






Quality Metrics to Guide Visual Analysis of High Dimensional Genomics Data

S. Johansson Fernstad¹ , A. Macquisten¹ , J. Berrington² , N. Embleton²  and C. Stewart³ 

¹School of Computing, Newcastle University, Newcastle-upon-Tyne, UK

²Newcastle Neonatal Service, Royal Victoria Infirmary, Newcastle-upon-Tyne, UK

³Institute of Cellular Medicine, Newcastle University, Newcastle-upon-Tyne, UK

Abstract

Studies of genome sequenced data are increasingly common in many domains. Technological advances enable detection of hundreds of thousands of biological entities in samples, resulting in extremely high dimensional data. To enable exploration and understanding of such data, efficient visual analysis approaches are needed that take domain and data specific requirements into account. Based on a survey with bioscience experts, this paper suggests a categorisation and a set of quality metrics to identify patterns of interest, which can be used as guidance in visual analysis, as demonstrated in the paper.

CCS Concepts

• **Human-centered computing** → **Visual analytics**; • **Applied computing** → **Bioinformatics**;

1. Introduction

Innovation in bioscience is increasingly data-driven. Advances in genome sequencing techniques have made it possible to rapidly detect large numbers of biological entities in samples from various environments, making the study of such data increasingly common in many domains. These datasets can be extremely high dimensional with each sequence-read (or biological entity) corresponding to a data dimension. The high dimensionality is a major analysis challenge, and efficient methods for exploratory analysis and visualization are crucial for gaining insights from genomics data. Common visualization methods are able to efficiently handle moderately sized datasets, but with dimensionalities increasing to hundreds of thousands, alternative approaches are necessary. One approach is to use quality metrics (QM), or measures of interestingness, as an aid to guide users to data subsets of interest [BBK*18]. What is interesting in a dataset is, however, highly task and domain dependent. We argue that the definition of appropriate QM has to be done within a domain specific context. This paper presents the result of a survey with domain experts that identify patterns of relevance for studies of genomics data. Based on the survey, a set of QM are suggested, which aim to measure these patterns in context of biological entities. The QMs can be used to guide visual and interactive analysis, for instance by highlighting particularly interesting data, for extraction of data subsets for further investigation, or for ordering in visual representations to aid pattern identification. The utility of the QMs are demonstrated through a set of examples where data from a study of the gut microbiome of preterm infants [SEC*17] are visualized.

2. Background

This section describes some of the main features of genomics data, and cover relevant previous research in high dimensional data and QM in visualization. While this paper is focussed on genomics data, the suggested approaches would generally be equally applicable to other types of 'omics data.

2.1. Genomics Data

Data from genome sequencing studies can generally be defined as multivariate, with genome sequences or biological entities (such as bacterial species) as dimensions, and samples as data items. The data values are the counts of individual biological entities in samples, providing an abundance profile for each sample. The data is very high dimensional and may include thousands or even millions of unique biological entities. Meanwhile, the number of samples is often relatively small, leading to extremely sparse data spaces. The samples are often categorised into different groups, such as test-control, healthy-unhealthy, female-male and so on, with varying number of categories. Throughout this paper, genomics data dimensions are referred to as biological entities, data items are referred to as samples, and categories of samples are referred to as sample groups. Abundance refers to the count of a biological entity, relative abundance refers to the relative count of a biological entity within a sample, and prevalence refers to if a biological entity is detected or not in a sample. The abundance distribution is often strongly skewed in genomics data, with high abundance of an entity in a small number of samples and low abundance or no prevalence in a larger number of samples. Furthermore, commonly only a small

part of biological entities are highly abundant and prevalent. The visualization examples in this paper utilise data from a study of the gut microbiome of preterm infants [SEC*17], consisting of 516 biological entities across 867 samples. The samples are classified by *Birth Mode*, with *Cesarean Birth* and *Vaginal Birth* as sample groups. The biological entities are, in this case, Operational Taxonomic Units (OTUs), which are a close approximation to bacterial species, extracted through clustering of DNA sequences. OTUs have an associated hierarchical taxonomy through the biological classification system, and are typically converted into a genus for analysis as an OTU name generally has no biological meaning.

2.2. High Dimensional Data in Visualization

High dimensionality in visualization can be defined as when it becomes challenging to visually extract meaningful relations among dimensions [BTK11]. Common visualization methods for multivariate data, such as Parallel Coordinates (PC) [Ins85] and Scatter Plot Matrix (SPloM) [BC87], are useful for datasets with moderately high dimensionality, but their usability quickly decrease with increasing dimensionality. Extensive overview of recent visualization systems and methods for analysis of high dimensional data are available in Bertini et al. [BTK11], Johansson Fernstad et al. [JSJ13] and Liu et al. [LMW*17]. A common approach to analysis of high dimensional data is to apply dimension reduction, which may involve the projection of data to a new set of dimensions, including methods such as self-organizing maps [Koh98], multidimensional scaling and principal components analysis [Cox05]; or the selection of a subset of particularly interesting dimensions to retain for analysis. Projection methods may often be computationally efficient, but are disadvantaged by unintuitive relationships between the original and new set of dimensions. For analysis of genomics data, selection of interesting subsets of biological entities may be more straightforward than projection, since individual entities often are of interest.

The utilization of QM has been popular for tasks such as projection, ordering, abstraction and view optimization [BTK11, BBK*18]. Bertini et al. [BTK11] define QM as calculated metrics that capture data properties which are useful for the extraction of meaningful information about data. In context of high dimensional data visualization, a QM can be thought of as a measure of how interesting a dimension, a subset of dimensions or a dimension ordering is, or how well it represents the underlying data. As such it can help the data analyst to concentrate on the most interesting part of the data. The definition of what is interesting is domain and task dependent, and in many cases multiple measures may be relevant [JJ09]. This paper suggests a set of QM of particular relevance for the visual analysis of genomics data, based on interviews and surveys with domain experts. QM have been used previously to deal with high dimensionality in visualization. Johansson Fernstad et al. [JJA*11, JSJ13] represented dimensions in context of multiple QM, using a PC that is also used for interactive subset selection. Their approach were in spirit related to methods presented by Turkay et al. [TFH11, TPH12] and Krause et al. [KDFB16], who both link representations of dimension space and item space. Wang et al. [WLS19] provided subspace comparison through dimension aggregation and incremental anal-

ysis. Lehmann et al. [LHT15] identified a set of metrics that work similar to human perception, but concluded that further studies are needed to understand how perceptivity depends on the underlying data. Earlier studies [LAdS12, STMT12] have also shown that the success of a quality metric largely depends on the underlying dataset. Behrisch et al. [BBK*18] provide an extensive review and categorisation of the use of QM in visualization, separating the QM calculation into *Image Space*, *Data Space*, and *Hybrid*. The QM suggested in this paper are *Data Space* metrics, and can as such be considered visualization agnostic. They are based on tasks and patterns of relevance for studies of genomics data, taking into account typical features of this data.

3. Quality Metrics for Genomics Data

To address the high dimensionality challenge of genomics data, the QM presented here are focussed on identification of interesting biological entities or groups of entities. From visualization viewpoint, such QM can be used to highlight data of potential interest for further investigation, for selection of interesting subsets of biological entities to be analysed visually, and for ordering of entities in visualization to increase perceivability of interesting data patterns. Previous research into QM for studies of genomics data [JJA*11] define the abundance and prevalence of biological entities as QM of interest, as well as a confidence value for the taxonomic classification of entities. These QM were chosen based on informal interviews with bio-scientists. To provide a broader foundation, we asked 20 scientists with expertise in bioinformatics (5), microbiology (10) and other biology (5), within a range of application domains (medicine and health, pharma, agriculture, environment, and personal and home care), to answer an online questionnaire regarding which data patterns they find most interesting for studies of microbial ecology. To define a set of patterns to be used in the study, an initial set were selected based on our previous work and iterated with two of our microbiologist collaborators, resulting in the patterns listed in figure 1. The participants were asked to rank the patterns using a five point likert scale (1 = not interesting, 5 = very interesting), and were provided a free text option to add other patterns of interest. Figure 1 displays the result of the questionnaire in terms of percentage of participants that answered 4 (interesting, green colour) and 5 (very interesting, blue colour). Additional patterns suggested were: temporal relationships, predictive power, and phylogenetic structure in communities.

The patterns deemed most interesting (with a rank of 4 or 5 by more than 60% of participants) can be separated into five categories: **1) Individual entity values**: the abundance and prevalence of biological entities (first and second bar); **2) Sample group differences**: the difference in abundance and prevalence between sample groups (third and fourth bar); **3) Multivariate entity relationships**: the correlation and the similarity between biological entities (fifth and sixth bar); **4) Taxonomy**: certainty of taxonomic classification (seventh bar); and **5) Sample–entity relationship**: the relationship between biological entities and individual or groups of samples (two rightmost bars).

This paper presents a set of QM based on the first three of these categories. The QM are by no means intended as an exhaustive list of all possible metrics for these patterns. The certainty of taxo-

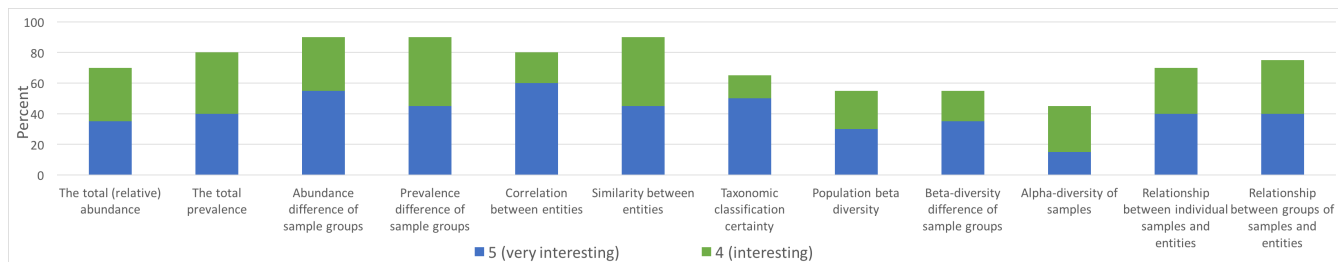


Figure 1: Percentage of participants ranking patterns as interesting (4) and very interesting (5) for studies of microbial ecology.

nomic classification, **category 4**, is not included as it is a measure extracted as part of the initial assignment of taxonomy. Furthermore, **category 5**, the relationship between biological entities and individual samples or groups of samples, is considered too complex to be successfully described as a single quantitative value, and are more meaningfully explored in an interactive visualization system. The suggested QM are designed mainly for dimension ranking, although they could be adapted to fit other purposes. Their utility is demonstrated through subset selection and ordering in PC and Scatter Plots, with sample polylines and points coloured by the two *Birth Mode* groups (*Vaginal Birth* represented by blue and *Cesarean Birth* by red). Additional examples are provided as supplemental material. The following notation is used: a genomics dataset X , includes M biological entities and N samples. \vec{x}_j and \vec{x}_k are biological entities where $j, k = 1, \dots, M$ and $x_{i,j}$ is the abundance or relative abundance of biological entity j in sample i .

Category 1 – Individual Entity Values: The abundance of a biological entity corresponds to the total count of that entity, that is detected in all samples. Logarithmic scaling is often applied, due to the skewness of the abundance distribution. An abundance QM for entity \vec{x}_j can, hence, be calculated as $Q_{ab}(\vec{x}_j) = \log(\sum_{i=1}^N x_{i,j})$. Prevalence, on the other hand, is the relative number of samples an entity has been detected in. The QM for prevalence for entity \vec{x}_j can then be defined as $Q_{pr}(\vec{x}_j) = \sum_{i=1}^N (1 : x_{i,j} > 0)$. Thus, high abundance or prevalence values are assigned to entities with high total abundance or prevalence. Figure 2 shows examples of using the metrics to select the ten most abundant and ten most prevalent entities for further examination using PC. Comparing the two PC it becomes apparent that the most abundant entities are not exactly the same as the most prevalent, for instance the three rightmost entities in figure 2a are detected at higher counts than the three rightmost in figure 2b, although the latter are detected in more samples.

Category 2 – Sample Group Differences: The difference between groups of samples is often of interest for analysis. Previous research [JJA*11] suggested QM based on the difference in average abundance and prevalence between all sample groups. Prevalence is a binary value, either an entity is prevalent in a sample, or not, and the prevalence value of a sample group can straightforwardly be described as a percentage (i.e. entity A is prevalent in 40% of samples in group X). The prevalence difference QM of entity \vec{x}_j can then be defined as the average difference in prevalence between sample groups, $Q_{D_{pr}}(\vec{x}_j) = (\sum_{a=1}^{G-1} \sum_{b=a+1}^G |Q_{pr}(\vec{x}_j, a) - Q_{pr}(\vec{x}_j, b)|) / (G - 1)$, where G is the number of groups and $Q_{pr}(\vec{x}_j, a)$ is the prevalence in group a , such

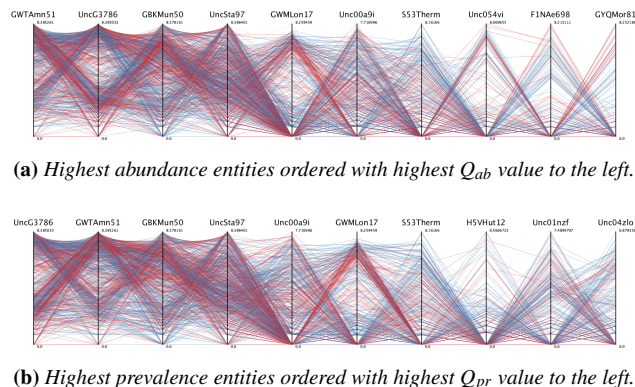


Figure 2: The 10 biological entities with highest Q_{ab} and Q_{pr} values, applying logarithmic scaling to the axes.

that $Q_{D_{pr}}(\vec{x}_j)$ is high when the prevalence difference between sample groups is high. In figure 3, $Q_{D_{pr}}(\vec{x}_j)$ is used to select the biological entities with highest prevalence difference. It is visible that the first and fourth entity from left, *HUJBact2* and *GFKSpe61* (both *Actinomyces*), are only prevalent in blue samples, while the second from right, *LcbSal24* (a *Lactobacillus*), is only prevalent in red samples. For abundance, which is a numerical measure, an issue with the approach by Johansson Fernstad et al. [JJA*11] is that sample groups may have a big difference in average abundance while still largely overlap. This is comparable to clustering where the centroids are relatively distant from each other, but the clusters are still not well separated. To address this, this paper suggest the use of cluster separation metrics to evaluate if groups of samples are well separated within a biological entity. In the examples provided here, the Davies-Bouldin index [DB79] is used, but other cluster separation measures, such as silhouette analysis [Rou87], could be used as well. The Davies-Bouldin index is based on a ratio between the within cluster scatter (S_a) and the separation between pairs of clusters ($C_{a,b}$). The goodness of clustering for a cluster pair is defined as $R_{a,b} = (S_a + S_b) / C_{a,b}$. The Davies-Bouldin index, which provides a goodness measure for the whole clustering, is then defined as $DB = (\sum_{a=1}^G (D_a)) / G$, where G is the number of clusters (sample groups) and $D_a = \max_{b \neq a} (R_{a,b})$ is the maximum cluster pair goodness value for cluster a . A low DB corresponds to a high cluster separation, thus the QM for entity \vec{x}_j is defined as $Q_{D_{ab}}(\vec{x}_j) = \max DB - DB(\vec{x}_j)$, where $\max DB$ is the highest Davies-Bouldin index calculated for the individual biological entities. This

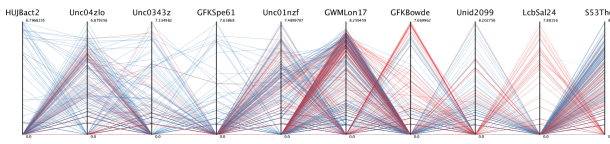
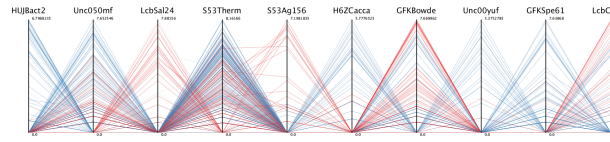
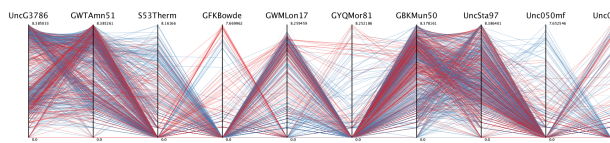


Figure 3: Entities with highest prevalence difference, applying logarithmic scaling and ordered with highest $Q_{D_{pr}}$ value to the left.



(a) Sample group difference identified using the Davies-Bouldin index.



(b) Sample group difference identified based on average abundance.

Figure 4: Biological entities with highest abundance difference between sample groups, using different metrics. Logarithmic scaling is applied and axes ordered with highest QM value to the left.

then results in a high $Q_{D_{ab}}$ for entities where the sample groups are well separated. Figure 4 displays the ten highest ranked entities based on cluster separation and average abundance difference. It is clearly visible that sample groups are more separated in figure 4a than in figure 4b, confirming that cluster separation may be a better QM than difference of averages. Identification of biological entities where sample groups are different can here help to understand differences in the microbiome that may be driven by *Birth Mode*.

Category 3 – Multivariate Entity Relationships: Similarity and correlation measures provide descriptions of relationships between pairs of biological entities. They can indicate coexistences and support identification of entities with potential symbiotic or antibiotic interaction. A range of similarity and correlation measures have been suggested for genomics analysis [KLL*10], including Pearson correlation, Chi-squared, Gower and Canberra distances, and Bray-Curtis dissimilarity [BC57]. Another group of similarity measures are the UniFrac distance [LLK*11] which takes the phylogenetic similarity of entities into account. Pearson correlation is used in the examples in this paper, but in principle, any pairwise correlation or similarity metric could be used, including output from analysis tools such as QIIME2 [BRD*19], mothur [SWR*09] or Bioconductor [LHP*13]. As a basis, $Q_{Sim}(\vec{x}_j, \vec{x}_k)$ is defined as the correlation or similarity (C) of a pair of biological entities \vec{x}_j and \vec{x}_k . It can then be used for ordering of variables, extraction of pairs with high or low similarity, or summarised to extract individual entities with high similarity to other entities. Since both positive and negative correlation can be of interest, a high QM is assigned irrespective of the sign of the correlation, hence, $Q_{Sim}(\vec{x}_j, \vec{x}_k) = |C_{Cor}(\vec{x}_j, \vec{x}_k)|$. Where a dissimilarity measure

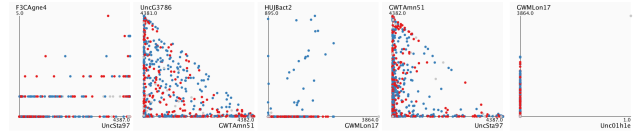


Figure 5: The five entity pairs with highest Q_{Sim} .

is used, such as Bray-Curtis [BC57], the metric is calculated as $Q_{Sim}(\vec{x}_j, \vec{x}_k) = (1 - C_{Dis}(\vec{x}_j, \vec{x}_k))$. The abundance distribution in entities is often highly skewed, with a large number of entities with very low prevalence. These entities are mathematically similar, but they are not interesting for identifying coexistence and similar patterns. Thus, we suggest combining the metric with a prevalence threshold t_p , setting $Q_{Sim}(\vec{x}_j, \vec{x}_k) = 0$ for $\vec{x}_j, \vec{x}_k \leq t_p$. Figure 5 display scatter plots of the five pairs of biological entities that were highest ranked by $Q_{Sim}(\vec{x}_j, \vec{x}_k)$ based on Pearson correlation. While it is clear from the figure that correlation patterns are relatively noisy in this dataset, with a large number of samples near the axes meaning they are detected at low levels or not detected at all for that biological entity, some potentially interesting patterns are still visible. For instance, the first and third plot show that samples with higher abundance of the entities represented by the y-axes, also tend to have higher abundance of the entities represented by the x-axes, indicating a potential pattern of symbiosis or co-existence of those entities. The second and fourth plot, on the other hand, display what in part could be described as a negative correlation, where no samples have high abundance of both biological entities concurrently, which could indicate a possible antibiotic pattern. The pairwise metric can be useful for identifying these kind of patterns, as well as for ordering of entities in multivariate visualization, using approaches such as the correlation based ordering described in Johansson and Johansson [JJ09]. In situations where a single value per entity is beneficial, such as when ranking entities for subset selection, a summarised QM can be useful, which can be calculated as $Q_{Sim_{sum}}(\vec{x}_j) = \sum_{k=1, k \neq j}^M (Q_{Sim}(\vec{x}_j, \vec{x}_k))$. The supplemental material includes further examples of visualization where the above QMs are utilised.

4. Conclusions and Future Work

Visual analysis of high dimensional data is particularly challenging in studies of genomics data, where rapid technological advances generate thousands of dimensions. Quality metrics are commonly used in high dimensional data analysis to guide extraction of subsets of particularly interesting data or for dimension ordering. The relevance of a quality metric is however often task and domain dependent. We identified patterns of interest for the analysis of genomics data, through a survey with bioscience experts. A set of quality metrics were suggested to support identification of these patterns. The utility of the metrics was demonstrated through dimension selection and ordering of data from a gut microbiome study, visualized using parallel coordinates and scatter plots. In the future, the metrics will be incorporated in visual analytics systems to provide semi-automated guidance. Their usability will be evaluated through quantitative usability studies as well as through qualitative user testing with domain experts.

References

- [BBK*18] BEHRISCH M., BLUMENSCHNEIN M., KIM N. W., SHAO L., EL-ASSADY M., FUCHS J., SEEBACHER D., DIEHL A., BRANDES U., PEISTER H., ET AL.: Quality metrics for information visualization. In *Computer Graphics Forum* (2018), vol. 37, Wiley Online Library, pp. 625–662. 1, 2
- [BC57] BRAY J. R., CURTIS J. T.: An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs* 27, 4 (October 1957), 325–349. 4
- [BC87] BECKER R. A., CLEVELAND W. S.: Brushing scatterplots. *Technometrics* 29, 2 (May 1987), 127–142. 2
- [BRD*19] BOLYEN E., RIDEOUT J. R., DILLON M. R., BOKULICH N. A., ABNET C. C., AL-GHALITH G. A., ALEXANDER H., ALM E. J., ARUMUGAM M., ASNICAR F., ET AL.: Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature biotechnology* 37, 8 (2019), 852–857. 4
- [BTK11] BERTINI E., TATU A., KEIM D.: Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2203–2212. 2
- [Cox05] COX T.: *Introduction to Multivariate Analysis*. Hodder Arnold, 2005. 2
- [DB79] DAVIES D. L., BOULDIN D. W.: A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 2 (1979), 224–227. 3
- [Ins85] INSELBERG A.: The plane with parallel coordinates. *The Visual Computer* 1, 4 (1985), 69–91. 2
- [JJ09] JOHANSSON S., JOHANSSON J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 993–1000. 2, 4
- [JJA*11] JOHANSSON FERNSTAD S., JOHANSSON J., ADAMS S., SHAW J., TAYLOR D.: Visual exploration of microbial populations. In *Proceedings of IEEE Symposium on Biological Data Visualization* (October 2011), IEEE, pp. 127–134. 2, 3
- [JSJ13] JOHANSSON FERNSTAD S., SHAW J., JOHANSSON J.: Quality-based guidance for exploratory dimensionality reduction. *Information Visualization* 12, 1 (Jan 2013), 44–64. 2
- [KDFB16] KRAUSE J., DASGUPTA A., FEKETE J.-D., BERTINI E.: Seekaview: An intelligent dimensionality reduction strategy for navigating high-dimensional data spaces. In *LDV 2016-IEEE 6th Symposium on Large Data Analysis and Visualization* (2016). 2
- [KLL*10] KUCZYNSKI J., LIU Z., LOZUPONE C., McDONALD D., FIERER N., KNIGHT R.: Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nature methods* 7, 10 (2010), 813. 4
- [Koh98] KOHONEN T.: The self-organizing map. *Neurocomputing* 21, 1–3 (1998), 1–6. 2
- [LAdS12] LEWIS J. M., ACKERMAN M., DE SA V. R.: Human cluster evaluation and formal quality measures: A comparative study. In *CogSci* (2012), pp. 1870–1875. 2
- [LHP*13] LAWRENCE M., HUBER W., PAGES H., ABOYOUN P., CARLSON M., GENTLEMAN R., MORGAN M. T., CAREY V. J.: Software for computing and annotating genomic ranges. *PLoS computational biology* 9, 8 (2013), e1003118. 4
- [LHT15] LEHMANN D. J., HUNDT S., THEISEL H.: A study on quality metrics vs. human perception: Can visual measures help us to filter visualizations of interest? *it-Information Technology* 57, 1 (2015), 11–21. 2
- [LLK*11] LOZUPONE C., LLADSER M. E., KNIGHTS D., STOMBAUGH J., KNIGHT R.: Unifrac: an effective distance metric for microbial community comparison. *The ISME journal* 5, 2 (2011), 169. 4
- [LMW*17] LIU S., MALJOVEC D., WANG B., BREMER P.-T., PASCUCCI V.: Visualizing high-dimensional data: Advances in the past decade. *IEEE Transactions on Visualization and Computer Graphics* 23, 3 (2017), 1249–1268. 2
- [Rou87] ROUSSEEUW P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65. 3
- [SEC*17] STEWART C. J., EMBLETON N. D., CLEMENTS E., LUNA P. N., SMITH D. P., FOFANOVA T. Y., NELSON A., TAYLOR G., ORR C. H., PETROSINO J. F., ET AL.: Cesarean or vaginal birth does not impact the longitudinal development of the gut microbiome in a cohort of exclusively preterm infants. *Frontiers in microbiology* 8 (2017), 1008. 1, 2
- [STMT12] SEDLMAIR M., TATU A., MUNZNER T., TORY M.: A taxonomy of visual cluster separation factors. In *Computer Graphics Forum* (2012), vol. 31, Wiley Online Library, pp. 1335–1344. 2
- [SWR*09] SCHLOSS P. D., WESTCOTT S. L., RYABIN T., HALL J. R., HARTMANN M., HOLLISTER E. B., LESNIEWSKI R. A., OAKLEY B. B., PARKS D. H., ROBINSON C. J., ET AL.: Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 23 (2009), 7537–7541. 4
- [TFH11] TURKAY C., FILZMOSER P., HAUSER H.: Brushing dimensions – a dual visual analysis model for high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2591–2599. 2
- [TPH12] TURKAY C., PARULEK J., HAUSER H.: Dual analysis of dna microarrays. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies* (2012), pp. 26:1–26:8. 2
- [WLS19] WANG J., LIU X., SHEN H. W.: High-dimensional data analysis with subspace comparison using matrix visualization. *Information Visualization* 18, 1 (2019), 94–109. doi:10.1177/1473871617733996. 2