M. Kebschull^{1†}, R.T. Demmer²,
B. Grün³, P. Guarnieri^{4,5}, P. Pavlidis⁶,
and P.N. Papapanou^{1*}

¹Division of Periodontics, Section of Oral and Diagnostic Sciences, College of Dental Medicine, Columbia University, New York, NY, USA; ²Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA; ³Department of Applied Statistics, University of Linz, Linz, Austria; ⁴Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY, USA; ⁵Department of Systems Biology, Columbia University, New York, NY, USA; ⁶UBC Centre for High-Throughput Biology, Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada; and [†]*Present address*: Department of Periodontology, Operative and Preventive Dentistry, University of Bonn, Bonn, Germany; *corresponding author, pp192@cumc.columbia.edu

J Dent Res 93(5):459-468, 2014

ABSTRACT

The currently recognized principal forms of periodontitis—chronic and aggressive—lack an unequivocal, pathobiology-based foundation. We explored whether gingival tissue transcriptomes can serve as the basis for an alternative classification of periodontitis. We used cross-sectional whole-genome gene expression data from 241 gingival tissue biopsies obtained from sites with periodontal pathology in 120 systemically healthy nonsmokers with periodontitis, with available data on clinical periodontal status, subgingival microbial profiles, and serum IgG antibodies to periodontal microbiota. Adjusted model-based clustering of transcriptomic data using finite mixtures generated two distinct clusters of patients that did not align with the current classification of chronic and aggressive periodontitis. Differential expression profiles primarily related to cell proliferation in cluster 1 and to lymphocyte activation and unfolded protein responses in cluster 2. Patients in the two clusters did not differ with respect to age but presented with distinct phenotypes (statistically significantly different whole-mouth clinical measures of extent/severity, subgingival microbial burden by several species, and selected serum antibody responses). Patients in cluster 2 showed more extensive/severe disease and were more often male. The findings suggest that distinct gene expression signatures in pathologic gingival tissues translate into phenotypic differences and can provide a basis for a novel classification.

KEY WORDS: gene expression, genomics, cluster analysis, classification, diagnosis, periodontal diseases.

DOI: 10.1177/0022034514527288

Received December 8, 2013; Last revision February 1, 2014; Accepted February 3, 2014

A supplemental appendix to this article is published electronically only at <http://jdr.sagepub.com/supplemental>.

© International & American Associations for Dental Research

Gingival Tissue Transcriptomes Identify Distinct Periodontitis Phenotypes

INTRODUCTION

There are few readily discernible phenotypic differences between the two currently recognized principal forms of periodontitis: chronic (CP) and aggressive (AgP) (“International Workshop for the Classification of Periodontal Diseases and Conditions,” 1999). Early onset of the disease and rapid progression rate that leads to extensive and severe tissue destruction are more frequently associated with AgP than CP, without being pathognomonic for either form. The resulting diagnostic dilemma based on clinically identifiable traits (Armitage and Cullinan, 2010) is further sustained by the presence of common microbiological, immunologic, and histopathologic features of the two entities (Armitage, 2010; Ford *et al.*, 2010; Smith *et al.*, 2010).

To further explore potential differences in the pathophysiology of the two forms, we recently carried out a “class validation” analysis by studying whole-genome transcriptomic profiles in CP and AgP gingival lesions (Kebschull *et al.*, 2013). Although our data demonstrated differential enrichment in a number of gene ontology pathways, the highly variable performance of the supervised machine learning algorithms used, as well as the need to account for the expression of thousands of genes to effectively distinguish between CP and AgP, further support the notion that the current classification has an imperfect biological foundation.

In this study, we hypothesized that gene expression profiles in gingival tissues obtained from patients with periodontitis can form the basis for an alternative classification of the disease. To address this hypothesis, we employed the same data set as in our previous work (Kebschull *et al.*, 2013) in a “class discovery” analytical approach and used clustering, an unsupervised learning method, to identify *de novo* groups of periodontitis patients with common gene expression features. Subsequently, we examined the periodontitis-related phenotypes of the emerging clusters of patients, with respect to clinical characteristics, subgingival bacterial profiles, and serum IgG antibody levels to periodontal microbiota.

MATERIALS & METHODS

The design and procedures of the study were approved by the Columbia University Institutional Review Board (AAAB0896). The study complies with the STROBE guidelines (see <http://stroke-statement.org/>).

We analyzed gene expression data generated with Affymetrix HG-U133Plus 2.0 microarrays (GEO accession No. GSE16134) from 241 gingival tissue biopsies obtained from sites with periodontal pathology (bleeding on probing, probing pocket depth [PPD] ≥ 4 mm, and clinical attachment loss ≥ 3 mm), harvested from systemically healthy patients with periodontitis ($n = 65$ with CP; $n = 55$ with AgP), as previously described (Demmer *et al.*, 2008; Kebschull and Papapanou, 2010; Kebschull *et al.*, 2013). The characteristics of the study sample are described in Appendix Table 1, and those of the individual tissue biopsies in Appendix Table 2. Excellent correlation of array expression data and confirmatory quantitative polymerase chain reaction was shown previously (Papapanou *et al.*, 2009).

Model-based Clustering Using Finite Mixtures

Quality-controlled, robust multiarray algorithm–normalized data were corrected for batch effects with ComBat (Johnson *et al.*, 2007) and analyzed through unsupervised learning algorithms. At first, an exploratory analysis was performed according to consensus clustering of individual biopsies (see Appendix Materials & Methods) to identify major determinants of cluster assignment. Based on these results (Appendix Results), we proceeded with a finite mixture of regression models with varying and fixed regression coefficients using the flexmix package v. 2.3-1 (Grün and Leisch, 2008). The PPD of the individual biopsy was modeled as a fixed effect, while the study participant was used as a grouping variable, i.e., cluster membership was assumed to be the same for all observations from the same study participant. A model with $k = 2$ clusters was fitted based on the evaluation of the goodness of fit, assessed by the Bayesian information criterion (for k ranging from 1 to 10).

Clustering was based on a list of features (i.e., probes) ranked by overall total variance across the data set. To preclude sex chromosome–linked probes from emerging as the principal determinants of the clustering, these probes were excluded from the analysis.

Stability of the clustering was assessed through two approaches. First, we compared the cluster assignments of participants after using varying feature numbers (100 to 53,423; i.e., up to all autosomal probes on the array). Second, after having selected 5000 features for further analyses, we performed 200 iterations of bootstrapping and assessed the percentage of iterations where a participant was assigned to the same cluster as when using the full data set.

Alignment of the identified disease clusters with the current classification scheme was tested by calculating the Hubert-Arabie adjusted Rand index (Hubert and Arabie, 1985) using the mclust (Fraley *et al.*, 2012) package v. 4.2, and visualized using the heatmap.2 function in the gplots package v. 2.11.3 (Warnes *et al.*, 2009).

Intercluster Gene Expression Profiles

The differential expression of probes between the clusters was visualized with heatmap.2 and assessed with limma v. 3.16.8 (Smyth, 2004; Smyth *et al.*, 2005). In these analyses, the patient

was modeled as a random factor to account for the multiple dependent gingival tissue biopsies harvested from each participant. We accounted for multiple comparisons using both false discovery rate (Benjamini and Hochberg, 1995) and Bonferroni corrections.

The functional relevance of the obtained differentially expressed probes was investigated with ermineJ v. 3.0 (Gillis *et al.*, 2010).

Phenotype Analyses

The obtained patient clusters were subsequently examined with respect to clinical, microbiological, and serologic characteristics. Specifically, we compared full-mouth clinical measures of periodontitis, subgingival microbial burden, serum antibodies to periodontal species, and “infection ratios” (Picolos *et al.*, 2005) via unpaired 2-tailed t tests assuming equal variances or Fisher exact tests in R. More detailed information is provided in the Appendix.

RESULTS

Cluster Identification

Exploratory cluster analyses of transcriptomic profiles of the individual biopsies showed that samples with similar PPD tended to cluster together (see Appendix Results, Appendix Figure 2, Supplemental Table 1, Appendix Table 3). Using model-based clustering with correction for the patient as a grouping variable and the pocket depth as a fixed effect, we identified two clusters that were fairly stable with respect to patient allocation over a wide range of feature numbers employed (Appendix Figure 3). For further analysis, we selected the 5000 most variable probes, resulting in allocation of the 120 patients in two clusters of relatively similar size. Thus, 68 patients (57%) formed cluster 1, and 52 (43%) formed cluster 2 (Appendix Figure 4). The identified clusters did not align with the current classification of CP and AgP (adjusted Rand index = 0.0143) or their extent-based subdivisions (Table 3) nor with an extent-based grouping irrespective of primary diagnosis (localized vs. generalized periodontitis, adjusted Rand index = 0.0095; Figure 1). The identified clusters were highly stable over 200 iterations of bootstrapping, with an average of 98.5% of all cluster assignments being identical (range = 73.5%-100%).

Gene Expression Profiles and Functional Groups Between Clusters

A total of 2,837 probes were differentially expressed between the clusters (false discovery rate < 0.05 , $\log_2\text{fc} > 0.25$), with fold changes up to 2.48-fold (Table 1, Supplemental Table 2a, Appendix Figure 5).

One hundred seventy-four functional groups were significantly enriched (corrected $p < .05$) in cluster 1 and 196 groups in cluster 2 (Table 2, Supplemental Table 2b, 2c). These included groups related to cellular metabolism and proliferation (notably proliferation of keratinocytes in cluster 1) and groups

with roles in the endoplasmic reticulum stress pathway, lymphocyte proliferation, and immune activation in cluster 2.

Identified Clusters Display Distinct Phenotypes

The two clusters displayed distinct clinical phenotypes (Table 3). Patients in both clusters were of similar age (39.6 vs. 40.1; $p = .8129$), but males were overrepresented in cluster 2 (65.4% vs. 39.7%; $p = .0061$). Cluster 2 patients showed periodontitis of higher extent and severity than their cluster 1 counterparts, as indicated by statistically significantly higher mean PPD (4.3 vs. 3.8 mm; $p = .0004$), higher mean clinical attachment loss (4.8 vs. 4.1 mm; $p = .0006$), and higher mean number of sites with pocket depth ≥ 5 mm (70.4 vs. 51.3; $p = .0002$) or clinical attachment loss ≥ 5 mm (71.0 vs. 53.0; $p = .0027$). In addition, patients in cluster 2 had a higher frequency of sites that bled on probing (75.4 vs. 64.2%; $p = .0013$), despite comparable levels of plaque (92.6 vs. 89.9%; $p = .077$).

With respect to bacterial colonization levels, patients in cluster 2 were heavier colonized by most of the assessed “red” and “orange complex” species, though not by *Aggregatibacter actinomycetemcomitans*. Conversely, colonization level by the health-associated *Actinomyces naeslundii* was significantly higher in cluster 1.

With respect to serum IgG, patients in cluster 2 showed approximately fivefold higher mean antibody levels to *Camphylobacter rectus* (585.9 vs. 112.1 $\mu\text{g/mL}$; $p = .0001$), higher IgG levels to *Treponema denticola* and *Actinomyces naeslundii*, and a tendency for higher IgG levels to *P. gingivalis* ($p = .0557$). Last, when the level of responsiveness to periodontal microbiota was assessed via the ratio of antibody titer over the mean level of colonization by the homologous species, cluster 2 patients showed significantly higher “infection ratios” against *C. rectus* and *A. naeslundii*.

DISCUSSION

To our knowledge, this is the first attempt to develop an alternative classification of periodontitis based on molecular profiles in periodontitis-affected gingival tissues. We used an unsupervised machine learning approach to mine whole-genome gene expression data from a well-phenotyped cohort of patients with either CP or AgP, and we identified two *de novo* clusters with high similarity in transcriptional profiles. Importantly, the clusters did not align with the currently accepted periodontitis classification, but their distinct molecular signatures translated into

Figure 1. Comparison of cluster assignment. Transcriptomic profile-based cluster assignments of study participants are compared with the current primary classification of periodontitis—chronic (CP) or aggressive (AgP)—as well as with an extent-based subdivision based on the 1999 “International Workshop” criteria (localized or generalized periodontitis). The degree of concordance among the alternative ways of classifying periodontitis was assessed via the Herbert-Arabie adjusted Rand index. Perfect alignment is indicated by an adjusted Rand index of 1; random alignment by an adjusted Rand index of 0.

discernible periodontitis-related characteristics: the two groups differed significantly with respect to extent and severity of periodontitis, level of subgingival colonization, and intensity of serum antibody responses to periodontal microbiota. These data suggest that molecular profiling of gingival tissues can indeed form the basis for the development of an alternative, pathobiology-based classification of periodontitis that correlates well with phenotypic features of the disease.

The study of tissue molecular profiles has led to significant insights in disease pathophysiology in several fields of medicine, most notably in oncology, where molecular profile-based classification of tumors showing different treatment response and prognosis has eventually translated into therapeutic decisions in clinical practice (Witt *et al.*, 2011; Sadanandam *et al.*, 2013). A similar approach in the pathobiology and classification of human periodontitis has not been adopted to this date. It is important to emphasize that to extract sound, biologically meaningful data from transcriptional databases such as the one developed by our group, a clustering algorithm that can appropriately account for the underlying structure of the actual data must be utilized. Our data set originated from 120 patients, most of whom contributed multiple periodontitis-affected gingival tissue samples. Consequently, the whole-genome expression profiles obtained from individual biopsies within the same patient were not independent from each other. Therefore, our clustering analyses could not be based on commonly used algorithms such as hierarchical or *k*-means clustering, since these techniques do

Table 1. Differentially Expressed Genes Between the Two Clusters

Symbol	Name	FC	<i>p</i>	BH Adjusted <i>p</i>
(a) Overexpressed probes in Cluster 1 vs. 2 ^a				
XIST	X inactive specific transcript (nonprotein coding)	2.48	3.86E-04	1.62E-03
KRT2	Keratin 2	1.90	3.14E-07	3.09E-06
CYP3A5	Cytochrome P450, family 3, subfamily A, polypeptide 5	1.90	1.06E-22	3.93E-20
EXPH5	Exophilin 5	1.87	2.89E-24	1.74E-21
KRT1	Keratin 1	1.86	2.02E-07	2.10E-06
DSC1	Desmocollin 1	1.79	9.91E-10	1.80E-08
CHL1	Cell adhesion molecule with homology to L1CAM (close homolog of L1)	1.71	1.10E-16	9.03E-15
LOR	Loricrin	1.70	1.16E-05	7.57E-05
AADACL2	Arylacetamide deacetylase-like 2	1.69	2.82E-09	4.62E-08
KRT3	Keratin 3	1.69	6.86E-12	2.00E-10
MAMDC2	MAM domain containing 2	1.66	1.96E-14	1.00E-12
FLG	Filaggrin	1.66	4.00E-08	4.95E-07
SLC27A6	Solute carrier family 27 (fatty acid transporter), member 6	1.66	1.86E-17	1.84E-15
VAV3	Vav 3 guanine nucleotide exchange factor	1.66	4.98E-16	3.59E-14
TNFRSF19	Tumor necrosis factor receptor superfamily, member 19	1.65	2.03E-17	1.99E-15
PCDH7	Protocadherin 7	1.63	1.83E-17	1.82E-15
FOXP2	Forkhead box P2	1.62	1.16E-18	1.58E-16
SOX6	SRY (sex determining region Y)-box 6	1.62	2.49E-19	3.99E-17
TYRP1	Tyrosinase-related protein 1	1.61	1.87E-04	8.61E-04
HPGD	Hydroxyprostaglandin dehydrogenase 15-(NAD)	1.61	4.77E-17	4.24E-15
PPM1L	Protein phosphatase, Mg2+/Mn2+ dependent, 1L	1.60	2.91E-19	4.57E-17
ABI3BP	ABI family, member 3 (NESH) binding protein	1.60	1.21E-14	6.46E-13
ELOVL4	ELOVL fatty acid elongase 4	1.59	3.52E-17	3.23E-15
MAP2	Microtubule-associated protein 2	1.59	1.78E-17	1.78E-15
CORIN	Corin, serine peptidase	1.57	4.63E-12	1.41E-10
LINC00478	Long intergenic nonprotein coding RNA 478	1.57	4.56E-20	8.83E-18
HLF	Hepatic leukemia factor	1.56	8.70E-18	9.30E-16
ANXA1	Annexin A1	1.55	1.13E-08	1.60E-07
NPR3	Natriuretic peptide receptor C/guanylate cyclase C (atrionatriuretic peptide receptor C)	1.55	7.82E-18	8.47E-16
DSG1	Desmoglein 1	1.53	4.00E-11	9.83E-10
CAPNS2	Calpain, small subunit 2	1.53	1.95E-19	3.21E-17
C6orf132	Chromosome 6 open reading frame 132	1.53	1.01E-12	3.53E-11
NEBL	Nebulette	1.53	2.96E-23	1.32E-20
NEFL	Neurofilament, light polypeptide	1.53	2.64E-08	3.43E-07
GPR27	G protein-coupled receptor 27	1.52	9.88E-17	8.17E-15
KRT15	Keratin 15	1.52	1.01E-14	5.48E-13
FAM83C	Family with sequence similarity 83, member C	1.52	4.27E-13	1.61E-11
CXCL14	Chemokine (C-X-C motif) ligand 14	1.52	8.63E-13	3.04E-11
MUC15	Mucin 15, cell surface associated	1.51	3.81E-17	3.47E-15
CALML5	Calmodulin-like 5	1.51	5.18E-08	6.22E-07
RPTN	Repetin	1.50	4.90E-05	2.68E-04
BBOX1	Butyrobetaine (gamma), 2-oxoglutarate dioxygenase (gamma-butyrobetaine hydroxylase) 1	1.50	3.95E-21	1.01E-18
TM4SF1	Transmembrane 4 L six family member 1	1.50	4.14E-10	8.17E-09
ZNF273	Zinc finger protein 273	1.50	1.17E-16	9.53E-15
TRPS1	Trichorhinophalangeal syndrome 1	1.50	1.92E-16	1.51E-14
RORA	RAR-related orphan receptor A	1.50	1.45E-17	1.49E-15
FLG2	Filaggrin family member 2	1.50	2.29E-05	1.38E-04
SLC16A9	Solute carrier family 16, member 9 (monocarboxylic acid transporter 9)	1.50	3.41E-15	2.06E-13
SH3GL3	SH3-domain GRB2-like 3	1.50	1.81E-12	6.00E-11
ARAP2	ArfGAP with RhoGAP domain, ankyrin repeat and PH domain 2	1.49	6.89E-21	1.65E-18
(b) Overexpressed probes in Cluster 2 vs. 1 ^b				
IGKC	Immunoglobulin kappa constant	2.43	1.14E-25	1.19E-22
MZB1	Marginal zone B and B1 cell-specific protein	2.42	1.91E-24	1.24E-21

(continued)

Table 1. (continued)

Symbol	Name	FC	p	BH Adjusted p
ANKRD36BP2	Ankyrin repeat domain 36B pseudogene 2	2.29	1.76E-18	2.25E-16
FCRL5	Fc receptor-like 5	2.27	4.51E-27	6.66E-24
FCRLA	Fc receptor-like A	2.26	1.17E-21	3.34E-19
TNFRSF17	Tumor necrosis factor receptor superfamily, member 17	2.24	9.97E-23	3.71E-20
SPAG4	Sperm associated antigen 4	2.24	5.95E-28	1.48E-24
SLAMF7	SLAM family member 7	2.17	2.12E-23	1.01E-20
FKBP11	FK506 binding protein 11, 19 kDa	2.13	4.17E-29	1.75E-25
IRF4	Interferon regulatory factor 4	2.12	3.96E-24	2.26E-21
IGHM	Immunoglobulin heavy constant mu	2.12	7.17E-16	5.02E-14
CD79A	CD79a molecule, immunoglobulin-associated alpha	2.10	1.52E-27	3.01E-24
LOC96610	BMS1 homolog, ribosome assembly protein (yeast) pseudogene	2.09	5.75E-20	1.08E-17
SEL1L3	Sel-1 suppressor of lin-12-like 3 (<i>C. elegans</i>)	2.07	3.16E-29	1.72E-25
DERL3	Derlin 3	2.04	3.99E-29	1.75E-25
PNOC	Prepronociceptin	1.98	4.30E-32	5.88E-28
IGHD	Immunoglobulin heavy constant delta	1.97	1.67E-17	1.68E-15
CD38	CD38 molecule	1.97	1.66E-27	3.03E-24
POU2AF1	POU class 2 associating factor 1	1.97	1.08E-22	3.98E-20
FAM46C	Family with sequence similarity 46, member C	1.96	1.58E-21	4.29E-19
IGLV1-44	Immunoglobulin lambda variable 1-44	1.96	1.13E-17	1.18E-15
CPNE5	Copine V	1.95	2.12E-32	5.80E-28
IGLC1	Immunoglobulin lambda constant 1 (Mcg marker)	1.94	5.73E-15	3.30E-13
ST6GAL1	ST6 beta-galactosamide alpha-2,6-sialyltransferase 1	1.94	6.14E-18	6.95E-16
KLHL6	Kelch-like family member 6	1.94	5.58E-20	1.06E-17
LAX1	Lymphocyte transmembrane adaptor 1	1.92	5.71E-22	1.77E-19
IGJ3	Immunoglobulin lambda joining 3	1.92	9.39E-21	2.18E-18
KCNN3	Potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	1.90	5.20E-28	1.48E-24
ZBP1	Z-DNA binding protein 1	1.90	1.93E-23	9.27E-21
IGK@	Immunoglobulin kappa locus	1.89	1.21E-20	2.72E-18
EAF2	ELL associated factor 2	1.89	1.24E-22	4.40E-20
IGLL5	Immunoglobulin lambda-like polypeptide 5	1.88	3.56E-18	4.24E-16
CD27	CD27 molecule	1.86	2.85E-23	1.28E-20
CHST2	Carbohydrate (N-acetylglucosamine-6-O) sulfotransferase 2	1.82	1.21E-26	1.61E-23
MEI1	Meiosis inhibitor 1	1.81	1.35E-30	1.05E-26
IGHG1	Immunoglobulin heavy constant gamma 1 (G1m marker)	1.81	2.79E-15	1.72E-13
PPAPDC1B	Phosphatidic acid phosphatase type 2 domain containing 1B	1.81	1.97E-27	3.36E-24
CECR1	Cat eye syndrome chromosome region, candidate 1	1.79	8.35E-28	1.83E-24
HLA-DOB	Major histocompatibility complex, class II, DO beta	1.78	2.10E-29	1.28E-25
HSPA13	Heat shock protein 70kDa family, member 13	1.77	3.66E-22	1.21E-19
RHOH	Ras homolog family member H	1.76	1.65E-14	8.58E-13
CYTIP	Cytohesin 1 interacting protein	1.73	3.56E-20	7.20E-18
NCF4	Neutrophil cytosolic factor 4, 40kDa	1.73	5.39E-26	6.14E-23
XBP1	X-box binding protein 1	1.73	1.38E-24	9.81E-22
RPS4Y1	Ribosomal protein S4, Y-linked 1	1.73	1.79E-03	6.12E-03
PTP4A3	Protein tyrosine phosphatase type IVA, member 3	1.73	1.22E-19	2.10E-17
LINC00525	Long intergenic non-protein coding RNA 525	1.73	1.51E-14	7.92E-13
GLCCI1	Glucocorticoid induced transcript 1	1.73	3.56E-23	1.56E-20
EIF1AY	Eukaryotic translation initiation factor 1A, Y-linked	1.71	4.10E-04	1.71E-03

BH, Benjamini-Hochberg; FC, fold change.

^aTop 50 probe sets overexpressed in periodontitis-affected gingival tissues of cluster 1 vs. cluster 2 patients (false discovery rate < 0.05, absolute fold change ≥ 1.19). A total of 1812 probe sets were significantly differentially expressed at false discovery rate (Benjamini-Hochberg-adjusted *p* value) < .05, and 1574 probe sets met the Bonferroni wide significance threshold (unadjusted *p* value < 9.15E-07). Note that several probes may map to the same gene. In this table, only the data for the probe with the highest fold change are reported. For the complete list of differentially expressed probes and detailed information on the individual probes, see Supplemental Table 2a.

^bTop 50 probe sets overexpressed in periodontitis-affected gingival tissues of cluster 2 vs. cluster 1 patients (false discovery rate < 0.05, absolute fold change ≥ 1.19). A total of 1025 probe sets were significantly differentially expressed at false discovery rate (Benjamini-Hochberg-adjusted *p* value) < 0.05, and 936 probe sets met the Bonferroni genome-wide significance threshold (unadjusted *p* value < 9.15E-07). Note that several probes may map to the same gene. In this table, only the data for the probe with the highest fold change are reported. For the complete of differentially expressed probes and detailed information on the individual probes, see Supplemental Table 2a.

Table 2. Functional Groups Enriched in the Two Clusters

Name	ID	Corrected <i>p</i>
(a) Enriched groups in Cluster 1 vs. 2 ^a		
RNA 3'-end processing	GO:0031123	6.03E-06
mRNA 3'-end processing	GO:0031124	5.96E-06
Keratinocyte differentiation	GO:0030216	4.30E-06
Translational elongation	GO:0006414	5.68E-06
Translational termination	GO:0006415	1.23E-05
Cellular protein complex disassembly	GO:0043624	4.61E-05
Early endosome to late endosome transport	GO:0045022	4.26E-05
Viral transcription	GO:0019083	4.15E-05
Stem cell maintenance	GO:0019827	4.27E-05
Protein ubiquitination involved in ubiquitin-dependent protein catabolic process	GO:0042787	4.03E-05
Asymmetric protein localization	GO:0008105	4.13E-05
Viral genome expression	GO:0019080	5.17E-05
Nuclear export	GO:0051168	1.34E-04
Protein monoubiquitination	GO:0006513	2.64E-04
Cytokinesis	GO:0000910	4.34E-04
Nucleus organization	GO:0006997	5.76E-04
Regulation of microtubule-based process	GO:0032886	5.53E-04
Primary neural tube formation	GO:0014020	7.48E-04
Histone ubiquitination	GO:0016574	7.17E-04
DNA-dependent transcription, termination	GO:0006353	6.81E-04
Membrane disassembly	GO:0030397	7.75E-04
Nuclear envelope disassembly	GO:0051081	7.75E-04
Neural tube closure	GO:0001843	1.10E-03
Neural tube formation	GO:0001841	1.61E-03
DNA-dependent DNA replication	GO:0006261	1.97E-03
Termination of RNA polymerase II transcription	GO:0006369	2.04E-03
Tube closure	GO:0060606	2.15E-03
Mitotic nuclear envelope disassembly	GO:0007077	2.12E-03
RNA polyadenylation	GO:0043631	2.26E-03
Nuclear pore organization	GO:0006999	2.27E-03
Anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	GO:0031145	2.34E-03
Establishment of cell polarity	GO:0030010	2.28E-03
Postreplication repair	GO:0006301	2.26E-03
Nuclear envelope organization	GO:0006998	2.22E-03
RNA export from nucleus	GO:0006405	2.19E-03
Protein modification by small protein removal	GO:0070646	2.30E-03
Apical protein localization	GO:0045176	3.17E-03
Positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	GO:0051437	3.29E-03
Regulation of centrosome cycle	GO:0046605	3.28E-03
DNA geometric change	GO:0032392	3.65E-03
mRNA polyadenylation	GO:0006378	3.66E-03
Regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	GO:0051439	3.88E-03
Lateral sprouting from an epithelium	GO:0060601	4.33E-03
mRNA export from nucleus	GO:0006406	4.99E-03
Deoxyribonucleotide biosynthetic process	GO:0009263	5.88E-03
Regulation of Rac protein signal transduction	GO:0035020	5.89E-03
Regulation of centrosome duplication	GO:0010824	6.32E-03
Histone monoubiquitination	GO:0010390	6.61E-03
DNA biosynthetic process	GO:0071897	6.83E-03
Nuclear import	GO:0051170	7.16E-03
(b) Enriched groups in Cluster 2 vs. 1 ^b		
Protein N-linked glycosylation	GO:0006487	<2.20E-16
Peptidyl-asparagine modification	GO:0018196	<2.20E-16
Protein N-linked glycosylation via asparagine	GO:0018279	<2.20E-16
Activation of signaling protein activity involved in unfolded protein response	GO:0006987	5.02E-13

(continued)

Table 2. (continued)

Name	ID	Corrected <i>p</i>
Cellular response to topologically incorrect protein	GO:0035967	8.86E-13
Cellular response to unfolded protein	GO:0034620	7.85E-13
Endoplasmic reticulum unfolded protein response	GO:0030968	2.06E-12
Positive regulation of nuclease activity	GO:0032075	2.74E-11
ER-nucleus signaling pathway	GO:0006984	5.73E-11
Regulation of nuclease activity	GO:0032069	3.64E-08
Regulation of B cell proliferation	GO:0030888	3.40E-06
B cell receptor signaling pathway	GO:0050853	8.10E-06
Regulation of B cell activation	GO:0050864	4.12E-05
Regulation of leukocyte mediated cytotoxicity	GO:0001910	2.35E-04
Retrograde protein transport, ER to cytosol	GO:0030970	2.60E-04
Regulation of dendritic cell differentiation	GO:2001198	2.59E-04
Positive regulation of mononuclear cell proliferation	GO:0032946	2.90E-04
Regulation of cell killing	GO:0031341	3.00E-04
Positive regulation of lymphocyte proliferation	GO:0050671	4.31E-04
ER-associated protein catabolic process	GO:0030433	5.28E-04
Negative regulation of lymphocyte activation	GO:0051250	7.42E-04
Respiratory burst involved in defense response	GO:0002679	8.07E-04
Regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	GO:0002822	8.16E-04
Protein secretion	GO:0009306	7.91E-04
Mononuclear cell proliferation	GO:0032943	7.93E-04
Regulation of lymphocyte mediated immunity	GO:0002706	9.52E-04
Blood vessel remodeling	GO:0001974	9.43E-04
Lymphocyte proliferation	GO:0046651	2.31E-03
Negative regulation of B cell proliferation	GO:0030889	2.27E-03
Positive regulation of lymphocyte differentiation	GO:0045621	3.04E-03
Regulation of T cell mediated cytotoxicity	GO:0001914	2.97E-03
Regulation of granulocyte differentiation	GO:0030852	2.89E-03
Dolichol-linked oligosaccharide biosynthetic process	GO:0006488	2.94E-03
Negative regulation of mononuclear cell proliferation	GO:0032945	3.29E-03
Negative regulation of lymphocyte proliferation	GO:0050672	3.29E-03
Leukocyte chemotaxis	GO:0030595	3.23E-03
Leukocyte proliferation	GO:0070661	3.47E-03
Positive regulation of leukocyte mediated cytotoxicity	GO:0001912	5.33E-03
Learning	GO:0007612	5.61E-03
Regulation of phospholipase C activity	GO:1900274	5.67E-03
Regulation of alpha-beta T cell activation	GO:0046634	5.57E-03
Positive regulation of B cell proliferation	GO:0030890	5.45E-03
Neutral amino acid transport	GO:0015804	5.35E-03
Immune response—inhibiting cell surface receptor signaling pathway	GO:0002767	5.54E-03
Vesicle targeting, to, from or within Golgi	GO:0048199	5.43E-03
Respiratory burst	GO:0045730	5.62E-03
Positive regulation of B cell activation	GO:0050871	5.61E-03
Neuromuscular process	GO:0050905	5.54E-03
Potassium ion transport	GO:0006813	5.55E-03
Positive regulation of cell killing	GO:0031343	5.71E-03

^aTop 50 functional GO groups. For a complete list and detailed information, see Supplemental Table 2b.

^bTop 50 functional GO groups. For a complete list and detailed, see Supplemental Table 2c.

not account for sample interdependency. Instead, we employed a two-step approach: first, a consensus clustering algorithm based on *k* means in an exploratory analysis to identify tissue sample properties that could influence the clustering of patients

into *de novo* groups (our ultimate goal). Indeed, these analyses revealed that clustering of individual biopsies based on expression profiles was influenced by the severity of periodontitis affecting the particular tissue sample, i.e., the PPD of the pocket

Table 3. Clinical, Microbiological, and Serologic Phenotypes of Patients in the Two Clusters

	Cluster 1	Cluster 2	<i>p</i>
Age	39.6 ± 12.5	40.1 ± 13.2	.8129
Male, %	39.7	65.4	.0061
Hispanic, %	67.2	84.3	.0537
Periodontal diagnosis, %			
Localized chronic periodontitis	38.2	23.1	.1126
Generalized chronic periodontitis	22.1	23.1	1.0000
Localized aggressive periodontitis	14.7	15.4	1.0000
Generalized aggressive periodontitis	25.0	38.5	.1622
Periodontal parameters			
Number of teeth	28.5 ± 2.6	28.1 ± 3.0	.5346
Plaque, %	89.9 ± 8.2	92.6 ± 8.5	.0770
Bleeding-on-probing, %	64.2 ± 17.9	75.4 ± 19.1	.0013
Pocket depth	3.8 ± 0.8	4.3 ± 0.9	.0004
No. of sites with pocket depth ≥ 5 mm	51.3 ± 26.2	70.4 ± 28.3	.0002
Clinical attachment level	4.1 ± 0.9	4.8 ± 1.3	.0006
No. of sites with clinical attachment loss ≥ 5 mm	53.0 ± 30.0	71.0 ± 34.1	.0027
Subgingival bacterial level, cell count × 10 ⁴			
Aa	52.5 ± 74.1	56.5 ± 62.6	.0576
Pg	244.8 ± 332.0	645.1 ± 1287.0	.0001
Td	159.6 ± 260.4	270.0 ± 302.2	.0004
Tf	455.1 ± 485.7	746.3 ± 743.4	.0009
Fn	279.9 ± 192.1	266.2 ± 127.0	.2946
Pi	446.3 ± 509.7	572.7 ± 630.3	.0320
Cr	93.6 ± 206.4	95.8 ± 117.6	.0219
Mm	180.5 ± 177.5	207.8 ± 140.2	.0199
Ec	80.8 ± 118.8	56.5 ± 57.3	.7896
Vp	121.6 ± 202.6	92.2 ± 131.0	.6843
An	539.4 ± 874.0	240.7 ± 193.9	.0087
Serum IgG antibody levels, µg/mL			
Aa	614.4 ± 828.5	1267.2 ± 4125.2	.1527
Pg	1351.5 ± 2278.6	2326.1 ± 3946.5	.0557
Td	102.7 ± 86.3	160.4 ± 133.4	.0059
Tf	97.0 ± 117.4	141.4 ± 180.0	.2480
Fn	63.5 ± 111.0	76.6 ± 78.5	.0872
Pi	507.5 ± 448.2	589.4 ± 485.3	.2015
Cr	112.1 ± 130.1	585.9 ± 1311.2	.0001
Mm	219.6 ± 312.7	171.4 ± 136.9	.6317
Ec	71.6 ± 56.8	77.6 ± 68.4	.7778
Vp	34.4 ± 34.6	61.4 ± 113.7	.0615
An	407.9 ± 657.6	845.0 ± 1143.8	.0324
Infection ratio, µg/mL × 10 ⁴ cells			
Aa	33.8 ± 36.0	28.2 ± 32.6	.8047
Pg	24.3 ± 75.3	6.1 ± 9.2	.8911
Td	2.6 ± 3.8	1.4 ± 2.2	.6054
Tf	0.7 ± 1.3	0.3 ± 0.5	.0647
Fn	0.3 ± 0.3	0.3 ± 0.4	.0995
Pi	2.5 ± 4.0	1.8 ± 2.1	.8253
Cr	5.0 ± 7.2	22.6 ± 105.9	.0033
Mm	2.1 ± 4.5	1.2 ± 1.6	.9557
Ec	2.5 ± 3.0	2.5 ± 2.6	.7870
Vp	1.1 ± 2.5	2.0 ± 6.8	.1107
An	2.1 ± 4.3	5.3 ± 8.0	.0083

Aa, *Aggregatibacter actinomycetemcomitans*; An, *Actinomyces naeslundii*; Cr, *Campylobacter rectus*; Ec, *Eikenella corrodens*; Fn, *Fusobacterium nucleatum*; Mm, *Micromonas micros*; Pg, *Porphyromonas gingivalis*; Pi, *Prevotella intermedia*; Td, *Treponema denticola*; Tf, *Tannerella forsythia*; Vp, *Veillonella parvula*.

Full-mouth clinical, microbiological, and serologic parameters of the patients in the two clusters were compared by 2-tailed unpaired *t* tests assuming equal variances for continuous variables or by Fisher exact test for categorical variables. Statistical testing of bacterial counts and serological data was performed after ln-transformation. Data are presented as mean values ± SD. Bold font indicates *p* < .05.

associated with the gingival biopsy. Consequently, in a second step, our definitive analyses employed a model-based clustering approach using finite mixtures (Grün and Leisch, 2008) that considered the individual patient as a grouping variable and the PPD of the particular gingival tissue sample as a fixed effect. Mixture models have been shown to outperform other clustering algorithms in recovering the true data set structure (de Souto *et al.*, 2008). However, the number of features (probes) employed by the clustering algorithm is critically important: use of a low number of features can result in a situation where few genes with strong variability “drive” the cluster formation, irrespective of their relevance from a pathobiological perspective. Conversely, apart from being computationally demanding, use of a large number of features likely results in infusion of considerable random noise that may obscure the formation of clearly distinguishable clusters. In our final analyses, the selection of the number of features used (5,000) was partly informed by findings from our earlier work that identified differentially expressed genes in gingival tissues between states of periodontal health and disease (Demmer *et al.*, 2008). However, we have also analyzed clustering results based on a range of number of features, to examine consistency in cluster formation. Indeed, as illustrated in Appendix Figure 3, two fairly robust clusters emerged across a number of employed features, suggesting that our final clustering was rather stable. We based our clustering solely on autosomal genes, as incorporation of X-/Y-linked genes would likely result in sex being a major determinant of patient aggregation into different groups. Of note, use of only two clusters resulted in the best fit based on the Bayesian information criterion, although it is conceivable that in larger data sets more clusters could potentially emerge. Since our primary aim was to identify molecular subtypes of periodontitis, our clustering was exclusively based on transcriptomic information from gingival tissue samples showing signs of pathology. It is conceivable that inclusion of healthy gingival tissues into our computations could have led to a better performance of the clustering algorithm, as it would also reflect expression differences between states of periodontal health and disease. However, given the availability of healthy tissue samples from only 69 of the 120 patients, we opted not to follow this approach.

The transcriptomic profiles that distinguished the two clusters are of interest: first, the number of the differentially expressed probes between clusters 1 and 2 was more than 10 times higher than that between AgP and CP lesions in our earlier “class validation” publication (Kebischull *et al.*, 2013). Second, a number of differentially expressed pathways between the two clusters are compatible with the more extensive/severe periodontitis found in cluster 2. Specifically, functional groups related to chemotaxis, proliferation, differentiation, and activation of immune cells—most prominently of B cells—were enriched in cluster 2. These functions are in line with the notion that most periodontal tissue destruction is host-derived and results from the recruitment and activation of immune cells (Hernandez *et al.*, 2011). Indeed, we identified pathways mediating cytotoxicity to be enriched in this cluster, corroborating prior evidence implicating the activation of natural killer cells and

natural killer T cells in periodontal tissue destruction (Chashu *et al.*, 2012; Kramer *et al.*, 2013; Nowak *et al.*, 2013). Furthermore, a strong enrichment of endoplasmic reticulum stress groups was observed in cluster 2. Endoplasmic reticulum stress, also known as the unfolded protein response, is a cellular stress response related to the accumulation of misfolded proteins in the endoplasmic reticulum, eventually leading to the activation of apoptotic cascades. Interestingly, several links between this stress response pathway and immune activation by bacterially-induced pathways have been identified (Claudio *et al.*, 2013). On the other hand, cluster 1 was characterized by enrichment of functions related to cellular maintenance and proliferation, cellular growth, mitosis, or progression of the cell cycle.

Importantly, the two clusters presented with periodontitis-related phenotypic differences. Patients in cluster 2 had higher extent and severity of periodontitis, were more heavily colonized by traditional periodontal pathogens, and presented with a number of differences with respect to antibody responsiveness to selected microbiota. Despite the fact that we actively removed sex chromosome-linked genes from the feature list used in the clustering process, two-thirds of the patients in cluster 2 were male. This skewed sex distribution between the two clusters is in line with the well-established notion of higher periodontitis severity in males (Eke *et al.*, 2012).

Several additional steps are obviously necessary toward the development of an alternative, molecular-based classification of periodontitis. First, examining whether incorporation of post-transcriptional elements into the clustering process, such as miRNA expression profiles (Stoecklin-Wasmer *et al.*, 2012), affects the discrimination between the two clusters will be informative. We also need to develop algorithms based on robust phenotypic surrogates of the transcriptional profiles—for example, by using clinical, microbiological, or immunologic features of periodontitis, which will allow cluster assignment without the need for invasive diagnostic procedures or whole-genome transcriptomic analyses. Ultimately, a tentative alternative periodontitis classification will have to be externally validated in new patient cohorts as well as in longitudinal observational and intervention studies.

ACKNOWLEDGMENTS

The authors thank Anne-Laure Boulesteix, Munich, Germany, for helpful advice. This research was supported by the National Institute of Dental and Craniofacial Research (DE-015649 and DE-021820), the National Center for Advancing Translational Sciences (UL1-TR000040), and Colgate-Palmolive Co, New Jersey, USA. MK was supported by the German Research Foundation (KFO208, TP6 and TP9), the German Society for Periodontology, the German Society for Oral and Maxillofacial Sciences, and the Neue Gruppe; RTD by the National Institutes of Health (R00 DE-018739); BG by the Austrian Science Fund (V170-N18); and PP by the Michael Smith Foundation for Health Research. The authors declare no potential conflicts of interest with respect to the authorship and/or publication of this article.

REFERENCES

- Armitage GC (2010). Comparison of the microbiological features of chronic and aggressive periodontitis. *Periodontol* 2000 53:70-88.
- Armitage GC, Cullinan MP (2010). Comparison of the clinical features of chronic and aggressive periodontitis. *Periodontol* 2000 53:12-27.
- Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc B* 57:289-300.
- Chaushu S, Wilensky A, Gur C, Shapira L, Elboim M, Halftek G, et al. (2012). Direct recognition of *Fusobacterium nucleatum* by the NK cell natural cytotoxicity receptor NKp46 aggravates periodontal disease. *PLoS Pathog* 8:e1002601.
- Claudio N, Dalet A, Gatti E, Pierre P (2013). Mapping the crossroads of immune activation and cellular stress response pathways. *EMBO J* 32:1214-1224.
- de Souto M, Costa I, de Araujo D, Ludermitr T, Schliep A (2008). Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics* 9:497.
- Demmer RT, Behle JH, Wolf DL, Handfield M, Kebschull M, Celenti R, et al. (2008). Transcriptomes in healthy and diseased gingival tissues. *J Periodontol* 79:2112-2124.
- Eke PI, Dye BA, Wei L, Thornton-Evans GO, Genco RJ; CDC Periodontal Disease Surveillance Workgroup (2012). Prevalence of periodontitis in adults in the United States: 2009 and 2010. *J Dent Res* 91:914-920.
- Ford PJ, Gamonal J, Seymour GJ (2010). Immunological differences and similarities between chronic periodontitis and aggressive periodontitis. *Periodontol* 2000 53:111-123.
- 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 597. Seattle, WA: University of Washington, Department of Statistics.
- Gillis J, Mistry M, Pavlidis P (2010). Gene function analysis in complex data sets using ErmineJ. *Nat Protoc* 5:1148-1159.
- Grün B, Leisch F (2008). FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *J Stat Software* 28:1-35.
- Hernandez M, Dutzan N, García-Sesnich J, Abusleme L, Dezerrega A, Silva N, et al. (2011). Host-pathogen interactions in progressive chronic periodontitis. *J Dent Res* 90:1164-1170.
- Hubert L, Arabie P (1985). Comparing partitions. *J Classif* 2:193-218.
- International workshop for a classification of periodontal diseases and conditions 1999, Oak Brook, Illinois. (1999). *Ann Periodontol* 4:1-112.
- Johnson WE, Li C, Rabinovic A (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8:118-127.
- Kebschull M, Papapanou PN (2010). The use of gene arrays in deciphering the pathobiology of periodontal diseases. *Methods Mol Biol* 666:385-393.
- Kebschull M, Guarnieri P, Demmer RT, Boulesteix AL, Pavlidis P, Papapanou PN (2013). Molecular differences between chronic and aggressive periodontitis. *J Dent Res* 92:1081-1089.
- Kramer B, Kebschull M, Nowak M, Demmer RT, Haupt M, Korner C, et al. (2013). Role of the NK cell-activating receptor CRACC in periodontitis. *Infect Immun* 81:690-696.
- Nowak M, Kramer B, Haupt M, Papapanou PN, Kebschull J, Hoffmann P, et al. (2013). Activation of invariant NK T cells in periodontitis lesions. *J Immunol* 190:2282-2291.
- Papapanou PN, Behle JH, Kebschull M, Celenti R, Wolf DL, Handfield M, et al. (2009). Subgingival bacterial colonization profiles correlate with gingival tissue gene expression. *BMC Microbiol* 9:221.
- Picolas DK, Lerche-Sehm J, Abron A, Fine JB, Papapanou PN (2005). Infection patterns in chronic and aggressive periodontitis. *J Clin Periodontol* 32:1055-1061.
- Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, Wulfschlegel S, et al. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat Med* 19:619-625.
- Smith M, Seymour GJ, Cullinan MP (2010). Histopathological features of chronic and aggressive periodontitis. *Periodontol* 2000 53:45-54.
- Smyth GK (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments [published online February 12, 2004]. *Stat Appl Genet Mol Biol* 3:article 3.
- Smyth GK, Michaud J, Scott HS (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* 21:2067-2075.
- Stoecklin-Wasmer C, Guarnieri P, Celenti R, Demmer RT, Kebschull M, Papapanou PN (2012). MicroRNAs and their target genes in gingival tissues. *J Dent Res* 91:934-940.
- Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al. (2009). gplots: various R programming tools for plotting data. *R Package Version* 2(4).
- Witt H, Mack SC, Ryzhova M, Bender S, Sill M, Isserlin R, et al. (2011). Delineation of two clinically and molecularly distinct subgroups of posterior fossa ependymoma. *Cancer Cell* 20:143-157.