Published in final edited form as:

Methods Mol Biol. 2017 ; 1537: 347–364. doi:10.1007/978-1-4939-6685-1_20.

Exploring Genome-Wide Expression Profiles Using Machine Learning Techniques

Moritz Kebschull and Panos N. Papapanou

Abstract

Although contemporary high-throughput –omics methods produce high-dimensional data, the resulting wealth of information is difficult to assess using traditional statistical procedures. Machine learning methods facilitate the detection of additional patterns, beyond the mere identification of lists of features that differ between groups.

Here, we demonstrate the utility of (1) supervised classification algorithms in class validation, and (2) unsupervised clustering in class discovery. We use data from our previous work that described the transcriptional profiles of gingival tissue samples obtained from subjects suffering from chronic or aggressive periodontitis (1) to test whether the two diagnostic entities were also characterized by differences on the molecular level, and (2) to search for a novel, alternative classification of periodontitis based on the tissue transcriptomes.

Using machine learning technology, we provide evidence for diagnostic imprecision in the currently accepted classification of periodontitis, and demonstrate that a novel, alternative classification based on differences in gingival tissue transcriptomes is feasible. The outlined procedures allow for the unbiased interrogation of high-dimensional datasets for characteristic underlying classes, and are applicable to a broad range of –omics data.

Keywords

Periodontal disease; Aggressive periodontitis; Chronic periodontitis; Gene expression; Transcriptome; Gingiva; Classification; Machine learning

1 Introduction

The high-dimensional data produced by contemporary –omics methodology (see Chapter 18 by Kebschull et al. of this volume) provide a wealth of information that is difficult to analyze using traditional statistical methods. In Chapter 18 of this volume, we have presented a workflow for the identification of features in the dataset that differ between a priori-defined subgroups of samples, e.g., based on clinical diagnosis or experimental treatment allocation. These analyses produce lists of features, and, subsequently, of ontology groups that are differentially expressed after correction for multiple hypothesis testing. Nevertheless, differential expression of features between groups does not necessarily imply that these groups are distinguishable based on characteristic patterns of these features. In addition, differential expression analyses can only assess dissimilarities between already defined groups, while novel groupings based on characteristic patterns in the data are impossible to generate.

To address these problems, –omics researchers have ventured into the field of machine learning, a genre of computer science that uses artificial intelligence for pattern recognition and computational learning. Specifically, both supervised and unsupervised learning approaches proved useful for the analysis of –omics data. Supervised learning encompasses the training of a learning algorithm on labeled samples and the subsequent use of the learned algorithm to predict the labels for new, unlabeled samples. This approach for the classification of samples based on patterns recognized by the learner is commonly used for class validation, i.e., the evaluation of the learnability of a class distinction, e.g., two different diagnoses. In contrast, unsupervised learning entails the subdivision of a set of samples with no prior allocation, into two or more novel classes, based on characteristic similarities of their encompassing features.

Our group has used machine learning techniques to study the classification of periodontal diseases based on the transcriptomes of 240 disease-affected gingival tissue samples from 120 subjects with chronic or aggressive periodontitis. First, we performed a “class validation” analysis to evaluate whether supervised classification algorithms were able to distinguish chronic from aggressive periodontitis based on the tissue transcriptomes. Indeed, the best-performing algorithms were able to reach high diagnostic accuracy in the differentiation between chronic and aggressive periodontitis. However, to do so, the algorithms had to utilize the expression of thousands of genes as diagnostic features, rendering the classifier very computationally demanding, and the results likely less generalizable. In addition, we found a substantial heterogeneity in classifier performance, despite the use of generally accepted, robust methods and otherwise identical procedures, which is strongly suggestive of diagnostic imprecision in the current classification of periodontitis [1]. We subsequently sought to detect novel classes of periodontitis patients based on characteristic transcriptomic patterns in their diseased gingival tissues using unsupervised clustering. Since disease severity at the particular gingival unit was shown to be a major determinant of local gene expression, and given our major goal to allocate the *patient*, rather than the *individual tissue sample*, we utilized a model-based clustering approach. Specifically, we used mixture models implemented in the flexmix package in R [2], and corrected for both the severity of periodontitis at the particular gingival tissue sample (i.e., the maximum probing depth adjacent to the sample) and the interdependency of multiple tissue samples obtained from the same patient. Our approach identified two novel clusters of periodontitis patients that did not only differ substantially in their defining underlying transcriptomic features, but also in their whole-mouth clinical and microbiological profiles, as well as in serological markers of periodontitis. We suggested that, after appropriate validation steps in independent cohorts and in longitudinal studies, these findings could support a novel, pathobiology-based classification of periodontitis [3].

In this chapter, we provide an overview of the practical application of machine learning techniques on high-dimensional –omics data. First, we delineate the steps necessary to perform a class validation analysis of a dataset obtained using the information provided in our accompanying chapters in this volume (see Chapters 18 and 19 by Kebschull et al.), utilizing the CMA package in R. Then, we describe the unsupervised, mixture model-based clustering of the same dataset using the flexmix package.

2 Materials

2.1 Hardware

1. A computer with x86–64 compatible processor(s) running either Linux, Mac OS X, or Windows (*see* Note 1). RAM 16GB.

2.2 Software

1. The R statistical environment, including the Bioconductor framework, and the following libraries.
 - CMA [4].
 - reshape [5].
 - ConsensusClusterPlus [6].
 - flexmix [2].
 - limma [7].
 - gplots [8].
 - mclust [9].
2. (Optional, but highly recommended) An integrated programming environment (IDE) for R, e.g., RStudio, or a programming editor, e.g., GNU Emacs/ESS.
3. (Optional, but highly recommended) A version control system, e.g., git.

2.3 Data

1. Quality-controlled, preprocessed mRNA expression profiling data from microarray or RNASeq experiments (*see* Chapters 18 and 19, both by Kebschull et al. of this volume). For this case study, we utilize a hypothetical dataset:
 - mRNA expression profiles generated using microarrays from clinically “diseased” gingival tissue biopsies.
 - 200 Subjects with periodontitis, 1 sample per subject = 200 samples in total (*see* Note 2).
 - For each subject, a diagnosis of chronic or aggressive periodontitis [10, 11] was assigned by consensus.
 - For each tissue biopsy, there exist clinical and microbiological data.
 - Expression data were quality-controlled, normalized, and batch-corrected (*see* Note 3).

3 Methods

3.1 Use of Supervised Learning Algorithms for the Distinction of Aggressive and Chronic Periodontitis Based on mRNA Expression

1. Preprocessing of data for use with the CMA package.

Use batch-corrected, preprocessed, and normalized data from microarray or sequencing experiments (*see* Chapters 18 and 19, both by Kebschull et al. of this volume). After the steps described in (*see* Chapter 18 by Kebschull et al. of this volume), the data are usually in the form of a large array with thousands of rows for the different features (i.e., genes, transcripts, CpG islands, etc.) and columns for the individual samples.

In this example, we assume that only samples associated with periodontal disease are present (edata_aff, a subset of the edata expression data matrix generated in Chapter 18 of this volume, with only affected samples remaining). In R, we format the data according to the specifications of the CMA R package we intend to use for the supervised analysis.

```
#label with diagnosis
>colnames(edata_aff) <- pheno_aff$Diagnosis
#rotate (samples -> rows, features -> columns)
>edata_aff_rot <- as.data.frame(t(edata_aff))
#generate a factor containing the diagnosis information with two
levels, in this example 'chronic' and 'aggressive'
>diseased <- rownames(edata_aff_rot)
>diseasedY <- as.factor(diseased)
#change back into matrix format, label rows with subject number
>diseasedX <- as.matrix(edata_aff_rot)
>labels <- pheno_aff$Patient
>rownames(diseasedX) <- labels
```

2. Generate training and evaluation sets.

For the evaluation of the performance of a learning algorithm in our dataset, we perform an internal validation procedure (*see* Note 4).

```
#load CMA package
>library(CMA)
#set a random seed - important to keep constant for reproducible
results
>set.seed(651)
#generate learning sets of the same size comprising on average
about 2/3 of the different available samples by bootstrapping
(sampling with replacement) or other methods (see Note 5) for a
high number of iterations, e.g. 1000 different sets (see Notes 2
and 5). This step needs to be adjusted in cases of multiple
samples per subject (see Note 7).
>datboot <- GenerateLearningsets(y=disease dY, method="bootstrap",
niter=1000, ntrain=floor(0.66*length(diseasedY)), strat=TRUE)
```

3. Feature selection.

For each learning set, characteristic features are identified using statistical tests comparing the predefined groups. Here, we use the “moderated” *t*-test implemented in the limma R package that was introduced in Chapter 18 (see Note 6).

```
>varsel <-GeneSelection(X=diseasedX, y=diseasedY,
learningsets=datboot, method="limma")
```

4. Supervised learning analysis.

```
#perform classification of the evaluation sets generated in (b) by
different (see Note 8) classifier algorithms, using the best nb-
gene features identified by the feature selection process in (c)
#these procedures produce warnings from the R system (see Note 9)
>class_svm <- classification(X=diseasedX, y=as.factor(diseasedY),
learningsets=datboot, genesel=varsel, nb-gene=250, tuninglist =
list(grid = list()), classifier=svmCMA, probability=TRUE)
>class_lda <- classification(X=diseasedX, y=as.factor(diseasedY),
learningsets=datboot, classifier = dldaCMA, genesel=varsel,
nbgene=250)
>[...] more classifiers
```

5. Presentation and interpretation of results.

The performance of the classifiers can be assessed by different measures (see Note 10), including the Area under the Receptor Operating Curve (AUC) that plots the false positive by the true positive rates.

The performance data can then be plotted, either for all 1000 iterations (Figs. 1 and 2a), or for a random iteration (Fig. 2b).

```
>auc_svm<-evaluation(class_svm, measure="auc", scheme="iter")
>auc_lda<-evaluation(class_lda, measure="auc", scheme="iter")
#plot the AUC for the different classifiers and the different
iterations
>boxplot(attributes(auc_lda)$score, attributes(auc_svm)$score,
names=c("DLDA", "SVM"), main="AUC")
```

3.2 Identification of Novel Classes of Periodontitis Based on mRNA Expression Profiles Using Unsupervised Clustering

3.2.1 Preprocessing of Data for Use with the flexmix package—As in Subheading 3.1 (1), we use a dataset of expression data from periodontally affected subjects (edata_aff, a matrix of >50,000 features [rows] × 200 samples [columns], with a

correspondent data frame with phenotypical information, pheno_aff). In this unsupervised analysis, the data are not labeled.

```
#set a random seed - important to keep constant for reproducible results
>set.seed(651)
#take top genes used for clustering and number of bootstrap iterations (see
Note 11)
>numbertop <- 5000
```

X- and Y-linked genes can lead to undesired phenomena during clustering and should be removed (*see* Note 12).

```
#get X-/Y-linked genes (example code for Affymetrix arrays)
>x <- hgu133plus2CHR
>mapped_probes <- mappedkeys(x)
>xx <- as.data.frame(x[mapped_probes])
>is.X <- subset(xx, chromosome=="X")
>is.y <- subset(xx, chromosome=="Y")
>sexChr <- rbind(is.X, is.y)
>sexChr <- sexChr[,1]
>rownames(edata_aff) -> allGenes
>overlap <- allGenes %in% sexChr
>edata_aff <- edata_aff[!overlap,]
#take top genes by median absolute deviation
>mads<-apply(edata_aff,1,mad,na.rm=TRUE)
>edata_aff = edata_aff[order(mads, decreasing=TRUE) [1:numbertop],]
#scale data
>edata_aff = sweep(edata_aff,1, apply(edata_aff,1,median,na.rm=TRUE))
#combine top genes with probing depth information for each sample
>data <- t(edata_aff)
>data <- cbind(data, patient, ppd)
>data <- as.data.frame(data)
#convert data into 'long' format
>library(reshape)
>data.long <- melt(data, id=c("patient", "ppd"), variable_name="probe")
```

3.2.2 Assess Influence of Phenotypical Variables—Depending on the nature of the samples, there may be phenotypes that primarily drive expression. In our example, the gene expression of gingival tissue biopsies was found to be strongly related to the maximum probing depth associated with the biopsy (Fig. 3).

This information is important, because if uncorrected, the strong influence of probing depth would have led to a separation of deep from shallow lesions by the clustering algorithm Fig. 4.

```
> ppd <- pheno_aff$PDMax
# color code samples according to their associated maximum probing depth
> colors <- as.character(factor(ppd, 4:12, c("green4", "green3", "green2",
"orchid", "orchid", "orchid1", "orchid2", "orchid3", "orchid4")))
# MDS plot of individual biopsies using all genes, color-coded for probing
depth plotMDS(edata_aff, gene.selection="common", labels=ppd, col=colors)
# Perform exploratory clustering analysis without accounting for other
factors, e.g. by consensus clustering [6], a resampling approach to standard
hierarchical clustering (see Fig. 4). The resulting clusters can then be
assessed for differences in main phenotypical variables
> library(ConsensusClusterPlus)
> results <- ConsensusClusterPlus(edata_ aff,maxK=10,
reps=1000,pItem=0.8,pFeature=1, title="exploratory_clustering.pdf",
clusterAlg="hc", distance="pearson",seed=65 1,plot="pdf")
```

3.2.3 Perform Mixture Model-Based Clustering with Correction for Probing

Depth—The flexmix R package performs model-based clustering based on finite mixtures of regressions that allows for the inclusion of fixed and random factors, e.g., the maximum probing depth associated with a gingival biopsy as a fixed factor, and the subject as a random factor which ensures that cluster membership is determined on the subject level (not relevant in this example, since all subjects only contributed a single biopsy). The resulting models for several numbers of clusters can be assessed for the goodness-of-fit using standard measures, such as the Akaike Information Criterion (AIC).

```
> library(flexmix)
> model2_ppd <- flexmix(value ~ probe | patient, model = FLXMRglmfix(fixed =
~ 0+ppd, k = 2), k=2, data=data.long, control = list(verbose = 1, iter.max =
20, minprior = 0))
# extract cluster assignments for the different samples
> model2_ppd@cluster -> cluster2_ppd
```

3.2.4 Phenotypic Analysis of the New Clusters

1. Differences in expression of features and the corresponding biological groups between the novel clusters can be identified using the methodology introduced in Chapter 18 (this volume), i.e., limma [7] for a differential expression analysis and ermineJ [12] for a subsequent ontology analysis.
2. Often, the separation of the obtained clusters is illustrated using a heatmap (Fig. 5), e.g., utilizing heatmap.2 in the gplots R package [8].

```

# extract most informative features from the model and format
> pars <- parameters(model2_ppd)
> ordering <- order(abs(apply(pars[2:101,], 1, diff)), decreasing
= TRUE)
> data_corrected <- as.matrix(data[, 1:100] - data$ppd * pars[1,
1])
> ordered <- data_corrected[,ordering][, 1:100]
> ordered <- ordered[order(data$cluster),]
# add sideline to the heatmap to illustrate the different clusters
in red and green.
> sideline <- rep(c("red", "green"), table(data$cluster))
# load the gplots package and plot heatmap
> library(gplots)
> heatmap.2(t(ordered), Colv=FALSE, dendrogram="none",
trace="none", col=redgreen, ColSideColors=sideline)

```

3. To compare cluster assignments, e.g., the similarity of the novel classes identified by the unsupervised analysis and “traditional” groupings, measures like the Hubert-Arabie adjusted Rand index [13] (ranging from 0 indicating entirely random overlaps, and 1 indicating perfect agreement) can be used (Fig. 6).

```

# load the mclust [9] library and compare the novel, mixture model
clustering based classes and the 1999 classification
> library(mclust)
> clusterComp <- adjustedRandIndex(cluster2_ppd, pheno_aff
$Diagnosis)

```

4. Finally, it is of high interest to assess whether the novel classes identified by mixture model-based clustering would also display disparate phenotypes. To do so, standard parametrical or nonparametrical tests can be employed.

```

# t-tests
pval <- t.test(NoOfDeepSites ~ cluster2_ppd, data=clusters, var.equal
= TRUE, na.action="na.omit")
# contingency tables and Fisher test
> gendertable <- xtabs(~ cluster2_ppd + Gender, data=clusters)
> genderfisher <- fisher.test(gendertable)

```


4 Notes

1. When running Windows and/or a graphical user interface, several methods of script optimization to use multiple computing cores of the CPU(s) may not be available.
2. Multiple samples from the same subject may deliver more stable insights into the pathobiology of disease than single samples [14], but also present with several challenges from a statistical perspective: Most algorithms for supervised (e.g., CMA [4]) and unsupervised learning (e.g., ConsensusClusterPlus [6]) do not account for multiple statistically dependent samples and therefore cannot be utilized for the analysis of such datasets without modification. Specifically, in the case of classifier algorithms, care must be taken to avoid training an algorithm using one sample from a particular individual, and then evaluating the classifier using another sample from the same individual. Since multiple samples from the same individual are correlated, this would lead to artificially high performance of the classifier. Similarly, in unsupervised clustering analyses that aim to identify classes of subjects based on –omics profiles of multiple samples, methods such as model-based clustering need to be employed to correct for the correlation of the samples, and possibly also for their inherent characteristics. By no means should multiple samples be coerced into a single sample (e.g., by averaging of the expression values) to facilitate the analysis.
3. Batch effects introduced by different chemistries, operators, but also by normal day-to-day variation are of particular concern for machine learning applications, because samples within a particular batch correlate with each other irrespective of their underlying biology and may lead to bias [15]. In Chapter 18, we describe a workflow to (a) visually assess data for evidence of batch effects, e.g., by multidimensional scaling plots, and (b) correct for these effects using the ComBat procedure in the sva R package [16].
4. Ideally, the performance of a classifier algorithm trained on a labeled dataset should be assessed using a second, independent evaluation set. For example, this external validation can involve a different cohort of subjects. This way, the true generalizability of the learner can be assessed and the inherent danger of overfitting a classifier that is highly specific only to one particular dataset can be mitigated. In case independent datasets are not available, an internal validation is performed by splitting the dataset into pairs of learning and evaluation sets by procedures such as bootstrapping or other cross-validation methods (*see* Note 5). The learner is then trained using the learning set, and evaluated using the corresponding evaluation set. The performance of the algorithm for that particular pair of learning and evaluation sets is then recorded. The average of a high number of iterations (e.g., 1000) reduces the variance of the error estimator, and thereby over-optimism, a typical problem of not-well controlled machine learning applications in –omics research [17].

5. The GenerateLearningsets procedure offers several different options for the generation of splits into learning and evaluation sets, including the bootstrap that was used in our example, but also k-fold cross-validation, Monte-Carlo cross-validation, and k-fold cross-validation. The strat = TRUE argument ensures a similar distribution of the classes in the original dataset and the obtained splits.
6. For the selection of features, the CMA package offers a variety of different methods. First, there are filter methods comparing the predefined classes in the learning set using statistical testing to rank differentially expressed features. The limma method [18] used in our examples falls in this category. Second, methods such as the random forest variable importance measure rank features according to their discriminatory potential. Third, methods such as elastic nets or the lasso exist that are classification algorithms themselves selecting sparse sets of variables that can also be used by other algorithms.
7. As indicated in Note 2, CMA does not account for multiple samples. Therefore, if planning to use this package for the supervised classification analysis of data with multiple, statistically dependent samples, as our group did in our paper testing the 1999 classification of periodontitis [1], the datboot object generated using GenerateLearningsets needs to be modified. Specifically, we randomly sampled subjects, rather than individual samples, and assigned all samples belonging to a drawn individual to the learning sets. In so doing, we avoided to test an algorithm on a sample from an individual that had already had provided a sample for the learning phase, as outlined in Note 2.
8. In a class validation situation, to obtain reliable information about the distinction between two or more classes, it is highly recommended to perform classification using several different algorithms. To avoid bias, sound scientific practice mandates the reporting of all results obtained, rather than only the best one [19].
9. Note that these function calls result in warnings when using the GenerateLearningsets option bootstrap—one warning about duplicate entries for each iteration of the learning set generation (which is perfectly normal for a sampling procedure with replacement).
10. In addition to the area-under-the-ROC-curve (AUC), several other measures for the classifier performance can be obtained using the evaluation function. These include sensitivity and specificity (for $K = 2$, as in our example). In addition, and also in the case of more than two classes, the Brier score, the mis-classification rate, and the average probability of correct classification can be employed.
11. The optimal choice of features used for the unsupervised clustering is subject to an ongoing debate. Generally, by selecting a low proportion of available features one may overlook critical biological information that could have informed the unsupervised analysis. In contrast, when using a very high proportion or all features present, the considerable noise may lead to suboptimal results. It is, however, not entirely clear what numbers or proportions are “too few” or “too many.” Therefore, it is advisable to assess the stability of the clusters obtained

using a range of feature numbers. Relatively stable cluster assignments, as depicted in Fig. 7, support the notion of robust novel classes.

12. Often, the features used for clustering are ranked using a measure of variation within the dataset, with those features showing the highest variation being interpreted as the likely most informative. However, in a dataset that comprises samples from both male and female subjects, those highly variable features include genes with links to sex chromosomes. Depending on the issue under investigation, it may be preferable to exclude these features from the analysis.

Acknowledgments

This work was supported by grants from the German Society for Periodontology (DG PARO) and the German Society for Oral and Maxillo-Facial Sciences (DGZMK) to M.K., and by grants from NIH/NIDCR (DE015649 and DE024735) and by an unrestricted gift from Colgate-Palmolive Inc. to P.N.P. The authors thank Prof. Anne-Laure Boulesteix (Munich, Germany) and Prof. Bettina Grün (Linz, Austria) for their support with the CMA and flexmix packages, respectively.

References

1. Kebschull M, Guarnieri P, Demmer RT, Boulesteix AL, Pavlidis P, Papapanou PN (2013) Molecular differences between chronic and aggressive periodontitis. *J Dent Res* 92:1081–1088 [PubMed: 24122488]
2. Grün B, Leisch F (2008) FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *J Stat Softw* 28:1–35 [PubMed: 27774042]
3. Kebschull M, Demmer RT, Grün B, Guarnieri P, Pavlidis P, Papapanou PN (2014) Gingival tissue transcriptomes identify distinct periodontitis phenotypes. *J Dent Res* 93:459–468 [PubMed: 24646639]
4. Slawski M, Daumer M, Boulesteix AL (2008) CMA: a comprehensive bioconductor package for supervised classification with high dimensional data. *BMC Bioinformatics* 9:439 [PubMed: 18925941]
5. Wickham H (2007) Reshaping data with the reshape package. *J Stat Software* 21:1–20
6. Wilkerson MD, Hayes DN (2010) ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 26:1572–1573 [PubMed: 20427518]
7. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43:e47 [PubMed: 25605792]
8. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B (2009) gplots: various R programming tools for plotting data. *R Package Version* 2(4)
9. Fraley C, Raftery AE, Murphy TB, Scrucca L (2012) MCLUST version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report no. 597, Department of Statistics, University of Washington, USA
10. Armitage GC (1999) Development of a classification system for periodontal diseases and conditions. *Ann Periodontol* 4:1–6 [PubMed: 10863370]
11. Armitage GC, Cullinan MP (2010) Comparison of the clinical features of chronic and aggressive periodontitis. *Periodontol* 2000 53:12–27
12. Gillis J, Mistry M, Pavlidis P (2010) Gene function analysis in complex data sets using ErmineJ. *Nat Protoc* 5:1148–1159 [PubMed: 20539290]
13. Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2:193–218
14. Papapanou PN, Abon A, Verbitsky M, Picoles D, Yang J, Qin J, Fine JB, Pavlidis P (2004) Gene expression signatures in chronic and aggressive periodontitis: a pilot study. *Eur J Oral Sci* 112:216–223 [PubMed: 15154918]

15. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 11:733–739 [PubMed: 20838408]
16. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28:882–883 [PubMed: 22257669]
17. Boulesteix AL (2010) Over-optimism in bioinformatics research. *Bioinformatics* 26:437–439 [PubMed: 19942585]
18. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:Article3
19. Boulesteix AL, Strobl C (2009) Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. *BMC Med Res Methodol* 9:85 [PubMed: 20025773]

Fig. 1.

Microarray-based classification of AP and CP gingival lesions. Four different microarray classifier algorithms were trained to distinguish gingival lesions from AP or CP patients based on their whole-transcriptome expression profiles. For each of the 1000 splittings into training/evaluation sets that accounted for multiple tissue samples per participant, variable selection was performed based on the training set using a mixed-effects linear model. Subsequently, four different classifier algorithms [diagonal linear discriminant analysis (DLDA), partial least square analysis combined with linear discriminant analysis (PLS-LDA), shrunken centroids discriminant analysis (scDA), or a support vector machines (SVM)] were trained on the training set to distinguish between AP and CP gingival lesions based on 250 genes (DLDA, PLS-LDA, and SVM). Performance of the algorithms in the classification of the corresponding evaluation sets was then assessed using the sensitivity and specificity of AP detection, as well as (ROC) area-under-the-curve. With permission from Sage Publishing, reprinted from [1]

Fig. 2. Microarray classifier distinction of gingival lesions from AP and CP–SVM algorithm using different feature set sizes. For each of the 1000 splittings into training/evaluation sets, a support vector machine (SVM) classifier algorithm was trained based on the training set to distinguish AP from CP gingival lesions using either 5, 10, 50, 100, 250, 500, 750, 1000, 2500, or 5000 genes. Performance of the algorithms in the classification of the corresponding evaluation datasets was then assessed using the sensitivity and specificity of AP detection, as well as (ROC) area-under-the-curve (a). In addition, for each number of features, a ROC curve was generated for a representative iteration (b). The SVM algorithm showed improving performance with increasing signature size. With permission from Sage Publishing, reprinted from [1]

Fig. 3. MDS plot of gene expression profiles of individual tissue samples according to probing depth. Multidimensional scaling plot of transcriptomic profiles from 241 tissue samples, based on all autosomal probes on the array. Each individual tissue sample was labeled with the maximal pocket depth associated with the particular biopsy. Samples with PD of 4–6 mm are coded in different shades of green, and samples with of 8–12 mm in different shades of purple. Note that most biopsies aggregate according to their local disease severity, with most shallow pockets on the right side of the plot, and most deep sites on the left. With permission from Sage Publishing, reprinted from [3]

Fig. 4.

Consensus Clustering ($k = 2$) of individual biopsies. Gene expression profiles from 241 periodontitis-affected tissue samples from 120 patients with periodontitis were subjected to consensus clustering, based on the 5000 most variable probes across the entire dataset. Consensus clustering was run for 1000 iterations with a resampling rate of 80 %. This exploratory analysis disregarded the interdependency of samples from the same donor and was meant to identify phenotypic features of the individual tissue samples that determined clustering allocation. With permission from Sage Publishing, reprinted from [3]

Fig. 5.

Heatmap of cluster-defining genes. The cluster-defining genes were sorted by their importance for the model, and the preprocessed expression data were corrected for the probing depth of the individual sample. Data were visualized using the `heatmap.2` function in the `gplots` package v.2.11.3 [8]. Note the clear separation of gene expression between the two clusters. With permission from Sage Publishing, reprinted from [3]

Fig. 6. Stability of cluster assignments from the model-based clustering approach using finite mixtures over a wide range of feature numbers. Cluster assignments obtained with model-based clustering of 241 transcriptomic profiles of periodontitis-affected gingival tissue samples from 120 patients, using different numbers of features. All autosomal probes on the microarray were sorted by absolute variance across the whole dataset, and the top 100–53,243 probes were employed by the clustering algorithm. The graph shows robust clustering, with most patients assigned to the same cluster in all situations. Cluster #1 is represented in blue color, Cluster #2 in red. When using small (<1000 features) or very large (>10,000 features) sets, some “promiscuous” patients change clusters. This behavior is expected, since very small sets tend to lack all information required for correct cluster assignment, and very large sets add a considerable amount of nonspecific noise. With permission from Sage Publishing, reprinted from [3]

Fig. 7.

Comparison of cluster assignment. Transcriptomic profile-based cluster assignments of study participants (panel a) are compared with the current primary classification of periodontitis [chronic (CP) or aggressive (AgP); (panel b)], as well as with an extent-based subdivision based on the 1999 International Workshop criteria [localized or generalized periodontitis; (panel c)]. The degree of concordance between the three ways of classifying periodontitis was assessed using the Herbert-Arabie adjusted Rand index (ARI). Perfect alignment is indicated by an ARI of 1, random alignment by an ARI of 0. Each column across panels a–c represents a patient. With permission from Sage Publishing, reprinted from [3]