SCIENTIFIC REPORTS
nature research

OPEN

# Principal Component Analysis applied directly to Sequence Matrix

Tomokazu Konishi[1*], Shiori Matsukuma[2], Hayami Fuji[1], Daiki Nakamura[1], Nozomi Satou[1] & Kunihiro Okano[1]

**Sequence data is now widely used to observe relationships among organisms. However, understanding structure of the qualitative data is challenging. Conventionally, the relationships are analysed using a dendrogram that estimates a tree shape. This approach has difficulty in verifying the appropriateness of the tree shape; rather, horizontal gene transfers and mating can make the shape of the relationship as networks. As a connection-free approach, principal component analysis (PCA) is used to summarize the distance matrix, which records distances between each combination of samples. However, this approach is limited regarding the treatment of information of sequence motifs; distances caused by different motifs are mixed up. This hides clues to figure out how the samples are different. As any bases may change independently, a sequence is multivariate data essentially. Hence, differences among samples and bases that contribute to the difference should be observed coincidentally. To archive this, the sequence matrix is transferred to boolean vector and directly analysed by using PCA. The effects are confirmed in diversity of Asiatic lion and human as well as environmental DNA. Resolution of samples and robustness of calculation is improved. Relationship of a direction of difference and causative nucleotides has become obvious at a glance.**

Nucleotide sequence could be desired information for classification of organisms such as performed in phylogenetics, as genetic information is highly specific to individuals, easy to obtain with accuracy, and may reflect biological characters of samples. Indeed, amplifying specific fragments of DNA and obtaining nucleotide sequences have become an ubiquitous tool for this purpose[1,2]. However, there is no simple solution to estimate relationships among the sequences. As they are qualitative data, numerical conversion is required for any calculation for estimating the relationships in a quantitative way. Additionally, a nucleotide sequence is a multivariate data with huge number of independent items that are recorded as form of bases. Therefore, differences among samples consist of multiple dimensions, and are difficult to be understood. In this sense, estimation of sample relationships is a question in multivariate analysis in essence.

Conventionally, relationships among nucleotide sequences are summarized using a dendrogram. Two classes of approaches are available to estimate the dendrogram (Fig. 1): 1. targets the distance matrix, which records distances among samples, and both the estimation of the distances and the manner in which they are summarized to make the dendrogram are based on specific mathematical models, and 2. targets directly each of the sites (a set of nucleotides or amino acids in a same position of aligned sequences) and calculates the score for a certain tree model; the estimation of the score is based on a given model that assumes several parameters (Fig. 1, maximum parsimony, likelihood and Bayesian). Via heavy computational calculations, optimized values are found for the parameters[2,3].

Despite the efforts to find better dendrograms, the results essentially lack objectivity. Although the mathematical models are designed taking evolution into consideration, they are based on many assumptions that could never be verified by evidence. In fact, most individuals that participate in the process of evolution have been lost, and evidence of periods of time are only rarely available. The topology of a tree, i.e., which combination of samples should be connected, belongs to models, and not to parameters; such assumptions of the tree shape can generate errors, referred to as *long-branch attraction*[3], that are difficult to identify. Moreover, knowledge obtained from such dendrograms cannot be integrated beyond different models. Accordingly, the resulting dendrograms are based on certain ideas that are not falsifiable; in this sense, their standpoint does not fall within conventional science, as is the case of superstring theory[4].

[1]Department of Biological Environment, Faculty of Biresource Sciences, Akita Prefectural University, Shimoshinjyo Nakano, Akita, 010-0195, Japan. [2]Biodiversity Promotion Division, Tokyo Branch, Regional Environmental Planning Inc. NDS Building, 2-22-3 Sakurashinmachi Setagaya-ku, Tokyo, 154-0015, Japan. *email: konishi@akita-pu.ac.jp

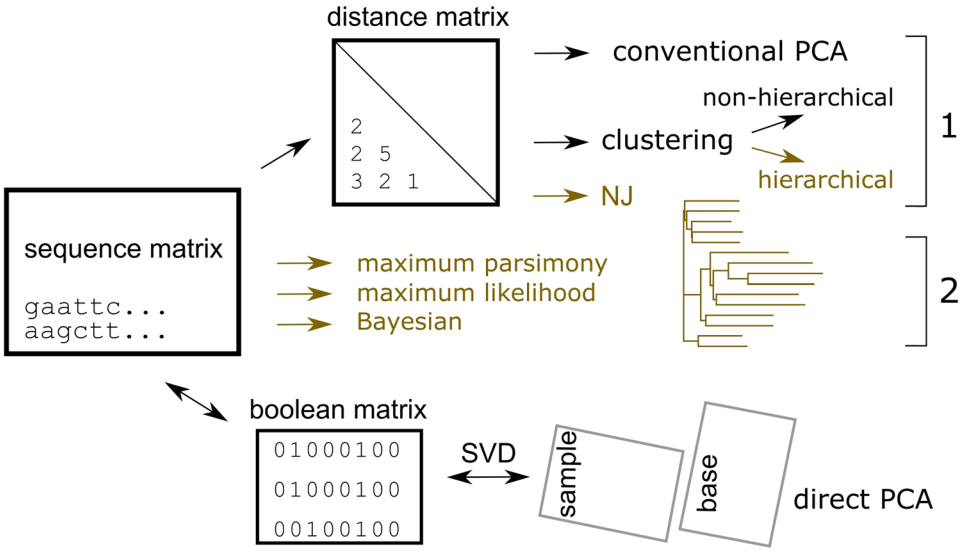



**Figure 1.** Schematic diagram of the methods used in the study. Two distinct approaches are used: one estimates the distances among samples, to form a distance matrix using one of various models (Yang, 2006), and then finds the relationships based on the distances. The other finds the relationship directly from the sequence matrix. Only direct PCA shows the relationships of samples and bases coincidently. Methods that estimate a tree-shaped relationship are coloured in oak.

Mathematical models often assume several conditions. For example, those that produce a dendrogram assume that all the subjected samples share a tree-shaped relationship (in Fig. 1, Methods, coloured in oak). In general, such universal propositions can be disproved by showing an opposing example. Below is a simple example of the tree-shape assumption.

In speciation, it was long believed that accumulating mutations generate new species; this makes a linkage between the new and ancestral species, forming a branch of a phylogenic tree. However, whole genome sequencings of prokaryotic species have revealed that evolution is not limited to such hierarchical way, but genes may be horizontally transferred among species[5]. Indeed, such transfer may occur among eukaryotes[6]; additionally, remixing of genetical information often occurs by sexual mating. Such transfer connects distant branches, and hence the relationships among organisms would be a complex network, not a tree. As Neighbor Joining and other clustering method summarize the relationships to estimate phylogenic tree, if the real structure was a network, the resulted estimation will be distorted. Any scientific research should not depend on assumptions that cannot be verified, since such assumptions critically damage the objectivity of any conclusions.

In the classification of samples, other non-hierarchical methods have been used, in addition to the tree-based approach (Fig. 1). A typical example is k-means; for the given number of clusters, this absorbs neighbour samples of randomly specified starting samples. The self-organizing map is another example[7]; this is an artificial neural network that organizes samples with low-dimensional nodes, which is preferentially designated by the analyst. The two approaches basically do not indicate the distances among samples. Moreover, the results are heavily dependent on the given numbers or dimensions, and reproducibility of the results cannot be expected.

Scientific data analyses must fulfil several specific requirements, such as objectivity and reproducibility[8]. Science is an attempt at integrating the contributions of many researchers. Knowledge without objectivity or reproducibility ruins the integration process. The approaches shown above are not objective, because their assumptions are never falsified by the real data. Moreover, they are not reproducible, because many arbitrary options can drastically change the results. Such a choice includes the condition of alignment and the definition of distances; the effects of the former will be demonstrated in a subsequent section. In science, reproducibility is necessary to guarantee the objectivity of the results; i.e., they do not depend on the experimenter. Of course, it is erroneous to keep using inappropriate models, because they would lead to the wrong conclusions.

It is not easy to estimate a phylogenic network, and an appropriate answer would not be provided by an automated method. Short DNA sequences would be insufficient to estimate relationships among organisms; the result may include reading errors, and not falsifiable; it is sensitive to serious errors provoked by horizontal DNA transfer. Estimations using several different sequences may yield conflicting results. Therefore, we need to be aware of the limitations of the existing information. However, as the data structure could be complex, i.e., the base sequence may have the same dimension as its length, a mathematical aid for observing the structure of data is required. A method that can observe the sequence matrix data without harming its structure is preferred; if we can preserve the structure, we would be able to integrate the results to estimate the network structure when sufficient information becomes available.

To maintain objectivity, we should not connect samples based on unverifiable assumptions; rather, they should be kept pendulous. Analysing the distance matrix using Principal Component Analysis (PCA) would satisfy this criterion because it does not assume a specific structure of data (Fig. 1, conventional PCA). Rather, it rotates

the matrix and projects it to sets of diagonal axes; it finds directions of differences and presents them by using the axes. The relationships are presented within the found axes with a simpler vision than the original matrix. Both the reproducibility and objectivity of the calculations are satisfactory. This method seems to be used rather spontaneously and spread, as PCA is a preferred method for analysing matrix data such as the distance matrix[9].

However, as this method targets the distance matrix, all information for the sites is lost and is not included in the results; i.e., which site contributes to which difference and how it occurs. This is a serious disadvantage in the estimation of the role of sequence motifs or finer relationships between samples. As each sample of sequence data are made of multiple bases, they can be recognized as a point located in multiple dimensions of the same number of bases. Difference between a pair of samples defines a vector in the dimensions, which consists of both length and direction. Direction could be important as well as distance, therefore, it is better to analyse information of bases and samples coincidently.

To archive this, sequence data have to be transferred to numeral. Here they are replaced as boolean vector that is presented by 0 and 1. For example, a base of DNA is recorded by a set of five digits (A, T, G, C, and -); this transformation of recording format is known as 'one-hot encoding of information' in computer architecture[10]. This transformation has merits that no information is lost and so it is completely reversible; also, it is applicable to both nucleotide and amino acid sequences. Differences between two samples are defined by subtraction, distance is found by using Euclidean length, and average sequence is found as the arithmetic mean of samples. Some of the options for adjustments for weights, parallel, convergent, or back substitutions may be still possible, as will be discussed later. Then, the quantified sequence matrix is rotated using singular value decomposition (SVD), to identify the principal components (PC) (Fig. 1, direct PCA, mathematical explanation is in Materials and Methods).

The transformed sequence matrix is further analyses by PCA[11]. Basing on Euclidean metric space, it is compatible with randomly occurring mutations or a Brownian model. From this point of view, it is completely different from several methods proposed for phylogenetics, such as size-correction or phylogenic PCA, both would be recognized as derivatives of the conventional one. Those methods were developed to fit non-Brownian evolution; however, they had to assume many conditions that are never verified through analysis. Hence, they are outside the scope of this article, which introduces a scientific method. The method is different from any of other PCA applications shown in Fig. 1[9,12]. In direct PCA, both PC for samples and bases are calculated, whereas only the former is given in the conventional PCA[11]. The latter are loadings given the full length; they indicate which nucleotide base (or amino acid residue) contributes to the scores of the former, and vice versa. By normalizing for the size of the matrix, they show absolute values that can be compared among studies. Here, the scaled PCs are termed sPCs and their magnitudes are comparable among experiments. According to the nature of the PCA, the structure of the data is maintained, whereas other methods distort the structure by applying certain estimations. Moreover, the calculations are light because they do not require comparisons among models. They are reversible and no information will be lost.

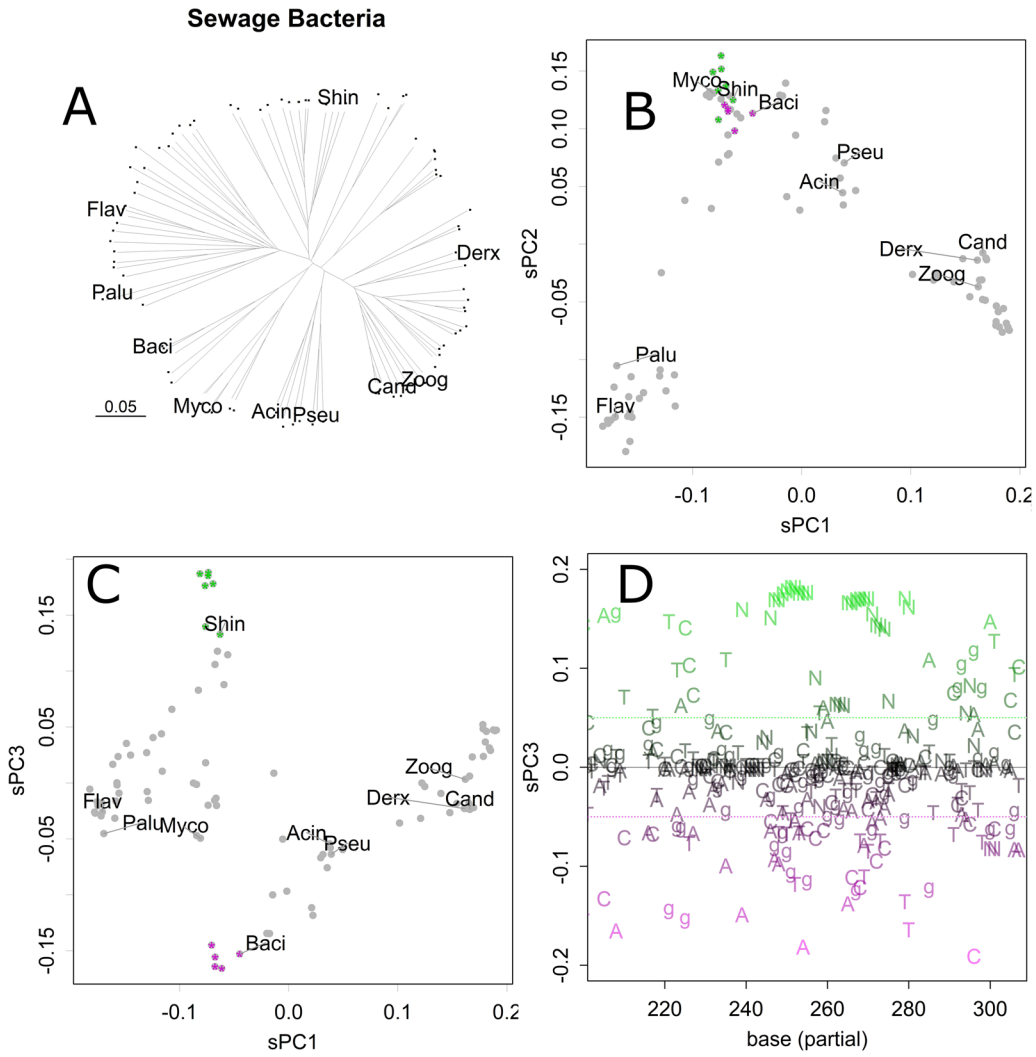
Here, the method was applied to observe the divergence in major histocompatibility complex (MHC) genes and to classify environmental DNA. Each axis of PCA presented differences of samples and bases coincidentally: which base or motif was characteristic to which group of samples and how. This helped understanding differences among multiple samples. The directivity of each axis improved robustness to outlying bases, which appear rarely among samples, and improved resolution of samples when applied to phylogenetic purposes.

## Results

**Structure and presentation of data.** First, let us examine an example that may indicate a characteristic of the network structure. In Fig. 2A, relationships among 94 samples of 16S ribosomal DNA taken from sewage bacteria in active sludge were estimated with some reference species by using distance matrix and NJ method. This method finds the tree structure from distance matrix of the sequences, estimating minimum evolution to connect the whole samples. The resulted tree shows several clusters connected rather complex ways, and some of them are apart from reference species. Of course, the contributions of sequence motifs do not appear in the presentation.

In contrast, each axis of PCA (Fig. 1 and Materials and Methods) would regard specific connections, presenting the structure easier to understand and fewer distortions. The 1st and 2nd axes of PCA separated samples into four major clusters, each of which is related to the references (Fig. 2B), finding bases common to the clusters. The samples were further separated or connected in the lower axes that regarded to other connections of the network structure of species. For example, a cluster found in sPC1 and sPC2 (Fig. 2B, bottom centre) was further separated into three groups that include *Bacillus*, *Mycobacterium*, and *Shinella*, in sPC3 (Figs. 2C, 3D version is available in Supplementary Information Fig. S2_BC). Hence, *Bacillus* and *Shinella* were similar in part, whereas they were different in other parts; they were gathered and separated accordingly in the corresponding axes of sPC for samples. The clusters of samples had different bases at specific positions, dispersed to entire range of the sequence, as obvious in sPC for bases (Fig. 2D,E); bases that presented higher magnitude had more contributions to separate samples.

The relationship of sPC for samples and bases appears in Fig. 2C (*y*-axis) and 2D, both of which present the PC3 axis. In Fig. 2D, positive bases (green) contributed to pulling up *Shinella* in Fig. 2C, and negative bases (magenta) contributed to pulling down *Bacillus*; i.e., the green bases tended to appear in *Shinella* samples, while the magenta bases tended to appear in *Bacillus* samples. Bases that had small contributions are shown in grey, and those that had no contribution were omitted. Here, Fig. 2D is presented to provide a view of the whole; to specify the bases, scores should be checked on tables. Panel E presents a part of summarized table output, showing sPC3 positive and negative bases and some samples including *Shinella* and *Bacillus* sequences (indicated in panels B and C as green and magenta asterisks). Although they are similar as they showed likely sPC1 and sPC2 values, they differed in several bases in sPC3, especially in the sequential indels. As an example of the table output, data are supplied as Supporting Information (Table S1 for bacteria and Table S2 for lion samples).




E

```
base            210       220       230       240       250       260       270       280       290       300
PC3 positive    A--G-T-------T---T-TACCC-------T---N------NNNNNNNNNN-NTA--NNNNNNNNNNNNNN----NN----A------CGGNNG-G-AT---C
Shinella        AGGGTTGTAAAGCTCTTTCACCG---GTGAAGA----------------------TAAT----------------------GACGGTAACCGG---AGAAGAAGC
LC279136        AGGGTTGTAAAGCTCTTTCGCCC---GCGAAGA----------------------TAAT----------------------GACTGT-AGCGG---GAGAAGAAGC
LC279138        AGGGTCGTAAAGCTCTTTTACCA---GGGATGA----------------------TAAT----------------------GACAGT-ACCTG---GAGAATAAGC
LC279149        AGGGTTGTAAAGCTCTTTTACCC---GGGATGATA-----------------A----------------------T-GACAGT-ACCGG---GAGAATAAGC
LC279157        AGGGTCGTAAAGCTCTTTTACCA---GGGATGA----------------------TAAT----------------------GACAGT-ACCTG---GAGAATAAGC
LC279162        AGGGTTGTAAAGCTCTTTTACCA---GGGATGA----------------------TAAT----------------------GACAGT-ACCTG---GAGAATAAGC
LC279163        AGGGTTGTAAAGCTCTTTTACCC---GAGATGA----------------------TAAT----------------------GACAGT-ATCGG---GAGAATAAGC
LC279166        AGGGTCGTAAAGCTCTTTTACCA---GGGATGA----------------------TAAT----------------------GACAGT-ACCTG---GAGAATAAGC
LC279092        CGGGTCGTAAAGCTCTTTTACCA---GGGATGA----------------------TAAT----------------------GACAGT-ACCTG---GAGAATAAGC

PC3 negative    C--A-C------AC---G-GGGTA--------A---A------CGTGGGTTAC-C-CG--G-ACGCTTGCTC-----GT-----G------T--ACC-T-NN---G
Bacillus        CGGGTCGTAAAACTCTGTTGTTA---GGGAAGA-A-------CAAGTGCTAGTTGAAT---AAGCTGGCACC----TT-GACGGT-ACCTAACCAGA--AAGC
LC279137        CGGATCGTAAAACTCTGTTATCA---GGGAAGA-A-----CAAATGTGTAAG--TAACTGTGCACAT---------CTTGACGGT-ACCTGATCAGA--AAGC
LC279146        CGGATCGTAAAACTCTGTTGGTA---GAGAAGA-A-------CGTTGGTGAG---AGTGGAAAGCTCATCAA---GT-GACGGTAACTAC-CCAGA--AAGG
LC279156        CGGATCGTAAAACTCTGTTATTA---GGGAAGA-A-----CAAATGTGTAAG--TAACTATGCACGTC--------TT-GACGGT-ACCTAATCAGA--AAGC
LC279164        CGGATCGTAAAACTCTGTTGGTA---GAGAAGA-A-------CGTTGGTGAG---AGTGGAAAGCTCATCAA---GT-GACGGTAACTAC-CCAGA--AAGG
LC279118        CGGATCGTAAAGCTCTGTTGTAA---GTCAAGA-A-------CGAGTGTGAG---AGTGGAAAGTTCACACT---GT-GACGGT-AGCTTACCAGA--AAGG
```

**Figure 2.** Analysis of sewage bacteria. Neighbor Joining tree. Each dot presents a sample. All the samples are connected to the tree by a single line. Abbreviations are the first four letters of reference samples for *Acinetobacter indicus, Bacillus anthracis, Candidatus Accumulibacter phosphatis, Derxia gummosa, Flavobacterium pectinovorum, Mycobacterium sp. strain FED_F02A, Paludibacter propionicigenes, Pseudomonas punonensis, Shinella kummerowiae*, and *Zoogloea resiniphila*, respectively (**A**). Scatter plot of scaled PC, sPC1 and sPC2 for samples (**B**). Scatter plot for sPC1 and sPC3 (**C**). A 3D version is available in Supplementary information, Fig. 2_BC.html. sPC3 for nucleotides (partial) (**D**). Positive and negative bases are coloured green and magenta, respectively; those had low contributions are coloured grey, and none contribution are omitted. *Shinella*, which appeared as a positive in the same sPC3 axis (**C**), tended to contain bases indicated in green, whereas *Bacillus* tended to contain bases indicated in magenta. Such tendencies are obvious in a summary of table output of $sPC_n$ (**E**). Bases positive ($>0.05$) and negative ($<-0.05$) in sPC3 are shown with sequences of several samples that are indicated in panels **B** and **C** by green or magenta asterisks.

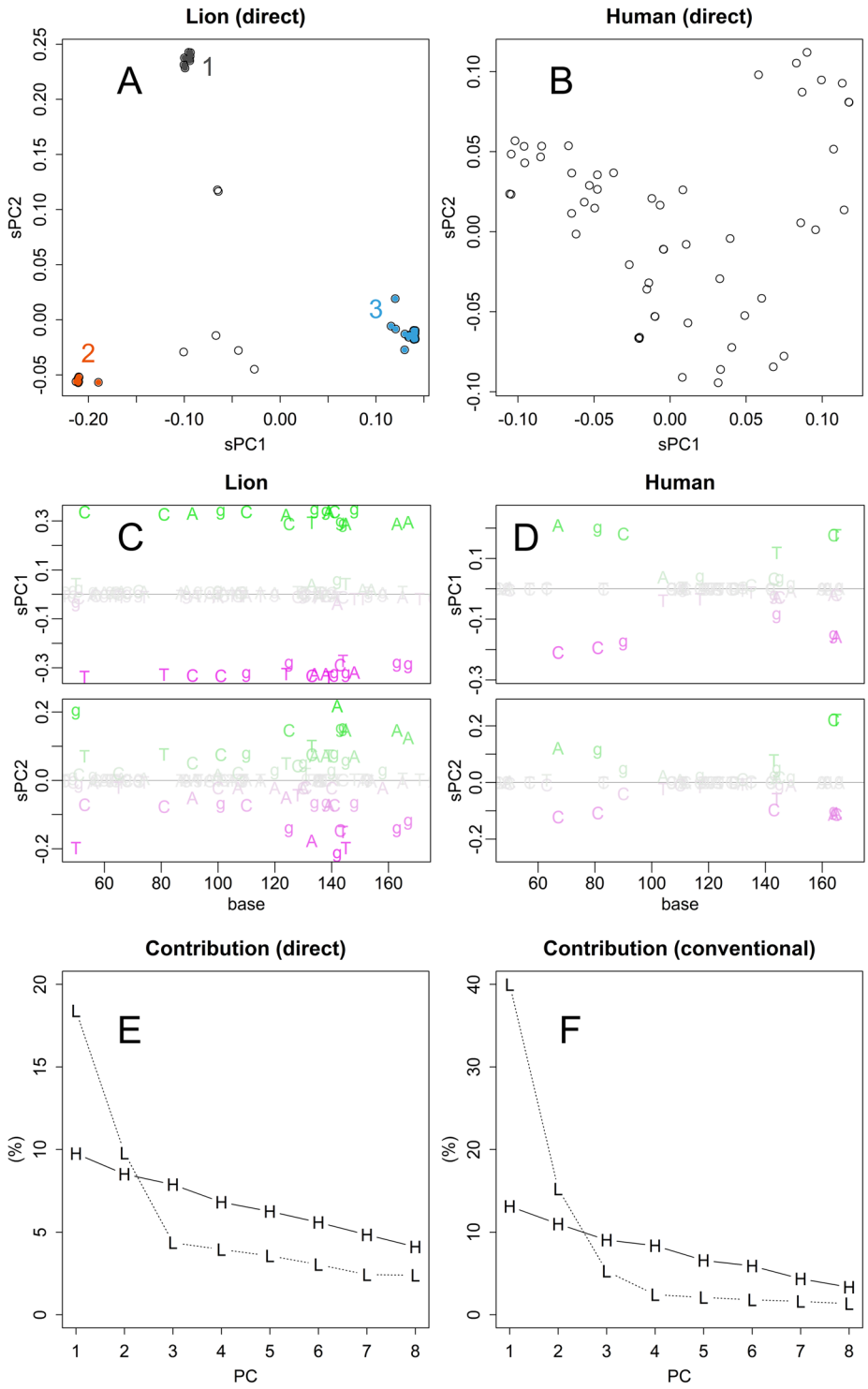



**Figure 3.** Direct PCA on MHC samples. Scatter plot of sPC1 and sPC2 for samples of Asiatic lions (**A**). PCA was performed directory to the Nucleotide sequence matrix. Scatter plot for samples of human data (**B**). sPC1 and sPC2 for nucleotide bases of lion data (**C**). PCs for bases of human data (**D**). Contributions (%) for sPC1 to sPC8 of PCA applied to sequence matrix directory (**E**) and distance matrix (**F**), respectively.

**Three distinct clusters found in Asiatic lions.** Next, let us examine another example of the grouping structure of an engendered species. Nucleotide sequences for major histocompatibility protein (MHC) of Asiatic lion[13], and corresponding sequence of human samples were directly analysed by PCA (Fig. 3). As the figures depict the scaled PC, they present the magnitude of differences in a uniform scale among the human and lion data (Fig. 3A,B), although it should be noted that the contributions of the PCs should be different. Among the lion's samples, there were three clusters distinctively separated (Fig. 3A). Although differences among the three

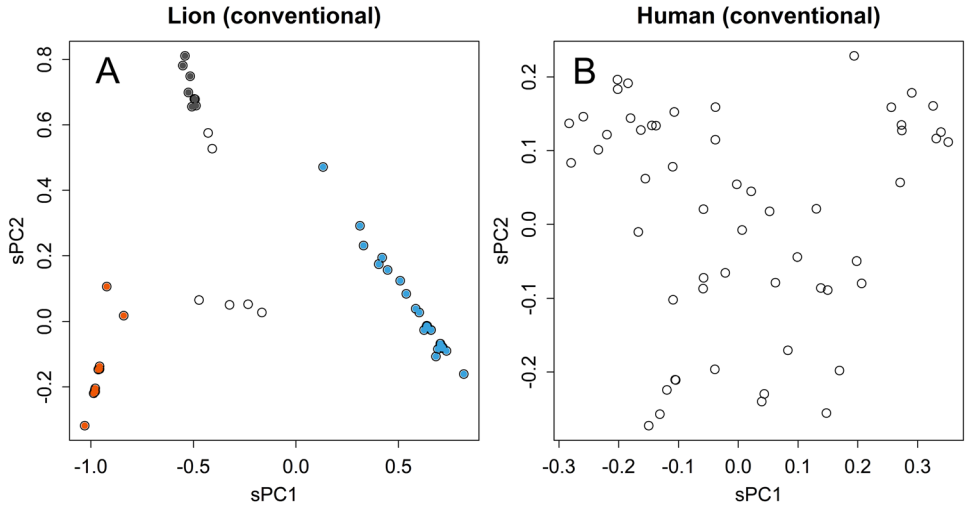



**Figure 4.** PCA by using distance matrix. Scatter plot of conventional PCA that applied to distance matrix of lion (**A**) and human (**B**) samples.

clusters were comparable to variations among human samples (Fig. 3B), differences within the clusters were quite small, and this is contrastive to the well dispersed dots, although a small cluster is separated at the upper right position in human samples. Incidentally, the standard deviation in the distances among the lion groups 1–3 was 0.014, 0.006, and 0.006, respectively, while that of the total lion samples was 0.091, which is comparable to that of human, 0.114; those values indicate the mean distances of samples from the center of groups.

The different grouping structure between lion and human also appeared in sPC for bases. sPC1 of lion took extreme and even values (Fig. 3C). This shows that each of the tree groups have strong propensity to specific bases at defined position. sPC1 for lion presents the difference among the groups, and it contributed 20% of the total differences (Fig. 3E). On the other hand, bases took various values both in sPC1 and 2, and the number of bases presented is much less in human samples (Fig. 3D). Bases differs along larger number of underlying directions; hence contribution of sPC1 and 2 were fewer than those of lion and decreased gently (Fig. 3E).

Bases that were common to all the samples could be identified, as they had null PC scores in all the axes. Moreover, bases that were specific to a group appeared in characteristic axes. For example, the group A in lion samples appeared as bases that were positive in sPC2. Therefore, those bases were automatically selected by specifying the threshold; quantile–quantile plots may suggest the value of the threshold (Fig. S1). These analyses can be performed on a spreadsheet software; however, they can also be performed systematically on R (scripts are in the Supporting Information S1 Doc). As an example of an output, aligned lion sequences are presented with the consensus and characteristic bases in the FASTA format (Table S6 and trial kit).

**Conventional and direct PCA: Resolution of samples.** In the conventional method, the analysed matrix had been summarized by concentrating to information of distance. Therefore, variations in data could be bundled within fewer axes (Fig. 3F). However, the summarizing process would mix up information oriented in various directions. Actually, the groups appeared in PC for samples became more dispersed (Fig. 4). The only difference between the two methods, conventional and direct PCA, was the subject of the SVD: distance and sequence matrices, respectively. Therefore, the dispersing was caused by using the distance matrix. This is a critical characteristic in the classification of samples; the dispersed samples hamper the separation and definition of groups. The distances presented were much longer than those of the direct method. This enlargement was partly provoked by the higher contributions as described above. Another reason is differences in source of distances: in the conventional method, distances between samples are indicated, while those between the center and samples are indicated in the direct method.

**Response to outliers.** The direct method was robust to outlying sample. A human sample selected by chance contained some unique motifs; the motifs gave long distance to other samples in the distance matrix (Fig. S2B). The axes for sPC1 and sPC2 were mainly determined by this sample in the conventional PCA (Fig. 5A); the sample behaved as an outlier, a sample that changes the whole results much.

By distinguishing the exceptional motifs, the direct method improves robustness to the outlier. In PCA of nucleotide matrix, the sample does not appear as an outlier in both sPC1 and sPC2 (Fig. 5B). Indeed, the whole shape of sPCs was almost identical to that calculated without the sample (Fig. 3B). The unique motifs gave specific direction that appeared in sPC5 and sPC6 (Fig. 5C,D).

**Response to altered alignment.** Several methods are available for the alignment of sequences, and the results of the alignment would vary slightly among the methods. However, if the default setting of the parameters was used, they did not alter the results of the direct PCA much (Figs. S3–S5). By reducing the gap penalty to one-third, the average number of gaps per sample increased from 0 to 13 in human data. Such immoderate intervention affected the limited axes; in the case of humans, it altered sPC2 and sPC3, but not other axes

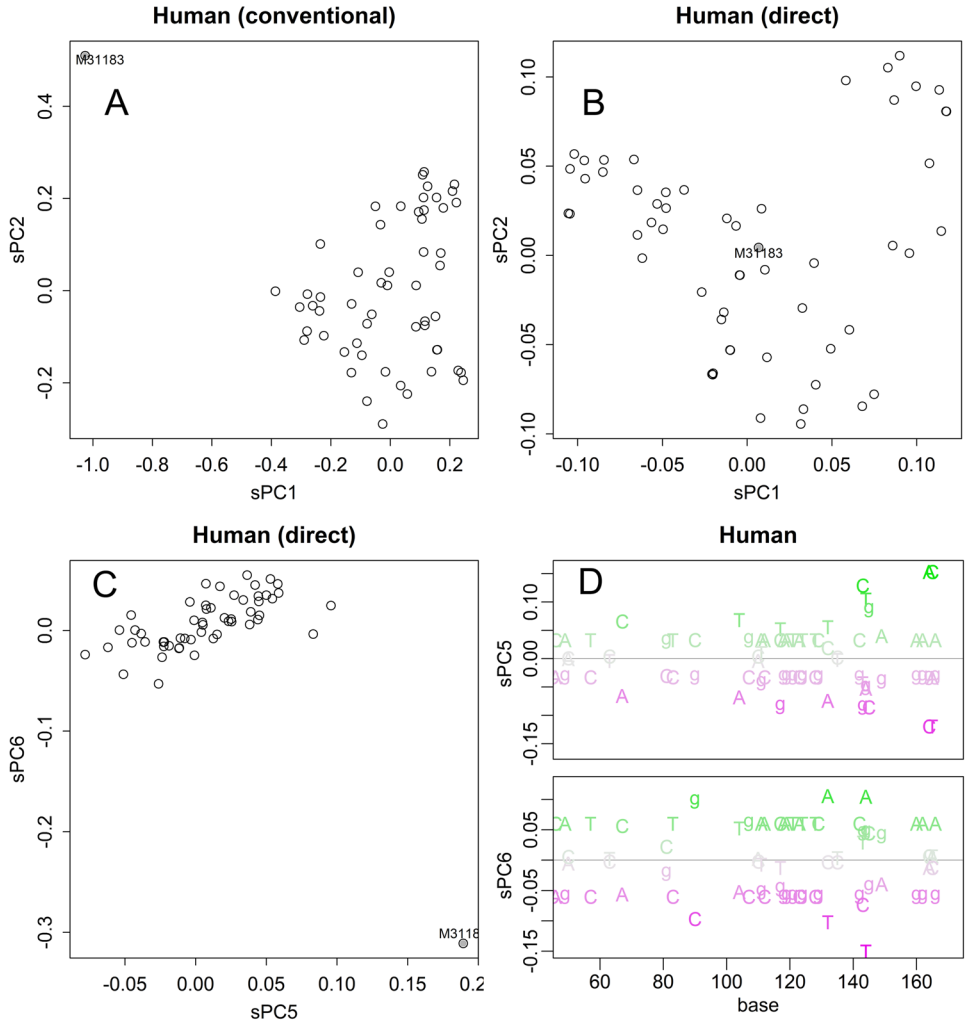



**Figure 5.** Effect of an outlier sample. Results of PCA of samples that include an outlier. Conventional method applied to the distance matrix (**A**). Direct method applied to the nucleotide matrix (**B**,**C**). The bases specific to the outlier (magenta, **D**).

(Figs. 6A,B, and S6). Conversely, the gathering of multiple clusters was severely altered in the results of the NJ method (Fig. 6C). This shows the sensitivity of the NJ method to arbitrary parameters, which would render the results of the analysis dependent on the specific analyst of the data.

**Distance adjustment by JC69 model.** The JC69 model will enlarge distances in a matrix; it makes a lattice grid to have spindle shaped skew, and could alter PCA, which finds directions and length solely in the applied matrix. In comparison with Fig. 3A, the presented case (Fig. 7A) a bit enlarged the differences among clusters; the differences are clear in the distribution of the compensation ratios (Fig. 7B). However, as samples in a same cluster would be compensated likely manner, alterations to the relationships were rather limited. Indeed, this did not change much the results of human (Fig. S7).

**Relationships between samples and attenuation patterns of the contributions.** The attenuation patterns of $\mathbf{\Sigma}$ and their contributions indicate the property of a sample group. If a sequence is randomly altered and forms a group, the PC for samples will be randomly distributed and the attenuation occurs at a constant rate (Fig. S9). This would mimic a group with a long history and various directions of evolution. In contrast, if a sequence is altered sequentially in a random manner, the trajectory can be recognized as a random walk with multiple dimensions. In this case, the PC for samples can be recognized as a set of Fourier series that oscillates at a constant wavelength (Fig. S9)[14]. In this case, attenuation occurs rapidly and may correlate with $\tan\theta$ ($1/2\,\pi > \theta \geq 0$, where $\theta$ is an arithmetic sequence with the same number to the dimensions), as a rule of thumb (Fig. S8). This would mimic a group presenting a history of evolution. For example, in the case of the Asiatic lion, the attenuation was rapid and correlated with the tangent value (Fig. S4,G,H). The contribution of PC1 and PC2 appeared higher than the linear regression line, indicating that they absorbed the large differences between the three groups. In contrast, those of bacteria (Fig. S3,G,H) and humans (Fig. S5,G,H) exhibited the character of random data as well (Fig. S8).

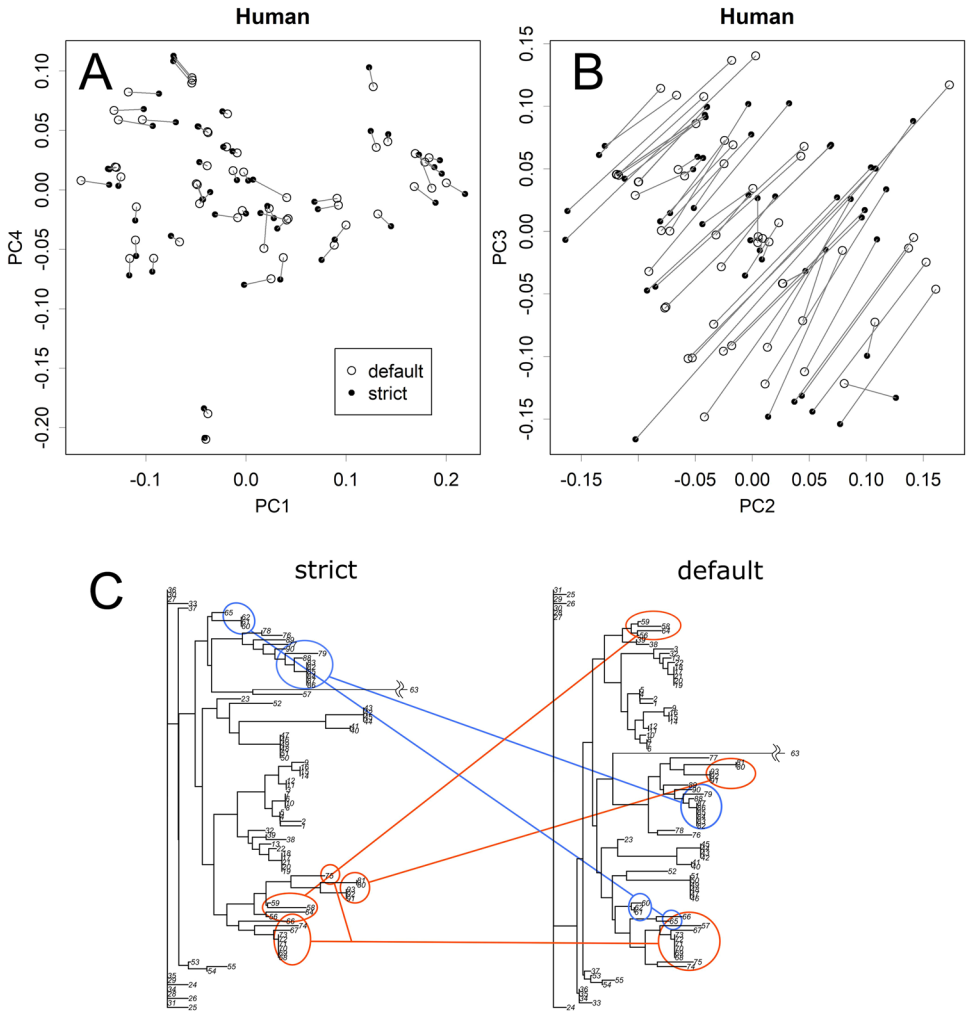



**Figure 6.** Effect of arbitrary parameters on the alignment. In the alignment process, the gap parameter of the default condition was decreased to obtain strictly aligned sequences. Results of direct PCA (**A**,**B**). Identical samples were connected. The effect appeared only in limited axes (**B** and Fig. S7). The connection of clusters was totally altered in the results of the NJ method (**C**).

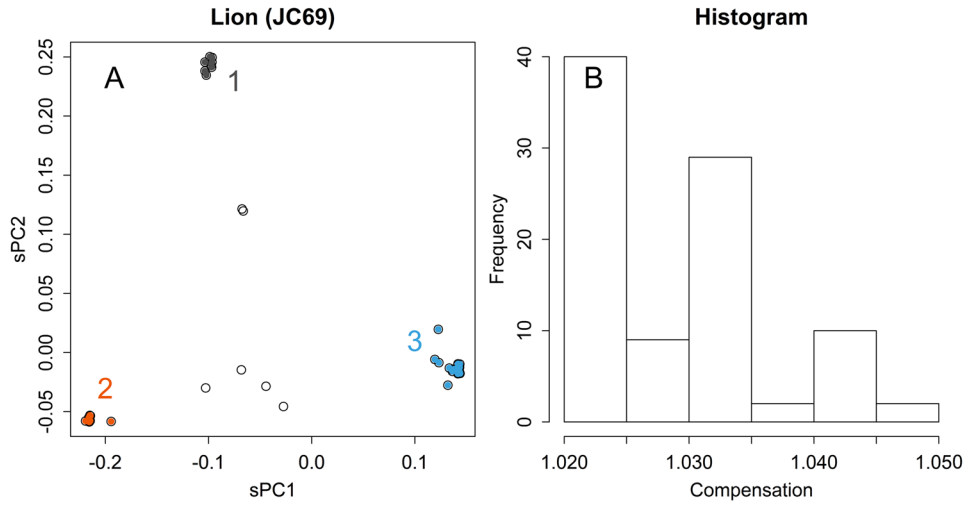



**Figure 7.** Adjustment of distances. Results of PCA applied to lion samples directory, with an adjustment according to JC69 (**A**). Histogram of ratios of the compensated and original distances (**B**).

## Discussion

In Fig. 3A, direct PCA clearly showed that the three clusters existed in the sanctuary's lions. PCA is certainly not a method of clustering that classifies samples according to specific indices because PCA does not artificially judge separations among samples. However, if the hidden data structure includes gatherings of samples that could be caused by several mechanisms, the PCA will present the structural characteristics using several axes. In the lion's case, the first two axes detected distinctive differences, which might be caused by three separate crossing relationships. Ironically, the PCA showed these relationships in a much clearer way than did the clustering results (Fig. 2B,C).

In practice, the classification of biological samples of species can be achieved by considering various characteristics at different levels. In principle, classification is a field of natural history rather than of science; e.g., various opinions are debated, but none are conclusively true. The information provided by the nucleotide sequence of a gene is just a small part of those characteristics. Therefore, information obtained by clustering does not mean much from a natural history perspective because objective feedback to the present classification systems would be required. In this case, PCA would also be beneficial.

Groups that were separated from others in some axes were further separated in other axes (Fig. 2B,C, and Supplementary Information Fig. S2_BC). This indicates that they shared certain motifs and differed in others; the shared and different motifs would appear in sPC for bases at corresponding axes. This is a typical characteristic of data with a network structure, in which samples are connected by multiple relationships with each other. Direct PCA presents each of the relationships in an axis; the appearance in an axis may be a necessary condition for a real relationship. Hence, similarity in an axis does not always prove genetic relationship; however, if such relationship really exists and if the sequence matrix has information that is sufficient for detection, the relationship will appear in one of the axes. Such multiple relationships will get lumped together and replaced by an average distance in the distance matrix. Hence, conventional PCA will show the relationships in a smaller number of dimensions (Fig. 3E,F); however, the presented structure will not show the real structure and the resolution will be damaged (Figs. 3A and 4A). Moreover, the tree-based methods can give a completely wrong structure; for example, those that appear in Fig. 6C are both wrong. By contrast, direct PCA maintains the original information of the sequence matrix, thus presenting the structure of the data in a rotated form.

PC for bases shows which base may have contributed to the differences observed among samples. The values increase and decrease across the base positions (Tables S1 and S2). High scores are not always randomly dispersed: they gather in hot spots, such as that found in Fig. 3C (sPC1), which indicate motifs that are specific to some groups. This inhomogeneous character also indicates the limitation of phylogenetic estimations performed using a small fragment of the genome; the separation of samples will be altered in the presence or absence of these specific motifs. Asiatic lions were separated into three clusters, and each group has quite small fluctuations (Fig. 3A). Differences among the clusters had provoked by specific patterns of sequences (Fig. 3C). Most of the differences appeared in the first and second axes of sPC, which represented differences of the three clusters (Fig. 3E). Unlike the case of human (Fig. 3B), Asiatic lions may have propagated from a bottle neck that was formed when they were almost extinct. The distinctively separated clusters may evoke the recently recognized species of giraffe[15]. If the clusters indicate subspecies, each of them has only limited genetic variations. However, there were several samples that showed values in between the clusters (Fig. 3A); they seemed to be offspring of hybrids among the clusters, which were not observed in giraffe's case. If the hybrids are fertile, although the clusters seemed to be isolated with each other, inter cluster mating could retrieve certain magnitude of variety. Such clear separation of sample groups should not be expected in conventional PCA (Fig. 4A), and the human samples also became more dispersed (Fig. 4B). The decreased resolution occurred because of the conventional method observed in the distance matrix, which mixed up various directions of differences. In this way, the results obtained by direct PCA cannot be negated by those obtained via other methods, which can introduce errors and distortions. It should be noted that direct PCA just rotates the matrix of sequence data; hence, the original structure of the data is maintained.

For identifying unknown samples, both classification of a set of references and relegating the sample into appropriate class are required. In comparison with hierarchical clustering methods, PCA is superior for this purpose in the following points. First, it can separate the classification of references and relegation of a sample; these could be separately performed, as finding of the unitary matrix and applying it to a sample as a multiple regression analysis[11]. Not only reducing calculations, this can protect the system from effect of outliers (Fig. 5A) or parameters (Fig. 6). Second, the presentation would become easier to understand, especially when number of samples are large. Not limited in the bacterial case shown in Fig. 2A, observing the whole structure of NJ, shown within a single dimension, will become rather difficult when numbers of samples are large (for example, that for lion and human's cases are in Supplementary Data as Fig. S2). As PCA does not mix up the dimensions, each of the presentations will become simpler. This characteristic is advantageous in classification in two ways: better resolution (Figs. 3A and 4A) and the specification of the base that allowed their separation (Fig. 3C). Moreover, the low reproducibility observed indicates a disadvantage in automatic classifications using an inappropriate model (Fig. 6C).

Superiority of the direct PCA than the conventional PCA became clear in separation of samples and robustness (Figs. 3–5, and S3–S5). These base on intactness of information recorded in multivariate data. Additionally, information of base will supply useful cues to understand relationship of sample differences and function of the sequence; this will become clearly apparent in analysing functional motifs of amino acid sequences.

It should be noted that the robust character of the method may connect to insensitivity to differences that are potentially important. Actually, PCA is a method that observe a matrix from certain direction, and differences among other directions will be ignored. The main purpose of the conventional PCA is to summarize most of differences within the first two axes[9]. The direct PCA to sequence matrix may not be a good tool for this purpose; actually, the contribution of the first two axes were much higher in the conventional method (Fig. 3E,F). Rather,

the direct method is preferable when observing toward lower PC axes, both for samples and nucleotides, to find magnitude of differences and how. Of course, such summarization, i.e., the presentation of a structure using smaller number dimensions, will inevitably introduce distortions in the presented structure.

In the subjects studied, sites that contributed highly to the separation of samples appeared scattered throughout the sequences, rather than gathered to form sequential motifs (Figs. 2D, 3C,D, and 5D). The characteristics of the sites would appear in each of the axes of sPC for samples; which samples are separated and how this occurs is indicated on the axes. The sites may have contributed to the separation, with other sites with similar scores.

The compensation for distance by JK69 did not alter much to the results (Figs. 7 and S7). This is partly because of the magnitude of distance, which was less than 0.01 in $\hat{p}$ in the presented data. However, this may not be very short in comparing homologous sequences, such as found from database searching. Other more sophisticated Marcov models[1,2] may be applicable to **C**, however, effects would be rather limited. Additionally, although Markov chain models presume vertical evolution: i.e., changing each base of a sequence by chance, diversity of sexually reproducing organisms are made by exchanging chromosomes, and those of bacteria are accelerated by horizontal gene transfer. Markov chain models could conflict with those extensive changes. As the compensations have only limited effects and they stand on assumptions that cannot be verified, there is no reason to recommend such additional calculations.

Since bases may have different weights in physical meanings; for example, the third base of a codon should be lighter than the first, because of the redundancy of the codon. Therefore, **C** could be scaled to equalize weights of positions, by estimating the weight from standard deviations of each column, for example. However, it should be noted that sometimes technical errors may occur in sequencing process. This may give a unique motif among samples; scaling by standard deviation will enlarge artefacts caused by such errors.

The use of direct PCA is not a robust alternative for clustering methods; being free from unverifiable assumptions, it can be used in science that insists falsifiability[16]. It does not alter the structure of data: it just changes the format and sorting of sequence information. No information is added or lost. In contrast, to show a tree-shaped relationship, we have to alter the data structure according to assumed characteristics. These assumptions are unverifiable and could be entirely wrong.

If one wishes to handle phylogenetic problems in a field of science, direct PCA is the only choice at present. This would reflect the philosophy in this field: e.g., many methods have been produced under ad hoc assumptions and the methods have become more complex with further assumptions. However, the appropriateness of such assumptions cannot be confirmed. In contrast, direct PCA is a simple method with no arbitral choices. In addition, it does not assume that the relationships among samples should form a phylogenetic tree. Different assumptions, as well as sensitivity to outliers or parameters, will produce a variety of results that enable "cherry-picking" of the most convenient one. It is true that not all research areas have to be strictly scientific. However, sequence data are capable of giving valuable information that directly influences human welfare: for example, classifying of viruses and validation of their changes. Those areas should be objective if we are to share knowledge validly. Although here we compared our method with previous ones, this is not a recommendation of such ad-hoc selection of methodologies. Here, the comparisons were performed to overcome the difficulty of discussion beyond intellectual frameworks that differed in falsifiability and objectivity. We hope this method will be a fair contribution to science.

## Materials and Methods

Information for MHC1 of 92 Asiatic lions[13], which lives only in the Gir forest sanctuary of Gujarat State in India, had been retrieved from DNA Data Bank of Japan (DDBJ) database[17]. As those lion sequences were not full-length, corresponding region in human gene, HMC1, sequences were also analysed for comparison; as the human database seemed to have a bias in nationality, 93 were selected out from 217 registered samples, avoiding repetition of contributors of data. One of the samples, accession ID of M31183, differed from others much (Figs. 5A,D, and S2B); it was selected by chance, and used as an example of outlier.

Nucleotide sequences for bacteria 16S ribosomal DNA were obtained in a student training course held in our department using clone library method or colony direct PCA. Environmental DNA were isolated from active sludge taken from Rinkai sewage plant Akita using ISOIL (NIPPON GENE, Japan) according to the manufactures instruction as instructed in its accompanying manual. 16S ribosomal DNA amplified by polymerase chain reaction using eubacteria specific primer set[15]. And a TOPO TA Cloning Kit (Invitorogen, USA) was used to prepare a genomic library of the PCR amplified genes. Nucleotide sequence of a single colony was determined in Biotechnology Center of Akita Prefectural University, and the sequences were determined by Sangar method[18]. The 16S ribosomal DNA sequence used to analysis were registered under DDBJ[17] accession No. LC279073 to LC279194.

The nucleotide sequences were aligned by using MUSCLE[19]. Other methods (ClustalW[20], MAFFT[21], and DECHIPER[22]) were used to identify the effect of the alignment methodologies. Default parameters used for alignment were applied unless otherwise specified. For the strict and default conditions of DECHIPER, gapOpening was -5 and -17, and gapExtension was -1 and -1.5, respectively. The sequences that were aligned using MUSCLE with accession IDs are presented as Supporting Information (Tables S3–S5).

**Digitizing the nucleotide sequences.** Calculations were performed in R[23]. All the required scripts, sample data, and a trial kit are in the Supplement Data. Aligned nucleotides of length $l$ were digitized using a set of five Boolean vectors of 0 and 1, each with the length $l$. Each of the vectors correspond to one of the nucleotides, A, T, G, C, or "–"; -); in this notation, T can be expressed as "01000". These vectors were connected to create $\overrightarrow{x}$, a vector of length $5l$, which contained the complete information about the nucleotides. Then the vectors were piled to make a sequence matrix **X**. For example, that contains two samples having TAGC and GAGC is,

$$\mathbf{X} = \begin{pmatrix} \text{A} & \text{T} & \text{G} & \text{C} & - \\ 0100 & 1000 & 0010 & 0001 & 0000 \\ 0100 & 0000 & 1010 & 0001 & 0000 \end{pmatrix}.$$

This matrix possesses mutual compatibility with the nucleotide sequence matrix. Any base position consists of the five columns, and the degree of freedom for the five columns is one. As each base is independent, length of differences was estimated in Euclidean length, and then scaled by the number of bases[11].

PCA rotates a matrix to fit the directions of differences to diagonal axes[24]. To do this, center of the rotation have to be defined in prior to the rotation. Mean center of the matrix is a preferable choice, but it could be found from specific sets of reference samples[11]. Samples apart from the center would gain larger moment, and so would have larger influence in finding directions of the differences. The average center is found as means of each columns of the matrix; hence, in the presented example **X**, it is,

$$\vec{\mathrm{m}} = (0100\ 1/2\ 000\ 1/2\ 010\ 0001\ 0000).$$

The **X** could be centered as $\vec{\mathrm{d}} = \vec{\mathrm{x}} - \vec{\mathrm{m}}$ at any rows of **X** to give matrix **D**. The **D** records differences of each sample from the center sequence at each position of the bases.

**Distance.** In direct PCA, differences among samples were estimated by using Euclidean length, as $\widehat{e_{1,2}} = \sqrt{\Sigma(\vec{\mathrm{x}_1} - \vec{\mathrm{x}_2})^2/2}$ , where $\vec{\mathrm{x}_1}$ and $\vec{\mathrm{x}_2}$ are the subjected vectors, scaling the double-counted substitutions at corresponding two nucleotides. This is further scaled by the number of bases $l$, after estimating the principal components. It is noteworthy that the scaled component presents square root of the conventional substitution rate, $\hat{p} = \Sigma|\vec{\mathrm{x}_1} - \vec{\mathrm{x}_2}|/2l$.

**Adjusting by a Markov chain model.** Distance from the center, which is recorded in **D**, was optionally adjusted the rate of parallel or back substitutions by a Markov chain model, JC69, which estimate a same frequency of changing for any bases and any positions[25]. For this, scale of $\vec{\mathrm{d}}$ was adjusted to fit the evolutionary distance, by multiplying a compensation ratio. Raw distance of each sample from the average sequence was found as $\hat{p} = \sum|d|/2l$, where $d$ is elements of $\vec{\mathrm{d}}$. Corresponding evolutionary distance was estimated as $\hat{d} = -3/4\ \ln(1 - 4/3\hat{p})$. Then each item of **D** was adjusted as $\vec{\mathrm{d}}_\mathrm{a} = \vec{\mathrm{d}} \times (\hat{d}/\hat{p})^{0.5}$, where $\hat{d}/\hat{p}$ is the compensation ratio uniquely found for a sample.

**PCA.** The **D** or its adjusted version $\mathbf{D}_\mathrm{a}$, both of which was a centered nucleotide sequence matrix, was directly subjected for singular value decomposition, $\mathbf{D} = \mathbf{L\Sigma R}^*$, producing two unitary matrixes **L** and **R**, and a rectangular diagonal matrix $\mathbf{\Sigma}$ that records singular values[11]. The principal components for samples were estimated as $\mathbf{S} = \mathbf{L\Sigma} = \mathbf{DR}$, and for bases as $\mathbf{N} = \mathbf{R\Sigma} = \mathbf{D}^*\mathbf{L}$. Therefore, columns of matrixes **S** and **N** present sets of the orthogonal directions with distances that were recorded in the unitary matrixes and the singular values, respectively; note that those are rotation of **D**, without altering relationships among elements. Then they were scaled as $\hat{\mathbf{S}} = \mathbf{S}/\sqrt{l}$, and $\hat{\mathbf{N}} = \mathbf{N}/\sqrt{n}$, where $l$ is the nucleotide length and $n$ is the number of samples, to conform to the mean Euclidean distance; by this scaling, they would become comparable size. Here both $\hat{\mathbf{S}}$ and $\hat{\mathbf{N}}$ would be notated as sPC.

To comparison of methodologies, here the conventional PCA were applied to a distance matrix that uses the same system of distance: not on $\hat{p}$, but $\hat{e}$. The distance matrix is centered and applied to singular value decomposition. Then components for samples were estimated and scaled for the length of nucleotides; the resulted values were further adjusted for double recorded in the matrix (as it comprises both upper and lower triangles) as dividing by $\sqrt{2}$. The PC was further scaled for the length $l$ of the sequence, as $\hat{\mathbf{S}} = \mathbf{S}/\sqrt{l}$.

One of the viewpoints for the superiority of the methodology was robustness. It is an important characteristic when evaluating analytical methods. In some cases, arbitrary parameters used in analyses could change the global results; for example, gap penalty is a parameter of alignment, and can change the number of gaps that are inserted in the aligned sequences. If this parameter affects the results of the analysis significantly, the analysis lacks objectivity, because the analyst may control the results by choosing the parameter. Moreover, outlying samples would also change the results; by adding or removing the outliers, analysts may control the results. Therefore, robustness could be comprehended as a concept that indicates the resistivity of a method to such arbitrary choices.

The distance and the manner in which it is projected to the axes are explained in Fig. S8. For simplicity, the data are presented in 2D; in practice, the sizes of base pairs or samples, whichever is smaller, determine the number of dimensions. If the number of samples $n$ is smaller, then one of the degrees of freedom is used for the centre mean; hence, the effective number of PCs is $n$ - 1; if it is the number of samples $m$, then it determines the number of dimensions.

The rates of data differences recorded in a PC are presented as the contributions. The size of differences is recorded in $\mathbf{\Sigma}$ as a singular value, while the two unitary matrixes **L** and **R** record the directions of PC for samples and bases, respectively. The contribution is the rate of each singular value among the total values. It should be noted that the contribution is not a normalized value; for example, a larger dimension will give smaller contributions. Normalization is difficult because the attenuation patterns are affected by the property of the data. Therefore, contributions cannot be compared between analyses.

The R code and sample sequences used in this study are available from GitHub[26]. The data simulation was performed as follows. To mimic random mutations without any limitations, a random sequence with a position length of 1,000 (each position can be 1 or −1) was produced. It was subjected to 200 rounds of sequential trials.

In each trial, the sign of a position that was chosen randomly was reversed. The product was recorded and used in the next trial for the random walk. Therefore, the rate of mutation was constant, and some of the mutations may have reverted. The recorded trials formed a sequence matrix, which was subjected to PCA.

**Neighbor joining.** The distance matrix $\hat{e}$ was scaled by the same way of the conventional PCA, and further processed by *nj* function of *ape* package[27] of R[23].

Received: 29 January 2019; Accepted: 23 November 2019;
Published online: 17 December 2019

## References

1. San Mauro, D. & Agorreta, A. In *Cellular and Molecular Biology Letters* Vol. 15 311 (2010).
2. Yang, Z. *Computational Molecular Evolution*. (Oxford University Press (2006).
3. Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nat Rev Genet* **13**, 303–314, https://doi.org/10.1038/nrg3186 (2012).
4. Ellis, G. & Silk, J. Scientific method: Defend the integrity of physics. *Nature* **516**, 321–323, https://doi.org/10.1038/516321a (2014).
5. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA* **96**, 3801–3806 (1999).
6. Polz, M. F., Alm, E. J. & Hanage, W. P. Horizontal Gene Transfer and the Evolution of Bacterial and Archaeal Population Structure. *Trends Genet* **29**, 170–175, https://doi.org/10.1016/j.tig.2012.12.006 (2013).
7. Iwasaki, Y., Abe, T., Wada, K., Wada, Y. & Ikemura, T. A Novel Bioinformatics Strategy to Analyze Microbial Big Sequence Data for Efficient Knowledge Discovery: Batch-Learning Self-Organizing Map (BLSOM). *Microorganisms* **1**, 137–157, https://doi.org/10.3390/microorganisms1010137 (2013).
8. Reiss, J. & Sprenger, J. *Scientific Objectivity*, https://plato.stanford.edu/archives/win2017/entries/scientific-objectivity/ (2017).
9. Garrity, G. M., Bell, J. A. & Liburn, T. In *Bergey's Manual of Systematic Bacteriology* Vol. 2nd ed (eds. Don J. Brenner, Noel R. Krieg, James T. Staley, & George M. Garrity) (Springer-Verlag, (2005).
10. Harris, S. & Harris, D. *Digital Design and Computer Architecture*. 2 edn, (Morgan Kaufmann (2012).
11. Konishi, T. Principal component analysis for designed experiments. *BMC Bioinformatics* **16**(Suppl 18), S7, https://doi.org/10.1186/1471-2105-16-S18-S7 (2015).
12. Alexe, G. *et al*. PCA and clustering reveal alternate mtDNA phylogeny of N and M clades. *J Mol Evol* **67**, 465–487, https://doi.org/10.1007/s00239-008-9148-7 (2008).
13. Sachdev, M., Sankaranarayanan, R., Reddanna, P., Thangaraj, K. & Singh, L. Major histocompatibility complex class I polymorphism in Asiatic lions. *Tissue Antigens* **66**, 9–18, https://doi.org/10.1111/j.1399-0039.2005.00432.x (2005).
14. Antognini, J. M. & Sohl-Dickstein, J. PCA of high dimensional random walks with comparison to neural network training. *arXiv*, 1806.08805 (2018).
15. Fennessy, J. *et al*. Multi-locus analyses reveal four giraffe species instead of one. *Curr. Biol.* **26**, 2543–2549, https://doi.org/10.1016/j.cub.2016.07.036 (2016).
16. Thornton, S. "Karl Popper" In *The Stanford Encyclopedia of Philosophy* (ed. Edward N. Zalta) https://plato.stanford.edu/entries/popper/ (2018).
17. Mashima, J. *et al*. DNA Data Bank of Japan. *Nucleic Acids Res* **45**, D25–D31, https://doi.org/10.1093/nar/gkw1001 (2017).
18. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**, 441–448 (1975).
19. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797, https://doi.org/10.1093/nar/gkh340 (2004).
20. Larkin, M. A. *et al*. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948, https://doi.org/10.1093/bioinformatics/btm404 (2007).
21. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–3066 (2002).
22. Wright, E. S. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics* **16**, 322, https://doi.org/10.1186/s12859-015-0749-z (2015).
23. R_Core_Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing (2017).
24. Jolliffe, I. T. *Principal Component Analysis*. (Springer-Verlag (2002).
25. Jukes, T. H. & Cantor, C. R. In *Mammalian Protein Metabolism* Vol. 3 (ed. Munro, H. N.) 21–132 (Academic Press, (1969).
26. Konishi, T. *direct PCA for sequences*, https://github.com/TomokazuKonishi/direct-PCA-for-sequences (2019).
27. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

## Acknowledgements

Conflict of Interest: none declared.

## Author contributions

T.K. wrote the main manuscript. All the other authors contributed to data curation. The study was started by requests of S.M. and K.O. to help the lion and the bacterial studies, respectively.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41598-019-55253-0.

**Correspondence** and requests for materials should be addressed to T.K.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.