Nat Commun 8(2017): 2225.

27. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. GT. Science 369(2020): 1318-30.

28. Van Vranken JG, Bricker DK, Dephoure N, et al. SDHAF4 promotes mitochondrial succinate dehydrogenase activity and prevents neurodegeneration. Cell Metab 20(2014): 241-52.

29. Koide T, Banno M, Aleksic B, et al. Common variants in MAGI2 gene are associated with increased risk for cognitive impairment in schizophrenic patients. PLOS ONE 7(2012): e36836.

30. Kumar R, Corbett MA, Smith NJ, et al. Homozygous mutation of STXBP5L explains an autosomal recessive infantile-onset neurodegenerative disorder. Hum Mol Genet 24(2015): 2000-10.

31. Li W, Ju K, Li Z, et al. Significant association of GRM7 and GRM8 genes with schizophrenia and major depressive disorder in the Han Chinese population. Eur Neuropsychopharmacol 26(2016): 136-46.

32. Assoum M, Philippe C, Isidor B, et al. Autosomal-recessive mutations in AP3B2, adaptor-related protein complex 3 beta 2 subunit, cause an early-onset epileptic encephalopathy with optic atrophy. Am J Hum Genet. 99(2016): 1368-76.

33. Dikow N, Maas B, Karch S, et al. 3p25. 3 microdeletion of GABA transporters SLC6A1 and SLC6A11 results in intellectual disability, epilepsy and stereotypic behavior. Am J Med Genet A 164 (2014): 3061-8.

34. Lu H, Jiao Q, Wang Y, et al. The mental retardation-associated protein srGAP3 regulates survival, proliferation, and differentiation of rat embryonic neural stem/progenitor cells. Stem Cells Dev 22(2013): 1709-16.

35. Ambalavanan A, Girard SL, Ahn K, et al. De novo variants in sporadic cases of childhood onset schizophrenia. Eur J Hum Genet 24(2016): 944-8.

36. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. Nat Genet 43(2011): 969-76.

37. Schizophrenia Working Group of the Psychiatric Genomics. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511(2014): 421-7.

38. Håvik B, Degenhardt FA, Johansson S, et al. DCLK1 variants are associated across schizophrenia and attention deficit/hyperactivity disorder. PLOS ONE 7(2012): e35424.

39. Bowie E, Goetz SC. TTBK2 and primary cilia are essential for the connectivity and survival of cerebellar Purkinje neurons. eLife 9(2020): e51166.

40. Muñoz-Estrada J, Lora-Castellanos A, Meza I, et al. Primary cilia formation is diminished in schizophrenia and bipolar disorder: a possible marker for these psychiatric diseases. Schizophr Res 195(2018): 412-20.

41. Verma S, Du P, Nakanjako D, et al. Tuberculosis in advanced HIV infection is associated with increased expression of IFN-γ and its downstream targets. BMC Infect Dis 18(2018): 1-13.

42. McCaffrey TA, St Laurent G, 3rd, Shtokalo D, et al. Biomarker discovery in attention deficit hyperactivity disorder: RNA sequencing of whole blood in discordant twin and case-controlled cohorts. BMC Med Genomics 13 (2020): 160.

43. Tiedt S, Prestel M, Malik R, et al. RNA-Seq identifies circulating miR-125a-5p, miR-125b-5p, and miR-143-3p as potential biomarkers for acute ischemic stroke. Circ Res 121(2017): 970-80.

44. Liu R, Holik AZ, Su S, et al. Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. Nucleic Acids Res 43(2015): e97.

45. Voineagu I, Wang X, Johnston P, et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature 474(2011): 380-4.

46. Sorokina AM, Saul M, Goncalves TM, et al. Striatal transcriptome of a mouse model of ADHD reveals a pattern of synaptic remodeling. PLOS ONE 13(2018): e0201553.

Shoji T and Sato Y., J Bioinform Syst Biol 2023
DOI:10.26502/jbsb.5107049

47. Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(2009): R25.

48. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 9(2012): 357-9.

49. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 25(2009): 1754-60.

50. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 26(2010): 589-95.

51. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(2013):15-21.

52. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12 (2015): 357-60.

53. Kim D, Paggi JM, Park C, et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37(2019): 907-15.

54. Heimberg G, Bhatnagar R, El-Samad H, et al. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. Cell Syst 2(2016): 239-50.

55. Biswas S, Kerner K, Teixeira P, et al. Tradict enables accurate prediction of eukaryotic transcriptional states from 100 marker genes. Nat Commun 8(2017): 1.

56. Kobayashi-Kirschvink KJ, Nakaoka H, Oda A, et al. Linear regression links transcriptomic data and cellular Raman spectra. Cell Syst 7(2018): 104-117.

57. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible read trimming tool for Illumina NGS data. Bioinformatics 30(2014): 2114-20.

58. Liao Y, Smyth GK, Shi W. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res 41(2013): e108.

59. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 (2010): 139-40.

60. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res 12(2011): 2825-30.