

(19) World Intellectual Property Organization
International Bureau(43) International Publication Date
19 August 2010 (19.08.2010)(10) International Publication Number
WO 2010/093379 A1

(51) International Patent Classification:

C12Q 1/68 (2006.01) G01N 33/574 (2006.01)
C12N 15/11 (2006.01)

(21) International Application Number:

PCT/US2009/051743

(22) International Filing Date:

24 July 2009 (24.07.2009)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

61/152,541 13 February 2009 (13.02.2009) US

(71) Applicants (for all designated States except US): **FRED HUTCHINSON CANCER RESEARCH CENTER** [US/US]; 1100 Fairview Avenue North, J5-110, Seattle, WA 98109 (US). **THE UNIVERSITY OF WASHINGTON** [US/US]; 4311 11th Ave NE, Suite 500, Seattle, WA 98105 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **CHEN, Chu** [US/US]; Seattle, WA 98109 (US). **MENDEZ, Eduardo** [—/US]; Seattle, WA 98109 (US). **HOUCK, John** [—/US]; Seattle, WA 98109 (US). **FAN, Wenhong** [—/US]; Seattle, WA 98109 (US). **LOHAVANICH-BUTR, Pawadee** [—/US]; Seattle, WA 98109 (US). **YUEH, Bevan** [—/US]; Minneapolis, MN 55455 (US). **FUTRAN, Neal, D.** [—/US]; Seattle, WA 98195 (US). **SCHWARTZ, Stephen, M.** [—/US]; Seattle, WA 98109 (US). **ZHAO, Lue, Ping** [—/US]; Seattle, WA 98109 (US).(74) Agents: **ROSENMAN, Stephen, J.** et al.; Seed Intellectual Property Law Group PLLC, Suite 5400, 701 Fifth Avenue, Seattle, WA 98104-7064 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD,

[Continued on next page]

(54) Title: GENE EXPRESSION PROFILING IDENTIFIES GENES PREDICTIVE OF ORAL SQUAMOUS CELL CARCINOMA AND ITS PROGNOSIS

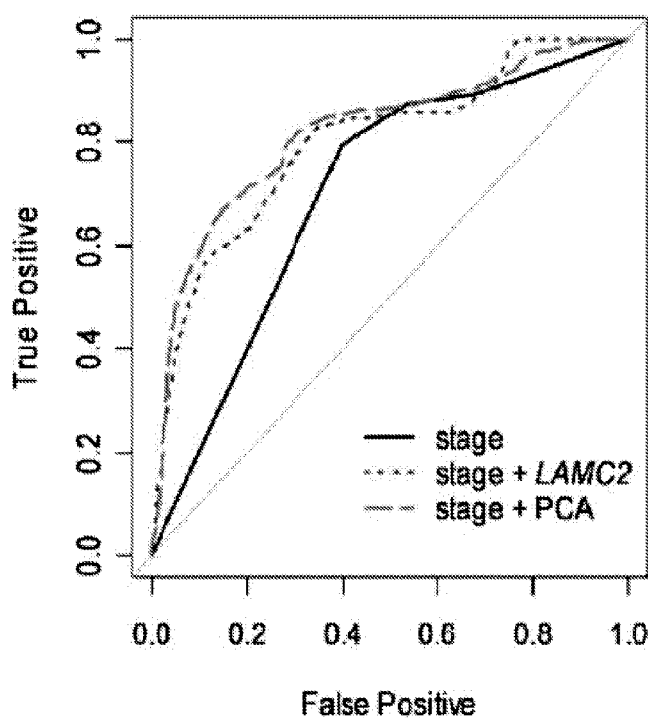


FIG. 13B

(57) Abstract: Methods are provided for determining the presence or risk of squamous cell carcinoma in a subject, including head and neck carcinoma, oral squamous cell carcinoma (OSCC), and pre-neoplastic dysplasia. Specifically, the invention relates to gene expression profiling of selected squamous cell carcinoma biomarker genes or gene indicator sets. In certain embodiments, the differential expression of selected biomarker genes or gene indicator sets in a subject sample as compared to a control sample identifies the presence or risk of squamous cell carcinoma, or an increased risk of oral squamous cell carcinoma (OSCC)-specific mortality in a subject having OSCC. Methods are also provided for differentiating between squamous cell carcinoma cells and pre-neoplastic dysplastic cells in a subject. Selected oligonucleotide probes are also provided for detecting the squamous cell carcinoma biomarker genes or gene indicator sets.



ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI,
NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD,
SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT,
TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,
MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM,
TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
ML, MR, NE, SN, TD, TG).

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE,

Published:

— *with international search report (Art. 21(3))*

— *with sequence listing part of description (Rule 5.2(a))*

GENE EXPRESSION PROFILING IDENTIFIES GENES PREDICTIVE OF ORAL SQUAMOUS CELL CARCINOMA AND ITS PROGNOSIS

CROSS-REFERENCE TO RELATED APPLICATION

This application claims the benefit under 35 U.S.C. § 119(e) of U.S.
5 Provisional Patent Application No. 61/152,541 filed February 13, 2009, where this
provisional application is incorporated herein by reference in its entirety.

STATEMENT OF GOVERNMENT INTEREST

This invention was made in part with government support under
Grant No. R01 CA095419 awarded by the National Cancer Institute/ National
10 Institutes of Health, National Research Service Award T32DC00018 from the
National Institute on Deafness and Other Communication Disorders, and trans-
NIH K12RR023265 Career Development Programs for Clinical Researchers.
The government has certain rights in this invention.

15 STATEMENT REGARDING SEQUENCE LISTING

The Sequence Listing associated with this application is provided in
text format in lieu of a paper copy, and is hereby incorporated by reference into
the specification. The name of the text file containing the Sequence Listing is
360056_402PC_SEQUENCE_LISTING.txt. The text file is 883 KB, was created
20 on July 22, 2009, and is being submitted electronically via EFS-Web.

BACKGROUND

Technical Field

The presently disclosed invention embodiments relate to
compositions and methods for the detection and treatment of cancer. In
25 particular, the present embodiments relate to identifying the presence of, or a risk
for having, squamous cell carcinoma including oral squamous cell carcinoma
(OSCC) and head-and-neck squamous cell carcinoma (HNSCC) in a subject, by

identifying differential expression of one or more squamous cell carcinoma indicator genes as described herein.

Description of the Related Art

Squamous cell carcinoma of the oral cavity and oropharynx (OSCC) is of considerable public health significance. In the United States, it is estimated that nearly 35,000 new OSCC cases were diagnosed in 2007, and approximately 7,550 OSCC deaths are estimated to occur (see <http://www.cancer.org>, web site of the American Cancer Society, Atlanta, GA). World-wide, OSCC is the 6th most common cancer, with an estimated 405,000 new cases and 211,000 deaths annually (<http://www-dep.iarc.fr>, International Agency for Research on Cancer, Lyon, France) (1). Despite considerable advances in surgical techniques, and the use of adjuvant treatment modalities, the 5-year survival rate for OSCC patients is about 60% for U.S. caucasians and 36% for U.S. African-Americans (<http://www.cancer.org>, American Cancer Society, Atlanta, GA). In addition, OSCC is often associated with loss of eating and speech function, disfigurement and psychological distress. As much as 20% of oral dysplasia undergoes malignant transformation to OSCC (2, 3).

Among post-surgical OSCC patients with histologic positive tumor margins, the likelihood of local recurrence is as high as 70 to 80%. Even among post-surgical OSCC patients with negative margins, the reported probability of recurrence is 30-40% (4), suggesting histologic examination alone is inadequate in predicting recurrence (4-6).

Clearly there is an urgent need to identify better ways to predict which patients with dysplastic precursor lesions are likely to develop OSCC, and to predict whether patients who have been surgically treated for OSCC are likely to relapse, so that high-risk patients can be selected for more rigorous treatment and follow-up. The present invention addresses these needs and provides other related advantages.

BRIEF SUMMARY

In certain embodiments the present invention provides a method for identifying a risk for having, or presence of, oral squamous cell carcinoma (OSCC) in a subject, the method comprising (a) determining a squamous cell carcinoma indicator gene set (SCCIGS) expression level in a biological sample from the subject that comprises at least one OSCC cell or at least one cell from an OSCC surgical margin; and (b) comparing the SCCIGS expression level of (a) to a reference SCCIGS expression level that is characteristic of epithelial cells from a control tissue that comprises normal oral epithelium known to be free of squamous cell carcinoma cells; wherein differential expression of the SCCIGS in the biological sample relative to the control tissue indicates the subject has, or is at risk for having, OSCC.

In certain further embodiments the SCCIGS comprises one or more SCC biomarker genes selected from SEQ ID NOS:1-200, or variants thereof that are differentially expressed in a squamous cell carcinoma cell as compared to a control tissue that comprises normal oral epithelium known to be free of squamous carcinoma cells. In certain embodiments the SCCIGS is one or more SCCIGS selected from (a) the SCCIGS consisting of a LAMC2 gene, (b) the SCCIGS consisting of LAMC2 and COL4A1 genes, and (c) the SCCIGS consisting of COL1A1 and PADI1 genes. In certain embodiments the SCCIGS is one or more SCCIGS selected from (a) the SCCIGS consisting of a LAMC2 gene, (b) the SCCIGS consisting of LAMC2 and COL4A1 genes, (c) the SCCIGS consisting of COL1A1 and PADI1 genes, (d) the SCCIGS consisting of a C21orf81 gene, (e) the SCCIGS consisting of KRT17 and PRSS3 genes, (f) the SCCIGS consisting of COL1A2 and EST 230740_1at genes, (g) the SCCIGS consisting of COL1A1 and XLKD1 genes, (h) the SCCIGS consisting of THY1, FLJ22671 (referred to as C2orf54 in the Affymetrix database) and HAS3 genes, (i) the SCCIGS consisting of POSTN and TIA2 (referred to as PDPN in the Affymetrix database) genes, (j) the SCCIGS consisting of MGC40368 (referred to as TCP11L2 in the Affymetrix database), GIP3 (referred to as IFI6 in the Affymetrix database) and COL27A1 genes, (k) the SCCIGS consisting of CDH3 and ELOVL6 genes, (l) the SCCIGS consisting of the COL4A1 gene and (m) the

SCCIGS consisting of genes identified by 131 probe sets as set forth in Figure 2A.

In certain embodiments the step of determining a SCCIGS expression level comprises (a) specifically hybridizing a detectable, squamous cell carcinoma indicator gene (SCCIG)-specific oligonucleotide probe to one or more of (i) all or a SCCIG-characteristic portion of a SCCIG transcript, (ii) a polynucleotide having a nucleotide sequence that is fully complementary to (i), and (iii) a nucleic acid amplification product of one or more of (i) and (ii); and (b) detecting the SCCIG-specific probe, and thereby determining the SCCIGS expression level. In certain further embodiments the SCCIG-specific oligonucleotide probe has a nucleotide sequence that is selected from the group consisting of SEQ ID NOS:201-331. In certain embodiments the biological sample comprises a biopsy tissue, which in certain further embodiments is selected from an excised tumor, a tumor-positive margin tissue, a tumor-negative margin tissue and a close margin tissue. In certain other embodiments the biological sample comprises one or a plurality of dysplastic cells.

In other embodiments there is provided a method for identifying a risk for having, or presence of, oral squamous cell carcinoma (OSCC) in a subject having oral epithelial dysplasia but no frank OSCC, the method comprising (a) determining a squamous cell carcinoma indicator gene set (SCCIGS) expression level in a biological sample from the subject that comprises at least one dysplastic oral epithelial cell; and (b) comparing the SCCIGS expression level of (a) to a reference SCCIGS expression level that is characteristic of OSCC cells; wherein substantial similarity of the SCCIGS expression level in the biological sample relative to the OSCC reference SCCIGS expression levels indicates the subject has, or is at risk for having, OSCC. In certain further embodiments the squamous cell carcinoma indicator gene set (SCCIGS) comprises any one or more of the genes shown in Figure 6. In certain other embodiments the step of determining a SCCIGS expression level comprises (a) specifically hybridizing a detectable, squamous cell carcinoma indicator gene (SCCIG)-specific oligonucleotide probe to one or more of: (i) all or a SCCIG-characteristic portion of a SCCIG transcript, (ii) a polynucleotide having

a nucleotide sequence that is fully complementary to (i), and (iii) a nucleic acid amplification product of one or more of (i) and (ii); and (b) detecting the SCCIG-specific probe, and thereby determining the SCCIGS expression level. In certain further embodiments the SCCIG-specific oligonucleotide probe has a nucleotide
5 sequence that is selected from the group consisting of the probes listed in Figure 9. In certain other embodiments the biological sample comprises a biopsy tissue. In certain other embodiments the subject has no detectable cancer and the biological sample comprises one or a plurality of dysplastic cells.

Turning to another embodiment, there is provided a method for
10 identifying a risk for having, or presence of, a squamous cell carcinoma (SCC) in a subject, wherein the SCC is selected from oral SCC (OSCC) and head-and-neck SCC (HNSCC), the method comprising (a) determining a squamous cell carcinoma indicator gene set (SCCIGS) expression level in a biological sample from the subject that comprises at least one OSCC cell or at least one cell from
15 an OSCC surgical margin; and (b) comparing the SCCIGS expression level of (a) to a reference SCCIGS expression level that is characteristic of epithelial cells from a control tissue that comprises normal oral epithelium known to be free of squamous cell carcinoma cells, wherein: if the biological sample comprises an OSCC cell then the control tissue comprises normal oral epithelium, and if the
20 first biological sample comprises a HNSCC cell then the control tissue comprises normal epithelium from oropharynx, hypopharynx, larynx or oral cavity; and wherein differential expression of the SCCIGS in the biological sample relative to the control tissue indicates the subject has, or is at risk for having, OSCC. In certain further embodiments the SCCIGS comprises one or more SCC biomarker
25 genes selected from SEQ ID NOS:1-200, or variants thereof that are differentially expressed in a squamous cell carcinoma cell as compared to a control tissue that comprises normal oral epithelium known to be free of squamous carcinoma cells.

In certain embodiments the SCCIGS is one or more SCCIGS selected from (a) the SCCIGS consisting of a LAMC2 gene, (b) the SCCIGS
30 consisting of LAMC2 and COL4A1 genes, and (c) the SCCIGS consisting of COL1A1 and PADI1 genes. In certain embodiments the SCCIGS is one or more SCCIGS selected from (a) the SCCIGS consisting of a LAMC2 gene, (b) the

SCCIGS consisting of LAMC2 and COL4A1 genes, (c) the SCCIGS consisting of COL1A1 and PADI1 genes, (d) the SCCIGS consisting of a C21orf81 gene, (e) the SCCIGS consisting of KRT17 and PRSS3 genes, (f) the SCCIGS consisting of COL1A2 and EST 230740_1at genes, (g) the SCCIGS consisting of COL1A1 and XLKD1 genes, (h) the SCCIGS consisting of THY1, FLJ22671 (referred to as C2orf54 in the Affymetrix database) and HAS3 genes, (i) the SCCIGS consisting of POSTN and TIA2 (referred to as PDPN in the Affymetrix database) genes, (j) the SCCIGS consisting of MGC40368 (referred to as TCP11L2 in the Affymetrix database), GIP3 (referred to as IFI6 in the Affymetrix database) and COL27A1, (k) the SCCIGS consisting of CDH3 and ELOVL6 genes, (l) the SCCIGS consisting of the COL4A1 gene and (m) the SCCIGS consisting of genes identified by 131 probe sets as set forth in Figure 2A.

In certain embodiments the biological sample comprises an OSCC cell and the control tissue comprises normal oral epithelium. In certain embodiments the biological sample comprises a HNSCC cell and the control tissue comprises normal epithelium from oropharynx, hypopharynx, larynx or oral cavity. In certain embodiments the step of determining a SCCIGS expression level comprises (a) specifically hybridizing a detectable, squamous cell carcinoma indicator gene (SCCIG)-specific oligonucleotide probe to one or more of: (i) all or a SCCIG-characteristic portion of a SCCIG transcript, (ii) a polynucleotide having a nucleotide sequence that is fully complementary to (i), and (iii) a nucleic acid amplification product of one or more of (i) and (ii); and (b) detecting the SCCIG-specific probe, and thereby determining the SCCIGS expression level. In certain further embodiments the SCCIG-specific oligonucleotide probe has a nucleotide sequence that is selected from the group consisting of SEQ ID NOS:201-331. In certain other further embodiments the biological sample comprises a biopsy tissue. In certain still further embodiments the biopsy tissue is selected from the group consisting of an excised tumor, a tumor-positive margin tissue, a tumor-negative margin tissue and a close margin tissue. In certain embodiments the biological sample comprises one or a plurality of dysplastic cells.

In certain other embodiments of the present invention there is provided a method for identifying an increased risk of oral squamous cell carcinoma (OSCC)-specific mortality in a subject having OSCC, the method comprising: (a) determining a squamous cell carcinoma indicator gene set (SCCIGS) expression level in a biological sample from the subject that comprises at least one OSCC cell or at least one cell from an OSCC surgical margin; (b) determining that the subject has, or is at risk for having, OSCC by comparing the SCCIGS expression level of (a) to a reference SCCIGS expression level that is characteristic of epithelial cells from a control tissue that comprises normal oral epithelium known to be free of squamous cell carcinoma cells, wherein a differentially expressed SCCIGS in the biological sample relative to the control tissue indicates the subject has, or is at risk for having, OSCC; and (c) identifying within said differentially expressed SCCIGS a presence or absence of a substantially up- or down-regulated SCCIGS subset (SCCIGSS), wherein presence of the substantially up- or down-regulated SCCIGSS indicates the subject has an increased risk of OSCC-specific mortality.

In certain further embodiments the SCCIGS comprises one or more SCC biomarker genes selected from SEQ ID NOS:1-200, or variants thereof that are differentially expressed in a squamous cell carcinoma cell as compared to a control tissue that comprises normal oral epithelium known to be free of squamous carcinoma cells. In certain embodiments the SCCIGSS is one or more SCCIGSS selected from the group consisting of (a) the SCCIGS consisting of a LAMC2 gene, (b) the SCCIGSS consisting of a LAMC2 gene, (c) the SCCIGSS consisting of OSMR, SERPINE1 and OASL genes, (d) the SCCIGSS consisting of a SLC16A1 gene, (e) the SCCIGSS consisting of a KLF7 gene, (f) the SCCIGSS consisting of THBS1 and SLC16A1 genes, (g) the SCCIGSS consisting of a HOMER3 gene, (h) the SCCIGSS consisting of a GRP68 gene, (i) the SCCIGSS consisting of a PDPN gene, (j) the SCCIGSS consisting of an ANKRD35 gene, and (k) the SCCIGSS consisting of CDH3 and EPS8L1 genes.

In certain embodiments (i) the SCCIGS is one or more SCCIGS selected from the group consisting of (a) the SCCIGS consisting of a LAMC2 gene, (b) the SCCIGS consisting of LAMC2 and COL4A1 genes, (c) the SCCIGS

consisting of COL1A1 and PADI1 genes, (d) the SCCIGS consisting of a C21orf81 gene, (e) the SCCIGS consisting of KRT17 and PRSS3 genes, (f) the SCCIGS consisting of COL1A2 and EST 230740_1at genes, (g) the SCCIGS consisting of COL1A1 and XLKD1 genes, (h) the SCCIGS consisting of THY1, 5 FLJ22671 and HAS3 genes, (i) the SCCIGS consisting of POSTN and TIA2(PDPN) genes, (j) the SCCIGS consisting of MGC40368(TCP11L2), GIP3(IFI6) and COL27A1 genes, (k) the SCCIGS consisting of CDH3 and ELOVL6 genes, (l) the SCCIGS consisting of the COL4A1 gene, and (m) the SCCIGS consisting of genes identified by 131 probe sets as set forth in Figure 10 2A, and wherein (ii) the SCCIGSS consists of one or more genes identified by a probe set as set forth in Table 2.

In certain further related embodiments at least one of the steps selected from the step of determining a SCCIGS expression level and the step of identifying a SCCIGSS comprises (a) specifically hybridizing a detectable, 15 squamous cell carcinoma indicator gene (SCCIG)-specific oligonucleotide probe to one or more of (i) all or a SCCIG-characteristic portion of a SCCIG transcript, (ii) a polynucleotide having a nucleotide sequence that is fully complementary to (i), and (iii) a nucleic acid amplification product of one or more of (i) and (ii); and (b) detecting the SCCIG-specific probe, and thereby determining the SCCIGS 20 expression level. In certain further embodiments the SCCIG-specific oligonucleotide probe has a nucleotide sequence that is selected from SEQ ID NOS:201-331. In certain embodiments the biological sample comprises a biopsy tissue, which in certain further embodiments is selected from an excised tumor, a tumor-positive margin tissue, a tumor-negative margin tissue and a close margin 25 tissue. In certain embodiments the biological sample comprises one or a plurality of dysplastic cells.

In certain related embodiments determining one or a plurality of SCCIGS expression levels comprises measuring one or more protein levels in the biological sample. In certain further embodiments the biological sample 30 comprises a biological fluid, which in certain still further embodiments is selected from saliva, blood, serum, plasma and lymph.

These and other aspects of the invention will be evident upon reference to the following detailed description and attached drawings. All of the U.S. patents, U.S. patent application publications, U.S. patent applications, foreign patents, foreign patent applications and non-patent publications referred to in this specification and/or listed in the Application Data Sheet, are
5 incorporated herein by reference in their entirety, as if each was incorporated individually. Aspects of the invention can be modified, if necessary, to employ concepts of the various patents, applications and publications to provide yet further embodiments of the invention.

10 BRIEF DESCRIPTION OF THE SEVERAL VIEWS OF THE DRAWINGS

Figure 1 shows the most prominently involved biological pathways in oral squamous cell carcinoma (OSCC). Top: IFN- γ signaling pathway. Bottom: JAK/STAT pathway. Analysis was performed using Ingenuity®Systems, version 4.0.

15 Figure 2 shows the selected biomarker genes and corresponding probe sets that were shown to be differentially expressed between OSCC and controls based on the criteria described in Example 1.

Figure 3 shows the top 10 squamous cell carcinoma indicator gene set (SSCIGS) models from the logistic regression analyses of the selected
20 biomarker genes from Figure 2. The predictive power of these SSCIGS models was validated using internal and external (GSE6791) controls, as measured by the area under the curve (AUC). An AUC of 0.5 represents a test that is no better than chance at discriminating between cases and controls, and an AUC of 1.0 provides perfect discrimination.

25 Figure 4 shows qRT-PCR results comparing RNA transcripts for four genes between OSCC cases and controls (see Example 2).

Figure 5 shows the tissue specificity of a squamous cell carcinoma gene set (SSCIGS) consisting of *LAMC2* and *COL4A1* (top) and a SSCIGS consisting of *COL1A1* and *PADI1* (bottom). The data from Example 5 are
30 represented in Box Whisker plots of logistic regression scores (y axis) for normal controls and cases in an internal testing set (N: normal, DYS: dysplasia, T:

OSCC), GEO GSE6791 head and neck normal controls (HNN) and cases (HNT), GEOGSE 6791 cervical normal controls (CN) and cases (CT), and GEO GSE6044 lung normal controls (LN), lung squamous cell carcinoma (LSCC), lung adenocarcinoma (LAD) and lung small cell cancer (LSC).

5 Figure 6 shows a list of biomarker genes that are differentially expressed between OSCC and dysplasia/normal controls, and which can be utilized, for example, to distinguish between frank OSCC and dysplasia.

 Figure 7 shows the sequence information, including GenBank accession numbers and descriptive annotations for SEQ ID NOS:1-200, a list of
10 differentially expressed SCC biomarker genes that may be used to identify the risk or presence of SCC in a subject.

 Figure 8 shows the sequence identifiers for SEQ ID NOS:201-331, a list of exemplary Affymetrix probes (Affymetrix Corp., Santa Clara, CA) that specifically hybridize to certain of the SCC biomarker genes or gene sets
15 described herein. The middle column shows the Affymetrix probe identifier and the far right column shows the corresponding biomarker gene to which the probe specifically hybridizes.

 Figure 9 shows a selected set of exemplary Affymetrix probes that specifically hybridize to certain of the SCC biomarker genes that can be utilized
20 to discriminate between OSCC tumor cells and dysplastic epithelial cells. The middle column shows the Affymetrix probe identifier and the far right column shows the corresponding biomarker gene to which the probe specifically hybridizes.

 Figure 10 shows supervised hierarchical cluster analysis of the
25 gene expression data. The 131 probe sets were clustered as described in the text. The bar underneath the heat map codes the samples according to tissue phenotype: normal, dysplasias and tumors. Each column in the heat map represents the expression levels for all genes in a particular sample, whereas each row represents the relative expression of a particular gene across all
30 samples. The expression level of any gene in any given sample (relative to the mean expression level of that gene across all samples) was also recorded along a color scale (not shown) in which red represents transcription up-regulation,

green represents down-regulation, and the color intensity indicates the magnitude of deviation from the mean. Cluster 1 refers to a group of probe sets which appear to be only fully downregulated in a group of 45 patients labeled with a bar at the bottom of the heat map.

5 Figure 11 shows Principal Component Analysis (PCA) using the 131 probe sets. The first principal component (PC) is plotted on the x-axis and captures 63.28 % of the variance. The second PC is plotted on the y-axis and captures 5.66 % of the variance.

 Figure 12 shows survival and OSCC-specific mortality estimates in
10 OSCC patients. The two groups were identified with hierarchical clustering analysis using the 131 differentially expressed genes in invasive OSCC as described in the text. 12A. Kaplan-Meier analysis of all-cause mortality. Vertical marks represent censored events. 12B. Cumulative incidence of OSCC-specific mortality.

15 Figure 13 shows Receiver Operating Characteristic Analysis of 2-year Survival Comparing the Prognostic Ability of Stage with Gene Expression Data. 13A. ROC Curves for 2-year survival for , 'stage', 'LAMC2' and 'PCA'. 13B. ROC Curves for 2-year survival for models 'stage', 'stage and LAMC2' and 'stage and PCA'. 13C. Area Under the Curve (AUC) and bootstrapped 95%
20 Confidence Intervals for all five models.

 Figure 14 shows three-dimensional plot of the first and second principal components and the risk scores from the top Cox regression model ($0.59151 \cdot LAMC2$). The samples are color coded according to vital status. Diamonds ($\diamond$) are used to show the overlap between risk scores and samples
25 from either the group of 45 (red diamonds) or the group of 74 (blue diamonds).

DETAILED DESCRIPTION

 Embodiments of the present invention relate generally to the use of highly predictive gene expression profiling, based on differentially expressed
30 biomarker genes or gene sets, to detect the presence or risk of squamous cell

carcinoma (SCC) in a subject, and in certain further embodiments to identify an increased risk of oral squamous cell carcinoma (OSCC)-specific mortality in a subject having OSCC. In certain embodiments, the methods provided herein may be used to identify a variety of head and neck squamous cell carcinomas (HNSCC), including oral squamous cell carcinomas (OSCC).

Gene expression profiling is a useful way to distinguish between cells that express different phenotypes, and may be used in particular embodiments to distinguish between cancer cells and normal cells, or in other embodiments to distinguish between different types of cancer cells, and/or in certain further embodiments to identify aggressively neoplastic OSCC cells in a method for identifying an increased risk of oral squamous cell carcinoma (OSCC)-specific mortality in a subject having OSCC. Gene expression profiling according to the methods provided herein relates generally to measurements of selected biomarker genes or gene sets shown to be differentially expressed in various types of SCC, such as HNSCC and OSCC. In certain embodiments, differentially expressed biomarker genes or gene sets that may be used to identify SCC are referred to herein as squamous cell carcinoma indicator gene sets (SCCIGS), which are exemplified in Figures 2 and 7 and detailed below. In certain embodiments, subsets of such differentially expressed biomarker genes or gene sets that may be used to identify aggressively neoplastic SCC that, as described herein, may be indicators of an increased risk of OSCC-specific mortality, are referred to herein as squamous cell carcinoma indicator gene set subsets (SCCIGSS), which are exemplified in Tables 2 and 4 and which are described in greater detail below.

According to certain embodiments of the present invention, determination and comparison of the expression levels of one or more selected biomarker genes or gene sets, such as a given SCCIGS, provide novel and useful parameters for diagnosing the risk or presence of SCC in a subject. As such, certain embodiments of the present invention relate to methods for identifying the risk or presence of SCC in a subject by comparing the expression levels of selected biomarker genes or gene sets in a biological sample from that subject to the expression levels of those same biomarker genes or gene sets in

an appropriate control, such as a normal tissue known to be free of SCC (e.g., a control tissue or a reference tissue). The presence or risk of SCC may be identified by the differential expression of the selected biomarker genes or gene sets in the subject sample as compared to the control.

5 Certain embodiments may also include a simple genetic test based on the gene expression profile of one or more selected biomarker genes, such as a selected SCCIGS. In these and other aspects, a simple genetic test may employ selected probes or probe sets that are specific for one or more SCCIGS as provided herein, including but not limited to the Affymetrix oligonucleotide
10 probes exemplified herein and variants thereof, to measure the gene expression levels of a SCCIGS or other biomarker gene, and to compare those levels to a reference SCCIGS expression level in an appropriate control (e.g., a control tissue or a reference tissue). Such a simple genetic test provides advantages over other previous methods in the art, which have failed to utilize rigorous
15 statistical analyses to identify highly predictive and readily detectable sets of biomarker genes or gene sets. As described herein, also provided is a corresponding set of biomarker gene probes, for use in diagnosing the presence of SCC in a subject and/or the risk of developing SCC. In similar fashion and as described herein, certain embodiments further contemplate identifying substantial
20 down-regulation (e.g., expression that is reduced in a statistically significant manner by at least 50%, 60%, 70%, 80%, 90%, 95% or more, relative to an appropriate control group) of a subset of SCCIGS, referred to as SCCIGSS, where such down-regulation may indicate an increased risk of OSCC-specific mortality relative to other OSCC cases identified according to the disclosure
25 herein.

 In certain embodiments of the present invention, without wishing to be bound by any one theory, it is believed that patients who develop local recurrence and/or second primary oral tumors are those whose surgical margins or uninvolved buccal mucosa harbor molecular changes that are found in oral
30 dysplasia or invasive OSCC. In these and related embodiments, the predictive models provided herein may be used to test biopsies of histologically normal surgical margins and clinically normal oral mucosa of OSCC patients, in order to

identify a risk for having, or the presence of, local recurrence and/or second primary oral cancer.

The strong predictive power of the presently disclosed biomarker genes and/or gene sets when used according to the methods described herein
5 may be exploited generally to differentiate between normal cells, pre-neoplastic cells (*i.e.*, dysplasia), and SCC tumor cells, including a variety of HNSCC tumor cells, such as OSCC tumor cells. Among other uses apparent to a person skilled in the art based on the present disclosure, this predictive power may find use in a clinical setting, for example, to identify or monitor subjects having SCC, or
10 subjects at risk for developing SCC, and may also find use in research settings, for instance, to further characterize the underlying biological bases of SCC oncogenesis and pathology.

Squamous cell carcinoma (SCC) generally includes malignant tumors of squamous epithelium (*i.e.*, epithelium that shows squamous cell
15 differentiation), which may occur in many different organs, including the skin, lips, mouth, esophagus, urinary bladder, prostate, lungs, vagina, and cervix. Squamous cells form the surface tissue layer (*i.e.*, epithelium) of much of the body, and include cells of the skin and mucous membranes. SCC is thought to derive from keratinizing or malpighian epithelial cells. One hallmark of squamous
20 cell carcinoma is the presence of keratin, or "keratin pearls," on histologic evaluation. These keratin formations typically relate to well-formed desmosome attachments and intracytoplasmic bundles of keratin tonofilaments. SCC is morphologically variable, and may appear, by way of non-limiting example, as plaques, nodules, or verrucae.

25 SCC usually begins as surface lesions with erythema and slight elevation, often termed erythroplasia. Some early SCC lesions may appear to be pure white, and are referred to as leukoplakia, but only a small percentage of leukoplakia lesions represent carcinoma *in situ* or invasive carcinoma. Erythroplasia, or early red lesions, are typically asymptomatic and may represent
30 either carcinoma *in situ* or invasive carcinoma. Tender, painful lesions usually are suggestive of perineural invasions. When lesions become palpable masses, symptoms such as a vague persistent sore throat or ear infection typically occur.

In more advanced cases, dissemination to ipsilateral submandibular and jugulodigastric nodes is common, and a subject suspected of having SCC may present with a mass in the neck. When lymph node or remote bone and organ metastases are associated with an early oral primary lesion, often a
5 second, more advanced primary upper aero-digestive or lung cancer is responsible for the metastases.

Head-and-neck squamous cell carcinoma (HNSCC) refers generally to a group of biologically similar squamous cell carcinomas originating from the upper aero-digestive tract, including the lip, oral cavity (mouth), nasal cavity,
10 paranasal sinuses, pharynx, and larynx, among others. HNSCC often spreads to the lymph nodes of the neck, which may represents the first, and sometimes only, manifestation of the disease at the time of diagnosis.

HNSCCs are typically characterized by their originating tissues. For example, HNSCCs may arise from the salivary glands, which produce saliva, the
15 fluid that keeps mucosal surfaces in the mouth and throat moist. The major salivary glands may be found in the floor of the mouth and near the jawbone. The paranasal sinuses are small hollow spaces in the bones of the head surrounding the nose. The nasal cavity is the hollow space inside the nose.

HNSCCs may also originate in the pharynx. The pharynx is
20 essentially a hollow tube common to the upper digestive and respiratory tracts, originating behind the nose, forming the throat lumen and leading to the esophagus and the trachea. The pharynx has three parts, the nasopharynx, the oropharynx, and the hypopharynx. Nasopharyngeal cancer arises in the nasopharynx, the region in which the nasal cavities and the Eustachian tubes
25 connect with the upper part of the throat. Oropharyngeal cancer often begins in the oropharynx, the middle part of the throat that includes the soft palate, the base of the tongue, and the tonsils. The hypopharynx includes the pyriform sinuses, the posterior pharyngeal wall, and the postcricoid area. Tumors of the hypopharynx frequently have an advanced stage at diagnosis, and have the most
30 adverse prognoses of pharyngeal tumors, tending to metastasize early due to the extensive lymphatic network around the larynx.

HNSCC may also originate in the larynx, or "voice box." Such cancers may occur on the vocal folds themselves (*i.e.*, "glottic" cancer), or on tissues above and below the true cords (*i.e.*, "supraglottic" and "subglottic" cancers, respectively). Laryngeal cancer is strongly associated with tobacco smoking. In general, HNSCC is highly curable if detected early, usually with some form of surgery, although chemotherapy and radiation therapy may also play an important role.

SCC of the oral cavity, or mouth, may represent one particular aspect of HNSCC, and is typically referred to as oral squamous cell carcinoma (OSCC). OSCC is associated with substantial mortality and morbidity. OSCC relates generally to the formation of SCC in the area extending from the vermilion border of the lips to a plane between the junction of the hard and soft palate superiorly and the circumvallate papillae of the tongue inferiorly. This area includes, for example, the front two thirds of the tongue, the gingiva (gums), the buccal mucosa (the lining of the inside of the cheeks, the floor (bottom) of the mouth under the tongue, the hard palate (the roof of the mouth), and/or the retromolar trigone (the small area behind the wisdom teeth).

OSCC typically spreads primarily by either local extension or by the lymphatic system. The extent of tumor invasion depends upon the anatomic site, the tumor's biologic aggressiveness, and host response factors.

The lymphatic system is the most important and frequent route of metastasis in OSCC. Typically, the ipsilateral cervical lymph nodes are the primary site for metastatic deposits, but occasionally contralateral or bilateral metastatic deposits may be detected. The risk for lymphatic spread is greater for posterior lesions of the oral cavity, possibly because of delayed diagnosis or increased lymphatic drainage at those sites, or both. Cervical lymph nodes with metastatic deposits tend to appear as firm-to-hard, nontender enlargements. Once the tumor cells perforate the nodal capsule and invade the surrounding tissue, these lymph nodes often become fixed and non-mobile.

Metastatic spread of tumor deposits from oral carcinoma usually occurs in an orderly pattern, beginning with the uppermost lymph nodes and spreading down the cervical chain. Because of this pattern of spread, the jugulo-

digastric nodes are most prone to early metastasis. Carcinomas involving the lower lip and floor of the mouth are an exception, as they tend to spread to the submental nodes. Although lymph node metastasis is not an early event, many individuals with oral cancer nonetheless present at diagnosis with nodal

5 metastasis. Hematogenous spread of tumor cells is infrequent in the oral cavity but may occur because of direct vascular invasion or seeding from surgical manipulation.

Many patients with OSCC will present initially with highly confined localized disease stages. These patients may be treated with curative intent, usually involving surgery, radiation therapy, or both. Only about 20-40% of patients will develop a local or regional tumor recurrence. However, over subsequent years, these "cured" patients are often at higher risk for developing a second malignancy than for developing a recurrence of the initial tumor. Tumor recurrences most often occur during the first 2 years after therapy; later

10 recurrences are rare. Second malignancies, on the other hand, may be observed at a steady rate-perhaps 3-5% per year. Thus, with sufficient follow-up time, second malignancies or other medical diseases become greater problems than recurrence of the primary disease. The methods provided herein may be utilized to monitor such tumor recurrences or second malignancies.

20 The term dysplasia, or dysplastic cell, refers generally to a maturation abnormality of cells within a tissue, which often involves the expansion of immature cells and a corresponding decrease in the number of mature cells at a given site. Dysplasia is often indicative of an early or pre-neoplastic process. The term dysplasia is typically used when the cellular

25 abnormality is restricted to the originating tissue, as in the case of an early, *in-situ* neoplasm. Dysplasia is often considered the earliest form of pre-cancerous lesion recognizable in a biopsy, and dysplasia relevant to HNSCC or OSCC typically relates to dysplasia of epithelial cells. Dysplasia may be further characterized as "low grade" or "high grade." The risk of low grade dysplasia

30 transforming into high grade dysplasia and, eventually, cancer is low. High grade dysplasia represents a more advanced progression towards malignant transformation, with increased risk of developing a carcinoma *in situ*.

Carcinoma *in situ*, meaning "cancer in place," represents generally the transformation of a neoplastic lesion to one in which cells undergo essentially no maturation, and thus may be considered cancer-like. In this state, cells are often considered to have lost their tissue identity, and have reverted to a primitive
5 cell form that grows rapidly and without regulation. This form of cancer, however, often remains localized, and has not invaded into tissues below the surface.

Invasive carcinoma refers generally to a cancer that has invaded beyond the original tissue layer or basement membrane and may be able to spread to other parts of the body (*i.e.*, metastasize). The molecular events
10 involved in the development of squamous dysplasia and subsequent carcinoma are poorly understood.

According to certain embodiments, a subject may include any animal, any mammal, and particularly any human individual having, at risk for having, or suspected of having, a SCC tumor cell, such as a HNSCC tumor cell,
15 an OSCC tumor cell, and/or a pre-neoplastic growth, such as a dysplastic cell. Such a subject may have previously undergone treatment for SCC and may be at risk for developing another case of SCC, or may be newly suspected of having SCC or SCC-related dysplasia. Subjects may be identified according to routine clinical techniques described herein and known in the art, including, for example,
20 by clinical examination of the head and neck, skin, mouth, or other relevant tissue (see, *e.g.*, Epstein *et al.*, *Can Fam Physician*. 54:870-5, 2008; and Robinson *et al.*, *Otolaryngol Clin North Am*. 39:295-306, 2006), and/or by various imaging technologies (see, *e.g.*, Piatta *et al.*, *Acta Otorhinolaryngol Ital*. 28:49-54, 2008; and Isles *et al.*, *Clin Otolaryngol*. 33:210-22, 2008). Toluidine blue (vital staining)
25 also provides a useful adjunct to clinical examination. The mechanism of vital staining is based on selective binding of the dye to dysplastic or malignant cells in the oral epithelium (*e.g.*, Helsper, *CA Canc. J. Clin*. 22:172, 1972). According to non-limiting theory, it may be that toluidine blue selectively stains for acidic tissue components and thus binds more readily to DNA, which is increased in neoplastic
30 cells. Vital staining can also help to determine the most appropriate biopsy sites and to surgically delineate margins.

Diagnostic imaging evaluation, such as either computer tomography (CT) scanning or magnetic resonance imaging (MRI), may also be used to identify a subject at risk for SCC, and further to assess the extent of local and/or regional tumor spread, the depth of invasion, and the extent of lymphadenopathy.

5 CT is often considered superior in detecting early bone invasion and lymph node metastasis, but MRI is typically preferred for assessing the extent of soft tissue involvement and for providing a three-dimensional display of the tumor. MRI is also the preferred technique for imaging carcinoma of the nasopharynx or lesions involving paranasal sinuses or the skull base.

10 Typical symptoms associated with SCC such as HNSCC or OSCC may include, for example, a sore on the lip or in the mouth that does not heal, a lump or thickening on the lips or gums or in the mouth, a white or red patch on the gums, tongue, tonsils, or lining of the mouth, bleeding, pain, or numbness in the lip or mouth, change in voice, loose teeth or dentures that no longer fit well,
15 trouble chewing or swallowing or moving the tongue or jaw, swelling of jaw, and/or sore throat or feeling that something is caught in the throat.

Merely by way of non-limiting example, a subject having or suspected of having SCC may be identified by oral lesions that appear in areas of erythroplakia or leukoplakia, and which may be exophytic or ulcerated. Both the
20 latter variants are typically indurated and firm with a rolled border. Tonsillar carcinoma in a subject usually presents as an asymmetric swelling and sore throat in which pain often radiates to the ipsilateral ear; a metastatic mass in the neck may be the first symptom. OSCC associated lesions are described in *Detecting Oral Cancer, A Guide for Health Care Professionals*, U.S. Department
25 of Health and Human Services, National Institutes of Health, Bethesda, MD.

Risk factors for HNSCC and/or OSCC that may be used to identify a subject according to certain herein disclosed embodiments may include, for example, tobacco product use, heavy alcohol use, exposure to sunlight (e.g., lower lip SCC), being male, and being infected with human papillomavirus (HPV)
30 or Epstein-Barr virus (EBV). Environmental exposures to paint fumes, plastic by-products, wood dust, asbestos, and/or gasoline fumes have also been implicated as risk factors. Gastroesophageal reflux disease is thought to be a significant

risk factor for cancer of the larynx, and especially the anterior two thirds of the vocal cords. Irritation from poorly fitting dentures also has been implicated. Additional risk factors for HNSCC or SCC are described, for example, in Napier *et al. J Oral Pathol Med.* 37:1-10, 2008.

5 Biological samples may be provided by obtaining a blood sample, biopsy specimen, tissue explant, organ culture or any other tissue or cell preparation from a subject or a biological source, including tissue extracts or lysates derived from biopsies, cell extracts or lysates, nucleic acid extracts (e.g., RNA or DNA), and/or protein extracts and/or biological fluids including body
10 fluids. The subject or biological source may be a human or non-human animal, a primary cell culture or culture adapted cell line including but not limited to genetically engineered cell lines that may contain chromosomally integrated or episomal recombinant nucleic acid sequences, immortalized or immortalizable cell lines, somatic cell hybrid cell lines, differentiated or differentiable cell lines,
15 transformed cell lines and the like. In certain preferred embodiments of the invention, the subject or biological source may be suspected of having or being at risk for having SCC, and in certain preferred embodiments of the invention the subject or biological source may be known to be free of a risk or presence of such a condition according to current art-accepted criteria with which the skilled
20 person will be familiar.

 A biological sample may include any type of cell-containing or cell- or tissue-derived sample that may be isolated, obtained, or derived from a subject and utilized to determine whether the cells from that subject show the differential SCCIGS expression profile that is characteristic of a SCC tumor cell as provided
25 herein, or of a dysplastic cell as provided herein. In certain preferred embodiments, the expression level(s) may be determined for one or more SCC-related biomarker genes or SCCIGS as identified herein, using the methods described herein and molecular biology techniques as known in the art. A suitable biological sample (e.g., a biological sample as provided herein) typically
30 may be suspected of comprising a SCC tumor cell or a dysplastic cell, such as a dysplastic epithelial cell. A biological sample may also include whole cells or fixed cells. Other typical sources of biological samples include cell cultures, as

noted above, including but not limited to those in which gene expression states may be manipulated to explore the relationship among genes (including, *e.g.*, SCCIGS).

Biopsy tissues may be obtained, for example, using surgical
5 scalpels, needles, biopsy punches or other means, and typically can be performed under local anesthesia. Incisional biopsy typically refers to the removal of a representative sample of the lesion; excisional biopsy typically refers to the complete removal of the lesion, with a border of normal tissue. A clinician may obtain multiple biopsy specimens of suspicious lesions to define the extent
10 of the primary disease and to evaluate the patient for the presence of possible synchronous second malignancies. Useful adjuncts to biopsies include vital staining, exfoliative cytology, fine needle aspiration biopsy, routine dental radiographs and other plain films, and imaging with magnetic resonance imaging (MRI) or computed tomography (CT).

15 Biopsy tissues may include, by way of non-limiting example, excised tumors or suspected tumors, tumor-positive margin tissues, tumor-negative margin tissues, and/or close margin tissues. Margin tissues refer generally to SCC-related surgical margins, which in turn relate to the area of tissue around the clinical border of a SCC tumor that should be surgically
20 removed to reduce the chance of tumor recurrence at the margins of skin excision. Merely by way of illustrative example, SCC surgical margins can range from about 3 mm to about 1 cm or more around the histologically established border of the SCC tumor, the size of which may be based in part on the staging by TNM classification to determine whether the tumor is considered a low-risk or
25 high-risk tumor (see, *e.g.*, Wittekind, Ch; Sobin, L. H. (2002). *TNM classification of malignant tumours*. New York: Wiley-Liss). High-grade tumors typically afford larger surgical margins, whereas low-grade tumors typically afford smaller surgical margins. In certain embodiments contemplated herein, surgical margin tissues may be monitored for the differential expression of SCCIGS or SCC gene
30 biomarkers after initial excision, such as for post-operative confirmation of tumor-negative margins, or during a follow-up period, such as for monitoring the potential recurrence of SCC tumor-positive cells in the surgical margin tissues

(see, e.g., de Visscher *et al.*, *International Journal of Oral and Maxillofacial Surgery* 31:154-157, 2002).

Examples of biological fluids include body fluids such as blood, serum and serosal fluids, plasma, lymph, urine, cerebrospinal fluid, saliva, mucosal secretions of the secretory tissues and organs, vaginal secretions, ascites fluids such as those associated with non-solid tumors, fluids of the pleural, pericardial, peritoneal, abdominal and other body cavities, and the like. Biological fluids may also include liquid solutions contacted with a subject or biological source, for example, cell and organ culture medium including cell or organ conditioned medium, lavage fluids and the like. In certain preferred embodiments the biological sample is saliva, and in certain other highly preferred embodiments the biological sample is blood or a fluid fraction thereof (e.g., serum, plasma), or lymph. In other preferred embodiments the biological sample is a cell-free liquid solution.

Certain embodiments of the present invention relate to the identification and use of one more selected Squamous Cell Carcinoma Gene Sets (SCCIGS). Among other uses as may be apparent to a person skilled in the art based on the present disclosure, the SCCIGSs provided herein may be used to detect SCC in a subject, to identify a risk of developing SCC in a subject, and/or to monitor for the recurrence of SCC in a subject. SSCIGSs represent one or more biomarker genes, alone or in selected combinations (*i.e.*, a biomarker gene set), that identify a risk for having, or the presence of, SCC in a subject when the expression levels of the SSCIGS in a suspected biological sample reflect differential expression (e.g., in a statistically significant manner) compared to the expression levels of the same SSCIGS in control epithelial cells that are known to be free of SCC cells.

A gene relates generally to a unit of inheritance that occupies a specific locus on a chromosome and includes transcriptional and/or translational regulatory sequences and/or a coding region (e.g., a polypeptide encoding region or a region encoding a structural RNA such as tRNA or rRNA, or a functional RNA such as a miRNA) and/or non-translated sequences (*i.e.*, introns, 5' and 3' untranslated sequences). The individual genes in a given SSCIGS may be either

over-expressed or under-expressed (*i.e.*, statistically significant higher or lower expression levels) in the biological sample comprising an SCC cell when compared to the control (non-cancer) cell. Typically, it is the differential expression of one or more selected SCCIGS that may identify the presence or risk of SCC in a subject.

In certain embodiments, a SCCIGS may be selected from the preferred SCCIGS models exemplified in Figure 3 (see SCCIGS models 1-10). In these and other embodiments, Figures 2 and 7 also provide a list of predictive SCC biomarker genes one or more of which may be selected to generate a suitable SCCIGS (see, *e.g.*, SEQ ID NOS:1-200) for identifying the risk or presence of SCC in a subject.

The SCC biomarker genes and SCCIGS gene sets described herein were identified according to the exemplary processes described in Example 1 below and in Chen *et al.*, *Cancer Epidemiol Biomarkers Prev* 17(8) Published Online July 30, 2008. As a brief summary of such an exemplary process, to identify potential biomarkers for early detection of invasive OSCC, the data provided herein were generated by comparing gene expression in samples of (i) incident primary OSCC, (ii) oral dysplasia, and (iii) clinically normal oral tissue from surgical patients without head and neck cancer or pre-neoplastic oral lesions (controls), using Affymetrix U133 2.0 Plus arrays. Selected differentially expressed probe sets and their corresponding biomarker genes were identified using a training set of 119 OSCC patients and 35 controls (see Figures 2 and 7; and SEQ ID NOS:1-200).

To identify certain preferred biomarker gene sets, or preferred SCCIGSs, forward and stepwise logistic regression analyses identified successive combinations of genes the expression of which differentiated OSCC from controls. One preferred SCCIGS model included the *LAMC2* gene, encoding laminin gamma 2 chain, and the *COL4A1* gene, encoding type IV collagen, alpha 1 chain. Subsequent SCCIGS modeling without these two markers showed that differential expression of the *COL1A1* gene, encoding type I collagen, alpha 1 chain, and of the *PADI1* gene, encoding type 1 peptidyl arginine deiminase, can also distinguish OSCC from controls. These two models were

validated using an internal independent testing set of 48 invasive OSCC and 10 controls and an external testing set of 42 head and neck squamous cell carcinoma (HNSCC) cases and 14 controls (GEO GSE6791), with sensitivity and specificity above 95% (see Figure 3). These two models were also able to distinguish oral epithelial dysplasia (n=17) from control (n=35) tissue. Differential expression of these four genes was confirmed by qRT-PCR (see Figure 4).

Certain exemplary SCCIGSs that may be utilized according to the present methods include, for example, a SCCIGS that includes or consists essentially of the *LAMC2* (SEQ ID NOS:18 and 19) and *COL4A1* (SEQ ID NO:113) genes; a SCCIGS that includes or consists essentially of the *COL1A1* (SEQ ID NOS:20 and 21) and *PADI1* (SEQ ID NO:145) genes; a SCCIGS that includes or consists essentially of the C21orf81 gene (SEQ ID NOS:186 and 187); a SCCIGS that includes or consists essentially of the *KRT17* (SEQ ID NO:59) and *PRSS3* (SEQ ID NOS:123 and 124) genes; a SCCIGS that includes or consists essentially of the *COL1A2* (SEQ ID NO:22) and EST 230740_1at (SEQ ID NO:198) genes; a SCCIGS that includes or consists essentially of the *COL1A1* and *XLKD1* (SEQ ID NO:142) genes; a SCCIGS that includes or consists essentially of the *THY1* (SEQ ID NOS:97 and 126), *FLJ22671* FLJ22671 (also referred to as C2orf54) (SEQ ID NO:143), and *HAS3* (SEQ ID NOS:155 and 156) genes; a SCCIGS that includes or consists essentially of the *POSTN* (SEQ ID NOS:6 and 108) and *TIA2* (also referred to as *PDPN*; SEQ ID NOS:264, 274, 52-55 and 150-153) genes; a SCCIGS that includes or consists essentially of the *MGC40368* (also referred to as TCP11L2; SEQ ID NOS:3, 314), *GIP3* (also referred to as *IFI6*; SEQ ID NOS:34-36, 206) and *COL27A1* (SEQ ID NOS:157 and 158) genes; a SCCIGS that includes or consists essentially of the *CDH3* (SEQ ID NO:24) and *ELOVL6* (SEQ ID NO:109) genes; and a SCCIGS that includes or consists essentially of the *COL4A1* (SEQ ID NO:113) gene (see Figures 3 and 7). Additional SCC biomarker genes providing strong predictive powers that may be utilized according to the present methods include those described in Figures 2 and 7.

Also included within the meaning of a SCC biomarker gene or SSCIGS are "variants" of the biomarker or SSC indicator genes described herein

(see, e.g., SEQ ID NOS:1-200), such as splice variants, isoforms, allelic variants (same locus), homologs (different locus), and orthologs (different organism), or biological functional equivalents of such genes. Such variants may also include biomarker genes comprising polynucleotide sequences having 30, 40, 50, 60, 70, 75, 80, 85, 90, 95, 97, 98, 99% identity to a sequence set forth in SEQ ID NOS:1-200, and which are differentially expressed in a SCC cell as compared to a control cell known to be free of SCC. Such variants also encompass polynucleotide sequences that are distinguished from a reference polynucleotide by the addition (e.g., insertion), deletion or substitution of at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45 or 50 nucleotides. Accordingly, the terms "polynucleotide variant" and "variant" include polynucleotides in which one or more nucleotides have been added, inserted or deleted, or replaced with different nucleotides.

A variety of methods known in the art and described herein may be utilized to determine an SCCIGS expression level, according to the methods provided herein. In certain aspects, measurements of SCCIGS expression levels may include determination in a biological sample from a subject as provided herein, of ribonucleic acid (RNA) and/or protein abundances, or protein activity levels. For example, the expression levels of a biomarker gene or gene set, such as a SCCIGS, may be determined according to the RNA transcript levels of the individual genes within that gene set, such as by measuring the levels of a specific mRNA that is the transcription product of a given SCCIGS gene. RNA transcripts may include, but are not limited to, pre-mRNA nascent transcript(s), transcript processing intermediates, mature mRNA(s), and degradation products, in addition to nucleic acid amplification products of such sequences (e.g., cDNAs).

Methods for isolating total mRNA, and for determining levels of specific mRNA transcripts therein, are well known to those of skill in the art. For example, methods of isolation and purification of nucleic acids from a biological sample are described in detail in Chapter 3 of Laboratory Techniques in Biochemistry and Molecular Biology: Hybridization With Nucleic Acid Probes, Part I. Theory and Nucleic Acid Preparation, P. Tijssen, ed. Elsevier, N.Y. (1993) and

Chapter 3 of Laboratory Techniques in Biochemistry and Molecular Biology: Hybridization With Nucleic Acid Probes, Part I. Theory and Nucleic Acid Preparation, P. Tijssen, ed. Elsevier, N.Y. (1993)). See also Ausubel et al. (Eds.), 2007 *Curr. Protocols Molec. Biol.*, Wiley, N.Y.

5 RNA levels may be determined according to techniques known in the art and exemplified herein. As non-limiting examples, RNA levels can be measured by utilizing arrays, such as RNA microarray-based techniques known in the art (see, e.g., Goley et al., *BMC Cancer* 4:20, 2004, performing RNA microarray analysis on needle core biopsies of tumors), and described herein
10 (see Example 1), or by utilizing quantitative reverse-transcriptase polymerase chain reactions (qRT-PCR) (see Example 2). As other examples, RNA levels can be determined by relying on other quantitative RNA assays known in the art (see, e.g., Lee et al., *Analytical Biochemistry* 357:299-301, 2006). RNA levels may, for instance, be determined by reverse transcribing the mRNA transcript of a given
15 gene to form a cDNA molecule, optionally amplifying the cDNA molecule, and measuring the levels of the DNA molecule, such as by quantitative real-time PCR (qRT-PCR, see, e.g., VanGuilder et al., *Biotechniques*. 44:619-26, 2008).

Examples of other useful techniques for determining the amount of nucleic acid target sequences (e.g., a mRNA transcript of a biomarker gene)
20 present in a sample based on specific hybridization of an oligonucleotide primer or probe to the target sequence include specific amplification of target nucleic acid sequences and quantification of amplification products, including but not limited to polymerase chain reaction (PCR, Gibbs et al., *Nucl. Ac. Res.* 17:2437, 1989), transcriptional amplification systems, strand displacement amplification
25 and self-sustained sequence replication (3SR, Gingeras et al., *J. Infect. Dis.* 164:1066, 1991). Examples of other useful techniques include ligase chain reaction (e.g., Landegren et al., *Science* 241:1077, 1988; Nickerson et al., *Proc. Natl. Acad. Sci. USA* 87:8923 1990; Barany, *Proc. Natl. Acad. Sci. USA* 88:189, 1991; Wu et al., *Genomics* 4:560, 1989), cycled probe technology and solid-
30 phase DNA-binding assays such as those disclosed in U.S. Patent No. 6,340,566, as well as other suitable methods that will be known to those familiar with the art.

In situ hybridization (ISH) using oligonucleotide probes also represents a widely used technique to measure mRNA levels, and thereby determine gene expression levels in tissue samples. ISH may be employed in respect to distribution as well as in respect to quantification of gene expression levels of a SCC biomarker (see, e.g., Erdtmann-Vourliotis *et al.* (*Brain Research Protocols*. 4:82-91, 1999).

Protein levels may also be determined according to certain embodiments, to determine the expression level(s) of an individual gene or genes in an SCCIGS, and thereby determine the SCCIGS expression level. Protein levels may be measured either directly, such as by measuring the amount of protein in an extract, cell, tissue, or other biological sample, or indirectly, such as by measuring the amount of protein activity in biological sample.

Protein levels may be measured from cell or tissue extracts, or from whole cells or tissues. Protein extraction from cell samples may be performed according to any of a number of methodologies with which the skilled person will be familiar. For instance, depending on whether the collection of particular cell fractions is desired, cell samples may be typically lysed using one or more of a hypotonic buffer, urea, a chaotrope (e.g., guanidine-HCl) and buffers containing various detergents, such as Nonidet™ NP-40, Triton™ -X100, Tween™-20, alkyl glucosides, betaine-containing surfactants, sodium dodecyl sulfate (SDS) or other detergents recognized in the art for this purpose, and further processed to be compatible with an intended assay. See, e.g., Sambrook *et al.* (*Molecular Cloning: A Laboratory Manual* (3rd Edition, 2001, Cold Spring Harbor Press, Cold Spring Harbor, NY). For protein level analysis in whole cells or tissues, cells or tissues can be either placed in an appropriate analysis buffer for live analysis, or fixed with various fixation agents, such as formaldehyde, paraformaldehyde, methanol, or ethanol, among others, followed by further processing according to the requirements of the intended assay (e.g., flow cytometry, immunohistochemistry, etc.).

Methods for measuring protein levels are known in the art. For example, protein or proteomic arrays may be used in certain embodiments, such as antibody microarrays, in which antibodies specific for one or more proteins of

interest are spotted onto a protein chip and are used as capture molecules to detect and quantify the proteins from biological samples, such as cell lysate solutions (see generally, Jones *et al.* *Nature* 439:168-174, 2006; and Chen *et al.* *Curr Opin Chem Biol* 10:28-34, 2006 for protein arrays). Antibodies specific for
5 those proteins expressed by the SCC biomarker genes or gene sets described herein may be generated using techniques known in the art (see, e.g., Harlow and Lane, eds. (1988) *Antibodies: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY; Freshney, (Ed.), *Culture of Animal Cells-5th Ed.*, 2005, Wiley-Liss, NY; ;Masters (Ed.), *Animal Cell Culture-3rd*
10 *Edition*, 2000 Oxford Univ. Press, NY).

Other examples of proteomic array based techniques that may be applied to clinical tissue samples include tissue microarrays and surface-enhanced laser desorption/ionization (SELDI-TOF) (see, e.g., Bertucci *et al.*, *Mol Cell Proteomics*. 5:1772-86, 2006; and Bollard *et al.*, *Proteomics: Clinical*
15 *Applications*, 1:934-954, 2007). Certain aspects may employ other techniques to measure protein levels, including, but not limited to, western blotting, radio-immunoprecipitation, proteomics, and flow cytometry (see, e.g., Prinz *et al.* *Proteomics*. 8:1179-96,2008; and Peterson *et al.*, *Toxicol Pathol*. 36:117-32, 2008). Immunohistochemistry-based assessment of protein expression also
20 provides a natural validation method of expression-profiling data that is easily performed on tissue samples (see, e.g., Sullivan *et al.*, *Clin Colorectal Cancer*. 7:172-177, 2008). A person skilled in the art will appreciate that the above-described methods are merely exemplary, and that any quantitative or semi-quantitative methods to determine gene expression levels may be employed to
25 practice the methods provided herein.

Generally, whether used to detect nucleic acid levels (e.g., mRNA levels, cDNA levels, etc.) or protein levels, an array typically comprises a solid support with peptide or nucleic acid-based probes attached covalently or non-covalently to the support, but the presently contemplated embodiments need not
30 be so limited and may also encompass assays based on fluid-phase interactions. In certain embodiments arrays typically comprise a plurality of different nucleic acid or peptide probes that are coupled to a surface of a substrate in discrete,

known locations. These arrays, also described as "microarrays" or "chips" have been generally described in the art, for example, in U.S. Pat. Nos. 5,143,854, 5,445,934, 5,744,305, 5,677,195, 6,040,193, 5,424,186 and Fodor *et al.*, *Science*, 251:767 777 (1991). Among other methods known in the art, these and
5 similar arrays may be produced using mechanical synthesis methods or light directed synthesis methods, incorporating a combination of photolithographic methods and solid phase synthesis methods. Techniques for the synthesis of these arrays using mechanical synthesis methods are described, for example, in U.S. Pat. Nos. 5,384,261, and 6,040,193. Arrays may be fabricated on a surface
10 of virtually any shape or even a multiplicity of surfaces. In certain embodiments, an array is fabricated on a planar array surface. Arrays may in other embodiments take the form of peptides or nucleic acids on beads, gels, polymeric surfaces, fibers such as fiber optics, glass or any other appropriate substrate, see U.S. Pat. Nos. 5,770,358, 5,789,162, 5,708,153, 6,040,193 and 5,800,992.

15 When utilizing RNA arrays or microarrays, quantitative real-time polymerase chain reaction (qRT-PCR), or other nucleic acid based protocols to determine the expression levels of a SCCIGS in a biological sample, certain embodiments may include the use of selected oligonucleotide probes. Oligonucleotide probes refer generally to polymers composed of a multiplicity of
20 nucleotide residues (deoxyribonucleotides or ribonucleotides, or related structural variants or synthetic analogues thereof) linked *via* phosphodiester bonds (or related structural variants or synthetic analogues thereof). Thus, while the term oligonucleotide probe typically refers to a nucleotide polymer in which the nucleotide residues and linkages between them are naturally occurring, it will be
25 understood that the term also includes within its scope various analogues including, but not restricted to, peptide nucleic acids (PNAs), phosphoramidates, phosphorothioates, methyl phosphonates, 2-O-methyl ribonucleic acids, and the like.

The exact size of an oligonucleotide probe can vary depending on
30 the particular application. An oligonucleotide is typically rather short in length, generally from about 10 to 30 nucleotide residues (*e.g.*, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 nucleotide residues), but

the term can refer to molecules of any length, including oligonucleotides from about 30 to about 100 or more nucleotide residues in length (*e.g.*, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 65, 70, 75, 80, 85, 90, 95, 100, or more, including all
5 integers in between). Among other features, the length of an oligonucleotide probe may affect its ability specifically to bind or hybridize to its intended target sequence (*e.g.*, a complementary sequence, according to well-established principles of Watson-Crick base pairing), which refers generally to its ability to bind more strongly to its intended target sequence than to any other sequences
10 in a given sample, and, thus, to discriminate between its intended target and the other sequences present in the sample. In certain preferred embodiments, an oligonucleotide probe may be about 25 nucleotide residues in length, including, for example, certain Affymetrix® probes (see Example 1).

Oligonucleotide probes may be selected or designed according to
15 routine techniques known in the art and described herein for their ability to specifically bind or hybridize to an intended target sequence. In particular, polynucleotides or oligonucleotide probes may be designed or selected to specifically hybridize or bind to the individual genes in certain preferred SSCIGSs (see Figure 3), or may be designed or selected to specifically hybridize or bind
20 more to one or more SCC gene biomarkers, as described herein (see Figures 2 and 7; and SEQ ID NOS:1-200). A polynucleotide or oligonucleotide probe may also be selected to specifically bind or hybridize to variants, whether naturally-occurring or otherwise (*e.g.*, allelic variants, splice variants), of the SCC biomarker genes described herein.

25 To specifically bind to its intended target sequence, an oligonucleotide probe typically comprises a polynucleotide sequence that is complementary to at least a portion of the polynucleotide sequence of a target gene, *e.g.*, the exemplary target genes of SEQ ID NOS:1-200. For polynucleotides, complementary or complementarity refers generally to
30 polynucleotides related by the well-known nucleotide base-pairing rules. For example, the sequence "A-G-T," is complementary to the sequence "T-C-A." Complementarity may be "partial," in which only some of the nucleic acids' bases

are matched according to the base pairing rules, or there may be "full" or "total" complementarity between the nucleic acids. The degree of complementarity between nucleic acid strands has significant effects on the efficiency and strength of hybridization between nucleic acid strands.

5 By "corresponds to" or "corresponding to" is typically meant a polynucleotide having a nucleotide sequence that is substantially identical or complementary to all or a portion of a reference polynucleotide sequence. In certain preferred embodiments, an oligonucleotide probe is fully complementary to at least a portion of the polynucleotide sequence of a target gene, since
10 simultaneous consideration of the percent similarity ($\leq 90\%$), the length of identical sequence stretches (≤ 20 bases), and the binding free energy (≥ -35 kcal mol⁻¹) may be predictive of probe specificity (see, e.g., Liebich *et al.*, *Appl Environ Microbiol.* 72:1688–1691, 2006). A polynucleotide or probe is fully complementary when there are no base mismatches between the probe and the
15 relevant portion of the target sequence.

In designing an oligonucleotide probe or other polynucleotide sequence that is complementary to the polynucleotide sequence of the target gene, certain characteristics of such probes or sequences may be considered to optimize the ability of the probe to specifically hybridize and detect the target
20 sequence. For example, to avoid false positives, if there is substantial sequence information available for a given source organism (e.g., human) or cell type (e.g., oral epithelial cell), oligonucleotide probes may be chosen that are not similar to any other expressed sequences in that organism or cell type.

It is also known in the art that certain polynucleotide sequences, such as those containing inverted repeats, may be able to self-hybridize and form
25 secondary structures that interfere with specific detection of target sequences. Typically, such sequences may be avoided to improve probe specificity.

As another useful feature of probe design, in certain embodiments oligonucleotide probes of "high complexity," as opposed to probes of "low
30 complexity," may provide more specific target sequence detection. One example of a probe with low complexity includes "AAAAAAA GGAGTTTTTTTTT CAAAAAAGCTTTT" (SEQ ID NO:332). One example of a probe

with higher complexity includes "CGTGACTGA CAGCTGACTGC TAGCCATGCAAC" (SEQ ID NO:333). A fast and flexible approach to oligonucleotide probe design for genomes and gene families may be found, for example, in Feng *et al.* (*Bioinformatics* 23:1195-1202, 2007) and He *et al.*

5 (*Applied and Environmental Microbiology*, 71:3753-3760, 2005). One example of rational oligonucleotide probe design for *in situ* hybridization protocols may be found in Erdtmann-Vourliotis *et al.* (*Brain Research Protocols*. 4:82-91, 1999).

Certain exemplary oligonucleotide probes that may be used in the methods described herein include the Affymetrix oligonucleotide probes
10 described in SEQ ID NOS:201-331 (see, e.g., Figures 2 and 8). Also contemplated are polynucleotide variants of these probes that are capable of specifically hybridizing or binding to a SCC biomarker gene, including one or more genes identified as a SSCIGS. Polynucleotide variants refer to either a polynucleotide that displays substantial sequence identity with a reference
15 polynucleotide sequence (e.g., at least 70, 75, 80, 85, 90, 91, 92, 93, 94, 95, 96, 97, 98, or 99% identity to SEQ ID NOS:201-331) or a polynucleotide that hybridizes with a reference sequence, or its complementary sequence, under moderate or stringent conditions that are described hereinafter.

Oligonucleotide probes may be modified according to techniques
20 known in the art, such as to improve stability or facilitate detection. For example, oligonucleotide probes may be modified by directly attaching thereto one or more detectable molecules, as described below. Oligonucleotide probes may also be modified by attaching thereto one or more ligand molecules, such as biotin, that may be used to indirectly attach a detectable molecule, such as a
25 detectable molecule that is bound to one or more avidin molecules.

A variety of detectable molecules may be used to render an oligonucleotide probe detectable, such as a radioisotopes, fluorochromes, dyes, enzymes, nanoparticles, chemiluminescent markers, biotin, or other monomer known in the art that can be detected directly (e.g., by light emission) or indirectly
30 (e.g., by binding of a fluorescently-labeled antibody).

Radioisotopes provide examples of detectable molecules that can be utilized in certain aspects of the present invention. Several radioisotopes can

be used as detectable molecules for labeling nucleotides or proteins, including, for example, ^{32}P , ^{33}P , ^{35}S , ^3H , and ^{125}I . These radioisotopes have different half-lives, types of decay, and levels of energy which can be tailored to match the needs of a particular protocol. For example, ^3H is a low energy emitter which results in low background levels, however this low energy also results in long time periods for autoradiography. Radioactively labeled ribonucleotides, deoxyribonucleotides and amino acids are commercially available. Nucleotides are available that are radioactively labeled at the first, or α , phosphate group, or the third, or γ , phosphate group. For example, both [α - ^{32}P] dATP and [γ - ^{32}P] dATP are commercially available. In addition, different specific activities for radioactively labeled nucleotides are also available commercially and can be tailored for different protocols.

Other examples of detectable molecules that can be utilized to detect an oligonucleotide probe include fluorophores. Several fluorophores can be used for labeling nucleotides including, for example, fluorescein, tetramethylrhodamine, Texas Red, and a number of others (e.g., Haugland, *Handbook of Fluorescent Probes - 9th Ed.*, 2002, Molec. Probes, Inc., Eugene OR; Haugland, *The Handbook: A Guide to Fluorescent Probes and Labeling Technologies-10th Ed.*, 2005, Invitrogen, Carlsbad, CA).

Non-radioactive and non-fluorescent detectable molecules are also available. As noted above, biotin can be attached directly to nucleotides and detected by specific and high affinity binding to avidin or streptavidin which has been chemically coupled to an enzyme catalyzing a colorimetric reaction (such as phosphatase, luciferase, or peroxidase). Digoxigenin labeled nucleotides can also similarly be used for non-isotopic detection of nucleic acids. Biotinylated and digoxigenin-labeled nucleotides are commercially available.

Very small particles, termed nanoparticles, also can be used to label oligonucleotide probes. These particles range from 1-1000 nm in size and include diverse chemical structures such as gold and silver particles and quantum dots. When irradiated with angled incident white light, silver or gold nanoparticles ranging from 40-120 nm will scatter monochromatic light with high intensity. The wavelength of the scattered light is dependent on the size of the

particle. Four to five different particles in close proximity will each scatter monochromatic light, which when superimposed will give a specific, unique color. The particles are being manufactured by companies such as Genicon Sciences (Carlsbad, CA). Derivatized silver or gold particles can be attached to a broad
5 array of molecules including, proteins, antibodies, small molecules, receptor ligands, and nucleic acids. For example, the surface of the particle can be chemically derivatized to allow attachment to a nucleotide.

Other types of nanoparticles that can be used for detection of a detectable molecule include quantum dots. Quantum dots are fluorescing
10 crystals 1-5 nm in diameter that are excitable by light over a large range of wavelengths. Upon excitation by light having an appropriate wavelength, these crystals emit light, such as monochromatic light, with a wavelength dependent on their chemical composition and size. Quantum dots such as CdSe, ZnSe, InP, or InAs possess unique optical properties; these and similar quantum dots are
15 available from a number of commercial sources (e.g., NN-Labs, Fayetteville, AR; Ocean Nanotech, Fayetteville, AR; Nanoco Technologies, Manchester, UK; Sigma-Aldrich, St. Louis, MO).

Many dozens of classes of particles can be created according to the number of size classes of the quantum dot crystals. The size classes of the
20 crystals are created either 1) by tight control of crystal formation parameters to create each desired size class of particle, or 2) by creation of batches of crystals under loosely controlled crystal formation parameters, followed by sorting according to desired size and/or emission wavelengths. Two examples of references in which quantum dots are embedded within intrinsic silicon epitaxial
25 layers of semiconductor light emitting/detecting devices are United States Patent Nos. 5,293,050 and 5,354,707 to Chapple Sokol, *et al.*

In certain embodiments, oligonucleotide probes may be labeled with one or more light-emitting dyes. The light emitted by the dyes can be visible light or invisible light, such as ultraviolet or infrared light. In exemplary embodiments,
30 the dye may be a fluorescence resonance energy transfer (FRET) dye; a xanthene dye, such as fluorescein and rhodamine; a dye that has an amino group in the alpha or beta position (such as a naphthylamine dye, 1-

dimethylaminonaphthyl-5-sulfonate, 1-anilino-8-naphthalene sulfonate and 2-p-touidiny-6-naphthalene sulfonate); a dye that has 3-phenyl-7-isocyanatocoumarin; an acridine, such as 9-isothiocyanatoacridine and acridine orange; a pyrene, a bensoxadiazole and a stilbene; a dye that has 3-(ϵ -carboxypentyl)-3'-ethyl-5,5'-dimethyloxacarbocyanine (CYA); 6-carboxy fluorescein (FAM); 5&6-carboxyrhodamine-110 (R110); 6-carboxyrhodamine-6G (R6G); N,N,N',N'-tetramethyl-6-carboxyrhodamine (TAMRA); 6-carboxy-X-rhodamine (ROX); 6-carboxy-4',5'-dichloro-2',7'-dimethoxyfluorescein (JOE); ALEXA Fluor™; Cy2; Texas Red and Rhodamine Red; 6-carboxy-2',4,7,7'-tetrachlorofluorescein (TET); 6-carboxy-2',4,4',5',7,7'-hexachlorofluorescein (HEX); 5-carboxy-2',4',5',7'-tetrachlorofluorescein (ZOE); NAN; NED; Cy3; Cy3.5; Cy5; Cy5.5; Cy7; and Cy7.5; Alexa Fluor 350; Alexa Fluor 488; Alexa Fluor 532; Alexa Fluor 546; Alexa Fluor 568; Alexa Fluor 594; or Alexa Fluor 647.

A detectable molecule can be directly attached to a nucleotide using methods well known in the art. Nucleotides can also be chemically modified or derivatized in order to attach a detectable molecule. For example, a fluorescent monomer such as a fluorescein molecule can be attached to dUTP (deoxyuridine-triphosphate) using a four-atom aminoalkynyl group. In this example, each detectable molecule may be attached to a nucleotide making a detectable molecule: nucleotide complex. Amine-reactive and thiol-reactive fluorophores are available and may be used for labeling nucleotides and biomolecules.

As one example, nucleotides may be fluorescently labeled during chemical synthesis, since incorporation of amines or thiols during nucleotide synthesis permit addition of fluorophores. Fluorescently labeled nucleotides are commercially available. For example, uridine and deoxyuridine triphosphates are available that are conjugated to ten different fluorophores that cover the spectrum. Fluorescent dyes that can be bound directly to nucleotides can also be utilized as detectable molecules. For example, FAM, JOE, TAMRA, and ROX are amine reactive fluorescent dyes that have been attached to nucleotides and are used in automated DNA sequencing. These fluorescently labeled

nucleotides, for example, ROX-ddATP, ROX-ddCTP, ROX-ddGTP and ROX-ddUTP, are commercially available.

As noted herein, the terms specifically binds or specifically hybridizes refer generally to an oligonucleotide probe or polynucleotide sequence that not only binds to its intended target gene sequence in a sample under selected hybridization conditions, but does not bind significantly to other target sequences in the sample, and thereby discriminates between its intended target and all other targets in the target pool. A probe that specifically hybridizes to its intended target sequence may also detect concentration differences under the selected hybridization conditions.

An intended target sequence refers typically to a polynucleotide or nucleic acid sequence, which refers generally to mRNA, RNA, cRNA, cDNA or DNA (*i.e.*, polymeric forms of nucleotides of at least 10 bases in length, either ribonucleotides or deoxynucleotides or a modified form of either type of nucleotide). In certain embodiments, an oligonucleotide probe may specifically bind or specifically hybridize to at least a portion of one polynucleotide having a sequence selected from SEQ ID NOS:1-200, including variants thereof (*e.g.*, allelic variants, splice variants, etc.) that is differentially expressed in SCC cells (*e.g.*, OSCC) compared to control cells are known to be free of SCC cells.

In certain embodiments, an oligonucleotide probe specifically hybridizes to one or more of a polynucleotide having all or a SCCIG-characteristic portion of a SCCIG transcript (*e.g.*, mRNA transcript of a SCC biomarker gene, or a fragment thereof), a polynucleotide having a nucleotide sequence that is fully complementary to a SCCIG-characteristic portion of a SCCIG transcript, and/or a nucleic acid amplification product of the above noted polynucleotides (*e.g.*, cDNA of a SCC biomarker gene or SSCIGS, or fragment thereof). A nucleic acid amplification product may be obtained from a biological sample, for example, by performing RT-PCR on a sample cell extract that contains a polynucleotide, such as an mRNA transcript, having all or a SCCIG-characteristic portion of a SCCIG transcript. A SCCIG-characteristic portion of a SSCIG transcript refers to a segment, stretch, domain, region, portion or the like of any one of the polynucleotides set forth as SEQ ID NOS: 1-200 (or the full complement thereof),

which comprises less than the full-length polynucleotide of the respective one of SEQ ID NOS: 1-200, and which has a nucleotide sequence that is unique to that particular sequence among all polynucleotide transcript sequences found in the species from which the SSCIG set is obtained (e.g., the human transcriptome)

5 such that an oligonucleotide probe that hybridizes specifically to the SSCIG-characteristic portion does not exhibit full complementarity to any other transcript in the subject transcriptome. Accordingly, for example, a SSCIG-characteristic portion of any one of SEQ ID NOS: 1-200 may be derived from the biomarker gene sequences set forth in SEQ ID NOS:1-200, or from variants thereof.

10 Nucleic acid hybridization conditions include those described herein and known in the art for controlled, detectable annealing of a first oligonucleotide or polynucleotide sequence to a second oligonucleotide or polynucleotide sequence (see Examples 1 and 2), and will often vary depending on the particular application. The term "hybridizes under low stringency, medium
15 stringency, high stringency, or very high stringency conditions" refers generally to conditions for hybridization and washing. Guidance for performing hybridization reactions can be found in Sections 6.3.1-6.3.6 of Ausubel *et al.*, ("Current Protocols in Molecular Biology", John Wiley & Sons Inc, 1994-1998, Chapter 15). Aqueous and non-aqueous methods are described in that reference, and either
20 class of method can be used according to embodiments contemplated herein.

Low stringency conditions referred to herein may include and encompass from at least about 1% v/v to at least about 15% v/v formamide and from at least about 1 M to at least about 2 M salt for hybridization at 42°C, and at least about 1 M to at least about 2 M salt for washing at 42°C. Low stringency
25 conditions also may include 1% Bovine Serum Albumin (BSA), 1 mM EDTA, 0.5 M NaHPO₄ (pH 7.2), 7% SDS for hybridization at 65°C, and (i) 2 × SSC, 0.1% SDS; or (ii) 0.5% BSA, 1 mM EDTA, 40 mM NaHPO₄ (pH 7.2), 5% SDS for washing at room temperature. One embodiment of low stringency conditions includes hybridization in 6 × sodium chloride/sodium citrate (SSC) at about 45°C,
30 followed by two washes in 0.2 × SSC, 0.1% SDS at least at 50°C (the temperature of the washes can be increased to 55° C for low stringency conditions).

In certain embodiments, medium stringency conditions may include and encompass from at least about 16% v/v to at least about 30% v/v formamide and from at least about 0.5 M to at least about 0.9 M salt for hybridization at 42°C, and at least about 0.1 M to at least about 0.2 M salt for washing at 55°C.

- 5 Medium stringency conditions also may include 1% Bovine Serum Albumin (BSA), 1 mM EDTA, 0.5 M NaHPO₄ (pH 7.2), 7% SDS for hybridization at 65°C, and (i) 2 × SSC, 0.1% SDS; or (ii) 0.5% BSA, 1 mM EDTA, 40 mM NaHPO₄ (pH 7.2), 5% SDS for washing at 60-65°C. One embodiment of medium stringency conditions includes hybridizing in 6 × SSC at about 45°C, followed by one or
10 more washes in 0.2 × SSC, 0.1% SDS at 60°C.

- In certain embodiments, high stringency conditions may include and encompass from at least about 31% v/v to at least about 50% v/v formamide and from about 0.01 M to about 0.15 M salt for hybridization at 42°C, and about 0.01 M to about 0.02 M salt for washing at 55°C. High stringency conditions also may
15 include 1% BSA, 1 mM EDTA, 0.5 M NaHPO₄ (pH 7.2), 7% SDS for hybridization at 65°C, and (i) 0.2 × SSC, 0.1% SDS; or (ii) 0.5% BSA, 1 mM EDTA, 40 mM NaHPO₄ (pH 7.2), 1% SDS for washing at a temperature in excess of 65° C. One embodiment of high stringency conditions includes hybridizing in 6 × SSC at about 45°C, followed by one or more washes in 0.2 × SSC, 0.1% SDS at 65°C.
20 One embodiment of very high stringency conditions includes hybridizing in 0.5 M sodium phosphate, 7% SDS at 65°C, followed by one or more washes in 0.2 × SSC, 1% SDS at 65°C.

- Other stringency conditions are well known in the art and skilled persons will recognize that various factors can be manipulated to optimize the
25 specificity of the hybridization. Optimization of the stringency of the final washes can serve to ensure a high degree of hybridization. For detailed examples, see Ausubel *et al.*, *supra* at pages 2.10.1 to 2.10.16 and Sambrook *et al.* (*Molecular Cloning: A Laboratory Manual* (3rd Edition, 2001)), and Maniatis *et al.* (*Molecular Cloning: A Laboratory Manual* (1982)).

- 30 While stringent washes are typically carried out at temperatures from about 42°C to about 68°C, one skilled in the art will appreciate that other temperatures may be suitable for stringent conditions. Maximum hybridization

rate typically occurs at about 20°C to 25°C below the T_m for formation of a DNA-DNA hybrid. It is well known in the art that the T_m is the melting temperature, or temperature at which two complementary polynucleotide sequences dissociate. Methods for estimating T_m are well known in the art (see Ausubel *et al.*, *supra* at
5 page 2.10.8).

In general, the T_m of a perfectly matched duplex of DNA may be predicted as an approximation by the formula: $T_m = 81.5 + 16.6 (\log_{10} M) + 0.41 (\%G+C) - 0.63 (\% \text{ formamide}) - (600/\text{length})$ wherein: M is the concentration of Na^+ , preferably in the range of 0.01 molar to 0.4 molar; %G+C is the sum of
10 guanosine and cytosine bases as a percentage of the total number of bases, within the range between 30% and 75% G+C; % formamide is the percent formamide concentration by volume; length is the number of base pairs in the DNA duplex. The T_m of a duplex DNA decreases by approximately 1° C with every increase of 1% in the number of randomly mismatched base pairs.
15 Washing is generally carried out at $T_m - 15^\circ\text{C}$ for high stringency, or $T_m - 30^\circ\text{C}$ for moderate stringency.

In one example of a hybridization procedure, a membrane (*e.g.*, a nitrocellulose membrane or a nylon membrane) or chip containing immobilized DNA is hybridized overnight at 42°C in a hybridization buffer (50% deionized
20 formamide, 5 × SSC, 5 × Denhardt's solution (0.1% ficoll, 0.1% polyvinylpyrrolidone and 0.1% bovine serum albumin), 0.1% SDS and 200 mg/mL denatured salmon sperm DNA) containing a labeled probe. The membrane is then subjected to two sequential medium stringency washes (*i.e.*, 2 × SSC, 0.1% SDS for 15 min at 45°C, followed by 2 × SSC, 0.1% SDS for 15 min at 50°C),
25 followed by two sequential higher stringency washes (*i.e.*, 0.2 × SSC, 0.1% SDS for 12 min at 55°C followed by 0.2 × SSC and 0.1% SDS solution for 12 min at 65-68°C).

For RNA microarrays, hybridization can be performed according to the Examples provided herein (see Example 1) or per other protocols known in
30 the art. For example, a cRNA may be hybridized to an Affymetrix U133 2.0 Plus GeneChip array and scanned using an Affymetrix GeneChip array Scanner 3000 7G per Affymetrix protocols. (Affymetrix Corp., Santa Clara, CA).

For PCR based protocols, such as qRT-PCR, hybridization can be performed according to the Examples provided herein (see Example 2) or per other protocols known in the art (see, e.g., Skrzypski *et al.*, *Lung Cancer*. 59:147-54, 2008; and Bustin *et al.*, *Clin Sci (Lond)*. 109:365-79, 2005). The use of
5 particular hybridization temperatures and PCR-based reaction buffers according to the characteristics of one or more selected oligonucleotide probe(s) are well known in the art (see, e.g., Sambrook *et al.*, *supra*). A person skilled in the art will appreciate that these and other similar protocols may be used to specifically hybridize an oligonucleotide probe to its intended target sequence, and to
10 measure the biomarker gene or SSCIGS expression levels therefrom.

In certain embodiments, the methods of identifying a risk for having, or presence of, SCC in a subject may include comparing SSCIGS expression levels in (i) a reference sample that is from a subject known to be free from SCC, with (ii) SSCIGS expression levels in a suspected biological sample, wherein the
15 differential expression of the SSCIGS indicates the subject has, or is at risk for having, SCC. Differential expression of a SSCIGS refers generally to a statistically significant difference, in a biological sample from a subject that is suspected of having or being at risk for having SCC, in one or more gene expression levels of SCC biomarker(s) or SSCIGS members(s) as compared to
20 the expression levels of the same SSC biomarker(s) or SSCIGS member(s) in an appropriate cancer-free control. The statistically significant difference may relate to either an increase or a decrease in expression levels, as measured by RNA levels, protein levels, protein function, or any other relevant measure of gene expression such as those described herein.

25 A result is typically referred to as statistically significant if it is unlikely to have occurred by chance. The significance level of a test or result relates traditionally to a frequentist statistical hypothesis testing concept. In simple cases, statistical significance may be defined as the probability of making a decision to reject the null hypothesis when the null hypothesis is actually true (a
30 decision known as a Type I error, or "false positive determination"). This decision is often made using the p-value: if the p-value is less than the significance level, then the null hypothesis is rejected. The smaller the p-value, the more significant

the result. Bayes factors may also be utilized to determine statistical significance (see, e.g., Goodman S., *Ann Intern Med* 130:1005-13, 1999).

In more complicated, but practically important cases, the significance level of a test or result may reflect an analysis in which the probability of making a decision to reject the null hypothesis when the null hypothesis is actually true is *no more than* the stated probability. This type of analysis allows for those applications in which the probability of deciding to reject may be much smaller than the significance level for some sets of assumptions encompassed within the null hypothesis.

10 In certain exemplary embodiments, statistically significant differential expression may include situations wherein the expression level of a given SSCIGS provides at least about a 1.2X, 1.3X, 1.4X, 1.5X, 1.6X, 1.7X, 1.8X, 1.9X, 2.0X, 2.2X, 2.4X, 2.6X, 2.8X, 3.0X, 4.0X, 5.0X, 6.0X, 7.0X, 8.0X, 9.0X, 10.0X, 15.0X, 20.0X, 50.0X, 100.0X, or greater difference in expression (*i.e.*,
15 differential expression that may be higher or lower expression) in a suspected biological sample as compared to an appropriate control, including all integers and decimal points in between (e.g., 1.24X, 1.25X, 2.1X, 2.5X, 60.0X, 75.0X, etc.). In certain embodiments, statistically significant differential expression may include situations wherein the expression level of a given SSCIGS provides at
20 least about 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 percent (%) or greater difference in expression (*i.e.*, differential expression that may be higher or lower) in a suspected biological sample as compared to an appropriate control, including all integers and decimal points in between.

25 As an additional example, differential expression may also be determined by performing Z-testing, *i.e.*, calculating an absolute Z score, as described herein and known in the art (see Example 1). Z-testing is typically utilized to identify significant differences between a sample mean and a population mean. For example, as compared to a standard normal table (e.g., a
30 control tissue), at a 95% confidence interval (*i.e.*, at the 5% significance level), a Z-score with an absolute value greater than 1.96 indicates non-randomness. For a 99% confidence interval, if the absolute Z is greater than 2.58, it means that

$p < .01$, and the difference is even more significant—the null hypothesis can be rejected with greater confidence. In these and related embodiments, an absolute Z-score of 1.96, 2, 2.58, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more, including all decimal points in between (e.g., 10.1, 10.6, 11.2, etc.), may provide a strong measure of statistical significance. In certain embodiments, an absolute Z-score of greater than 6 may provide exceptionally high statistical significance.

In certain embodiments, such as when using an Affymetrix Microarray to measure the expression levels of a SSCIGS, differential expression may also be determined by the mean expression value summarized by Affymetrix Microarray Suite 5 software (Affymetrix, Santa Clara, CA), or other similar software, typically with a scaled mean expression value of 1000.

A control tissue or reference tissue refers generally to a cell-based sample, such as an epithelial cell sample, that is known to be free of SCC cells according to currently accepted diagnostic criteria such as those described herein, and may also in certain embodiments relate to an epithelial cell-containing sample that is free of dysplastic cells. “Epithelial” cells refers generally to any one or more of many types of closely packed cells that form the epithelium covering the body (e.g., skin) and the linings of body cavities, for instance, membranous mucosal tissue covering internal organs and other internal surfaces of the body (e.g., the inside of the mouth, the respiratory tract, the gastrointestinal tract, etc.). A person skilled in the art can readily obtain a control sample and determine if it is free of SCC or dysplasia according to techniques known in the art and described herein, e.g., by using pathological, cytological, or molecular biological techniques (see, e.g., Mehrotra *et al.*, *Mol Cancer*. 5:11, 2006; and Civantos *et al.*, *J Surg Oncol*. 97:683-90, 2008). As one example, control tissues may be obtained from individuals undergoing tonsillectomy or oral surgery for treatment of diseases other than cancer, such as obstructive sleep apnea (see Example 1). Control tissues, or an appropriate extract thereof, may also be obtained from commercially available sources.

A control tissue or reference tissue may be either internal (i.e., from the same subject as the biological sample) or external (i.e., from a source or

subject that is different from the biological sample). In certain embodiments, such as those wherein the biological sample is suspected of containing an OSCC cell, a control cell or reference cell may be obtained as, or derived from, an oral epithelial cell. In certain embodiments, such as those wherein the biological sample is suspected of containing an HNSCC cell, a control cell or reference cell may be obtained as, or derived from, a normal epithelium from the pharynx, hypopharynx, larynx, oral cavity, sinus tissue, or other appropriate control tissue.

Certain embodiments of the present invention relate to the use of the herein described SCC biomarker genes or SSCIGS that can differentiate between SCC tumor cells (*e.g.*, OSCC tumor cells) and dysplastic cells, *i.e.*, dysplasia, a pre-neoplastic cellular state typically considered at risk for developing into SCC carcinoma *in situ* or invasive carcinoma, although not necessarily destined to do so. Certain of these and related embodiments include methods for identifying the risk or presence of OSCC in a subject having oral epithelial dysplasia but no frank OSCC (*i.e.*, OSCC is not readily apparent), as described herein, such as by comparing the expression levels of one or more selected SSCIGSs in a biological sample from that subject with the expression levels of a reference SSCIGS that is characteristic of a OSCC tumor cell, wherein the substantial similarity of the selected SSCIGS between the biological sample and the reference sample indicates the presence or risk of OSCC. A subject having no frank OSCC may include a subject having no clinically detectable SCC carcinoma, typically as determined by standard diagnostic techniques known in the art and described herein.

Substantial similarity relates generally to the lack of a statistically significant difference in the expression levels between the biological sample and the reference control. Examples of substantially similar expression levels may include situations wherein the expression level of a given SSCIGS provides less than about a .05X, 0.1X, 0.2X, 0.3X, 0.4X, 0.5X, 0.6X, 0.7X, 0.8X, 0.9X, 1.0X, 1.1X, 1.2X, 1.3X, or 1.4X difference in expression (*i.e.*, differential expression that may be higher or lower expression) in a suspected biological sample as compared to an OSCC reference sample, including all decimal points in between (*e.g.*, .15X, 0.25X, 0.35X, etc.). In certain embodiments, differential expression

may include situations wherein the expression level of a given SSCIGS provides less than about 0.25, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50 percent (%) difference in expression (*i.e.*, differential expression that may be higher or lower) in a suspected biological sample as compared to a reference sample, including all decimal points in between.

In certain embodiments, the particular SCC biomarker genes (*e.g.*, SCCIG) described in Figure 6, including variants thereof, may be utilized to differentiate between SCC tumor cells and dysplastic cells, since as described herein these genes are differentially expressed between SCC tumor cells and dysplastic/control cells. Similarly, certain embodiments may include the use of oligonucleotide probes in the methods for differentiating between SCC tumor cells and dysplastic cells, such as the exemplary probes described in Figures 6 and 9, including variants thereof that can specifically hybridize to a differentially expressed SCC biomarker or SSCIG.

When employing selected oligonucleotide probes to differentiate between SCC tumor cells and dysplastic cells, the oligonucleotide probe may be specifically hybridized to one or more of (i) a polynucleotide having all or a SCCIG-characteristic portion of a SCCIG transcript as provided herein (*e.g.*, mRNA transcript of a SCC biomarker gene, or a fragment thereof), (ii) a polynucleotide having a nucleotide sequence that is fully complementary to a SCCIG-characteristic portion of a SCCIG transcript, and/or (iii) a nucleic acid amplification product of the above noted polynucleotides (*e.g.*, cDNA of a SCC biomarker gene or SSCIGS, or fragment thereof). As noted above, a nucleic acid amplification product may be obtained from a biological sample, for example, by performing reverse transcription-polymerase chain reaction amplification (RT-PCR) of polynucleotide transcription products such as mRNAs that are present in a cell extract and that have all or a SCCIG-characteristic portion of a SCCIG transcript. A SCCIG-characteristic portion of a SSCIG transcript may be derived from the biomarker gene sequences or SSCIGS set forth in SEQ ID NOS:1-200, and in particular, in the biomarker genes referred to in Figure 6, including variants thereof that are differentially expressed in a SCC tumor cell as compared to a dysplastic cell.

* * * * *

5 All of the U.S. patents, U.S. patent application publications, U.S. patent applications, foreign patents, foreign patent applications and non-patent publications referred to in this specification are incorporated herein by reference in their entireties.

10 Although the foregoing invention has been described in some detail to facilitate understanding, it will be apparent that certain changes and modifications may be practiced within the scope of the appended claims. Accordingly, the described embodiments, including the Examples below, are to be considered as illustrative and not restrictive, and the invention is not to be limited to the details given herein, but may be modified within the scope and equivalents of the appended claims.

EXAMPLES

EXAMPLE 1

IDENTIFICATION OF SQUAMOUS CELL CARCINOMA INDICATOR GENE SETS (SCCIGS)

To identify biomarkers or indicator gene sets for squamous cell carcinoma (SCC), tissues were obtained from patients either having or suspected of having oral SCC (OSCC) and from control patients. Eligible cases were patients with their first primary OSCC scheduled for surgical resection or biopsy between December 1, 2003 and April 17, 2007 at the University of Washington Medical Center, Harborview Medical Center and the VA Puget Sound Health Care System in Seattle, Washington. Patients with diagnosed dysplastic lesions were also enrolled at these medical centers during the same period. Eligible controls were patients who had tonsillectomy or oral surgery for treatment of diseases other than cancer, *e.g.*, obstructive sleep apnea, at the same institutions and during the same time periods in which the OSCC cases were treated.

Among 244 eligible OSCC patients, consent was obtained from 187 patients. Of these, 171 patients gave permission for medical chart abstraction and provided sufficient tissue to yield GeneChip arrays results that passed quality control (QC) criteria (see below). Among 21 eligible dysplasia cases, 15 provided consent for the study. Of these, the GeneChip results from 11 patients passed QC checks. One dysplasia patient provided dysplasia tissues from two different sites. One OSCC patient provided one piece of cancer tissue and one piece of dysplasia tissue, and assay results from this latter tissue were grouped with the dysplasia patients. Four of the eligible patients originally believed to have OSCC had a final pathology report of dysplasia, and the results from these patients were included in the dysplasia group, and not in the OSCC group for analyses. In total, 17 dysplasia samples were used for analysis. During the case recruitment period, 47 of 55 eligible controls consented to participate. Samples from two controls failed QC checks, leaving 45 control samples for analysis.

Each participant was interviewed using a structured questionnaire regarding demographic, medical, functional, quality of life, and lifestyle history,

including tobacco and alcohol use. Tumor characteristics (site, stage) were obtained from medical records. This study was conducted with written informed consent and Institutional Review Office approvals.

Tumor tissue was obtained at time of resection or biopsy from
5 patients with a primary OSCC, or dysplasia. Clinically normal tissue from the oral cavity or oropharynx was obtained from controls. For the small number of controls (~30%) with tonsillitis or tonsil hypertrophy, only mucosa tissue from tonsillar pillar was obtained to avoid potential influence of inflammation on the results. Immediately after surgical removal, the tissue was immersed in
10 RNALater (Applied Biosystems, Inc. Foster City, CA) for a minimum of 12 hours at 4°C before being transferred to long term storage at 80°C prior to use.

For DNA Microarray analysis, total RNA was extracted, purified, processed using a GeneChip Expression 3'-Amplification Reagents Kit (Affymetrix, Santa Clara, CA), and interrogated with an Affymetrix U133 2.0 Plus
15 GeneChip arrays. Specifically, the RNA and DNA from each specimen was simultaneously extracted using the TRIzol method (Invitrogen, Carlsbad, CA). To increase DNA purity, the DNA extraction protocol was modified to include the use of a "back extraction buffer" (4 M guanidine thiocyanate, 50 mM sodium citrate, and 1 M Tris, pH 8.0). RNA was further purified with the use of an RNeasy mini
20 kit (Qiagen, Valencia, CA) per Affymetrix (Santa Clara, CA) recommendations.

For expression array analysis, 1.0 to 2.5 µg of total RNA was used to generate biotin-labeled cRNA using the GeneChip Expression 3'-Amplification Reagents Kit (Affymetrix) per manufacturer's protocol. The cRNA was hybridized to an Affymetrix U133 2.0 Plus GeneChip array and scanned using an
25 Affymetrix GeneChip array Scanner 3000 7G in the Fred Hutchinson Cancer Research Center's Genomics Shared Resources per Affymetrix protocols. At least one clinically normal tissue sample from a control subject was processed in tandem with every seven to eight tumor tissue samples from OSCC cases.

Two rounds of QC checks were conducted to evaluate whether to
30 include results from each of the GeneChips. In the first round, recommendations made by Affymetrix were followed. In the second round, the "affyQCReport" and "affyPLM" software in the Bioconductor package (<http://www.bioconductor.org>)

were used to search for poor quality chips. In total, 172 chips from 165 patients (119 OSCC patients, 35 controls and 11 dysplasia patients passed two rounds of QC evaluation.

Preprocessing and probe set filtering was performed on the
5 GeneChip arrays that passed QC checks. To this end, the gcRMA algorithm from Bioconductor was used to extract gene expression values and perform normalization. Next, to limit the multiple testing penalty in the statistical testing step, probe sets were eliminated that either showed no variation across the samples being compared (*i.e.*, inter quartile range (IQR) of expression levels
10 less than 0.1 on log₂ scale) or were expressed at very low magnitude (*i.e.*, any probe set in which the maximum expression value for that probe set in any of the samples was less than 3 on log₂ scale). After these criteria were applied, ~21,000 probe sets remained for differential expression analyses.

To examine differential gene expression and to build prediction
15 models, the samples were divided into a training set of 119 OSCC cases and 35 controls and a testing set of 48 OSCC cases and 10 controls. The division of study subjects into training and testing sets was based on the calendar date that patients were enrolled into the study. Gene expression values from gcRMA were analyzed using a regression-based, estimating equations approach implemented
20 in GenePlus software (7, 8). Age and sex were included as covariates in the analyses of the training set. To control type I errors, particular group of genes were declared either "upregulated/overexpressed" or "downregulated/underexpressed" based on a fixed number of false discoveries (NFD), *i.e.*, the number of false discoveries in a list of discovered genes is
25 controlled at the pre-specified NFD (9). The choice of NFD, with an appropriate account for the number of genes under investigation (J), dictated the threshold for individual gene-specific p-values as NFD/J . Using $NFD < 1$ as a statistical testing criterion, 7,604 candidate probe sets were identified as being differentially expressed between controls and OSCC cases.

30 To build predictive models and substantially reduce the number of comparisons, this list of candidate probe sets was further narrowed using the following criteria, which retained only those probe sets that showed a significantly

large difference in signal intensity between cases and controls: 1) absolute Z-score of greater than 6 in the differential gene expression analysis, implying exceptionally high statistical significance; 2) a 1.5-fold or greater difference in gene expression between controls and cases and, 3) the mean expression value summarized by Affymetrix Microarray Suite 5.0 across samples >300 (with the scaled mean expression value of 1000). Probe sets with such expression values are more likely to be suitable for validation by alternative methodologies such as qRT-PCR. A selected number of probe sets and their corresponding biomarker genes, including variants thereof, were selected by these three criteria (see Figures 2 and 7).

The selected probe sets were analyzed using both forward and hybrid of forward-backward logistic regression procedures (SAS PROC LOGISTIC). For the one OSCC case with results from 5 replicate tissues and one control with results from duplicate tissues, the respective average of the replicate results was used. In the forward stepwise selection, probe sets were processed in the logistic regression model: one probe set at a time until no probe set could be added based on the significance level of 0.01. When the hybrid stepwise selection was adopted, the probe set with the smallest p-values and $p < 0.01$ entered first, and significance levels for other selected probe sets was evaluated for possible removal if their p-values were greater than 0.05 in the current model.

The performance of the two models (*i.e.*, results from the forward and hybrid stepwise procedures) was compared using receiver operating characteristic (ROC) curves. An ROC curve is a plot of true positive rate (sensitivity) on the Y-axis against false positive rate (1-specificity) on the X-axis for each possible value (*i.e.*, the logistic score for each individual for a given model) representing a positive test. A model with perfect discrimination between cases from controls will have a ROC curve that passes through the upper left corner, with 100% sensitivity, 100% specificity, and an area under the curve (AUC) of 1. An AUC=0.5 represents a test that is no better than chance at discriminating between cases and controls (10-12).

The selected prediction models were validated using an internal independent testing set of 48 invasive OSCC and 10 controls and an external testing set of 42 head and neck squamous cell carcinoma (HNSCC) cases and 14 controls (Gene Expression Omnibus (GEO) GSE6791,

5 www.ncbi.nlm.nih.gov/geo) (13). CEL files from these datasets were extracted using the gcRMA algorithm. ROC curves were drawn by applying the expression results to the prediction models.

Results:

The cases in both the training and testing sets tended to be older
10 than the controls. Compared to controls, cases were more likely to be male, white, and current smokers. Approximately two thirds of the cases had AJCC stage III or IV disease with about 50% of the cases presenting with metastasis to the neck. Oral cavity tumors accounted for 74% and 60% of the OSCC cases in the training and testing sets, respectively. Oropharyngeal tumors accounted for
15 26% and 40% of the OSCC cases in the training and testing sets, respectively. Most of the dysplasia subjects were white males whose lesions were located in the oral cavity.

Figures 2 and 7 list the selected probe sets and their corresponding biomarker genes that were differentially expressed between OSCC and controls
20 based on the criteria described above. Included among these probe sets are transforming growth factor (*TGFB1*), cell signaling molecule (*STAT1*), immune markers (*IL1B*), chemokines (*CXCL2*, *CXCL3*, *CXCL9*), and genes encoding for extracellular matrix proteins and collagens that have previously been shown to be involved in the motility and invasion of tumor cells.

25 Hierarchical clustering of gene expression using the selected probe sets showed that invasive OSCC and normal control formed two main clusters. About half the dysplasia tissues clustered with OSCC samples and half clustered with the controls. Compared to invasive OSCC, oral dysplasia tissue appeared to have a set of genes that were not yet upregulated and another set of genes that
30 were not yet downregulated.

Figure 3 shows the top 10 SCCIGS models from the logistic regression analyses of the selected probe sets in the training data set. The model with *LAMC2* (probe set 207517_at, encoding laminin- γ 2; SEQ ID NOS:91 and 92) and *COL4A1* (probe set 211980_at, encoding collagen type IV α 1; SEQ ID NO:113) provided the most discriminating power to separate OSCC from controls (AUC=0.99952). The power to distinguish OSCC from controls was still significant if expression of only one of these two probe sets was used (AUC=0.99424 with *COL4A1* alone).

After removing *LAMC2* and *COL4A1* from subsequent modeling, *COL1A1* (probe set 202310_s_, encoding for collagen type I α 1; SEQ ID NOS:20 and 21) and *PADI1* (probe set 220962_s_, encoding for peptidyl arginine deiminase type 1; SEQ ID NO:145) emerged as the next set of markers that best separated OSCC from controls (AUC=0.99976).

When the expression values from the testing datasets were applied to the predictive models derived from the training dataset, the model with *LAMC2* and *COL4A1* provided the most discriminating power to separate OSCC from controls: AUC=0.997 in the independent, internal testing set and AUC=0.976 in the external testing set (GEO GSE6791) (see Figure 3). The model with *COL1A1* and *PADI1* was also strongly predictive: AUC=0.99167 in the independent, internal testing set and AUC=0.97789 in the external GEO GSE6791 data set (see Figure 3). Results from the testing of the other eight models against the internal and external datasets indicate that they also performed well in distinguishing OSCC from controls (see Figure 3).

Discussion:

Described herein are selected probe sets, corresponding to a variety of known genes (see SEQ ID NOS:1-200), which are highly effective in distinguishing invasive OSCC and normal oral tissue. Also described herein is a list of genes that may be involved in the transformation of normal oral tissue to dysplasia, as well as a list of genes that may be involved in the transformation of oral dysplasia to invasive OSCC (see Example 3; and Figure 8). Although prior studies have described global changes in gene transcription that distinguish

normal oral epithelium from carcinoma, there is considerable heterogeneity among the lists of genes that have been reported, and few studies have utilized rigorous statistical testing and validation with independent datasets to produce a limited combinations of genes, as described herein, with high sensitivity and specificity in distinguishing OSCC from normal oral tissue, and which provide accurate prediction models (14). Embodiments of the present invention provide prediction models that were generated using rigorous statistical analyses, and the differences in gene expression detected using microarray technology was validated not only by qRT-PCR, but by testing against independent internal and external genome-wide gene expression datasets. The result has been to generate candidate markers and indicator gene sets that can be easily applied to the testing of biopsies or surgical margins to aid diagnosis and prognosis of OSCC.

It is believed that the prediction models and the biomarker genes identified herein may provide utility in predicting local recurrence at surgical margins or the development of second primary cancer of OSCC patients, or for selective screening of individuals who are at high risk of OSCC. If otherwise histologically-negative margins harbor microscopic original tumors as residual disease, the gene expression profiles of such margins would more likely resemble those of the resected invasive OSCC, such that the measurement of one or more of the biomarker genes or SCCIGs identified herein, and/or application of one of the predictive models described herein, could potentially be of use for the detection of residual tumor cells.

Also, for individuals who are at high risk of OSCC, their oral epithelium could contain cells that are molecularly abnormal and primed for the development of cancer. As such, the molecular profile might be more similar to that of a pre-neoplastic oral lesion than that of an invasive OSCC. The list of genes that distinguishes invasive OSCC from dysplasia and controls could be used to gauge the malignant potential of these molecular changes. Recently, p53 and eIF4E have been evaluated to augment histologic assessment of surgical margins (4, 15). eIF4E expression, but not P53 mutation and overexpression, in histologically negative surgical margins was a significant

predictor of recurrence and shorter disease-free survival of HNSCC patients (16-18).

According to certain embodiments of the present invention, the expression patterns of two pairs of genes (LAMC2 and COL4A1; COL1A1 and PADI1) may be particularly effective in distinguishing OSCC from normal oral tissue in independent testing sets. The sensitivity and specificity were close to 100%. Because of the stringent criteria applied to select candidate markers (*e.g.*, SCCIGS), there may be other probe sets among the biomarker genes and exemplary probe sets with a similar predictive property.

The results provided herein were adjusted for age and sex. Although life style characteristics, such as tobacco use and infection with human papillomavirus (HPV) play an important role in OSCC development, no appreciable difference in gene expression on the genome-wide level was observed according to smoking status (*i.e.*, former/current vs. never) or HPV status (*i.e.*, positive vs. negative).

Laminin binds to Type IV collagen and to many cell types via cell surface laminin receptors (24). Following attachment to laminin in the basement membrane, tumor cells secrete collagenase IV that specifically breaks down type IV collagen and thereby facilitates cell spreading and migration (25). In addition, laminin fragments generated by post-translational proteolytic cleavage bind to cell surface integrins and other proteins to trigger and modulate cellular motility (26). Increased levels of laminin have been associated with a number of carcinoma (27-35). In some of these studies, laminin was associated with tumor aggressiveness, metastasis and poor prognosis. Results from mouse models showed that tumor cells with high levels of laminin and low level of unoccupied laminin receptor are resistant to killing by natural cytotoxic T cells and are highly malignant (36), and that treatment with low concentrations of laminin receptor binding fragments of laminin blocked lung metastasis of hematogenously introduced tumor cells (37). A large number of unoccupied laminin receptors have been observed for breast and colon cancer cells (25); no similar reports have appeared on OSCC or HNSCC cells.

The gene products of *COL4A1* and *COL4A2* are assembled into type IV collagen that form the scaffold of basement membrane integrating other extracellular molecules, including laminin, to produce a highly organized structural barrier. Collagen IV also plays an important role in the interaction of basement membrane with cells (38, 39). Immune cells, migrating endothelial cells and metastatic tumor cells have been reported to produce and tightly regulate type IV collagen-specific collagenase (40-42).

Peptidyl arginine deiminases (EC 3.5.3.15) catalyze post-translational modification of proteins through conversion of arginine residues to citrullines. Although their physiological functions are not well understood, these deiminases have been implicated in the genesis of multiple sclerosis, rheumatoid arthritis, and psoriasis (43). The isoform peptidyl arginine deiminases type 1 (PADI1) is present in the keratinocytes of all layers of human epidermis. It has been reported that deimination of filaggrin by PADI1 is necessary for epidermal barrier function and deimination of keratin K1 may lead to ultrastructural changes of the extracellular matrix (43). The expression of PADI1 is downregulated in both dysplasia and OSCC when compared to controls. If deimination of arginine residues of proteins in the keratinocytes of oral mucosa by PADI1 forms an epidermis barrier, downregulation of PADI1 may allow the growth, expansion and movement of tumor cells.

EXAMPLE 2

VALIDATION OF SQUAMOUS CELL CARCINOMA INDICATOR GENE SETS BY qRT-PCR

The expression of *LAMC2*, *COL4A1*, *COL1A1*, and *PADI1* in OSCC compared to controls was validated by quantitative real-time PCR (qRT-PCR). Generally, qRT-PCR was performed in triplicate on a subset of 30 OSCC cases and 30 controls and bioinformatically validated.

Specifically, each sample containing 7.5 ng purified total RNA was assayed in triplicate in 10 µl reaction volumes using the QuantiTect™ SYBR Green RT-PCR kit (Qiagen, Valencia, CA) and bioinformatically validated QuantiTect primers (Qiagen, Valencia, CA) on a 7900HT Sequence Detection System (ABI, Foster City, CA). The cycling conditions were as follows: 30

minutes at 50°C, 15 minutes at 95°C, and 40 cycles of 15 seconds at 94°C, 30 seconds at 55°C, and 30 seconds at 72°C.

For *COL1A1* (NM_000088), a 118-bp amplicon spanning exons 1 and 2 was amplified. For *COL4A1* (NM_001845), a 119-bp amplicon spanning
5 exons 6, 7, 8, and 9 was amplified. For *LAMC2* (NM_005562), a 74-bp amplicon spanning exons 18 and 19 was amplified. For *PADI1* (NM_013358), an 80-bp amplicon spanning exons 3, 4, and 5 was amplified. ACTB was used as the reference gene and amplified a 146-bp amplicon that spanned exons 3 and 4.

Ten-point standard curves were generated using Universal Human
10 Reference RNA (Stratagene, La Jolla, CA) for all genes except *PADI1*, for which Normal Adjacent Esophagus Total RNA (Ambion, Austin, TX) was used. The linear correlation coefficient (R²) was 0.99 or greater for all runs. The mean threshold cycle (Ct) values were calculated from the triplicate Ct values. Mean Ct values were further normalized in relation to the mean Ct value of the ACTB
15 gene.

Results of qRT-PCR on *LAMC2*, *COL4A1*, *COL1A1* and *PADI1* confirmed the differential expression of these genes between OSCC and controls at the transcript level (see Figure 4).

EXAMPLE 3

20 COMPARISON OF GENE EXPRESSION PROFILES IN CONTROLS, DYSPLASTIC LESIONS AND INVASIVE CANCER.

While the expression of some genes may be continuously increasing or decreasing from the moment normal oral tissue begins its oncogenic process, it is also possible that some genes get turned on or off during
25 the conversion from dysplasia to invasive cancer. To test this hypothesis, and to identify genes that may be specific for the conversion of dysplasia to OSCC, the gene expression profiles of invasive cancers (n=167) were compared with the profiles of a combination of normal oral tissue (from 45 controls) and dysplastic lesions (n=17) using ~21,000 filtered probe sets. From those probe sets that
30 were differentially expressed between OSCC samples and the combination of controls and dysplastic lesions, those that were differentially expressed between

controls and dysplasia were further excluded using as the number of false discoveries (discussed above), $NFD=1$. The resulting gene list contained the genes that were up- or downregulated in OSCC but not in dysplasia.

Conversely, the dysplastic lesions and OSCC samples were
5 combined and compared with the controls. For those probe sets showing differential expression, the genes that were also differentially expressed between dysplasia and cancer were excluded. The resulting gene list contained genes that showed up- or downregulation relative to normal tissue as early as dysplasia.

Comparison of gene expressions of invasive cancer with those of
10 normal oral tissue (from controls) and dysplasia combined using ~21,000 filtered probe sets, followed by elimination of those probe sets that were differentially expressed between dysplasia and controls, showed the differential expression of 6544 probe sets, including 3988 up-regulated and 2666 down-regulated probe sets in invasive OSCC. From the original biomarker genes and corresponding
15 probe sets used to identify OSCC from controls, Figure 6 lists the biomarker genes and probe sets that may be specific for the conversion of oral dysplasia to OSCC. Further, selected probe sets that were specific for the development of dysplasia from normal tissue were also identified.

Accordingly, a set has been identified herein of genes that are
20 involved in, and specific for, the malignant transformation of oral dysplasia into invasive OSCC. Included genes encode for proteins having known roles in cell-ECM (extracellular matrix) and cell-cell interactions, and/or in cellular motility, migration and/or invasion, such as *LAMC2* and *SERPINE1* (PAI-1); for directed-cellular movement, such as *CXCL2*, *CXCL3*, and *CXCL9*, as well as for immune
25 function, such as *IL1 β* and *IFIT3*. Among the biomarker genes and probe sets described herein, a large number of collagen genes were among the sets that may be associated with the conversion of oral tissue to dysplasia (Figure 2) and were absent among the probe sets that may be involved in the conversion of dysplasia to invasive OSCC (Figure 6). These observations suggested that
30 collagen genes may play an important role early in the oral carcinogenesis process.

EXAMPLE 4

ANALYSIS OF BIOLOGICAL PATHWAYS IN OSCC

Biological pathway analyses and hierarchical clustering of differentially expressed genes were performed. To this end, the 7,604 probe sets
5 that were differentially expressed between OSCC and controls (see Example 1, *supra*) were analyzed using Ingenuity® Pathway Analysis 4.0 software (Ingenuity® Systems, www.ingenuity.com). Using Affymetrix GeneSpring™ software GX7.3.1, hierarchical clustering of all the samples was also performed based on the expression of the SCC biomarker genes and corresponding probe
10 sets.

Results obtained with the Ingenuity Pathway Analyses tool showed that the JAK/STAT signaling pathway and the IFN-γ signaling pathway were the two biological pathways that could be most strongly associated with the differentially expressed genes. Figure 1 shows genes that were up- or down-
15 regulated in these two pathways in the training dataset.

A wide array of cytokines and growth factors, including epidermal growth factor receptor (EGFR), can transmit signals through the JAK/STAT pathway (44, 45). EGFR over-expression has been reported in up to 90% of HNSCC tumors (46). Single modality therapeutics that target and negatively
20 regulate EGFR, such as small molecule tyrosine kinase inhibitors, monoclonal antibodies, antisense therapy or immunotoxin conjugates, however, were only effective in 5-15% of patients with advanced HNSCC (47). These observations suggest that there are other proteins and pathways driving the growth of some of these tumors. The results described herein are believed to be the first to show a
25 strong association between the IFN-γ signaling pathway and OSCC, noting that IFN-γ signaling also involves the JAK/STAT pathway (44, 48).

EXAMPLE 5

COMPARISON OF THE PREDICTION MODELS IN DIFFERENT CANCER TISSUES

To test the predictive power of the *LAMC2/COL4A1* and
30 *COL1A1/PADI1* indicator gene sets for other cancers, gene expression data were

downloaded from GEO GSE6791 for normal and tumor cervical tissue samples, and from GSE6044 for normal and tumor lung samples (see also Example 1). These datasets were chosen because: 1) they were generated using the same Affymetrix U133 GeneChip platform utilized herein, facilitating the testing of the tissue specificity of the predictive models described herein; and 2) OSCC share some of the same risk factors as cervical cancer and lung cancer (e.g., human papillomavirus for cervical cancer and cigarette smoking for both cervical and lung cancer). Gene expression values were extracted using gcRMA, and scores for each of the above-noted prediction models were calculated for each tissue type.

Results:

The model containing *LAMC2* and *COL4A1* distinguished head and neck squamous cell carcinoma (HNSCC) from controls, but distinguished neither cervical cancer nor lung cancer from their respective controls (Figure 5, top panel). The *COL1A1* and *PADI1* predictive model also performed well for HNSCC and, to a lesser extent, for lung cancer (Figure 5, bottom panel). Furthermore, these results showed that the models described herein could not only distinguish invasive cancer from controls, but could also distinguish oral dysplasia from controls. The respective AUC was 0.98 for *LAMC2* and *COL4A1* and 0.99477 for *COL1A1* and *PADI1*. The effect observed for the model *LAMC2* and *COL4A1* was primarily driven by *COL4A1*, suggesting that *COL4A1* up-regulation occurred earlier than *LAMC2* up-regulation in oral carcinogenesis.

Literature Cited

1. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. *CA Cancer J Clin* 2005;55:74-108.
2. Silverman SJ, Gorsky M, Lozada F. Oral leukoplakia and
5 malignant transformation. A follow-up study of 257 patients. *Cancer*
1984;53:563-568.
3. Reibel J. Prognosis of oral pre-malignant lesions: significance of
clinical, histopathological, and molecular biological characteristics. *Crit Rev Oral*
Biol Med 2003;14:47-62.
- 10 4. Upile T, Fisher C, Jerjes W et al. The uncertainty of the surgical
margin in the treatment of head and neck cancer. *Oral Oncol* 2007;43:321-6.
5. Batsakis JG. Surgical excision margins: a pathologist's
perspective. *Adv Anatomic Path* 1999;6:140-8.
6. Brandwein-Gensler M, Teixeira MS, Lewis CM et al. Oral
15 squamous cell carcinoma: histologic risk assessment, but not margin status, is
strongly predictive of local disease-free and overall survival. *Am J Surg Pathol*
2005;29:167-178.
7. Thomas JG, Olson JM, Tapscott SJ, Zhao LP. An efficient and
robust statistical modeling approach to discover differentially expressed genes
20 using genomic expression profiles. *Genome Res* 2001;11:1227-36.
8. Zhao LP, Prentice R, Breeden L. Statistical modeling of large
microarray data sets to identify stimulus-response profiles. *Proc Natl Acad Sci*
USA 2001;98:5631-6.
9. Xu XL, Olson JM, Zhao LP. A regression-based method to
25 identify differentially expressed genes in microarray time course studies and its
application in an inducible Huntington's disease transgenic model. *Hum Mol*
Genet 2002;11:1977-85.
10. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med*
1978;8:283-98.

11. Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. Principles and applications. *Ann Intern Med* 1981;94:557-92.
12. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;39:561-77.
13. Pyeon D, Newton M.A, Lambert PF et al. Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. *Cancer Res* 2007;67:4605-4619.
14. Choi P, Chen C. Genetic Expression Profiles and Biologic Pathway Alterations in Head and Neck Squamous Cell Carcinoma. *Cancer* 2005;104:1113-28.
15. Black C, Marotti J, Zarovnya E, Paydarfar J. Critical evaluation of frozen section margins in head and neck cancer resections. *Cancer* 2006;107:2792-2800.
16. Nathan CO, Franklin S, Abreo FW, Nassar R, De Benedetti A, Glass J. Analysis of surgical margins with the molecular marker eIF4E: a prognostic factor in patients with head and neck cancer. *J Clin Oncol* 1999;17:2909-2914.
17. Nathan CO, Sanders K, Abreo FW, Nassar R, Glass J. Correlation of p53 and the protooncogene eIF4E in larynx cancers: prognostic implications. *Cancer Res* 2000;60:3599-604.
18. Nathan CA, Amirghahri N, Rice C, Abreo FW, Shi R, Stucker FJ. Molecular analysis of surgical margins in head and neck squamous cell carcinoma patients, *Laryngoscope* 2002;112:2129-40.
19. Mendez E, Cheng C, Farwell DG et al. Transcriptional expression profiles of oral squamous cell carcinomas. *Cancer* 2002;95:1482-94.
20. Patel V, Aldridge K, Ensley JF et al. Laminin-gamma2 overexpression in head-and-neck squamous cell carcinoma. *Int J Cancer* 2002;99:583-8.

21. Lindberg P, Larsson A, Nielsen BS. Expression of plasminogen activator inhibitor-1, urokinase receptor and laminin gamma-2 chain is an early coordinated event in incipient oral squamous cell carcinoma. *Int J Cancer* 2006;118:2948-56.
- 5 22. Ziober AF, Kirtesh RP, Faizan A et al. Identification of a Gene Signature for Rapid Screening of Oral Squamous Cell Carcinoma. *Clin Cancer Res* 2006;12:5960-71.
23. Gonzalez HE, Gujrati M, Frederick M et al. Identification of 9 genes differentially expressed in head and neck squamous cell carcinoma. *Arch*
10 *Otolaryngol Head Neck Surg* 2003;129:754-9.
24. McCarthy JB, Basara ML, Palm SL, Sas DF, Furcht LT. The role of cell adhesion proteins- -laminin and fibronectin--in the movement of malignant and metastatic cells. *Cancer Metastasis Rev* 1985;4:125-52.
25. Liotta LA, Wewer U, Rao NC et al. Biochemical mechanisms of
15 tumor invasion and metastases. *Prog Clin Biol Res* 1988;256:3-16.
26. Hintermann E, Quaranta V. Epithelial cell motility on laminin-5: regulation by matrix assembly, proteolysis, integrins and erbB receptors. *Matrix Biol* 2004;23:75-85.
27. Yamamoto H, Itoh F, Iku S, Hosokawa M, Imai K. Expression of
20 the gamma(2) chain of laminin-5 at the invasive front is associated with recurrence and poor prognosis in human esophageal squamous cell carcinoma. *Clin Cancer Res* 2001;7:896-900.
28. Kagesato Y, Mizushima H, Koshikawa N et al. Sole expression of laminin gamma 2 chain in invading tumor cells and its association with stromal
25 fibrosis in lung adenocarcinomas, *Jpn J Cancer Res* 2001;92:184-192.
29. Barsky SH, Rao CN, Hyams D, Liotta LA. Characterization of a laminin receptor from human breast carcinoma tissue. *Breast Cancer Res Treat* 1984;4:181-8.
- 30 30. Haslam SZ, Woodward TL. Host microenvironment in breast cancer development: epithelial-cell-stromal-cell interactions and steroid hormone action in normal and cancerous mammary gland. *Breast Cancer Res* 2003;5:208-15.

31. Kaklamani VG, Gradishar WJ. Gene expression in breast cancer. *Current Treat Options Oncol* 2006;7:123-8.
32. Aishima S, Matsuura S, Terashi T et al. Aberrant expression of laminin gamma 2 chain and its prognostic significance in intrahepatic
5 cholangiocarcinoma according to growth morphology. *Mod Pathol* 2004;17:938-45.
33. Soini Y, Maatta M, Salo S, Tryggvason K, Autio-Harmainen H. Expression of the laminin gamma 2 chain in pancreatic adenocarcinoma. *J Pathol* 1996;180:290-4.
- 10 34. Olsen J, Kirkeby LT, Brorsson MM et al. Converging signals synergistically activate the LAMC2 promoter and lead to accumulation of the laminin gamma 2 chain in human colon carcinoma cells. *Biochem J* 2003;371:1-21.
35. Gontero P, Banisadr S, Freja B, Brausi M. Metastasis markers in
15 bladder cancer: a review of the literature and clinical considerations. *Eur Urol*, 2004;46:296-311.
36. Malinoff HL, McCoy JP Jr, Varani J, Wicha MS. Metastatic potential of murine fibrosarcoma cells is influenced by cell surface laminin. *Int J Cancer* 1984;33:651-655.
- 20 37. Barsky SH, Rao CN, Williams JE, Liotta LA. Laminin molecular domains which alter metastasis in a murine model. *J Clin Invest* 1984;74:843-8.
38. Kuhn K. Basement membrane (type IV) collagen. *Matrix Biol* 1995;14:439-445.
39. Van Waes C, Carey TE. Overexpression of the A9
25 antigen/alpha 6 beta 4 integrin in head and neck cancer. *Otolaryngol Clin North Am* 1992;25:1117-39.
40. Bosman FT, Havenith M, Cleutjens JP. Basement membranes in cancer. *Ultrastruct Pathol* 1985;8:291-304.
41. Schmidt C, Pollner R, Poschl E, Kuhn K. Expression of human
30 collagen type IV genes is regulated by transcriptional and post-transcriptional mechanisms. *FEBS Lett* 1992;312: 174-8.

42. Schmidt C, Fischer G, Kadner H, Genersch E, Kuhn K, Poschl E. Differential effects of DNA-binding proteins on bidirectional transcription from the common promoter region of human collagen type IV genes COL4A1 and COL4A2. *Biochim Biophys Acta* 1993;1174: 1-10.
- 5 43. Nachat R, Mechin MC, Takahara H et al. Peptidylarginine deiminase isoforms 1-3 are expressed in the epidermis and involved in the deimination of K1 and filaggrin. *J Invest Dermatol* 2005;124:384-93.
44. O'Shea JJ, Gadina M, Schreiber RD. Cytokine signaling in 2002: new surprises in the Jak/Stat pathway. *Cell* 2002;109:S121-31.
- 10 45. Rawlings JS, Rosler KM, Harrison DA. The JAK/STAT signaling pathway. *J Cell Sci* 2004;117:8-3.
46. Kalyankrishna S, Grandis JR. Epidermal growth factor receptor biology in head and neck cancer. *J Clin Oncol* 2006;24:2666-72.
47. Choong NW, Cohen EE. Epidermal growth factor receptor
15 directed therapy in head and neck cancer. *Crit Rev Oncol-Hematol* 2006;57:25-43.
48. Hebenstreit D, Horejs-Hoeck J, Duschl A. JAK/STAT-dependent gene regulation by cytokines. *Drug News Persp* 2005;18:243-249.
49. Kondoh N, Ohkura S, Arai M et al. Gene expression signatures
20 that can discriminate oral leukoplakia subtypes and squamous cell carcinoma. *Oral Oncol* 2007;43:455-462.

EXAMPLE 6

25 GENETIC EXPRESSION PROFILE ASSOCIATED WITH ORAL CANCER IDENTIFIES A GROUP OF PATIENTS AT HIGH-RISK OF POOR SURVIVAL

In this Example, OSCC was further sub-classified on the basis of 131 probe sets (108 known genes) which were previously found (Examples 1-5)
30 to be differentially expressed between OSCC and normal controls; this sub-classification was shown to be associated with survival. In particular, the Example shows that 1) there were significant survival differences in cluster

analysis-defined OSCC subgroups; 2) this classification was independently associated with overall and OSCC-specific survival after adjustment for potential confounders such as age, sex and stage; and 3) genetic expression data and AJCC stage combined predicted survival of OSCC patients better than AJCC stage alone. This Example provides the first description of an association between gene expression profiling and OSCC-specific survival in manner that is prospective and predictive of clinical outcomes.

To determine if gene expression signature of invasive oral squamous cell carcinoma (OSCC) can sub-classify OSCC on the basis of survival, hierarchical clustering was performed on the expression of 131 genes in 119 OSCC, 35 normal and 17 dysplastic mucosae to identify cluster-defined subgroups. Multivariate Cox regression ascertained association between gene expression and survival. By stepwise Cox regression the top predictive models of OSCC-specific survival were determined, and then compared by Receiver Operating Characteristics (ROC) analysis.

As described in detail below, the 3-year overall mean survival ($\pm$ SE) for a cluster of 45 OSCC patients was $38.7 \pm 0.09\%$, compared to $69.1 \pm 0.08\%$ for the remaining patients. Multivariate analysis adjusted for age, sex and stage showed that the 45 OSCC cluster patients had worse overall and OSCC-specific survival (HR=3.31, 95% CI: 1.66, 6.58; HR=5.43, 95% CI: 2.32, 12.73, respectively). Stepwise Cox regression on the 131 probe sets revealed that a model with a term for *LAMC2* (laminin, gamma 2) gene expression best identified patients with worst OSCC-specific survival. A Cox model was fit with a term for a principal component analysis-derived risk-score marker ('PCA') and two other models that combined stage with either *LAMC2* or PCA. ROC analysis Area Under the Curve for models combining stage with either *LAMC2* or PCA was 0.80 or 0.82, respectively, compared to 0.70 for stage alone ($p=0.013$ and 0.008 , respectively). This Example thus shows that OSCC can be sub-classified on the basis of gene expression (e.g., using OSCCIGSS). Gene expression and cancer stage combined predicted survival of OSCC patients better than stage alone.

As a brief background, although advances in surgical techniques and the use of adjuvant treatment modalities have led to some site-specific improvements in survival of patients with oral squamous cell carcinoma (OSCC), the overall prognosis for advanced stage disease has not improved significantly in the past two decades (1). One of the impediments to the effective management of OSCC patients is the limited ability to predict the natural history of individual lesions. Unfortunately, the current head and neck cancer staging system is inadequate for predicting survival outcomes, and there seems to be significant clinical and molecular heterogeneity within stages (2,3). However, to date, there are no molecular markers that are used clinically to stratify OSCC and other head and neck cancer patients. Recently, many studies have utilized high-throughput microarray technology in an attempt to identify the different genetic pathways involved in the carcinogenic process and to relate gene expression signatures to clinical outcomes (4-7). Gene expression profiling of OSCC would be most useful if it could add to the existing staging system to predict clinical outcomes more accurately, yet no studies to date have addressed this question.

As described above, 131 probe sets were identified (corresponding to 108 known genes) which were differentially expressed between OSCC and normal oral mucosa (8). In this Exemplar, hierarchical clustering and principal component analyses of OSCC, dysplasia and normal oral mucosa using these 131 probe sets revealed that oral dysplasias appeared to have varied expression patterns such that some clustered with OSCC and others with normal oral mucosae. According to non-limiting theory, data described in this Example indicated that there may be a spectrum of oral carcinogenesis on the basis of these 131 probe sets, and that OSCC that were least 'dysplasia-like' in gene expression were those that were further along in the carcinogenic process and, thus, were associated with poor survival rates.

MATERIALS AND METHODS

Study population: As described in Chen et al., English-speaking patients 18 year of age or older were identified with a first, primary OSCC or dysplasia undergoing surgery or biopsy between December 16th, 2003 and April

17th, 2007 at one of the three University of Washington-affiliated hospitals: University of Washington Medical Center, Harborview Medical Center and the Puget Sound Veterans Affairs Health Care System (VA). Eligible controls were patients who were scheduled to undergo surgery of the oral cavity or oropharynx for non-cancer treatment, such as tonsillectomy or sleep apnea, at the
5 aforementioned institutions during the same time period the cases were recruited. All patients recruited to the study were interviewed in person using a structured lifestyle and medical history questionnaire. Data regarding tumor characteristics, such as stage, were abstracted from medical records. Comorbidity scores were
10 calculated using Adult Comorbidity Evaluation-27 Test (9,10). Patients were followed actively through phone contact and passively through review of medical records and linkage to the U.S. Social Security Death Index. If a patient had died, the death was classified as due to OSCC or not due to OSCC based on review of medical records and death certificates. All participants gave informed
15 consent, and all study procedures were approved by the Institutional Review Boards of the Fred Hutchinson Cancer Research Center, University of Washington, and the VA.

For 150 of the 187 OSCC patients that were recruited, Affymetrix U133 2.0 Plus array data that had passed quality control criteria and at least 4
20 months of follow-up time. Array results from 17 dysplasias (11 dysplasia patients and an additional six dysplastic lesions from five OSCC patients were also included. Out of these five OSCC patients, only one invasive tumor tissue was included among the 150 OSCC cases and 35 normal oral mucosa from controls. All samples were collected, processed, and hybridized onto Affymetrix HG-U133
25 Plus 2.0 oligonucleotide arrays as described above and in Chen et al (8).

Generation of the 131-probe set list: The 131 probe set list was obtained by comparing the differential gene expression between 119 OSCC cases and 35 normal controls as described in the preceding Examples and in Chen et al (8).

30 *Hierarchical Clustering and Principal Component Analysis:*
Supervised hierarchical clustering analysis and principal component analysis (PCA) of the expression data from 119 OSCCs and 35 controls used to generate

the 131 probe set list in from the preceding Examples (e.g., Fig. 2A), plus an additional 17 dysplasias, were performed using GeneSpring GX Software v7.3.1 (Silicone Genetics, CA).

Differential Gene Expression Among OSCC: To identify sub-groups of the 119 OSCC cases based on differential gene expression values from GC Robust Multi-array Average (GCRMA), a regression-based approach was implemented in GenePlus software (11). For this comparison, the number of false discoveries (NFD) was used as the type I error selection criterion (12). Gene ontology and pathway analysis for the resultant list of genes were performed using Ingenuity Pathway Analysis software version 6.

Survival Analysis: Follow-up time for analyses of survival for the 119 OSCC cases was calculated from the date of surgery to the date of death, loss-to-follow-up, or April 30, 2007, whichever came first, according to the Kaplan-Meier method. Differences between groups were assessed with the log-rank test. OSCC-specific Kaplan Meier survival estimates were not computed because of possible informative censoring due to death from other causes. Rather, OSCC-specific cumulative mortality was estimated using methods described by Kalbfleisch and Prentice, which account for competing risk events (13,14). Cox-proportional hazards regression model was used to estimate associations with cluster-defined OSCC sub-group status adjusted for age, sex, stage and co-morbidity score. Dummy variables were created for cluster-defined OSCC sub-group status, sex, stage, and co-morbidity scores. These statistical analyses were conducted using STATA software version 9.2.

Prediction model building for OSCC-specific mortality: For this analysis, a total of 150 OSCC cases were used: 119 cases which had been used to derive the 131 probe sets in the preceding Examples (e.g., Fig. 2A) (8), plus an additional 31 cases that were recruited thereafter for which vital status information was obtained and at least 4 months follow-up. Stepwise Cox-proportional hazards regression was used based on the 131 probe sets previously found (preceding Examples) to be differentially expressed between OSCC cases and controls (SAS version 9.2) (8). For the stepwise regression, the significance level for both entrance and exit were each set at $\alpha=0.01$. To

obtain the top 10 models, ten sequential stepwise regression procedures were conducted, with each successive procedure eliminating the selected probe set(s) from the previous procedure. Individual risk scores from the top probe set Cox regression model were compared graphically to risk scores from Cox models with terms for the first and second principal components (PC) from PCA of the 131 probe sets, using Matlab version R2006b.

Comparing survival prediction models with TNM stage: To assess whether a survival model which incorporates gene expression data is better than one without it, an adapted Receiver Operating Characteristics (ROC) analysis was used (15). Risk scores were calculated for 5 models. The first three models contained the terms: 'stage'; 'gene(s) from top prediction model'; and 'PCA' (principal component analysis) – a score representing the expression of the entire 131 probe sets as summarized by the combination of the first and second Principal Components (PCs). The other two models combined the term 'stage' with either of the other two terms. For each model, ROC curves were constructed for predicting two year all-cause survival. At each level of the model-derived risk score, the nearest 10% (using Nearest Neighbor Estimation) was used to estimate true positive and false positive rates. The survival ROC package (<http://faculty.washington.edu/heagerty/Software/SurvROC>), available for R-project software, was used to implement these methods. The Area Under the Curve (AUC) was calculated to quantify the ability of each model to predict two year survival. One thousand bootstrap samples were generated to estimate standard errors and 95% confidence intervals for AUC estimates, and to obtain *p*-values for testing the null hypothesis that specific gene expression values or PCA do not add to ability of stage to predict survival.

In order to reduce the over-optimism of ROC and AUC estimates due to using the same data both to estimate and assess the predictive ability of risk scores, a jackknife leave-one-out analysis (16) was performed. Parameter estimates for the risk model were obtained excluding one subject, and the resulting risk model was used to estimate a risk score based on the excluded subject's gene expression and/or stage characteristics. This process was repeated until risk scores were assigned to each subject. ROC and AUC

estimates were calculated for these jackknife risk scores as they were for the original risk scores.

Validation of LAMC2, OSMR, SERPINE1, OASL, by qRT-PCR:

qRT-PCR was performed to validate the expression of the four genes found to be
 5 related to survival in the top two models. Sixty samples were chosen at random
 for testing. Each sample was assayed in triplicate in 10 µl reaction volumes
 using the QuantiTect SYBR Green RT-PCR kit (Qiagen, Valencia, CA) and
 bioinformatically validated QuantiTect primers (Qiagen, Valencia, CA) on a
 7900HT Sequence Detection System (ABI, Foster City, CA). The cycling
 10 conditions were as follows: 30 minute incubation at 50° C, 15 minute incubation
 at 95° C, and 40 cycles each of 15 seconds at 94° C, 30 seconds at 55° C, and
 30 seconds at 72° C. The fragment amplified included: 1) For *LAMC2*
 (NM_005562) a 74-bp amplicon spanning exons 18 and 19 ; 2) For *OASL*
 (NM_003733), a 98-bp amplicon spanning exons 4 and 5; 3) for *OSMR*
 15 (NM_003999) a 113-bp amplicon spanning exons 13 and 14 ;4) For *SERPINE1*
 (NM_000602), a 105-bp amplicon spanning exons 3 and 4; and as the reference
 gene, ACTB, a 146-bp amplicon spanning exons 3 and 4. Ten point standard
 curves were generated using Universal Human Reference RNA (Stratagene, La
 Jolla, CA) for all genes. The linear correlation coefficient (R^2) was 0.99 or greater
 20 for all runs. The mean threshold cycles (Ct) values were calculated from the
 triplicate Ct values. Samples that had Ct values with standard deviation greater
 than 0.35 in their triplicate run were repeated. Mean Ct values were standardized
 to the mean Ct value of ACTB.

25 RESULTS:

Study population: The characteristics of the study participants are
 shown in Table 1. In general, the OSCC cases tended to be older and male, and
 they were more likely to be current smokers when compared to controls. The
 majority of the OSCC cases had advanced stage disease (approximately two
 30 thirds with AJCC stage III and IV).

Table 1. Clinical Characteristics of Study Participants

	Cases		Controls		Dysplasias	
	n=(150)		n=(35)		n=(16)	
	n	%	n	%	n	%
Age (years)						
19-39	6	4	14	40	0	0
40-49	23	15.3	8	22.9	2	12.5
50-59	51	34	4	11.4	5	31.3
60-88	70	46.7	9	25.7	9	56.3
Gender						
Male	109	72.7	25	71.4	10	62.5
Female	41	27.3	10	28.6	6	37.5
Race						
White	133	88.7	24	68.6	15	93.8
Nonwhite	11	7.3	11	31.4	1	6.3
Unknown	6	4	0	0	0	0
Smoking						
Never/former	77	51.3	25	71.4	8	50
Current	73	48.7	10	28.6	8	50
Alcohol use						
Never/former	52	34.7	9	25.7	9	56.3
Current	96	64	26	74.3	7	43.8
Unknown	2	1.3	0	0	0	0
AJCC Stage						
I	38	25.3	0	0	N/A	N/A
II	16	10.7	0	0	N/A	N/A
III	18	12	0	0	N/A	N/A
IV	78	52	0	0	N/A	N/A
Site						
Oral	103	68.7	1	2.9	14	87.5
Oropharynx	47	31.3	34	97.1	2	12.5
Unknown	0	0	0	0	0	0
Tumor size						
T1/T2	101	67.3	0	0	N/A	N/A
T3/T4	49	32.7	0	0	N/A	N/A
Nodal status						

N0	71	47.3	N/A	0	N/A	N/A
N1	79	52.7	N/A	0	N/A	N/A

* One patient contributed to two separate samples.

Hierarchical Cluster and Principal Component Analysis: Results from a supervised hierarchical cluster analysis of the 119 OSCC cases, 35 normal controls and 17 dysplastic lesions using the 131 probe sets are shown in Figure 10. Although OSCC cases largely clustered separately from controls, 7 OSCC cases clustered with the controls. One cluster of genes (cluster 1, Figure 1) appeared to show an increasing gradient of down-regulation progressing from normal to dysplastic to invasive lesions. Notably, neither the dysplasias nor those OSCC that misclassified with the normal controls demonstrated consistent down-regulation of these genes. In particular, all members of this group of 12 probe sets (corresponding to nine genes) were completely down-regulated in a subset of 45 OSCC (cluster 1, Figure 1 and Table S2). It therefore appeared according to non-limiting theory that the gene expression signature of this group of 45 OSCC represented one end of a continuum of gene expression that was characteristic of increasingly aggressive neoplastic behavior.

Figure 11 shows the results of a principal component analysis (PCA) on the 131 probe set expression data based on the samples' phenotype (normal, dysplasia or cancer). The first principal component (PC), which accounted for the greatest amount of variability, captured 60.26% of the variance, whereas the second PC captured 6.31%. On the basis of these two components alone, the controls and OSCC cases were at opposite ends of the spectrum with dysplasia samples in between (Fig. 11). In addition, the same group of 45 OSCC samples identified in the hierarchical cluster analysis was at one extreme on the basis of the first PC scores (Fig. 11). Although some dysplastic lesions had first PC scores that overlapped with OSCC, none reached the first PC scores of the group of 45 OSCC samples.

Differential Expression of the 45 sample sub-cluster: This cluster-defined OSCC subgroup was initially identified largely based on a qualitative analysis of the expression of a group of 12 down-regulated probe sets (Fig. 10,

cluster 1). A linear regression model was used to more rigorously determine which probe sets were differentially expressed in this sub-cluster, compared to the rest of the OSCC cases. After adjusting for age and sex, 62 out of the 131 probe sets were detectably shown to be differentially expressed between these two groups (NFD = 1) (Table 2). Therefore, although the 12 down-regulated probe sets each represented a readily apparent change in expression in these 45 samples, nearly one-half of the 131 probe sets showed a distinctive signature in this sub-cluster.

10 **Table 2.** Differentially expressed probe sets in group of 45 OSCC sub-cluster out of the 131 probe set list (NFD=1).

Up-regulated in group of 45 sub-cluster		
Gene Name	Gene Symbol	Description
241872_at	DKFZp761D221	SH3-domain GRB2-like (endophilin) interacting protein 1
229947_at	PI15	peptidase inhibitor 15
229225_at	NRP2	neuropilin 2
226997_at		CDNA FLJ10196 fis, clone HEMBA1004776
225292_at	COL27A1	collagen, type XXVII, alpha 1
222344_at	C5orf13	Chromosome 5 open reading frame 13
217519_at	MACF1	Glycine-rich protein (GRP3S)
213869_x_at	THY1	Thy-1 cell surface antigen
212489_at	COL5A1	collagen, type V, alpha 1
212488_at	COL5A1	collagen, type V, alpha 1
212473_s_at	MICAL2	microtubule associated monooxygenase, calponin and LIM domain containing 2
212472_at	MICAL2	microtubule associated monooxygenase, calponin and LIM domain containing 2
212354_at	SULF1	sulfatase 1
212344_at	SULF1	sulfatase 1
207850_at	CXCL3	chemokine (C-X-C motif) ligand 3
205574_x_at	BMP1	bone morphogenetic protein 1
204879_at	PDPN	podoplanin
204715_at	PANX1	pannexin 1
204647_at	HOMER3	homer homolog 3 (Drosophila)
204619_s_at	CSPG2	chondroitin sulfate proteoglycan 2 (versican)

204114_at	NID2	nidogen 2 (osteonidogen)
204051_s_at	SFRP4	secreted frizzled-related protein 4
203325_s_at	COL5A1	collagen, type V, alpha 1
202267_at	LAMC2	laminin, gamma 2
202238_s_at	NNMT	nicotinamide N-methyltransferase
201108_s_at	THBS1	thrombospondin 1
1568765_at	SERPINE1	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1
1555420_a_at	KLF7	Kruppel-like factor 7 (ubiquitous)
1554008_at	OSMR	oncostatin M receptor
Down-regulated in group of 45 sub-cluster		
Gene Name	Gene Symbol	Description
243718_at		Transcribed locus, weakly similar to NP_060190.1 signal-transducing adaptor protein-2; brk kinase substrate [Homo sapiens]
242417_at	LOC283278	hypothetical protein LOC283278
241233_x_at	C21orf81	chromosome 21 open reading frame 81
231118_at	ANKRD35	ankyrin repeat domain 35
230740_at		Transcribed locus
230104_s_at	TPPP	brain-specific protein p25 alpha
227782_at		BTB/POZ-zinc finger protein-like
225548_at	SHRM	shroom
223541_at	HAS3	hyaluronan synthase 3
221665_s_at	EPS8L1	EPS8-like 1
220149_at	FLJ22671	hypothetical protein FLJ22671
218885_s_at	GALNT12	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 12 (GalNAc-T12)
218779_x_at	EPS8L1	EPS8-like 1
213421_x_at	PRSS3	protease, serine, 3 (mesotrypsin)
212224_at	ALDH1A1	aldehyde dehydrogenase 1 family, member A1
211980_at	COL4A1	collagen, type IV, alpha 1

209975_at	CYP2E1	cytochrome P450, family 2, subfamily E, polypeptide 1
209800_at	KRT16	keratin 16 (focal non-epidermolytic palmoplantar keratoderma)
207206_s_at	ALOX12	arachidonate 12-lipoxygenase
207126_x_at	UGT1A10 ; UGT1A8 ; UGT1A7 ; UGT1A6 ; UGT1A5 ; UGT1A9 ; UGT1A4 ; UGT1A1 ; UGT1A3	UDP glucuronosyltransferase 1 family, polypeptide A10 ; UDP glucuronosyltransferase 1 family, polypeptide A8 ; UDP glucuronosyltransferase 1 family, polypeptide A7 ; UDP glucuronosyltransferase 1 family, polypeptide A6 ; UDP glucuronosyltransferase 1 family, polypeptide A5 ; UDP glucuronosyltransferase 1 family, polypeptide A9 ; UDP glucuronosyltransferase 1 family, polypeptide A4 ; UDP glucuronosyltransferase 1 family, polypeptide A1 ; UDP glucuronosyltransferase 1 family, polypeptide A3
206094_x_at	UGT1A6	UDP glucuronosyltransferase 1 family, polypeptide A6
206093_x_at	TNXB	tenascin XB
205319_at	PSCA	prostate stem cell antigen
205200_at	CLEC3B	C-type lectin domain family 3, member B
205157_s_at	KRT17	keratin 17
204754_at	HLF	Hepatic leukemia factor

204532_x_at	UGT1A10 ; UGT1A8 ; UGT1A7 ; UGT1A6 ; UGT1A5 ; UGT1A9 ; UGT1A4 ; UGT1A1 ; UGT1A3	UDP glucuronosyltransferase 1 family, polypeptide A10 ; UDP glucuronosyltransferase 1 family, polypeptide A8 ; UDP glucuronosyltransferase 1 family, polypeptide A7 ; UDP glucuronosyltransferase 1 family, polypeptide A6 ; UDP glucuronosyltransferase 1 family, polypeptide A5 ; UDP glucuronosyltransferase 1 family, polypeptide A9 ; UDP glucuronosyltransferase 1 family, polypeptide A4 ; UDP glucuronosyltransferase 1 family, polypeptide A1 ; UDP glucuronosyltransferase 1 family, polypeptide A3
203921_at	CHST2	carbohydrate (N-acetylglucosamine-6-O) sulfotransferase 2
202404_s_at	COL1A2	collagen, type I, alpha 2
202310_s_at	COL1A1	collagen, type I, alpha 1
201497_x_at	MYH11	myosin, heavy polypeptide 11, smooth muscle
1553212_at	K5B	keratin 5b
1552283_s_at	ZDHHC11	zinc finger, DHHC-type containing 11

11 probe sets from "cluster 1" in Figure 1 are in bold. One probe set, TCP11L2 t-complex 11 (mouse) like 2 (1553861_at), out of the 12 probe sets in "cluster 1" was not down-regulated with statistical significance.

5 *Survival Analysis:* The patient characteristics for this sub-cluster compared to those of the rest of the cases are shown in Table 3. Patients in the 45-sample sub-cluster had more advanced disease, as determined by both tumor size and nodal metastasis. The range of follow-up time for patients known to be alive at the end of the study was 10.7 to 38.7 months, with a median of 22.2

10 months.

Table 3. Characteristics of two cluster-defined OSCC subgroups

Group of 74		Group of 45	
(n=74)		(n=45)	
n	(%)	n	(%)

Age (years)				
20-39	1	-1.4	4	-8.9
40-49	10	-13.5	8	-17.8
50-59	28	-37.8	12	-26.7
60-88	35	-47.3	21	-46.7
Gender				
Male	55	-74.3	29	-64.4
Female	19	-25.7	16	-35.6
AJCC Stage				
I	21	-28.4	7	-15.6
II	11	-14.9	3	-6.7
III	9	-12.2	7	-15.6
IV	33	-44.6	28	-62.2
Tumor Size				
T1/T2	56	-75.7	22	-48.9
T3/T4	18	-24.3	23	-51.1
Nodal Status				
N0	40	-54.1	15	-33.3
N1	34	-45.9	30	-66.7
Vital Status				
Alive	59	-79.7	21	-46.7
Dead-OSCC	9	-12.2	17	-37.8
Dead-non OSCC	6	-8.1	3	-6.7
Dead -unknown cause	0	0	4	-8.9

To test the hypothesis that this 45 sample sub-cluster had a more aggressive phenotype, Kaplan-Meier survival curves for overall survival were compared (Figure 12A). The 3-year mean ($\% \pm \text{SE}$) overall survival for the 45 sample sub-cluster was $38.7 \pm 0.09\%$ compared to $69.1 \pm 0.08\%$ ($p=0.0001$) for the other 74 samples. The estimated cumulative mortality ($\% \pm \text{SE}$) due to OSCC at 3 years for the 45 sample sub-cluster was $45.7 \pm 0.09\%$ compared to $16.8 \pm 0.06\%$ ($p=0.0003$) for the other 74 samples (Figure 3B). Hazard ratios (HR) were estimated for both overall and OSCC-specific mortality, adjusting for AJCC stage, age and sex. Patients from the 45 sample sub-cluster had a significantly higher rate of death, both overall (HR=3.31, 95% CI: 1.66, 6.58) and due to OSCC (HR=5.43, 95% CI: 2.32, 12.73). HR were then adjusted for tumor size, nodal status and tumor site (oral vs. oropharyngeal) separately and the results did not change

appreciably. Although a higher co-morbidity score was statistically significantly associated with both mortality outcomes, it neither confounded nor improved the precision of the association with OSCC sub-cluster when included in the Cox regression model.

5 *Prediction models:* A stepwise Cox-proportional hazards regression was performed based on these 131 probe sets to determine which, if any, were associated with OSCC-specific survival. Out of 150 patients, 109 were alive and 41 had died at the end of the follow-up period (Table 1). Among these, there were 27 OSCC-specific deaths and 14 deaths of either unknown or non-OSCC-specific causes. A model containing *LAMC2* (laminin, gamma 2) was preferred for identifying patients with the worst OSCC-specific survival. The subsequent nine models that were identified through the stepwise approach are shown in Table 4.

15

Table 4. 10 preferred multivariate Cox regression models of OSCC-specific survival

Model	Gene Symbol (Affymetrix Probe set ID)	Model coefficients
1	<i>LAMC2</i> (207517_at)	0.59151*LAMC2
2	<i>OSMR</i> (1554008_at) <i>SERPINE1</i> (1568765_at) <i>OASL</i> (210797_at)	0.42485*OSMR + 0.40482*SERPINE1 + 0.33483*OASL
3	<i>SLC16A1</i> (209900_s_at)	0.81478*SLC16A1
4	<i>KLF7</i> (1555420_at)	0.60694*KLF7
5	<i>THBS1</i> (201108_s_at) <i>SLC16A1</i> (202235_at)	0.44241*THBS1 + 0.43257*SLC16A1
6	<i>HOMER3</i> (204647_at)	0.66632*HOMER3
7	<i>GRP68</i> (229055_at)	0.63313*GRP68
8	<i>PDPN</i> (204879_at)	0.51904*PDPN
9	<i>ANKRD35</i> (231118_at)	0.58503*ANKRD35
10	<i>CDH3</i> (203256_at) <i>EPS8L1</i> (218779_x_at)	0.75146* CDH3 – 0.50956* EPS8L1

The 3-D plot (Figure 14) shows that the risk score from the preferred model ($0.59151 \times LAMC2$) were highly correlated with the risk scores from models containing terms for the first and second PCs from the analysis of the 131 probe sets. In addition, those patients with the highest risk scores from either the preferred model or the PC models were mostly the ones in the cluster-defined group of 45 patients.

Comparing survival prediction models with AJCC stage: Results from the ROC analysis for each of the 5 models described above are shown in Figure 13. The AUCs for models with either gene expression alone or in combination with stage were higher than for a model with stage alone (Fig. 13C). The differences in the AUCs between models with 'stage' plus either 'LAMC2' or 'PCA' and stage alone were statistically significant ($p = 0.013$ and 0.008 , respectively). The AUCs from the jackknife leave-one-out analyses (0.81 for the model with 'stage' and 'LAMC2' and 0.79 for the model with 'stage' and 'PCA') were virtually the same as those estimated using conventional methods.

Validation of LAMC2, OSMR, SERPINE1, OASL, by qRT-PCR:

The correlation coefficients for *LAMC2*, *OSMR*, *SERPINE1* and *OASL* between microarray and qRT-PCR expression data for the 60 samples assayed were 0.65 , 0.14 , 0.74 , and 0.89 , respectively. Thus, with the exception of *OSMR*, qRT-PCR results were well-correlated with those from microarray analyses.

Discussion:

In this Example, from expression levels of 131 probe sets that were previously found to be highly associated with OSCC (8), it has been shown that OSCC can be further sub-classified on the basis of gene expression signatures. Moreover, this classification was independently associated with overall and OSCC-specific survival after adjustment for potential confounders such as age, sex and stage. None of the dysplastic lesions overlapped with the group of 45 OSCC cases on the basis of the first PC (Fig. 11) suggesting that this 45 sample sub-cluster represented a more invasive phenotype. From these findings it was investigated whether there was a trend of differential expression of these 131 probe sets in OSCC, such that the varying degrees of up- or down-regulation of

some genes might be of prognostic significance. The observation that the score ('PCA') that summarized the expression levels of all 131 probe sets as a combination of the first and second PCs was significantly associated with OSCC-specific survival supports this hypothesis.

5 Another important finding of this Example is that similar results were obtained between summary measures of all 131 probe sets and the preferred model containing only one gene. The risk scores from models with each of the first two principal components and the preferred, highest-scoring gene-specific model ($0.59151 \cdot LAMC2$) were highly correlated (Fig. 14). This result
10 underscored the possibility for reducing the dimensionality of the data not only to the summary of principal components, but to one single probe set, without substantial loss of information. This efficiency was important because it will be easier to implement molecular tests for clinical use if fewer molecules need to be measured.

15 Two previous studies by Chung et al have shown an association between microarray-derived expression data and clinical outcomes in head and neck squamous cell carcinomas (HNSCC) (4,5). In the first study, a 582 gene set from 60 HNSCC classified the tumors into 4 different subclasses with statistically significant differences in recurrence-free survival. In a separate study using
20 formalin-fixed tissues, these authors identified a second 950 gene signature from unsupervised analysis and a 75-gene list from supervised PCA that was predictive of recurrence. Using the Unigene ID numbers for each gene, we determined that these two previous signatures have 39 genes in common.

 In contrast, comparison of the presently disclosed 131 probe sets
25 with the gene lists from the previous two studies revealed only one gene shared in all three lists: solute carrier family 16, member 1 (*SLC16A1*). This gene comprises the third-highest scoring model for OSCC-specific survival according to the data presently available (Table 4).

 Pramana et al tested 42 genes with known function out of the 75-
30 gene list from Chung et al , and showed that these genes were predictive of locoregional control in their own data set (5,6). Only two genes overlapped

between the presently disclosed 131 probe set list and these 42 genes: Glycine-rich protein (GRP3S) (*MACF1*) and collagen type V, alpha 1 (*COL5A1*).

There are likely to be many reasons for the lack of more substantial overlap between the presently disclosed gene lists and those of Chung et al. and Pramana et al. For example, their 950- and 75-gene lists were derived from formalin-fixed samples and a different array platform (5). In addition, the samples in the studies of Chung et al were from multiple head and neck sites, whereas the samples used to generate the data in the present Example were limited to the oral cavity and oropharynx. The end points also differed between studies, since overall and OSCC-specific survival were analyzed in the present Example, while Chung et al. examined recurrence-free survival (5,8). The statistical approaches to derive these gene signatures were also substantially different (5,8,11). Given all these issues and without wishing to be bound by theory, it appears that overlapping genes in particular should be further investigated for their potential generalizability in predicting clinical outcomes.

Among the 131 genes used in the supervised cluster analysis described here, 62 probe sets were differentially expressed between the group of 45 patients and the remaining OSCC. Ingenuity Pathway Analysis of the 62 probe sets showed an overrepresentation of genes involved in cell migration; cell-to-cell signaling and interaction; and cellular growth and proliferation. In addition, five of the genes in the 10 highest-scoring models that are presented here as predictive of OSCC-specific mortality, such as *LAMC2*, *SERPINE1*, *THBS1*, *PDPN* and *CDH3*, play a role in cell motility and cell-to-cell signaling, implying that expression of genes involved in the process of invasion and metastasis is an important determinant of outcome in patients with these malignancies (17-23). Specifically, the proteins encoded by these genes reside in the extracellular matrix and are believed to function, for example, in angiogenesis, platelet aggregation and/or cell movement.

For instance, *THBS1* and *PDPN* have both been ascribed a role in platelet aggregation and may be involved in tumor metastasis by facilitating tumor cell-platelet interactions and platelet-facilitated tumor cell metastasis (20). It is also known that *THBS1* binds with members of the tenacin family and

SPARC/osteonectin (20). In fact, tenascin C and SPARC were found to be significantly upregulated at both the gene expression and protein levels in this and other studies by our group (2,24). In addition, P-cadherin (*CDH3*), a component of the 10th model described here, is associated with cell-to-cell
5 signaling, and the *CDH3* gene has previously been shown to be significantly downregulated in metastatic tumors cell isolated from lymph nodes (25). The findings in this Example showing that the dysregulation of these genes' expression was associated with OSCC-specific survival were consistent with the non-limiting theory that tumor proliferation and metastasis may be mediated by
10 complex interactions between extracellular matrix proteins and cell-surface receptors.

The functions of other genes in the 10 top-scoring models described in this Example are less well understood. *OASL* appears to be a member of a family of Trips (Thyroid hormone-interacting proteins) and may thus
15 be involved in signal transduction in the presence of thyroid hormone (26). Oncostatin M receptor (*OSMR*) is a member of the IL6 cytokine family and is thought to be involved in signal transduction and proliferation (27,28).

This Example is believed to provide the first demonstration of an association between a gene signature and OSCC-specific survival, and in
20 particular of the use of gene expression data to improve upon AJCC stage in predicting survival. As described above, regression models that combined stage with gene expression had significantly higher AUCs than stage alone (Figure 13). Given the recent emphasis on genome-wide gene expression studies to find signatures predictive of clinical outcomes, these and related embodiments will
25 permit integration of meaningful genetic data into clinical practice.

References:

1. Carvalho AL, Nishimoto IN, Califano JA, Kowalski LP. Trends in incidence and prognosis for head and neck cancer in the United States:
30 A site-specific analysis of the SEER database. Int J Cancer 2005;114:806-16.
2. Mendez E, Cheng C, Farwell DG et al. Transcriptional expression profiles of oral squamous cell carcinomas. Cancer 2002;95:1482-94.

3. Baatenburg de Jong RJ, Hermans J, Molenaar J, Briaire JJ, le Cessie S. Prediction of survival in patients with head and neck cancer. *Head Neck* 2001;23:718-24.
4. Chung CH, Parker JS, Karaca G et al. Molecular
5 classification of head and neck squamous cell carcinomas using patterns of gene expression. *Cancer Cell* 2004;5:489-500.
5. Chung CH, Parker JS, Ely K et al. Gene Expression Profiles Identify Epithelial-to-Mesenchymal Transition and Activation of Nuclear Factor- κ B Signaling as Characteristics of a High-risk Head and Neck
10 Squamous Cell Carcinoma. *Cancer Res* 2006;66:8210-8.
6. Pramana J, van den Brekel MW, Van Velthuisen ML et al. Gene expression profiling to predict outcome after chemoradiation in head and neck cancer. *Int J Radiat Oncol Biol Phys* 2007;69:1544-52.
7. Ginos MA, Page GP, Michalowicz BS et al.
15 Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck. *Cancer Res* 2004;64:55-63.
8. Chen C, Mendez E, Houck J et al. Gene expression profiling identifies genes predictive of oral squamous cell carcinoma. *Cancer Epidemiology, Biomarkers and Prevention*, In Press 2008
9. Piccirillo JF. Importance of comorbidity in head and
20 neck cancer. *Laryngoscope* 2000;110:593-602.
10. Piccirillo JF, Creech C, Zequeira R, Anderson S, Johnston AS. Inclusion of Comorbidity into Oncology Data Registries. *Journal of Registry Management* 1999;26:66-70.
11. Thomas JG, Olson JM, Tapscott SJ, Zhao LP. An
25 efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res* 2001;11:1227-36.
12. Xu XL, Olson JM, Zhao LP. A regression-based
30 method to identify differentially expressed genes in microarray time course studies and its application in an inducible Huntington's disease transgenic model. *Hum Mol Genet* 2002;11:1977-85.

13. Kalbfleisch, J. D. and Prentice, R. L. The Statistical Analysis of Failure Time Data. New York: John Wiley and Sons, 1980.
14. Satagopan JM, Ben-Porat L, Berwick M et al. A note on competing risks in survival data analysis. Br J Cancer 2004;91:1229-35.
- 5 15. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2000;56:337-44.
16. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. N Engl J Med
10 1985;313:793-9.
17. Schenk S, Hintermann E, Bilban M et al. Binding to EGF receptor of a laminin-5 EGF-like fragment liberated during MMP-dependent mammary gland involution. J Cell Biol 2003;161:197-209.
18. Bajou K, Masson V, Gerard RD et al. The
15 plasminogen activator inhibitor PAI-1 controls in vivo tumor vascularization by interaction with proteases, not vitronectin. Implications for antiangiogenic strategies. J Cell Biol 2001;152:777-84.
19. Pedersen H, Brunner N, Francis D et al. Prognostic impact of urokinase, urokinase receptor, and type 1 plasminogen activator
20 inhibitor in squamous and large cell lung cancer tissue. Cancer Res 1994;54:4671-5.
20. Bornstein P. Diversity of function is inherent in matricellular proteins: an appraisal of thrombospondin 1. J Cell Biol 1995;130:503-6.
- 25 21. Yee KO, Streit M, Hawighorst T, Detmar M, Lawler J. Expression of the type-1 repeats of thrombospondin-1 inhibits tumor growth through activation of transforming growth factor-beta. Am J Pathol 2004;165:541-52.
22. Kato Y, Sasagawa I, Kaneko M et al. Aggrus: a
30 diagnostic marker that distinguishes seminoma from embryonal carcinoma in testicular germ cell tumors. Oncogene 2004;23:8552-6.

23. Sandler MA, Zhang JN, Westerhausen DRJr, Billadello JJ. A novel protein interacts with the major transforming growth factor-beta responsive element in the plasminogen activator inhibitor type-1 gene. *J Biol Chem* 1994;269:21500-4.
- 5 24. Choi P, Jordan CD, Mendez E et al. Examination of oral cancer biomarkers by tissue microarray analysis. *Arch Otolaryngol Head Neck Surg* 2008;134:539-46.
25. Mendez E, Fan W, Choi P et al. Tumor-specific genetic expression profiles of metastatic oral squamous cell carcinoma. *Head Neck* 2007;29:803-14.
- 10 26. Lee JW, Choi HS, Gyuris J, Brent R, Moore DD. Two classes of proteins dependent on either the presence or absence of thyroid hormone for interaction with the thyroid hormone receptor. *Mol Endocrinol* 1995;9:243-54.
- 15 27. Mosley B, De Imus C, Friend D et al. Dual oncostatin M (OSM) receptors. Cloning and characterization of an alternative signaling subunit conferring OSM-specific receptor activation. *J Biol Chem* 1996;271:32635-43.
- 20 28. Gearing DP, Comeau MR, Friend DJ et al. The IL-6 signal transducer, gp130: an oncostatin M receptor and affinity converter for the LIF receptor. *Science* 1992;255:1434-7.

CLAIMS

What is claimed is:

1. A method for identifying a risk for having, or presence of, oral squamous cell carcinoma (OSCC) in a subject, the method comprising:

(a) determining a squamous cell carcinoma indicator gene set (SCCIGS) expression level in a biological sample from the subject that comprises at least one OSCC cell or at least one cell from an OSCC surgical margin; and

(b) comparing the SCCIGS expression level of (a) to a reference SCCIGS expression level that is characteristic of epithelial cells from a control tissue that comprises normal oral epithelium known to be free of squamous cell carcinoma cells; wherein differential expression of the SCCIGS in the biological sample relative to the control tissue indicates the subject has, or is at risk for having, OSCC.

2. The method of claim 1 wherein the SCCIGS comprises one or more SCC biomarker genes selected from SEQ ID NOS:1-200, or variants thereof that are differentially expressed in a squamous cell carcinoma cell as compared to a control tissue that comprises normal oral epithelium known to be free of squamous carcinoma cells.

3. The method of claim 1 or claim 2 wherein the SCCIGS is one or more SCCIGS selected from the group consisting of:

- (a) the SCCIGS consisting of a LAMC2 gene,
- (b) the SCCIGS consisting of LAMC2 and COL4A1 genes, and
- (c) the SCCIGS consisting of COL1A1 and PADI1 genes.

4. The method of claim 1 or claim 2 wherein the SCCIGS is one or more SCCIGS selected from the group consisting of:

- (a) the SCCIGS consisting of a LAMC2 gene,
- (b) the SCCIGS consisting of LAMC2 and COL4A1 genes,
- (c) the SCCIGS consisting of COL1A1 and PADI1 genes,

- (d) the SCCIGS consisting of a C21orf81 gene,
- (e) the SCCIGS consisting of KRT17 and PRSS3 genes,
- (f) the SCCIGS consisting of COL1A2 and EST 230740_1 at genes,
- (g) the SCCIGS consisting of COL1A1 and XLKD1 genes,
- (h) the SCCIGS consisting of THY1, FLJ22671 and HAS3 genes,
- (i) the SCCIGS consisting of POSTN and TIA2(PDPN) genes,
- (j) the SCCIGS consisting of MGC40368(TCP11L2), GIP3(IFI6) and COL27A1 genes,
- (k) the SCCIGS consisting of CDH3 and ELOVL6 genes,
- (l) the SCCIGS consisting of the COL4A1 gene, and
- (m) the SCCIGS consisting of genes identified by 131 probe sets as set forth in Figure 2A.

5. The method of any one of claims 1-4 wherein the step of determining a SCCIGS expression level comprises:

(a) specifically hybridizing a detectable, squamous cell carcinoma indicator gene (SCCIG)-specific oligonucleotide probe to one or more of:

- (i) all or a SCCIG-characteristic portion of a SCCIG transcript,
 - (ii) a polynucleotide having a nucleotide sequence that is fully complementary to (i), and
 - (iii) a nucleic acid amplification product of one or more of (i) and (ii); and
- (b) detecting the SCCIG-specific probe, and thereby determining the SCCIGS expression level.

6. The method of claim 5 wherein the SCCIG-specific oligonucleotide probe has a nucleotide sequence that is selected from the group consisting of SEQ ID NOS:201-331.

7. The method of claim 5 wherein the biological sample comprises a biopsy tissue.
8. The method of claim 7 wherein the biopsy tissue is selected from the group consisting of an excised tumor, a tumor-positive margin tissue, a tumor-negative margin tissue and a close margin tissue.
9. The method of claim 5 wherein the biological sample comprises one or a plurality of dysplastic cells.
10. A method for identifying a risk for having, or presence of, oral squamous cell carcinoma (OSCC) in a subject having oral epithelial dysplasia but no frank OSCC, the method comprising:
 - (a) determining a squamous cell carcinoma indicator gene set (SCCIGS) expression level in a biological sample from the subject that comprises at least one dysplastic oral epithelial cell; and
 - (b) comparing the SCCIGS expression level of (a) to a reference SCCIGS expression level that is characteristic of OSCC cells; wherein substantial similarity of the SCCIGS expression level in the biological sample relative to the OSCC reference SCCIGS expression levels indicates the subject has, or is at risk for having, OSCC.
11. The method of claim 10 wherein the squamous cell carcinoma indicator gene set (SCCIGS) comprises any one or more of the genes shown in Figure 6.
12. The method of claim 10 wherein the step of determining a SCCIGS expression level comprises:
 - (a) specifically hybridizing a detectable, squamous cell carcinoma indicator gene (SCCIG)-specific oligonucleotide probe to one or more of:

- (i) all or a SCCIG-characteristic portion of a SCCIG transcript,
 - (ii) a polynucleotide having a nucleotide sequence that is fully complementary to (i), and
 - (iii) a nucleic acid amplification product of one or more of (i) and (ii); and
- (b) detecting the SCCIG-specific probe, and thereby determining the SCCIGS expression level.

13. The method of claim 12 wherein the SCCIG-specific oligonucleotide probe has a nucleotide sequence that is selected from the group consisting of the probes listed in Figure 9.

14. The method of claim 10 wherein the biological sample comprises a biopsy tissue.

15. The method of claim 10 wherein the subject has no detectable cancer and the biological sample comprises one or a plurality of dysplastic cells.

16. A method for identifying a risk for having, or presence of, a squamous cell carcinoma (SCC) in a subject, wherein the SCC is selected from oral SCC (OSCC) and head-and-neck SCC (HNSCC), the method comprising:

- (a) determining a squamous cell carcinoma indicator gene set (SCCIGS) expression level in a biological sample from the subject that comprises at least one OSCC cell or at least one cell from an OSCC surgical margin; and
- (b) comparing the SCCIGS expression level of (a) to a reference SCCIGS expression level that is characteristic of epithelial cells from a control tissue that comprises normal oral epithelium known to be free of squamous cell carcinoma cells, wherein: if the biological sample comprises an OSCC cell then the control tissue comprises normal oral epithelium, and if the first biological sample comprises a HNSCC cell then the control tissue comprises normal

epithelium from oropharynx, hypopharynx, larynx or oral cavity; and wherein differential expression of the SCCIGS in the biological sample relative to the control tissue indicates the subject has, or is at risk for having, OSCC.

17. The method of claim 16, wherein the SCCIGS comprises one or more SCC biomarker genes selected from SEQ ID NOS:1-200, or variants thereof that are differentially expressed in a squamous cell carcinoma cell as compared to a control tissue that comprises normal oral epithelium known to be free of squamous carcinoma cells.

18. The method of claim 15 or claim 16 wherein the SCCIGS is one or more SCCIGS selected from the group consisting of:

- (a) the SCCIGS consisting of a LAMC2 gene,
- (b) the SCCIGS consisting of LAMC2 and COL4A1 genes, and
- (c) the SCCIGS consisting of COL1A1 and PADI1 genes.

19. The method of claim 15 or claim 16 wherein the SCCIGS is one or more SCCIGS selected from the group consisting of:

- (a) the SCCIGS consisting of a LAMC2 gene,
- (b) the SCCIGS consisting of LAMC2 and COL4A1 genes,
- (c) the SCCIGS consisting of COL1A1 and PADI1 genes,
- (d) the SCCIGS consisting of a C21orf81 gene,
- (e) the SCCIGS consisting of KRT17 and PRSS3 genes,
- (f) the SCCIGS consisting of COL1A2 and EST 230740_1at genes,
- (g) the SCCIGS consisting of COL1A1 and XLKD1 genes,
- (h) the SCCIGS consisting of THY1, FLJ22671 and HAS3 genes,
- (i) the SCCIGS consisting of POSTN and TIA2(PDPN) genes,
- (j) the SCCIGS consisting of MGC40368(TCP11L2), GIP3(IFI6) and COL27A1,
- (k) the SCCIGS consisting of CDH3 and ELOVL6 genes,

- (l) the SCCIGS consisting of the COL4A1 gene and
- (m) the SCCIGS consisting of genes identified by 131 probe sets as set forth in Figure 2A.

20. The method of any one of claims 16-19 wherein the biological sample comprises an OSCC cell and the control tissue comprises normal oral epithelium.

21. The method of any one of claims 16-19 wherein the biological sample comprises a HNSCC cell and the control tissue comprises normal epithelium from oropharynx, hypopharynx, larynx or oral cavity.

22. The method of any one of claims 16-19 wherein the step of determining a SCCIGS expression level comprises:

(a) specifically hybridizing a detectable, squamous cell carcinoma indicator gene (SCCIG)-specific oligonucleotide probe to one or more of:

- (i) all or a SCCIG-characteristic portion of a SCCIG transcript,
- (ii) a polynucleotide having a nucleotide sequence that is fully complementary to (i), and
- (iii) a nucleic acid amplification product of one or more of (i) and (ii); and

(b) detecting the SCCIG-specific probe, and thereby determining the SCCIGS expression level.

23. The method of claim 22 wherein the SCCIG-specific oligonucleotide probe has a nucleotide sequence that is selected from the group consisting of SEQ ID NOS:201-331.

24. The method of claim 22 wherein the biological sample comprises a biopsy tissue.

25. The method of claim 24 wherein the biopsy tissue is selected from the group consisting of an excised tumor, a tumor-positive margin tissue, a tumor-negative margin tissue and a close margin tissue.

26. The method of claim 22 wherein the biological sample comprises one or a plurality of dysplastic cells.

27 A method for identifying an increased risk of oral squamous cell carcinoma (OSCC)-specific mortality in a subject having OSCC, the method comprising:

(a) determining a squamous cell carcinoma indicator gene set (SCCIGS) expression level in a biological sample from the subject that comprises at least one OSCC cell or at least one cell from an OSCC surgical margin;

(b) determining that the subject has, or is at risk for having, OSCC by comparing the SCCIGS expression level of (a) to a reference SCCIGS expression level that is characteristic of epithelial cells from a control tissue that comprises normal oral epithelium known to be free of squamous cell carcinoma cells, wherein a differentially expressed SCCIGS in the biological sample relative to the control tissue indicates the subject has, or is at risk for having, OSCC; and

(c) identifying within said differentially expressed SCCIGS a presence or absence of a substantially up- or down-regulated SCCIGS subset (SCCIGSS), wherein presence of the substantially up- or down-regulated SCCIGSS indicates the subject has an increased risk of OSCC-specific mortality.

28. The method of claim 27 wherein the SCCIGS comprises one or more SCC biomarker genes selected from SEQ ID NOS:1-200, or variants thereof that are differentially expressed in a squamous cell carcinoma cell as compared to a control tissue that comprises normal oral epithelium known to be free of squamous carcinoma cells.

29. The method of claim 27 or claim 28 wherein the SCCIGSS is one or more SCCIGSS selected from the group consisting of:

- (a) the SCCIGS consisting of a LAMC2 gene,
- (b) the SCCIGSS consisting of a LAMC2 gene,
- (c) the SCCIGSS consisting of OSMR, SERPINE1 and OASL genes,
- (d) the SCCIGSS consisting of a SLC16A1 gene,
- (e) the SCCIGSS consisting of a KLF7 gene,
- (f) the SCCIGSS consisting of THBS1 and SLC16A1 genes,
- (g) the SCCIGSS consisting of a HOMER3 gene,
- (h) the SCCIGSS consisting of a GRP68 gene,
- (i) the SCCIGSS consisting of a PDPN gene,
- (j) the SCCIGSS consisting of an ANKRD35 gene, and
- (k) the SCCIGSS consisting of CDH3 and EPS8L1 genes.

30. The method of claim 27 or claim 28 wherein (i) the SCCIGS is one or more SCCIGS selected from the group consisting of:

- (a) the SCCIGS consisting of a LAMC2 gene,
- (b) the SCCIGS consisting of LAMC2 and COL4A1 genes,
- (c) the SCCIGS consisting of COL1A1 and PADI1 genes,
- (d) the SCCIGS consisting of a C21orf81 gene,
- (e) the SCCIGS consisting of KRT17 and PRSS3 genes,
- (f) the SCCIGS consisting of COL1A2 and EST 230740_1at genes,
- (g) the SCCIGS consisting of COL1A1 and XLKD1 genes,
- (h) the SCCIGS consisting of THY1, FLJ22671 and HAS3 genes,
- (i) the SCCIGS consisting of POSTN and TIA2(PDPN) genes,
- (j) the SCCIGS consisting of MGC40368(TCP11L2), GIP3(IFI6) and COL27A1 genes,
- (k) the SCCIGS consisting of CDH3 and ELOVL6 genes,

(l) the SCCIGS consisting of the COL4A1 gene, and
(m) the SCCIGS consisting of genes identified by 131 probe sets as set forth in Figure 2A,
and wherein (ii) the SCCIGSS consists of one or more genes identified by a probe set as set forth in Table 2.

31. The method of any one of claims 27-30 wherein at least one of the steps selected from the step of determining a SCCIGS expression level and the step of identifying a SCCIGSS comprises:

(a) specifically hybridizing a detectable, squamous cell carcinoma indicator gene (SCCIG)-specific oligonucleotide probe to one or more of:

- (i) all or a SCCIG-characteristic portion of a SCCIG transcript,
 - (ii) a polynucleotide having a nucleotide sequence that is fully complementary to (i), and
 - (iii) a nucleic acid amplification product of one or more of (i) and (ii); and
- (b) detecting the SCCIG-specific probe, and thereby determining the SCCIGS expression level.

32. The method of claim 31 wherein the SCCIG-specific oligonucleotide probe has a nucleotide sequence that is selected from the group consisting of SEQ ID NOS:201-331.

33. The method of claim 31 wherein the biological sample comprises a biopsy tissue.

34. The method of claim 33 wherein the biopsy tissue is selected from the group consisting of an excised tumor, a tumor-positive margin tissue, a tumor-negative margin tissue and a close margin tissue.

35. The method of claim 31 wherein the biological sample comprises one or a plurality of dysplastic cells.

36. The method of any one of claims 1-4, 10, 16, 17, 27 and 28 wherein determining one or a plurality of SCCIGS expression levels comprises measuring one or more protein levels in the biological sample.

37. The method of claim 36 wherein the biological sample comprises a biological fluid.

38. The method of claim 37 wherein the biological fluid is selected from the group consisting of saliva, blood, serum, plasma and lymph.

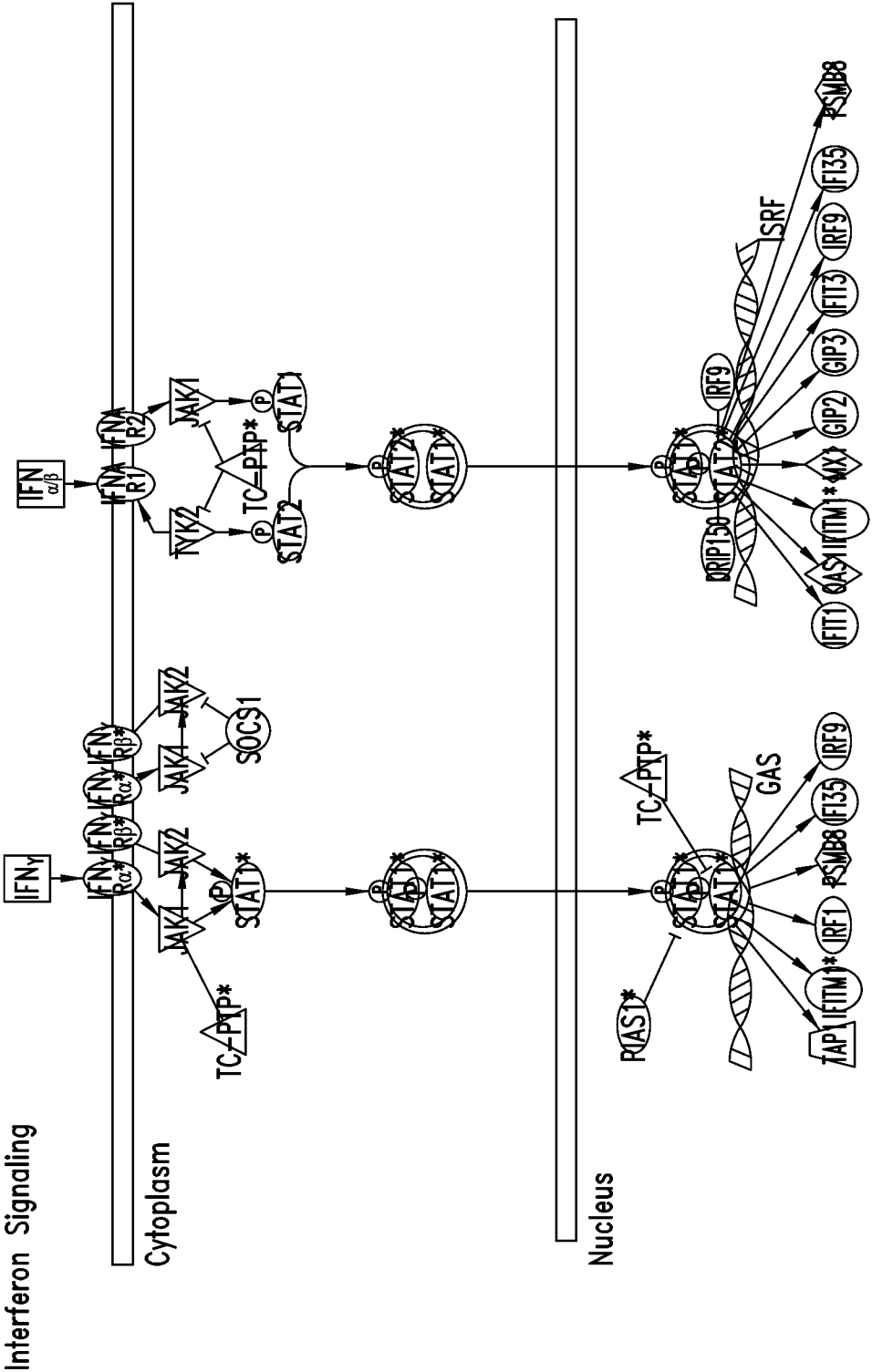


FIG. 1

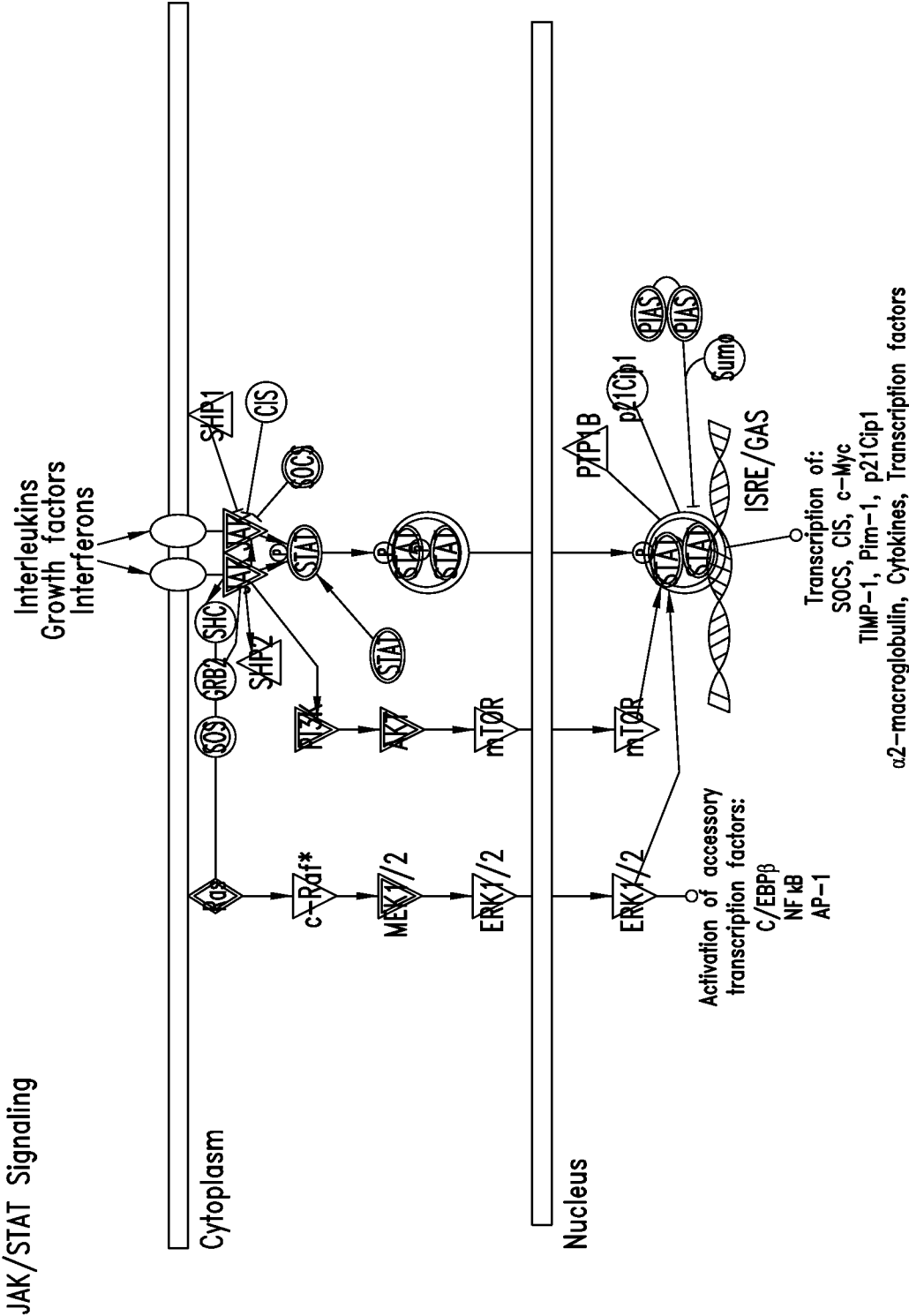


FIG. 1 (Continued)

3/34

Up-regulation in OSCC				Down-regulation in OSCC				
Probe set	Gene	Z score	Probe set	Gene	Z score	Probe set	Gene	Z score
202311_s_at	COL1A1	22.3	212365_at	MYO1B	12.0	1553212_at	KRT78	-18.6
202404_s_at	COL1A2	20.6	212012_at	PXDN	11.9	220149_at	FLJ22671	-18.1
202310_s_at	COL1A1	20.3	229860_x_at	LOC401115	11.8	241233_x_at	C21orf81	-16.4
211980_at	COL4A1	19.5	219211_at	USP18	11.8	1569608_x_at	LOC643187	-15.2
202267_at	LAMC2	18.8	219863_at	HERC5	11.8	220962_s_at	PADI1	-14.5
204415_at	IFI6	18.6	204619_s_at	CSPG2	11.7	210868_s_at	ELOVL6	-14.0
225681_at	CTHRC1	18.0	203968_s_at	CDC6	11.5	205319_at	PSCA	-13.7
212488_at	COL5A1	17.0	208156_x_at	EPPK1	11.5	204754_at	HLF	-13.7
211924_s_at	PLAUR	16.8	210797_s_at	OASL	11.4	218779_x_at	EPS8L1	-13.5
203256_at	CDH3	16.3	1568765_at	SERPINE1	11.4	221665_s_at	EPS8L1	-13.2
221729_at	COL5A2	16.2	204972_at	OAS2	11.2	220016_at	AHNAK	-12.8
213869_x_at	THY1	15.8	223541_at	HAS3	11.2	218885_s_at	GALNT12	-12.7
217312_s_at	COL7A1	15.7	218888_s_at	NETO2	11.1	231118_at	ANKRD35	-12.7
1555778_a_at	POSTN	15.6	209949_at	NCF2	11.1	225548_at	SHROOM3	-12.1

3/34

FIG. 2A

212489_at	COL5A1	15.3	204779_s_at	HOXB7	11.0	206094_x_at	UGT1A6	-11.9
221730_at	COL5A2	15.2	41037_at	TEAD4	11.0	206093_x_at	TNXB	-11.9
212354_at	SULF1	15.1	209800_at	KRT16	10.9	218935_at	EHD3	-11.9
207517_at	LAMC2	15.0	217519_at	MACF1	10.8	207126_x_at	UGT1A4	-11.8
212344_at	SULF1	15.0	202238_s_at	NNMT	10.7	230740_at	—	-11.7
204715_at	PANX1	14.7	221898_at	PDPN	10.7	204532_x_at	UGT1A4	-11.6
208851_s_at	THY1	14.2	201108_s_at	THBS1	10.7	242417_at	LOC283278	-11.5
222693_at	FNDCC3B	13.9	209969_s_at	STAT1	10.4	213421_x_at	PRSS3	-10.6
204647_at	HOMER3	13.9	203921_at	CHST2	10.2	205200_at	CLEC3B	-10.5
213668_s_at	SOX4	13.7	204103_at	CCL4	10.2	1552283_s_at	ZDHHHC11	-10.4
205574_x_at	BMP1	13.3	241872_at	SGIP1	10.1	220037_s_at	XLKD1	-10.1
1555420_a_at	KLF7	13.3	207850_at	CXCL3	10.0	1553861_at	TCP11L2	-10.0
217430_x_at	COL1A1	13.3	204747_at	IFIT3	9.7	204378_at	BCAS1	-9.7
210809_s_at	POSTN	13.2	219725_at	TREM2	9.6	242009_at	—	-9.7
205157_s_at	KRT17	12.9	203915_at	CXCL9	9.6	207206_s_at	ALOX12	-9.6

4/34

FIG. 2A (Continued)

5/34

203695_s_at	DFNA5	12.9	204879_at	PDPN	9.6	205730_s_at	ABLIM3
203325_s_at	COL5A1	12.9	1554008_at	OSMR	9.3	238715_at	ARHGAP27
209900_s_at	SLC16A1	12.8	204051_s_at	SFRP4	9.2	205428_s_at	CALB2
203085_s_at	TGFB1	12.7	227697_at	SOCS3	9.1	1565661_x_at	FUT6
229225_at	NRP2	12.7	210001_s_at	SOCS1	9.0	208609_s_at	TNXA /TNXB
225288_at	—	12.5	235276_at	—	8.5	227782_at	ZBTB7C
202235_at	SLC16A1	12.5	222344_at	C5orf13	8.4	226303_at	PGM5
204114_at	NID2	12.5	225520_at	MTHFD1L	8.4	240000_at	—
229554_at	—	12.4	218404_at	SNX10	8.3	201497_x_at	MYH11
214453_s_at	IFI44	12.4	229055_at	GPR68	8.1	227419_x_at	PLAC9
212472_at	MICAL2	12.3	209774_x_at	CXCL2	8.0	230104_s_at	TPPP
205483_s_at	ISG15	12.2	39402_at	IL1B	7.2	212224_at	ALDH1A1
226997_at	—	12.2	219300_s_at	CNTNAP2	6.9	243718_at	—
212473_s_at	MICAL2	12.0	229947_at	PI15	6.7	209975_at	CYP2E1
225292_at	COL27A1	12.0	238581_at	GBP5	6.3		

FIG. 2B

6/34

Validation of predictive models using internal and external (GSE6791) testing data sets.

SSCGIS model with gene name and Affymetrix probe set ID	Model from logistic regression	AUC	
		Internal control	GSE6791 control
SSCIGS Model 1			
LAMC2, 207517_at			
COL4A1, 211980_at	7.8739*LAMC2+7.6269*COL4A1	0.99792	0.97619
SSCIGS Model 2			
COL1A1, 202310_s_at			
PADI1, 220962_s_at	2.4377*COL1A1-2.8841*PADI1	0.99167	0.97789
SSCIGS Model 3			
C21orf81, 241233_x_	-2.1042*C21orf81	0.98540	0.97450
SSCIGS Model 4			
KRT17, 205157_s_at			
PRSS3, 213421_x_at	2.5638*KRT17-2.4506*PRSS3	0.97710	0.97450
SSCIGS Model 5			
COL1A2, 202404			
EST, 230740_at	1.9345*COL1A2-1.5931*230740_at	0.98960	0.95920
SSCIGS Model 6			
COL1A1, 202311_s_at			
XLKD1, 220037_s_at	2.2372*COL1A1-1.3377*XLKD1	0.99170	0.95070

FIG. 3A

7/34

SSCIGS Model 7

THY1,
208851_s_at

FLJ22671,
220149_at

HAS3, 223541_at $2.4643 \cdot \text{THY1} - 1.6340 \cdot \text{FLJ22671} + 1.5310 \cdot \text{HAS3}$ 0.99790 0.96260

SSCIGS Model 8

POSTN,
1555778_a_at

TIA2, 221898_at $1.4909 \cdot \text{POSTN} + 1.8340 \cdot \text{TIA2}$ 0.98960 0.90820

SSCIGS Model 9

MGC40368,
1553861_at

GIP3, 204415_at

COL27A1, — 0.97290 0.95410
225288_at $2.2659 \cdot \text{MGC40368} + 1.0718 \cdot \text{GIP3} + 1.7854 \cdot \text{COL27A1}$

SSCIGS Model 10

CDH3, 203256_at

ELOVL6, $1.9861 \cdot \text{CDH3} - 2.1743 \cdot \text{ELOVL6}$ 0.99380 0.89800
210868_s_at

FIG. 3B

8/34

qRT-PCR results comparing RNA transcripts for four genes between OSCC cases and controls

	Mean (SD) Ct [*]	95% CI [‡]	<i>P</i>
<i>LAMC2</i>			
Case	2.83 (1.02)	2.44-3.21	≤ 0.0001
Control	7.38 (0.54)	7.18-7.59	
<i>COL4A1</i>			
Case	5.13 (0.86)	4.81-5.45	≤ 0.0001
Control	8.58 (0.78)	8.29-8.87	
<i>COL1A1</i>			
Case	2.28 (1.14)	1.85-2.71	≤ 0.0001
Control	6.94 (0.64)	6.70-7.18	
<i>PADI1</i>			
Case	10.86 (2.34)	9.99-11.73	≤ 0.0001
Control	5.06 (0.94)	4.71-5.41	

* Ct (threshold cycle) values are inversely associated with the amount of RNA transcripts in the sample. Based on analyses of 30 OSCC cases and 30 controls.

‡CI, confidence interval.

FIG. 4

Tissue specificity of model LAMC2 and COL4A1 (top) and model COLaA1 and PADI1 (bottom)

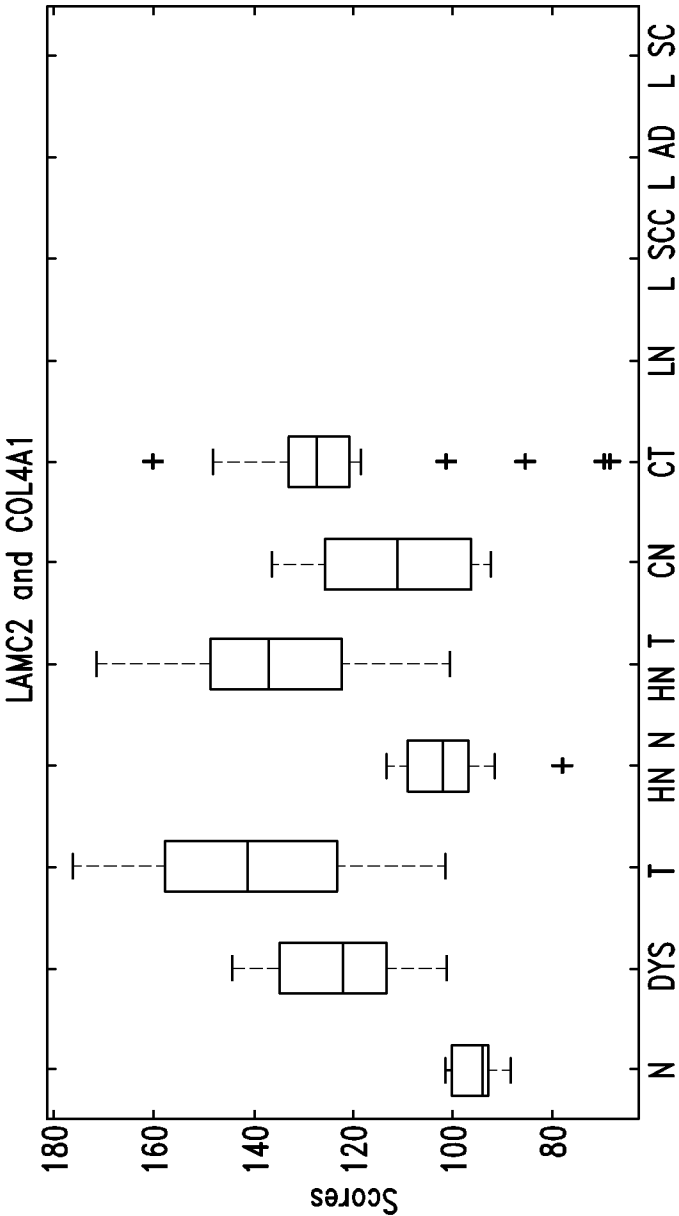


FIG. 5

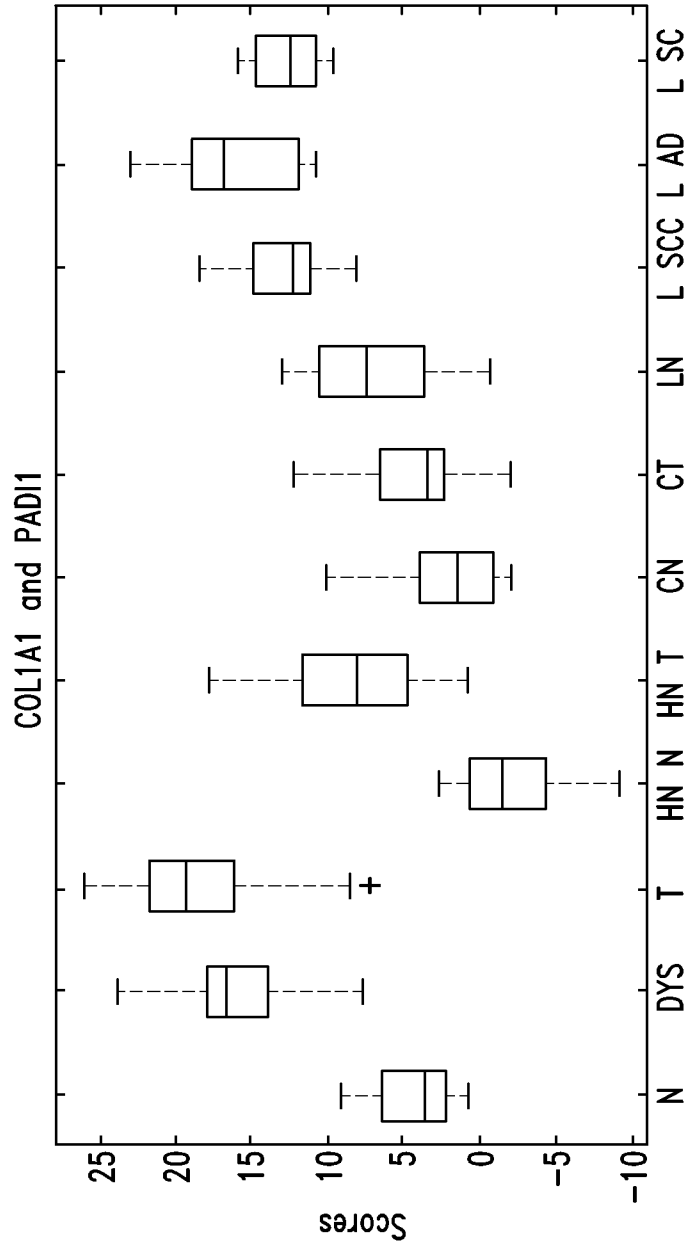


FIG. 5 (Continued)

11/34

Differentially expressed genes between OSCC and dysplasia/normal controls.

Up-regulation in OSCC			Down-regulation in OSCC		
Probe set	Gene	Z score	Probe set	Gene	Z score
202267_at	<i>LAMC2</i>	17.6	205200_at	<i>TNA</i>	-14.7
207517_at	<i>LAMC2</i>	17.5	231118_at	<i>FLJ25124</i>	-14.0
1568765_at	<i>SERPINE1</i>	14.9	206093_x_at	<i>TNXB</i>	-13.1
1555420_a_at	<i>KLF7</i>	13.8	227782_at	<i>ZBTB7C</i>	-11.3
222693_at	<i>FAD104</i>	13.6	1552283_s_at	<i>ZDHHC11</i>	-11.2
207850_at	<i>CXCL3</i>	12.8	213421_x_at	<i>PRSS3</i>	-10.8
210001_s_at	<i>SOCS1</i>	11.9	238715_at	<i>ARHGAP27</i>	-10.6
229225_at	<i>NRP2</i>	11.8	208609_s_at	<i>TNXB</i>	-10.4
227697_at	<i>SOCS3</i>	11.6	242009_at		-10.3
209949_at	<i>NCF2</i>	11.5	226303_at	<i>PGM5</i>	-9.9
204103_at	<i>CCL4</i>	11.5	207206_s_at	<i>ALOX12</i>	-9.7
218404_at	<i>SNX10</i>	11.4	230104_s_at	<i>TPPP</i>	-9.6
203695_s_at	<i>DFNA5</i>	11.4	220037_s_at	<i>XLKD1</i>	-8.6
212354_at	<i>SULF1</i>	11.2	243718_at		-8.6
229860_x_at	<i>LOC4115</i>	10.9	227419_x_at	<i>PLAC9</i>	-8.5
209774_x_at	<i>CXCL2</i>	10.8	201497_x_at	<i>MYH11</i>	-8.0
241872_at	<i>SGIP1</i>	10.7	204532_x_at	<i>UGT1A10</i>	-7.8
203968_s_at	<i>CDC6</i>	10.6	206094_x_at	<i>UGT1A1,1A4,1A6</i>	-7.6
225520_at	<i>FTHFSDC1</i>	10.1	207126_x_at	<i>UGT1A10</i>	-7.5
229947_at	<i>PI15</i>	9.3	212224_at	<i>ALDH1A1</i>	-6.4
204747_at	<i>IFIT3</i>	8.9			
39402_at	<i>IL1B</i>	8.9			
235276_at	<i>EPSTI1</i>	8.8			
203915_at	<i>CXCL9</i>	8.7			
204779_s_at	<i>HOXB7</i>	8.6			
219211_at	<i>USP18</i>	8.5			
238581_at	<i>GBP5</i>	7.6			
219300_s_at	<i>CNTNAP2</i>	5.0			
204051_s_at	<i>SFRP4</i>	4.2			

FIG. 6

12/34

SEQ ID NO: 1	1552283_s_at NM_024786 gi 13376149 ref NM_024786.1	Homo sapiens zinc finger, DHHC-type containing 11 (ZDHHC11), mRNA
SEQ ID NO: 2	1553212_at NM_173352 gi 89357931 ref NM_173352.2	Homo sapiens keratin 78 (KRT78), mRNA
SEQ ID NO: 3	1553861_at NM_152772 gi 22749516 ref NM_152772.1	Homo sapiens t-complex 11 (mouse)-like 2 (TCP11L2), mRNA
SEQ ID NO: 4	1554008_at NM_003999 gi 4557039 ref NM_003999.1	Homo sapiens oncostatin M receptor (OSMR), mRNA
SEQ ID NO: 5	1555420_a_at NM_003709 gi 142345029 ref NM_003709.2	Homo sapiens Kruppel-like factor 7 (ubiquitous) (KLF7), mRNA
SEQ ID NO: 6	1555778_a_at NM_006475 gi 5453833 ref NM_006475.1	Homo sapiens periostin, osteoblast specific factor (POSTN), mRNA
SEQ ID NO: 7	1565661_x_at NM_000150 gi 103472034 ref NM_000150.2	Homo sapiens fucosyltransferase 6 (alpha (1,3) fucosyltransferase) (FUT6), transcript variant 1, mRNA
SEQ ID NO: 8	1565661_x_at NM_001040701 gi 103472032 ref NM_001040701.1	Homo sapiens fucosyltransferase 6 (alpha (1,3) fucosyltransferase) (FUT6), transcript variant 2, mRNA
SEQ ID NO: 9	1568765_at NM_000602 gi 169790801 ref NM_000602.2	Homo sapiens serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1 (SERPINE1), mRNA
SEQ ID NO: 10	1569608_x_at XM_001716758 gi 169215259 ref XM_001716758.1	PREDICTED: Homo sapiens hypothetical LOC643187 (LOC643187), mRNA
SEQ ID NO: 11	201108_s_at NM_003246 gi 40317625 ref NM_003246.2	Homo sapiens thrombospondin 1 (THBS1), mRNA
SEQ ID NO: 12	201497_x_at NM_001040113 gi 92091582 ref NM_001040113.1	Homo sapiens myosin, heavy chain 11, smooth muscle (MYH11), transcript variant SM2B, mRNA
SEQ ID NO: 13	201497_x_at NM_001040114 gi 92091585 ref NM_001040114.1	Homo sapiens myosin, heavy chain 11, smooth muscle (MYH11), transcript variant SM1B, mRNA
SEQ ID NO: 14	201497_x_at NM_002474 gi 92091587 ref NM_002474.2	Homo sapiens myosin, heavy chain 11, smooth muscle (MYH11), transcript variant SM1A, mRNA
SEQ ID NO: 15	201497_x_at NM_022844 gi 92091584 ref NM_022844.2	Homo sapiens myosin, heavy chain 11, smooth muscle (MYH11), transcript variant SM2A, mRNA
SEQ ID NO: 16	202235_at NM_003051 gi 115583684 ref NM_003051.3	Homo sapiens solute carrier family 16, member 1 (monocarboxylic acid transporter 1) (SLC16A1), mRNA
SEQ ID NO: 17	202238_s_at NM_006169 gi 62953139 ref NM_006169.2	Homo sapiens nicotinamide N-methyltransferase (NNMT), mRNA
SEQ ID NO: 18	202267_at NM_005562 gi 157419137 ref NM_005562.2	Homo sapiens laminin, gamma 2 (LAMC2), transcript variant 1, mRNA
SEQ ID NO: 19	202267_at NM_018891 gi 157419139 ref NM_018891.2	Homo sapiens laminin, gamma 2 (LAMC2), transcript variant 2, mRNA

FIG. 7A

SEQ ID NO: 20	202310_s_at NM_000088 gi 110349771 ref NM_000088.3	Homo sapiens collagen, type I, alpha 1 (COL1A1), mRNA
SEQ ID NO: 21	202311_s_at NM_000088 gi 110349771 ref NM_000088.3	Homo sapiens collagen, type I, alpha 1 (COL1A1), mRNA
SEQ ID NO: 22	202404_s_at NM_000089 gi 48762933 ref NM_000089.3	Homo sapiens collagen, type I, alpha 2 (COL1A2), mRNA
SEQ ID NO: 23	203085_s_at NM_000660 gi 63025221 ref NM_000660.3	Homo sapiens transforming growth factor, beta 1 (TGFB1), mRNA
SEQ ID NO: 24	203256_at NM_001793 gi 189083715 ref NM_001793.4	Homo sapiens cadherin 3, type 1, P-cadherin (placental) (CDH3), mRNA
SEQ ID NO: 25	203325_s_at NM_000093 gi 89276750 ref NM_000093.3	Homo sapiens collagen, type V, alpha 1 (COL5A1), mRNA
SEQ ID NO: 26	203695_s_at NM_004403 gi 116734721 ref NM_004403.2	Homo sapiens deafness, autosomal dominant 5 (DFNA5), transcript variant 1, mRNA
SEQ ID NO: 27	203915_at NM_002416 gi 4505186 ref NM_002416.1	Homo sapiens chemokine (C-X-C motif) ligand 9 (CXCL9), mRNA
SEQ ID NO: 28	203921_at NM_004267 gi 71067350 ref NM_004267.3	Homo sapiens carbohydrate (N-acetylglucosamine-6-O) sulfotransferase 2 (CHST2), mRNA
SEQ ID NO: 29	203968_s_at NM_001254 gi 51944959 ref NM_001254.3	Homo sapiens cell division cycle 6 homolog (S. cerevisiae) (CDC6), mRNA
SEQ ID NO: 30	204051_s_at NM_003014 gi 170784837 ref NM_003014.3	Homo sapiens secreted frizzled-related protein 4 (SFRP4), mRNA
SEQ ID NO: 31	204103_at NM_002984 gi 90704850 ref NM_002984.2	Homo sapiens chemokine (C-C motif) ligand 4 (CCL4), transcript variant 1, mRNA
SEQ ID NO: 32	204114_at NM_007361 gi 144953894 ref NM_007361.3	Homo sapiens nidogen 2 (osteonidogen) (NID2), mRNA
SEQ ID NO: 33	204378_at NM_003657 gi 191251776 ref NM_003657.2	Homo sapiens breast carcinoma amplified sequence 1 (BCAS1), mRNA
SEQ ID NO: 34	204415_at NM_002038 gi 94538326 ref NM_002038.3	Homo sapiens interferon, alpha-inducible protein 6 (IFI6), transcript variant 1, mRNA
SEQ ID NO: 35	204415_at NM_022872 gi 94538329 ref NM_022872.2	Homo sapiens interferon, alpha-inducible protein 6 (IFI6), transcript variant 2, mRNA
SEQ ID NO: 36	204415_at NM_022873 gi 94538330 ref NM_022873.2	Homo sapiens interferon, alpha-inducible protein 6 (IFI6), transcript variant 3, mRNA
SEQ ID NO: 37	204532_x_at NM_000463 gi 45827762 ref NM_000463.2	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A1 (UGT1A1), mRNA
SEQ ID NO: 38	204532_x_at NM_001072 gi 132626744 ref NM_001072.3	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A6 (UGT1A6), transcript variant 1, mRNA
SEQ ID NO: 39	204532_x_at NM_007120 gi 45827763 ref NM_007120.2	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A4 (UGT1A4), mRNA

14/34

SEQ ID NO: 40	204532_x_at NM_019075 gi 45827761 ref NM_019075.2	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A10 (UGT1A10), mRNA
SEQ ID NO: 41	204532_x_at NM_019076 gi 45827768 ref NM_019076.4	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A8 (UGT1A8), mRNA
SEQ ID NO: 42	204532_x_at NM_021027 gi 45827769 ref NM_021027.2	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A9 (UGT1A9), mRNA
SEQ ID NO: 43	204532_x_at NM_205862 gi 45827766 ref NM_205862.1	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A6 (UGT1A6), transcript variant 2, mRNA
SEQ ID NO: 44	204619_s_at NM_001126336 (Formerly CSPG2) gi 187607299 ref NM_001126336.1	Homo sapiens versican (VCAN), transcript variant 2, mRNA
SEQ ID NO: 45	204619_s_at NM_004385 gi 187607292 ref NM_004385.3	Homo sapiens versican (VCAN), transcript variant 1, mRNA
SEQ ID NO: 46	204647_at NM_004838 gi 21361156 ref NM_004838.2	Homo sapiens homer homolog 3 (Drosophila) (HOMER3), mRNA
SEQ ID NO: 47	204715_at NM_015368 gi 39995063 ref NM_015368.3	Homo sapiens pannexin 1 (PANX1), mRNA
SEQ ID NO: 48	204747_at NM_001031683 gi 72534657 ref NM_001031683.1	Homo sapiens interferon-induced protein with tetratricopeptide repeats 3 (IFIT3), mRNA
SEQ ID NO: 49	204747_at NM_001549 gi 142349653 ref NM_001549.3	Homo sapiens interferon-induced protein with tetratricopeptide repeats 3 (IFIT3), mRNA
SEQ ID NO: 50	204754_at NM_002126 gi 94983916 ref NM_002126.4	Homo sapiens hepatic leukemia factor (HLF), mRNA
SEQ ID NO: 51	204779_s_at NM_004502 gi 85068579 ref NM_004502.3	Homo sapiens homeobox B7 (HOXB7), mRNA
SEQ ID NO: 52	204879_at NM_001006624 gi 54792059 ref NM_001006624.1	Homo sapiens podoplanin (PDPN), transcript variant 3, mRNA
SEQ ID NO: 53	204879_at NM_001006625 gi 54792061 ref NM_001006625.1	Homo sapiens podoplanin (PDPN), transcript variant 4, mRNA
SEQ ID NO: 54	204879_at NM_006474 gi 54792055 ref NM_006474.4	Homo sapiens podoplanin (PDPN), transcript variant 1, mRNA
SEQ ID NO: 55	204879_at NM_198389 gi 54792057 ref NM_198389.2	Homo sapiens podoplanin (PDPN), transcript variant 2, mRNA
SEQ ID NO: 56	204972_at NM_001032731 gi 74229020 ref NM_001032731.1	Homo sapiens 2'-5'-oligoadenylate synthetase 2, 69/71kDa (OAS2), transcript variant 3, mRNA
SEQ ID NO: 57	204972_at NM_002535 gi 74229016 ref NM_002535.2	Homo sapiens 2'-5'-oligoadenylate synthetase 2, 69/71kDa (OAS2), transcript variant 2, mRNA
SEQ ID NO: 58	204972_at NM_016817 gi 74229018 ref NM_016817.2	Homo sapiens 2'-5'-oligoadenylate synthetase 2, 69/71kDa (OAS2), transcript variant 1, mRNA
SEQ ID NO: 59	205157_s_at NM_000422 gi 4557700 ref NM_000422.1	Homo sapiens keratin 17 (KRT17), mRNA

FIG. 7C

SEQ ID NO: 60	205200_at NM_003278 gi 156627578 ref NM_003278.2	Homo sapiens C-type lectin domain family 3, member B (CLEC3B), mRNA
SEQ ID NO: 61	205319_at NM_005672 gi 83641882 ref NM_005672.3	Homo sapiens prostate stem cell antigen (PSCA), mRNA
SEQ ID NO: 62	205428_s_at NM_001740 gi 153946408 ref NM_001740.3	Homo sapiens calbindin 2 (CALB2), transcript variant CALB2, mRNA
SEQ ID NO: 63	205428_s_at NM_007087 gi 153946407 ref NM_007087.2	Homo sapiens calbindin 2 (CALB2), transcript variant CALB2b, mRNA
SEQ ID NO: 64	205428_s_at NM_007088 gi 153946406 ref NM_007088.2	Homo sapiens calbindin 2 (CALB2), transcript variant CALB2c, mRNA
SEQ ID NO: 65	205483_s_at NM_005101 gi 193083170 ref NM_005101.3	Homo sapiens ISG15 ubiquitin-like modifier (ISG15), mRNA
SEQ ID NO: 66	205574_x_at NM_001199 gi 4502420 ref NM_001199.1	Homo sapiens bone morphogenetic protein 1 (BMP1), transcript variant BMP1-1, mRNA
SEQ ID NO: 67	205574_x_at NM_006128 gi 5902807 ref NM_006128.1	Homo sapiens bone morphogenetic protein 1 (BMP1), transcript variant BMP1-2, mRNA
SEQ ID NO: 68	205574_x_at NM_006129 gi 5902809 ref NM_006129.2	Homo sapiens bone morphogenetic protein 1 (BMP1), transcript variant BMP1-3, mRNA
SEQ ID NO: 69	205730_s_at NM_014945 gi 151301062 ref NM_014945.2	Homo sapiens actin binding LIM protein family, member 3 (ABLM3), mRNA
SEQ ID NO: 70	206093_x_at NM_019105 gi 188528647 ref NM_019105.6	Homo sapiens tenascin XB (TNXB), transcript variant XB, mRNA
SEQ ID NO: 71	206093_x_at NM_032470 gi 188528649 ref NM_032470.3	Homo sapiens tenascin XB (TNXB), transcript variant XB-S, mRNA
SEQ ID NO: 72	206093_x_at NR_001284 gi 166063969 ref NR_001284.2	Homo sapiens tenascin XA pseudogene (TNXA) on chromosome 6
SEQ ID NO: 73	206094_x_at NM_000463 gi 45827762 ref NM_000463.2	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A1 (UGT1A1), mRNA
SEQ ID NO: 74	206094_x_at NM_001072 gi 132626744 ref NM_001072.3	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A6 (UGT1A6), transcript variant 1, mRNA
SEQ ID NO: 75	206094_x_at NM_007120 gi 45827763 ref NM_007120.2	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A4 (UGT1A4), mRNA
SEQ ID NO: 76	206094_x_at NM_019075 gi 45827761 ref NM_019075.2	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A10 (UGT1A10), mRNA
SEQ ID NO: 77	206094_x_at NM_019076 gi 45827768 ref NM_019076.4	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A8 (UGT1A8), mRNA
SEQ ID NO: 78	206094_x_at NM_019077 gi 41282212 ref NM_019077.2	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A7 (UGT1A7), mRNA

SEQ ID NO: 79	206094_x_at NM_019078 gi 46249403 ref NM_019078.1	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A5 (UGT1A5), mRNA
SEQ ID NO: 80	206094_x_at NM_019093 gi 41282228 ref NM_019093.2	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A3 (UGT1A3), mRNA
SEQ ID NO: 81	206094_x_at NM_021027 gi 45827769 ref NM_021027.2	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A9 (UGT1A9), mRNA
SEQ ID NO: 82	206094_x_at NM_205862 gi 45827766 ref NM_205862.1	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A6 (UGT1A6), transcript variant 2, mRNA
SEQ ID NO: 83	206094_x_at NM_000463 gi 45827762 ref NM_000463.2	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A1 (UGT1A1), mRNA
SEQ ID NO: 84	206094_x_at NM_001072 gi 132626744 ref NM_001072.3	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A6 (UGT1A6), transcript variant 1, mRNA
SEQ ID NO: 85	206094_x_at NM_007120 gi 45827763 ref NM_007120.2	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A4 (UGT1A4), mRNA
SEQ ID NO: 86	206094_x_at NM_019075 gi 45827761 ref NM_019075.2	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A10 (UGT1A10), mRNA
SEQ ID NO: 87	206094_x_at NM_019076 gi 45827768 ref NM_019076.4	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A8 (UGT1A8), mRNA
SEQ ID NO: 88	206094_x_at NM_021027 gi 45827769 ref NM_021027.2	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A9 (UGT1A9), mRNA
SEQ ID NO: 89	206094_x_at NM_205862 gi 45827766 ref NM_205862.1	Homo sapiens UDP glucuronosyltransferase 1 family, polypeptide A6 (UGT1A6), transcript variant 2, mRNA
SEQ ID NO: 90	207206_s_at NM_000697 gi 154426291 ref NM_000697.2	Homo sapiens arachidonate 12- lipoxygenase (ALOX12), mRNA
SEQ ID NO: 91	207517_at NM_005562 gi 157419137 ref NM_005562.2	Homo sapiens laminin, gamma 2 (LAMC2), transcript variant 1, mRNA
SEQ ID NO: 92	207517_at NM_018891 gi 157419139 ref NM_018891.2	Homo sapiens laminin, gamma 2 (LAMC2), transcript variant 2, mRNA
SEQ ID NO: 93	207850_at NM_002090 gi 54144649 ref NM_002090.2	Homo sapiens chemokine (C-X-C motif) ligand 3 (CXCL3), mRNA
SEQ ID NO: 94	208156_x_at NM_031308 gi 13876385 ref NM_031308.1	Homo sapiens epiplakin 1 (EPPK1), mRNA
SEQ ID NO: 95	208609_s_at NM_019105 gi 188528647 ref NM_019105.6	Homo sapiens tenascin XB (TNXB), transcript variant XB, mRNA
SEQ ID NO: 96	208609_s_at NM_032470 gi 188528649 ref NM_032470.3	Homo sapiens tenascin XB (TNXB), transcript variant XB-S, mRNA

17/34

SEQ ID NO: 97	208851_s_at NM_006288 gi 19923361 ref NM_006288.2	Homo sapiens Thy-1 cell surface antigen (THY1), mRNA
SEQ ID NO: 98	209774_x_at NM_002089 gi 148298657 ref NM_002089.3	Homo sapiens chemokine (C-X-C motif) ligand 2 (CXCL2), mRNA
SEQ ID NO: 99	209800_at NM_005557 gi 163659905 ref NM_005557.3	Homo sapiens keratin 16 (KRT16), mRNA
SEQ ID NO: 100	209900_s_at NM_003051 gi 115583684 ref NM_003051.3	Homo sapiens solute carrier family 16, member 1 (monocarboxylic acid transporter 1) (SLC16A1), mRNA
SEQ ID NO: 101	209949_at NM_000433 gi 189083740 ref NM_000433.3	Homo sapiens neutrophil cytosolic factor 2 (NCF2), transcript variant 1, mRNA
SEQ ID NO: 102	209969_s_at NM_007315 gi 189458859 ref NM_007315.3	Homo sapiens signal transducer and activator of transcription 1, 91kDa (STAT1), transcript variant alpha, mRNA
SEQ ID NO: 103	209969_s_at NM_139266 gi 189458858 ref NM_139266.2	Homo sapiens signal transducer and activator of transcription 1, 91kDa (STAT1), transcript variant beta, mRNA
SEQ ID NO: 104	209975_at NM_000773 gi 75709190 ref NM_000773.3	Homo sapiens cytochrome P450, family 2, subfamily E, polypeptide 1 (CYP2E1), mRNA
SEQ ID NO: 105	210001_s_at NM_003745 gi 4507232 ref NM_003745.1	Homo sapiens suppressor of cytokine signaling 1 (SOCS1), mRNA
SEQ ID NO: 106	210797_s_at NM_003733 gi 38016933 ref NM_003733.2	Homo sapiens 2'-5'-oligoadenylate synthetase-like (OASL), transcript variant 1, mRNA
SEQ ID NO: 107	210797_s_at NM_198213 gi 38016929 ref NM_198213.1	Homo sapiens 2'-5'-oligoadenylate synthetase-like (OASL), transcript variant 2, mRNA
SEQ ID NO: 108	210809_s_at NM_006475 gi 5453833 ref NM_006475.1	Homo sapiens periostin, osteoblast specific factor (POSTN), mRNA
SEQ ID NO: 109	210868_s_at NM_024090 gi 195539341 ref NM_024090.2	Homo sapiens ELOVL family member 6, elongation of long chain fatty acids (FEN1/Elo2, SUR4/Elo3-like, yeast) (ELOVL6), transcript variant 1, mRNA
SEQ ID NO: 110	211924_s_at NM_001005376 gi 53829378 ref NM_001005376.1	Homo sapiens plasminogen activator, urokinase receptor (PLAUR), transcript variant 2, mRNA
SEQ ID NO: 111	211924_s_at NM_001005377 gi 53829380 ref NM_001005377.1	Homo sapiens plasminogen activator, urokinase receptor (PLAUR), transcript variant 3, mRNA
SEQ ID NO: 112	211924_s_at NM_002659 gi 53829377 ref NM_002659.2	Homo sapiens plasminogen activator, urokinase receptor (PLAUR), transcript variant 1, mRNA
SEQ ID NO: 113	211980_at NM_001845 gi 148536824 ref NM_001845.4	Homo sapiens collagen, type IV, alpha 1 (COL4A1), mRNA
SEQ ID NO: 114	212012_at NM_012293 gi 109150415 ref NM_012293.1	Homo sapiens peroxidase homolog (Drosophila) (PXDN), mRNA

FIG. 7F

SEQ ID NO: 115	212224_at NM_000689 gi 25777722 ref NM_000689.3	Homo sapiens aldehyde dehydrogenase 1 family, member A1 (ALDH1A1), mRNA
SEQ ID NO: 116	212344_at NM_015170 gi 189571635 ref NM_015170.2	Homo sapiens sulfatase 1 (SULF1), transcript variant 3, mRNA
SEQ ID NO: 117	212354_at NM_015170 gi 189571635 ref NM_015170.2	Homo sapiens sulfatase 1 (SULF1), transcript variant 3, mRNA
SEQ ID NO: 118	212365_at NM_012223 gi 194328683 ref NM_012223.3	Homo sapiens myosin IB (MYO1B), transcript variant 2, mRNA
SEQ ID NO: 119	212472_at NM_014632 gi 41281417 ref NM_014632.2	Homo sapiens microtubule associated monooxygenase, calponin and LIM domain containing 2 (MICAL2), mRNA
SEQ ID NO: 120	212473_s_at NM_014632 gi 41281417 ref NM_014632.2	Homo sapiens microtubule associated monooxygenase, calponin and LIM domain containing 2 (MICAL2), mRNA
SEQ ID NO: 121	212488_at NM_000093 gi 89276750 ref NM_000093.3	Homo sapiens collagen, type V, alpha 1 (COL5A1), mRNA
SEQ ID NO: 122	212489_at NM_000093 gi 89276750 ref NM_000093.3	Homo sapiens collagen, type V, alpha 1 (COL5A1), mRNA
SEQ ID NO: 123	213421_x_at NM_002771 gi 170014742 ref NM_002771.3	Homo sapiens protease, serine, 3 (PRSS3), transcript variant 2, mRNA
SEQ ID NO: 124	213421_x_at NM_007343 gi 170296789 ref NM_007343.2	Homo sapiens protease, serine, 3 (PRSS3), transcript variant 1, mRNA
SEQ ID NO: 125	213668_s_at NM_003107 gi 30179901 ref NM_003107.2	Homo sapiens SRY (sex determining region Y)-box 4 (SOX4), mRNA
SEQ ID NO: 126	213869_x_at NM_006288 gi 19923361 ref NM_006288.2	Homo sapiens Thy-1 cell surface antigen (THY1), mRNA
SEQ ID NO: 127	214453_s_at NM_006417 gi 166706910 ref NM_006417.4	Homo sapiens interferon-induced protein 44 (IFI44), mRNA
SEQ ID NO: 128	217312_s_at NM_000094 gi 157389010 ref NM_000094.3	Homo sapiens collagen, type VII, alpha 1 (epidermolysis bullosa, dystrophic, dominant and recessive) (COL7A1), mRNA
SEQ ID NO: 129	217430_x_at NM_000088 gi 110349771 ref NM_000088.3	Homo sapiens collagen, type I, alpha 1 (COL1A1), mRNA
SEQ ID NO: 130	218404_at NM_013322 gi 23111022 ref NM_013322.2	Homo sapiens sorting nexin 10 (SNX10), mRNA
SEQ ID NO: 131	218779_x_at NM_017729 gi 148539854 ref NM_017729.3	Homo sapiens EPS8-like 1 (EPS8L1), transcript variant 2, mRNA
SEQ ID NO: 132	218779_x_at NM_133180 gi 148539852 ref NM_133180.2	Homo sapiens EPS8-like 1 (EPS8L1), transcript variant 1, mRNA
SEQ ID NO: 133	218885_s_at NM_024642 gi 112807220 ref NM_024642.3	Homo sapiens UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 12 (GalNAc-T12) (GALNT12), mRNA
SEQ ID NO: 134	218888_s_at NM_018092 gi 24041025 ref NM_018092.3	Homo sapiens neuropilin (NRP) and tolloid (TLL)-like 2 (NETO2), mRNA
SEQ ID NO: 135	218935_at NM_014600 gi 7657055 ref NM_014600.1	Homo sapiens EH-domain containing 3 (EHD3), mRNA

19/34

SEQ ID NO: 136	219211_at NM_017414 gi 156071510 ref NM_017414.3	Homo sapiens ubiquitin specific peptidase 18 (USP18), mRNA
SEQ ID NO: 137	219300_s_at NM_014141 gi 89213248 ref NM_014141.4	Homo sapiens contactin associated protein-like 2 (CNTNAP2), mRNA
SEQ ID NO: 138	219725_at NM_018965 gi 189458827 ref NM_018965.2	Homo sapiens triggering receptor expressed on myeloid cells 2 (TREM2), mRNA
SEQ ID NO: 139	219863_at NM_016323 gi 110825981 ref NM_016323.2	Homo sapiens hect domain and RLD 5 (HERC5), mRNA
SEQ ID NO: 140	220016_at NM_001620 gi 61743953 ref NM_001620.1	Homo sapiens AHNAK nucleoprotein (AHNAK), transcript variant 1, mRNA
SEQ ID NO: 141	220016_at NM_024060 gi 55769540 ref NM_024060.2	Homo sapiens AHNAK nucleoprotein (AHNAK), transcript variant 2, mRNA
SEQ ID NO: 142	220037_s_at NM_006691 gi 151301201 ref NM_006691.3	Homo sapiens lymphatic vessel endothelial hyaluronan receptor 1 (LYVE1), mRNA (formerly XLKD1)
SEQ ID NO: 143	220149_at NM_001085437 gi 146198548 ref NM_001085437.1	Homo sapiens chromosome 2 open reading frame 54 (C2orf54), transcript variant 1, mRNA (formerly FLJ22671)
SEQ ID NO: 144	220149_at NM_024861 gi 146198667 ref NM_024861.2	Homo sapiens chromosome 2 open reading frame 54 (C2orf54), transcript variant 2, mRNA
SEQ ID NO: 145	220962_s_at NM_013358 gi 122056684 ref NM_013358.2	Homo sapiens peptidyl arginine deiminase, type I (PADI1), mRNA
SEQ ID NO: 146	221665_s_at NM_017729 gi 148539854 ref NM_017729.3	Homo sapiens EPS8-like 1 (EPS8L1), transcript variant 2, mRNA
SEQ ID NO: 147	221665_s_at NM_133180 gi 148539852 ref NM_133180.2	Homo sapiens EPS8-like 1 (EPS8L1), transcript variant 1, mRNA
SEQ ID NO: 148	221729_at NM_000393 gi 89363016 ref NM_000393.3	Homo sapiens collagen, type V, alpha 2 (COL5A2), mRNA
SEQ ID NO: 149	221730_at NM_000393 gi 89363016 ref NM_000393.3	Homo sapiens collagen, type V, alpha 2 (COL5A2), mRNA
SEQ ID NO: 150	221898_at NM_001006624 gi 54792059 ref NM_001006624.1	Homo sapiens podoplanin (PDPN), transcript variant 3, mRNA
SEQ ID NO: 151	221898_at NM_001006625 gi 54792061 ref NM_001006625.1	Homo sapiens podoplanin (PDPN), transcript variant 4, mRNA
SEQ ID NO: 152	221898_at NM_006474 gi 54792055 ref NM_006474.4	Homo sapiens podoplanin (PDPN), transcript variant 1, mRNA
SEQ ID NO: 153	221898_at NM_198389 gi 54792057 ref NM_198389.2	Homo sapiens podoplanin (PDPN), transcript variant 2, mRNA
SEQ ID NO: 154	222693_at NM_022763 gi 27477058 ref NM_022763.2	Homo sapiens fibronectin type III domain containing 3B (FNDC3B), mRNA
SEQ ID NO: 155	223541_at NM_005329 gi 20302152 ref NM_005329.2	Homo sapiens hyaluronan synthase 3 (HAS3), transcript variant 1, mRNA
SEQ ID NO: 156	223541_at NM_138612 gi 20302154 ref NM_138612.1	Homo sapiens hyaluronan synthase 3 (HAS3), transcript variant 2, mRNA
SEQ ID NO: 157	225288_at NM_032888 gi 32140759 ref NM_032888.2	Homo sapiens collagen, type XXVII, alpha 1 (COL27A1), mRNA

FIG. 7H

20/34

SEQ ID NO: 158	225292_at NM_032888 gi 32140759 ref NM_032888.2	Homo sapiens collagen, type XXVII, alpha 1 (COL27A1), mRNA
SEQ ID NO: 159	225520_at NM_015440 gi 40018634 ref NM_015440.3	Homo sapiens methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1-like (MTHFD1L), mRNA
SEQ ID NO: 160	225548_at NM_020859 gi 110835740 ref NM_020859.2	Homo sapiens shroom family member 3 (SHROOM3), mRNA
SEQ ID NO: 161	225681_at NM_138455 gi 34147546 ref NM_138455.2	Homo sapiens collagen triple helix repeat containing 1 (CTHRC1), mRNA
SEQ ID NO: 162	226303_at NM_021965 gi 68299819 ref NM_021965.3	Homo sapiens phosphoglucomutase 5 (PGM5), mRNA
SEQ ID NO: 163	226997_at NM_030955 gi 51558723 ref NM_030955.2	Homo sapiens ADAM metalloproteinase with thrombospondin type 1 motif, 12 (ADAMTS12), mRNA
SEQ ID NO: 164	227419_x_at NM_001012973 gi 61175221 ref NM_001012973.1	Homo sapiens placenta-specific 9 (PLAC9), mRNA
SEQ ID NO: 165	227697_at NM_003955 gi 45439351 ref NM_003955.3	Homo sapiens suppressor of cytokine signaling 3 (SOCS3), mRNA
SEQ ID NO: 166	227782_at NM_001039360 gi 86262148 ref NM_001039360.1	Homo sapiens zinc finger and BTB domain containing 7C (ZBTB7C), mRNA
SEQ ID NO: 167	229055_at NM_003485 gi 74316010 ref NM_003485.3	Homo sapiens G protein-coupled receptor 68 (GPR68), mRNA
SEQ ID NO: 168	229225_at NM_003872 gi 41872532 ref NM_003872.2	Homo sapiens neuropilin 2 (NRP2), transcript variant 2, mRNA
SEQ ID NO: 169	229225_at NM_018534 gi 41872543 ref NM_018534.3	Homo sapiens neuropilin 2 (NRP2), transcript variant 4, mRNA
SEQ ID NO: 170	229225_at NM_201264 gi 41872556 ref NM_201264.1	Homo sapiens neuropilin 2 (NRP2), transcript variant 6, mRNA
SEQ ID NO: 171	229225_at NM_201266 gi 41872561 ref NM_201266.1	Homo sapiens neuropilin 2 (NRP2), transcript variant 1, mRNA
SEQ ID NO: 172	229225_at NM_201267 gi 41872566 ref NM_201267.1	Homo sapiens neuropilin 2 (NRP2), transcript variant 5, mRNA
SEQ ID NO: 173	229225_at NM_201279 gi 41872571 ref NM_201279.1	Homo sapiens neuropilin 2 (NRP2), transcript variant 3, mRNA
SEQ ID NO: 174	229860_x_at XM_001716080 gi 169166560 ref XM_001716080.1	PREDICTED: Homo sapiens hypothetical gene supported by BC038466; BC062790 (LOC401115), mRNA
SEQ ID NO: 175	229860_x_at XM_379250 gi 113415447 ref XM_379250.4	PREDICTED: Homo sapiens hypothetical gene supported by BC038466; BC062790 (LOC401115), mRNA
SEQ ID NO: 176	229860_x_at XM_937968 gi 113415961 ref XM_937968.2	PREDICTED: Homo sapiens hypothetical gene supported by BC038466; BC062790 (LOC401115), mRNA
SEQ ID NO: 177	229947_at NM_015886 gi 66932906 ref NM_015886.3	Homo sapiens peptidase inhibitor 15 (PI15), mRNA

FIG. 71

21/34

SEQ ID NO: 178	230104_s_at NM_007030 gi 149192856 ref NM_007030.2	Homo sapiens tubulin polymerization promoting protein (TPPP), mRNA
SEQ ID NO: 179	231118_at NM_144698 gi 121582654 ref NM_144698.3	Homo sapiens ankyrin repeat domain 35 (ANKRD35), mRNA
SEQ ID NO: 180	235276_at NM_001002264 gi 50428916 ref NM_001002264.1	Homo sapiens epithelial stromal interaction 1 (breast) (EPSTI1), transcript variant 1, mRNA (formerly unnamed)
SEQ ID NO: 181	235276_at NM_033255 gi 50428918 ref NM_033255.2	Homo sapiens epithelial stromal interaction 1 (breast) (EPSTI1), transcript variant 2, mRNA
SEQ ID NO: 182	238581_at NM_052942 gi 31377630 ref NM_052942.2	Homo sapiens guanylate binding protein 5 (GBP5), mRNA
SEQ ID NO: 183	238715_at XM_001724434 gi 169211446 ref XM_001724434.1	PREDICTED: Homo sapiens hypothetical protein LOC646014 (LOC646014), mRNA (formerly ARHGAP27)
SEQ ID NO: 184	238715_at XM_001726150 gi 169211897 ref XM_001726150.1	PREDICTED: Homo sapiens hypothetical protein LOC646014 (LOC646014), mRNA
SEQ ID NO: 185	238715_at XM_928978 gi 169211018 ref XM_928978.3	PREDICTED: Homo sapiens hypothetical protein LOC646014 (LOC646014), mRNA
SEQ ID NO: 186	241233_x_at XR_040134 gi 169218103 ref XR_040134.1	PREDICTED: Homo sapiens misc_RNA (C21orf81), miscRNA
SEQ ID NO: 187	241233_x_at XR_040135 gi 169215161 ref XR_040135.1	PREDICTED: Homo sapiens misc_RNA (C21orf81), miscRNA
SEQ ID NO: 188	242009_at NM_001045 gi 145553964 ref NM_001045.3	Homo sapiens solute carrier family 6 (neurotransmitter transporter, serotonin), member 4 (SLC6A4), mRNA (formerly unnamed)
SEQ ID NO: 189	242417_at XM_495868 gi 89033832 ref XM_495868.2	PREDICTED: Homo sapiens hypothetical protein LOC283278 (LOC283278), mRNA
SEQ ID NO: 190	39402_at NM_000576 gi 27894305 ref NM_000576.2	Homo sapiens interleukin 1, beta (IL1B), mRNA
SEQ ID NO: 191	41037_at NM_003213 gi 46411151 ref NM_003213.2	Homo sapiens TEA domain family member 4 (TEAD4), transcript variant 1, mRNA
SEQ ID NO: 192	41037_at NM_201441 gi 46411155 ref NM_201441.1	Homo sapiens TEA domain family member 4 (TEAD4), transcript variant 2, mRNA
SEQ ID NO: 193	41037_at NM_201443 gi 46411158 ref NM_201443.1	Homo sapiens TEA domain family member 4 (TEAD4), transcript variant 3, mRNA

FIG. 7J

22/34

SEQ ID NO: 194	217519_at AI246331 (Formerly <i>MACF1</i>) gi 3841728 gb AI246331.1 AI246331 qu39f10.x1 NCI_CGAP_Lym5	Homo sapiens cDNA clone IMAGE:1967179 3' similar to contains Alu repetitive element;, mRNA sequence
SEQ ID NO: 195	241872_at AI149963 (Formerly <i>SGIP1</i>) gi 3678432 gb AI149963.1 AI149963 qf38f12.x1 Soares_testis_NHT	Homo sapiens cDNA clone IMAGE:1752335 3' similar to contains MER3.b2 MER3 repetitive element ;, mRNA sequence
SEQ ID NO: 196	222344_at AW972765 (Formerly <i>C5orf13</i>) gi 8162611 gb AW972765.1 AW972765 EST384859 MAGE resequenes, MAGL	Homo sapiens cDNA, mRNA sequence
SEQ ID NO: 197	229554_at AI141861 gi 3649318 gb AI141861.1 AI141861 qa66e06.x1 Soares_fetal_heart_NbHH19W	Homo sapiens cDNA clone IMAGE:1691746 3', mRNA sequence
SEQ ID NO: 198	230740_at AW576212 gi 7247751 gb AW576212.1 AW576212 UI-HF-BN0-afs-g-10-0-UI.s1 NIH_MGC_50	Homo sapiens cDNA clone IMAGE:3067963 3', mRNA sequence
SEQ ID NO: 199	240000_at AA621615 gi 2525554 gb AA621615.1 AA621615 af48a03.s1 Soares_total_fetus_Nb2HF8_9w	Homo sapiens cDNA clone IMAGE:1034860 3' similar to contains L1.b2 L1 repetitive element ;, mRNA sequence
SEQ ID NO: 200	243718_at AA846756 gi 2932896 gb AA846756.1 AA846756 aj99d09.s1 Soares_parathyroid_tumor_NbHPA	Homo sapiens cDNA clone IMAGE:1404593 3', mRNA sequence

FIG. 7K

23/34

SEQUENCE IDENTIFIER	AFFYMETRIX IDENTIFIER	GENE TARGET
SEQ ID NO: 201	202311_s_at	COL1A1
SEQ ID NO: 202	202404_s_at	COL1A2
SEQ ID NO: 203	202310_s_at	COL1A1
SEQ ID NO: 204	211980_at	COL4A1
SEQ ID NO: 205	202267_at	LAMC2
SEQ ID NO: 206	204415_at	IFI6
SEQ ID NO: 207	225681_at	CTHRC1
SEQ ID NO: 208	212488_at	COL5A1
SEQ ID NO: 209	211924_s_at	PLAUR
SEQ ID NO: 210	203256_at	CDH3
SEQ ID NO: 211	221729_at	COL5A2
SEQ ID NO: 212	213869_x_at	THY1
SEQ ID NO: 213	217312_s_at	COL7A1
SEQ ID NO: 214	1555778_a_at	POSTN
SEQ ID NO: 215	212489_at	COL5A1
SEQ ID NO: 216	221730_at	COL5A2
SEQ ID NO: 217	212354_at	SULF1
SEQ ID NO: 218	207517_at	LAMC2
SEQ ID NO: 219	212344_at	SULF1
SEQ ID NO: 220	204715_at	PANX1
SEQ ID NO: 221	208851_s_at	THY1
SEQ ID NO: 222	222693_at	FNDC3B
SEQ ID NO: 223	204647_at	HOMER3
SEQ ID NO: 224	213668_s_at	SOX4
SEQ ID NO: 225	205574_x_at	BMP1
SEQ ID NO: 226	1555420_a_at	KLF7
SEQ ID NO: 227	217430_x_at	COL1A1
SEQ ID NO: 228	210809_s_at	POSTN
SEQ ID NO: 229	205157_s_at	KRT17
SEQ ID NO: 230	203695_s_at	DFNA5
SEQ ID NO: 231	203325_s_at	COL5A1
SEQ ID NO: 232	209900_s_at	SLC16A1
SEQ ID NO: 233	203085_s_at	TGFB1
SEQ ID NO: 234	229225_at	NRP2
SEQ ID NO: 235	225288_at	(COL27A1)
SEQ ID NO: 236	202235_at	SLC16A1
SEQ ID NO: 237	204114_at	NID2
SEQ ID NO: 238	229554_at	(transcribed locus)
SEQ ID NO: 239	214453_s_at	IFI44
SEQ ID NO: 240	212472_at	MICAL2
SEQ ID NO: 241	205483_s_at	ISG15
SEQ ID NO: 242	226997_at	(ADAMTS12)
SEQ ID NO: 243	212473_s_at	MICAL2
SEQ ID NO: 244	225292_at	COL27A1
SEQ ID NO: 245	212365_at	MYO1B
SEQ ID NO: 246	212012_at	PXDN
SEQ ID NO: 247	229860_x_at	LOC401115
SEQ ID NO: 248	219211_at	USP18
SEQ ID NO: 249	219863_at	HERC5
SEQ ID NO: 250	204619_s_at	CSPG2 (VCAN?)
SEQ ID NO: 251	203968_s_at	CDC6
SEQ ID NO: 252	208156_x_at	EPPK1
SEQ ID NO: 253	210797_s_at	OASL
SEQ ID NO: 254	1568765_at	SERPINE1
SEQ ID NO: 255	204972_at	OAS2

FIG. 8A

24/34

SEQ ID NO: 256	223541_at	HAS3
SEQ ID NO: 257	218888_s_at	NETO2
SEQ ID NO: 258	209949_at	NCF2
SEQ ID NO: 259	204779_s_at	HOXB7
SEQ ID NO: 260	41037_at	TEAD4
SEQ ID NO: 261	209800_at	KRT16
SEQ ID NO: 262	217519_at	MACF1
SEQ ID NO: 263	202238_s_at	NNMT
SEQ ID NO: 264	221898_at	PDPN
SEQ ID NO: 265	201108_s_at	THSB1
SEQ ID NO: 266	209969_s_at	STAT1
SEQ ID NO: 267	203921_at	CHST2
SEQ ID NO: 268	204103_at	CCL4
SEQ ID NO: 269	241872_at	SGIP1 (transcribed locus)
SEQ ID NO: 270	207850_at	CXCL3
SEQ ID NO: 271	204747_at	IFIT3
SEQ ID NO: 272	219725_at	TREM2
SEQ ID NO: 273	203915_at	CXCL9
SEQ ID NO: 274	204879_at	PDPN
SEQ ID NO: 275	1554008_at	OSMR
SEQ ID NO: 276	204051_s_at	SFRP4
SEQ ID NO: 277	227697_at	SOCS3
SEQ ID NO: 278	210001_s_at	SOCS1
SEQ ID NO: 279	235276_at	(EPSTI1)
SEQ ID NO: 280	222344_at	C5orf13
SEQ ID NO: 281	225520_at	MTHFD1L
SEQ ID NO: 282	218404_at	SNX10
SEQ ID NO: 283	229055_at	GPR68
SEQ ID NO: 284	209774_x_at	CXCL2
SEQ ID NO: 285	39402_at	IL1B
SEQ ID NO: 286	219300_s_at	CNTNAP2
SEQ ID NO: 287	229947_at	PI15
SEQ ID NO: 288	238581_at	GBP5
SEQ ID NO: 289	1553212_at	KRT78
SEQ ID NO: 290	220149_at	FLJ22671 (C2orf54)
SEQ ID NO: 291	241233_x_at	C21orf81
SEQ ID NO: 292	1569608_x_at	LOC643187
SEQ ID NO: 293	220962_s_at	PADI1
SEQ ID NO: 294	210868_s_at	ELOVL6
SEQ ID NO: 295	205319_at	PSCA
SEQ ID NO: 296	204754_at	HLF
SEQ ID NO: 297	218779_x_at	EPS8L1
SEQ ID NO: 298	221665_s_at	EPS8L1
SEQ ID NO: 299	220016_at	AHNAK
SEQ ID NO: 300	218885_s_at	GALNT12
SEQ ID NO: 301	231118_at	ANKRD35
SEQ ID NO: 302	225548_at	SHROOM3
SEQ ID NO: 303	206094_x_at	UGT1A6
SEQ ID NO: 304	206093_x_at	TNXB
SEQ ID NO: 305	218935_at	EHD3
SEQ ID NO: 306	207126_x_at	UGT1A4

FIG. 8B

25/34

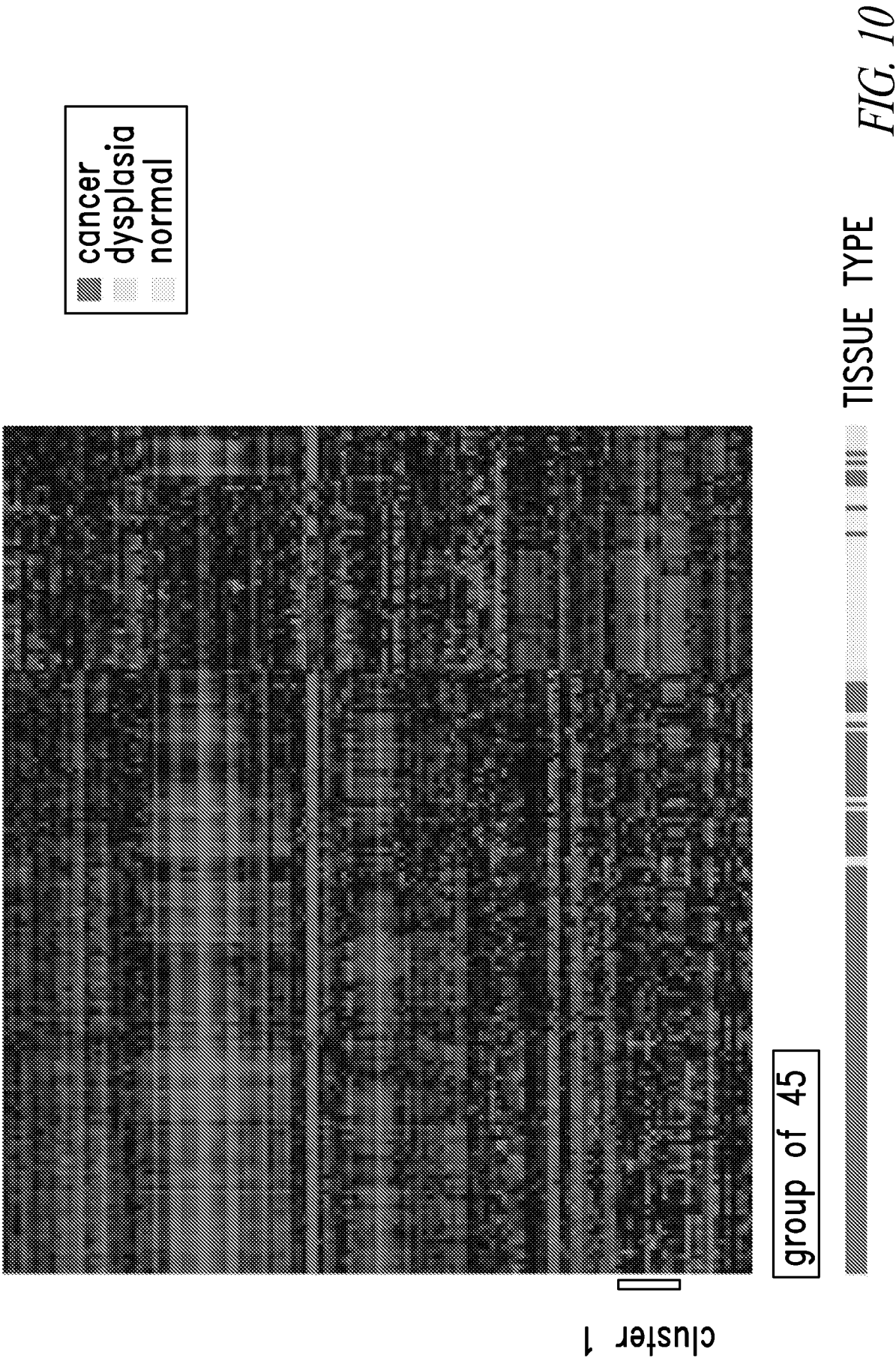
SEQ ID NO: 307	230740_at	(Transcribed locus)
SEQ ID NO: 308	204532_x_at	UGT1A4
SEQ ID NO: 309	242417_at	LOC283278
SEQ ID NO: 310	213421_x_at	PRSS3
SEQ ID NO: 311	205200_at	CLEC3B
SEQ ID NO: 312	1552283_s_at	ZDHHC11
SEQ ID NO: 313	220037_s_at	XLKD1 (LYVE1?)
SEQ ID NO: 314	1553861_at	TCP11L2
SEQ ID NO: 315	204378_at	BCAS1
SEQ ID NO: 316	242009_at	(SLC6A4)
SEQ ID NO: 317	207206_s_at	ALOX12
SEQ ID NO: 318	205730_s_at	ABLIM3
SEQ ID NO: 319	238715_at	ARHGAP27 (LOC646014?)
SEQ ID NO: 320	205428_s_at	CALB2
SEQ ID NO: 321	1565661_x_at	FUT6
SEQ ID NO: 322	208609_s_at	TNXA/TNXB
SEQ ID NO: 323	227782_at	ZBTB7C
SEQ ID NO: 324	226303_at	PGM5
SEQ ID NO: 325	240000_at	(Transcribed locus)
SEQ ID NO: 326	201497_x_at	MYH11
SEQ ID NO: 327	227419_x_at	PLAC9
SEQ ID NO: 328	230104_s_at	TPPP
SEQ ID NO: 329	212224_at	ALDH1A1
SEQ ID NO: 330	243718_at	(Transcribed locus)
SEQ ID NO: 331	209975_at	CYP2E1

FIG. 8C

26/34

SEQ ID NO: 205	202267_at	LAMC2
SEQ ID NO: 218	207517_at	LAMC2
SEQ ID NO: 254	1568765_at	SERPINE1
SEQ ID NO: 226	1555420_a_at	KLF7
SEQ ID NO: 222	222693_at	FNDC3B
SEQ ID NO: 270	207850_at	CXCL3
SEQ ID NO: 278	210001_s_at	SOCS1
SEQ ID NO: 234	229225_at	NRP2
SEQ ID NO: 277	227697_at	SOCS3
SEQ ID NO: 258	209949_at	NCF2
SEQ ID NO: 268	204103_at	CCL4
SEQ ID NO: 282	218404_at	SNX10
SEQ ID NO: 230	203695_s_at	DFNA5
SEQ ID NO: 217	212354_at	SULF1
SEQ ID NO: 247	229860_x_at	LOC401115
SEQ ID NO: 284	209774_x_at	CXCL2
SEQ ID NO: 269	241872_at	SGIP1 (transcribed locus)
SEQ ID NO: 251	203968_s_at	CDC6
SEQ ID NO: 281	225520_at	MTHFD1L
SEQ ID NO: 287	229947_at	PI15
SEQ ID NO: 271	204747_at	IFIT3
SEQ ID NO: 285	39402_at	IL1B
SEQ ID NO: 279	235276_at	(EPST11)
SEQ ID NO: 273	203915_at	CXCL9
SEQ ID NO: 259	204779_s_at	HOXB7
SEQ ID NO: 248	219211_at	USP18
SEQ ID NO: 288	238581_at	GBP5
SEQ ID NO: 286	219300_s_at	CNTNAP2
SEQ ID NO: 276	204051_s_at	SFRP4
SEQ ID NO: 311	205200_at	CLEC3B
SEQ ID NO: 301	231118_at	ANKRD35
SEQ ID NO: 304	206093_x_at	TNXB
SEQ ID NO: 323	227782_at	ZBTB7C
SEQ ID NO: 312	1552283_s_at	ZDHC11
SEQ ID NO: 310	213421_x_at	PRSS3
SEQ ID NO: 319	238715_at	ARHGAP27 (LOC646014?)
SEQ ID NO: 322	208609_s_at	TNXA/TNXB
SEQ ID NO: 316	242009_at	(SLC6A4)
SEQ ID NO: 324	226303_at	PGM5
SEQ ID NO: 317	207206_s_at	ALOX12
SEQ ID NO: 328	230104_s_at	TPPP
SEQ ID NO: 313	220037_s_at	XLKD1 (LYVE1?)
SEQ ID NO: 330	243718_at	(Transcribed locus)
SEQ ID NO: 327	227419_x_at	PLAC9
SEQ ID NO: 326	201497_x_at	MYH11
SEQ ID NO: 308	204532_x_at	UGT1A4
SEQ ID NO: 303	206094_x_at	UGT1A6
SEQ ID NO: 306	207126_x_at	UGT1A4
SEQ ID NO: 329	212224_at	ALDH1A1

FIG. 9



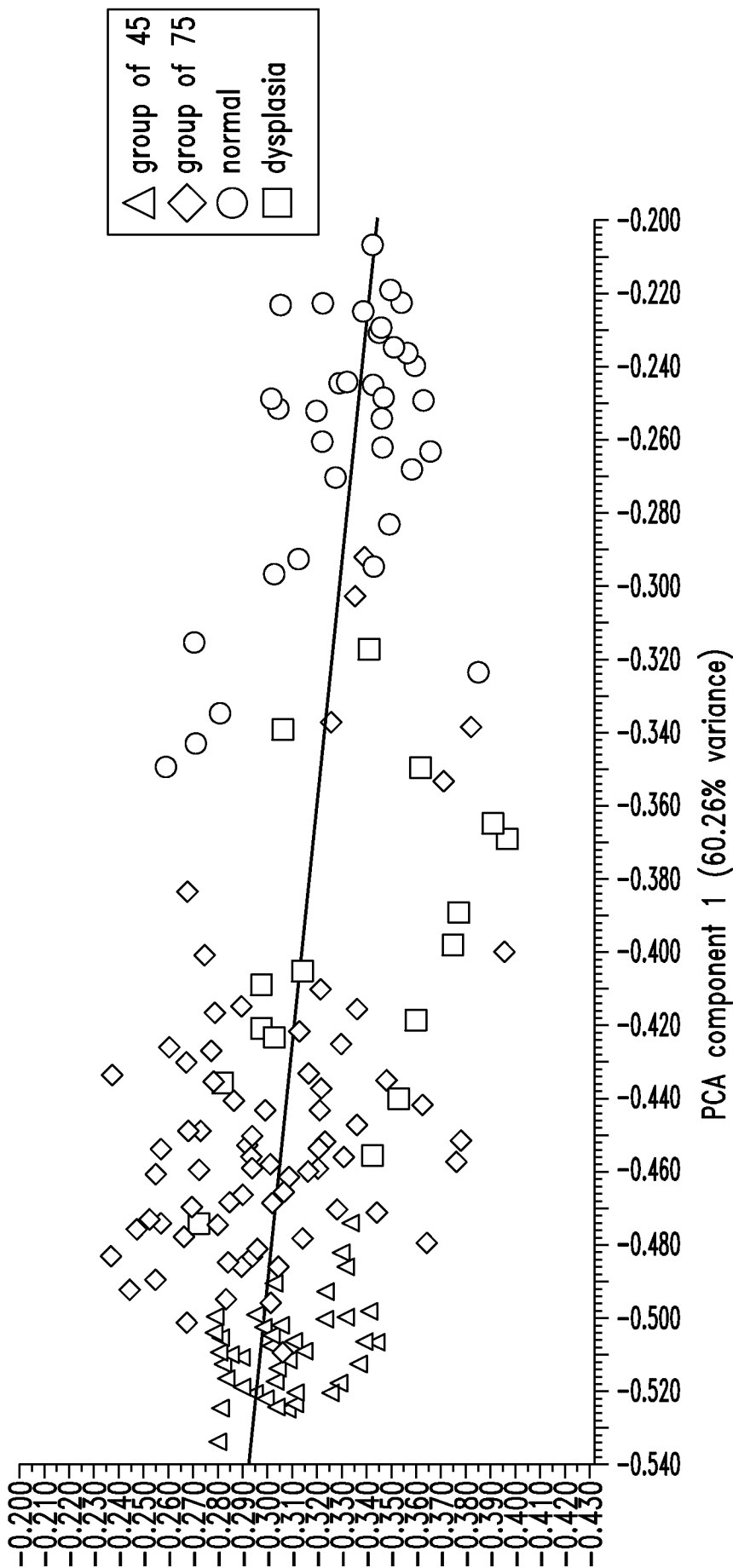


FIG. 11

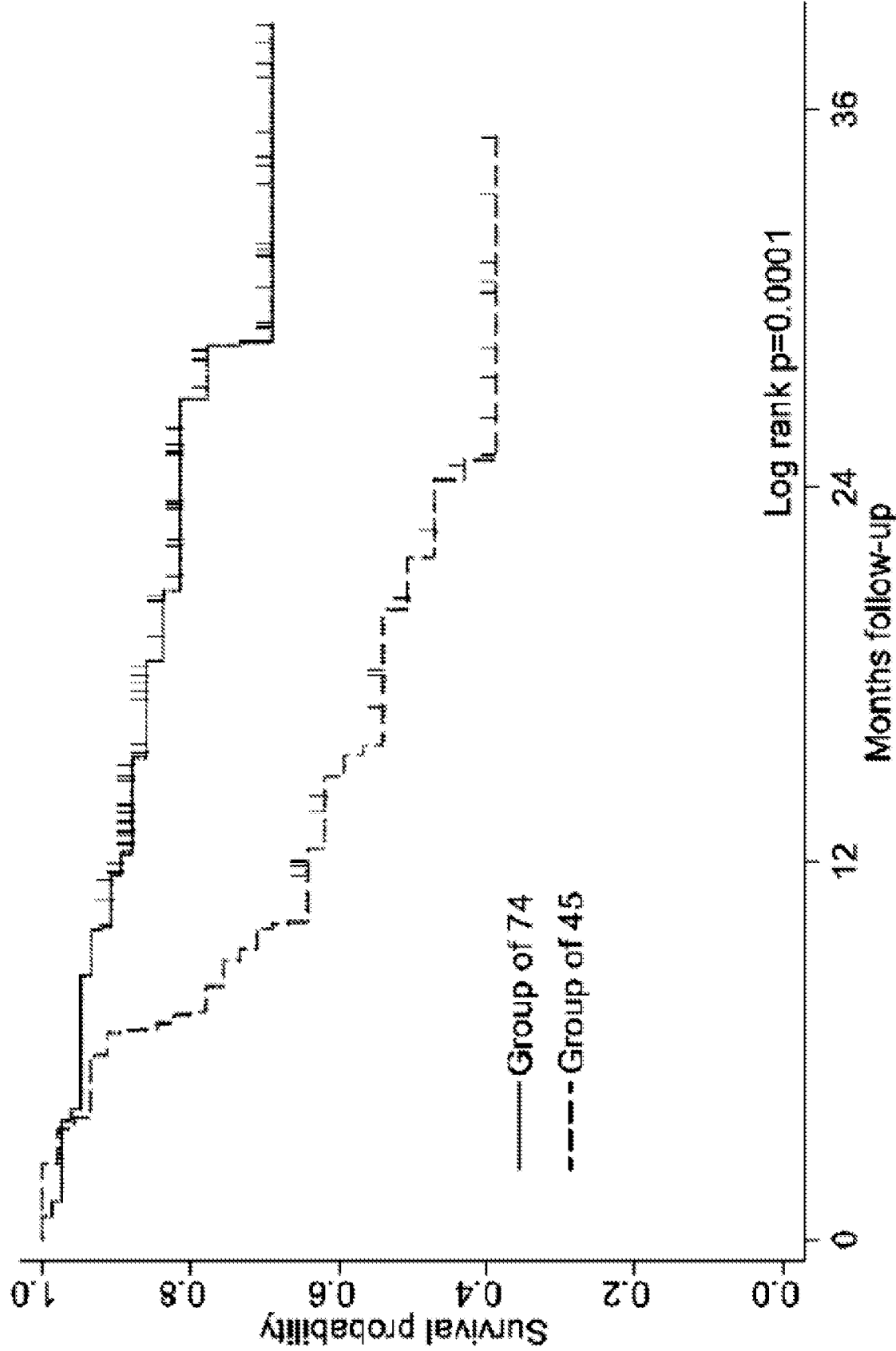


FIG. 12A

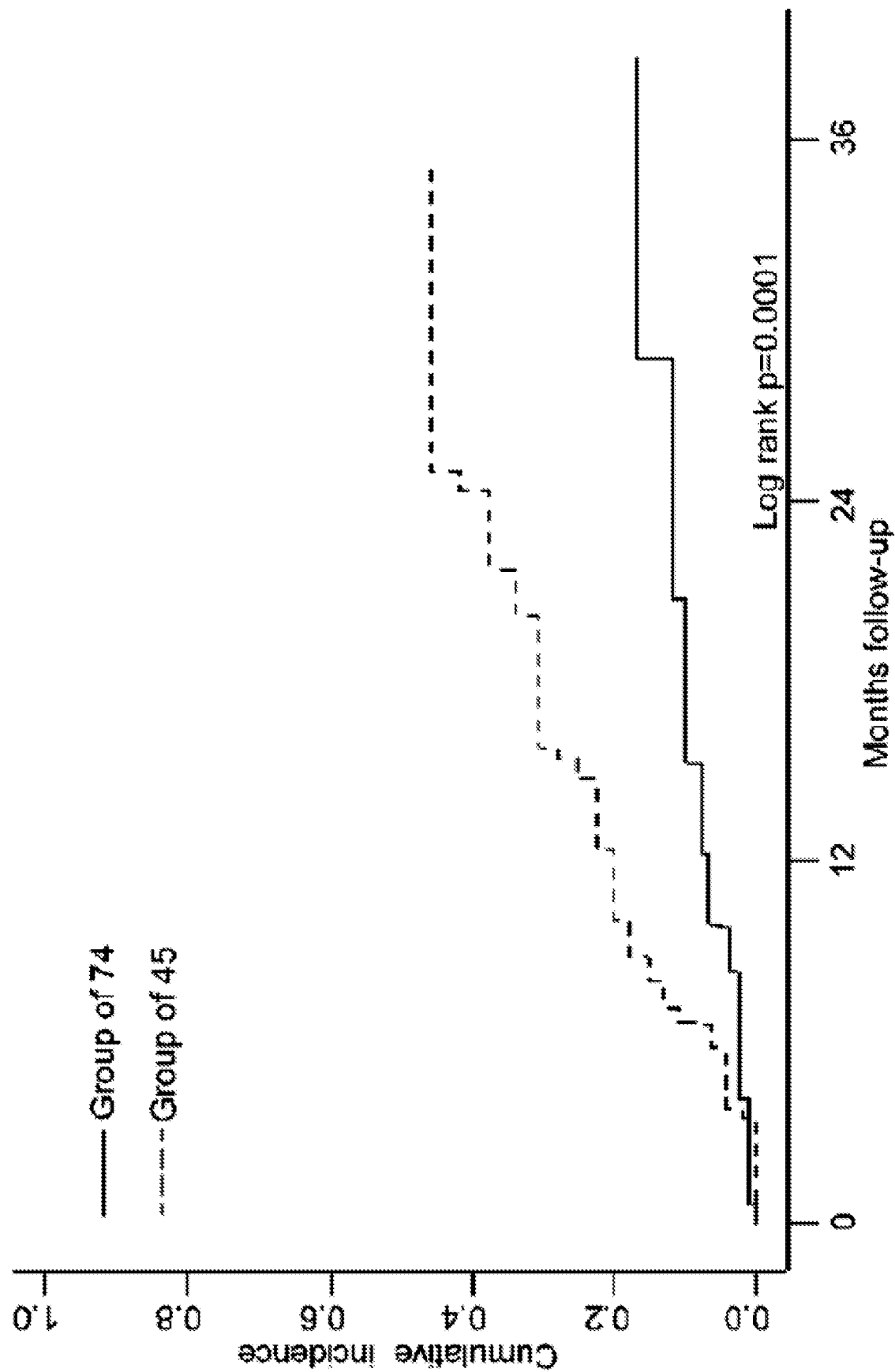


FIG. 12B

31/34

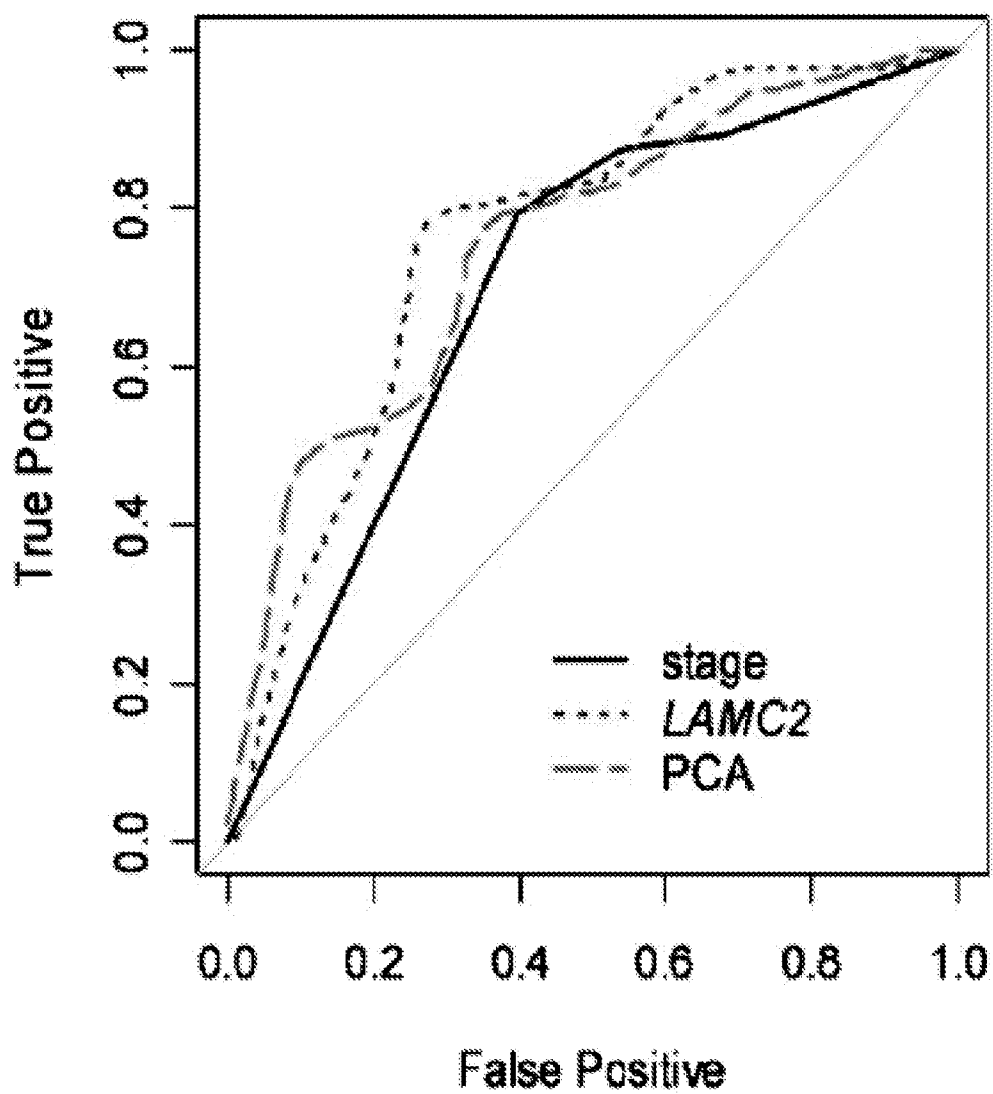


FIG. 13A

32/34

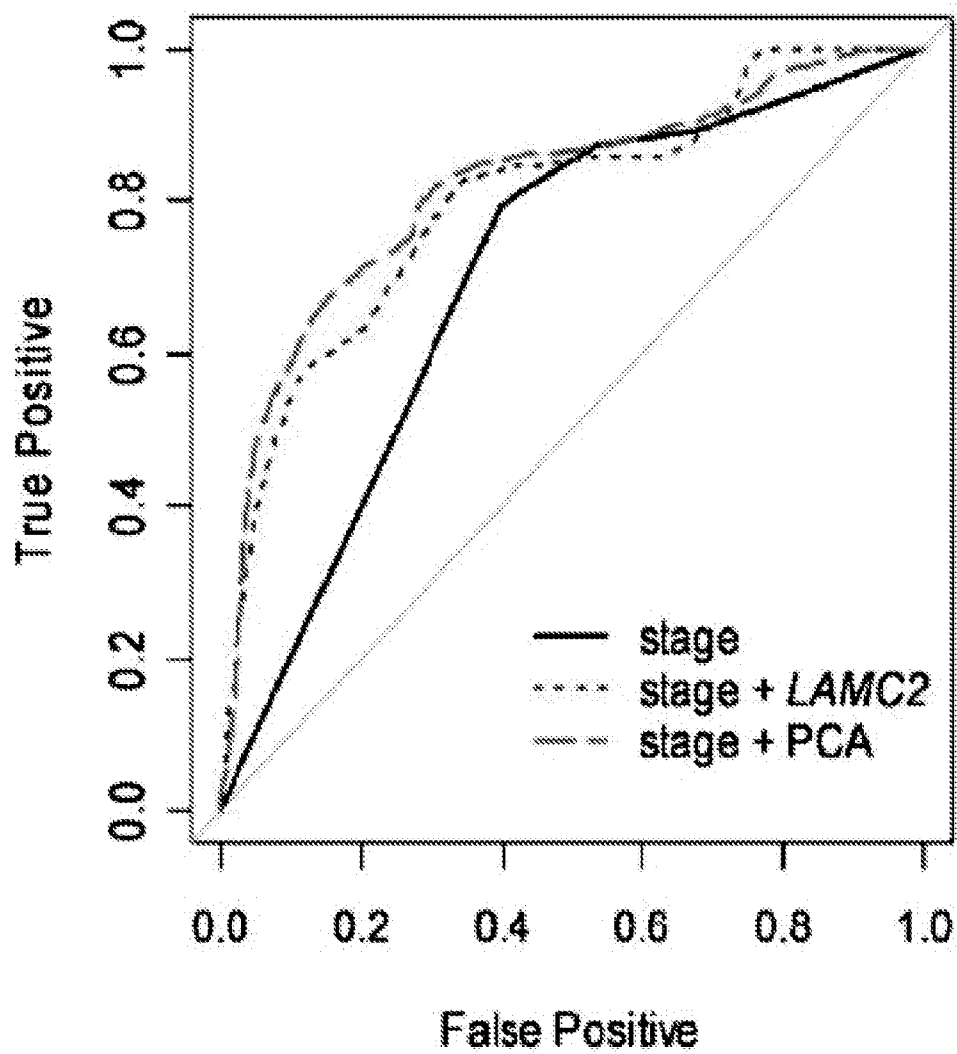


FIG. 13B

33/34

Model terms	AUC	bootstrapped 95% CI
stage	0.70	(0.60, 0.80)
<i>LAMC2</i>	0.77	(0.67, 0.86)
PCA	0.76	(0.64, 0.87)
stage and <i>LAMC2</i> [*]	0.80	(0.69, 0.91)
stage and PCA [†]	0.82	(0.70, 0.93)

^{*}p=0.0013 for 'stage' vs 'stage + *LAMC2*'[†]p=0.0008 for 'stage' vs. 'stage + PCA'*FIG. 13C*

34/34

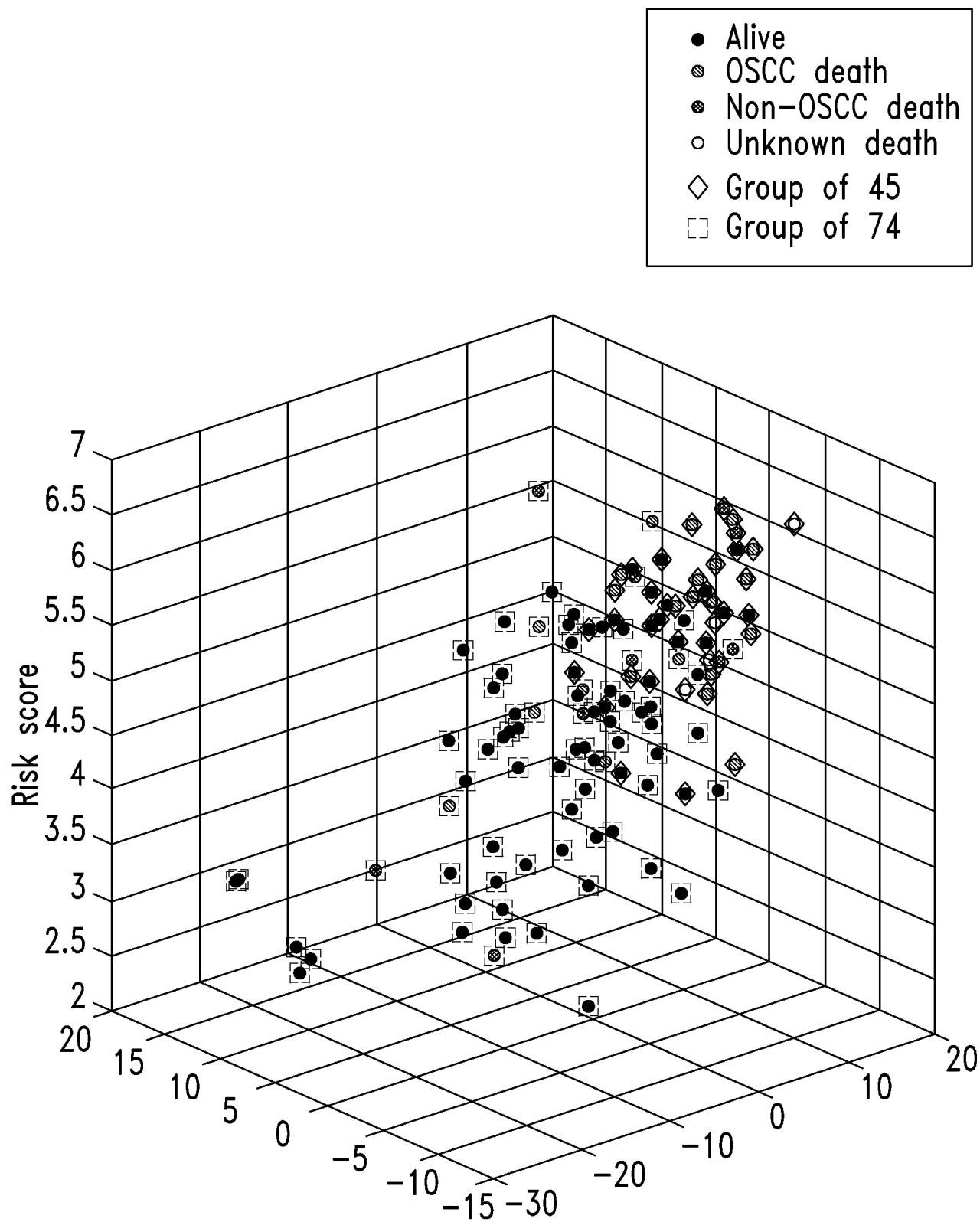


FIG. 14

A. CLASSIFICATION OF SUBJECT MATTER*C12Q 1/68(2006.01)i, C12N 15/11(2006.01)i, G01N 33/574(2006.01)i*

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC C12Q 1/68

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

NCBI PubMed, eKOMPASS (KIPO internal), "OSCC, gene expression profile, marker gene, comparing, microarray, etc."

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X ---- A	Chen et al., "Gene expression profiling identifies genes predictive of oral squamous cell carcinoma" Cancer Epidemiol. Biomarkers Prev., Vol. 17, No. 8, pp2152-2162, Epub 31 Jul 2008 - see the whole document, especially Abstract; Tables 1 and 4; Figure 3.	1-4 ---- 27-30
X ---- A	Gasparoni et al., "Prognostic value of differential expression of Laminin-5 gamma2 in oral squamous cell carcinomas: correlation with survival" Oncol. Rep., Vol. 18, No. 4, pp793-800, Oct 2007 - see the whole document, especially Abstract; Figures 1-8.	27-30 ---- 1-4
X ---- A	Franz et al., "A quantitative co-localization analysis of large unspliced tenascin-C(L) and laminin-5/gamma2-chain in basement membranes of oral squamous cell carcinoma by confocal laser scanning microscopy" Histochem. Cell Biol., Vol. 126, No. 1, pp125-131, 13 Dec 2005 - see the whole document, especially Abstract; Figures 1-3.	1-4 ---- 27-30
X ---- A	Patel et al., "Laminin-gamma2 overexpression in head-and-neck squamous cell carcinoma" Int. J. Cancer, Vol. 99, No. 4, pp583-588, 01 Jun 2002 - see the whole document, especially Abstract; Figures 4-5.	1-4 ---- 27-30 (all partial)



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

28 MAY 2010 (28.05.2010)

Date of mailing of the international search report

28 MAY 2010 (28.05.2010)

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
Government Complex-Daejeon, 139 Seonsa-ro, Seo-
gu, Daejeon 302-701, Republic of Korea

Facsimile No. 82-42-472-7140

Authorized officer

Lee, Jun Hyuk

Telephone No. 82-42-481-8115



INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2009/051743

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.b of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of :

a. a sequence listing filed or furnished

☐

on paper

☒

in electronic form

b. time of filing or furnishing

☒

contained in the international application as filed

☐

filed together with the international application in electronic form

☐

furnished subsequently to this Authority for the purposes of search

2. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that in the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2009/051743**Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)**

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☒ Claims Nos.: 6-9, 32-35, 37, 38
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

The subject matter of claims 6-9, 32-35, 37 and 38 is so unclear and indefinite that no meaningful search could be made because these claims are drafted in reference to multiple dependent claims such as claims 5, 31 and 36 (see the category Box No. II-3 below).
3. ☒ Claims Nos.: 5, 31, 36
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

See an Extra sheet.

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.
3. ☒ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
1-9, 27-38(all partial)
4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☒ No protest accompanied the payment of additional search fees.

This ISA found at least 649 inventions in this application as follows:

1) Groups I-IV

-**Group I:** Claims 1-9 and 36-38(partial) are drawn to an invention featuring a method for identifying a risk for having, or presence of, oral squamous cell carcinoma (OSCC) in a subject.

-**Group II:** Claims 10-15 and 18-19(partial) are drawn to an invention featuring a method for identifying a risk for having OSCC in a subject having oral epithelial dysplasia but no frank OSCC.

-**Group III:** Claims 16, 17, 18-19(partial), 20-26 are drawn to an invention featuring a method for identifying a risk for having, or presence of, squamous cell carcinoma (SCC) in a subject.

-**Group IV:** Claims 27-35 and 36-38(partial) are drawn to an invention featuring a method for identifying an increased risk of OSCC-specific mortality in a subject having OSCC.

The single general concept linking Groups I-IV together is merely "identifying OSCC in a subject by gene expression profiles". This common concept is, however, described in "Cancer Epidemiol. Biomarkers Prev., Vol. 17, No. 8, pp2152-2162 (Epub 31 Jul 2008)" and, therefore, does not represent any contribution over the prior art.

2) Groups "I1"- "I200", "II1-II49", "III1"- "III200" and "IV1"- "IV200"

-**Group I1-I200:** Claims 1-9 and 36-38 (all partial) are drawn to 200 inventions featuring a method for identifying a risk for having, or presence of, oral squamous cell carcinoma (OSCC) in a subject comprising determining expression profile of at least one gene selected from SEQ ID NOS:1-200.

-**Group II1-II49:** Claims 10-15 and 18-19 (all partial) are drawn to 49 inventions featuring a method for identifying a risk for having OSCC in a subject having oral epithelial dysplasia but no frank OSCC comprising determining expression profile of at least one gene selected from Figure 6 showing 49 genes.

-**Group III1-III200:** Claims 16-26(all partial) are drawn to 200 inventions featuring a method for identifying a risk for having, or presence of, squamous cell carcinoma (SCC) in a subject comprising determining expression profile of at least one gene selected from SEQ ID NOS:1-200.

-**Group IV1-IV200:** Claims 27-38(all partial) are drawn to 200 inventions featuring a method for identifying an increased risk of OSCC-specific mortality in a subject having OSCC comprising determining expression profile of at least one gene selected from SEQ ID NOS:1-200.

The single general concept linking Groups "I1"- "I200", "II1-II49", "III1"- "III200" and "IV1"- "IV200" together is merely "identifying OSCC in a subject by LAMC2 gene expression profiles". This common concept is, however, described in "Cancer Epidemiol. Biomarkers Prev., Vol. 17, No. 8, pp2152-2162 (Epub 31 Jul 2008)" and, therefore, does not represent any contribution over the prior art.

Hence the inventions do not relate to a single general inventive concept under PCT Rule 13.1 because the above mentioned common concept fails to make a contribution over the prior art within the meaning of PCT Rule 13.2.