



(51) International Patent Classification:

C07H 21/02 (2006.01) *G01N* 33/58 (2006.01)
C07H 21/04 (2006.01) *G01N* 33/48 (2006.01)
C12Q 1/68 (2006.01) *G01N* 33/574 (2006.01)
C12Q 1/70 (2006.01)

(21) International Application Number:

PCT/US2017/045898

(22) International Filing Date:

08 August 2017 (08.08.2017)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/494,527 11 August 2016 (11.08.2016) US
 62/602,529 26 April 2017 (26.04.2017) US

(71) Applicant: **THE TRUSTEES OF THE UNIVERSITY OF PENNSYLVANIA** [US/US]; 3160 Chestnut Street, Suite 200, Philadelphia, Pennsylvania 19104 (US).

(72) Inventors: **ROBERTSON, Erle S.**; 419 Shortridge Drive, Wynnewood, Pennsylvania 19096 (US). **ALWINE, James C.**; 108 Manchester Road, Media, Pennsylvania 19063 (US).

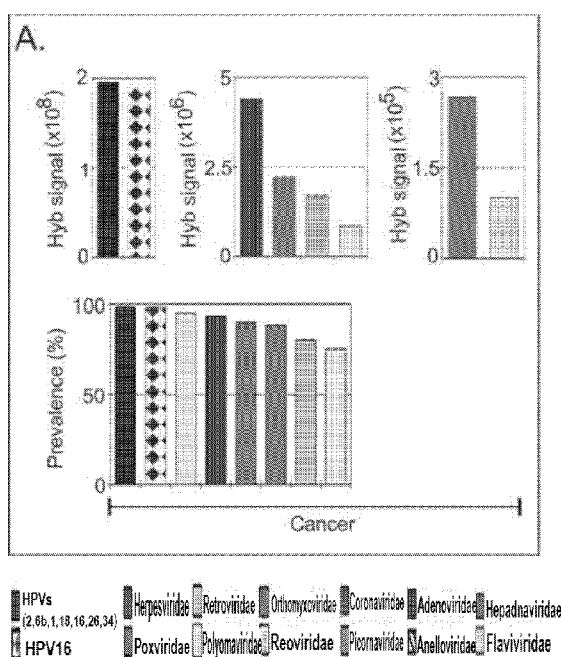
(74) Agent: **DOYLE, Kathryn** et al.; Saul Ewing LLP, Centre Square West, 1500 Market Street, 38th Floor, Philadelphia, Pennsylvania 19102 (US).

(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, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) Title: COMPOSITIONS AND METHODS FOR DETECTING ORAL SQUAMOUS CELL CARCINOMAS

(57) Abstract: The present invention includes compositions and methods for the detection of oral squamous cell carcinoma. Compositions and methods are provided for detecting a metagenomic signature in a tissue sample from a subject that indicates the subject has oral squamous cell carcinoma.



Declarations under Rule 4.17:

- *of inventorship (Rule 4.17(iv))*

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*
- *with sequence listing part of description (Rule 5.2(a))*

TITLE OF THE INVENTION
COMPOSITIONS AND METHODS FOR DETECTING ORAL
SQUAMOUS CELL CARCINOMAS

CROSS-REFERENCE TO RELATED APPLICATIONS

The present application claims priority under 35 U.S.C. § 119(e) to U.S. Provisional Patent Application No. 62/494,527, filed, August 11, 2016 and U.S. Provisional Patent Application No. 62/602,529, filed April 26, 2017, which are incorporated herein by reference in their entireties.

BACKGROUND OF THE INVENTION

Cancer remains the second most common cause of death in the US preceded by heart disease, accounting for nearly 1 of every 4 deaths. Oral cancer is one of the most common cancers worldwide, and incidence rates are higher in men compared to women. The predicted new oral cancer cases in 2016 will be 48,250 in the US, with predicted new cases annually exceeding 450,000, worldwide. Oral cancer is newly diagnosed in about 115 new individuals each day in the US alone, and 1 person dies from it every hour. Oral squamous cell carcinoma (OSCC) is the most common oral cancer, comprising about 90% of all the oral cancers. In the US, 3% of cancers in men and 2% in women are OSCC, most of which occur after age 50. The majority of cases are diagnosed at the late stage of cancer, and this accounts for the very high death rate of about 50% at five years from diagnosis. However, if diagnosed at early stages of development, the survival rate for oral cancer is relatively high at 80- 90%. A 70-80% risk factor for oral cancer has been linked to tobacco and alcohol usage and more recently about 20% to 25% to HPV16 infection. Less than 7% of oral cancers are not linked to a specific cause and can be attributed to genetic susceptibility or to infections or dysregulation of the oral microbiome.

The 5 year survival post-diagnosis of OSCC is directly related to the stage at diagnosis. Therefore, early detection efforts have the potential to increase the survival rate. Notably, during the early stage, oral cancer lesions can go unnoticed, as it is asymptomatic and painless. Thus, discovering biomarkers for oral cancer will be useful for early diagnosis and increased survival rate. However, as of today there are no efficient biomarkers for oral cancer. Studies focused on associating bacterial flora with oral cancer have suggested that some salivary bacteria may be indicators of disease, which is

potentially useful in patient diagnosis, monitoring, and overall health evaluation. However, 35% to 50% of the oral microbiome that is uncultivable could be associated with oral health or disease. Most culture independent laboratory techniques, including next generation sequencing (NGS), bacterial microarrays, DNA hybridization, PCR, and quantitative PCR, are currently used to determine the association of bacteria with oral health and disease, but not as diagnostics.

A need thus exists for methods and compositions to detect oral squamous cell carcinoma (OSCC) in a sample from a subject. The present invention satisfies this need.

SUMMARY OF THE INVENTION

As described herein, the present invention relates to compositions and methods for detecting oral squamous cell carcinoma. In one aspect, the invention includes a method of detecting oral squamous cell carcinoma in a tumor tissue sample from a subject. The method comprises hybridizing a detectably-labeled nucleic acid from the tumor tissue sample to a PathoChip array to generate a first hybridization pattern and hybridizing a detectably-labeled nucleic acid from a reference sample to a PathoChip array to generate a second hybridization pattern. The reference sample is from an otherwise identical non-tumor tissue from a subject. The first and second hybridization patterns are compared. When the first hybridization pattern is substantially a microbial hybridization signature and the second hybridization pattern is substantially not a microbial hybridization signature, oral squamous cell carcinoma is detected in the tumor tissue sample.

In another aspect, the invention includes a method of detecting oral squamous cell carcinoma in a tumor tissue sample from a subject, comprising hybridizing a detectably-labeled nucleic acid from the tumor tissue sample to a first microarray comprising at least three nucleic acid probes from microbes selected from the group consisting of Human papillomavirus 16 (HPV16), *Eshcherichia*, *Rothia*, *Peptoniphilus*, *Brevundimonas*, *Comamonas*, *Alcaligenes*, *Arcanobacterium*, *Actinomyces*, *Aeromonas*, *Bordetella*, *Aerococcus*, *Pediococcus*, *Caulobacter*, *Cardiobacterium*, *Plesiomonas*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateuria*, *Rhodotorula*, *Geotrichum*, *Pneumocystis*, *Hymenolepis*, *Centrocestus*, *Trichinella*, *Acinetobacter*, *Actinobacillus*, *Veillonella*, *Fonsecaea*, *Mobiluncus*, *Propionibacterium*, *Mycobacterium*, *Malassezia*, *Pleistophora*, *Prosthodendrium*, *Contracaecum*, *Toxocara*, Parvovirus 2, Human herpesvirus 6A, Adenovirus 1, Swine fever virus, JC polyomavirus, Moloney murine leukemia virus, Torque teno virus 2, Human papillomavirus 26, Tomato chlorotic leaf distortion virus,

and *Prevotella* to generate a first hybridization pattern and hybridizing a detectably-labeled nucleic acid from a reference sample to a second microarray comprising at least three nucleic acid probes from microbes selected from the group consisting of Human papillomavirus 16 (HPV16), *Eshcherichia*, *Rothia*, *Peptoniphilus*, *Brevundimonas*,
 5 *Comamonas*, *Alcaligenes*, *Arcanobacterium*, *Actinomyces*, *Aeromonas*, *Bordetella*, *Aerococcus*, *Pediococcus*, *Caulobacter*, *Cardiobacterium*, *Plesiomonas*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateruria*, *Rhodotorula*, *Geotrichum*, *Pneumocystis*, *Hymenolepis*, *Centrocestus*, *Trichinella*, *Acinetobacter*, *Actinobacillus*, *Veillonella*, *Fonsecaea*, *Mobiluncus*, *Propionibacterium*, *Mycobacterium*, *Malassezia*, *Pleistophora*,
 10 *Prosthodendrium*, *Contracaecum*, *Toxocara*, Parvovirus 2, Human herpesvirus 6A, Adenovirus 1, Swine fever virus, JC polyomavirus, Moloney murine leukemia virus, Torque teno virus 2, Human papillomavirus 26, Tomato chlorotic leaf distortion virus, and *Prevotella* to generate a second hybridization pattern. The reference sample is from an otherwise identical non-tumor tissue from a subject. The first and second hybridization
 15 patterns are compared. When the first hybridization pattern is substantially a microbial hybridization signature and the second hybridization pattern is substantially not a microbial hybridization signature, oral squamous cell carcinoma is detected in the tumor tissue sample.

Another aspect of the invention includes a composition comprising at least three
 20 nucleic acid probes selected from the group consisting of SEQ ID NOS: 1-76. Yet another aspect of the invention includes a microarray comprising at least three nucleic acid probes selected from the group consisting of SEQ ID NOS: 1-76.

Yet another aspect of the invention includes a microarray comprising at least three nucleic acid probes selected from the group of microbes consisting of Human
 25 papillomavirus 16 (HPV16), *Eshcherichia*, *Rothia*, *Peptoniphilus*, *Brevundimonas*, *Comamonas*, *Alcaligenes*, *Arcanobacterium*, *Actinomyces*, *Aeromonas*, *Bordetella*, *Aerococcus*, *Pediococcus*, *Caulobacter*, *Cardiobacterium*, *Plesiomonas*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateruria*, *Rhodotorula*, *Geotrichum*, *Pneumocystis*, *Hymenolepis*, *Centrocestus*, *Trichinella*, *Acinetobacter*, *Actinobacillus*, *Veillonella*,
 30 *Fonsecaea*, *Mobiluncus*, *Propionibacterium*, *Mycobacterium*, *Malassezia*, *Pleistophora*, *Prosthodendrium*, *Contracaecum*, *Toxocara*, Parvovirus 2, Human herpesvirus 6A, Adenovirus 1, Swine fever virus, JC polyomavirus, Moloney murine leukemia virus, Torque teno virus 2, Human papillomavirus 26, Tomato chlorotic leaf distortion virus, and *Prevotella*.

Another aspect of the invention includes a kit comprising at least three nucleic acid probes selected from the group consisting of SEQ ID NOS: 1-76, and instructional material for use thereof. Yet another aspect of the invention includes a kit comprising a microarray comprising at least three nucleic acid probes selected from the group consisting of SEQ ID NOS: 1-76, and instructional material for use thereof. Still another aspect of the invention includes a kit comprising a microarray comprising at least three nucleic acid probes selected from the group of microbes consisting of Human papillomavirus 16 (HPV16), *Eshcherichia*, *Rothia*, *Peptoniphilus*, *Brevundimonas*, *Comamonas*, *Alcaligenes*, *Arcanobacterium*, *Actinomyces*, *Aeromonas*, *Bordetella*, *Aerococcus*, *Pediococcus*, *Caulobacter*, *Cardiobacterium*, *Plesiomonas*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateuria*, *Rhodotorula*, *Geotrichum*, *Pneumocystis*, *Hymenolepis*, *Centrocestus*, *Trichinella*, *Acinetobacter*, *Actinobacillus*, *Veillonella*, *Fonsecaea*, *Mobiluncus*, *Propionibacterium*, *Mycobacterium*, *Malassezia*, *Pleistophora*, *Prosthodendrium*, *Contracaecum*, *Toxocara*, Parvovirus 2, Human herpesvirus 6A, Adenovirus 1, Swine fever virus, JC polyomavirus, Moloney murine leukemia virus, Torque teno virus 2, Human papillomavirus 26, Tomato chlorotic leaf distortion virus, and *Prevotella*, and instructional material for use thereof.

In various embodiments of the above aspects or any other aspect of the invention delineated herein, the microbial hybridization signature is generated by hybridization of the detectably-labeled nucleic acid from the tumor tissue sample to at least three nucleic acid probes on the PathoChip, wherein the probes are from microbes selected from the group consisting of Human papillomavirus 16 (HPV16), *Eshcherichia*, *Rothia*, *Peptoniphilus*, *Brevundimonas*, *Comamonas*, *Alcaligenes*, *Arcanobacterium*, *Actinomyces*, *Aeromonas*, *Bordetella*, *Aerococcus*, *Pediococcus*, *Caulobacter*, *Cardiobacterium*, *Plesiomonas*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateuria*, *Rhodotorula*, *Geotrichum*, *Pneumocystis*, *Hymenolepis*, *Centrocestus*, *Trichinella*, *Acinetobacter*, *Actinobacillus*, *Veillonella*, *Fonsecaea*, *Mobiluncus*, *Propionibacterium*, *Mycobacterium*, *Malassezia*, *Pleistophora*, *Prosthodendrium*, *Contracaecum*, *Toxocara*, Parvovirus 2, Human herpesvirus 6A, Adenovirus 1, Swine fever virus, JC polyomavirus, Moloney murine leukemia virus, Torque teno virus 2, Human papillomavirus 26, Tomato chlorotic leaf distortion virus, and *Prevotella*. In another embodiment, the at least three nucleic acid probes are selected from the group consisting of SEQ ID NOS: 1-76.

In another embodiment, the tumor tissue sample is selected from the group consisting of a biopsy, formalin-fixed, paraffin-embedded (FFPE) sample, or non-solid

tumor. In yet another embodiment, the subject is human. In still another embodiment, when oral squamous cell carcinoma is detected in the tumor tissue sample from a subject, the subject is provided with a treatment for oral squamous cell carcinoma. In another embodiment, the treatment comprises surgery, chemotherapy, or radiotherapy.

5 In yet another embodiment, the detectably-labeled nucleic acid is labeled with a fluorophore, radioactive phosphate, biotin, or enzyme. In yet another embodiment, the fluorophore is Cy3 or Cy5.

In still another embodiment, the nucleic acid probes are selected from about 10 to about 30 microbes and comprise about 3 to about 5 probes per microbe.

10 In another embodiment, the microarray is a biochip, glass slide, bead, or paper.

BRIEF DESCRIPTION OF THE DRAWINGS

The following detailed description of specific embodiments of the invention will be better understood when read in conjunction with the appended drawings. For the purpose of illustrating the invention, there are shown in the drawings exemplary
15 embodiments. It should be understood, however, that the invention is not limited to the precise arrangements and instrumentalities of the embodiments shown in the drawings.

FIGs. 1A-1F illustrate viral signatures detected in oral cancer and control samples. FIG. 1A shows the viral signatures that are detected with hybridization signal ($g-r>30$) by
20 PathoChip screen of 100 oral cancer samples ranked according to decreasing hybridization signal (weighted score sum of all the probes per accession) and prevalence. FIGs. 1B-1C show the hybridization signals and prevalence for the viral signatures detected in matched (MC) and non-matched (NC) controls respectively, ranked in descending order. FIG. 1D shows the association of different molecular signatures of viral
25 families with cancer and controls, represented as a Venn diagram, and a bar graph. FIG. 1E shows a heat map of hybridization signals detected by PathoChip screen of the HPV probes (Y-axis) with the oral cancer and control samples (x-axis). The hybridization signals of the cancer samples to each of these probes were compared to MCs and NCs. Samples were screened individually or in pools (marked with a •). FIG. 1F shows the
30 percentage of HPV16 probes detected with low ($g-r>30-300$), medium ($g-r>300-3000$) and high ($g-r>3000$) hybridization signal in 100 oral cancer samples screened individually and in pools (•) and 20 each of MCs and NCs screened in pools of 5.

FIGs. 2A-2F illustrate bacterial signatures detected in oral cancer samples. FIG. 2A is a series of pie charts showing the percentage of different groups and phyla of

bacteria detected in oral cancer, matched (MC) and non-matched controls (NC). FIGs. 2B-2D show the bacterial signatures that are detected with hybridization signal ($g-r>30$) by PathoChip screen of 100 oral cancer samples and in MCs and NCs ranked according to decreasing hybridization signal (weighted score sum of all the probes per accession) and prevalence. FIG. 2E shows a heat map of the hybridization signal for the bacterial probes of bacterial genera a-xyz, labeled in FIGs. 2B-2D, detected by PathoChip screen with the cancer, matched (MC) and non-matched control (NC) samples. Samples were screened individually and in pools (marked ■). FIG. 2F shows the association of molecular signatures of different bacterial genera with oral cancer and/or controls, represented as a Venn diagram as well as a bar graph.

FIGs. 3A-3J illustrate fungal (FIGs. 3A-3E) and parasitic (FIGs. 3F-3J) signatures detected in oral cancer samples. FIG. 3A shows the fungal signatures that are detected with hybridization signal ($g-r>30$) by PathoChip screen of 100 oral cancer samples ranked according to decreasing hybridization signal (weighted score sum of all the probes per accession) and prevalence. FIGs. 3B-3C show the fungal signatures detected in the matched (MC) and non-matched controls (NC) respectively, ranked according to decreasing hybridization signal and prevalence. FIG. 3D shows a heat map of the hybridization signal for the fungal probes of fungi a-g, labeled in FIG. 3A, detected by PathoChip screen with the cancer, matched (MC) and non-matched control (NC) samples. Samples were screened individually and in pools (marked ■). FIG. 3E shows the association of molecular signatures of different fungal genera with oral cancer and/or controls, represented as a Venn diagram and bar graph. FIG. 3F shows the parasitic signatures that are detected with hybridization signal ($g-r>30$) by PathoChip screen of 100 oral cancer samples ranked according to decreasing hybridization signal (weighted score sum of all the probes per accession) and prevalence. FIGs. 3G-3H show the parasitic signatures detected in the matched and non-matched controls (MC and NC) respectively, ranked according to decreasing hybridization signals and prevalence. FIG. 3I shows a heat map of the hybridization signal for the parasitic probes of parasites a-f, labeled in FIG. 3F, detected by PathoChip screen with the cancer, matched (MC) and non-matched control (NC) samples. Samples were screened individually and in pools (marked ■). FIG. 3J shows the association of molecular signatures of different parasitic genera with oral cancer and/or controls, represented as a Venn diagram and a bar graph.

FIGs. 4A-4C illustrate hierarachial clustering of 100 oral cancer samples. FIG. 4A shows hierarchial clustering by R program using Euclidean distance, complete linkage

and non-adjusted values. Samples marked (▪) were the samples that were screened in pools, the rest were screened individually. FIG. 4B shows clustering of the OSCC samples using NBClust software [CH (Calinski and Harabasz) index, Euclidean distance, complete linkage]. FIG. 4C shows topological analysis using Ayasdi software, using
 5 Euclidean (L2) metric and L-infinity centrality lenses. The OSCC samples that had similar detection for viral and microbial signatures formed the nodes, and those nodes are connected by an edge if the corresponding nodes have detection pattern in common with the first node. Nodes are color coded according to the detection of HPV 16.

FIGs. 5A-5D illustrate probe capture sequencing alignment for individual capture
 10 pools (HPV16, O, B, F and P). HPV16 capture probes were comprised of a set of HPV16 specific probes, O capture probes consisted of certain viral and bacterial probes, B pool was comprised of bacterial probes, F consisted of fungal probes and P was comprised of parasitic probes that are mentioned in Table 1. The hybridization signals of the HPV probes used for capture are shown as a heat map in FIG. 5A. Six pools of whole genome
 15 amplified DNA plus cDNA was hybridized to a set of biotinylated conserved and specific viral probes, then captured on streptavidin beads, and used for tagmentation library preparation and deep sequencing with paired –end 250-nt reads. FIGs. 5A-5D show the Miseq reads from individual capture when aligned with the metagenome of PathoChip (Chip probes), which cluster mostly at the capture probe regions. The genomic location
 20 along with the number of MiSeq reads are shown in the figure for each organism.

FIGs. 6A-6F illustrate microbial genomic integrations in the host chromosome. FIG. 6A is a series of bar graphs showing the number of viral (HPV16 and JC Polyoma viral) integration sites in host human chromosomes and the percentage of viral genomic sites for integration into host chromosomes. FIG. 6B is a circos plot highlighting fusion
 25 events with ≥ 20 reads support for the bacterial, fungal and parasitic insertions into individual human chromosomes. For the viral insertions, all the reads were taken into account. FIG. 6C is a karyogram plot of bacterial insertion sites (lines above chromosomes) in human chromosomes, cut off reads ≥ 20 . The number of insertion sites in each chromosome is mentioned in the figure before chromosome number. FIG. 6D is a
 30 karyogram plot of virus, parasite, and fungal, insertion sites in human chromosomes. The cutoff read for bacteria, fungus and parasite, ≥ 20 and for virus, all the insertion sites were included. The number of insertion sites in each chromosome is mentioned in the figure before the chromosome number. G-banding annotation for each chromosome is shown; gneg - Giemsa negative bands; The Giemsa positive bands have further been

subdivided into gpos25, gpos50, gpos75, and gpos100 with the higher number indicating a darker stain; acen - centromeric regions; gvar - variable length heterochromatic regions; stalk - tightly constricted regions on the short arms of the acrocentric chromosomes FIG. 6E is a schematic representation of viral and microbial genomic insertional sites in human chromosome 17. The genomic co-ordinates of the pathogens integrated and that of the host chromosome integration sites are mentioned. The co-ordinates for human chromosomes are from GRCh37/hf19 Assembly. FIG. 6F shows the association of host genes affected by viral/microbial genomic integrations to neoplasia of epithelial cells, analyzed by the Ingenuity Pathway Analysis (IPA) program that showed a p -value of $7.17E^{10}$ for such association.

FIGs. 7A-7G illustrate capture probe sequencing results. FIG. 7A is a series of histograms showing the percentage of reads >0 in each capture pool reaction (B for bacterial, F for fungal, HPV for papillomaviral, O for others, that included certain viral and bacterial capture probes and P for parasitic capture probes). The number of reads captured by individual capture pools co-related with the type of capture probes used. For example, 100 percent of the reads from B library aligned with the bacterial (B) capture probes. FIGs. 7B-7G show probe capture sequencing alignments post MiSeq. The MiSeq reads from individual capture when aligned with the metagenome of PathoChip (Chip probes) was found to cluster mostly at the capture probe regions. The genomic location along with the number of MiSeq reads are shown in the figures. The alignment for O capture pool library is shown in FIG. 7B; that of B library is shown Figure FIG. 7C-7D; that of F library is shown FIG. 7E; that of P library is shown in FIG. 7F-7G.

FIG. 8 is a schematic representation of viral, bacterial, fungal and parasitic genomic insertional sites in human chromosomes. The genomic co-ordinates of the pathogens integrated and that of the host chromosome integration sites are mentioned. The co-ordinates for human chromosomes are from GRCh37/hf19 Assembly.

FIGs. 9A-9B are a set of tables illustrating microbial genomic integration sites in the OCSCE host somatic chromosomes.

DETAILED DESCRIPTION OF THE INVENTION

Definitions

Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which the invention pertains. Although any methods and materials similar or equivalent to those

described herein can be used in the practice for testing of the present invention, exemplary materials and methods are described herein. In describing and claiming the present invention, the following terminology will be used.

5 It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to be limiting.

The articles “a”, “an”, and “the” are used herein to refer to one or to more than one (*i.e.*, to at least one) of the grammatical object of the article. By way of example, “an element” means one element or more than one element.

10 “About” as used herein when referring to a measurable value such as an amount, a temporal duration, and the like, is meant to encompass variations of $\pm 20\%$ or $\pm 10\%$, more preferably $\pm 5\%$, even more preferably $\pm 1\%$, and still more preferably $\pm 0.1\%$ from the specified value, as such variations are appropriate to perform the disclosed methods.

15 A “biomarker” or “marker” as used herein generally refers to a nucleic acid molecule, clinical indicator, protein, or other analyte that is associated with a disease. In certain embodiments, a nucleic acid biomarker is indicative of the presence in a sample of a pathogenic organism, including but not limited to, viruses, viroids, bacteria, fungi, helminths, and protozoa. In various embodiments, a marker is differentially present in a biological sample obtained from a subject having or at risk of developing a disease (e.g., an infectious disease) relative to a reference. A marker is differentially present if the
20 mean or median level of the biomarker present in the sample is statistically different from the level present in a reference. A reference level may be, for example, the level present in an environmental sample obtained from a clean or uncontaminated source. A reference level may be, for example, the level present in a sample obtained from a healthy control subject or the level obtained from the subject at an earlier timepoint, *i.e.*, prior to
25 treatment. Common tests for statistical significance include, among others, t-test, ANOVA, Kruskal-Wallis, Wilcoxon, Mann-Whitney and odds ratio. Biomarkers, alone or in combination, provide measures of relative likelihood that a subject belongs to a phenotypic status of interest. The differential presence of a marker of the invention in a subject sample can be useful in characterizing the subject as having or at risk of
30 developing a disease (e.g., an infectious disease), for determining the prognosis of the subject, for evaluating therapeutic efficacy, or for selecting a treatment regimen.

By “agent” is meant any nucleic acid molecule, small molecule chemical compound, antibody, or polypeptide, or fragments thereof.

By "alteration" or "change" is meant an increase or decrease. An alteration may be by as little as 1%, 2%, 3%, 4%, 5%, 10%, 20%, 30%, or by 40%, 50%, 60%, or even by as much as 70%, 75%, 80%, 90%, or 100%.

5 By "biologic sample" is meant any tissue, cell, fluid, or other material derived from an organism.

By "capture reagent" is meant a reagent that specifically binds a nucleic acid molecule or polypeptide to select or isolate the nucleic acid molecule or polypeptide.

10 As used herein, the terms "determining", "assessing", "assaying", "measuring" and "detecting" refer to both quantitative and qualitative determinations, and as such, the term "determining" is used interchangeably herein with "assaying," "measuring," and the like. Where a quantitative determination is intended, the phrase "determining an amount" of an analyte and the like is used. Where a qualitative and/or quantitative determination is intended, the phrase "determining a level" of an analyte or "detecting" an analyte is used.

15 By "detectable moiety" is meant a composition that when linked to a molecule of interest renders the latter detectable, via spectroscopic, photochemical, biochemical, immunochemical, or chemical means. For example, useful labels include radioactive isotopes, magnetic beads, metallic beads, colloidal particles, fluorescent dyes, electron-dense reagents, enzymes (for example, as commonly used in an ELISA), biotin, digoxigenin, or haptens.

20 A "disease" is a state of health of an animal wherein the animal cannot maintain homeostasis, and wherein if the disease is not ameliorated then the animal's health continues to deteriorate. In contrast, a "disorder" in an animal is a state of health in which the animal is able to maintain homeostasis, but in which the animal's state of health is less favorable than it would be in the absence of the disorder. Left untreated, a disorder
25 does not necessarily cause a further decrease in the animal's state of health.

"Effective amount" or "therapeutically effective amount" are used interchangeably herein, and refer to an amount of a compound, formulation, material, or composition, as described herein effective to achieve a particular biological result or provides a therapeutic or prophylactic benefit. Such results may include, but are not
30 limited to, anti-tumor activity as determined by any means suitable in the art.

"Encoding" refers to the inherent property of specific sequences of nucleotides in a polynucleotide, such as a gene, a cDNA, or an mRNA, to serve as templates for synthesis of other polymers and macromolecules in biological processes having either a defined sequence of nucleotides (*i.e.*, rRNA, tRNA and mRNA) or a defined sequence of

amino acids and the biological properties resulting therefrom. Thus, a gene encodes a protein if transcription and translation of mRNA corresponding to that gene produces the protein in a cell or other biological system. Both the coding strand, the nucleotide sequence of which is identical to the mRNA sequence and is usually provided in sequence listings, and the non-coding strand, used as the template for transcription of a gene or cDNA, can be referred to as encoding the protein or other product of that gene or cDNA.

By "fragment" is meant a portion of a nucleic acid molecule. This portion contains, preferably, at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, or 90% of the entire length of the reference nucleic acid molecule or polypeptide. A fragment may contain 5, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, or 100 nucleotides.

"Homologous" as used herein, refers to the subunit sequence identity between two polymeric molecules, *e.g.*, between two nucleic acid molecules, such as, two DNA molecules or two RNA molecules, or between two polypeptide molecules. When a subunit position in both of the two molecules is occupied by the same monomeric subunit; *e.g.*, if a position in each of two DNA molecules is occupied by adenine, then they are homologous at that position. The homology between two sequences is a direct function of the number of matching or homologous positions; *e.g.*, if half (*e.g.*, five positions in a polymer ten subunits in length) of the positions in two sequences are homologous, the two sequences are 50% homologous; if 90% of the positions (*e.g.*, 9 of 10), are matched or homologous, the two sequences are 90% homologous.

"Hybridization" means hydrogen bonding, which may be Watson-Crick, Hoogsteen or reversed Hoogsteen hydrogen bonding, between complementary nucleobases. For example, adenine and thymine are complementary nucleotides that pair through the formation of hydrogen bonds.

"Identity" as used herein refers to the subunit sequence identity between two polymeric molecules particularly between two amino acid molecules, such as, between two polypeptide molecules. When two amino acid sequences have the same residues at the same positions; *e.g.*, if a position in each of two polypeptide molecules is occupied by an Arginine, then they are identical at that position. The identity or extent to which two amino acid sequences have the same residues at the same positions in an alignment is often expressed as a percentage. The identity between two amino acid sequences is a direct function of the number of matching or identical positions; *e.g.*, if half (*e.g.*, five positions in a polymer ten amino acids in length) of the positions in two sequences are

identical, the two sequences are 50% identical; if 90% of the positions (*e.g.*, 9 of 10), are matched or identical, the two amino acids sequences are 90% identical.

As used herein, an "instructional material" includes a publication, a recording, a diagram, or any other medium of expression which can be used to communicate the usefulness of the compositions and methods of the invention. The instructional material of the kit of the invention may, for example, be affixed to a container which contains the nucleic acid, peptide, and/or composition of the invention or be shipped together with a container which contains the nucleic acid, peptide, and/or composition. Alternatively, the instructional material may be shipped separately from the container with the intention that the instructional material and the compound be used cooperatively by the recipient.

The terms "isolated," "purified," or "biologically pure" refer to material that is free to varying degrees from components which normally accompany it as found in its native state. "Isolate" denotes a degree of separation from original source or surroundings. "Purify" denotes a degree of separation that is higher than isolation. A "purified" or "biologically pure" protein is sufficiently free of other materials such that any impurities do not materially affect the biological properties of the protein or cause other adverse consequences. That is, a nucleic acid or peptide of this invention is purified if it is substantially free of cellular material, viral material, or culture medium when produced by recombinant DNA techniques, or chemical precursors or other chemicals when chemically synthesized. Purity and homogeneity are typically determined using analytical chemistry techniques, for example, polyacrylamide gel electrophoresis or high performance liquid chromatography. The term "purified" can denote that a nucleic acid or protein gives rise to essentially one band in an electrophoretic gel. For a protein that can be subjected to modifications, for example, phosphorylation or glycosylation, different modifications may give rise to different isolated proteins, which can be separately purified.

By "marker profile" is meant a characterization of the signal, level, expression or expression level of two or more markers (*e.g.*, polynucleotides).

By the term "microbe" is meant any and all organisms classed within the commonly used term "microbiology," including but not limited to, bacteria, viruses, fungi and parasites.

By the term "microarray" is meant a collection of nucleic acid probes immobilized on a substrate. As used herein, the term "nucleic acid" refers to deoxyribonucleotides, ribonucleotides, or modified nucleotides, and polymers thereof in single- or double-

stranded form. The term encompasses nucleic acids containing known nucleotide analogs or modified backbone residues or linkages, which are synthetic, naturally occurring, and non-naturally occurring. Nucleic acid molecules useful in the methods of the invention include any nucleic acid molecule that specifically binds a target nucleic acid (e.g., a nucleic acid biomarker). Such nucleic acid molecules need not be 100% identical with an endogenous nucleic acid sequence, but will typically exhibit substantial identity.

Polynucleotides having "substantial identity" to an endogenous sequence are typically capable of hybridizing with at least one strand of a double-stranded nucleic acid molecule. By "hybridize" is meant pair to form a double-stranded molecule between complementary polynucleotide sequences (e.g., a gene described herein), or portions thereof, under various conditions of stringency. (See, e.g., Wahl, G. M. and S. L. Berger (1987) *Methods Enzymol.* 152:399; Kimmel, A. R. (1987) *Methods Enzymol.* 152:507).

By the term "modulating," as used herein, is meant mediating a detectable increase or decrease in the level of a response in a subject compared with the level of a response in the subject in the absence of a treatment or compound, and/or compared with the level of a response in an otherwise identical but untreated subject. The term encompasses perturbing and/or affecting a native signal or response thereby mediating a beneficial therapeutic response in a subject, preferably, a human.

In the context of the present invention, the following abbreviations for the commonly occurring nucleic acid bases are used. "A" refers to adenosine, "C" refers to cytosine, "G" refers to guanosine, "T" refers to thymidine, and "U" refers to uridine.

"Parenteral" administration of an immunogenic composition includes, e.g., subcutaneous (s.c.), intravenous (i.v.), intramuscular (i.m.), or intrasternal injection, or infusion techniques.

As used herein, the terms "peptide," "polypeptide," and "protein" are used interchangeably, and refer to a compound comprised of amino acid residues covalently linked by peptide bonds. A protein or peptide must contain at least two amino acids, and no limitation is placed on the maximum number of amino acids that can comprise a protein's or peptide's sequence. Polypeptides include any peptide or protein comprising two or more amino acids joined to each other by peptide bonds. As used herein, the term refers to both short chains, which also commonly are referred to in the art as peptides, oligopeptides and oligomers, for example, and to longer chains, which generally are referred to in the art as proteins, of which there are many types. "Polypeptides" include, for example, biologically active fragments, substantially homologous polypeptides,

oligopeptides, homodimers, heterodimers, variants of polypeptides, modified polypeptides, derivatives, analogs, fusion proteins, among others. The polypeptides include natural peptides, recombinant peptides, synthetic peptides, or a combination thereof.

5 By "reference" is meant a standard of comparison. As is apparent to one skilled in the art, an appropriate reference is where an element is changed in order to determine the effect of the element. In one embodiment, the level of a target nucleic acid molecule present in a sample may be compared to the level of the target nucleic acid molecule present in a clean or uncontaminated sample. For example, the level of a target nucleic acid molecule present in a sample may be compared to the level of the target nucleic acid molecule present in a corresponding healthy cell or tissue or in a diseased cell or tissue (e.g., a cell or tissue derived from a subject having a disease, disorder, or condition).

As used herein, the term "sample" includes a biologic sample such as any tissue, cell, fluid, or other material derived from an organism.

15 By "specifically binds" is meant a compound (e.g., nucleic acid probe or primer) that recognizes and binds a molecule (e.g., a nucleic acid biomarker), but which does not substantially recognize and bind other molecules in a sample, for example, a biological sample.

By "substantially identical" is meant a polypeptide or nucleic acid molecule exhibiting at least 50% identity to a reference amino acid sequence (for example, any one of the amino acid sequences described herein) or nucleic acid sequence (for example, any one of the nucleic acid sequences described herein). Preferably, such a sequence is at least 60%, more preferably 80% or 85%, and more preferably 90%, 95%, 96%, 97%, 98%, or even 99% or more identical at the amino acid level or nucleic acid to the sequence used for comparison.

25 Sequence identity is typically measured using sequence analysis software (for example, Sequence Analysis Software Package of the Genetics Computer Group, University of Wisconsin Biotechnology Center, 1710 University Avenue, Madison, Wis. 53705, BLAST, BESTFIT, GAP, or PILEUP/PRETTYBOX programs). Such software matches identical or similar sequences by assigning degrees of homology to various substitutions, deletions, and/or other modifications. Conservative substitutions typically include substitutions within the following groups: glycine, alanine; valine, isoleucine, leucine; aspartic acid, glutamic acid, asparagine, glutamine; serine, threonine; lysine, arginine; and phenylalanine, tyrosine. In an exemplary approach to determining the

degree of identity, a BLAST program may be used, with a probability score between e^{-3} and e^{-100} indicating a closely related sequence.

By the term “substantially microbial hybridization signature” is a relative term and means a hybridization signature that indicates the presence of more microbes in a tumor sample than in a reference sample. By the term “substantially not a microbial hybridization signature” is a relative term and means a hybridization signature that indicates the presence of less microbes in a reference sample than in a tumor sample.

By “subject” is meant a mammal, including, but not limited to, a human or non-human mammal, such as a bovine, equine, canine, ovine, feline, mouse, or monkey. The term “subject” may refer to an animal, which is the object of treatment, observation, or experiment (e.g., a patient).

By “target nucleic acid molecule” is meant a polynucleotide to be analyzed. Such polynucleotide may be a sense or antisense strand of the target sequence. The term “target nucleic acid molecule” also refers to amplicons of the original target sequence. In various embodiments, the target nucleic acid molecule is one or more nucleic acid biomarkers.

A “target site” or “target sequence” refers to a genomic nucleic acid sequence that defines a portion of a nucleic acid to which a binding molecule may specifically bind under conditions sufficient for binding to occur.

The term “therapeutic” as used herein means a treatment and/or prophylaxis. A therapeutic effect is obtained by suppression, remission, or eradication of a disease state.

As used herein, the terms “treat,” “treating,” “treatment,” and the like refer to reducing or ameliorating a disorder and/or symptoms associated therewith. It will be appreciated that, although not precluded, treating a disorder or condition does not require that the disorder, condition or symptoms associated therewith be completely eliminated.

By the term “tumor tissue sample” is meant any sample from a tumor in a subject including any solid and non-solid tumor in the subject.

Ranges: throughout this disclosure, various aspects of the invention can be presented in a range format. It should be understood that the description in range format is merely for convenience and brevity and should not be construed as an inflexible limitation on the scope of the invention. Accordingly, the description of a range should be considered to have specifically disclosed all the possible subranges as well as individual numerical values within that range. For example, description of a range such as from 1 to 6 should be considered to have specifically disclosed subranges such as from 1 to 3, from 1 to 4, from 1 to 5, from 2 to 4, from 2 to 6, from 3 to 6 etc., as well as

individual numbers within that range, for example, 1, 2, 2.7, 3, 4, 5, 5.3, and 6. This applies regardless of the breadth of the range.

Description

5 The present invention features compositions and methods for the detection or diagnosis of oral squamous cell carcinoma (OSCC) in a tissue sample from a subject. Oral squamous cell carcinoma (OSCC) is meant to include, but is not limited to, oropharyngeal squamous cell carcinoma (OPSCC) and oral cavity squamous cell carcinoma (OCSCC). Metagenomic signatures comprising detecting genetic material
10 from a number of viral, bacterial, fungal, and parasitic microbes were identified that indicate that a subject has oral squamous cell carcinoma.

 The microbiome is fundamentally one of the most critical organs in the human body. Dysbiosis can result in critical inflammatory responses and can lead to neoplastic events. The dysbiotic oral microbiome and the pathobionts associated with cancers can
15 provide clues as to the major contributors to oral squamous cell carcinomas (OSCCs). As disclosed herein, a pan-pathogen array technology (PathoChip) coupled with next-generation sequencing was used to establish the microbial signatures in human OSCCs. Signatures for DNA and RNA viruses including oncogenic viruses, gram positive and negative bacteria, fungi and parasites were detected. Cluster and topological analyses
20 identified 2 distinct groups of microbial signatures related to OSCCs. A comprehensive map of integration sites and chromosomal hotspots for microorganism insertions was generated herein. Without wishing to be bound by specific theory, identification of these microbial signatures and their integration sites provided novel insights into their contribution to OSCC and to potential targeted therapies. Demonstrated herein is a pan-
25 pathogen array technology called PathoChip, along with a capture next-generation sequencing strategy, to identify the microbial signatures associated with OSCCs.

Oral Squamous Cell Carcinomas (OSCCs)

 About 20% of all cancers are associated with different infectious agents. Changes
30 in the human microbiome composition, from commensals to pathogenic, are often associated with human diseases including cancers. Microorganisms have been proposed to cause cancer either by inducing chronic inflammation, interfering with host cell cycle or signaling pathways or by causing mutagenesis. Oral cancers, are the eleventh most prevalent cancers in the world. Recent studies have focused on the changes in the oral

microbiome associated with oral cancer, with the hope of finding the role of microbial dysbiosis in oral cancer development and progression, and also to identify the microbiome as a potential cancer biomarker. The technology disclosed herein provides a platform for these studies that associate the complex microbiome with disease states.

5

Metagenomic Signatures and OSCCs

In the present invention, a pan-pathogen array technology referred to as PathoChip was used to detect metagenomics signatures of oral squamous cell carcinomas (OSCCs) including oropharyngeal squamous cell carcinoma (OPSCC) and oral cavity squamous cell carcinoma (OCSCC). PathoChip is comprised of oligonucleotide probes that can detect all sequenced viruses as well as known pathogenic bacteria, fungi and parasites, and family-specific conserved probes, providing a means for detecting previously uncharacterized members of a family (Baldwin et al., (2014) mBio 5, e01714-01714; Banerjee et al., (2015) Sci Rep 5, 15162).

10

15

Viral and other microbial signatures in triple negative breast cancer have been previously described (Banerjee et al., (2015) Sci Rep 5, 15162)(International Application No. PCT/US2016/028394, filed April 20,2016). However, in the present invention specific viral, bacterial, fungal and parasitic microbial signatures specifically associated with tissues obtained from OSCCs were analyzed. Tissues were predominantly oropharyngeal (OPSCC) with some number of buccal and tongue based cancers (OCSCC). These are collectively referred to as OSCC herein. Specific genetic signatures of microorganisms were observed across the kingdoms associated only with OSCCs as well as those associated uniquely with the non-cancer controls. A predominant HPV16 genetic signature was detected as well as other microbial (bacterial, fungal and parasitic) signatures associated specifically with the OSCC samples. The present study demonstrates microbial biomarkers for oral squamous cell carcinoma using the PathoChip platform.

20

25

30

As described herein, molecular signatures of HPV16 were detected, with the highest hybridization signal intensity, prevalent in greater than 98% of the OSCC samples. This was a unique observation, given the fact that only a 35% prevalence of HPV16 was reported in OSCC in previous studies (Agrawal et al., 2013; Syrjanen et al., 2011). Specific probes of other HPVs like HPV2, HPV6b, HPV1, HPV18, HPV26, HPV34 were detected less commonly. Genomic signatures of Herpesviridae, Poxviridae, Retroviridae and Polyomaviridae were also detected at a significant level in OSCC

samples screened, and were dramatically underrepresented in the non-matched healthy controls. Without wishing to be bound by specific theory, it is possible that infection with these viruses can lead to transformation and OSCC. Alternatively, it is possible that some of these viral agents may have found an amiable microenvironment to co-exist in the cancer tissues. In either event, these observations are of particular significance, as there are no detailed reports of the viral association with OSCC other than HPVs and herpesviruses (Metgud et al., (2012) Oncol Rev 6, e21).

Bacteria can infect epithelial cells, colonize and cause inflammation, which can have a possible role in cancer progression. Results presented herein showed a slight decrease in the abundance of Firmicutes and Actinobacteria in oral cancer samples compared to matched controls, whereas, little or no difference in the abundance of these members was observed when comparing the cancer with the matched and non-matched controls. Importantly, an increase in the detection of the members of Proteobacteria was observed in the cancers compared to the matched and non-matched controls. In fact, of the bacteria detected only in the cancer samples (not in the controls), 11/13 belong to Proteobacteria. The actinobacteria genus *Rothia*, was detected only in the OSCC cancer samples in the present study.

Also demonstrated herein, yeasts such as *Rhodotorula*, *Geotrichum* and *Pneumocystis* were found to be significantly associated with OSCC tumor, and not with adjacent matched tissue control samples or healthy non-matched controls. *Rhodotorula* and *Pneumocystis* are well-known opportunistic pathogens, and without wishing to be by specific theory, may find the cancer microenvironment amiable for survival. This might also transform harmless commensals to pathogenic oral mucosal micro-organisms, leading to increased morbidity and mortality in cancer patients.

In the present study, *Fonsecaea* was detected in both OSCC cancer and the adjacent normal matched control tissues, but not from non-matched controls. Without wishing to be bound by any specific theory, this is could be due to the spread of the infection to the adjacent non-cancerous tissues from the tumor site. As described herein, microsporidia *Pleistophora* was detected in cancer as well as in controls, the detection being much more significant in the cancer group compared to the controls. Fungi of low pathogenicity like *Malassezia* and *Absidia*, along with dermatatious aetiologic agents of chromoblastomycosis *Phialophora* and *Cladophialophora* associated significantly with the oral cancer patients as compared to both controls. The fungi that were detected only in

the controls and not in the cancer samples were common dermatatious, low pathogenic fungi.

Some parasitic worms of the human body, or parasites acquired by ingesting raw fish and meat can also increase the risk of developing certain cancers. As described herein, molecular signatures of the intestinal parasites, *Hymenolepis*, *Centrocestus* and *Trichinella* were detected in almost all the OSCC samples screened but not in the control samples. Without wishing to be bound by any specific theory, these organisms may find the cancer microenvironment favorable for their growth. Probes of both *Hymenolepis*, *Centrocestus* showed very high signals ($g-r > 3000$) when hybridized to the whole genome amplified products of the cancer samples described herein. In the present study, molecular signatures of *Toxocara* were found in cancer as well as in both the controls, although, the hybridization signal was significantly less in the controls than cancer.

Thus, results of the present study showed an association of certain viral and microbial signatures with cancer that can be used as microbial biomarkers for oral cancer (Table 3). The microbial signatures that were associated with cancer as well as adjacent matched control tissues also should not be ignored for their potential as microbial biomarkers (Table 3), given there is a possibility of the spread of infection from cancer cells to the adjacent non-cancer cells.

Thus, by screening OSCC samples as well as matched and non-matched controls, distinct viral and microbial signature patterns were shown herein to be associated with OSCC. Among those found herein to be associated only with OSCC and not controls are: Human papillomavirus 16 (HPV16) viral signatures, bacterial signatures of mostly Proteobacterias *Eshcherichia*, *Brevundimonas*, *Comamonas*, *Alcaligenes*, *Caulobacter*, *Cardiobacterium*, *Plesiomonas*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateuria* along with Actinobacteria *Rothia* and Bacteroidetes *Peptoniphilus*, fungal signatures of *Rhodotorula*, *Geotrichum*, *Pneumocystis* and parasitic signatures of *Hymenolepis*, *Centrocestus*, *Trichinella*. Described herein is a map of microbial association that can serve as biomarkers of OSCCs.

The present invention includes a method of detecting oral squamous cell carcinoma in a tumor tissue sample from a subject. The method comprises hybridizing a detectably-labeled nucleic acid from the tumor tissue sample to a PathoChip array to generate a first hybridization pattern, then hybridizing a detectably-labeled nucleic acid from a reference sample to a PathoChip array to generate a second hybridization pattern, wherein the reference sample is from an otherwise identical non-tumor tissue from a

subject. The first and second hybridization patterns are compared, wherein when the first hybridization pattern is substantially a microbial hybridization signature and the second hybridization pattern is substantially not a microbial hybridization signature, oral squamous cell carcinoma is detected in the tumor tissue sample.

5 In another aspect of the invention the method comprises wherein the microbial hybridization signature is generated by hybridization of the detectably-labeled nucleic acid from the tumor tissue sample to at least three nucleic acid probes on the PathoChip, wherein the probes are from microbes selected from the group consisting of Human papillomavirus 16 (HPV16), *Eshcherichia*, *Rothia*, *Peptoniphilus*, *Brevundimonas*,
10 *Comamonas*, *Alcaligenes*, *Arcanobacterium*, *Actinomyces*, *Aeromonas*, *Bordetella*, *Aerococcus*, *Pediococcus*, *Caulobacter*, *Cardiobacterium*, *Plesiomonas*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateruria*, *Rhodotorula*, *Geotrichum*, *Pneumocystis*, *Hymenolepis*, *Centrocestus*, *Trichinella*, *Acinetobacter*, *Actinobacillus*, *Veillonella*, and *Fonsecaea*.

15 Another aspect of the invention includes a method wherein the first hybridization pattern is generated by hybridization of the detectably-labeled nucleic acid from the tumor tissue sample to at least three nucleic acid probes on the PathoChip, wherein the probes are selected from the group consisting of SEQ ID NOS: 1-76.

In the methods disclosed herein, the tumor tissue sample can be a biopsy,
20 formalin-fixed, paraffin-embedded (FFPE) sample, or non-solid tumor. The detectably-labeled nucleic acid can be labeled with a fluorophore, radioactive phosphate, biotin, or enzyme and the fluorophore can be Cy3 or Cy5.

The methods can also include providing the subject with a treatment for oral squamous cell carcinoma when oral squamous cell carcinoma is detected in the tumor
25 tissue sample from the subject. Examples of treatments include, but are not limited to, surgery, chemotherapy, or radiotherapy.

Target Nucleic Acid Molecules

30 Methods and compositions of the invention are useful for the identification of a target nucleic acid molecule in a biological sample to be analyzed. Target sequences are amplified from any biological sample that comprises a target nucleic acid molecule. Such samples may comprise fungi, spores, viruses, or cells (e.g., prokaryotes, eukaryotes, including human). Such samples may comprise viral, bacterial, fungal, and parasitic nucleic acid molecules. In specific embodiments, compositions and methods of the

invention detect one or more nucleic acid sequences from one or more pathogenic organisms, including viruses, viroids, bacteria, fungi, helminths, and/or protozoa.

In one embodiment, a sample is a biological sample, such as a tissue or tumor sample. The level of one or more polynucleotide biomarkers (e.g., to detect or identify
5 viruses, viroids, bacteria, fungi, helminths, and/or protozoa) is measured in the biological sample. In one embodiment, the biological sample is a tissue sample that includes a tumor cell, for example, from a biopsy or formalin-fixed, paraffin-embedded (FFPE) sample. Exemplary test samples also include body fluids (e.g. blood, serum, plasma, amniotic fluid, sputum, urine, cerebrospinal fluid, lymph, tear fluid, feces, or gastric fluid), feces,
10 tissue extracts, and culture media (e.g., a liquid in which a cell, such as a pathogen cell, has been grown). If desired, the sample is purified prior to detection using any standard method typically used for isolating a nucleic acid molecule from a biological sample. In one embodiment, a target nucleic acid of a pathogen is amplified by primer oligonucleotides to detect the presence of the nucleic acid sequence of an infectious agent
15 in the sample. Such nucleic acid sequences may derive from pathogens including fungi, bacteria, viruses and yeast.

Target nucleic acid molecules include double-stranded and single- stranded nucleic acid molecules (e.g., DNA, RNA, and other nucleobase polymers known in the art capable of hybridizing with a nucleic acid molecule described herein). RNA molecules
20 suitable for detection with a detectable oligonucleotide probe or detectable primer/template oligonucleotide of the invention include, but are not limited to, double-stranded and single-stranded RNA molecules that comprise a target sequence (e.g., messenger RNA, viral RNA, ribosomal RNA, transfer RNA, microRNA and microRNA precursors, and siRNAs or other RNAs described herein or known in the art). DNA
25 molecules suitable for detection with a detectable oligonucleotide probe or primer/template oligonucleotide of the invention include, but are not limited to, double stranded DNA (e.g., genomic DNA, plasmid DNA, mitochondrial DNA, viral DNA, and synthetic double stranded DNA). Single-stranded DNA target nucleic acid molecules include, for example, viral DNA, cDNA, and synthetic single- stranded DNA, or other
30 types of DNA known in the art. In general, a target sequence for detection is between about 30 and about 300 nucleotides in length (e.g., 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300 nucleotides). In a specific embodiment the target sequence is about 60 nucleotides in length. A target sequence for detection may also have at least

about 70, 80, 90, 95, 96, 97, 98, 99, or even 100% identity to a probe sequence. Probe sequences may be longer or shorter than the target sequence. For example, a 60-nucleotide probe may hybridize to at least about 44 nucleotides of a target sequence.

In particular embodiments, a biomarker is a biomolecule (e.g., nucleic acid molecule) that is differentially present in a biological sample. For example, a biomarker is taken from a subject of one phenotypic status (e.g., having oral squamous cell carcinoma) as compared with another phenotypic status (e.g., not having oral squamous cell carcinoma). A biomarker is differentially present between different phenotypic statuses if the mean or median expression level of the biomarker in the different groups is calculated to be statistically significant. Common tests for statistical significance include, among others, t-test, ANOVA, Kruskal-Wallis, Wilcoxon, Mann-Whitney and odds ratio. Biomarkers, alone or in combination, provide measures of relative risk that a subject belongs to one phenotypic status or another. Therefore, they are useful as markers for characterizing a disease (e.g., having triple negative breast cancer).

Target Capture Probes

Demonstrated herein, probe-capture next generation sequencing (NGS) was used to further validate PathoChip screen results. Genomic regions of all biomarkers as well as the viral and microbial signatures detected in OSCC were pulled out using the probes that were detected positive in the PathoChip screen (FIGs. 5A-5D and FIGs. 7A-7G).

In various embodiments, the sets of probes used herein are based on the construction of a metagenome and its use to select probes that identify target nucleic acid molecules associated with an infectious agent. As used herein “metagenome” refers to genetic material from more than one organism, e.g., in an environmental sample. The metagenome is used to select the sets of probes and/or to validate probe sets. In some embodiments, the metagenome comprises the sequences or genomes of about 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 1000, 1500, 2000 or more organisms. In one example, the nucleic acid sequences of thousands of organisms were linked to generate a metagenome comprising 58 chromosomes.

One non-limiting example of discrete metagenome probe selection:

A. Download individual genomes, genes and partial sequences into a local database of accessions

B. Mask low complexity sequences using bioinformatic tools. In one example, low

complexity sequences are masked using mdust (<http://doc.bioperl.org/bioperl-run/lib/Bio/Tools/Run/Mdust.html>) followed by BLASTN 2.0MP-WashU31 identification of unique regions in viral accessions.

C. BLASTN sequence comparison of each accession against all other accessions

5 D. Identify specific target regions within each accession

1. 250-300bp regions
2. No more than 50 contiguous nucleotides with 70% or greater sequence homology to any other accession or to the human genome

E. Supplement specific targets

- 10
1. Identify any accessions with zero or one target region
 2. Relax stringency parameters to no more than 30 contiguous nucleotides with 50% or greater sequence homology to any other accession, but no more than 50 contiguous nucleotides with 70% or greater sequence homology to human genome

- 15
3. Re-run target region identification on accession subset from I.E.I.

F. Identify conserved target regions

1. 70-300bp regions that have 70% or greater homology with at least one other accession
 2. Remove conserved targets with 50 or more contiguous nucleotides with
- 20
- 70% or greater sequence homology to human genome

G. Choose probes

1. Run Agilent array CGH probe selection algorithm on specific and conserved target regions
 2. Rank probes by Agilent design score
- 25
3. Select 1-3 highest ranking probes from 1-5 specific target regions in each accession
 4. Select 1-3 highest ranking probes from each conserved target region

Concatenated metagenome probe selection

30 A. Download individual genomes, genes and partial sequences into a local database of accessions

B. Compile all accessions into a single concatenated metagenome to facilitate use of genomics bioinformatics tools

1. Place 100 nonspecific nucleotides ("N") as spacers between each accession

2. Join accessions and spacers into chromosomes of 6-10 million bases
- C. Run Agilent array CGH probe selection algorithm for specificity within the metagenome
- D. Filter probes for specificity against human, mouse, and/or other mammalian genomes
- 5 E. Choose specific probes
 1. Rank probes by Agilent design score
 2. Select 10-20 highest ranking probes from each accession
 3. Require at least 100 bp separation between probes
- F. Choose conserved probes
 - 10 1. Identify conserved regions as in I.F.
 2. Select 5-10 highest ranking probes from each conserved region
 3. Require at least 100 bp separation between probes
- G. Empirical probe selection
 1. Manufacture microarrays containing all specific and conserved probes
 - 15 2. Hybridize microarrays to labeled human DNA
 3. Select 5-10 specific probes from each accession with lowest cross-hybridization signal
 4. Select 3-5 conserved probes from each conserved regions with lowest cross-hybridization signal

20

In one embodiment, the invention includes at least three nucleic acid probes selected from the group consisting of SEQ ID NOS: 1-76. In another embodiment, the invention includes a kit comprising at least three nucleic acid probes selected from the group consisting of SEQ ID NOS: 1-76, and instructional material for use thereof. The nucleic acid probes can be selected from between about 10 to about 30 microbes and comprise about 3 to about 5 probes per microbe.

25

Sample Preparation

The invention provides a means for analyzing multiple types of nucleic acids present in a sample, including DNA and RNA. In various embodiments, sample preparation involves extracting a mixture of nucleic acid molecules (e.g., DNA and RNA). In other embodiments, sample preparation involves extracting a mixture of nucleic acids from multiple organisms, cell types, infectious agents, or any combination thereof. In one embodiment, sample preparation involves the workflow below.

30

- A. Fragment genomic DNA
- B. Convert total RNA to first strand cDNA by random-primed reverse transcriptase
- C. Label genomic DNA with biotin or fluorescent dye by chemical or enzymatic incorporation
- 5 D. Label cDNA with biotin or fluorescent dye by chemical or enzymatic incorporation
- E. Label a mixture of genomic DNA and cDNA in the same chemical or enzymatic reaction
- F. Mix C + D and co-hybridize to microarray of probes
- G. Hybridize E to microarray of probes
- 10 H. Amplify targeted genomic DNA
 - 1. Use whole-genome amplification (GE GenomiPhi, Sigma WGA, NuGEN Ovation DNA) to non-specifically amplify genomic DNA
 - 2. Use amplified products as input for 4.C, or 4.E.
- I. Amplify targeted total RNA
- 15 1. Use whole-transcriptome amplification (Sigma WTA, Ambion in vitro transcription, NuGEN Ovation RNA) to non-specifically amplify total RNA
- 2. Use amplified products as input.

20 The samples are hybridized to the microarray (e.g., PathoChip), and the microarrays are washed at various stringencies. Microarrays are scanned for detection of fluorescence. Background correction and inter-array normalization algorithms are applied. Detection thresholds are applied. The results are analyzed for statistical significance.

25 **Nucleic Acid Amplification**

Target nucleic acid sequences are optionally amplified before being detected. The term “amplified” defines the process of making multiple copies of the nucleic acid from a single or lower copy number of nucleic acid sequence molecule. The amplification of nucleic acid sequences is carried out in vitro by biochemical processes known to those of skill in the art. Prior to or concurrent with identification, the viral sample may be amplified by a variety of mechanisms, some of which may employ PCR. For example, primers for PCR may be designed to amplify regions of the sequence. For RNA viruses a first reverse transcriptase step may be used to generate double stranded DNA from the single stranded RNA. See, for example, PCR Technology: Principles and Applications

30

for DNA Amplification (Ed. H.A. Erlich, Freeman Press, NY, N.Y., 1992); PCR
Protocols: A Guide to Methods and Applications (Eds. Innis, et al., Academic Press, San
Diego, Calif., 1990); Mattila et al., Nucleic Acids Res. 19, 4967 (1991); Eckert et al.,
PCR Methods and Applications 1, 17 (1991); PCR (Eds. McPherson et al., IRL Press,
5 Oxford); and US Patent Nos 4,683,202, 4,683,195, 4,800,159 4,965,188, and 5,333,675.
The sample may be amplified on the array. See, for example, US Patent No 6,300,070 and
US Ser No 09/513,300.

Other suitable amplification methods include the ligase chain reaction (LCR) (for
example, Wu and Wallace, Genomics 4, 560 (1989), Landegren et al., Science 241, 1077
10 (1988) and Barringer et al. Gene 89:117 (1990)), transcription amplification (Kwoh et al.,
Proc. Natl. Acad. Sci. USA 86, 1173 (1989) and WO88/10315), self-sustained sequence
replication (Guatelli et al., Proc. Nat. Acad. Sci. USA, 87, 1874 (1990) and
WO90/06995), selective amplification of target polynucleotide sequences (US Patent No
6,410,276), consensus sequence primed PCR (CP-PCR) (US Patent No 4,437,975),
15 arbitrarily primed PCR (AP-PCR) (US Patent Nos 5,413,909, 5,861,245) and nucleic acid
based sequence amplification (NABSA) (see, US Patent Nos 5,409,818, 5,554,517, and
6,063,603). Other amplification methods that may be used are described in, US Patent
Nos 5,242,794, 5,494,810, 4,988,617 and in US Ser No 09/854,317.

Additional methods of sample preparation and techniques for reducing the
20 complexity of a nucleic acid sample are described in Dong et al., Genome Research 11,
1418 (2001), in US Patent Nos 6,361,947, 6,391,592 and US Ser Nos 09/916,135,
09/920,491 (US Patent Application Publication 20030096235), 09/910,292 (US Patent
Application Publication 20030082543), and 10/013,598.

25 **Detection of Biomarkers**

The biomarkers of this invention can be detected by any suitable method. The
methods described herein can be used individually or in combination for a more accurate
detection of the biomarkers. Methods for conducting polynucleotide hybridization assays
have been developed in the art. Hybridization assay procedures and conditions will vary
30 depending on the application and are selected in accordance with the general binding
methods known including those referred to in: Sambrook and Russell, Molecular
Cloning: A Laboratory Manual (3rd Ed. Cold Spring Harbor, N.Y., 2001); Berger and
Kimmel Methods in Enzymology, Vol. 152, Guide to Molecular Cloning Techniques
(Academic Press, Inc., San Diego, Calif., 1987); Young and Davism, P.N.A.S, 80: 1194

(1983). Methods and apparatus for carrying out repeated and controlled hybridization reactions have been described in US Patent Nos 5,871,928, 5,874,219, 6,045,996 and 6,386,749, 6,391,623. A data analysis algorithm (E-predict) for interpreting the hybridization results from an array is publicly available (see Urisman, 2005, Genome Biol 6:R78).

In one embodiment, the hybridized nucleic acids are detected by detecting one or more labels attached to, or incorporated within, the sample nucleic acids. The labels may be attached or incorporated by any of a number of means well known to those of skill in the art. In one embodiment, the label is simultaneously incorporated during the amplification step in the preparation of the sample nucleic acids. Thus, for example, PCR with labeled primers or labeled nucleotides will provide a labeled amplification product. In another embodiment, transcription amplification, as described above, using a labeled nucleotide (e.g. fluorescein-labeled UTP and/or CTP) incorporates a label into the transcribed nucleic acids. In another embodiment PCR amplification products are fragmented and labeled by terminal deoxytransferase and labeled dNTPs. Alternatively, a label may be added directly to the original nucleic acid sample (e.g., mRNA, polyA mRNA, cDNA, etc.) or to the amplification product after the amplification is completed. Means of attaching labels to nucleic acids are well known to those of skill in the art and include, for example, nick translation or end-labeling (e.g. with a labeled RNA) by kinasing the nucleic acid and subsequent attachment (ligation) of a nucleic acid linker joining the sample nucleic acid to a label (e.g., a fluorophore). In another embodiment label is added to the end of fragments using terminal deoxytransferase.

Detectable labels suitable for use in the present invention include any composition detectable by spectroscopic, photochemical, biochemical, immunochemical, electrical, optical or chemical means. Useful labels in the present invention include, but are not limited to: biotin for staining with labeled streptavidin conjugate; anti-biotin antibodies, magnetic beads (e.g., DynabeadsTM); fluorescent dyes (e.g., Cy3, Cy5, fluorescein, texas red, rhodamine, green fluorescent protein, and the like); radiolabels (e.g., ³H, ¹²⁵I, ³⁵S, ⁴C, or ³²P); phosphorescent labels; enzymes (e.g., horse radish peroxidase, alkaline phosphatase and others commonly used in an ELISA); and colorimetric labels such as colloidal gold or colored glass or plastic (e.g., polystyrene, polypropylene, latex, etc.) beads. Patents teaching the use of such labels include US Patent Nos 3,817,837, 3,850,752, 3,939,350, 3,996,345, 4,277,437, 4,275,149 and 4,366,241.

Means of detecting such labels are well known to those of skill in the art. Thus, for example, radiolabels may be detected using photographic film or scintillation counters; fluorescent markers may be detected using a photodetector to detect emitted light. Enzymatic labels are typically detected by providing the enzyme with a substrate and detecting the reaction product produced by the action of the enzyme on the substrate, and calorimetric labels are detected by simply visualizing the colored label.

Methods and apparatus for signal detection and processing of intensity data are disclosed in, for example, US Patent Nos 5,143,854, 5,547,839, 5,578,832, 5,631,734, 5,800,992, 5,834,758; 5,856,092, 5,902,723, 5,936,324, 5,981,956, 6,025,601, 6,090,555, 6,141,096, 6,185,030, 6,201,639; 6,218,803; and 6,225,625, in US Ser Nos 10/389,194, 60/493,495 and in PCT Application PCT/US99/06097 (published as WO99/47964).

Detection by Microarray

In aspects of the invention, a sample is analyzed by means of a microarray. The nucleic acid molecules of the invention are useful as hybridizable array elements in a microarray. Microarrays generally comprise solid substrates and have a generally planar surface, to which a capture reagent (also called an adsorbent or affinity reagent) is attached. Frequently, the surface of a biochip comprises a plurality of addressable locations, each of which has the capture reagent bound there.

The array elements are organized in an ordered fashion such that each element is present at a specified location on the substrate. Useful substrate materials include membranes, composed of paper, nylon or other materials, filters, chips, glass slides, and other solid supports. The ordered arrangement of the array elements allows hybridization patterns and intensities to be interpreted as expression levels of particular genes or proteins. Methods for making nucleic acid microarrays are known to the skilled artisan and are described, for example, in U.S. Pat. No. 5,837,832, Lockhart, et al. (Nat. Biotech. 14:1675-1680, 1996), and Schena, et al. (Proc. Natl. Acad. Sci. 93:10614-10619, 1996), herein incorporated by reference. US Patent Nos 5,800,992 and 6,040,138 describe methods for making arrays of nucleic acid probes that can be used to detect the presence of a nucleic acid containing a specific nucleotide sequence. Methods of forming high-density arrays of nucleic acids, peptides and other polymer sequences with a minimal number of synthetic steps are known. The nucleic acid array can be synthesized on a solid substrate by a variety of methods, including, but not limited to, light-directed chemical coupling, and mechanically directed coupling. For additional descriptions and methods

relating to resequencing arrays see US Patent Application Ser Nos 10/658,879, 60/417,190, 09/381,480, 60/409,396, and US Patent Nos 5,861,242, 6,027,880, 5,837,832, 6,723,503.

By "hybridize" is meant pair to form a double-stranded molecule between
5 complementary polynucleotide sequences (e.g., a gene described herein), or portions thereof, under various conditions of stringency. (See, e.g., Wahl, G. M. and S. L. Berger (1987) *Methods Enzymol.* 152:399; Kimmel, A. R. (1987) *Methods Enzymol.* 152:507). For example, stringent salt concentration will ordinarily be less than about 750 mM NaCl and 75 mM trisodium citrate, preferably less than about 500 mM NaCl and 50 mM
10 trisodium citrate, and more preferably less than about 250 mM NaCl and 25 mM trisodium citrate. Low stringency hybridization can be obtained in the absence of organic solvent, *e.g.*, formamide, while high stringency hybridization can be obtained in the presence of at least about 35% formamide, and more preferably at least about 50% formamide. Stringent temperature conditions will ordinarily include temperatures of at
15 least about 30° C, more preferably of at least about 37° C, and most preferably of at least about 42° C. Varying additional parameters, such as hybridization time, the concentration of detergent, *e.g.*, sodium dodecyl sulfate (SDS), and the inclusion or exclusion of carrier DNA, are well known to those skilled in the art. Various levels of stringency are accomplished by combining these various conditions as needed. In a
20 preferred embodiment, hybridization will occur at 30° C in 750 mM NaCl, 75 mM trisodium citrate, and 1% SDS. In a more preferred embodiment, hybridization will occur at 37° C in 500 mM NaCl, 50 mM trisodium citrate, 1% SDS, 35% formamide, and 100 µg/ml denatured salmon sperm DNA (ssDNA). In a most preferred embodiment, hybridization will occur at 42° C in 250 mM NaCl, 25 mM trisodium citrate, 1% SDS,
25 50% formamide, and 200 µg/ml ssDNA. Useful variations on these conditions will be readily apparent to those skilled in the art.

For most applications, washing steps that follow hybridization will also vary in stringency. Wash stringency conditions can be defined by salt concentration and by
30 temperature. As above, wash stringency can be increased by decreasing salt concentration or by increasing temperature. For example, stringent salt concentration for the wash steps will preferably be less than about 30 mM NaCl and 3 mM trisodium citrate, and most preferably less than about 15 mM NaCl and 1.5 mM trisodium citrate. Stringent temperature conditions for the wash steps will ordinarily include a temperature of at least about 25° C, more preferably of at least about 42° C, and even more preferably of at least

about 68° C In a preferred embodiment, wash steps will occur at 25° C in 30 mM NaCl, 3 mM trisodium citrate, and 0.1% SDS. In a more preferred embodiment, wash steps will occur at 42° C in 15 mM NaCl, 1.5 mM trisodium citrate, and 0.1% SDS. In a more preferred embodiment, wash steps will occur at 68° C in 15 mM NaCl, 1.5 mM trisodium citrate, and 0.1% SDS. Additional variations on these conditions will be readily apparent to those skilled in the art. Hybridization techniques are well known to those skilled in the art and are described, for example, in Benton and Davis (Science 196:180, 1977); Grunstein and Hogness (Proc. Natl. Acad. Sci., USA 72:3961, 1975); Ausubel et al. (Current Protocols in Molecular Biology, Wiley Interscience, New York, 2001); Berger and Kimmel (Guide to Molecular Cloning Techniques, 1987, Academic Press, New York); and Sambrook et al., Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, New York.

One embodiment of the invention includes a microarray comprising at least three nucleic acid probes selected from the group consisting of SEQ ID NOS: 1-76. The nucleic acid probes can be selected from about 10 to about 30 microbes and comprise about 3 to about 5 probes per microbe. In another embodiment, the microarray comprises at least three nucleic acid probes selected from the group of microbes consisting of Human papillomavirus 16 (HPV16), *Eshcherichia*, *Rothia*, *Peptoniphilus*, *Brevundimonas*, *Comamonas*, *Alcaligenes*, *Arcanobacterium*, *Actinomyces*, *Aeromonas*, *Bordetella*, *Aerococcus*, *Pediococcus*, *Caulobacter*, *Cardiobacterium*, *Plesiomonas*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateruria*, *Rhodotorula*, *Geotrichum*, *Pneumocystis*, *Hymenolepis*, *Centrocestus*, *Trichinella*, *Acinetobacter*, *Actinobacillus*, *Veillonella*, and *Fonsecaea*. The microarray can be a biochip, or on a glass slide, bead, or paper.

Detection by Nucleic Acid Biochip

In aspects of the invention, a sample is analyzed by means of a nucleic acid biochip (also known as a nucleic acid microarray). To produce a nucleic acid biochip, oligonucleotides may be synthesized or bound to the surface of a substrate using a chemical coupling procedure and an ink jet application apparatus, as described in PCT application W095/251116 (Baldeschweiler et al.). Alternatively, a gridded array may be used to arrange and link cDNA fragments or oligonucleotides to the surface of a substrate using a vacuum system, thermal, UV, mechanical or chemical bonding procedure. Exemplary nucleic acid molecules useful in the invention include polynucleotides that

specifically bind nucleic acid biomarkers to one or more pathogenic organisms, and fragments thereof.

A nucleic acid molecule (*e.g.* RNA or DNA) derived from a biological sample may be used to produce a hybridization probe as described herein. The biological samples are generally derived from a patient, *e.g.*, as a bodily fluid (such as blood, blood serum, plasma, saliva, urine, ascites, cyst fluid, and the like); a homogenized tissue sample (*e.g.*, a tissue sample obtained by biopsy); or a cell or population of cells isolated from a patient sample. For some applications, cultured cells or other tissue preparations may be used. The mRNA is isolated according to standard methods, and cDNA is produced and used as a template to make complementary RNA suitable for hybridization. Such methods are well known in the art. The RNA is amplified in the presence of fluorescent nucleotides, and the labeled probes are then incubated with the microarray to allow the probe sequence to hybridize to complementary oligonucleotides bound to the biochip.

Incubation conditions are adjusted such that hybridization occurs with precise complementary matches or with various degrees of less complementarity depending on the degree of stringency employed. For example, stringent salt concentration will ordinarily be less than about 750 mM NaCl and 75 mM trisodium citrate, less than about 500 mM NaCl and 50 mM trisodium citrate, or less than about 250 mM NaCl and 25 mM trisodium citrate. Low stringency hybridization can be obtained in the absence of organic solvent, *e.g.*, formamide, while high stringency hybridization can be obtained in the presence of at least about 35% formamide, and most preferably at least about 50% formamide. Stringent temperature conditions will ordinarily include temperatures of at least about 30°C, of at least about 37°C, or of at least about 42°C. Varying additional parameters, such as hybridization time, the concentration of detergent, *e.g.*, sodium dodecyl sulfate (SDS), and the inclusion or exclusion of carrier DNA, are well known to those skilled in the art. Various levels of stringency are accomplished by combining these various conditions as needed. In a preferred embodiment, hybridization will occur at 30°C in 750 mM NaCl, 75 mM trisodium citrate, and 1% SDS. In embodiments, hybridization will occur at 37°C in 500 mM NaCl, 50 mM trisodium citrate, 1% SDS, 35% formamide, and 100 µg/ml denatured salmon sperm DNA (ssDNA). In other embodiments, hybridization will occur at 42°C in 250 mM NaCl, 25 mM trisodium

citrate, 1% SDS, 50% formamide, and 200 µg/ml ssDNA. Useful variations on these conditions will be readily apparent to those skilled in the art.

The removal of nonhybridized probes may be accomplished, for example, by washing. The washing steps that follow hybridization can also vary in stringency. Wash stringency conditions can be defined by salt concentration and by temperature. As above, wash stringency can be increased by decreasing salt concentration or by increasing temperature. For example, stringent salt concentration for the wash steps will preferably be less than about 30 mM NaCl and 3 mM trisodium citrate, and most preferably less than about 15 mM NaCl and 1.5 mM trisodium citrate. Stringent temperature conditions for the wash steps will ordinarily include a temperature of at least about 25°C, of at least about 42°C, or of at least about 68°C. In embodiments, wash steps will occur at 25°C in 30 mM NaCl, 3 mM trisodium citrate, and 0.1% SDS. In a more preferred embodiment, wash steps will occur at 42 C in 15 mM NaCl, 1.5 mM trisodium citrate, and 0.1% SDS. In other embodiments, wash steps will occur at 68 C in 15 mM NaCl, 1.5 mM trisodium citrate, and 0.1% SDS. Additional variations on these conditions will be readily apparent to those skilled in the art.

Detection systems for measuring the absence, presence, and amount of hybridization for all of the distinct nucleic acid sequences are well known in the art. For example, simultaneous detection is described in Heller et al., Proc. Natl. Acad. Sci. 94:2150-2155, 1997. In embodiments, a scanner is used to determine the levels and patterns of fluorescence.

Diagnostic assays

The present invention provides a number of diagnostic assays that are useful for the identification or characterization of a disease or disorder (e.g., oral squamous cell carcinoma), or a propensity to develop such a condition. In one embodiment, oral squamous cell carcinoma (OSCC) is characterized by quantifying the level of one or more biomarkers from one or more pathogenic organisms, including viruses, viroids, bacteria, fungi, helminths, and protozoa. While the examples provided below describe specific methods of detecting levels of these markers, the skilled artisan appreciates that the invention is not limited to such methods. Marker levels are quantifiable by any standard method, such methods include, but are not limited to real-time PCR, Southern blot, PCR, and/or mass spectroscopy.

The level of any two or more of the markers described herein defines the marker profile of a disease, disorder, or condition. The level of marker is compared to a reference. In one embodiment, the reference is the level of marker present in a control sample obtained from a patient that does not have oral squamous cell carcinoma. In
5 another embodiment, the reference is a healthy tissue or cell (i.e., that is negative for oral squamous cell carcinoma). In another embodiment, the reference is a baseline level of marker present in a biologic sample derived from a patient prior to, during, or after treatment for oral squamous cell carcinoma. In yet another embodiment, the reference is a standardized curve. The level of any one or more of the markers described herein (e.g.,
10 a combination of viral, bacterial, fungal, helminth, and/or protozoan biomarkers) is used, alone or in combination with other standard methods, to characterize the disease, disorder, or condition (e.g., oral squamous cell carcinoma).

In certain embodiments, one or more organisms described herein may be isolated or extracted from a sample using a capture reagent (e.g., an antibody) and/or detected
15 using ELISA. In a particular embodiment, reagents for capturing the pathogenic organism include streptavidin bound magnetic beads and biotin labelled probes. Such techniques can be further used to obtain nucleic acids pathogenic organism detection using nucleic acid based probes or for direct sequencing (e.g., MiSeq; Illumina).

20 Kits

The invention provides kits for the detection of a biomarker, which is indicative of the presence of one or more biological sequences or agents associated with oral squamous cell carcinoma. The kits may be used for detecting the presence of multiple biological agents associated with triple negative breast cancer. The kits may be used for the
25 diagnosis or detection of oral squamous cell carcinoma. In some embodiments, the kit comprises a panel or collection of probes to nucleic acid biomarkers (e.g., PathoChip) delineated herein as specific for detection of oral squamous cell carcinoma. In additional or alternative embodiments, the kit comprises an antibody specific for a pathogenic organism associated with oral squamous cell carcinoma. Such antibodies may be used for
30 ELISA detection or for extraction of a pathogenic organism associated with oral squamous cell carcinoma (e.g., a biotin labelled antibody in conjunction with streptavidin bound magnetic beads).

In some embodiments, the kit comprises one or more sterile containers which contain the panel of probes, nucleic acid biomarkers, or microarray chip. Such containers

can be boxes, ampoules, bottles, vials, tubes, bags, pouches, blister-packs, or other suitable container forms known in the art. Such containers can be made of plastic, glass, laminated paper, metal foil, or other materials suitable for holding medicaments.

The instructions will generally include information about the use of the composition for the detection or diagnosis of triple negative breast cancer. In other embodiments, the instructions include at least one of the following: description of the therapeutic agent; dosage schedule and administration for treatment or prevention of oral squamous cell carcinoma or symptoms thereof; precautions; warnings; indications; counter-indications; overdose information; adverse reactions; animal pharmacology; clinical studies; and/or references. The instructions may be printed directly on the container (when present), or as a label applied to the container, or as a separate sheet, pamphlet, card, or folder supplied in or with the container.

One embodiment of the invention is a kit comprising at least three nucleic acid probes selected from the group consisting of SEQ ID NOS: 1-76. The kit can include probes from about 10-30 organisms with about 3-5 probes per organism. Another embodiment of the invention is a kit comprising a microarray with at least three nucleic acid probes selected from the group consisting of SEQ ID NOS: 1-76. In another embodiment, the kit comprises a microarray comprising at least three nucleic acid probes selected from the group of microbes consisting of Human papillomavirus 16 (HPV16), *Eshcherichia*, *Rothia*, *Peptoniphilus*, *Brevundimonas*, *Comamonas*, *Alcaligenes*, *Arcanobacterium*, *Actinomyces*, *Aeromonas*, *Bordetella*, *Aerococcus*, *Pediococcus*, *Caulobacter*, *Cardiobacterium*, *Plesiomonas*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateuria*, *Rhodotorula*, *Geotrichum*, *Pneumocystis*, *Hymenolepis*, *Centrocestus*, *Trichinella*, *Acinetobacter*, *Actinobacillus*, *Veillonella*, and *Fonsecaea*. The kits contain instructional materials for use thereof.

Microbial Integrations into Host Genomes

In the present study, NGS data was used to identify sites of integration of the identified pathogens within the host genome. Numerous open reading frames were detected within the host genome in which integration had occurred. Integration hotspots for HPV16 was identified as well as other identified integration sites for a number of viruses, including the JC polyomavirus, as well as other pathogenic and tumorigenic bacteria, fungi and parasites in these OSCC samples. Data shown herein strongly suggest

greater molecular intimacy between the host genome and the genetic elements of associated microbial agents in the tumor microenvironment.

In the present study, multiple integration sites for viral, bacterial, fungal and parasitic sequences were detected in the host genome, providing the potential for significant alteration in host gene expressions. Data herein showed that insertional sites for HPV16 were found in stochastic fashion in the chromosomes and were mostly intronic. The highest number of integration sites for HPV16 in the present study were seen in chr 17, followed by chr 5. Although, the HPV16 and JC Polyomavirus oncogenic genomic integrations detected in the study were mostly intronic and intergenic, host gene expression still may be altered. HPV16 genomic integration sites in the human genome were detected at the intronic/upstream/downstream region of certain genes associated with cancer. A distribution of HPV integration sites were detected throughout the genome, many of which have the potential to functionally alter critical cellular gene expression through integration.

JC Polyomavirus Large T antigen sequence insertions in the host chromosomes were detected. Without wishing to be bound by specific theory, these insertions may lead to transformations by inducing mutations of the target genes. Also detected herein were VP1, VP2 and VP3 viral genomic sequence insertion sites in multiple regions (intergenic/upstream/downstream) of host chromosomes. Little has been reported regarding polyomavirus insertional sites in the host chromosome. Without being bound by specific theory, these insertional sites at or near the host genes may have a role in regulating the host gene function and contribute to carcinogenesis.

Little is known about bacterial DNA integrations. In the present study, numerous bacterial insertion sites, were detected especially in the exons of host genes. Without wishing to be bound by any specific theory, bacterial DNA insertions in the exonic regions may alter the expression of tumor suppressors, proto-oncogenes and chromatin remodeling genes, suggesting a role in driving oncogenesis. Apart from exonic insertions of bacterial DNA, numerous insertional sites were also detected in the present study at the intronic, UTR, ncRNA, and upstream and downstream of host genes involved in many cellular functions that can contribute to neoplasia.

In the present study, numerous fungal, and parasitic-genomic fusion transcripts were detected, which demonstrated insertion of fungal and parasitic sequences in the host chromosomes. No previous reports are available for fungal genomic integrations. In the present study, the fungal genomic sequence insertions in the host genome was mostly

intergenic or intronic, and may have a prominent role in regulating exonic expression. Similarly, parasitic sequence insertions detected at the proximity of proto-oncogenes, tumor suppressors and miRNAs may also lead to altered expressions of these genes, and ultimately lead to oncogenesis.

5 Some of the host genes that had microbial genomic insertions supported by high sequence reads showed significant association with neoplasia of epithelial tissue, and thus these microbial insertions at or near cancer associated genes can be a contributing factor for OSCC development.

10 A microbial-host fusion map was identified which for the first time comprehensively identifies the integrations of microbial sequences across the kingdoms throughout the somatic human chromosomes. Without wishing to be bound by specific theory, this may lead to altered host gene expression, and co-operation of the associated microbiome to the initiation and development of OSCC.

15 The practice of the present invention employs, unless otherwise indicated, conventional techniques of molecular biology (including recombinant techniques), microbiology, cell biology, biochemistry and immunology, which are well within the purview of the skilled artisan. Such techniques are explained fully in the literature, such as, “Molecular Cloning: A Laboratory Manual”, fourth edition (Sambrook, 2012); “Oligonucleotide Synthesis” (Gait, 1984); “Culture of Animal Cells” (Freshney, 2010);
20 “Methods in Enzymology” “Handbook of Experimental Immunology” (Weir, 1997); “Gene Transfer Vectors for Mammalian Cells” (Miller and Calos, 1987); “Short Protocols in Molecular Biology” (Ausubel, 2002); “Polymerase Chain Reaction: Principles, Applications and Troubleshooting”, (Babar, 2011); “Current Protocols in Immunology” (Coligan, 2002). These techniques are applicable to the production of the polynucleotides
25 and polypeptides of the invention, and, as such, may be considered in making and practicing the invention. Particularly useful techniques for particular embodiments will be discussed herein.

30 The following examples are put forth so as to provide those of ordinary skill in the art with a complete disclosure and description of how to make and use the assay, screening, and therapeutic methods of the invention, and are not intended to limit the scope of what the inventors regard as their invention.

EXPERIMENTAL EXAMPLES

The invention is further described in detail by reference to the following experimental examples. These examples are provided for purposes of illustration only, and are not intended to be limiting unless otherwise specified. Thus, the invention should in no way be construed as being limited to the following examples, but rather, should be construed to encompass any and all variations which become evident as a result of the teaching provided herein.

Without further description, it is believed that one of ordinary skill in the art can, using the preceding description and the following illustrative examples, make and utilize the compounds of the present invention and practice the claimed methods. The following working examples therefore, specifically point out the exemplary embodiments of the present invention, and are not to be construed as limiting in any way the remainder of the disclosure.

The materials and methods employed in these experiments are now described.

PathoChip design:

The PathoChip Array design has been previously described (Baldwin et al., (2014) mBio 5, e01714-01714; Banerjee et al., (2015) Sci Rep 5, 15162). Briefly, the array was generated from a metagenome of 58 chromosomes in silico. It is comprised of 60,000 probe sets of sequenced microorganisms from Genbank, which are manufactured as SurePrint glass slide microarrays (Agilent Technologies Inc.), containing 8 replicate arrays per slide (Baldwin et al., (2014) mBio 5, e01714-01714). Each probe is a 60-nt DNA oligomer that targets multiple genomic regions of pathogenic viruses, prokaryotic, and eukaryotic microorganisms.

Sample preparation and Microarray processing:

PathoChip screening utilized both DNA and RNA extracted from formalin-fixed paraffin-embedded (FFPE) tumor tissues. 100 de-identified FFPE oral squamous cell carcinoma (OSCC) samples were received as 10 µm sections on non-charged glass slides, and 20 each of matched and non-matched control samples were provided as paraffin rolls. Matched controls were obtained from the adjacent non-cancerous oral tissue of the same patient from which the cancer tissues were obtained, while non-matched controls were oral tissues obtained from otherwise healthy individuals. DNA and RNA were extracted in parallel from rolls or mounted sections of each FFPE sample. The quality of extracted nucleic acids was determined by agarose gel electrophoresis and the A_{260/280} ratio. The extracted RNA and DNA samples were subjected to whole transcriptome amplification

(WTA) using 50 ng each of RNA and DNA as input. A total of 60 arrays were used to screen the 100 OSCC samples, with 48 individual and the rest pooled in groups of 4-5 samples. The 20 matched and 20 non-matched control samples were pooled for screening using 4 arrays for each set of controls. The WTA products were analyzed by agarose gel electrophoresis and showed a range of 200-400bp amplicon sizes. Human reference RNA and DNA were also extracted from the human B cell line, BJAB and 15ng of each were used for WTA. The WTA products were purified, (PCR purification kit, Qiagen, Germantown, MD, USA), and 2µg of the amplified products from the OSCC tissues was labelled with Cy3 and that from the human reference was labelled with Cy5 (SureTag labeling kit, Agilent Technologies, Santa Clara, CA). Human reference DNA and RNA was used to determine cross-hybridization of probes to human DNA. The labelled DNAs were purified and the efficiencies of labeling were determined by measuring absorbance at 550nm (for Cy3) and 650nm (for Cy5). The labelled samples (Cy3 plus Cy5) were hybridized to the PathoChip as described previously (Baldwin et al., (2014) mBio 5, e01714-01714; Banerjee et al., (2015) Sci Rep 5, 15162). The hybridization cocktail (CGH blocking agent and hybridization buffer), was added to each of the labeled test sample (Cy3) mixed with reference (Cy5), denatured and hybridized to the arrays in 8-chamber gasket slides. The slides were incubated at 65°C with rotation and washed, then scanned for visualization using an Agilent SureScan G4900DA array scanner.

Microarray Data Extraction and Statistical analysis:

Raw data from the microarray images were extracted using Agilent Feature Extraction software; normalization and data analyses were done in the Partek Genomics Suite (Partek Inc., St. Louis, MO, USA). Model-based analysis of tiling arrays (MAT), which utilized a sliding window to scroll through the entire metagenome of the array to detect positive hybridization signal, was used to detect positive regions in the metagenome for each tumor. Analysis at the individual probe level (both for specific and conserved probes), and at the accession level (taking into account all the probes per accession), were performed as previously described (Baldwin et al., (2014) mBio 5, e01714-01714; Banerjee et al., (2015) Sci Rep 5, 15162). Probes of the microorganisms (microbial signatures) were detected in the samples by both outlier analyses (detecting probes in few samples) and paired t-tests with False Discovery Rate (FDR) multiple correction (detecting probes of significance in the majority of the tumor samples analyzed). One sided t-tests were performed to determine if cancer samples have significant detection of the candidate signature of organisms compared to the control

(both matched and non-matched) samples. The cancer samples were also subjected to hierarchical clustering, based on the detection of microbial signatures in the samples, using the R program (Euclidean distance, complete linkage, non-adjusted values), and NBClust software [CH (Calinski and Harabasz) index, Euclidean distance, complete linkage] (Charrad et al., (2014) Journal of Statistical Software 61, 1-36). The significant differences between the clusters observed by these methods were determined using t-test. Additional topological-based data analyses were conducted using the Ayasdi software (Ayasdi, Inc.), (Correlation metric, and L-infinity centrality lenses) where statistical significance between different groups was determined using two-sided t- test.

Probe Capture and Next Generation Sequencing:

WTA products of the oral cancer samples were pooled together for hybridization with selected biotinylated probes that were identified for microbial signatures in the oral cancer samples by the PathoChip screen. The targeted sequences were then captured by Streptavidin coated magnetic beads and libraries were generated for NGS. The selected probes were synthesized as 5'-biotinylated DNA oligomers (Integrated DNA Technologies, Coralville, IA, USA), mixed as 5 pools of capture probes (pools 1-5) (FIGs. 5A-5D, Table 1), and hybridized to WTA pools of oral cancer samples. Capture probe pool 1 contained 19 selected probes associated with bacteria (B capture), pool 2 contained 12 selected probes associated with the fungi (F capture), pool 3 contained 14 selected probes associated with parasitic signatures (P capture), pool 4 contains 36 other probes associated with viral and some bacterial signatures (O capture), pool 5 contains 6 HPV16 probes (HPV16 capture) (Table 1). Each of the 5 capture probe pools was added separately to the pooled WTA of the oral cancer samples (150ng/ul) in 5 separate reaction mixtures containing 3 M tetra-methyl ammonium chloride, 0.1% Sarkosyl, 50 mM Tris-HCl, 4 mM EDTA, pH 8.0 (1XTMAC buffer). 5 target capture reactions were done (Table 1). The reaction mixtures were denatured (100°C for 10 mins) followed by a hybridization step (60°C for 3 hours). Streptavidin Dynabeads (Life Technologies, Carlsbad, CA, USA) were added with continuous mixing at room temperature for 2 hours, followed by three washes of the captured bead-probe-target complexes in 0.30 M NaCl plus 0.030 M sodium citrate buffer (2XSSC) and three washes with 0.1× SSC. Captured single-stranded target DNA was eluted in Tris-EDTA and used for library preparation using Nextera XT sample preparation kit (Illumina, San Diego, CA, USA) followed by NGS. The 5 libraries were examined for quality control and submitted for NGS using an Illumina MiSeq instrument with paired-end 250-nt reads. Adapters and low-quality

fragments of raw reads were first removed using the Trim Galore software (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The processed reads were then aligned to the metagenome and the human genome using Genomic Short-read Nucleotide Alignment Program (GSNAP) (Thorvaldsdottir et al., (2013) Brief Bioinform 14, 178-192; Wu and Nacu, (2010) Bioinformatics 26, 873-881) with default parameters. After alignment, featureCounts (Liao et al., (2013) Bioinformatics 30, 923-930) was employed to count how many reads aligned to each of the capture probe regions. The detailed results for these capture probes are visualized in IGV (FIGs 5A-5D and FIGs. 7A-7G).

10

Table 1: List of the capture probe sequences used for capture sequencing

Capture Probe Library	Microorganisms	Probe sequences used for capture 5'-3'	SEQ ID NO.
B (Bacterial pool)	[AJ250959.1] <i>Arcanobacterium</i> sp.	/5Biosg/GAATCCCTTAAAGCCAGTCTCAGTTCGGATTGGGGTCTGCAACTC	SEQ ID NO. 1
	[X73952.1] <i>Arcanobacterium haemolyticum</i>	/5Biosg/GAATCCCTTAAAGCCAGTCTCAGTTCGGATTGGGGTCTGCAACTC	SEQ ID NO. 2
	[X97049.1] <i>Arcanobacterium phocae</i>	/5Biosg/GAATCCCTTAAAGCCAGTCTCAGTTCGGATTGGGGTCTGCAACTC	SEQ ID NO. 3
	[AJ879697.1] <i>Arcanobacterium</i> sp.	/5Biosg/CTAATCACTAAAGTCTGTCTCAGTTCGGATTGGGGTCTGCAACTCGAC	SEQ ID NO. 4
	[AJ576084.1] <i>Mobiluncus curtisii</i>	/5Biosg/CAACTCGACCCCATGAAGTTGGAGTCGCTAGTAATCGCAGATCAG	SEQ ID NO. 5
	[J01859.1] <i>Escherichia coli</i>	/5Biosg/TAAAACTCAAATGAATTGACGGGGGCCCCGACAAAGCGGTGGAGCATGTGGTTTAATTCTGA	SEQ ID NO. 6
	[X78719.1] <i>Actinomyces radinqae</i>	/5Biosg/ATCGGGGTCTGCAACTCGACCTCGTGAAGGTGGAGTCGCTAGTAAT	SEQ ID NO. 7
	[X78720.1] <i>Actinomyces turicensis</i>	/5Biosg/ATATCGTAAGGTGGAGCGAATCACTTAAAGCCAGTCTCAGTTCGGATTG	SEQ ID NO. 8
	[X82451.1] <i>Actinomyces meyeri</i>	/5Biosg/GAATCCCTTAAAGCCAGTCTCAGTTCGGATTGGGGTCTGCAACTC	SEQ ID NO. 9
	[AJ717364.1] <i>Rothia dentocariosa</i>	/5Biosg/CTAATCCCTAAAGCCGGTCTCAGTTCGGATTGGGGTCTGCAACT	SEQ ID NO. 10
	[AJ937773.1] <i>Propionibacterium jensenii</i>	/5Biosg/GAATCTCAAAAAGCCAGTCTCAGTTCGGATTGGGGTCTGCAACTC	SEQ ID NO. 11
	[D14147.1] <i>Peptoniphilus indolicus</i>	/5Biosg/CCTGAGAGGGTGAACGGCCACATTGGAACCTGAGAAACGGTCCAAA	SEQ ID NO. 12
	[D14147.1] <i>Peptoniphilus indolicus</i>	/5Biosg/GATAGCAAGCGAATCTCAAAAAGCCTATCTCCTCGGATTGTCTCTGCAACT	SEQ ID NO. 13
	[AF480601.1] <i>Mycobacterium kansasii</i>	/5Biosg/CGGGATGAAGCTGGGAACTGGGTCTAATACCGGATAGGACCACT	SEQ ID NO. 14
	[AF480601.1] <i>Mycobacterium kansasii</i>	/5Biosg/CGAGGTGAAGCGAATCTTTTAAAGCCGGTCTCAGTTCGGATCGG	SEQ ID NO. 15
	[AJ16941.1] <i>Mycobacterium chelonae</i>	/5Biosg/ATAGCTGGGAACTGGGTCTAATACCGGATAGGACCACTCT	SEQ ID NO. 16
	[AY457068.1] <i>Mycobacterium fortuitum</i>	/5Biosg/CACCTTTGGGATAAGCTGGGAACTGGGTCTAATACCGAATAGGAC	SEQ ID NO. 17
	[FM180525.1] <i>Brevundimonas diminuta</i>	/5Biosg/GAGTTGACAGAGGACAATCCGAAGTGAACAATTTAAGGATTAACCTCTGTAG	SEQ ID NO. 18
	[AB091200.1] <i>Frateuria aurantia</i>	/5Biosg/GAAACTTACGCTAATACCGCATACGCTACCGGAGAGAAGCAGG	SEQ ID NO. 19
F (Fungal pool)	[AB126649.1] <i>Rhodotorula laryngis</i>	/5Biosg/CTTCAATCAATACCTGCCCTATCAACTTTCGATGGTAGGATAGAGGCCCTAC	SEQ ID NO. 20
	[DQ325447.1] <i>Geotrichum citri-aurantii</i>	/5Biosg/AGAGCTCTTTCATGATTTTTGGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGTGAT	SEQ ID NO. 21
	[EU192364.1] <i>Malassezia globosa</i>	/5Biosg/GTGGCGGAGTACTTCAATCAAAATATCTGCCCTATCAACTGTCTGATGGTAG	SEQ ID NO. 22
	[AY083223.1] <i>Malassezia furfur</i>	/5Biosg/TTCATCAAAATATCTGCCCTATCAACTGTCTGATGGTAGGATAGAGGCCCTACCAT	SEQ ID NO. 23
	[EU192366.1] <i>Malassezia pachydermatis</i>	/5Biosg/ATGCTTCATCAAAATATCTGCCCTATCAACTGTCTGATGGTAGGATAGAGGCCCTA	SEQ ID NO. 24
	[EU192367.1] <i>Malassezia restricta</i>	/5Biosg/CTGGCGGAGTCTTCAATCAAAATATCTGCCCTATCAACTGTCTGATGGTAG	SEQ ID NO. 25
	[EU192368.1] <i>Malassezia slooffiae</i>	/5Biosg/CGGGGGTGTCTTCAATCAAAATATCTGCCCTATCAACTGTCTGATGGTAG	SEQ ID NO. 26
	[EU192369.1] <i>Malassezia sympodialis</i>	/5Biosg/ATGCTTCATCAAAATATCTGCCCTATCAACTGTCTGATGGTAGGATAGAGGCCCTA	SEQ ID NO. 27
	[EF119339.1] <i>Pleistophora mulleri</i>	/5Biosg/GAGGAGAACTGTAAGCAGCGGTTAAGCAGGAGTAACTGCGCGGAGTCACT	SEQ ID NO. 28
	[EF119339.1] <i>Pleistophora mulleri</i>	/5Biosg/GCAGAAAGCGAGCGGTGCTTTTGAATCTCGATTCGACTCTCCTATCATG	SEQ ID NO. 29
	[AF044392.1] <i>Pleistophora ehrenbaumi</i>	/5Biosg/AAAGCAGGGAAGGAAGGTCAGAACAGTCAAGTGCATGCTCCCTTAC	SEQ ID NO. 30
	[AB266392.1] <i>Pneumocystis jirovecii</i>	/5Biosg/CTTGTCTGGCGATGGTCTTCAATCAAAATCTGCCCTATCAACTTTCGATG	SEQ ID NO. 31
	[AF124475.1] <i>Hymenolepis diminuta</i>	/5Biosg/CTTCAAAATGTCTGCCCTATCAACTTTCGATGGTAGGTCACCTGCTAC	SEQ ID NO. 32
	[AY245759.1] <i>Centrocestus formosanus</i>	/5Biosg/AAAGTCAGAGGTCGAAGACGATCAGATCCGCTCTAGTTCGACCATAC	SEQ ID NO. 33
	[AF151921.1] <i>Prosthodendrium longiforme</i>	/5Biosg/GTCTGGAACACGGACCAAGGAGAGTAACTGCGCGGAGTCACT	SEQ ID NO. 34
	[AF151922.1] <i>Prosthodendrium hurkovaee</i>	/5Biosg/GAGCGGTAGGCGCATCACCGGCCGCTCCATGACAATGTCTCTG	SEQ ID NO. 35
	[AF151922.1] <i>Prosthodendrium hurkovaee</i>	/5Biosg/CCGAAGCGGTCTGACGTGCAATCGATCGCCAGACATGAGTATA	SEQ ID NO. 36
	[AY851256.1] <i>Trichinella nativa</i>	/5Biosg/ACATCACTTTCAAGTGTCTGCCCTATCAACTTTCGATGGTAGTCTATGCGACTA	SEQ ID NO. 37
	[AY851257.1] <i>Trichinella britovi</i>	/5Biosg/ACATCACTTTCAAGTGTCTGCCCTATCAACTTTCGATGGTAGTCTATGCGACTA	SEQ ID NO. 38
P (Parasitic pool)	[AY851259.1] <i>Trichinella murrelli</i>	/5Biosg/ACATCACTTTCAAGTGTCTGCCCTATCAACTTTCGATGGTAGTCTATGCGACTA	SEQ ID NO. 39
	[AY851260.1] <i>Trichinella</i> sp.	/5Biosg/ACATCACTTTCAAGTGTCTGCCCTATCAACTTTCGATGGTAGTCTATGCGACTA	SEQ ID NO. 40
	[AY851262.1] <i>Trichinella</i> sp.	/5Biosg/ACATCACTTTCAAGTGTCTGCCCTATCAACTTTCGATGGTAGTCTATGCGACTA	SEQ ID NO. 41
	[AY851263.1] <i>Trichinella papuae</i>	/5Biosg/ACATCACTTTCAAGTGTCTGCCCTATCAACTTTCGATGGTAGTCTATGCGACTA	SEQ ID NO. 42
	[AY851264.1] <i>Trichinella zimbabwensis</i>	/5Biosg/ACATCACTTTCAAGTGTCTGCCCTATCAACTTTCGATGGTAGTCTATGCGACTA	SEQ ID NO. 43
	[AY702702.1] <i>Contracaecum microcephalum</i>	/5Biosg/GATCAGATACGCCCTAGTCTGACGCGAATCACTACCAATGACGCTT	SEQ ID NO. 44
	[J094382.1] <i>Toxocara canis</i>	/5Biosg/TAACCCGTTGCAAAATCTCCGTGATCGGATGGGAATTCGAATTAATT	SEQ ID NO. 45
	[NC_008186.1] Mouse parvovirus 2	/5Biosg/AAAGACGCTAAAGACTGGGAGGCAAGGTTGGTCACTACTTCTTAGAA	SEQ ID NO. 46
	[NC_001664.2] Human herpesvirus 6A	/5Biosg/TATAGACAGCCCGATGCCAACGCGACAGGCGTGAAGTACATAGAT	SEQ ID NO. 47
	[NC_001664.2] Human herpesvirus 6A	/5Biosg/CAATACAGAGGTTGTGTTGGCAAAATTAAGCAGCAATTTGTGTATCTAAATATTAAAG	SEQ ID NO. 48
	[NC_006879.1] Simian adenovirus 1	/5Biosg/TCTTGATCAGTCTCTCCACCGGAACCTCTGCTTCTCCCTGTCT	SEQ ID NO. 49
	[NC_006879.1] Simian adenovirus 1	/5Biosg/AGAGGCGAGAGAGGAGCAACTGGGCAAGCAGCTGAACAGCATC	SEQ ID NO. 50
	[NC_002657.1] Classical swine fever virus	/5Biosg/ATTGGGGAACCAAGCCGAGAGAGGAGGAGAGTTGGGAGAGTAAAG	SEQ ID NO. 51
	[NC_001699.1] JC polyomavirus	/5Biosg/TAACAGCCAGTAAACAAAGCACAAGGGGAAGTGGAAAGCAGCCAA	SEQ ID NO. 52
	[NC_001699.1] JC polyomavirus	/5Biosg/AAAGGGAGGGAACCTATATTTCTTTTGGCCACTATACACCCAAAGTATAGATGATG	SEQ ID NO. 53
	[NC_001699.1] JC polyomavirus	/5Biosg/TTATAGGGGTGACAAAGTTTGTATGAATGTGCACTCTAATGGCAAGCAACTCAT	SEQ ID NO. 54
	[NC_001664.2] Human herpesvirus 6A	/5Biosg/CTAAACCGTCTTTCCTCTCAAACTTTAATGAGAAAAAGCTTACGGGGTGAAC	SEQ ID NO. 55
	[NC_001664.2] Human herpesvirus 6A	/5Biosg/ACATTAATACAGGCAAGGACACAATTCACGGCATCTTCTGAAGCGTGTGAC	SEQ ID NO. 56
	[NC_001664.2] Human herpesvirus 6A	/5Biosg/CTATCAGAATTACGAACGCCCGATAAACCTAATCCCGCGATCCAAATTAATAAT	SEQ ID NO. 57
	[NC_001664.2] Human herpesvirus 6A	/5Biosg/AACTTTACTCATGGTTCTGACAGTCTCACTTTCAAGACATTCACTTTCCAG	SEQ ID NO. 58
O (Other pool)	[NC_001664.2] Human herpesvirus 6A	/5Biosg/AAGTGGTAACTTTAAATTCGGAGTGGTGTGTGCGCGTGGTGAATTTT	SEQ ID NO. 59
	[NC_001664.2] Human herpesvirus 6A	/5Biosg/CGATGACCCGCTAAATTTGATAAATAGTCTGCTCTCACTCGGATTCAAGAAA	SEQ ID NO. 60
	[NC_001664.2] Human herpesvirus 6A	/5Biosg/TTAGCAATTTTAAAGTTCTAACCCTGTTGGGTAGGAAAGACCTTAACCA	SEQ ID NO. 61
	[NC_001501.1] Moloney murine leukemia virus	/5Biosg/GATAGACAGGAGGAGAGCAAGGAGGTCCTCACTCGATCGCGAC	SEQ ID NO. 62
	[NC_001501.1] Moloney murine leukemia virus	/5Biosg/GTAAACCAATGGAGATCGGAGAGCGATGGGCAACTCTCGGCAAC	SEQ ID NO. 63
	[NC_014480.2] Torque teno virus 2	/5Biosg/AAACACAGAATCACAGTCAAGATAGGAGCCCTTAAACCTTTGAGGACAAG	SEQ ID NO. 64
	[NC_014480.2] Torque teno virus 2	/5Biosg/GAGACAGGAGACAGCCATAAACACCAACGCTAGACTGACACTAATATGC	SEQ ID NO. 65
	[NC_01583.1] Human papillomavirus type 26	/5Biosg/CTATGTATGTTAATTAATGTCAGCAAGCACTAATGAATGAAACCAAAATTTGCG	SEQ ID NO. 66
	[NC_01583.1] Human papillomavirus type 26	/5Biosg/TGGTGCAATGGGCGTTCGATCATGACATAACAGATGATAGTGAATTCATTAATATG	SEQ ID NO. 67
	[NC_015961.1] Tomato chlorotic leaf distortion virus	/5Biosg/TATGGCTCTCTCTTCTGTCAACGTGGCGCATATAGACCATGGTCTTAATAA	SEQ ID NO. 68
	[NC_010729.1] <i>Porphyromonas gingivalis</i>	/5Biosg/GGGGAGAGGTAAGAGGACTTAAGGAGCAGGAACTAATGAATGAAATAGAAATATATGCA	SEQ ID NO. 69
	[AY689230.1] <i>Prevotella nigrescens</i>	/5Biosg/CAACGGGCCACCAAGGCGACGATCAGTAGGGGTTCTGAGAGGAAG	SEQ ID NO. 70
	[AY689230.1] <i>Prevotella nigrescens</i>	/5Biosg/TGTTGTGAATTTAGGTGCTCAACATTTAACTTGCAGCGCGAACTGTCAGACTT	SEQ ID NO. 71
	[AY689230.1] <i>Prevotella nigrescens</i>	/5Biosg/TGTTGTGAATTTAGGTGCTCAACATCTTGAATCTGAACTCCGAGTCCGACT	SEQ ID NO. 72
HPV16 pool	[NC_001526.2] Human papillomavirus type 16	/5Biosg/TGCAAGGCGAGCAATGTAGCAAAATTTAAAGAGTTATACGGGGTGAGTTTTTCAGAAAT	SEQ ID NO. 73
	[NC_001526.2] Human papillomavirus type 16	/5Biosg/CAACGAAGTATCTCTCTGAAATTTATAGGCGAGCACTTGCCCAACCAAC	SEQ ID NO. 74
	[NC_001526.2] Human papillomavirus type 16	/5Biosg/CTTGGGACCGGAAACCAACACAGACGACTTCCAGCGACCAAGAT	SEQ ID NO. 75
	[NC_001526.2] Human papillomavirus type 16	/5Biosg/GGAATTTTGGTCTACAACTCTCCCAAGAGGACACACTAGAAGATA	SEQ ID NO. 76

Microbial Fusion Detection:

Prior to fusion detection, quality control of sequenced reads was applied. The Trim Galore software (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) was employed for quality trimming of raw reads in order to remove adapters and low-quality fragments. Virus-Clip (Ho et al. (2015) Oncotarget 6, 20959-20963) was used to identify the virus fusion sites in the human genome. Specifically, the virus genome was used as the primary read alignment target, and first aligned reads to the PathoChip genome. Some mapped reads may contain soft-clipped segments. Soft-clipped reads were then extracted from the alignment and mapped (containing sequences of potential pathogen-integrated human loci) to the human genome. Utilizing this mapping information, the exact human and pathogen integration breakpoints at single-base resolution can be identified. All the integration sites were then automatically annotated with the affected human genes and their corresponding gene regions.

Some of the host genes that supported viral/microbial genomic insertions by high sequence reads were subjected to Ingenuity Pathway Analysis (IPA) that helped to combine the host genes with knowledge extracted from the literatures to predict likely outcomes. IPA software provided a statistical significance of the association of those genes with the disease outcome.

The results of the experiments are now described.

Example 1: Microbial signatures detected in OSCCs

The PathoChip technology was used to screen 100 FFPE pathologically defined OSCC patient samples as well as 20 matched and 20 non-matched oral tissue control samples for distinct viral and microbial signatures associated with the tumor tissue. Samples analyzed in this study were carcinomas taken from tongue, base of tongue, tonsil, floor of mouth, cheek and predominantly oropharynx which are collectively referred to herein as OSCC (Table 2). To detect the microbiome associated with OSCC, both DNA and RNA were extracted from the samples, subjected to whole genome amplification, labelled and hybridized to the probes on the PathoChip.

Table 2: Tumor location and HPV16 positivity status of the OSCC samples screened by PathoChip.

Study Code	Tumor location	Tumor type	p16
ENT01	tongue/tonsil	Squamous Cell Carcinoma (SCC)	Positive
ENT02	tongue/tonsil	SCC	Positive
ENT03	tonsil	SCC	Positive
ENT04	tonsil/base of tongue	SCC	Positive
ENT05	tonsil	SCC	Positive
ENT06	tonsil	SCC with sarcomatoid growth	Positive
ENT07	tonsil	SCC	Positive
ENT08	tonsil	SCC	Positive
ENT09	tonsil/tongue base	SCC	Positive
ENT10	floor of mouth/cheek	SCC	Negative
ENT11	tonsil/tongue base	SCC	Positive
ENT12	tonsil/tongue base	SCC	Positive
ENT13	tonsil	SCC	Positive
ENT14	oropharynx/palatine tonsil	SCC	Positive
ENT15	base of tongue	SCC	Negative
ENT16	tongue base	SCC	Positive
ENT17	tonsil/pharynx	SCC	Positive
ENT18	tonsil	SCC	Positive
ENT19	tonsil	SCC	Positive
ENT20	tonsil	SCC	Positive
ENT21	tongue base	SCC	Positive
ENT22	tonsil	SCC	Positive
ENT23	tonsil	SCC	Positive
ENT24	tongue base	SCC	Positive
ENT25	tongue base	SCC	Negative
ENT26	tonsil	SCC	Positive
ENT27	palatine tonsil	SCC	Positive
ENT28	oropharynx/base of tongue including lingual tonsil	SCC	Positive
ENT29	oropharynx	SCC	Positive
ENT30	Retromolar trigone	SCC with clear cell change	Positive
ENT31	oropharynx/palatine tonsil	SCC	Positive
ENT32	tongue/floor of mouth	SCC	Negative
ENT33	Tonsil	SCC	Positive
ENT34	tonsil/tongue base	SCC	Positive

ENT35	tonsil	Poorly differentiated carcinoma	Positive
ENT36	retromolar trigone/palate	SCC	Positive
ENT37	pharynx	SCC	Negative
ENT38	buccal mucosa	SCC	Negative
ENT39	tonsil	SCC	Positive
ENT40	tonsil	SCC	Positive
ENT41	base of tongue including lingual tonsil	SCC	Positive
ENT42	tonsil	SCC	Positive
ENT43	tonsil	SCC	Positive
ENT44	oropharynx, base of tongue including lingual tonsil	SCC	Positive
ENT45	oropharynx, palatine tonsil	SCC	Positive
ENT46	base of tongue, including lingual tonsil	SCC	Positive
ENT47	base of tongue including lingual tonsil	SCC	Negative
ENT48	tongue	SCC	Negative
ENT49	oropharynx/palatine tonsil, base of tongue including lingual tonsil	SCC	Positive
ENT50	palatine tonsil, base of tongue including lingual tonsil	SCC	Positive
ENT51	oropharynx/palatine tonsil, base of tongue including lingual tonsil	SCC	Positive
ENT52	tonsillar pillar	SCC	Negative
ENT53	oropharynx/palatine tonsil, base of tongue including lingual tonsil	SCC	Positive
ENT54	tongue/floor of mouth	SCC	Negative
ENT55	oropharynx/palatine tonsil	SCC	Positive
ENT56	oropharynx/soft palate	SCC	Positive
ENT57	tonsil	SCC	Positive
ENT58	oropharynx/base of tongue including lingual tonsil	SCC	Positive
ENT59	oropharynx/base of tongue including lingual tonsil	SCC	Positive
ENT60	oropharynx/palatine tonsil, base of tongue including lingual tonsil/pharyngeal wall (posterior)	SCC	Positive
ENT61	tonsil	SCC	Positive

ENT62	base of tongue including lingual tonsil	SCC	Positive
ENT63	oropharynx base of tongue, including lingual tonsil	SCC	Positive
ENT64	oropharynx base of tongue, including lingual tonsil	SCC	Positive
ENT65	oropharynx base of tongue, including lingual tonsil	SCC	Positive
ENT66	palatine tonsil	SCC	Positive
ENT67	oropharynx palatine tonsil base of tongue, including lingual tonsil	SCC	Positive
ENT68	tonsil, base of tongue	SCC	Positive
ENT69	hard palate	SCC	Positive
ENT70	tonsil	SCC	Positive
ENT71	tonsil	SCC	Positive
ENT72	tongue base	SCC	Positive
ENT73	tonsil, tongue base	SCC	Negative
ENT74	tonsil and tongue base	SCC	Positive
ENT75	tonsil	SCC	Positive
ENT76	oropharynx base of tongue, including lingual tonsil, pharyngeal wall (posterior)	SCC	Positive
ENT77	tonsil ,tongue base, pharynx	SCC	Positive
ENT78	oropharynx palatine tonsil	SCC	Positive
ENT79	tonsil/tongue base	SCC	NA
ENT80	tongue base	SCC	NA
ENT81	tonsil/base of tongue	SCC	NA
ENT82	tongue base	SCC	Positive
ENT83	tonsil	SCC	Positive
ENT84	buccal mucosa	SCC	NA
ENT85	tongue	SCC	Negative
ENT86	oropharynx, tonsil	SCC	Positive
ENT87	oropharynx, tongue base	SCC	Positive
ENT88	tonsil/tongue base	SCC	Positive
ENT89	tonsil	SCC	Positive
ENT90	tonsillar pillar	SCC	Positive
ENT91	tongue base	SCC	Positive
ENT92	tongue base	SCC	Positive
ENT93	tonsil	SCC	Positive

ENT94	tonsil	SCC	NA
ENT95	left cheek and lip resection	SCC	NA
ENT96	Left radical tonsillectomy, tongue base resection	SCC	Positive
ENT97	Tongue, right	SCC	NA
ENT98	Left radical tonsillectomy, tongue base resection, limited pharyngectomy	SCC	Positive
ENT99	Left tonsil, left radical tonsillectomy, tongue base resection, limited pharyngectomy	SCC	Positive
ENT100	Right tonsillectomy, tongue base resection	SCC	NA

i.) Viral signatures associated with OSCC

RNA and DNA viruses associated with the cancer and control samples were identified (FIGs. 1A-1F, Table 3). Viral sequences belonging to Papillomaviridae showed the highest hybridization signal in the OSCC samples screened, followed by that of Herpesviridae, Poxviridae, Retroviridae and Polyomaviridae (FIG. 1A). Viral signatures belonging to these families were seen to be >75% prevalent among the 100 OSCC samples screened. Interestingly, Papillomaviridae was detected in 98% of the cases (FIG. 1A). The hybridization signal for all papillomaviruses was much higher in the OSCC samples compared to the matched and non-matched controls (FIGs. 1A-1C, 1E and Table 4). Importantly, HPV16 was detected with both high hybridization signal and prevalence (98%) only in the OSCC samples (FIGs. 1A, 1D, 1E & 1F). FIG. 1F shows the detection of almost all of the HPV16 specific probes in the PathoChip across the majority of the OSCC samples with medium to high hybridization signal, while the HPV16 probes were detected with significantly lower hybridization signals in both matched and non-matched controls (FIGs. 1B-1C). Signatures of Reoviridae, Herpesviridae, Poxviridae, Orthomyxoviridae, Retroviridae and Polyomaviridae were detected in OSCC samples with high prevalence and at hybridization signals that were 2-3 logs higher than in controls (FIG. 1A & Table 4). Notably, viral signatures of Coronaviridae, Picornaviridae, Adenoviridae, Anelloviridae, Hepadnaviridae and Flaviviridae were significantly detected in the controls along with signatures of non-HPV16 papillomaviridae (FIGs. 1B-1C). These data show that viral signature is significantly changed when compared specifically to the OSCC tissue.

Table 3. Microbial signatures detected in OSCC and control samples. The putative microbial biomarkers are in **bold**.

	Cancer	MC	NC	Cancer/MC	Cancer/NC	Cancer/MC/NC	NC/MC
Viral signatures	HPV16	Coronaviridae	Togaviridae			Papillomaviridae	Adenoviridae
						Poxviridae	Picomaviridae
						Retroviridae	Flaviviridae
						Polyomaviridae	Hepadnaviridae
						Herpesviridae	Parvoviridae
						Reoviridae	Anelloviridae
						Orthomyxoviridae	
Bacterial signatures	<i>Eshcherichia</i>	<i>Campylobacter</i>	<i>Porphyromonas</i>	<i>Arcanobacterium</i>	<i>Mobiluncus</i>	<i>Mycobacterium</i>	
	<i>Rothia</i>	<i>Streptobacillus</i>		<i>Actinomyces</i>	<i>Propionibacterium</i>	<i>Sphingobacterium</i>	
	<i>Peptoniphilus</i>			<i>Aeromonas</i>	<i>Prevotella</i>	<i>Streptococcus</i>	
	<i>Brevundimonas</i>			<i>Bordetella</i>	<i>Citrobacter</i>		
	<i>Comamonas</i>			<i>Aerococcus</i>	<i>Sphingomonas</i>		
	<i>Alcaligenes</i>			<i>Pediococcus</i>	<i>Peptostreptococcus</i>		
	<i>Caulobacter</i>			<i>Acinetobacter</i>			
	<i>Cardiobacterium</i>			<i>Actinobacillus</i>			
	<i>Plesiomonas</i>			<i>Veillonella</i>			
	<i>Serratia</i>						
	<i>Edwardsiella</i>						
	<i>Haemophilus</i>						
	<i>Frateuria</i>						
Fungal signatures	<i>Rhodotorula</i>	<i>Pseudallescheria</i>	<i>Piedraia</i>	<i>Fonsecaea</i>		<i>Malassezia</i>	<i>Microsporium</i>
	<i>Geotrichum</i>		<i>Trichosporon</i>			<i>Pleistophora</i>	<i>Alternaria</i>
	<i>Pneumocystis</i>		<i>Encephalitozoon</i>			<i>Absidia</i>	
			<i>Trichophyton</i>			<i>Phialophora</i>	
			<i>Paecilomyces</i>			<i>Cladophialophora</i>	
			<i>Madurella</i>			<i>Cladosporium</i>	
Parasitic signatures	<i>Hymenolepis</i>	<i>Echinococcus</i>	<i>Dipylidium</i>		<i>Prosthodendrium</i>	<i>Toxocara</i>	<i>Diphylobothrium</i>
	<i>Centrocestus</i>		<i>Anisakis</i>		<i>Contracaecum</i>		<i>Strongyloides</i>
	<i>Trichinella</i>		<i>Echinostoma</i>				

5 ii.) Bacterial signatures associated with OSCC

FIGs. 2A-2F and Table 3 show the variety of bacterial signatures found in OSCC, matched and non-matched control samples. These include Proteobacteria, Actinobacteria, Firmicutes, Bacteroidetes, and Fusobacteria. There were differences in gram-positive and gram-negative microbiota in OSCCs compared to control samples. In the non-matched controls about 55% of the organisms were gram-negative compared to 40% in the matched controls and 49% in the OSCC samples. 43%, 50% and 36% of the bacterial

agents were gram-positive in the OSCCs, matched and non-matched controls, respectively (FIG. 2A). Interestingly, Proteobacteria, one of the major gram negative phylum (includes *Escherichia*, *Vibrio* and *Salmonella*) was much more pronounced in OSCCs at 41% compared to matched and non-matched control at 25% and 18%, respectively (FIG. 2A). The Bacteroides were more pronounced in the non-matched controls at 27% compared to 4% and 5% in the OSCC and matched controls, respectively (FIG. 2A). The gram-positive phylum Actinobacteria were similar across all samples at 31%, 30% and 36% (FIG. 2A). The Firmicutes phylum of gram-positive bacteria was more pronounced in the matched controls at 35% compared to 24% and 18% in OSCCs and non-matched controls, respectively (FIG. 2A). Among the Proteobacteria that were detected in the OSCC samples were the genera of *Escherichia*, *Brevundimonas*, *Aeromonas*, *Bordetella*, *Comamonas*, *Alcaligenes*, *Caulobacter*, *Acinetobacter*, *Citrobacter*, *Sphingomonas*, *Plesiomonas*, *Actinobacillus*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateruia* and *Cardiobacterium*. Proteobacteria *Brevundimonas* and Actinobacteria *Mobiluncus* were the most prevalent (98%) among the OSCC samples; followed by the generas of *Frateruia*, *Caulobacter*, *Actinomyces*, and *Aeromonas*, that were detected in about 90% of the cancer cases. Probes of the Actinobacterias (*Arcanobacterium*, *Mobiluncus*, *Actinomyces*, *Rothia*, *Propionibacterium*, *Peptoniphilus*, *Mycobacterium*) detected in the OSCC samples had high hybridization signals (Table 4), the highest being that of *Arcanobacterium* (FIG. 2B). Probes of Proteobacteria generas *Escherichia* and *Brevundimonas* were detected in 88% and 98% of cancer cases, respectively with high hybridization signals (FIGs. 2B-2E, Table 4). The other Proteobacteria generas detected in cancer cases showed low to moderate hybridization signals, but interestingly, they were highly prevalent (>75%), except for the generas *Serratia*, *Plesiomonas*, *Edwardsiella*, *Citrobacter* (46-62%) (FIGs. 2B-2E).

The matched control samples shared some of the bacterial signatures that were detected in the cancer samples along with other bacterial signatures of normal oral flora (FIGs. 2B-2D). Table 3 shows the list of bacterial genera detected and shared among the cancer, matched and non-matched control samples. Bacterial signatures of the genera *Actinomyces* were detected with the highest prevalence (100%) and hybridization signal intensity in the matched controls (FIGs. 2B-2D). 8 of the 14 bacterial genera detected in matched controls were also detected in the OSCC samples (Table 3). They represented the genera of *Arcanobacterium*, *Actinomyces*, *Aeromonas*, *Bordetella*, *Aerococcus*, *Pediococcus*, *Acinetobacter*, and *Veillonella* (Table 3). Among the non-matched control

samples, bacterial signatures of generas *Mobiluncus* and *Mycobacterium* were detected in all and probes of generas *Citrobacter* and *Mycobacterium* showed high hybridization signal (FIGs. 2B-2D). Importantly, most of the bacterial signatures detected in the control samples are of the normal oral flora.

5 The Venn diagram (FIG. 2F) summarizes findings showing that bacterial signatures representing 13 genera are found to be specifically associated with OSCC samples and not with the matched or non-matched controls. These are the Proteobacteria including *Escherichia*, *Brevundimonas*, *Comamonas*, *Alcaligenes*, *Caulobacter*, *Plesiomonas*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateuria* and *Cardiobacterium*; and
10 the Actinobacteria genus *Rothia*, as well as the Bacteroidetes genus *Peptoniphilus*. As in the case of the viruses, the bacterial microbial signatures showed a significant divergence in the OSCC when compared to the normal signatures and were more robust.

15 Table 4. Significant detection of the probes of micro-organisms in cancer compared to the matched (MC) and non-matched control (NC) samples. Weighted score sum of the hybridization signals of all the probes of an organism was calculated in cancer and controls, and significance (p -value<0.05) was calculated using one sided t-tests.

Types	Phyla	Family/Genera	Hybridization Signal (weighted score sum)			p-value	
			Cancer	MC	NC	Cancer vs. MC	Cancer vs. NC
Viruses		HPVs (2, 6a, 1, 18, 26, 34)	195426486	8029	6839	2.31E-10	2.3E-10
		HPV16	191895193	6468	2862	4.02E-10	3.98E-10
		Pooviridae	22066931	3151	11698	9.95E-06	3.59E-05
		Retroviridae	1656699	7618	78807	0.002287	0.291564
		Polyomaviridae	847250	1153	89753	8.59E-05	0.427587
		Herpesviridae	4863563	9081	74503	7.57E-07	9.3E-05
		Reoviridae	101004	44	251	3.26E-09	7.59E-09
		Orthomyxoviridae	266120	235	159	1.4E-05	1.3E-05
Bacteria	Actinobacteria	Actinobacterium	4313000	0	81	8.44E-06	8.48E-06
	Actinobacteria	Molobius	2664602	58	229	2.63E-06	2.86E-06
	Actinobacteria	Actinomycetes	2095762	310	0	1.48E-05	1.41E-05
	Actinobacteria	Rothia	1253759	0	0	6.62E-06	6.62E-06
	Actinobacteria	Propionibacterium	1192264	0	75	4.33E-06	4.4E-06
	Actinobacteria	Mycobacterium	532030	1186	5901	0.000114	0.000808
	Proteobacteria	Brevundimonas	662587	0	0	0.007274	0.007274
	Proteobacteria	Cardiobacterium	53035	217	0	0.002417	0.001331
	Proteobacteria	Aeromonas	412231	4524	478	0.01211	0.00984
	Proteobacteria	Bordetella	581307	169	70	0.005886	0.005212
	Proteobacteria	Comamonas	372304	2567	483	0.000375	4.83E-05
	Proteobacteria	Alcaligenes	347736	1782	186	0.008933	0.005065
	Proteobacteria	Caulobacter	327628	37	0	0.002719	0.002682
	Proteobacteria	Acinetobacter	297486	168	248	0.012434	0.012718
	Proteobacteria	Citrobacter	231427	0	3448	0.030438	0.089252
	Proteobacteria	Sphingomonas	206703	0	112	0.02798	0.028068
	Proteobacteria	Plesiomonas	117585	31	4521	0.053936	0.311419
	Proteobacteria	Actinobacillus	115450	0	0	0.365984	3.95E-06
	Proteobacteria	Serratia	111747	0	2046	0.039465	0.125478
	Proteobacteria	Edwardsiella	81383	0	0	0.046451	0.045451
	Proteobacteria	Haemophilus	45838	0	37	0.015341	0.016189
	Proteobacteria	Frutescens	16590	0	36	0.006651	0.006758
	Proteobacteria	Escherichia	2278254	0	0	4.41E-06	4.41E-06
	Firmicutes	Peptoniphilus	846305	0	0	0.000918	0.000318
	Firmicutes	Aerococcus	338497	2586	68	0.003597	0.000795
	Firmicutes	Pedococcus	521750	288	72	0.000116	0.000101
	Firmicutes	Peptostreptococcus	65757	0	0	0.061028	0.060022
	Firmicutes	Veillonella	53035	217	0	0.005775	0.003531
	Firmicutes	Streptococcus	50613	228	379	0.03941	0.058804
	Bacteroidetes	Prevotella	290058	275	838	0.012998	0.012614
	Bacteroidetes	Sphingobacterium	352917	187	110	0.000286	0.000275
Fungi		Fonsecaea	17341168	27419	62314	3.1E-20	3.16E-18
		Rhodospirillum	12912140	0	0	2.1E-19	2.1E-19
		Cladophialaphora	11366448	11539	25248	7.71E-17	1.64E-16
		Geotrichum	8862809	0	0	1.42E-11	1.42E-11
		Malassezia	7033309	87	4425	5.87E-13	6.64E-10
		Cladosporium	4991102	185123	961631	0.052290	0.401344
		Pleurotheca	2163898	2504	21859	3.11E-16	8.26E-14
		Pneumocystis	1210215	0	0	9.75E-08	9.75E-08
		Abidia	288315	235	216	5.87E-13	5.67E-13
Parasite		Phialophora	74254	0	372	5.75E-10	5.01E-09
		Hymenolepis	26463750	0	0	3.15E-29	3.15E-29
		Centrocestus	19026989	0	0	3.40901E-29	3.40901E-29
		Dipylidium	16436583	2402	85928	1.38067E-19	1.78333E-17
		Prosthodendrium	10743239	16758	37938	2.52E-18	9.51E-18
		Trichinella	1814992	0	0	5.19E-09	5.19E-09
		Contracaecum	906180	0	0	1.19E-16	1.19E-16
Parasite		Toxoplasma	260321	0	0	6.87E-15	6.87E-15

iii.) Fungal signatures associated with OSCC

Among the fungal signatures detected in the OSCC samples were those typically seen in the normal oral flora as well as those that are opportunistic infectious fungi.

Molecular signatures of *Fonsecaea*, *Malassezia*, *Pleistophora*, *Rhodotorula*, *Cladophialophora* and *Cladosporium* were detected in all the OSCC samples screened, *Pneumocystis* was detected in 93% of the cancer samples, and signatures of *Geotrichum*, *Phialophora*, *Absidia* and *Prevotella* were detected in >75% of the cancer cases screened (FIG. 3A). Probes of *Fonsecaea*, *Rhodotorula*, *Cladophialophora*, *Geotrichum* and *Malassezia* were detected in the OSCC samples with high hybridization signal intensity, the highest being that of *Fonsecaea* (FIGs. 3A and 3D, Table 4).

For the control samples screened, the matched controls detected some of the common oral flora along with some fungal signatures that were detected in the cancer samples. All the matched control samples significantly detected probes of *Phialophora*, *Cladosporium*, *Fonsecaea*, *Alternaria* and *Cladophialophora* (FIG. 3B). Except for the probes of *Alternaria*, all others mentioned above were detected in the matched control samples with high hybridization signal intensity (FIG. 3B). Probes of *Absidia* were detected with low hybridization signal intensity in 75% of the matched control samples screened. Among the fungal signatures detected in the non-matched control samples, signatures of *Cladosporium*, *Phialophora*, *Cladophialophora*, *Piedraia*, *Pleistophora* and *Alternaria* were detected in all (FIG. 3C), with high hybridization signals except for the probes of *Alternaria* (FIGs. 3C-3D). The probes of *Cladosporium*, a common oral flora were detected with the highest hybridization signal intensity in both the matched and non-matched control samples, and were also detected at similar intensity in the cancer samples ($p>0.05$) (FIGs. 3A-3D, Table 4).

The Venn diagram shows that three fungal signatures. *Rhodotorula*, *Geotrichum* and *Pneumocystis*, were associated only with OSCCs (Table 3, FIG. 3E). Again, a significant change in the fungal biome of OSCC was observed when compared to control oral samples.

iv.) Parasitic signatures associated with OSCC

Distinct molecular signatures for parasites were detected in OSCCs (FIGs. 3F and 3I, Table 4). Probes from 28S and/or 18S rRNA of *Hymenolepis*, *Centrocestus* and *Prosthodendrium* were detected in all the OSCC samples with very high hybridization signal (FIGs. 3F and 3I, Table 4). Probes of *Contracaecum*, *Dipylidium*, *Trichinella* and

Toxocara were detected in >95% of the cancer samples with moderate hybridization signal intensity (FIGs. 3F and 3I, Table 4).

Signatures for *Toxocara*, which were detected in OSCC samples, were also detected in 50% of the matched control samples screened, with lower hybridization signals along with parasitic signatures of *Strongyloides* and *Diphyllobothrium*, which were detected in 75% and 50% of the matched control samples respectively (FIGs. 3G and 3I). In the non-matched control samples, probes of *Dipylidium* and *Prosthodendrium* were detected with high hybridization signal intensity in all the samples screened. The hybridization signals of *Prosthodendrium* in the non-matched controls were significantly lower than that in the OSCC samples (Table 4). Parasitic signatures of *Toxocara* were also detected in all the non-matched control samples screened with moderate hybridization signal intensity, along with the probes of *Diphyllobothrium*, and *Strongyloides* detected in 75% of the non-matched controls (FIGs. 3H and 3I).

The Venn diagram in FIG. 3J summarizes the findings of parasitic signature associations with cancer and control samples. Molecular signatures of *Hymenolepis*, *Centrocestus* and *Trichinella* were found to be associated only with OSCC and not with the controls. Signatures of *Echinococcus* was found to be associated only with matched control samples and that of *Anisakis* and *Echinostoma* was found to be associated only with non-matched control samples. Thus distinct signatures differentiate cancer, matched controls and non-matched controls.

Example 2: Hierarchical clustering of OSCC samples based on detection of microbial signatures

Hierarchical clustering was done based on the detection of the microbial signatures in the 100 OSCC samples. Signature of *Cladosporium* was ignored as it was not significantly detected in the cancer samples compared to the controls. Using hierarchical clustering analysis R program, the OSCC samples fell into 2 major groups (A and B) based on specific microbiome (FIG. 4A). Molecular signatures for HPV16 were detected in the 2 major groups identified (FIG. 4A). Apart from HPV16 probes, group A OSCC samples also showed signatures of other viral probes, primarily belonging to Orthomyxoviridae and Reoviridae (FIG. 4A). The bacterial signatures were broadly detected in group A samples, compared to the sporadic bacterial signatures detected in group B samples with *Rothia* and *Mobiluncus* detected in both A and B groups. Both group A and B had high detection of some fungal and parasitic signatures except for the

parasitic signature *Trichinella* which was either absent or sporadically detected in group B while detected in almost all the group A samples. Thus, higher detection of viral, bacterial and parasitic signatures of *Trichinella*, was observed in group A OSCC samples compared to group B. In group A samples, some had higher hybridization signal for the bacterial and parasitic probes (Sub-group A1) than the rest of the samples of group A (Sub-group A2). Each of the sub-group is again sub-clustered based on the detection of certain probes belonging to Retroviridae, Poxviridae and Polyomaviridae. In group B, some samples (Sub-group B1) had a lower level of detection of the molecular signatures compared to the majority of group B samples (Sub-group B2). Samples in Sub-group B2 are further clustered based on the low detection of bacterial signatures in some of them. However, both group A and group B OSCC samples were positive for bacterial signatures of *Frateriuria*, *Mobiluncus*, fungal signatures of *Cladophialophora*, *Fonsecaea* and *Rhodotorula*, and all the parasitic signatures except for *Trichinella* in addition to HPV16 signatures.

Clustering of the OSCC samples was visualized using NBClust software (FIG. 4B). Two distinct clusters, clusters 1 and 2, were observed similar to the one described above. While there were no significant differences between the two clusters for the signatures of HPV 6b, HPV 16, HPV 26, HHV 8, HHV 6B, HHV 5, retroviral signatures, certain pox viral signatures, parapox viral signatures and polyoma viral signatures, there were significant differences in the detection of some of the viral and all the bacterial, fungal and parasitic signatures between the two clusters, cluster 1 having higher detection than 2. Among the viral signatures were signatures of Orthomyxoviridae, Reoviridae, HPV 34, HHV 6A, Mouse mammary tumor virus-like (MMTV-like) and certain poxvirus that were detected significantly higher in cluster 1 than in cluster 2.

Additional analyses using a topological approach represented data by grouping cases with similar detection for viral and microbial signatures into nodes, and connecting those nodes by an edge if the corresponding node have detection pattern in common to the first node (FIG. 4C). Topological analysis visualized all the OSCC cases into two clusters, 'Group a' and 'Group b', along with certain cases that did not have common detection pattern (ungrouped or singletons) (FIG. 4C). The clusters were characterized by detection of microbial signature patterns that may provide clues as to the stage of the disease. The nodes were colored based on HPV16 detection. The two major groups a and b showed significant differences in detection of certain micro-organisms between them, comprising significant higher detection of certain bacterial signatures (*Actinobacillus*,

Acinetobacter, Actinomyces, Aerococcus, Aeromonas, Alcaligenes, Arcanobacterium, Bordetella, Brevundimonas, Cardiobacterium, Citrobacter, Comamonas, Edwardsiella, Escherichia, Frateuria, Haemophilus, Mobiluncus, Mycobacterium, Pediococcus, Peptoniphilus, Peptostreptococcus, Plesiomonas, Prevotella, Propionibacterium, Rothia, Serratia, Sphingobacterium, Sphingomonas, Streptococcus, Veillonella), parasitic signatures of *Trichinella, Contracaecum, Prosthodendrium* and *Toxocara*, fungal signature of *Pleistophora, Malassezia, Cladophialophora, Fonsecaea, Absidia, Rhodotorula, Pneumocystis* and *Phialophora* and viral signatures of Orthomyxoviridae and Reoviridae in 'Group b' than 'Group a'. However, significantly higher detection of HPV16 was found in 'Group a' compared to 'Group b'. The samples within 'Group b' ranged from having no to very high HPV16 signals. Out of the 6 singletons or un-grouped OSCC samples, 1 had all the signatures detected in the OSCC samples, the others had no detection of viral signatures for Herpesviridae, Retroviridae, Poxviridae, Polyomaviridae and very less detection of parasitic *Trichinella*. The ungrouped samples had significantly lower detection of the majority of bacterial signatures along with fungal signatures of *Malassezia, Geotrichum, Pneumocystis, Fonsecaea, Absidia, Cladophialophora, Phialophora, Rhodotorula*, and *Pleistophora*, viral signatures of Reoviridae, Orthomyxoviridae, Herpesviridae, Retroviridae (MMTV-like), Poxviridae, Papillomaviridae (Human Papillomavirus 1 and 34), and parasitic signatures of *Trichinella, Toxocara, Contracaecum, Centrocestus, Hymenolepis, Prosthodendrium, Dipylidium* compared to the rest of the samples clustered in 'group a' and 'group b'.

Example 3: Validation of PathoChip results of OSCC by probe capture and next generation sequencing

Conserved and sequence specific probes detected as positive in the PathoChip screen were identified and used to enrich for the pathogenic targets from amplified genomic DNA/cDNA pool of the OSCC samples. The probes were conjugated with biotin, and streptavidin beads were used to capture the biotinylated probe-DNA/cDNA complexes from the amplified genomic DNA/cDNA pool of the OSCC samples. The enriched targets were subjected to MiSeq, and the sequence reads were aligned to the PathoChip metagenome (obtained by concatemerization of sequences of accessions used in the PathoChip) (Baldwin et al., (2014) mBio 5, e01714-01714). The results showed that the sequence reads clustered around the genomic locations of the probes (FIGs. 5A-5D, FIGs. 7A-7E). However, regions of the target genome outside the capture probe

locations were also detected (eg. sequence reads of *Trichinella papuae*, FIGs. 5A-5D).

Four HPV16 specific capture probes from the E1, E2/E4 and L1 genes, used in the reaction pulled out genomic sequence of HPV16, that aligned with most of the HPV16 E1, E2/E4, L2 and L1 genes (FIGs. 5A-5D). The conserved probe for Polyomavirus from the regulatory region (182-226bp of NC_001699.1) and the specific probes of JC Polyomavirus were used to enrich genomic regions of late mRNA and VP2/VP3 and VP1 of the virus, respectively from the cancer samples. Importantly, the captured sequences were found to align at the capture probe regions as expected (FIGs. 5A-5D). Capture probe designed from 16S rRNA region of the bacteria *Rothia*, captured most of the genomic sequence of the bacteria. Thus the sequence reads aligned not only with the capture probe region, but also extended across the genome of the bacteria (FIGs. 5A-5D). Other bacterial sequence reads aligned with their respective capture probe regions, further validating the PathoChip screen results (FIGs. 7C-7D). Sequence reads of fungi were also found to align with sequences at or adjacent to their respective capture probe regions (FIGs. 5A-5D, and FIG. 7E). For example, 1432 sequence reads of *Pneumocystis*, aligned at the capture probe location in their genome as well as outside of it (FIGs. 5A-5D). 2057 sequence reads of *Pleistophora*, aligned at the capture probe location of its genome (FIG. 7E). High sequence reads (>4000) were obtained for the skin fungus *Malassezia* (FIG. 7E).

Sequence alignments of the reads of parasites *Hymenolepis* and *Trichinella* also extended beyond their respective capture probe locations, thus further confirming the presence of nucleic acids from these micro-organisms in the cancer samples. The number of reads were extremely high for *Trichinella* (>20,000) and *Prosthodendrium* (>9000) (FIGs. 5A-5D, FIG. 7F-7G).

These results strongly confirmed the presence of these microbes. Therefore by using probes that were detected positive in the PathoChip screen, as “bait”, it was possible to capture the microbial sequences from the OSCC tissue samples, further validating the results of the PathoChip screen.

Example 4: Insertions of a broad range of microbial genomic fragments identified in the host human chromosomes

Next, it was investigated whether or not microbial genetic elements detected in the samples had integrated into the human genomes. Such integrations may have pathogenic implications especially as it relates to cancer. This was examined by analyzing the

sequences captured by the conserved and sequence specific biotinylated probes for the microbes detected in the PathoChip screen with human genomic sequences that would suggest insertion for viral as well as other microbial agents. The analysis detected numerous microbial genomic insertional sites within the human chromosomes (FIGs. 6A-6F). 38,019 bacterial insertional sites, 125 fungal genomic insertional sites, 508 parasitic insertional sites and 79 viral insertional sites were identified (FIGs. 6B-6D). To simplify the data, reads >20 were focused on for bacterial, fungal and parasitic sequence fusion with host genome, and FIG. 6B represents the data in a Circos plot highlighting the insertions. Although the number of viral insertions were lower compared to the other microbial insertions, the 79 insertional sites for JC and HPV16 were represented on the Circos plot. The Circos plot shows insertions going from the inner concentric circle to outer circle in the order of fungus, JC Polyomavirus, HPV16, parasites and bacteria. This is then comprehensively shown with its represented colors in the outermost circle with all insertions (FIG. 6B). A karyotype plot also shows the representative bacterial and fungal, parasitic and viral insertional sites in each chromosome (FIGs. 6C-6D). The sites with >20 reads for bacterial, fungal and parasitic genomic insertions were considered, and all the viral integration sites were included. Bacterial insertions are shown for all chromosomes in FIG. 6C. The number of insertions for each chromosomes are shown to the left of each chromosome number. Interestingly, chromosomes 1, 2, 3, 6 and 8 showed over 50 insertions each, and the Y chromosome having the least insertions (FIG. 6C). Notably, the mitochondrial chromosome also showed 4 insertions in this analysis (FIG. 6C). Insertions for viral, fungal and parasitic agents although less frequent were seen in all chromosomes (FIG. 6D). Chromosomes 2, 5, 6, 10, 17 and 19 had more insertions and chromosome 9 only had 1 insertion (FIG. 6D). Interestingly, the sites for microbial insertions were exonic, intronic, upstream/downstream and at the UTRs or at the ncRNA region of host genes.

These data, based on a very sensitive analytical approach, suggest that there is a far greater intimacy between human and microbial genomes, at the level of integration, than previously observed.

i.) Viral integrations into host genomes

Genomic elements of HPV16 and JC Polyomavirus were found to be integrated in the human chromosomes of OSCC cells. For HPV16, 7 insertion sites were detected in chromosome 17 (chr17), 6 in chromosome 5, and between 1 and 4 in other the

chromosomes except for chromosomes 13, X and Y (FIG. 6A). The genomic fragment of HPV16 that was identified most frequently integrated in the human genome (59%) was at the genomic co-ordinates 4,172 (based on accession NC_001526.2), which is located around the polyA sequence of E5 gene. Additionally, 12% of the HPV16 integrations included HPV genomic co-ordinates 3,393-3,425 in coding sequence of the E1 gene; 10% at co-ordinates 7,206-7,627 near the polyA sequence of the L1 gene, 9% in the coding region of L1 from co-ordinates 6,030-6,715; as well as lower percentage integrations in the coding sequences of the E4 gene (3,358-3,394), the E2 gene (3,393-3,425), the E7 gene (674-693) and the L2 gene (5,201-5,221) (FIG. 6A).

JC polyoma (JC) viral genomic integration was observed in human chromosomes 1, 2, 5, 6, 10, 13, 15, 16, 17, 19, X and Y (FIG. 6A). The JC viral co-ordinates (accession NC_001699), that were integrated in the host chromosomes were mostly at the region of large T antigen (co-ordinates 2,623-2,653) (Frisque et al., (1984) J Virol 51, 458-469) that accounted for 62% of the JC polyomaviral genomic insertions detected in the study. This is followed by regions around the JVgp1 (co-ordinates 191-460 with 46%), VP1 (co-ordinates 1,479-2,361 with 38%) and VP2/VP3 (co-ordinates 1,278-1,318 at 15%), respectively (FIG. 6A).

Viral integrations were detected at many genomic regions. Examples of these insertions are represented in FIG. 6E, FIG. 8, and FIGs. 9A-9B. Viral integration sites were mostly intronic, followed by intergenic sites for integration. Viral genomic integrations were also observed upstream or downstream and at 3'UTR of genes as well as at the ncRNA intronic regions (FIGs. 6A-6F, FIG. 8). HPV16 genomic hotspots for integrations at co-ordinates 4,188-4,243 (Seedorf et al., (1985) Virology 145, 181-185) detected in this study was integrated mostly at the intronic regions (53%) of genes LAMA3, ATXN10, INADL, ABCA10, EVC2, WDR89, CADPS2, HAUS6, EPHA6, FAM179B, COL14A1, MRPS27, FUCA2, ADAMTS12, TRIOBP, CSMD1, KCNQ1, and at the intronic ncRNA gene (6%) of the FAM35BP gene (FIG. 6E and FIG. 8). Additional hotspots for integrations were seen at intergenic regions (15%), downstream (9%) of the genes NACAP1, GUCA2A/GUCA2B, RSPH1; upstream (12%) of genes IL12RB2, LOC388436, LOC79999, FCHO1, MRPL52, SLC7A7, while some were integrated between two genes (6% intergenic), upstream of gene SPECC1 and downstream of gene CCDC144CP, as well as at the upstream of gene SSTR3 and downstream of gene RAC2. 4 of the 7 HPV16 E1 integration sites detected in the study were in the intronic region of the genes SLC13A3, DLGAP1 and CCDC155 (FIG. 8); 1

was at the intronic region of ncRNA LOC100288637, and 2 at intergenic regions. HPV16 E2 and E4 DNA elements were also found integrated at the intronic region of LOC102724957, while 2 other integration sites for E4, and 1 site for E7 fragments were intergenic (FIG. 8). Human genomic integration sites for L1 fragments were also detected at intronic regions of genes PAFAH1B1 (2/5) and ncRNA LOC100506207 (2/5), and an intergenic region (1/5). L2 fragments were found integrated within the intronic region of gene SSH2. Also, the genomic DNA located at the PolyA sequence of HPV16 L1 were integrated at the intronic regions of gene DEPDC4, and within the intergenic regions and at the 3'UTR region of the MKLN1 gene. Thus, genome wide integrations of the HPV genome were observed at numerous sites within the host chromosome.

The JC Polyoma virus integration sites for large T antigenic regions were found to be within the intronic regions of host genes CMTR1 and ME1 on chromosome 6; gene CPO of chromosome 2, and within intergenic regions of chromosomes 1, 2 and 3. Elements of the VP1 ORF were also found to be integrated in the intergenic regions of human chromosome 10, about 41Kb downstream of the lncRNA gene SFTA1P which is known to be upregulated in carcinoma (Zhao et al., (2014) Sci Rep 4, 6591.) (FIG. 8). It was also seen integrated upstream of the ABCA9 gene in chromosome 17, known to be associated with melanoma (Hedditch et al., (2014) J Natl Cancer Inst 106(7). pii, djul49), and in the 3'UTR of the epigenetic regulator gene MECP2 located in chromosome X (Figure 6E). Genomic elements of VP2 and VP3 integration sites in the intronic region of the FAM13B gene on chromosome 5 and the PCCA gene on chromosome 13, respectively were also detected. Agnoprotein Jvcp1 DNA elements were detected in the intronic regions of the MSH3 gene on chromosome 5, and the PHLSB3 gene on chr19. Late mRNA transcripts 191-253 (NC_001699.1) integration sites were seen at the intergenic regions of chromosome 16, 97Kb downstream of the NPIPA7 gene, also at 99Kb upstream of the NPIPA5 gene, and at intronic region of the SSG5 gene on chromosome 15. FIGs. 9A-9B show the various integration sites for JCV in the human genome, again these insertions could affect gene expression in ways that would promote oncogenesis.

ii.) Integration of bacterial genomic elements in host chromosomes

Numerous insertional sites were observed for bacterial genomic fragments in exonic, intronic, intergenic, 3' and 5' UTR region, upstream and downstream regions of numerous genes of human chromosomes (FIGs. 6A-6F). About 890 bacterial sequence

insertional sites were detected at different exons of human chromosomes; 172 insertional sites for species of *Mycobacterium*, 174 for *Aeromonas*, 182 for *Bordetella*, 53 for *Sphingomonas*, 39 for *Acinetobacter*, 31 for *Citrobacter*, 91 for *Escherichia*, 18 for *Porphyromonas*, 32 for *Haemophilus*, 43 for *Campylobacter*, 21 for *Pediococcus* and 1 for *Brevundimonas* (FIGs. 6B-6E). Several particularly interesting inserts within human gene related to cancer are shown in FIGs. 9A-9B. For example, detected were:

Mycobacterium (NC_008595.1) genomic elements 24065-24105 insertions at the exonic regions of the tumor suppressor ADAMTSL1 gene on chromosome 9; *Aeromonas* (NC_008570.1) genomic elements insertion sites in the exon of the RASSF5 (1q32.1), a member of Ras association domain family that functions as a tumor suppressor and shown to be inactivated in a variety of cancers (van der Weyden and Adams, (2007) Biochim Biophys Acta 1776, 58-85); *Sphingomonas* (NC_009511.1) genomic elements insertions in exonic regions of chromatin re-modelling gene SRCAP in chromosome 16; *Bordetella* (NC_002929.2) genomic insertional site within the exon of the proto-oncogene WNT3 on chromosome 17 (FIGs. 6A-6F); *Escherichia coli* (NC_013008.1) genomic insertional site at the end of the SMURF2 gene, a tumor suppressor and regulator of the G1/S checkpoint (Blank et al., (2012) Nat Med 18, 227-234) (Fushimi et al., (2008) Radiother Oncol 89, 237-244).

Apart from the exonic insertional sites, numerous sites (514) were detected at the 3' and 5' UTR regions of genes. For example, *Bordetella* genomic insertions were seen at the 5'UTR of C1orf162, *Aeromonas* insertions at the 3'UTR of the SEC14L4 gene; and *Campylobacter* insertions at the 3'UTR of the COL1A1 gene.

The numerous bacterial insertional sites at the intronic region of genes were supported by high bacterial-genomic fusion sequence reads; *Bordetella* genomic insertions within the SLC9A9 gene, *Escherichia* genomic insertions within the CASP9 gene; *Brevundimonas* insertions within the RIC8B gene; *Mycobacterium* genomic insertions in the MAP3K1 gene; *Pediococcus* genomic insertions within the LYPD6B gene.

iii.) Integration of fungal DNA elements in host chromosomes

Genomic fragments of the fungal OSCC flora were also detected at 125 insertion sites in the intergenic (46%), intronic (42%), upstream or downstream of genes or, ncRNA but not exonic regions in the human chromosomes (FIGs. 6B-6E). 25 insertional sites were observed for the genomic fragments of *Malassezia* in the host genome, 24 each

for *Rhodotorula* and *Pleistophora*, 19 for *Absidia*, 15 for *Geotrichum*, 7 for *Cladophialophora*, 6 for *Phialophora*, 3 for *Fonsecaea* and 1 for *Pneumocystis*. 18S rRNA genomic fragment of *Geotrichum* (AB000641.1) was detected at the intergenic regions of chromosome 4, 560Kb upstream of the GABRG1 gene; Genomic fragments of *Pleistophora* (AJ295327.1) at the intronic regions of the negative regulator of tumor suppressor ITCH gene on chromosome 20 (FIG. 8) and at the intronic region of tumor suppressor MAGI1 on chromosome 3, that of *Phialophora* (AF050282.1) in the intronic region of the E3 ubiquitin protein ligase ZNRF2 on chr7, and that of *Rhodotorula* (AF444490.1) at the intronic region of the CADPS2 gene of chromosome 7.

iv.) Genomic insertions of parasitic DNA in host chromosomes

508 genomic insertional sites for the OSCC associated parasites were detected in human chromosomes (FIG. 6D). The majority of these insertions were at intergenic (202 sites) or intronic (198) regions. Insertional sites were also detected upstream and downstream, at the splice site, 3 and 5' UTR regions of certain genes, while only 2 insertional sites were detected in the exons. *Strongyloides* (AB272235.1) sequence insertions were seen in the exonic region of the MAPK suppressor gene ZNF383 on chromosome 19 and *Contracaecum* (AF411204.1) sequence insertion sites were seen at the exonic region of the RHD gene on chromosome 1. Sequences of a parasite were found to have multiple insertional sites within the host chromosomes. A large number of sequence reads were obtained for the 28S rRNA genomic fragment of *Prosthodendrium* (AF151921.1)-human genomic fusion at the intergenic region of chromosome 8, 37Kb upstream of the proto-oncogene Lyn. *Trichinella* (AY851263.1) sequence insertion sites were detected on chromosome 17 at the intronic region of the AKAP1 gene which is known to be associated with epithelial cancers (Sotgia et al., (2012) Cell Cycle 11, 4390-4401) and also within the intergenic region of chromosome 10, 353KB upstream of the NRG3 gene. Sequence of another species of *Trichinella* (AY851262.1) was also detected with high sequence reads at the intronic region of the EPS15L1 gene on chromosome 19; *Strongyloides* (AB272235.1) sequences at the intergenic region of chromosome 13 downstream of the SLC10A2 gene, and at the intronic site of the LNP1 gene on chromosome 3 were evidenced with high sequence depth. Sequence of *Hymenolepis* (AF124475.1) was found downstream of the MIR3648 gene on chromosome 21; *Diphyllobothrium* sequences at the ncRNA ANKRD30BL gene and in the intergenic region of chromosome 9 about 106Kb upstream of the TRIM49B gene. Mutations in

ANKRD30BL are known to be associated with cancer (Weinhold et al., (2014) Nat Genet 46, 1160-1165); as well as that of TRIM49B, one of the RING type E3 ubiquitin ligase involved in deregulation of tumor suppressors (Hatakeyama, (2011) Nat Rev Cancer 11, 792-804). *Echinococcus* (EGU27015) sequence insertion sites were observed at the intronic region of the chromatin re-modelling gene ATRX on the X chromosome, and mutation of which was shown to be associated with cancer (Lovejoy et al., (2012) PLoS Genet 8, e1002772). Besides the higher number of reads for parasite-host fusion regions, there were also lower reads for a number of other parasitic insertions. Nevertheless, the insertion sites are important as they might contribute to cancer. For example, a smaller number of reads were obtained for *Echinococcus* (NC_009460.1) DNA elements detected 21Kbp upstream of tumor suppressor FGFR2 gene (FIG. 8), and mutation or abnormal expression is known to lead to cancer development. *Prosthodendrium* (AF151921) insertion on chromosome chr17 at the intronic region of the de-ubiquitinating enzyme encoding gene USP32 (FIGs. 6A-6F). *Strongyloides* (NC_005143.1) 28s rRNA genomic fragment insertion sites were noted on chromosome 17 and at the intronic region of the tumor suppressor SPECC1 gene (FIG. 6E).

Thus, from these observations it is evident that numerous host genes may be deregulated by viral, bacterial, fungal and parasitic genomic integrations at different sites within the human genome. IPA analysis of some of the affected host genes showed that most of the host genes had a significant association (p -value= $7.17E-10$) with neoplasia of epithelial cells (FIG. 6F), thus suggesting that disruption of many of these genes can directly contribute to the development of OSCC.

Numerous genomic insertional sites for parasites were detected in the OSCC cell genomes (FIG. 6D). The majority of these insertions were at intergenic (202 sites) or intronic (198) regions. Insertional sites were also detected upstream and downstream, at the splice site, 3 and 5' UTR regions of certain genes, while only 2 insertional sites were detected in the exons. FIGs. 9A-9B highlights some of the integrations that may affect human genes involved in cancer.

The insertional data suggest that there may be far more integrations of viral and other microorganisms than previously expected, and IPA analysis of some of the affected host genes showed that they have a significant association with oncogenesis (p -value= $7.17E-10$) (FIG. 6F).

Other Embodiments

The recitation of a listing of elements in any definition of a variable herein includes definitions of that variable as any single element or combination (or subcombination) of listed elements. The recitation of an embodiment herein includes that
5 embodiment as any single embodiment or in combination with any other embodiments or portions thereof.

The disclosures of each and every patent, patent application, and publication cited herein are hereby incorporated herein by reference in their entirety. While this invention has been disclosed with reference to specific embodiments, it is apparent that other
10 embodiments and variations of this invention may be devised by others skilled in the art without departing from the true spirit and scope of the invention. The appended claims are intended to be construed to include all such embodiments and equivalent variations.

CLAIMS

What is claimed:

1. A method of detecting oral squamous cell carcinoma in a tumor tissue sample from a subject, the method comprising:

hybridizing a detectably-labeled nucleic acid from the tumor tissue sample to a PathoChip array to generate a first hybridization pattern;

hybridizing a detectably-labeled nucleic acid from a reference sample to a PathoChip array to generate a second hybridization pattern, wherein the reference sample is from an otherwise identical non-tumor tissue from a subject;

comparing the first and second hybridization patterns, wherein when the first hybridization pattern is substantially a microbial hybridization signature and the second hybridization pattern is substantially not a microbial hybridization signature, oral squamous cell carcinoma is detected in the tumor tissue sample.

2. The method of claim 1, wherein the microbial hybridization signature is generated by hybridization of the detectably-labeled nucleic acid from the tumor tissue sample to at least three nucleic acid probes on the PathoChip, wherein the probes are from microbes selected from the group consisting of Human papillomavirus 16 (HPV16), *Eshcherichia*, *Rothia*, *Peptoniphilus*, *Brevundimonas*, *Comamonas*, *Alcaligenes*, *Arcanobacterium*, *Actinomyces*, *Aeromonas*, *Bordetella*, *Aerococcus*, *Pediococcus*, *Caulobacter*, *Cardiobacterium*, *Plesiomonas*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateuria*, *Rhodotorula*, *Geotrichum*, *Pneumocystis*, *Hymenolepis*, *Centrocestus*, *Trichinella*, *Acinetobacter*, *Actinobacillus*, *Veillonella*, *Fonsecaea*, *Mobiluncus*, *Propionibacterium*, *Mycobacterium*, *Malassezia*, *Pleistophora*, *Prosthodendrium*, *Contracaecum*, *Toxocara*, Parvovirus 2, Human herpesvirus 6A, Adenovirus 1, Swine fever virus, JC polyomavirus, Moloney murine leukemia virus, Torque teno virus 2, Human papillomavirus 26, Tomato chlorotic leaf distortion virus, and *Prevotella*.

3. The method of claim 2, wherein the at least three nucleic acid probes are selected from the group consisting of SEQ ID NOS: 1-76.

4. A method of detecting oral squamous cell carcinoma in a tumor tissue sample from a subject, the method comprising:

hybridizing a detectably-labeled nucleic acid from the tumor tissue sample to a first microarray comprising at least three nucleic acid probes from microbes selected from the group consisting of Human papillomavirus 16 (HPV16), *Eshcherichia*, *Rothia*, *Peptoniphilus*, *Brevundimonas*, *Comamonas*, *Alcaligenes*, *Arcanobacterium*, *Actinomyces*, *Aeromonas*, *Bordetella*, *Aerococcus*, *Pediococcus*, *Caulobacter*, *Cardiobacterium*, *Plesiomonas*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateuria*, *Rhodotorula*, *Geotrichum*, *Pneumocystis*, *Hymenolepis*, *Centrocestus*, *Trichinella*, *Acinetobacter*, *Actinobacillus*, *Veillonella*, *Fonsecaea*, *Mobiluncus*, *Propionibacterium*, *Mycobacterium*, *Malassezia*, *Pleistophora*, *Prosthodendrium*, *Contracaecum*, *Toxocara*, Parvovirus 2, Human herpesvirus 6A, Adenovirus 1, Swine fever virus, JC polyomavirus, Moloney murine leukemia virus, Torque teno virus 2, Human papillomavirus 26, Tomato chlorotic leaf distortion virus, and *Prevotella* to generate a first hybridization pattern;

hybridizing a detectably-labeled nucleic acid from a reference sample to a second microarray comprising at least three nucleic acid probes from microbes selected from the group consisting of Human papillomavirus 16 (HPV16), *Eshcherichia*, *Rothia*, *Peptoniphilus*, *Brevundimonas*, *Comamonas*, *Alcaligenes*, *Arcanobacterium*, *Actinomyces*, *Aeromonas*, *Bordetella*, *Aerococcus*, *Pediococcus*, *Caulobacter*, *Cardiobacterium*, *Plesiomonas*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateuria*, *Rhodotorula*, *Geotrichum*, *Pneumocystis*, *Hymenolepis*, *Centrocestus*, *Trichinella*, *Acinetobacter*, *Actinobacillus*, *Veillonella*, *Fonsecaea*, *Mobiluncus*, *Propionibacterium*, *Mycobacterium*, *Malassezia*, *Pleistophora*, *Prosthodendrium*, *Contracaecum*, *Toxocara*, Parvovirus 2, Human herpesvirus 6A, Adenovirus 1, Swine fever virus, JC polyomavirus, Moloney murine leukemia virus, Torque teno virus 2, Human papillomavirus 26, Tomato chlorotic leaf distortion virus, and *Prevotella* to generate a second hybridization pattern, wherein the reference sample is from an otherwise identical non-tumor tissue from a subject;

comparing the first and second hybridization patterns, wherein when the first hybridization pattern is substantially a microbial hybridization signature and the second hybridization pattern is substantially not a microbial hybridization signature, oral squamous cell carcinoma is detected in the tumor tissue sample.

5. The method of claim 4, wherein the at least three nucleic acid probes are selected from the group consisting of SEQ ID NOS: 1-76.

6. The method of any one of claims 1-5, wherein the tumor tissue sample is selected from the group consisting of a biopsy, formalin-fixed, paraffin-embedded (FFPE) sample, or non-solid tumor.
7. The method of any one of claims 1-6, wherein the subject is human.
8. The method of any one of claims 1-7, wherein the detectably-labeled nucleic acid is labeled with a fluorophore, radioactive phosphate, biotin, or enzyme.
9. The method of claim 8, wherein the fluorophore is Cy3 or Cy5.
10. The method of any one of claims 1-9, further comprising wherein when oral squamous cell carcinoma is detected in the tumor tissue sample from a subject, the subject is provided with a treatment for oral squamous cell carcinoma.
11. The method of claim 10, wherein the treatment comprises surgery, chemotherapy, or radiotherapy.
12. A composition comprising at least three nucleic acid probes selected from the group consisting of SEQ ID NOS: 1-76.
13. A microarray comprising at least three nucleic acid probes selected from the group consisting of SEQ ID NOS: 1-76.
14. The microarray of claim 13, wherein the nucleic acid probes are selected from about 10 to about 30 microbes and comprise about 3 to about 5 probes per microbe.
15. A microarray comprising at least three nucleic acid probes selected from the group of microbes consisting of Human papillomavirus 16 (HPV16), *Eshcherichia*, *Rothia*, *Peptoniphilus*, *Brevundimonas*, *Comamonas*, *Alcaligenes*, *Arcanobacterium*, *Actinomyces*, *Aeromonas*, *Bordetella*, *Aerococcus*, *Pediococcus*, *Caulobacter*, *Cardiobacterium*, *Plesiomonas*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateuria*, *Rhodotorula*, *Geotrichum*, *Pneumocystis*, *Hymenolepis*, *Centrocestus*, *Trichinella*, *Acinetobacter*, *Actinobacillus*,

Veillonella, *Fonsecaea*, *Mobiluncus*, *Propionibacterium*, *Mycobacterium*, *Malassezia*, *Pleistophora*, *Prosthodendrium*, *Contracaecum*, *Toxocara*, Parvovirus 2, Human herpesvirus 6A, Adenovirus 1, Swine fever virus, JC polyomavirus, Moloney murine leukemia virus, Torque teno virus 2, Human papillomavirus 26, Tomato chlorotic leaf distortion virus, and *Prevotella*.

16. The microarray of any one of claims 13-15, wherein the microarray is a biochip, glass slide, bead, or paper.

17. A kit comprising at least three nucleic acid probes selected from the group consisting of SEQ ID NOS: 1-76, and instructional material for use thereof.

18. A kit comprising a microarray comprising at least three nucleic acid probes selected from the group consisting of SEQ ID NOS: 1-76, and instructional material for use thereof.

19. A kit comprising a microarray comprising at least three nucleic acid probes selected from the group of microbes consisting of Human papillomavirus 16 (HPV16), *Eshcherichia*, *Rothia*, *Peptoniphilus*, *Brevundimonas*, *Comamonas*, *Alcaligenes*, *Arcanobacterium*, *Actinomyces*, *Aeromonas*, *Bordetella*, *Aerococcus*, *Pediococcus*, *Caulobacter*, *Cardiobacterium*, *Plesiomonas*, *Serratia*, *Edwardsiella*, *Haemophilus*, *Frateuria*, *Rhodotorula*, *Geotrichum*, *Pneumocystis*, *Hymenolepis*, *Centrocestus*, *Trichinella*, *Acinetobacter*, *Actinobacillus*, *Veillonella*, *Fonsecaea*, *Mobiluncus*, *Propionibacterium*, *Mycobacterium*, *Malassezia*, *Pleistophora*, *Prosthodendrium*, *Contracaecum*, *Toxocara*, Parvovirus 2, Human herpesvirus 6A, Adenovirus 1, Swine fever virus, JC polyomavirus, Moloney murine leukemia virus, Torque teno virus 2, Human papillomavirus 26, Tomato chlorotic leaf distortion virus, and *Prevotella*, and instructional material for use thereof.

20. The kit of any one of claims 17-19, wherein the nucleic acid probes are selected from between about 10 to about 30 microbes and comprise about 3 to about 5 probes per microbe.

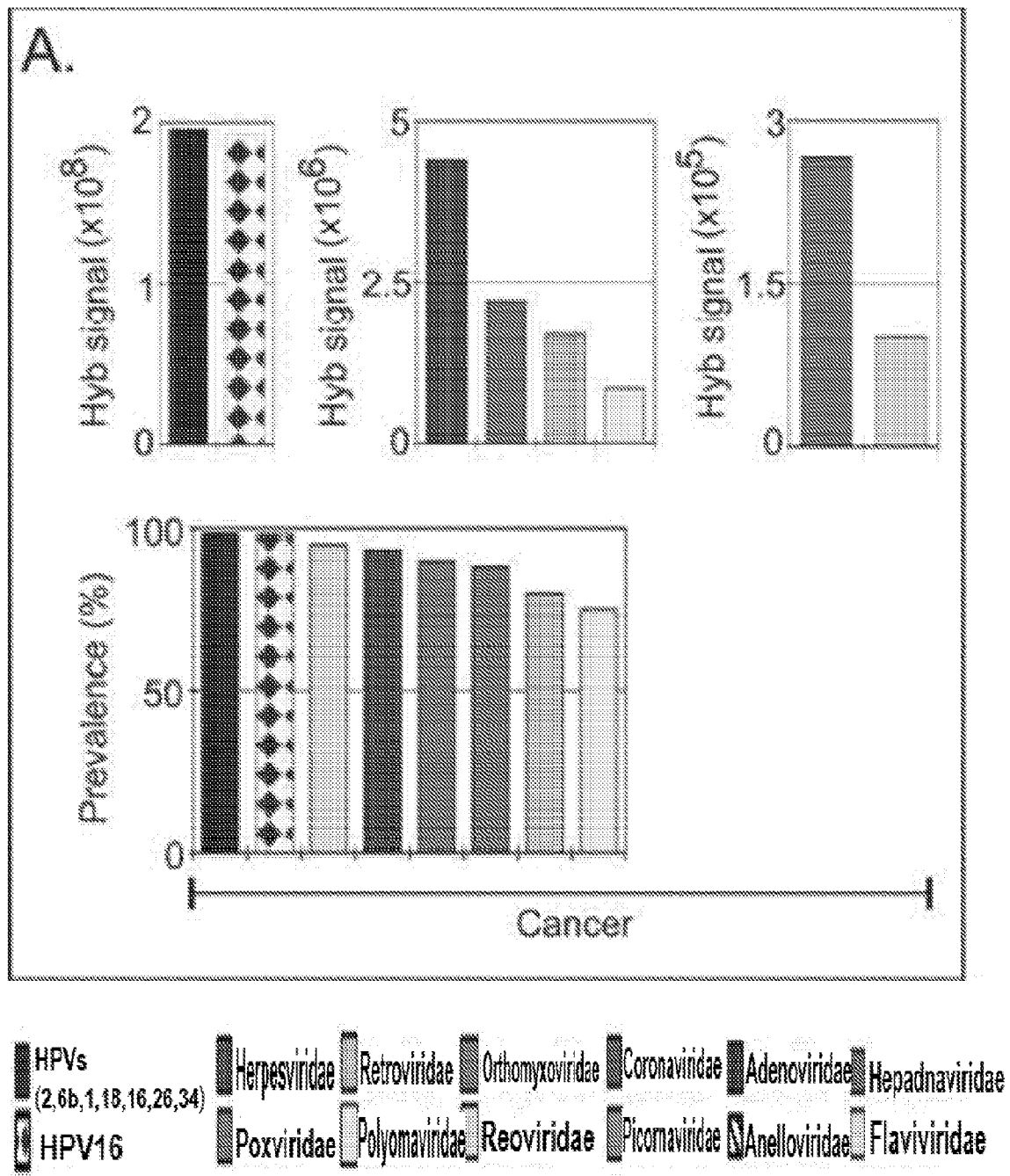


FIG. 1A

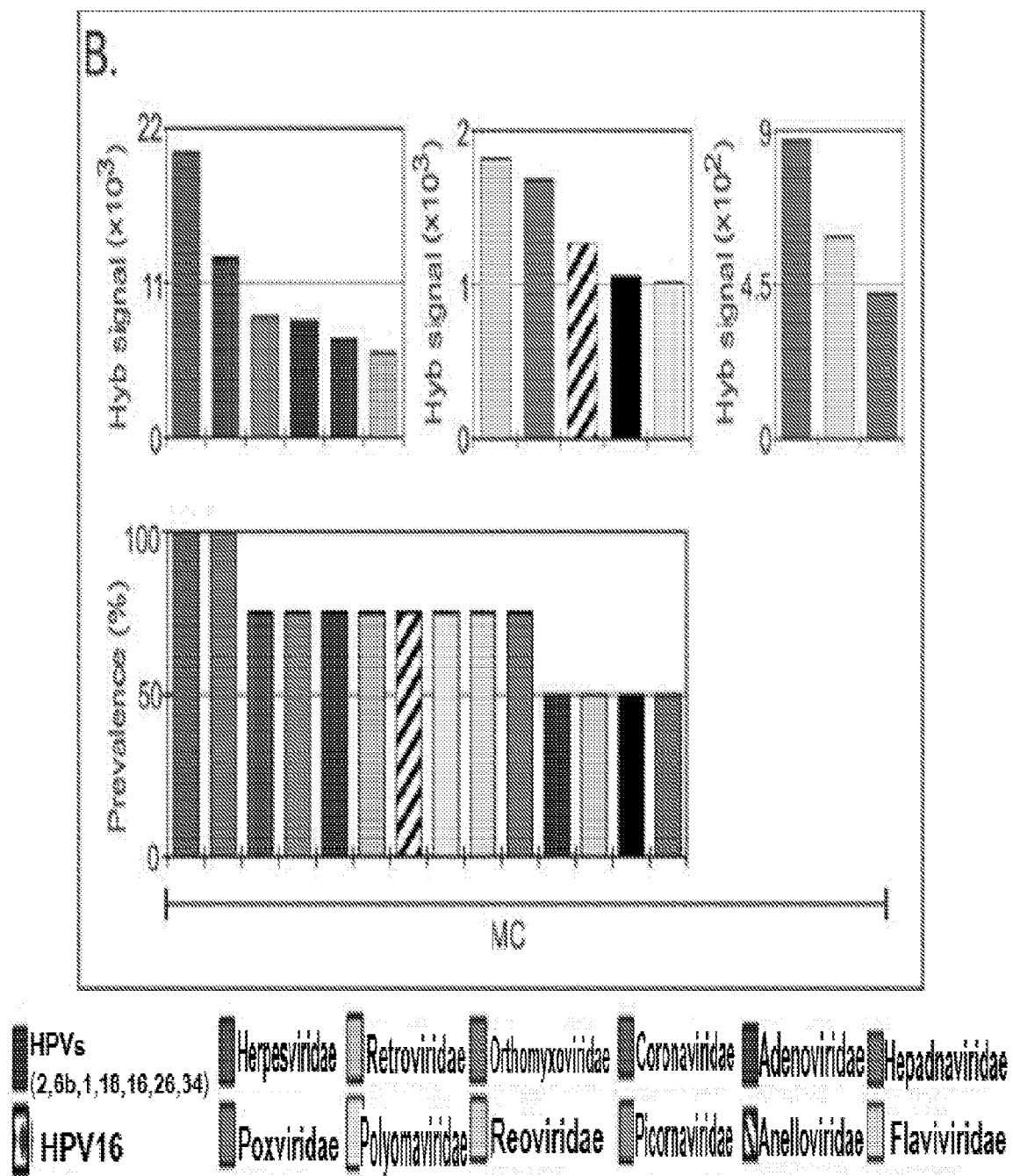


FIG. 1B

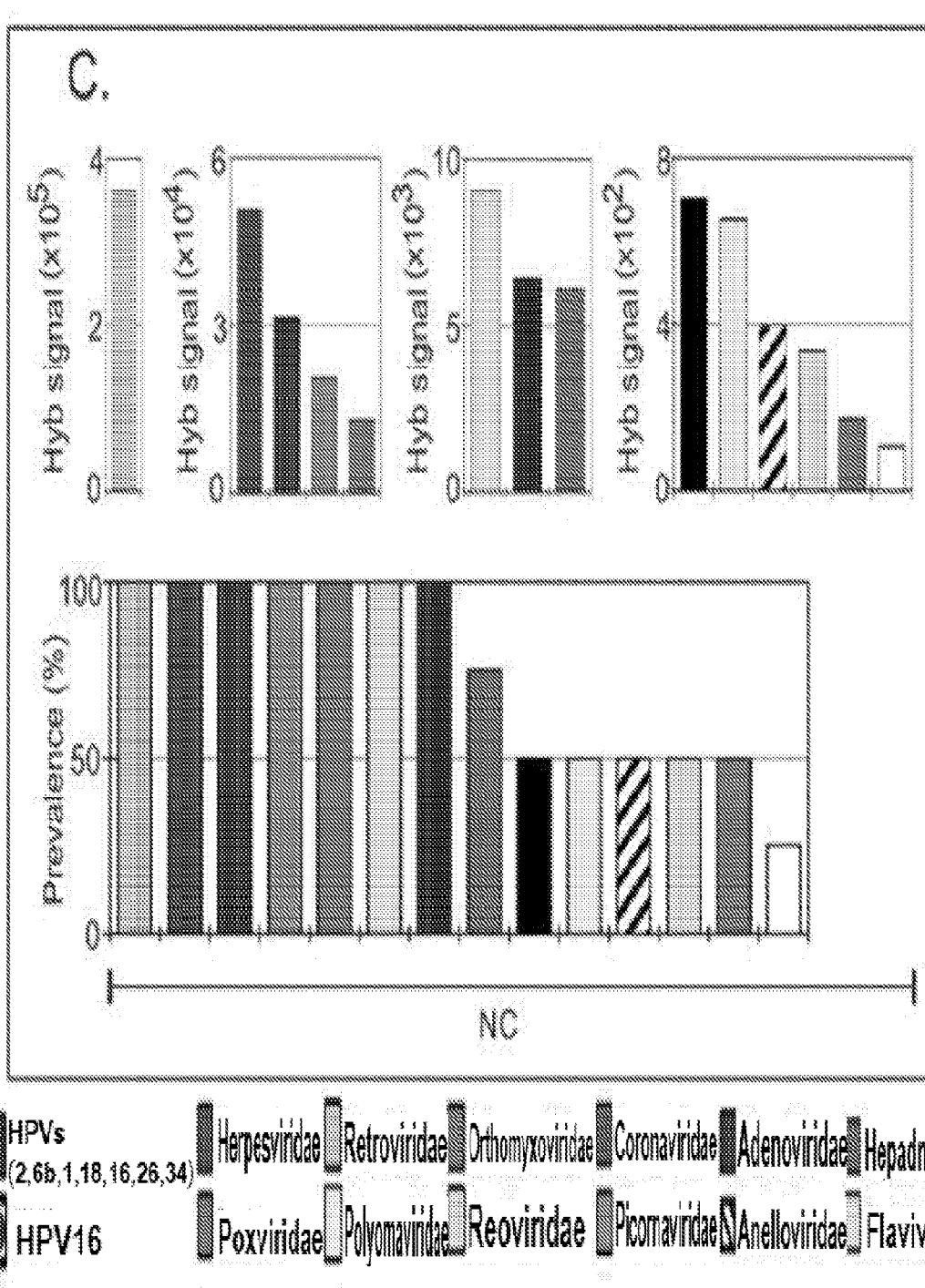


FIG. 1C

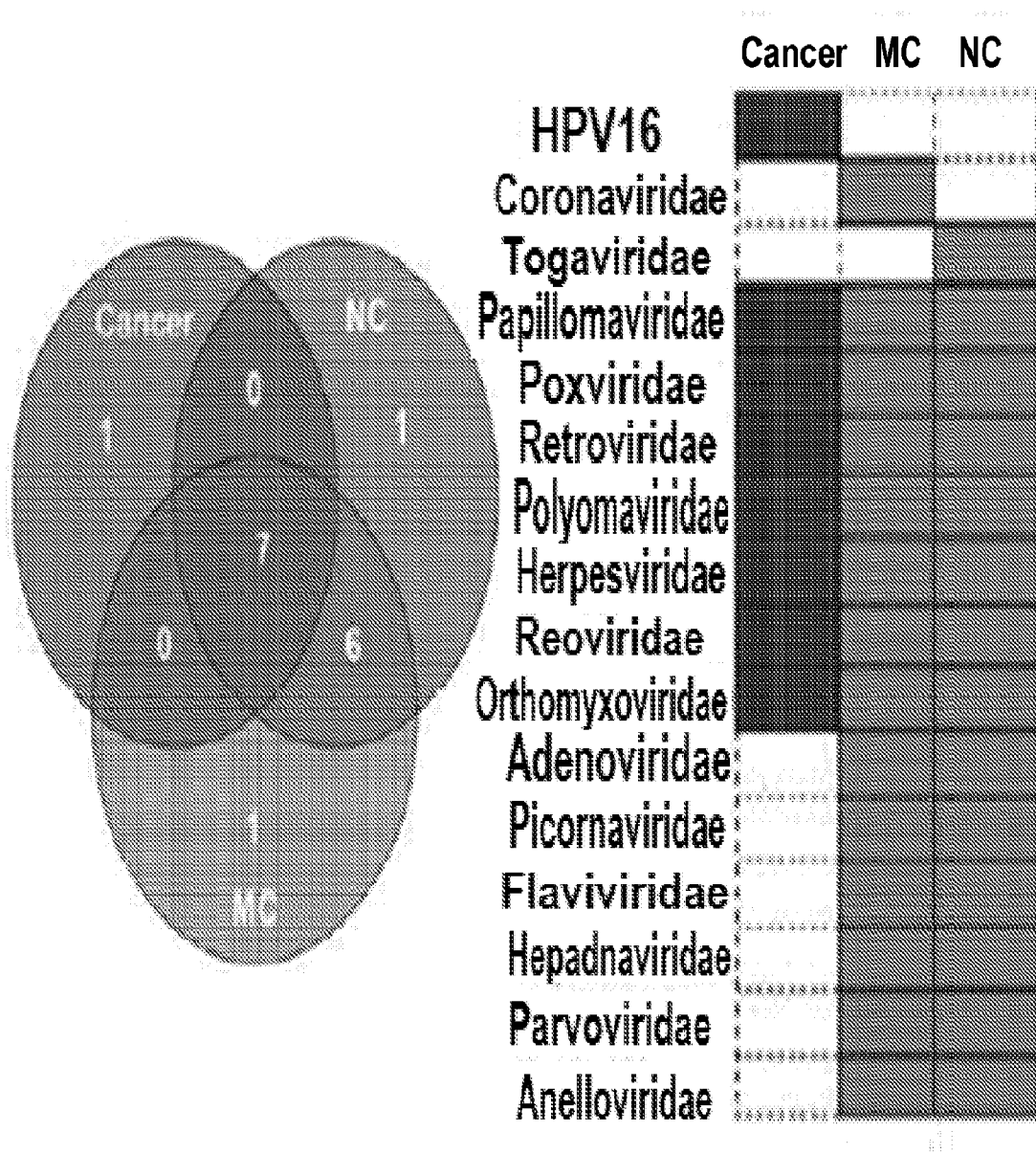


FIG. 1D

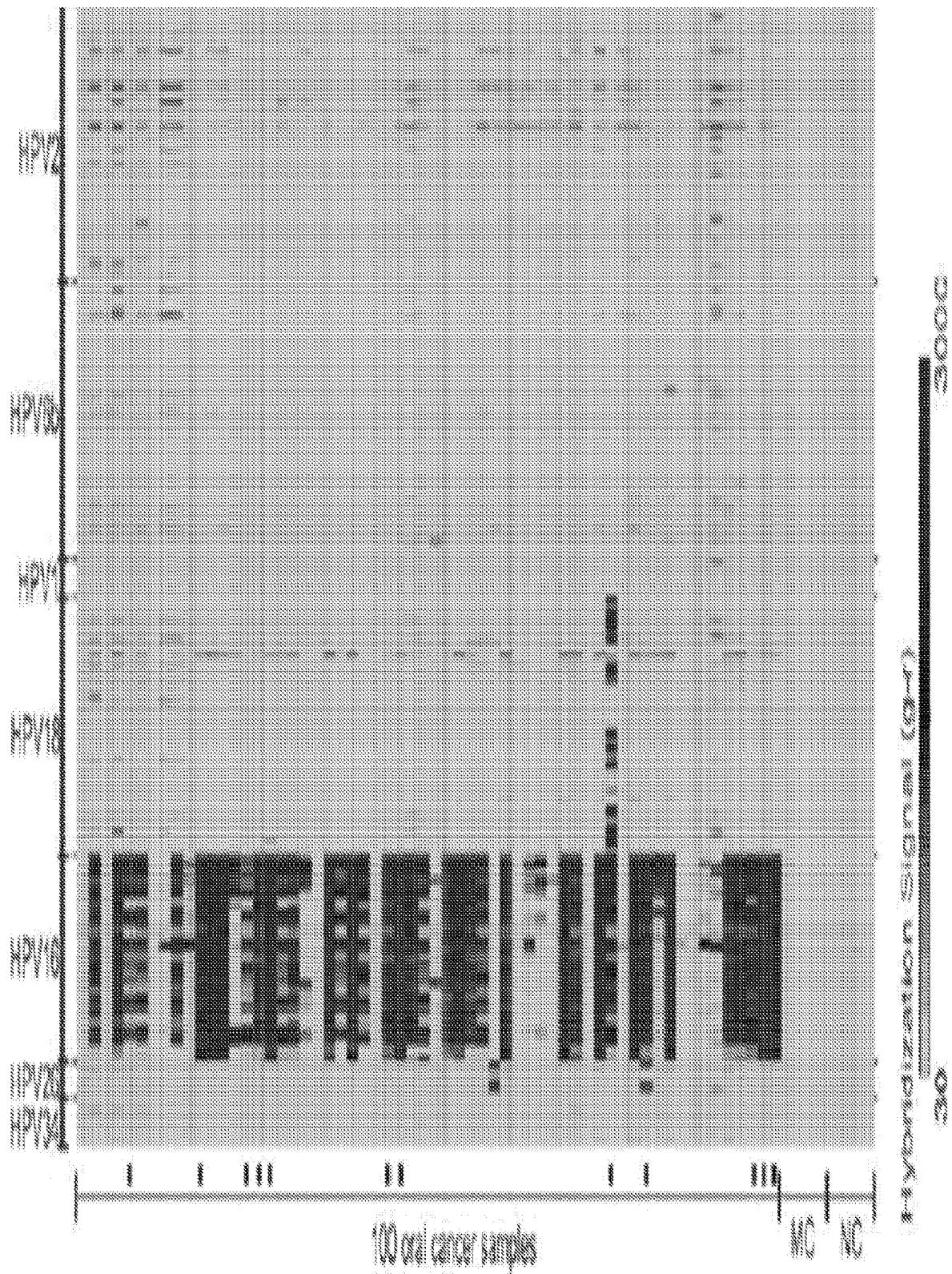


FIG. 1E

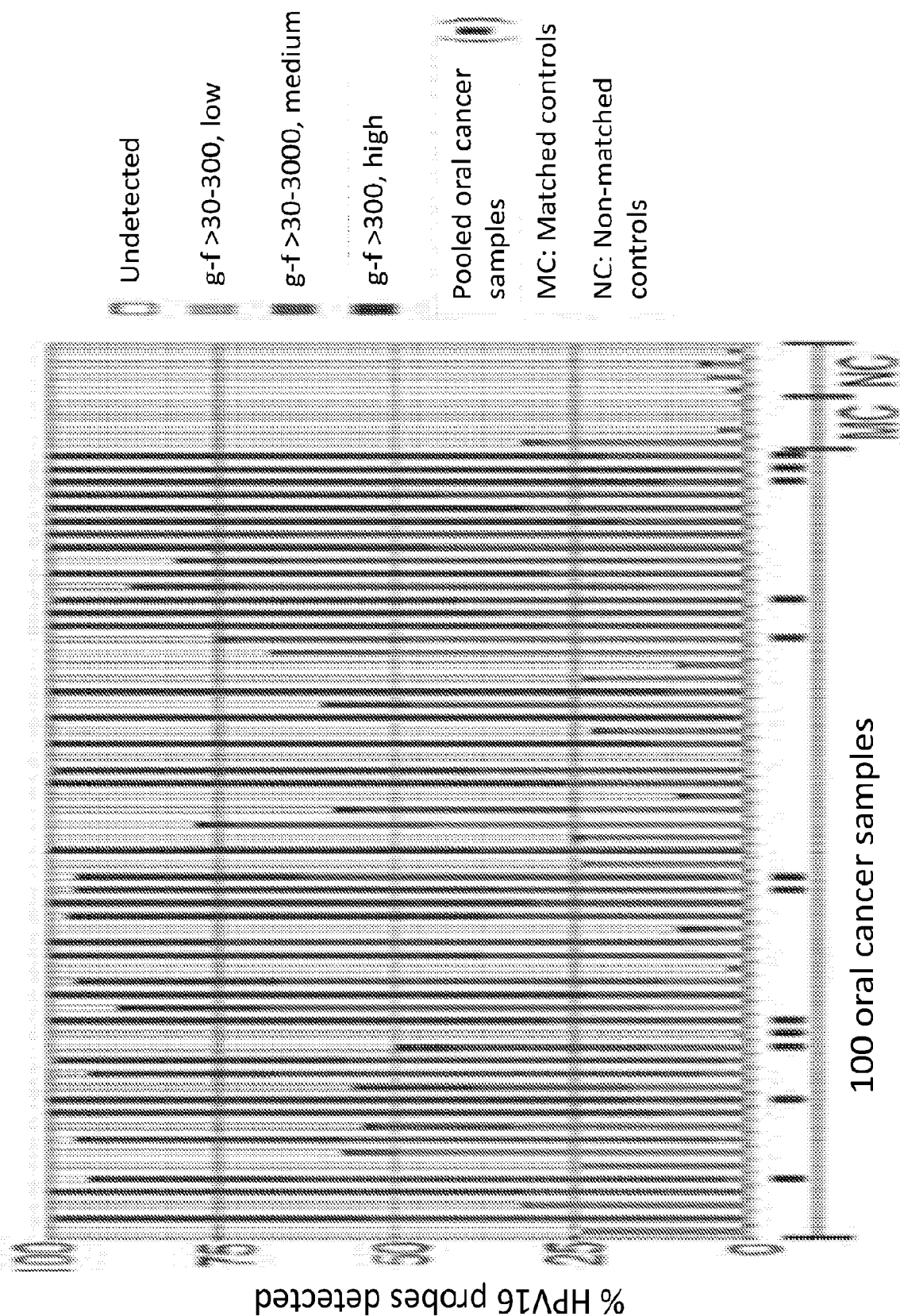


FIG. 1F

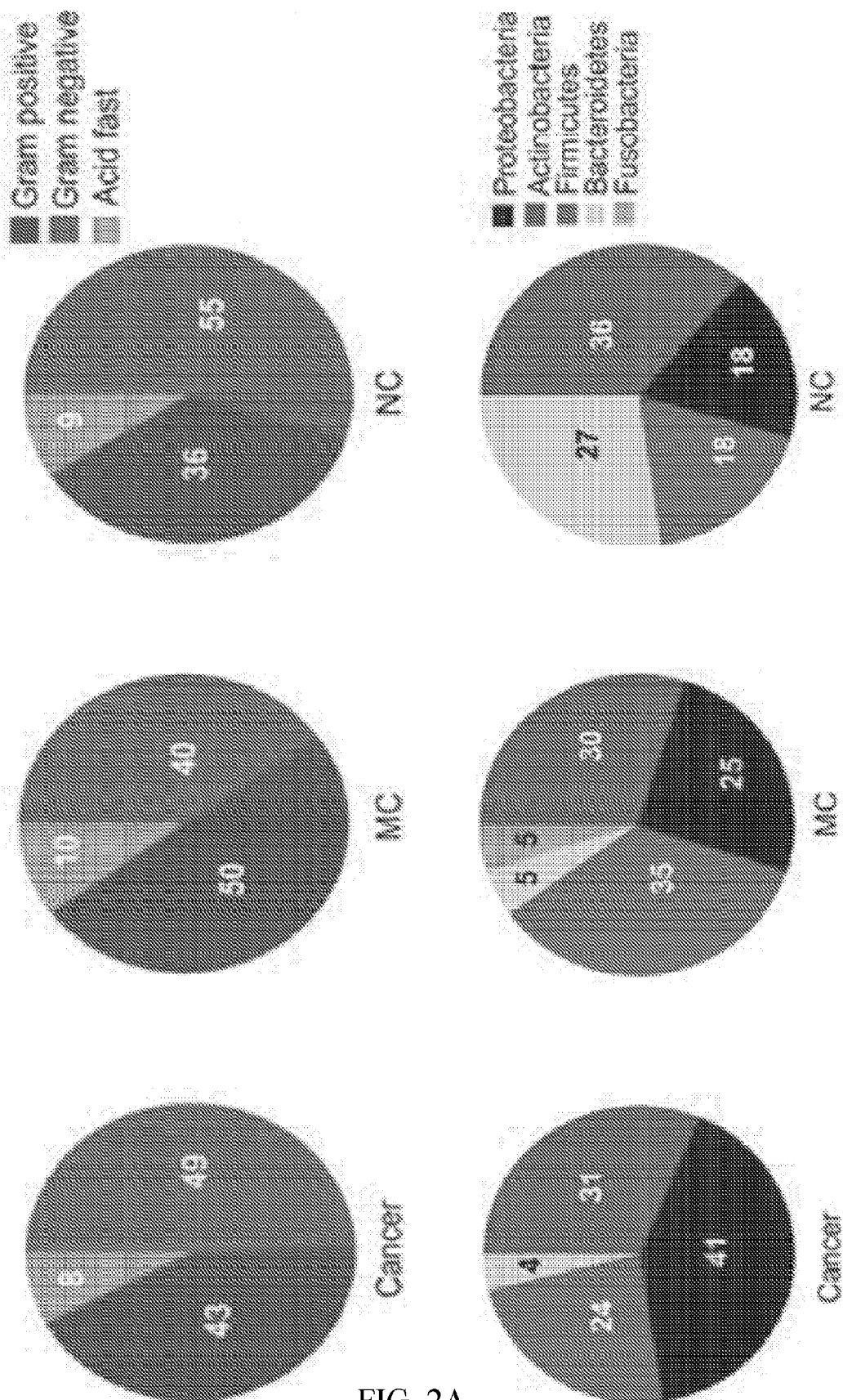


FIG. 2A

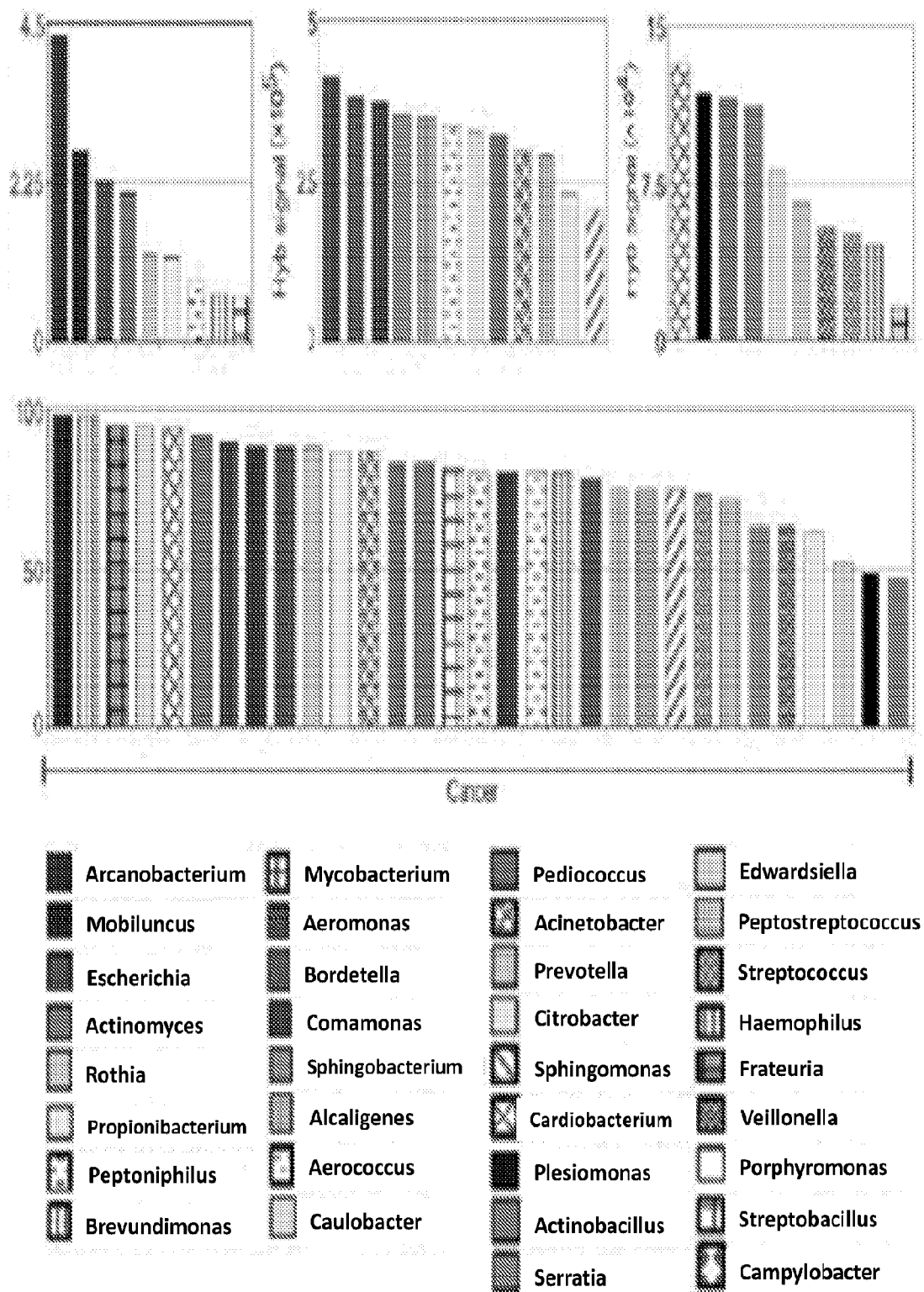


FIG. 2B

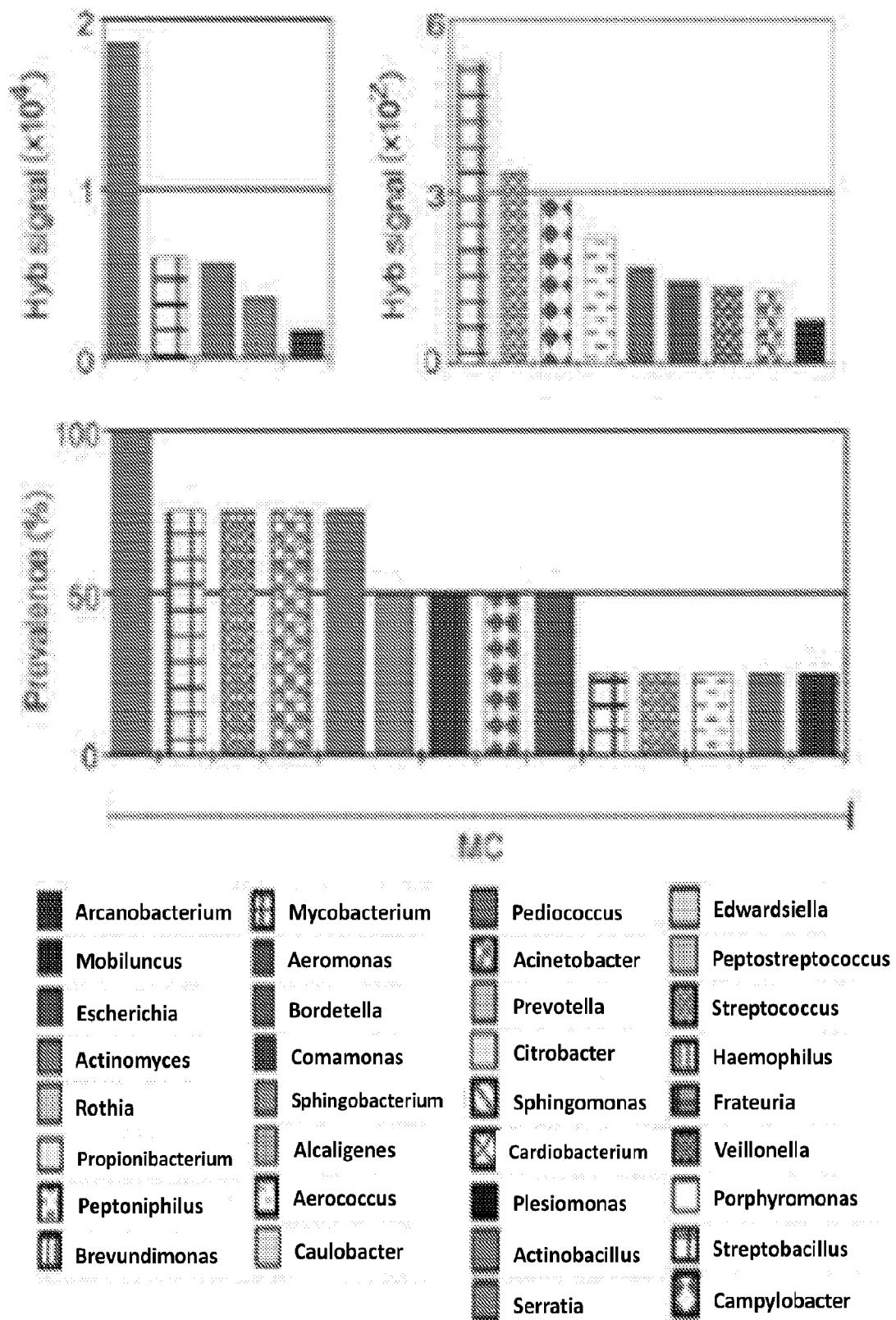


FIG. 2C

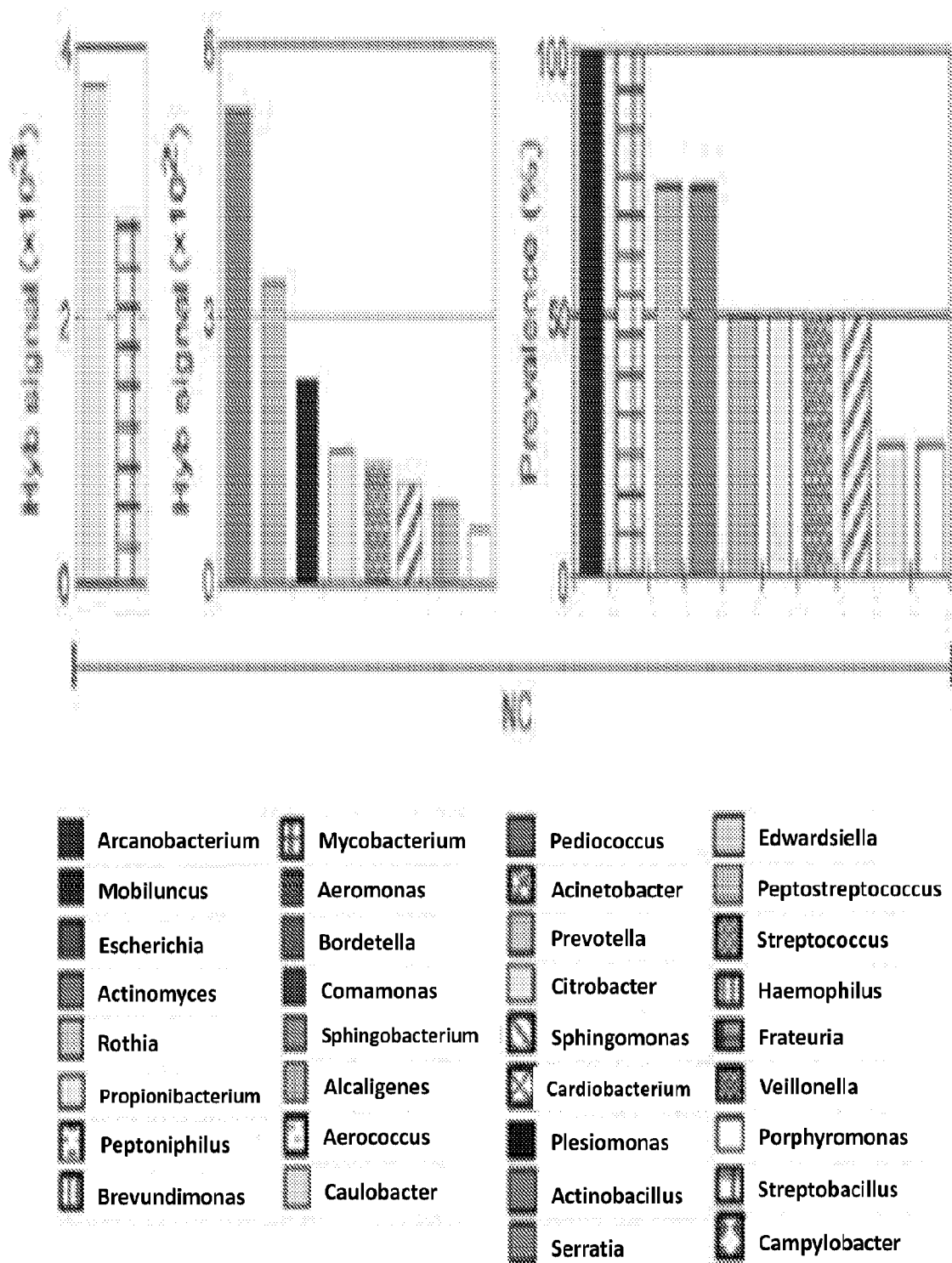
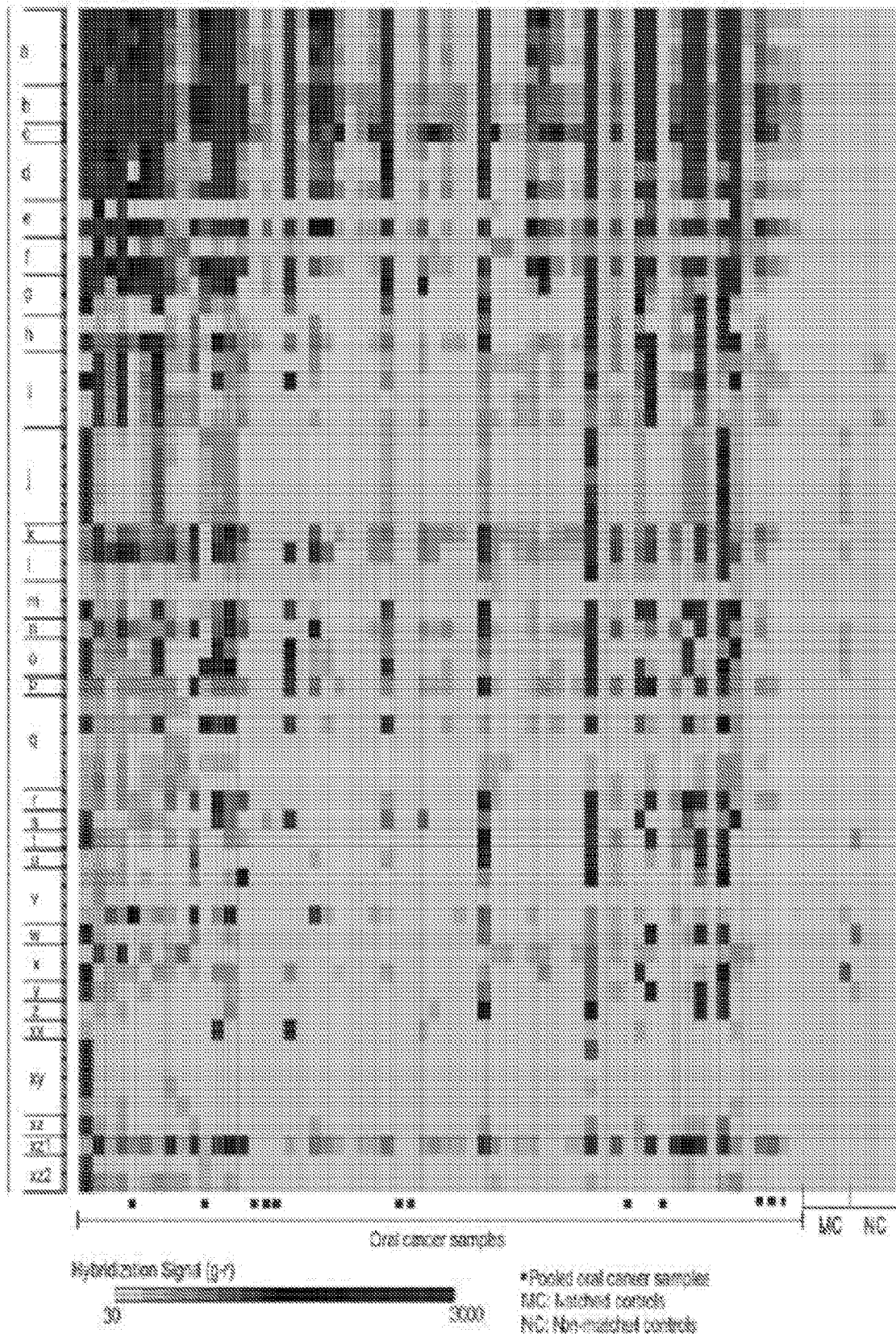


FIG. 2D



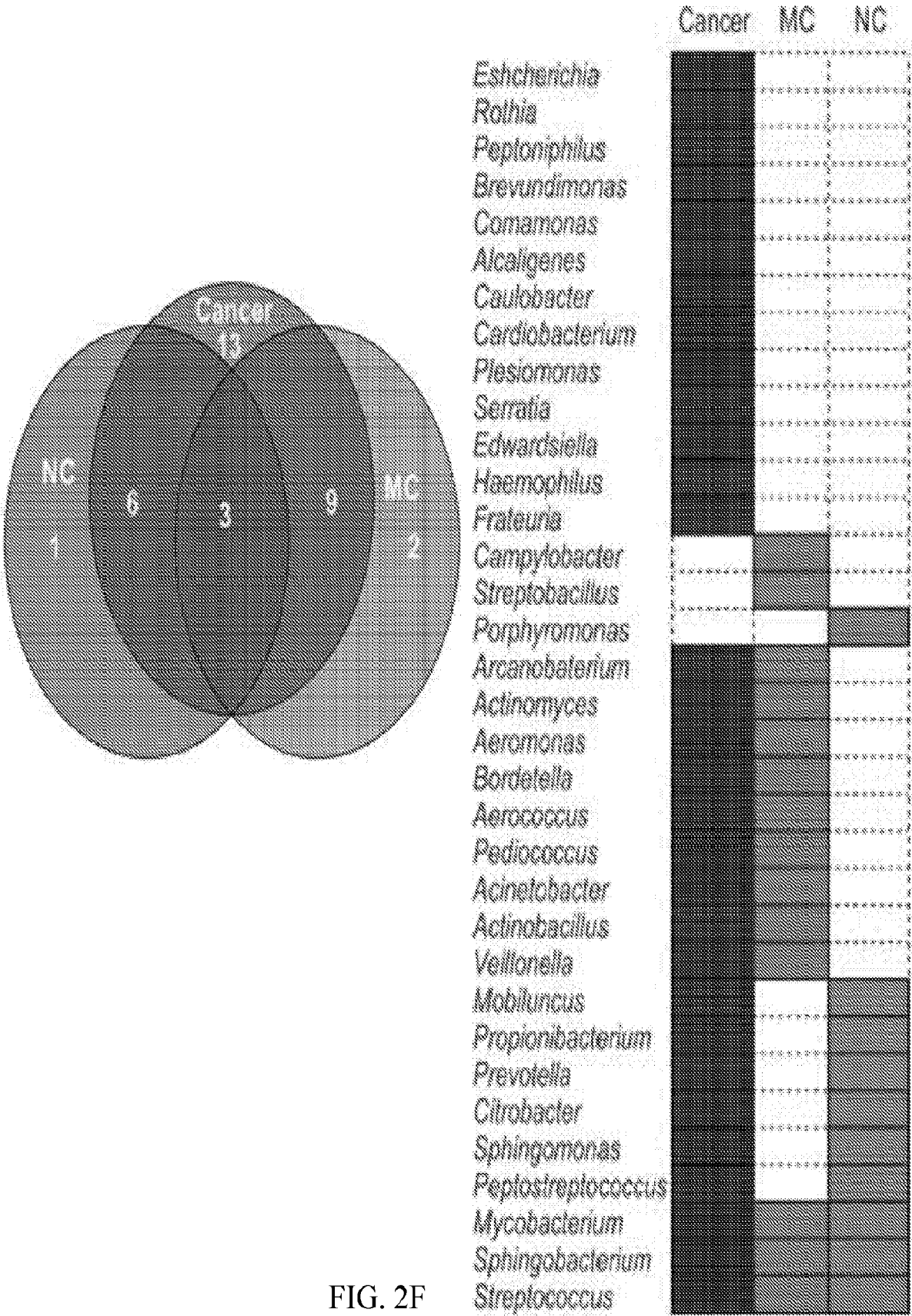


FIG. 2F

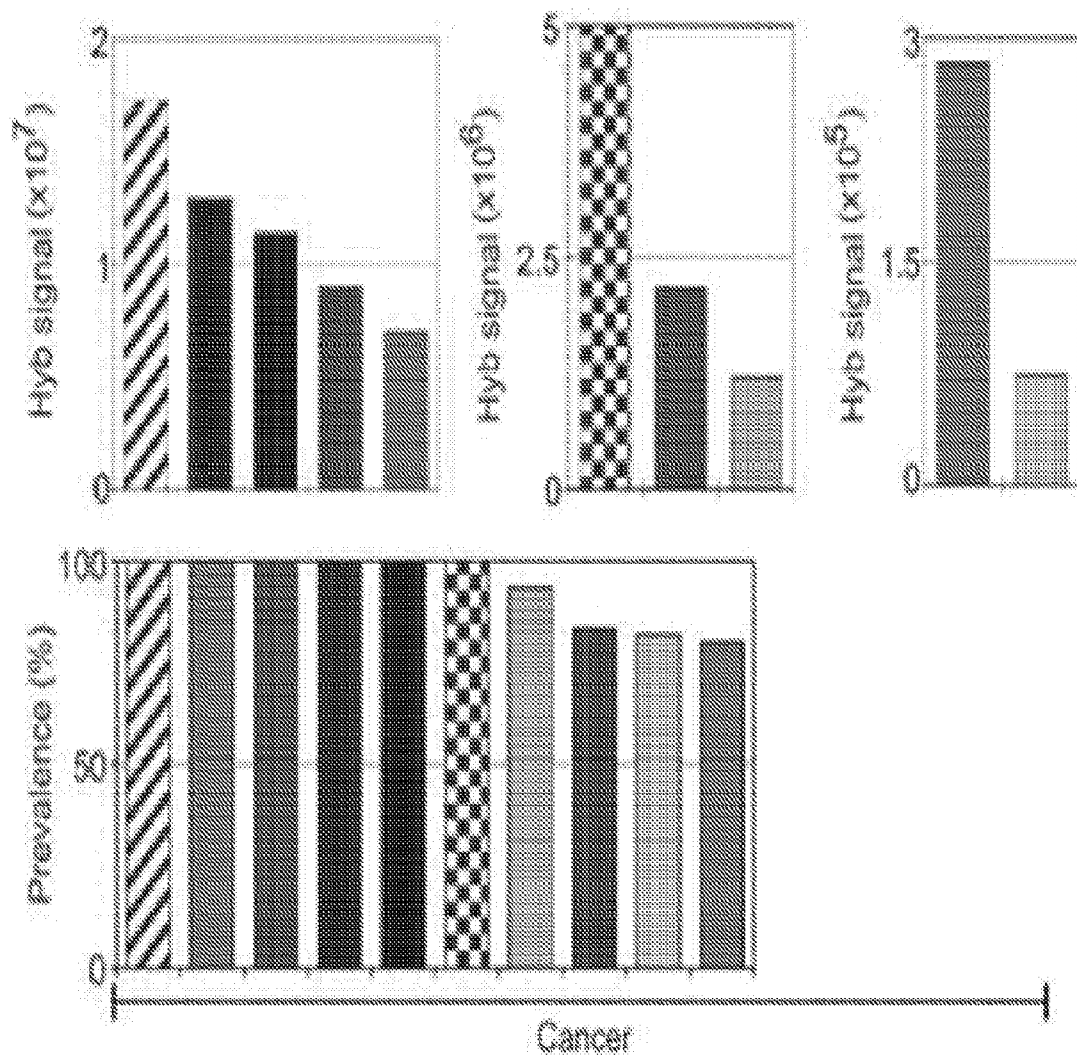


FIG. 3A

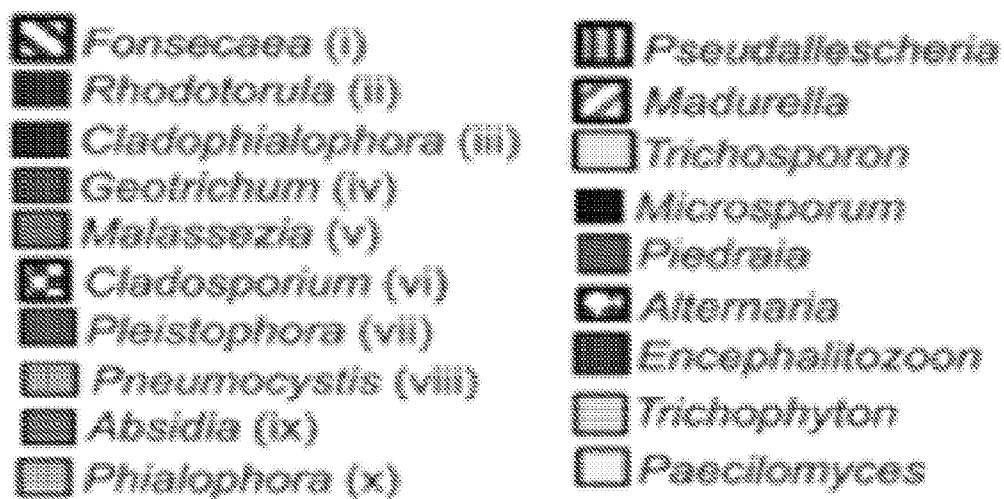
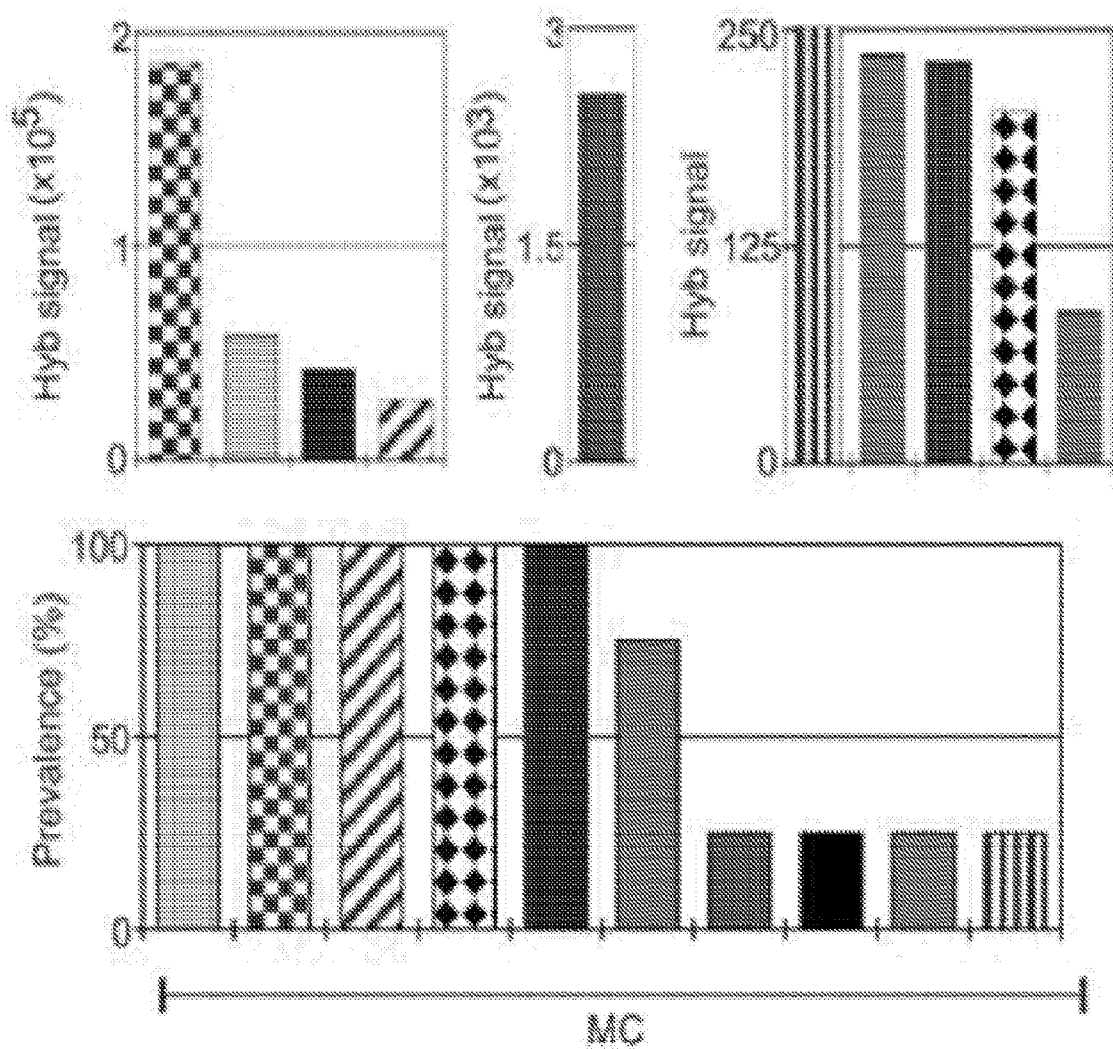


FIG. 3B

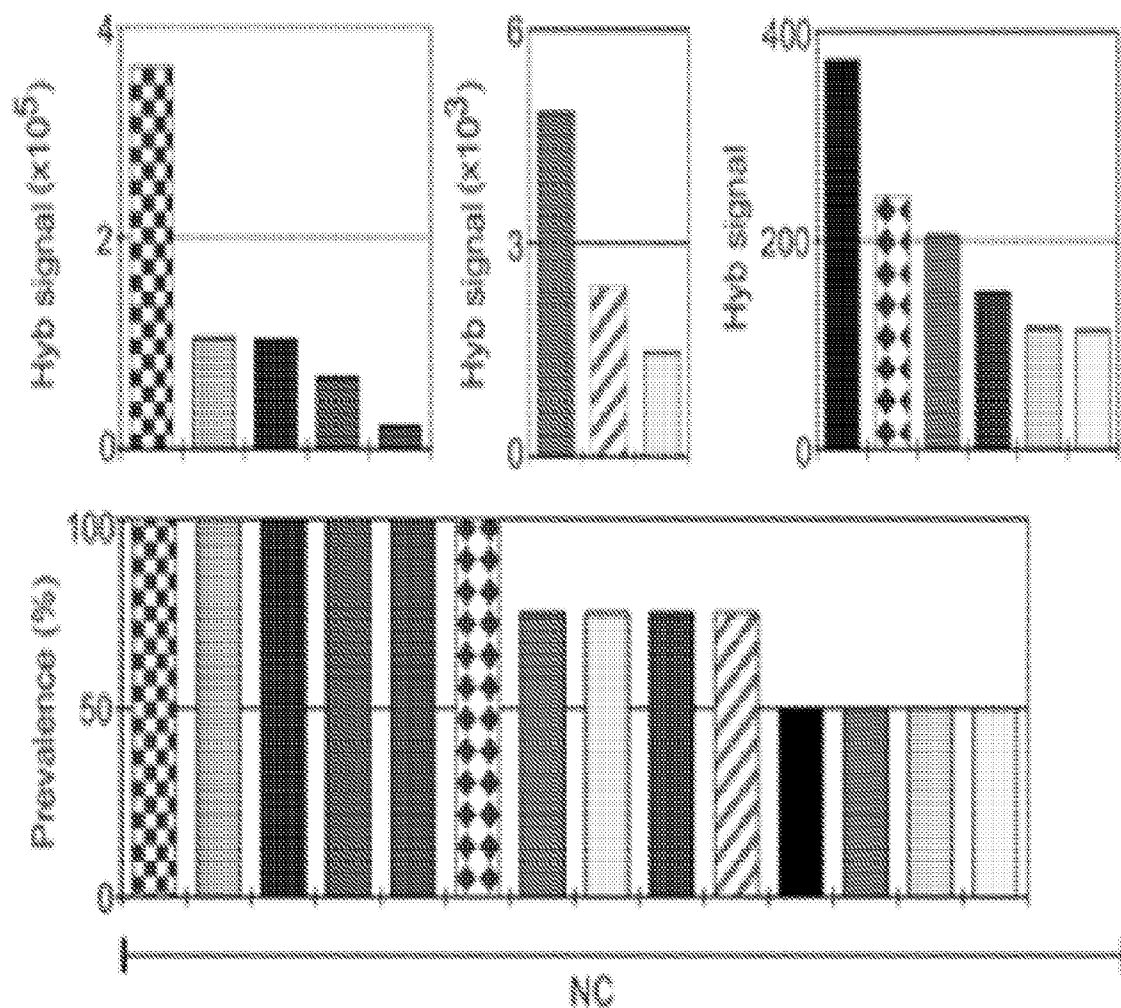


FIG. 3C

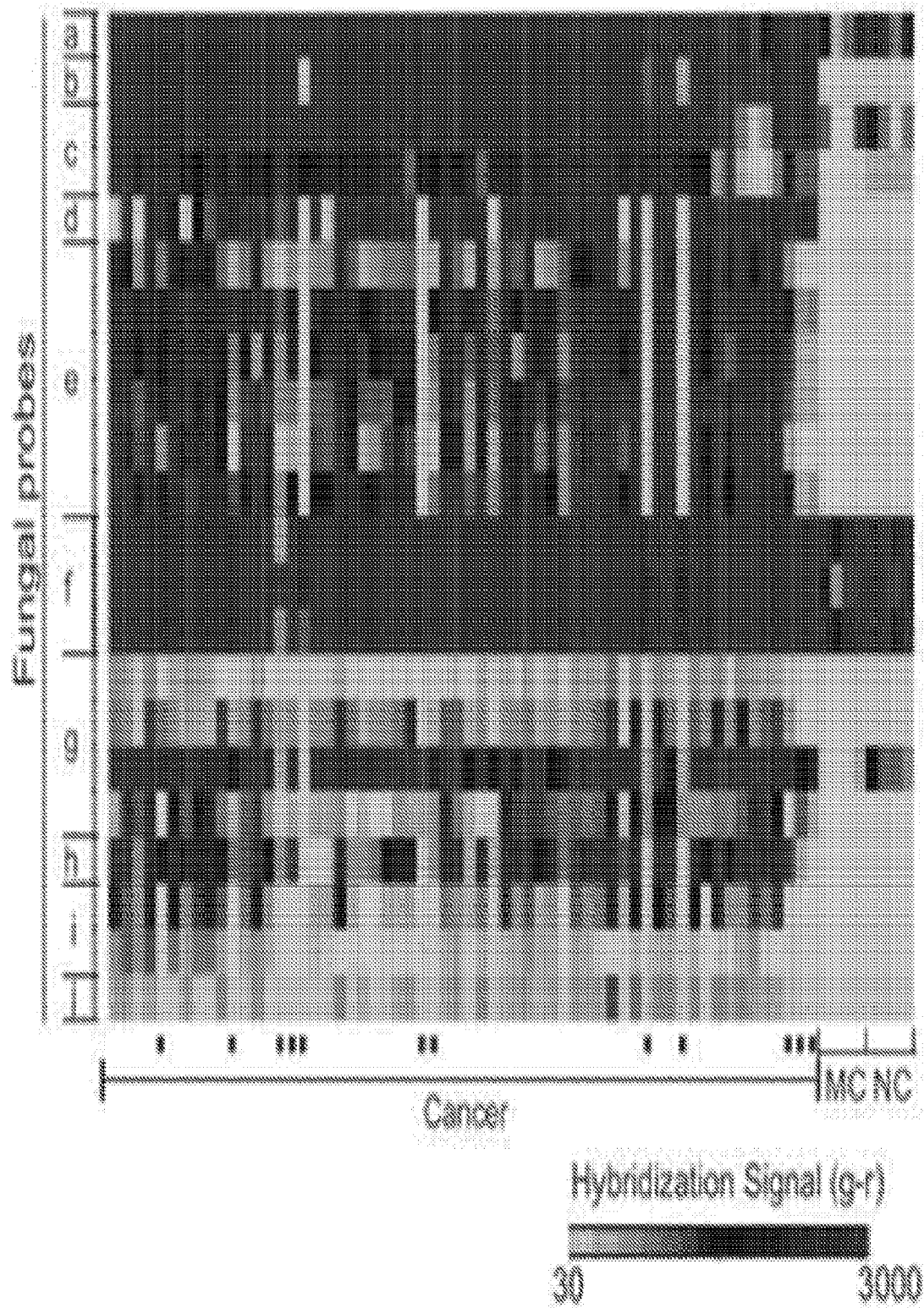


FIG. 3D

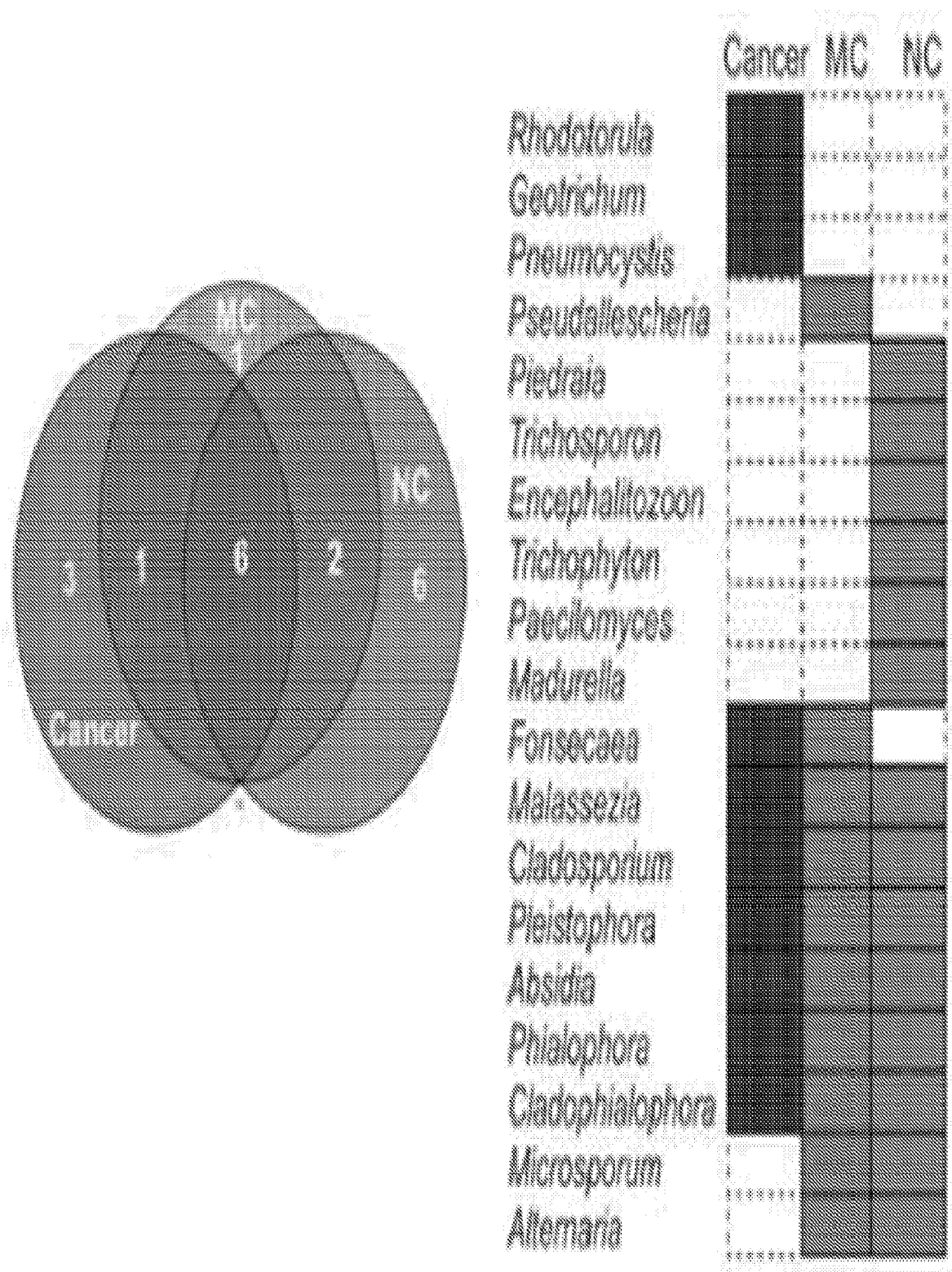


FIG. 3E

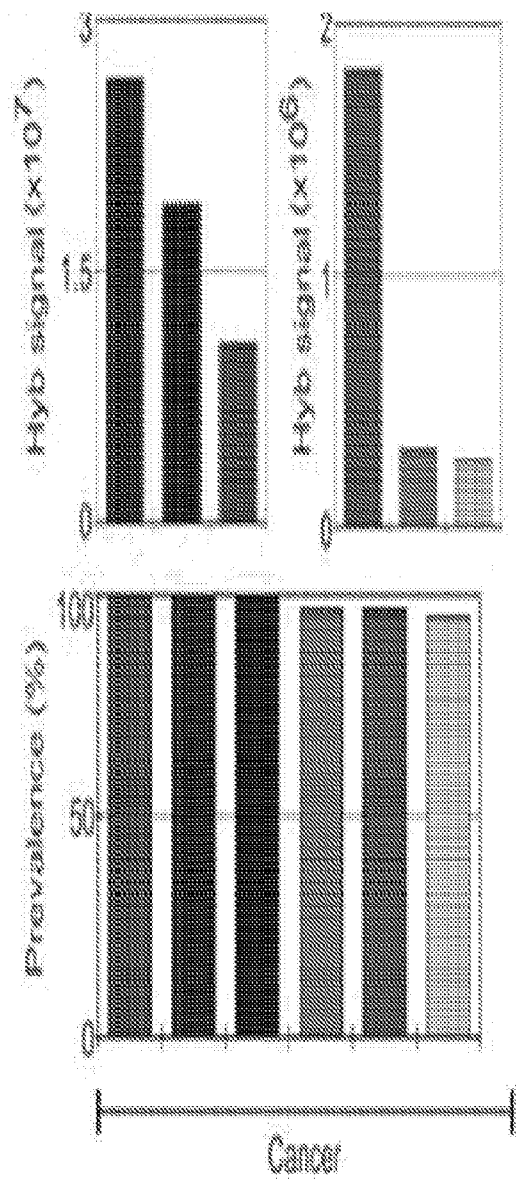


FIG. 3F

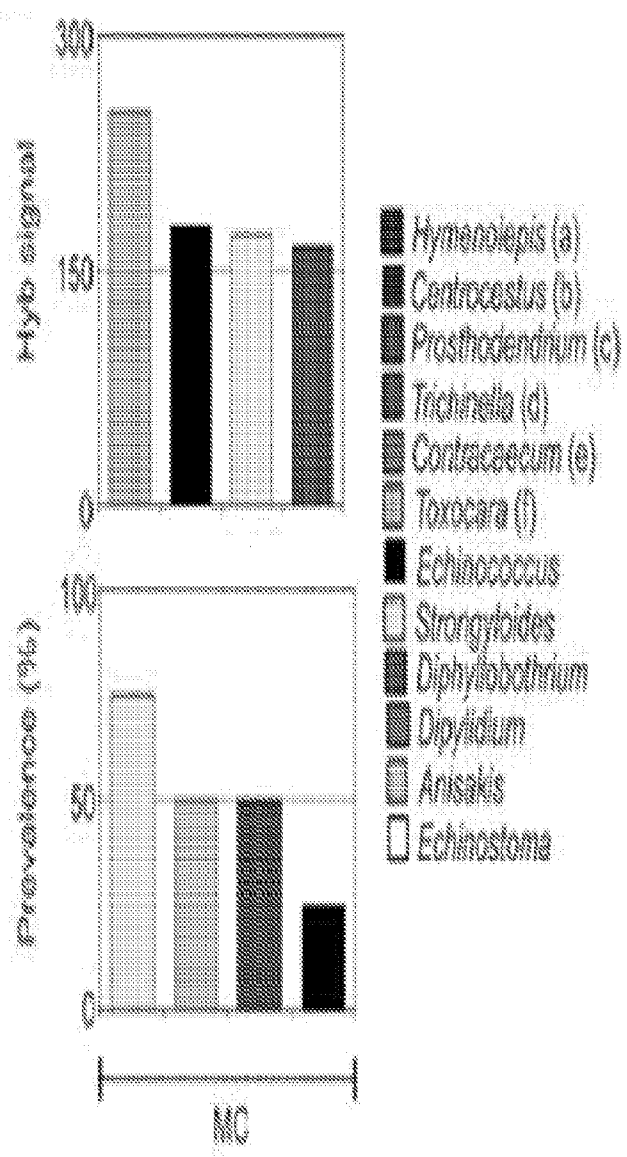


FIG. 3G

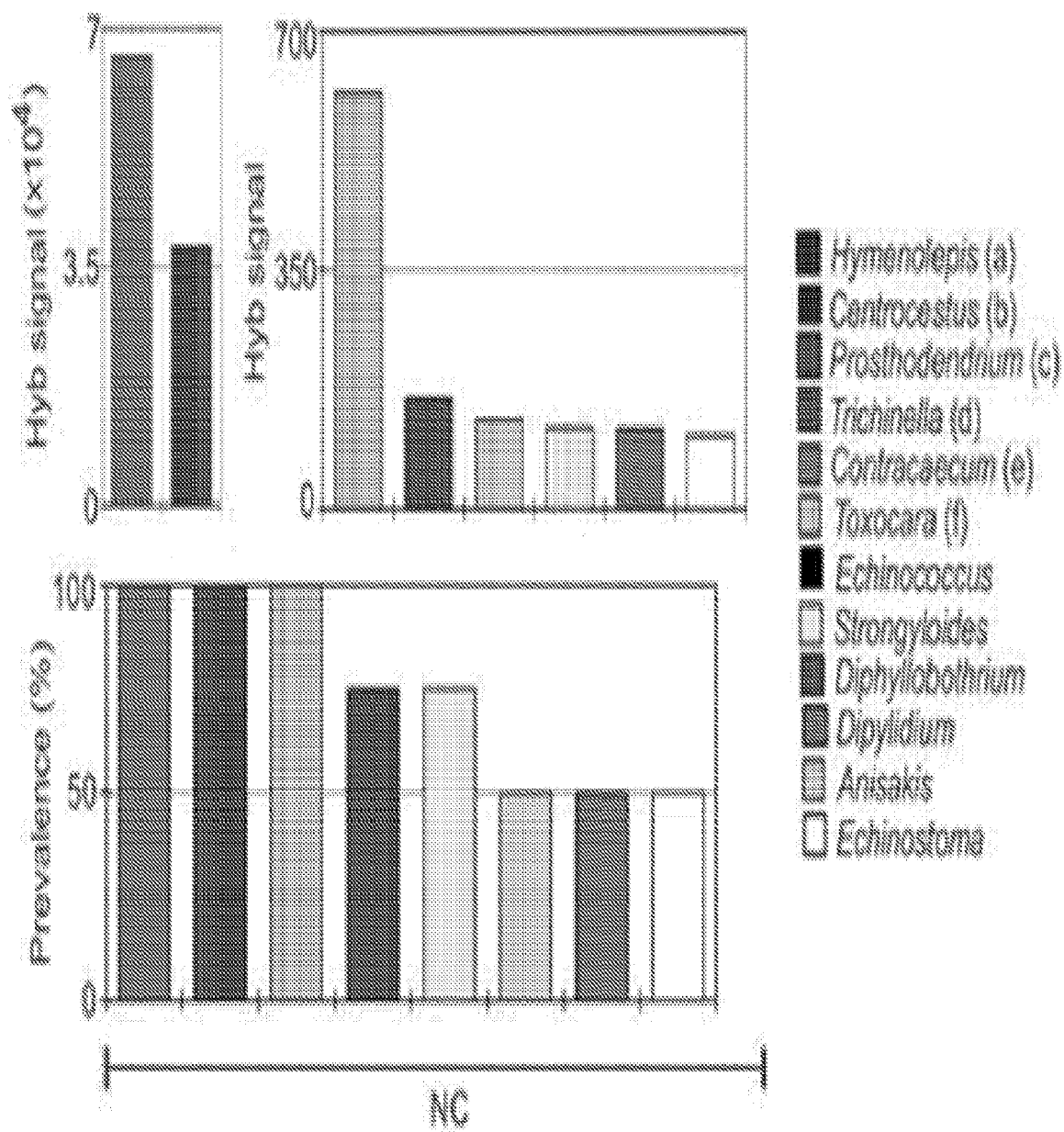


FIG. 3H

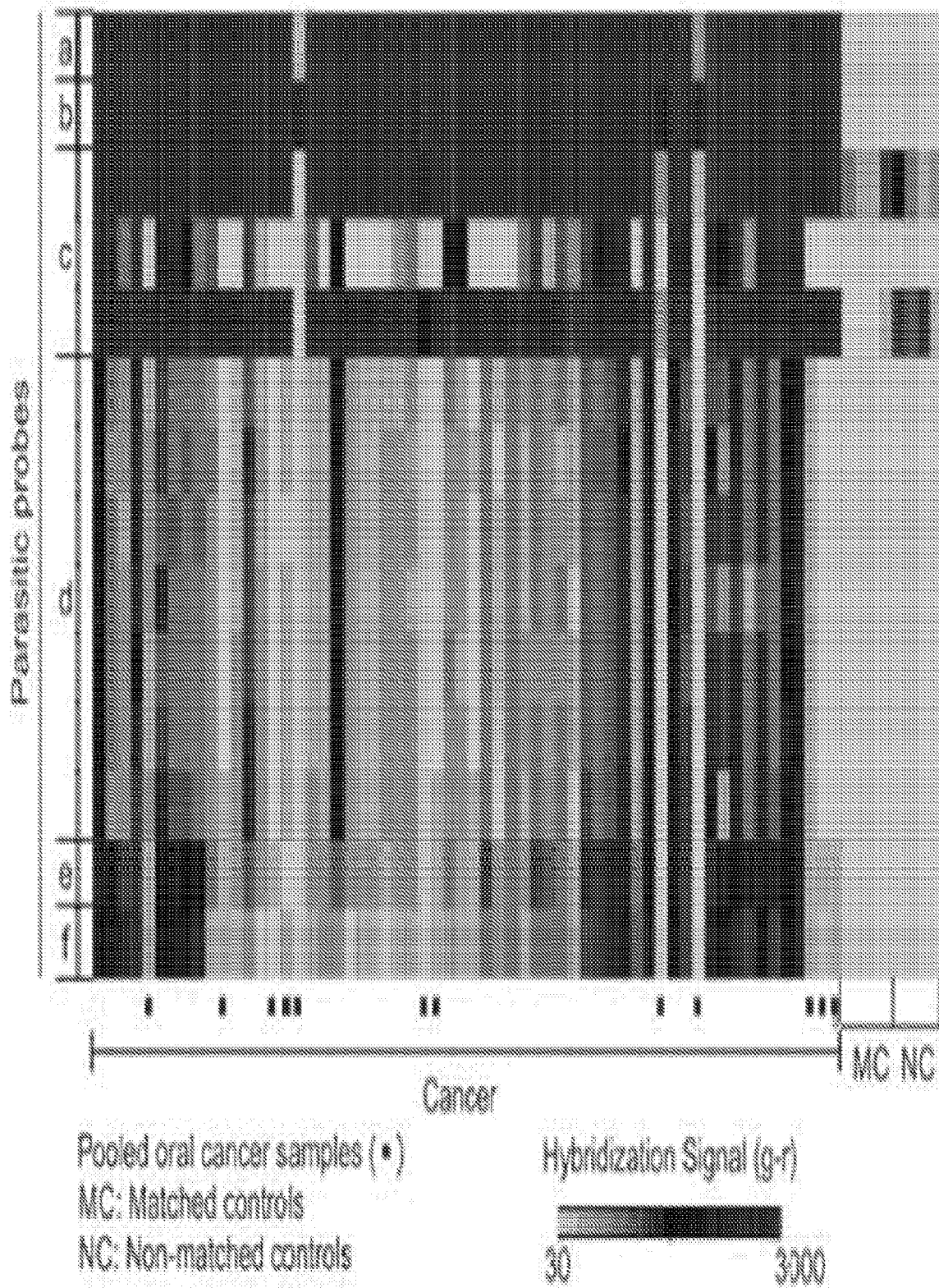


FIG. 3I

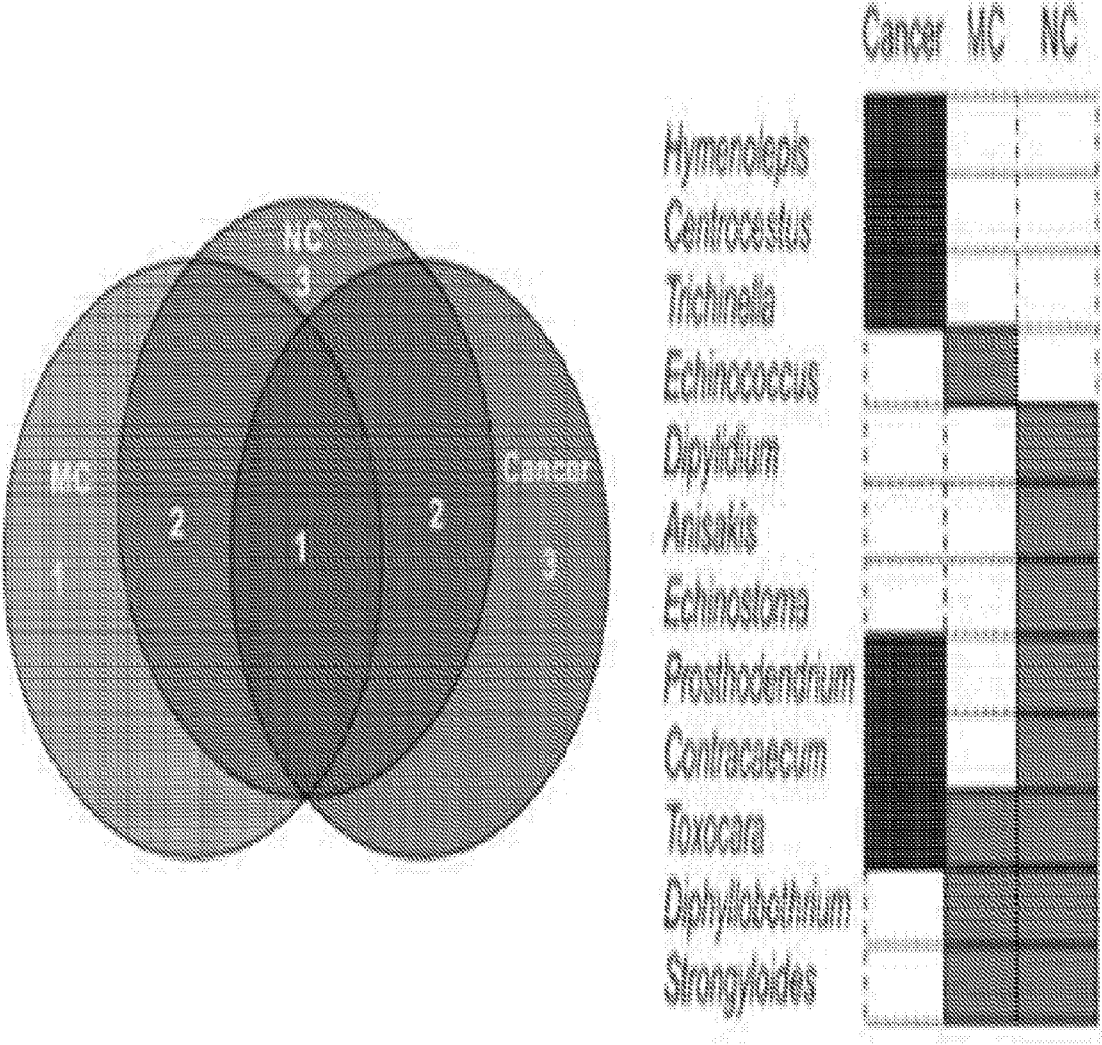


FIG. 3J



FIG. 4A

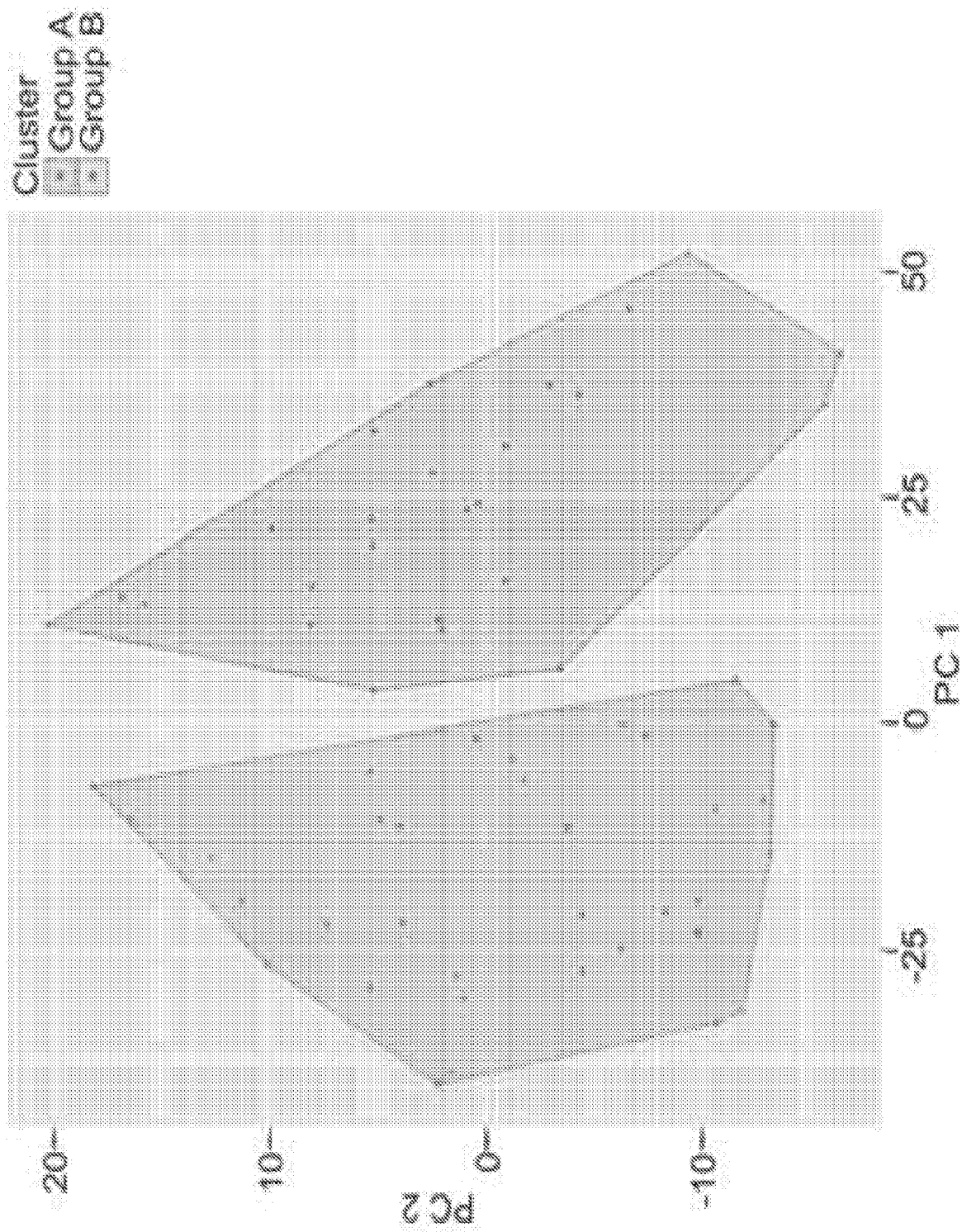


FIG. 4B

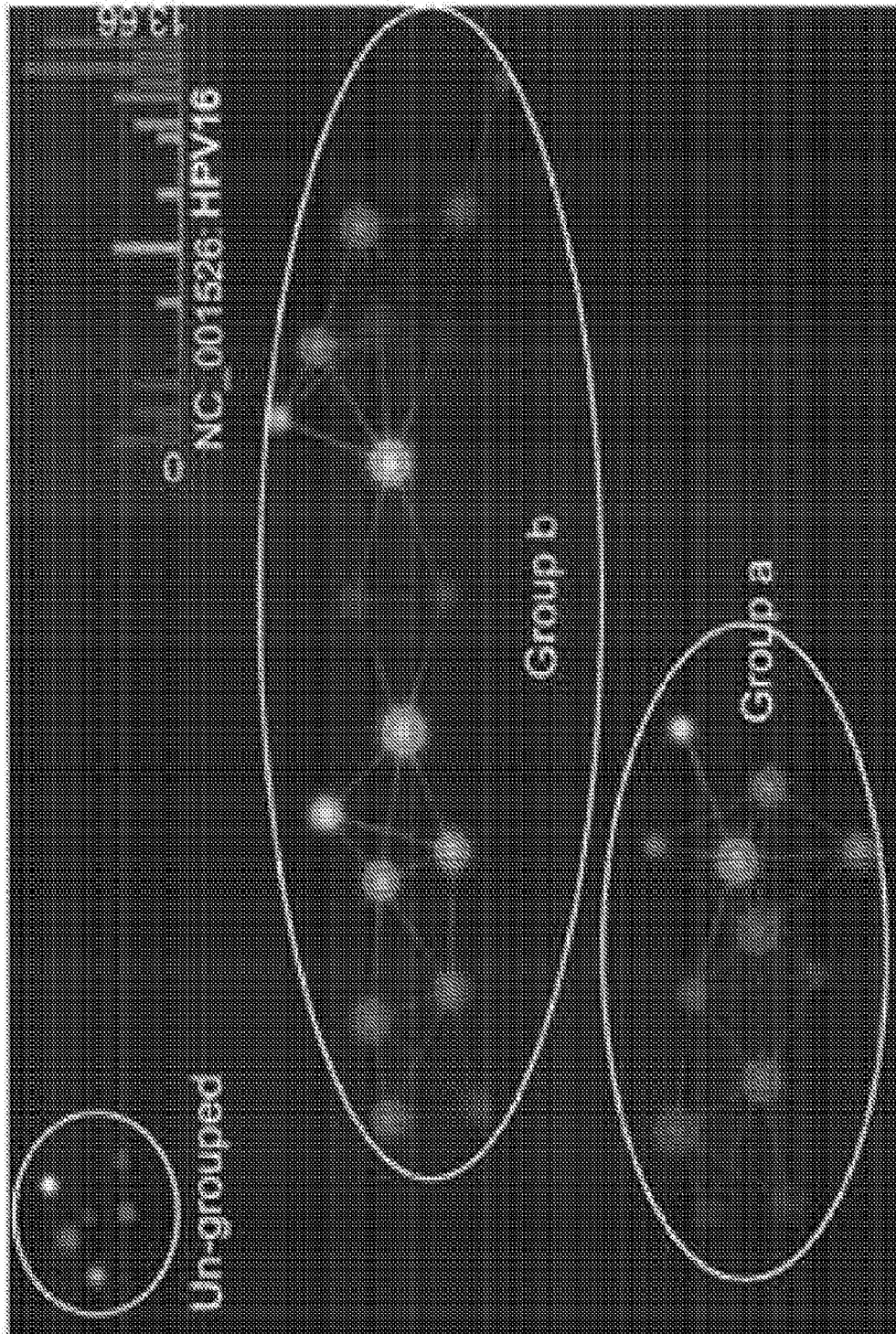


FIG. 4C

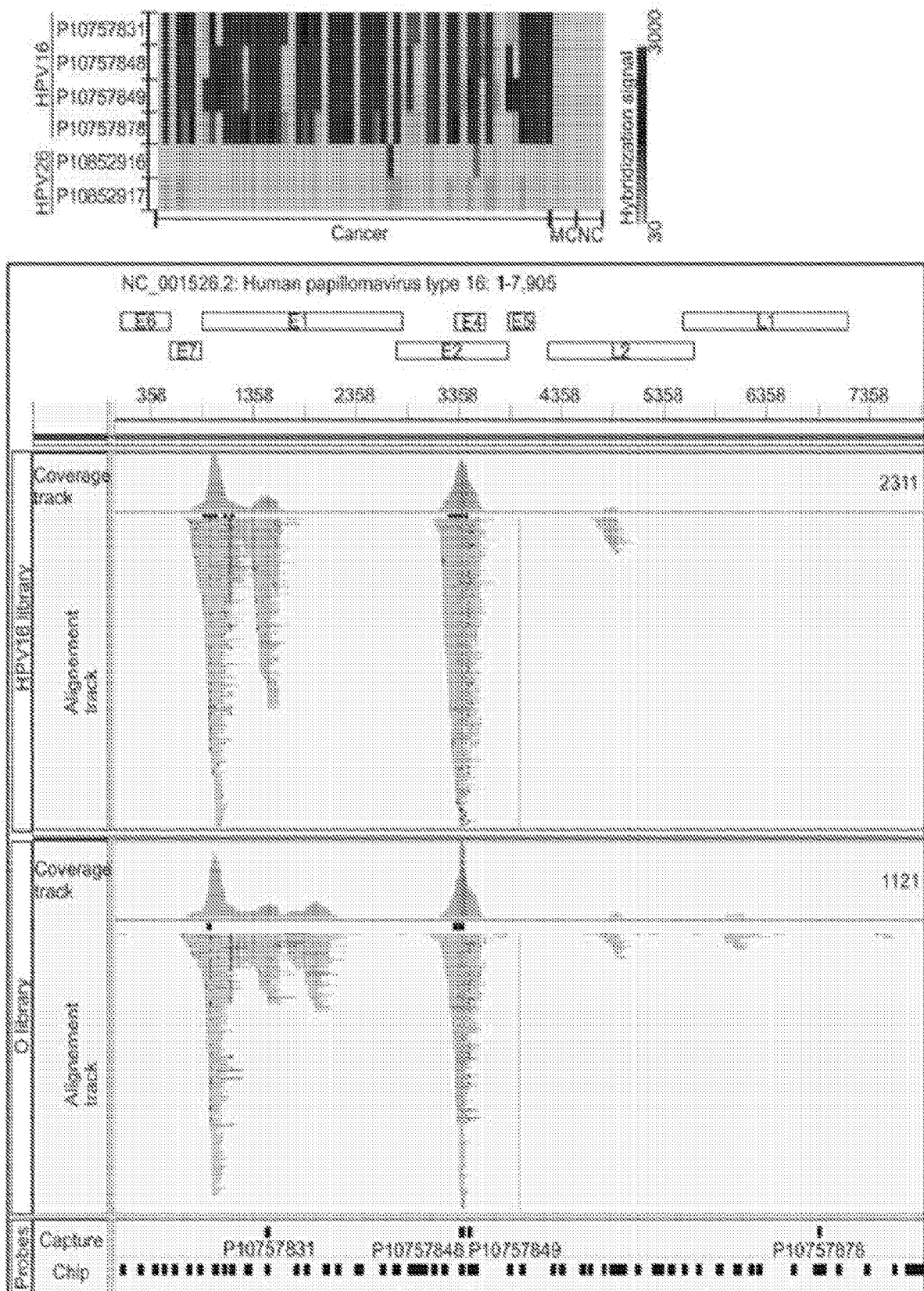


FIG. 5A

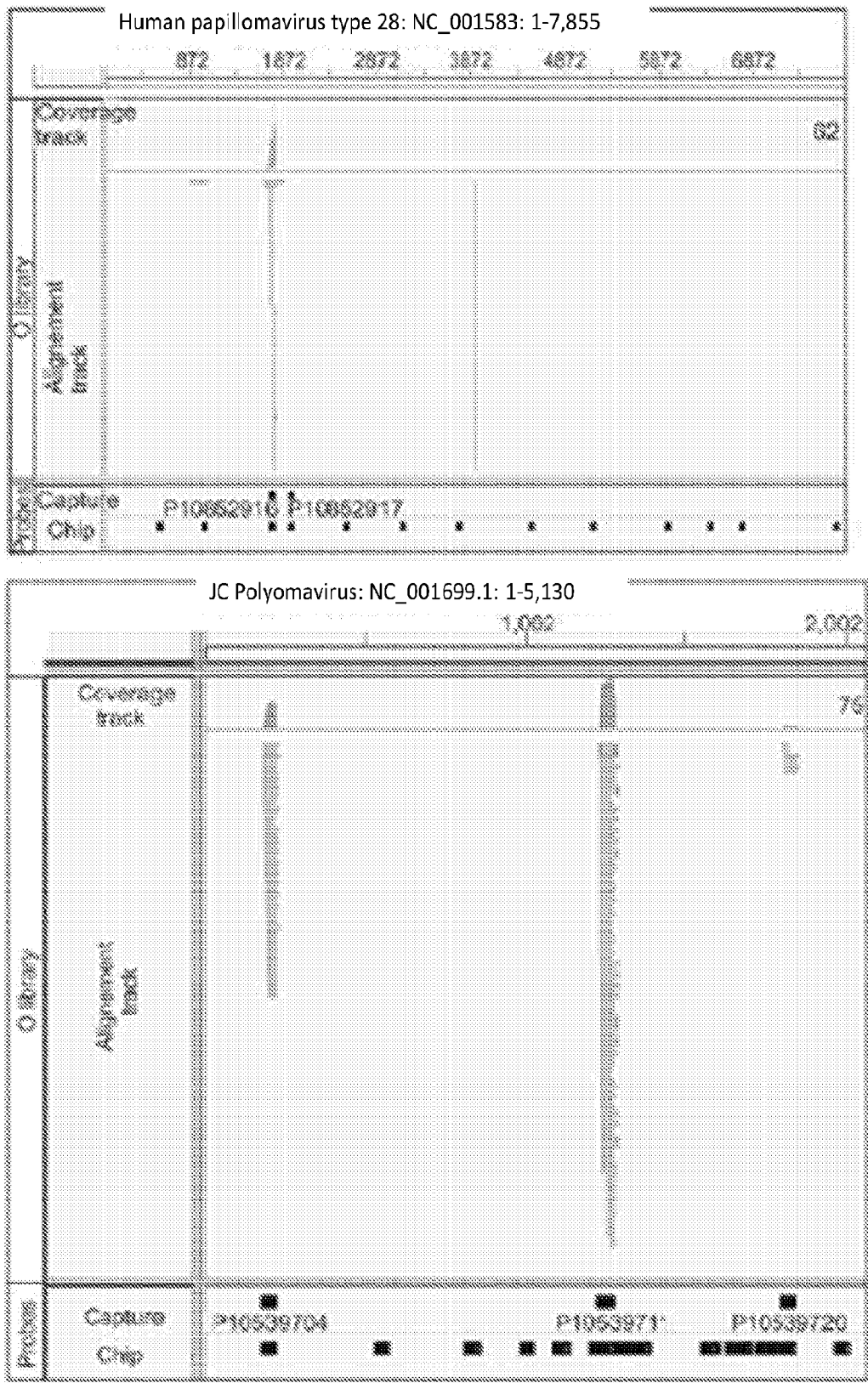


FIG. 5B

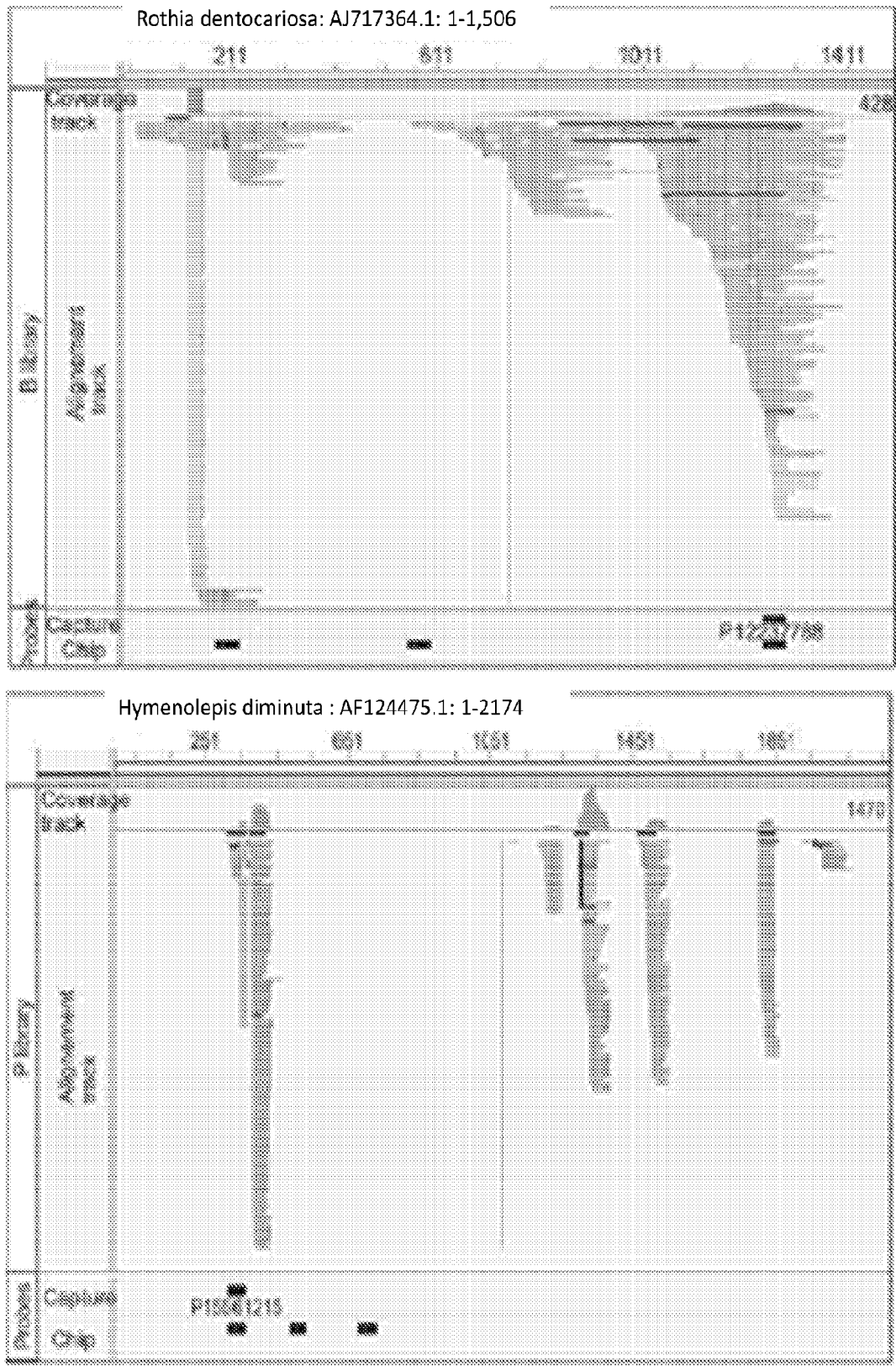


FIG. 5C

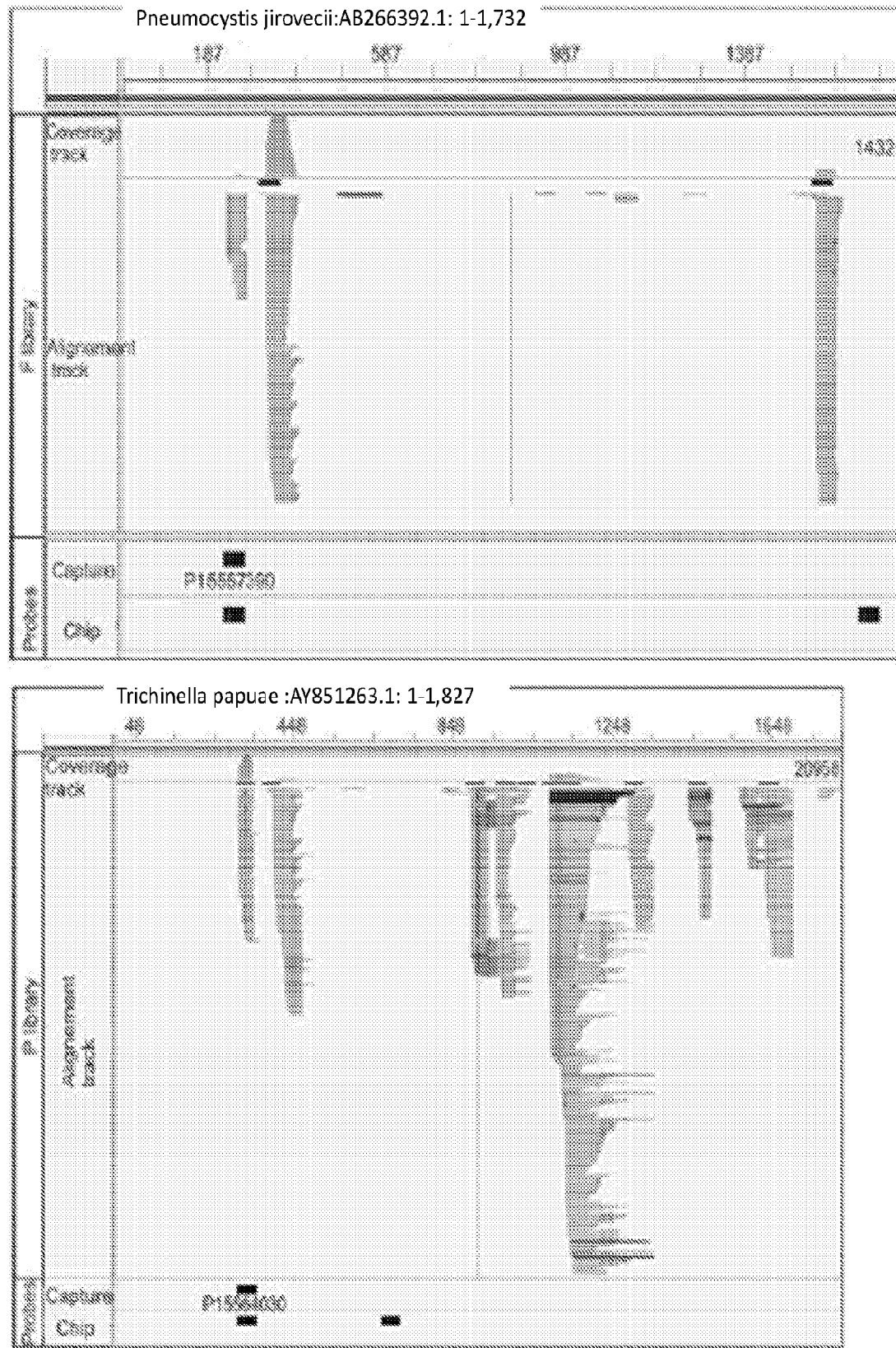


FIG. 5D

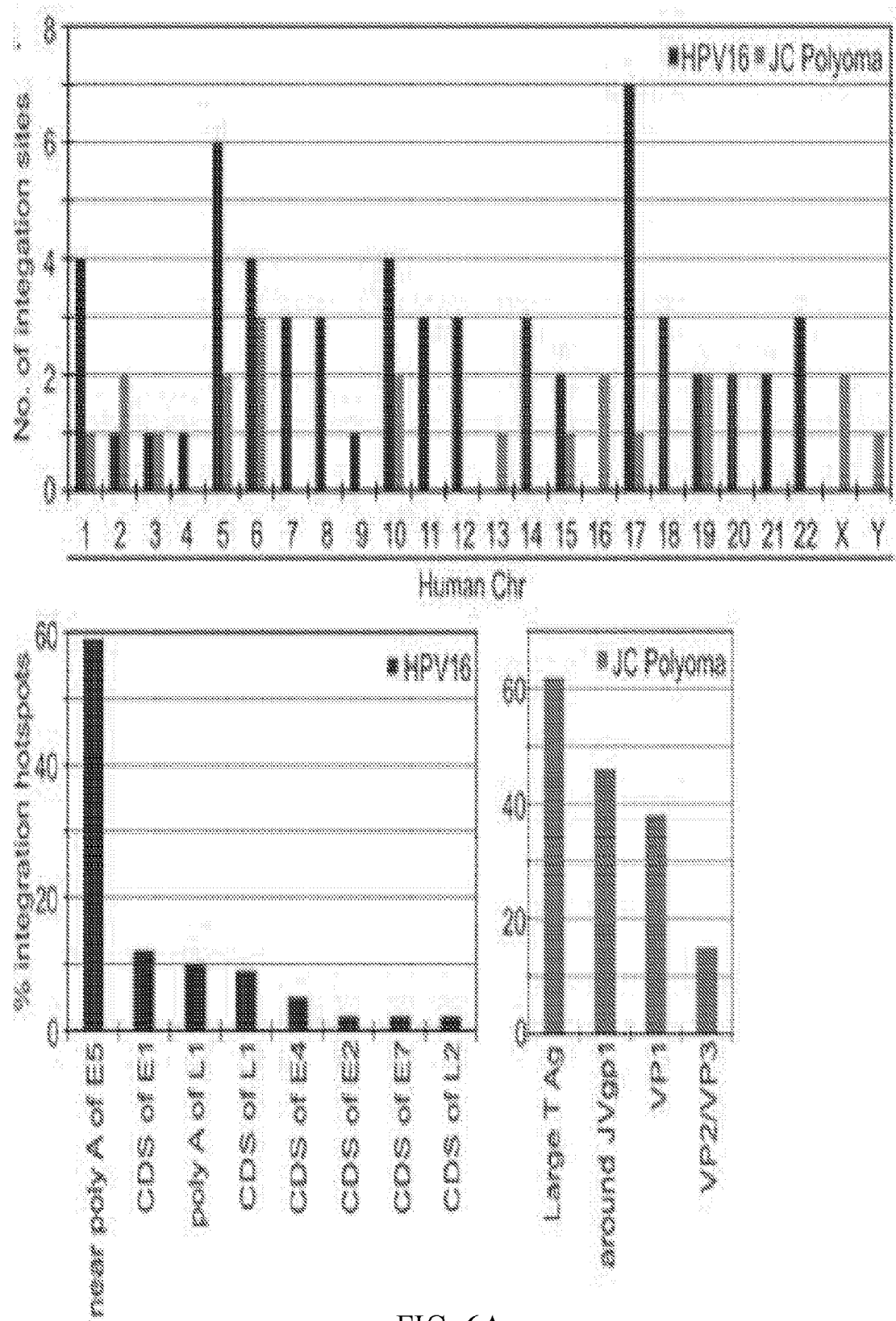


FIG. 6A

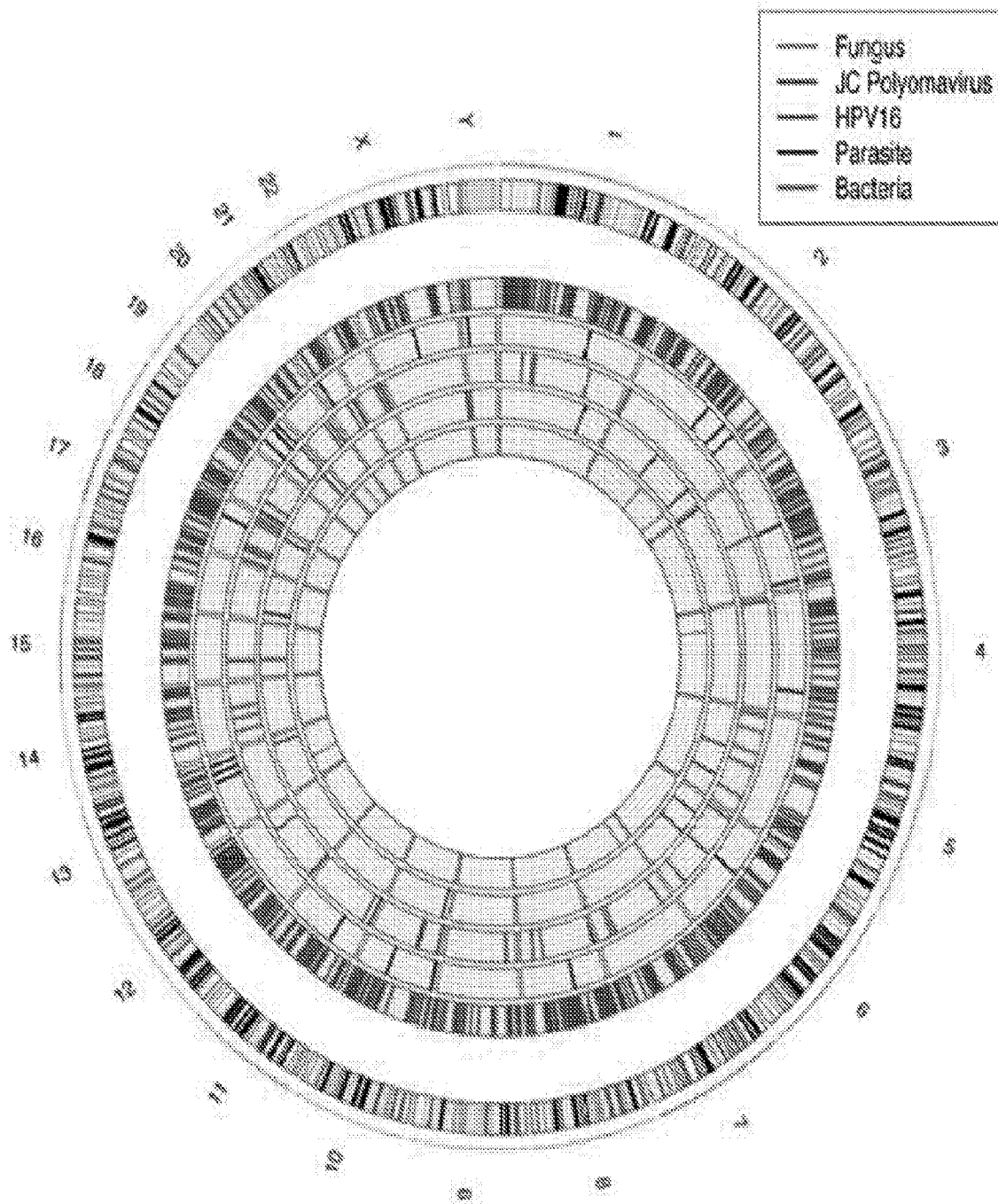


FIG. 6B

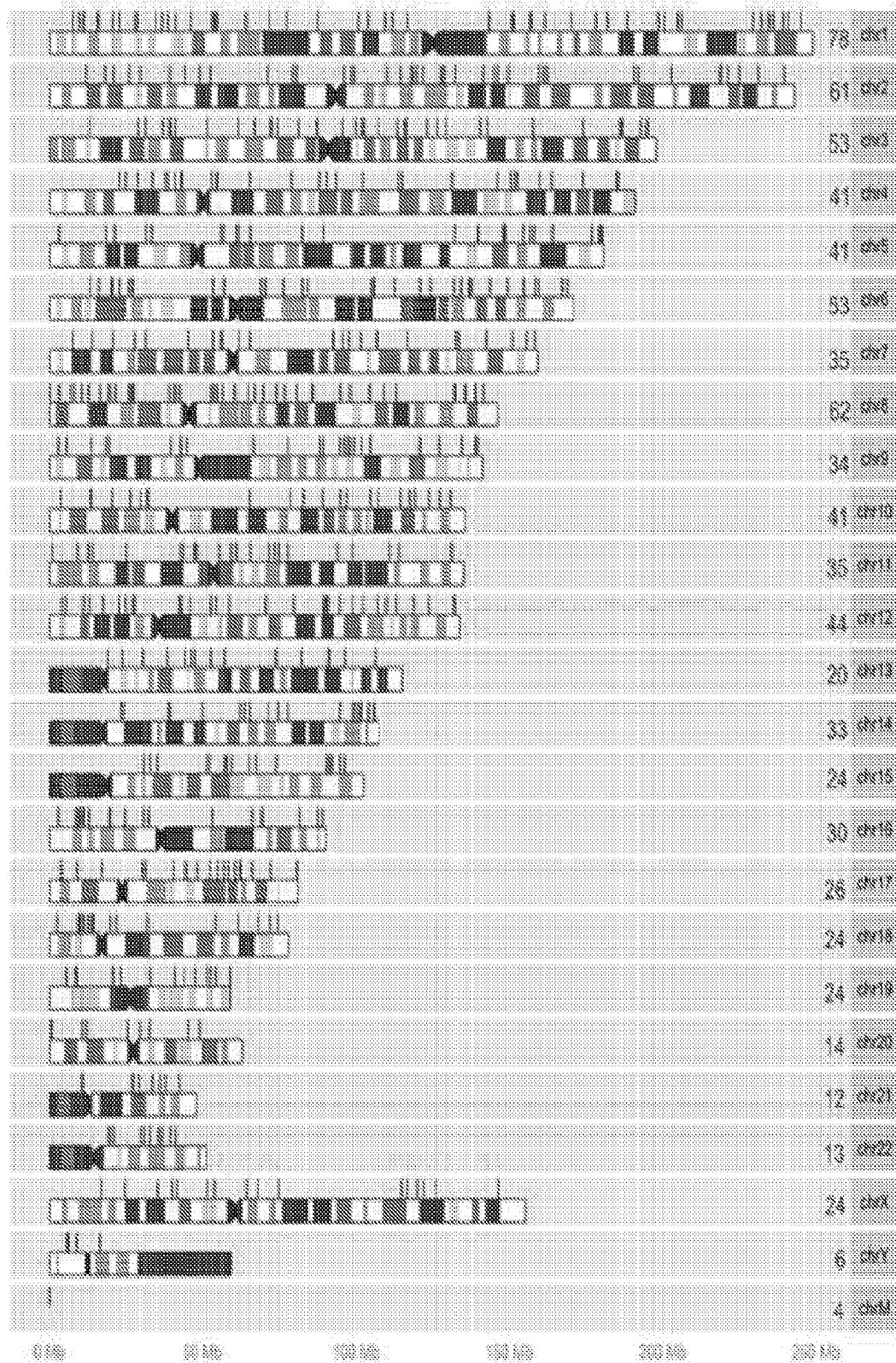


FIG. 6C

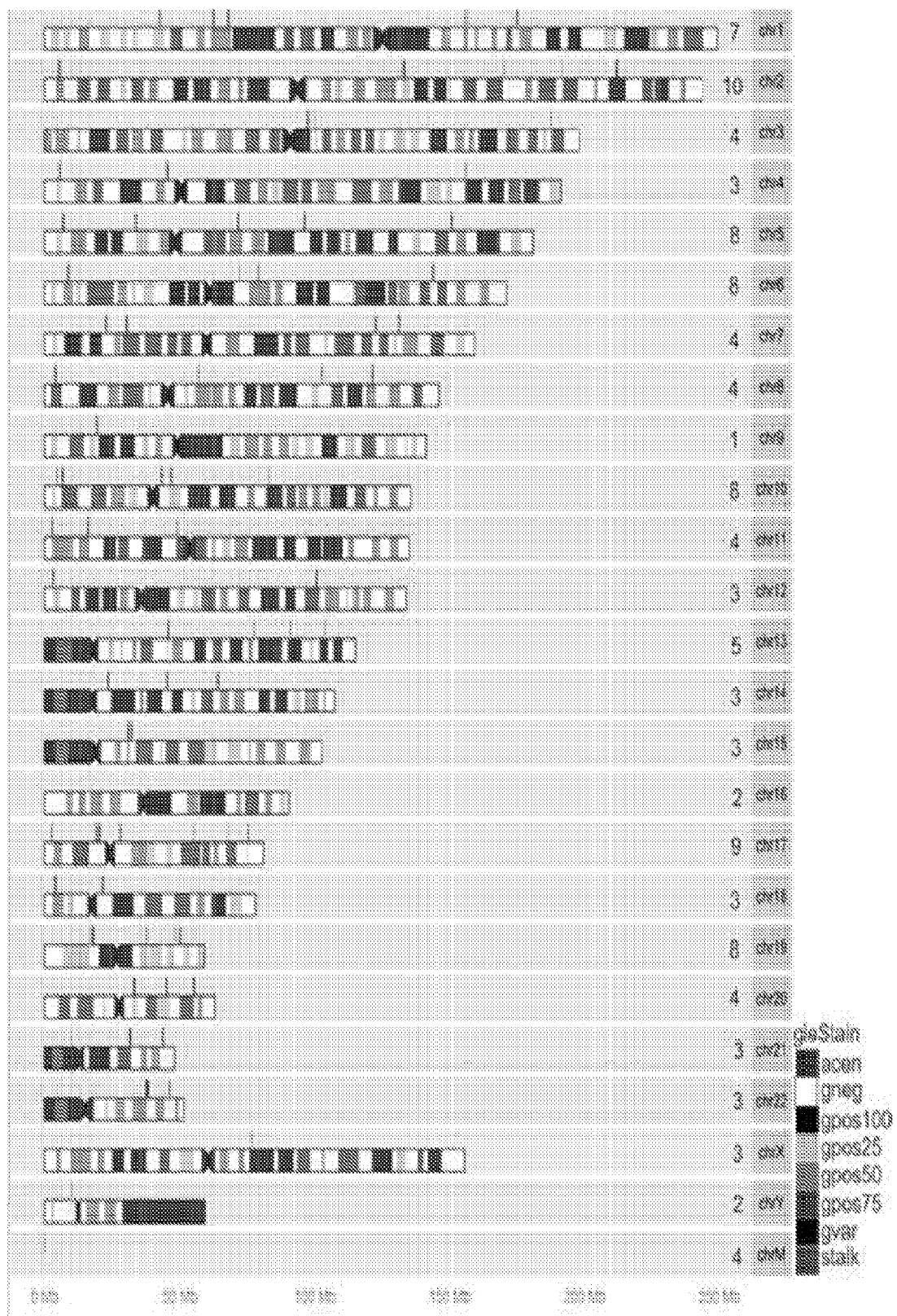
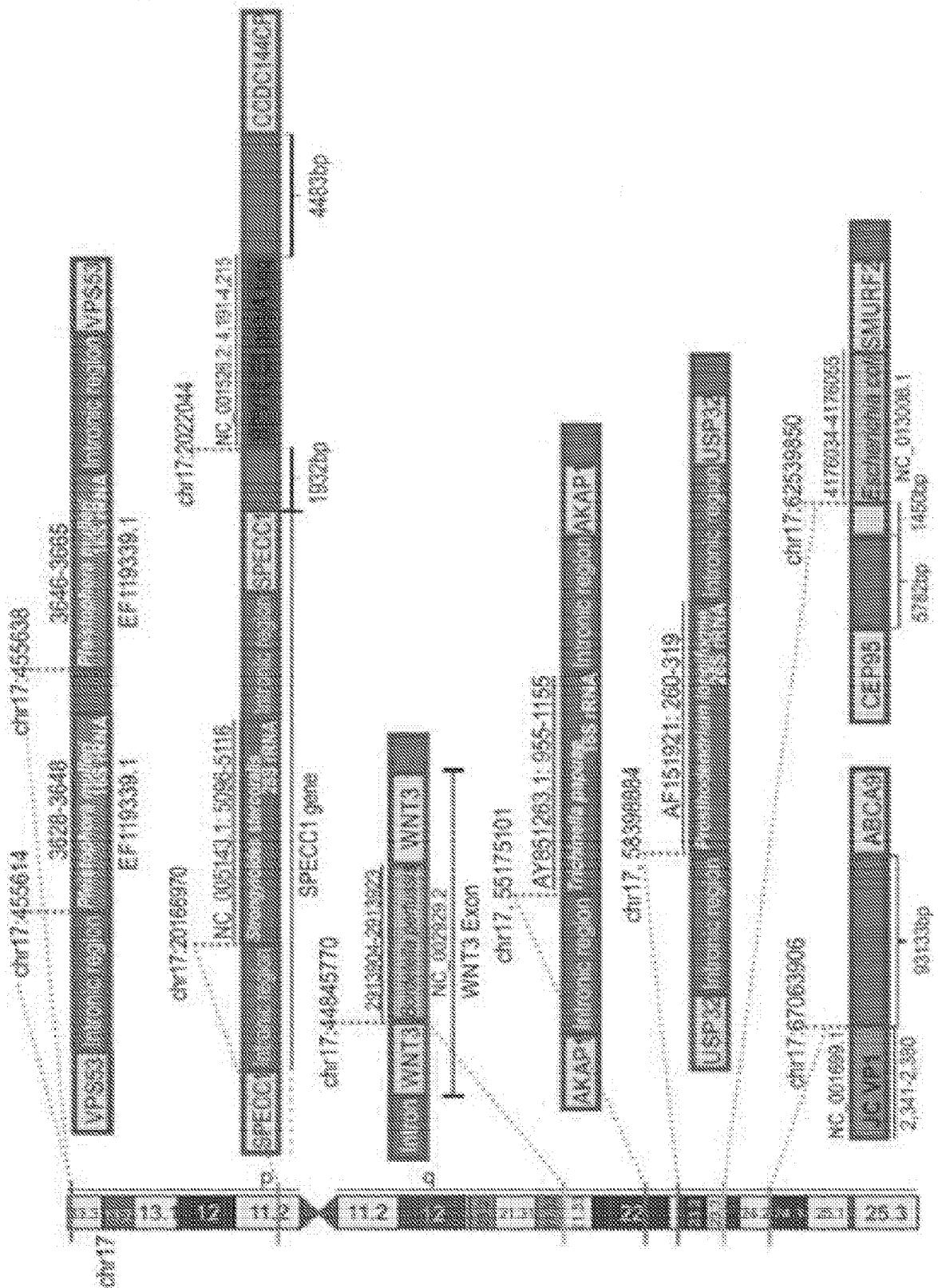


FIG. 6D



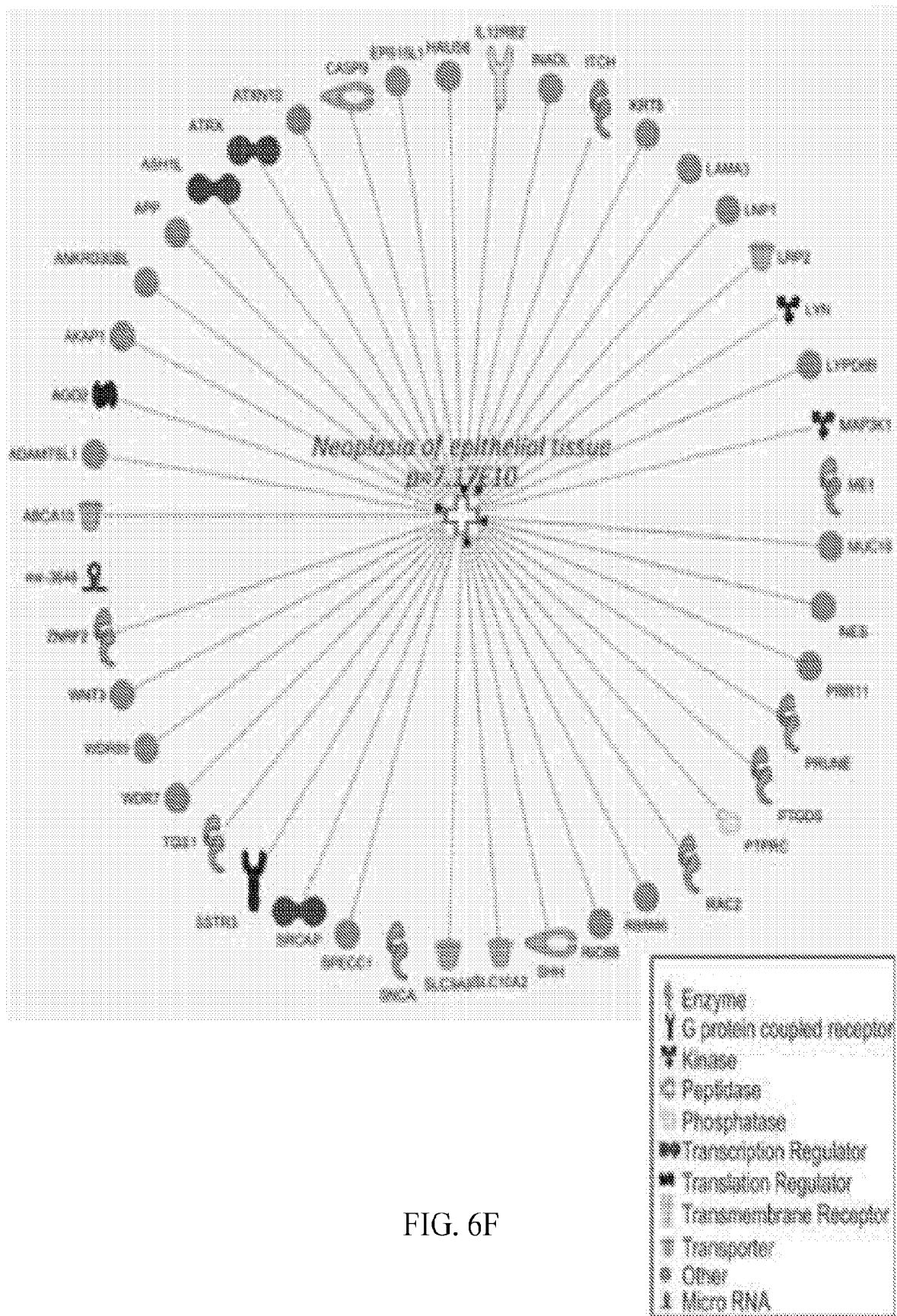


FIG. 6F

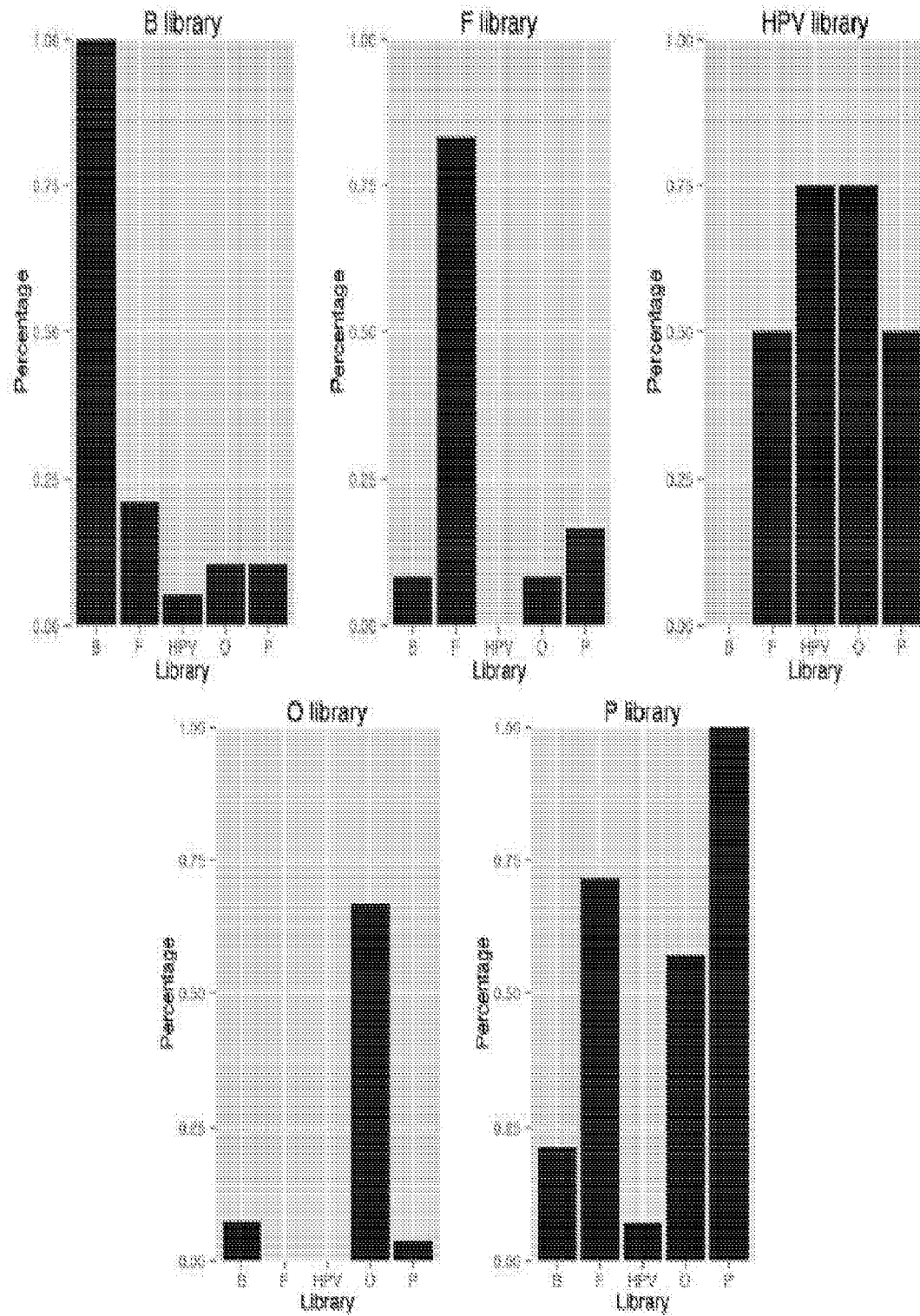


FIG. 7A

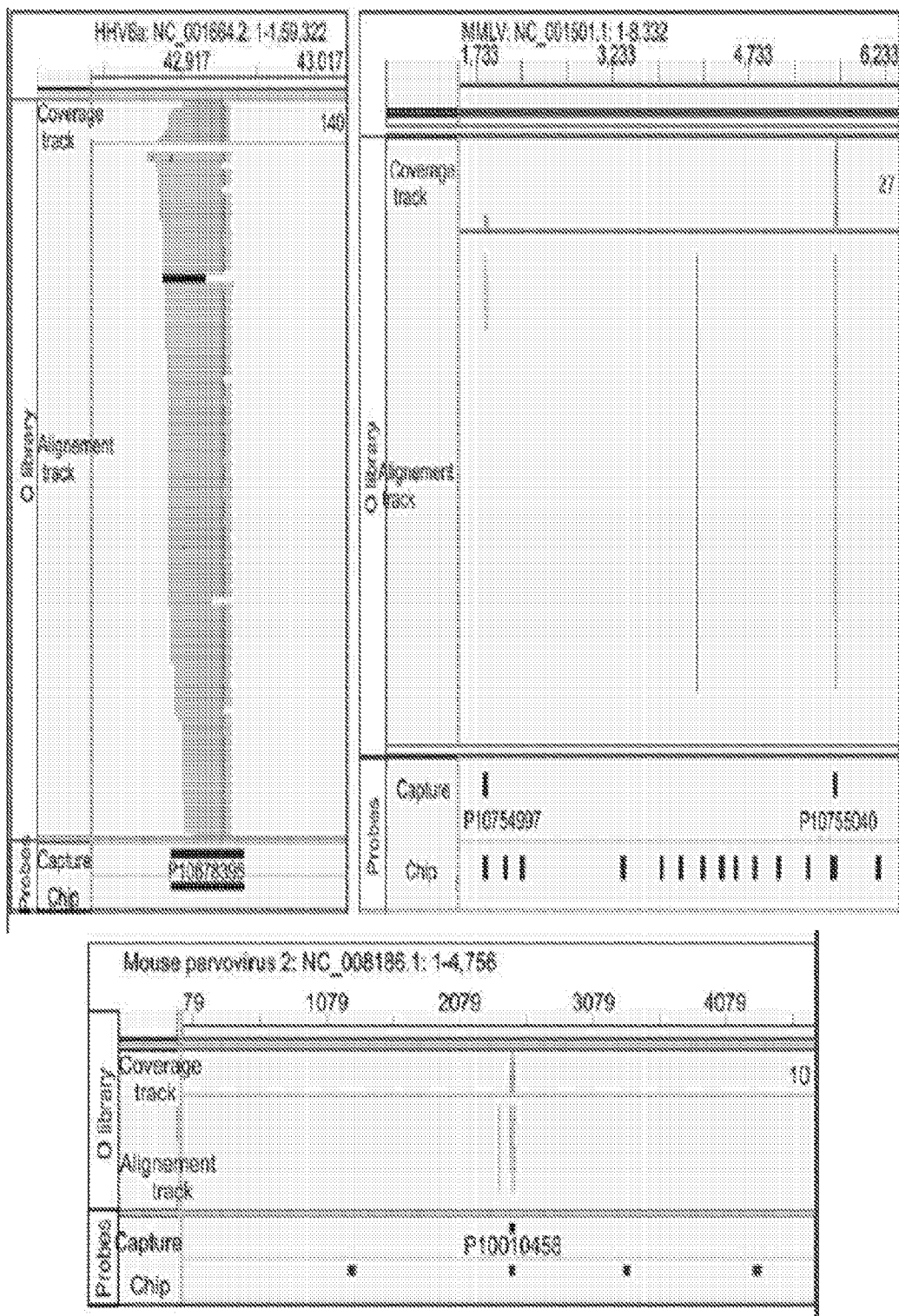


FIG. 7B

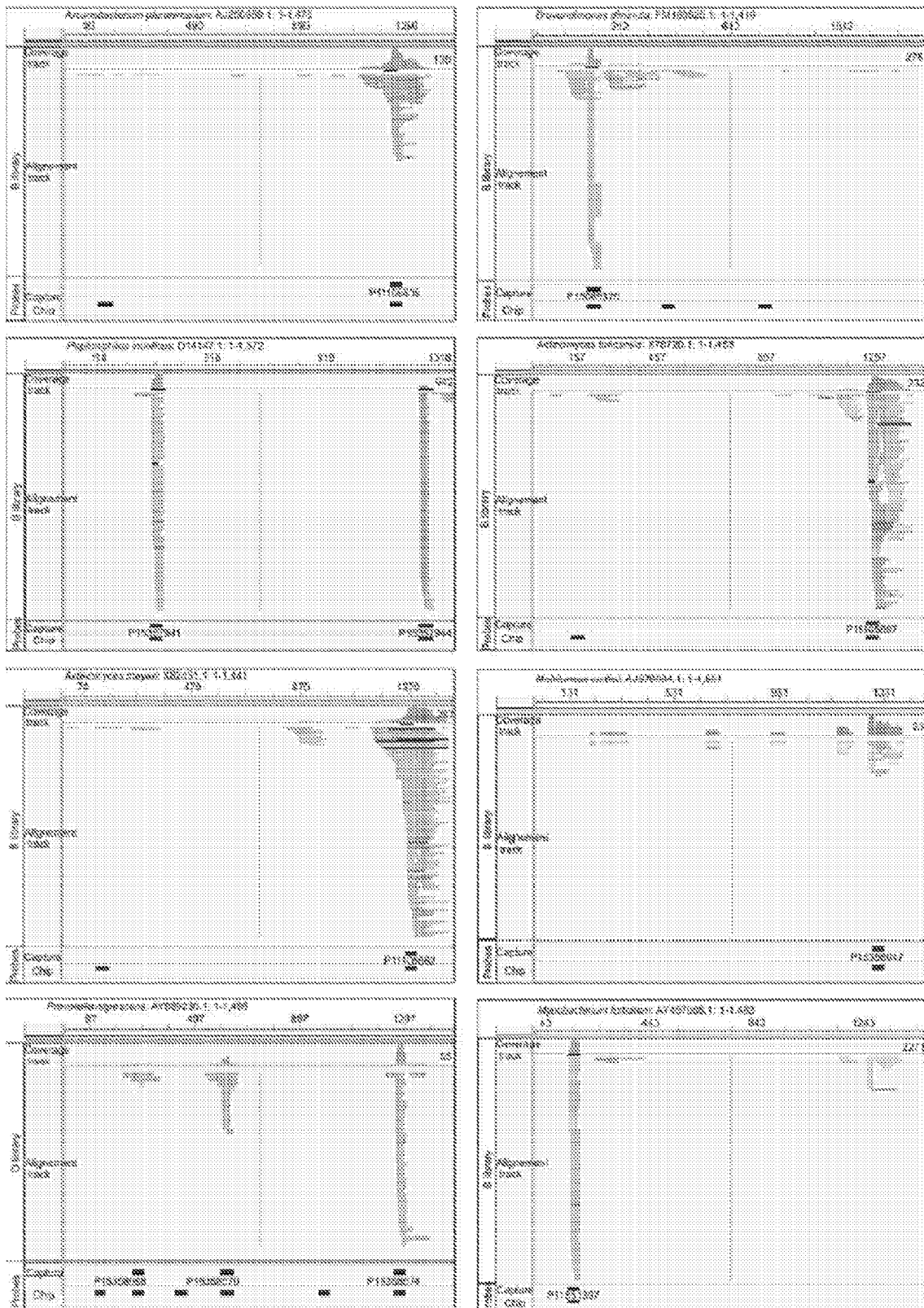


FIG. 7C

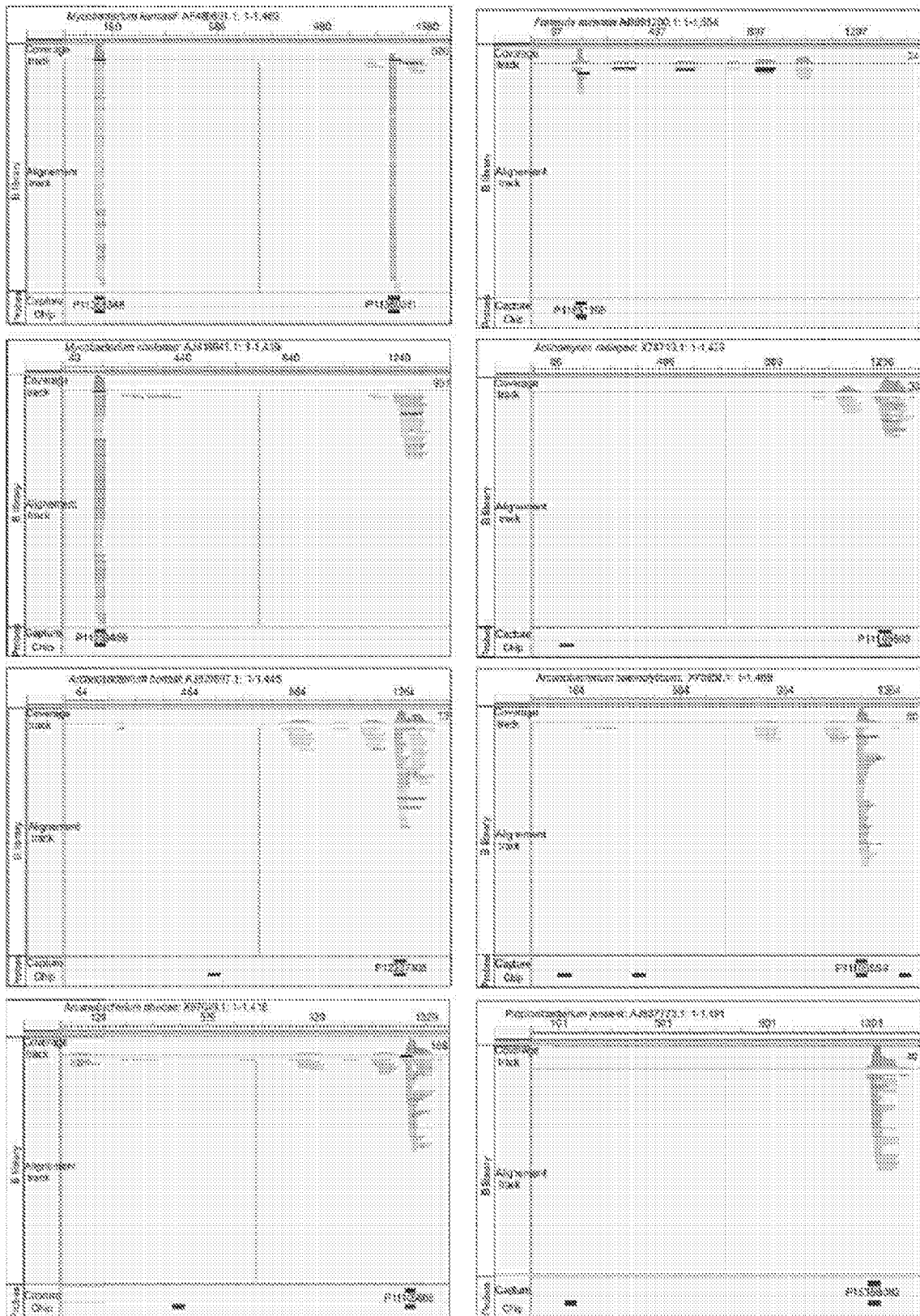


FIG. 7D

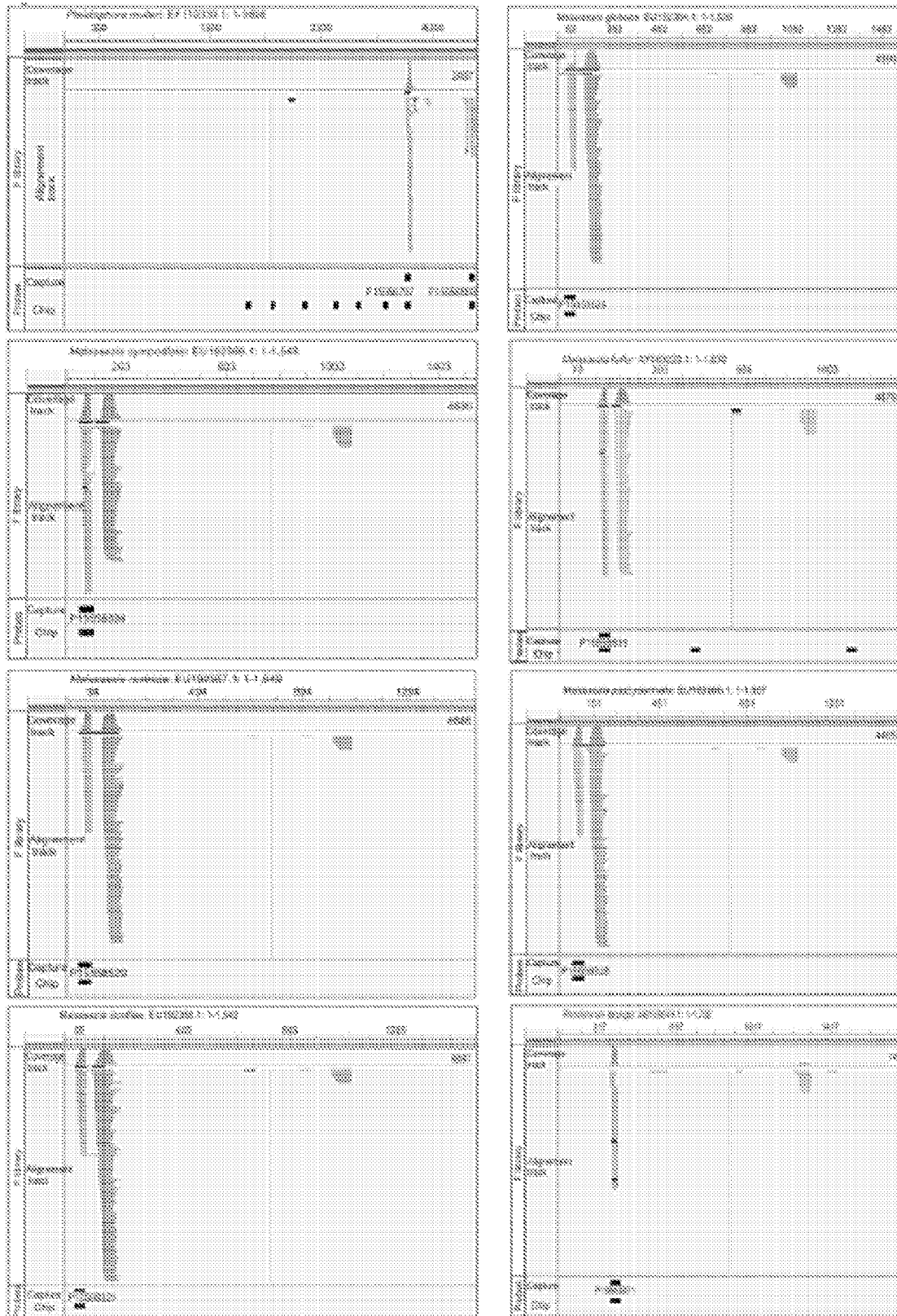


FIG. 7E

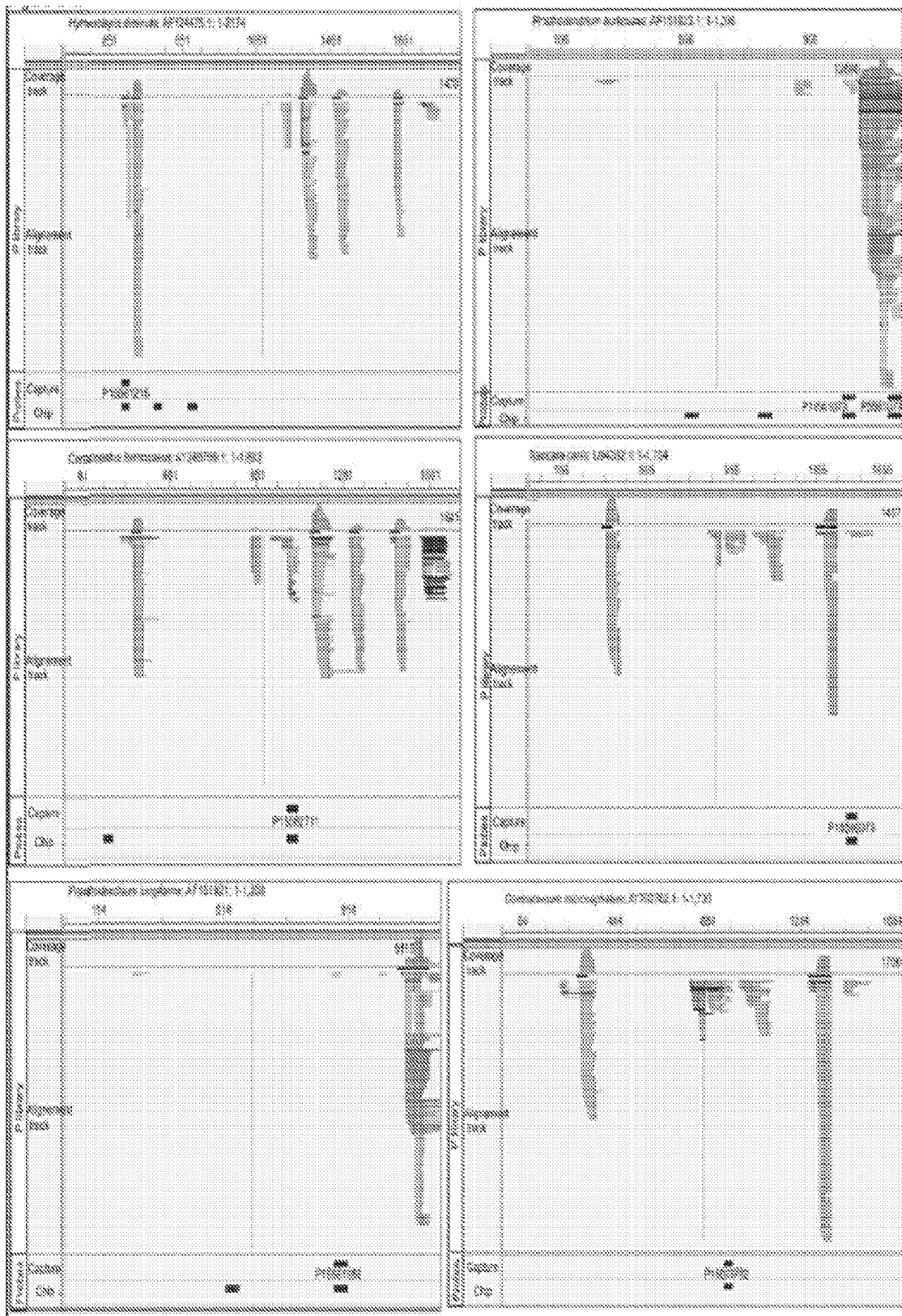


FIG. 7F

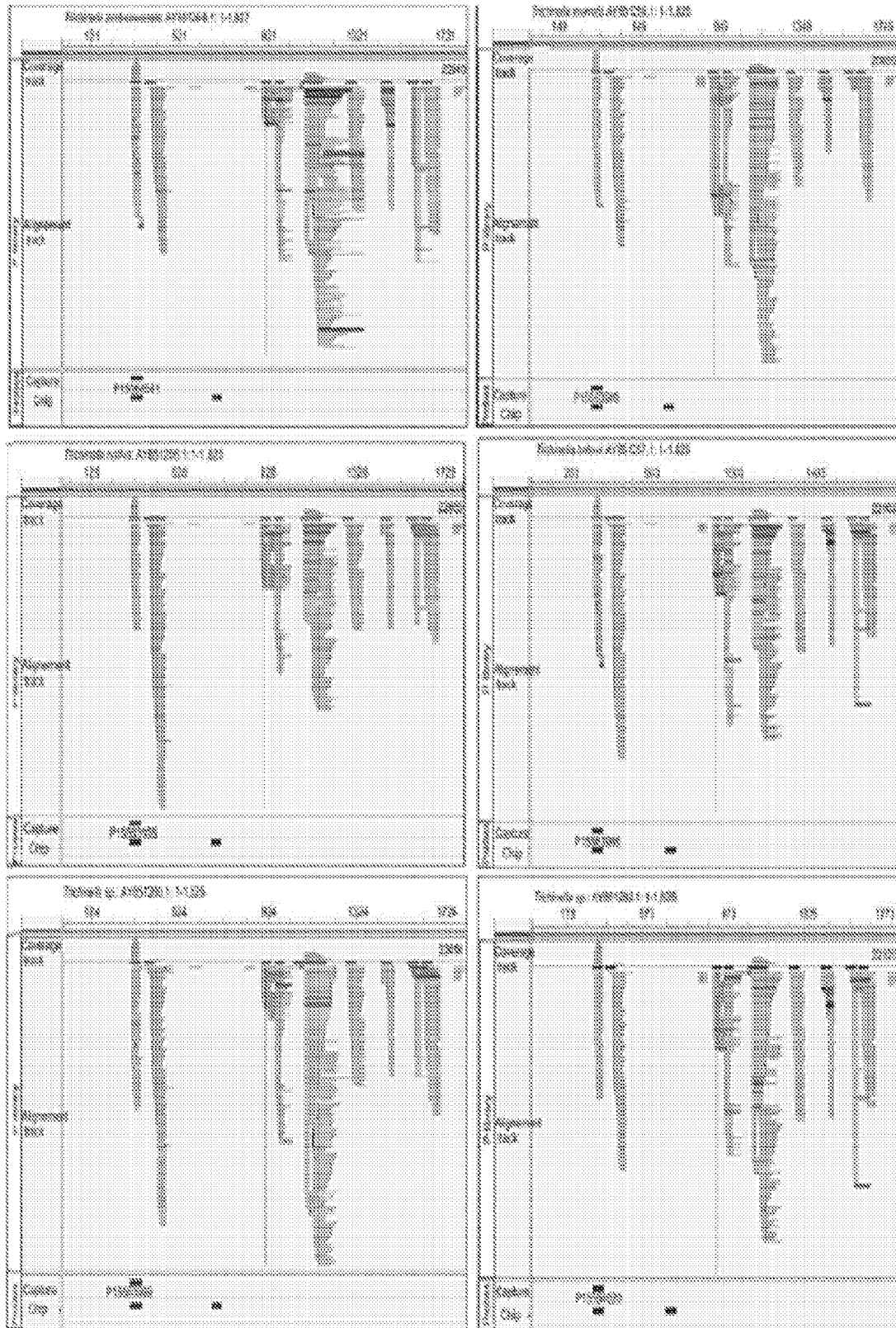


FIG. 7G

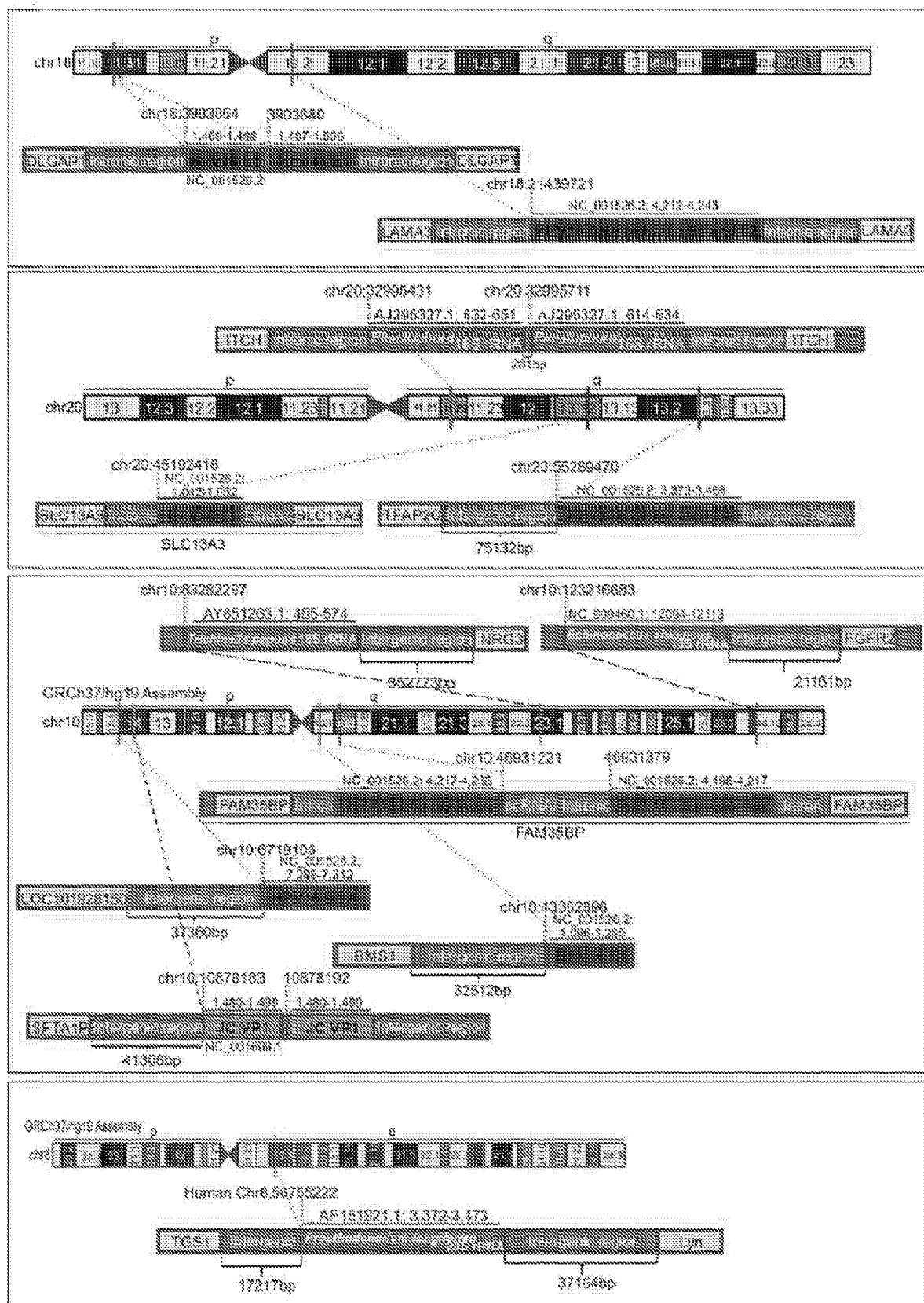


FIG. 8

Microbial insertion region	Human genomic integration sites
HPV16 insertions	
HPV16 4,188-4,243 (hotspot for integration)	Intronic (53% integrations) regions of LAMA3, ATXN10, INADL, ABCA10, EVC2, WDR89, CADPS2, HAU56, EPHA6, FAM179B, COL14A1, MRP927, FUCA2, ADAMTS12, TRIOBP, CSMD1, KCNQL
	Upstream (12%) of genes IL12RB2, LOC384436, LOC79999, FCHQ1, M8PL52, SLC7A7
	Downstream (8%) of the genes NACAP1, GUCA2A/GUCA2B, RSPH1
	Intronic ncRNA gene of the FAM35BP gene (6% integrations)
	Intergenic integrations (6%):
	-upstream of SPECC1; downstream of CCDC144CP -upstream of SSTR3 and downstream of RAC2.
HPV16 E1	Intronic region of SLC13A3, DLGAP1, CCDC155 and ncRNA LOC10028863
	Intergenic regions
HPV16 E2 and E4	Intronic region of LOC10272495
	Intergenic regions
HPV 16 L1	Intronic region of PAFAH1B1, ncRNA LOC10050620
	Intergenic region
L1 PolyA	Intronic regions of DEPDC4
	3'UTR region of the MKLN1
	Intergenic regions
HPV 16 L2	Intronic region of SSH2
JC Polyomavirus Insertions	
JC LT Ag	Intronic regions of CMTR1 and ME1 on chr6; CPO on chr2
	Intergenic regions of chromosomes 1, 2 and 3
VP1 ORF	Intergenic regions, 41Kb downstream of the lncRNA gene SFTA1P (chr10)
	Upstream of ABCA9 (chr17)
	3'UTR of the epigenetic regulator gene MECP2
VP2 and VP3	Intronic region FAM13B (chr 5) and PCCA (chr 13)
Agnoprotein Jcp1	Intronic regions MSH3 (chr 5) and PHLN3 (chr19)
Late coding region (191-258)	Intergenic regions of chromosome 16
	-97Kb downstream of NP1PA7
	-99Kb upstream of the NP1PA5 gene
	Intronic region of SSG5 (chr15)
Bacterial insertions	
Mycobacterium	Exon of ADAMTS11 (chr9)
	Intron of MAP3K1
Aeromonas	Exon of RASSF5 (1q32.1)
	3'UTR of SEC14L4
Sphingomonas	Exon of SRCAP (chr16)
	Exon of WNT3 (chr17)

FIG. 9A

Microbial insertion region	Human genomic integration sites
	5'UTR of C1orf162
	Intron of SLC9A9
<i>Escherichia coli</i>	3'-end of SMURF2
	Intron of CASP9
<i>Campylobacter</i>	3'UTR of COL1A1
<i>Brevundimonas</i>	Intron of RIC8B
<i>Pedococcus</i>	Intron of LYPD6B.
Fungal insertions	
<i>Geotrichum</i>	Intergenic- 550Kb upstream of the GASRC1 gene (chr4)
<i>Pleistophora</i>	Intron of TCH (chr20)
	Intron of MAG1 (chr3)
<i>Phialophora</i>	Intron of ZNF2 (chr7)
<i>Rhodotorula</i>	Intron of CADPS2 (chr7)
Parasitic insertions	
<i>Strongyloides</i>	Exon of ZNF383 (chr19)
	Intron of LNP1 (chr3)
	Intergenic- downstream of SLC10A2 (chr35)
	Intron of SPECC1 (chr17)
<i>Contrassectum</i>	Exon of RHD (chr1)
<i>Trichinella</i>	Intron of AKAP1 (chr17)
	Intron of EP515L1 (chr19)
	Intergenic- 553Kb upstream of NRG3 (chr10)
<i>Echinococcus</i>	Intron of ATEX (chrX)
	21Kb upstream of FGFR2
<i>Prosthodendrium</i>	Intron of USP32 (chr17)
	Intergenic region- 37Kb upstream of lyn gene
<i>Hymenolepis</i>	Downstream of MIR3648 (chr21)
<i>Diphyllobothrium</i>	Intergenic- 105Kb upstream of TRIM49B (chr9)
	in the ncRNA ANKRD308L gene

FIG. 9B

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US17/45898

A. CLASSIFICATION OF SUBJECT MATTER

IPC - C07H 21/02, 21/04; C12Q 1/68, 1/70; G01N 33/58, 33/48, 33/574, 33/58 (2017.01)

CPC - C07H 21/02, 21/04; C12Q 1/6816, 1/6825, 1/6876, 1/6886, 1/6888, 1/701; G01N 33/48, 33/574, 33/57407, 33/57484, 33/58

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)

See Search History document

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

See Search History document

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

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	✓ (BALDWIN, DA et al.) Metagenomic Assay for Identification of Microbial Pathogens in Tumor Tissues. MBio. 16 September 2014, Vol. 5, No. 5; e01714-14; pages 1-13; abstract; page 2, 1st column, 2nd and 3rd paragraphs and 2nd column, 1st and 3rd paragraphs; page 4, 1st column, 5th paragraph to 2nd column, 1st paragraph; page 7, 2nd column, 2nd paragraph; page 9, 2nd column, 1st paragraph; page 11, 1st column, 2nd and 4th paragraphs and 2nd column, 2nd and 3rd paragraphs; Figure 4; DOI: 10.1128/mBio.01714-14	15, 16/15, 19, 20/19
Y		1-2, 4, 6/1-2, 6/4
Y	US 2014/0094383 A1 (OHIO STATE INNOVATION FOUNDATION) 3 April 2014; paragraphs [0006], [0009], [0026], [0027], [0046], [0047], [0061], [0091], [0124], [0146]; Figure 12	1-2, 4, 6/1-2, 6/4
A	US 2015/0093360 A1 (SERES HEALTH, INC.) 2 April 2015; Table 1; paragraphs [0031], [0032], [0071], [0323], [0324]	3, 5, 6/3, 6/5, 12-14, 16/13-14, 17, 18, 20/17-18
A	✓ (LEE, YJ et al.) The PathoChip, a functional gene array for assessing pathogenic properties of diverse microbial communities. The ISME Journal. October 2013, Epub 13 June 2013, Vol. 7, No. 10; pages 1974-84; DOI: 10.1038/ismej.2013.88	1-5, 6/1-5, 12-15, 16/13-15, 17-19, 20/17-19
P, X	✓ (BANERJEE, S et al.) Microbial Signatures Associated with Oropharyngeal and Oral Squamous Cell Carcinomas. Scientific Reports. 22 June 2017, Vol. 7, No. 1; pages 1-20; DOI: 10.1038/s41598-017-03466-6	1-5, 6/1-5, 12-15, 16/13-15, 17-19, 20/17-19

☐ 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 another 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

30 November 2017 (30.11.2017)

Date of mailing of the international search report

02 JAN 2018

Name and mailing address of the ISA/

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Shane Thomas

PCT Helpdesk: 571-272-4300

PCT OSP: 571-272-7774

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US17/45898

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.:
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:
3. ☒ Claims Nos.: 7-11
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:

-Please see supplemental page-

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 additional fees, this Authority did not invite payment of additional fees.
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.:
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.:
Groups I+, Claims 1-6, 12-20 and SEQ ID NOs: 1-3

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.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/US17/45898

-Continued from Box No. III: Observations where unity of invention is lacking-

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fees must be paid.

Groups I+, Claims 1-6, 12-20 and SEQ ID NOs: 1-3 are directed toward a method of detecting oral squamous cell carcinoma in a tumor tissue sample from a subject; compositions, microarrays and kits therefor.

The method, compositions and kits will be searched to the extent they encompass a first probe nucleic acid encompassing SEQ ID NO: 1 (first exemplary probe sequence), a second probe nucleic acid encompassing SEQ ID NO: 2 (second exemplary probe sequence), and a third probe nucleic acid encompassing SEQ ID NO: 3 (third exemplary probe sequence). Applicant is invited to elect additional probe sequence(s), with specified SEQ ID NO: for each, to be searched. Additional probe sequence(s) will be searched upon the payment of additional fees. It is believed that claims 1, 2, 3 (in-part), 4, 5 (in-part), 6 (in-part), 12 (in-part), 13 (in-part), 14 (in-part), 15, 16 (in-part), 17 (in-part), 18 (in-part), 19 and 20 (in-part) encompass this first named invention and thus these claims will be searched without fee to the extent that they encompass SEQ ID NO: 1 (probe sequence), SEQ ID NO: 2 (probe sequence), and SEQ ID NO: 3 (probe sequence). Failure to clearly identify how any paid additional invention fees are to be applied to the "+" group(s) will result in only the first claimed invention to be searched/examined. An exemplary election would be a probe encompassing SEQ ID NO: 4 (first exemplary elected probe sequence).

No technical features are shared between the organisms and/or probe sequences of Groups I+ and, accordingly, these groups lack unity a priori.

Groups I+ share the technical features including: a method of detecting oral squamous cell carcinoma in a tumor tissue sample from a subject, the method comprising: hybridizing a detectably-labeled nucleic acid from the tumor tissue sample to a first microarray comprising at least three nucleic acid probes from microbes to generate a first hybridization pattern; hybridizing a detectably-labeled nucleic acid from a reference sample to a second microarray comprising at least three nucleic acid probes from said group of microbes to generate a second hybridization pattern, wherein the reference sample is from an otherwise identical non-tumor tissue from a subject; comparing the first and second hybridization patterns, wherein when the first hybridization pattern is substantially a microbial hybridization signature and the second hybridization pattern is substantially not a microbial hybridization signature, oral squamous cell carcinoma is detected in the tumor tissue sample; a composition comprising at least three nucleic acid probes; a microarray comprising at least three nucleic acid probes; a kit comprising at least three nucleic acid probes; and a kit comprising a microarray comprising at least three nucleic acid probes from microbes and instructional material for use thereof.

However, these shared technical features are previously disclosed by US 2014/0094383 A1 to Ohio State Innovation Foundation (hereinafter 'Ohio') in view of the article 'Metagenomic Assay for Identification of Microbial Pathogens in Tumor Tissues' by Baldwin et al. (hereinafter 'Baldwin').

Ohio discloses a method of detecting (a method of detecting; paragraph [0006]) squamous cell carcinoma (squamous cell carcinoma; paragraph [0046]) in a tumor tissue sample from a subject (in a biopsy (tumor tissue sample) from a subject; paragraphs [0006], [0048]), the method comprising: hybridizing a detectably-labeled nucleic acid from the tumor tissue sample (the method comprising: hybridizing a detectably-labeled nucleic acid from the tumor tissue sample; paragraphs [0006], [0048], [0060], [0061]) to a first microarray (to a first microarray; paragraph [0006]) comprising nucleic acid probes (comprising nucleic acid probes; paragraph [0078]) from microbes (from viruses or bacteria (from microbes); paragraph [0006]) to generate a first hybridization pattern (to generate a first hybridization pattern; paragraphs [0006], [0061]); a composition comprising nucleic acid probes (a composition comprising nucleic acid probes; abstract, paragraphs [0059], [0078]); a microarray (a microarray; paragraphs [0006], [0107]) comprising nucleic acid probes (comprising nucleic acid probes; paragraphs [0006], [0078]); a kit comprising nucleic acid probes (a kit comprising nucleic acid probes; paragraphs [0078], [0104]); and a kit (a kit; paragraph [0104]) comprising a microarray (a microarray; paragraphs [0006], [0107]) comprising nucleic acid probes (comprising nucleic acid probes; paragraphs [0006], [0078]) from microbes (from viruses or bacteria (from microbes); paragraph [0006]). Ohio further discloses the comparison of results from detection processes on patient and normal samples (comparison of results from detection processes on patient and normal samples; paragraphs [0026], [0129]).

Ohio does not disclose detecting oral squamous cell carcinoma; at least three nucleic acid probes; hybridizing a detectably-labeled nucleic acid from a reference sample to a second microarray comprising at least three nucleic acid probes from said group of microbes to generate a second hybridization pattern, wherein the reference sample is from an otherwise identical non-tumor tissue from a subject; comparing the first and second hybridization patterns, wherein when the first hybridization pattern is substantially a microbial hybridization signature and the second hybridization pattern is substantially not a microbial hybridization signature, oral squamous cell carcinoma is detected in the tumor tissue sample; and instructional material for use of kit materials.

Baldwin discloses screening samples from oral squamous cell carcinoma patients (screening samples from oral squamous cell carcinoma patients; page 4, first column, fifth paragraph - second column, first paragraph) using microarrays comprising 2,085 probes for microbes (using microarrays comprising 2,085 probes for microbes; page 2, first column, fourth paragraph - second column, first paragraph).

It would have been obvious to a person of ordinary skill in the art at the time of the invention was made to have provided instructional materials in a kit, as disclosed by Ohio, in order to better enable a user of the kit to properly perform detection methods associated with the kit, based on the disclosure of Ohio. It further would have been obvious to a person of ordinary skill in the art at the time of the invention was made to have modified the disclosure of Ohio to have included the use of at least three or more probes for the screening of cancer samples for the detection of oral squamous cell carcinoma, as disclosed by Baldwin, using probes specific for microbes, and comparing the pattern of hybridization to normal control samples, based on the disclosure of Ohio, in order to enable the determination of a pattern associated with a specific diagnosis of oral squamous cell carcinoma.

Since none of the special technical features of the Groups I+ inventions is found in more than one of the inventions, and since all of the shared technical features are previously disclosed by a combination of the Ohio and Baldwin references, unity of invention is lacking.