Application Note

Agilent
Trusted Answers

Development and Utilization of a PathoChip Array to Detect Distinct Pathogenic Microbial Signatures in FFPE Cancer Tissues

Authors

Erle S. Robertson
Department of Microbiology
University of Pennsylvania
Philadelphia, PA

Josh Zhiyong Wang
Agilent Technologies
Santa Clara, CA

Abstract

Microbiotic balance between beneficial and harmful microbes is being increasingly recognized as an important contributor to human health. This balance can affect metabolism and immune responses, while infectious pathogenic agents (including HPV and *H. pylori*) are one of the highest risk factors in cancer development. Due to these findings, screening for thousands of viruses, pathogenic bacteria/fungi, and parasites in human tumor tissues could help to uncover their role in cancer progression as well as therapeutic responses to treatment. Here we describe the development of a PathoChip array that can detect both DNA and RNA from thousands of viruses and pathogenic microbes using formalin-fixed paraffin-embedded (FFPE) tumor tissues. This assay includes several upstream sample preparation procedures and downstream data analysis procedures unique to pathogen detection. Furthermore, we show that this assay can successfully detect distinct microbial signatures in diverse cancer tissues including triple-negative breast cancer and oral/oropharyngeal squamous cell carcinoma. These cancers were identified with signature viruses and microbial pathogens that are suitable for validation using independent PCR or capture sequencing methods. These results demonstrate that PathoChip can be successfully used to obtain comprehensive pathogen information in challenging FFPE samples.

Introduction

The normal human microbiome comprises thousands of microbial species. Intense research has focused on tissue systems known to have resident microbiomes, including the gastrointestinal tract, skin, airway, and immune system. Infectious agents such as viruses, bacteria, and parasites can be major contributors to cancers in various tissues including liver, stomach, cervix, and blood.

Many methods, including traditional cultures, exist to detect microbes. However, metagenomic tools based on microbial genetic information are increasingly necessary to efficiently identify infectious agents associated with a particular disease. PCR amplification followed by 16s rRNA sequencing is a popular method to identify bacterial species, but is incompatible with viruses or eukaryotic microorganisms. Shotgun sequencing of total DNA from a sample allows unbiased detection but suffers from severely reduced efficiency due to high background from host human DNA.

Meanwhile, DNA microarrays have emerged due to their capability to quickly and economically screen large numbers of samples for broad microbial content. There are commercially available solutions, but the drawback is that they cover distinct (and, in some cases, overlapping) subsets of microbes, but none gets the full picture.

In this application note, we describe development of the PathoChip array (Figure 1) that is based on Agilent SurePrint microarray technology. The PathoChip array contains probes for all known publicly available virus sequences and hundreds of pathogenic bacteria, fungi, and helminths, providing wide coverage of microbial pathogens in an economical format. Where possible, multiple probes targeting independent regions of each genome are used to improve detection. Furthermore, while the PathoChip probe content was developed from sequences to known targets, the array maintains the capability to discover new strains or organisms. This feature was accomplished via the inclusion of probes to sequences that are conserved within and between viral families. A supporting workflow is described for profiling FFPE tumor samples, including simultaneous detection of DNA and RNA targets.

PathoChip arrays were successfully used to analyze hundreds of FFPE samples from triple-negative breast cancer and oral/oropharyngeal squamous cell carcinomas to detect microbial signature suitable for validation using independent PCR or capture sequencing methods.

Materials and Methods

Microarry Design

National Center for Biotechnology Information (NCBI) databases for genome, gene, and nucleotide accessions were queried (http://www.ncbi.nlm.nih.gov/pubmed) for all taxonomic virus annotations and for accessions from prokaryotic and eukaryotic human pathogen lists compiled by literature searches and Web resources (http://www.niaid.nih.gov: Emerging and Re-emerging Infectious Diseases, Category A, B, and C Priority Pathogens). The resulting accession sequences were assembled into a metagenome divided into 58 virtual “chromosomes” each composed of approximately 5 to 10 million nucleotides (nts) in length. Both unique and conserved regions in this metagenome were identified and included as described[6].

Probe sequences against both unique and conserved regions in the metagenome (maximum of 60 nts) were designed using the Agilent array comparative genomic hybridization (aCGH) design algorithm. These sequences were then filtered for low likelihood of cross hybridization to human genomic sequences to reduce hybridization background noise.

Probes mapping to unique or conserved regions of pathogen genomes, or any prokaryotic or eukaryotic pathogen accession, were added to the microarray design by default if fewer than 10 probes were available for the source accession. When more than 10 probes/accession were created, probes were filtered using following criteria: maximum 20 probes/accession, minimum inter probe spacing of 100 bp and even distribution covering the entirety of the sequence. Oncogenic viral agents were not restricted for number of probes, creating a saturation tiling probe set to cover the totality of them[6].

The Agilent SurePrint microarray 8x60K format was used to allow 8 samples per slide to be processed. Initially two arrays have been created to evaluate probe performance, PathoChip v2a and v2b, the arrays contained 60,000 probes targeting unique and conserved regions, respectively. High-performing probes from PathoChip v2a and v2b were then combined into PathoChip v3, which contains 37,704 probes to unique targets and 23,627 probes to conserved targets. This allows coverage of all oncogenic and pathogenic viral agents. PathoChip v3 was utilized to analyze the various FFPE cancer tissues described in this application note.

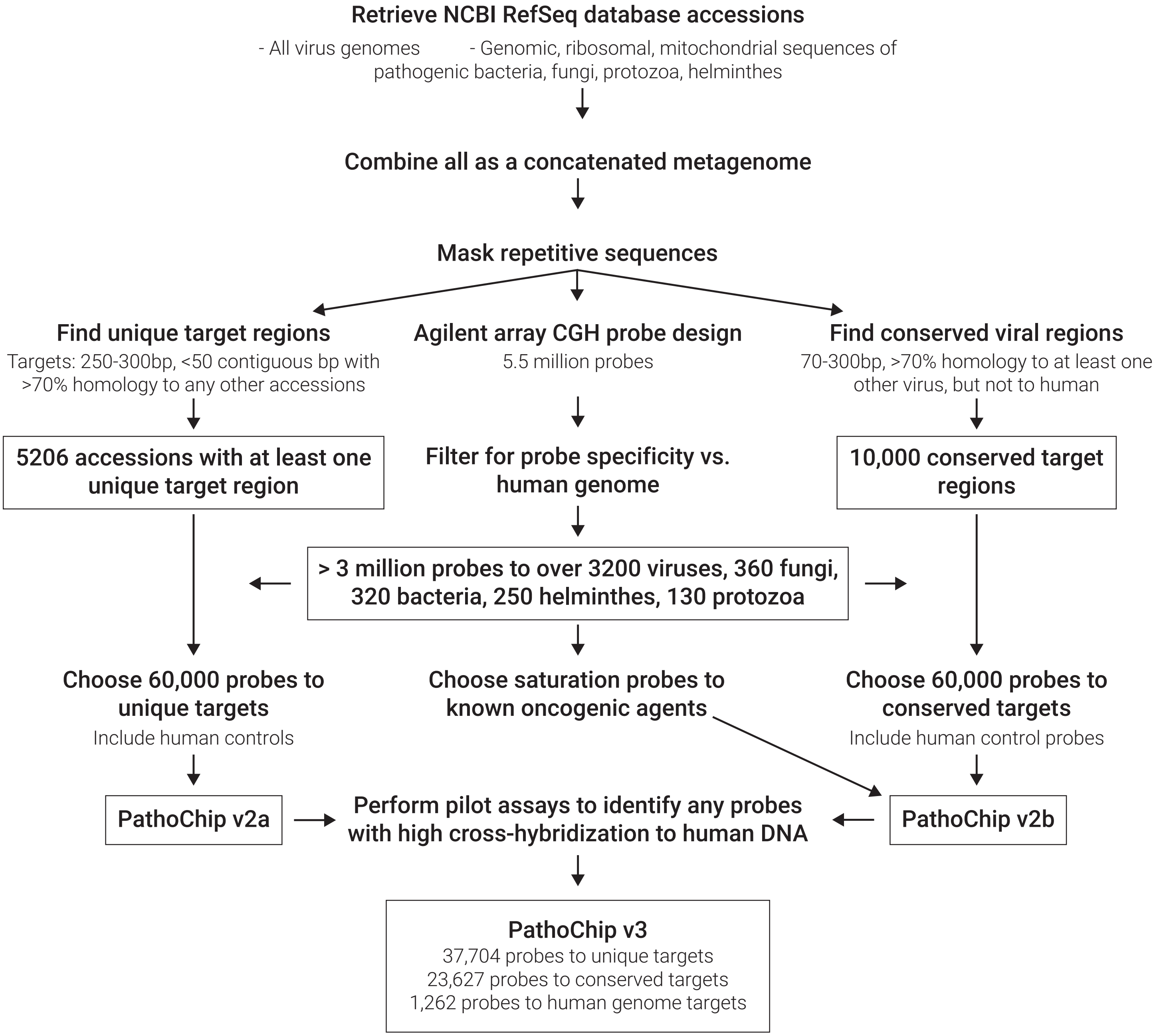

Retrieve NCBI RefSeq database accessions
- All virus genomes
- Genomic, ribosomal, mitochondrial sequences of pathogenic bacteria, fungi, protozoa, helminthes
Combine all as a concatenated metagenome
Mask repetitive sequences
Find unique target regions
Targets: 250-300bp, <50 contiguous bp with >70% homology to any other accessions
Agilent array CGH probe design
5.5 million probes
Find conserved viral regions
70-300bp, >70% homology to at least one other virus, but not to human
5206 accessions with at least one unique target region
Filter for probe specificity vs. human genome
10,000 conserved target regions
> 3 million probes to over 3200 viruses, 360 fungi, 320 bacteria, 250 helminthes, 130 protozoa
Choose 60,000 probes to unique targets
Include human controls
Choose saturation probes to known oncogenic agents
Choose 60,000 probes to conserved targets
Include human control probes
PathoChip v2a
Perform pilot assays to identify any probes with high cross-hybridization to human DNA
PathoChip v2b
PathoChip v3
37,704 probes to unique targets
23,627 probes to conserved targets
1,262 probes to human genome targets

Figure 1. PathoChip design scheme and design iterations

Samples

All samples were obtained from the Abramson Cancer Center's Tumor Tissue and Biospecimen Bank. A resident pathologist reviewed case history and confirmed tumor type and demarcation of the cancer cells. If significant adjacent normal tissue was present, sections were mounted on noncharged glass slides for dissection of tumor tissue using a template slide with a hematoxylin-and-eosin stained section and the cancer region clearly demarcated.

100 de-identified FFPE oral cavity and oropharyngeal squamous cell carcinoma samples (collectively referred to as OCSCC) were received as 10 µm sections on noncharged glass slides. 20 each of matched and non-matched control samples were received[7]. Clinically normal samples next to the cancers are referred here as "matched controls" as they were obtained from 20 cancer patients included in the study. Non-matched controls were oral tissues (uvula) obtained from otherwise healthy individuals.

100 de-identified FFPE triple-negative breast cancer samples were received in the form of 10 µm sections on noncharged glass slides. 17 matched and 20 non-matched control samples were provided as paraffin rolls[8]. Matched controls were obtained from the adjacent non-cancerous breast tissue of the same patient from which the cancer tissues are obtained. Non-matched controls were breast tissues obtained from healthy individuals.

Sample DNA/RNA Extraction, Amplification, Labeling, and Hybridization

While standard Agilent CGH assays use genomic DNA as starting material, we have sought to modify our procedure due to the fact viruses can use either DNA or RNA as genetic material. Specifically, FFPE tumors were used for sequential DNA and RNA extraction using the AllPrep DNA/RNA FFPE kit (QIAGEN). Nucleic acid quality control assessments included A260/280 ratios, yield, and size distribution by agarose gel electrophoresis. Although some FFPE RNA samples showed partial degradation, the RNA fragment sizes and extraction yields were sufficient for most samples, allowing total RNA to be directly used for cDNA generation.

50 ng of test sample genomic DNA and 50 ng of test sample total RNA were used together as input for TransPlex WTA-2 kit (Sigma-Aldrich). This kit enables the co-amplification of both genomic DNA and cDNA simultaneously. Amplification products were purified with the QIAquick PCR purification kit (QIAGEN), and 2 µg of purified amplification product were used for Cy3 dye labeling on a test sample with the SureTag labeling kit (Agilent). For the reference sample, 50 ng of a viral-negative reference genomic DNA (BJAB cell line) were amplified using the TransPlex
WTA-2 kit and 2 µg of purified amplification products were used for Cy5 dye labeling with the Agilent SureTag labeling kit. This reference sample served as a control to report probe cross-hybridization to human DNA.

Cy3- and Cy5-labeled DNAs were purified with spin columns included in the SureTag labeling kit. The specific activities of labeled DNAs were measured, then mixed for hybridization. Labeled DNAs were hybridized to PathoChip v3 8x60K arrays for 40 hours following Agilent recommendations of a 65°C hybridization temperature with 20-rpm rotation in an Agilent hybridization oven. Arrays were processed using the standard wash procedure and scanned on an Agilent SureScan microarray scanner C or D (p/n G2565AA & G4900DA).

Microarray Data Analysis

Scanned microarray images were analyzed using the Agilent Feature Extraction software to calculate average pixel intensity and subtract local background for each feature. Images were manually examined to note any arrays affected by high background, scratches, or other technical artifacts that exceed quality control (QC) thresholds as part of the QC process.

Feature intensities for Cy3 and Cy5 channels were imported into the Partek Genomics Suite (Partek Inc). The average intensity for human intergenic control probes was calculated for cohybridized test and human reference DNA samples. From this, a scale factor was determined to normalize the Cy5 human reference DNA signal average to the Cy3 signal average. The Cy5 intensities for all PathoChip probes were then multiplied by the scale factor to normalize for differences in dye performance. Cy3/Cy5 ratios and Cy3-Cy5 subtractions were calculated for each probe to provide input for dual-channel or single-channel analysis pipelines, respectively. Accession average (AccAvg) was defined as the average Cy3 or Cy5 intensity across all probes for one accession, and accession signal (AccSig) was defined as AccAvg(Cy3)-AccAvg(Cy5).

Model-based analysis of tiling arrays (MAT), as implemented in Partek, was used for sliding window analysis of probe signals (Cy3 minus Cy5) for each tumor sample. MAT parameters were a p-value cutoff of 0.99, a window of 5,000 bp, a minimum number of positive probes of 5, and a discard value of 0%. Candidate regions were classified by MAT scores of 30 to 300, 300 to 3,000, and >3,000. Partek analysis of variance (ANOVA) tools were used to perform paired t-tests with multiple testing correction using all tumor samples as replicates of the test condition and cohybridized human reference DNA replicates as the control condition. Comparisons were performed at the accession level using AccAvg(Cy3) versus AccAvg(Cy5) and at the individual probe level using Cy3 versus Cy5 intensity values. Significance thresholds were set at a step-up false discovery rate of <0.05 and fold difference of >2. An outlier analysis was also performed at accession and probe levels by calculating the standard deviation of AccSig or probe signal across all tumors and filtering for any values that were 2 or more standard deviations higher than the population mean[6].

In some data analysis, R-program was also used for normalization and data analyses[7]. The scale factor was calculated using the signals of green and red channels for human probes and scale factors are the sum of green/sum of red signal ratios of human probes. These scale factors were then used to obtain normalized signals for all other probes. For all probes except human probes, normalized signal is log2 transformed of green signals / scale-factor-modified red signals (log2 g – log2 scale factor * r). For normalized signals, t-tests were applied to select probes showing positive signal in cancer samples by comparing cancer samples versus controls (un-matched and matched controls) and to select probes significantly present in un-matched or matched controls versus cancer samples. The significance cutoff was log2 fold change > 0.5 and the adjusted p-value < 0.05. These adjusted p-values were obtained for multiple corrections using the Benjamini–Hochberg procedure, any probe was detected as significant in control sample under this adjusted p-value cutoff. The top ones in control with nominal p-value < 0.05 without any multiple comparison corrections were presented to have a comparison with the significant probes present in cancer samples. Prevalence was calculated based on the percent detection of the signatures in the cancer versus control samples.

Results and Discussion

Microarray Design Optimization

The assembled metagenome contains 5,206 accessions for over 4,200 viruses, bacteria, fungi, and parasites in 58 artificial chromosomes totaling 448.9 million bp. Roughly 5.5 million probes from this metagenome were identified by Agilent custom probe design algorithms built for CGH applications. Over 3 million of these probes were predicted to have low risk of cross-hybridization with a human genome sequence.

Initially, a subset of these probes that map to unique target regions of the selected pathogens was synthesized on PathoChip v2a microarrays, and a separate subset that covers conserved regions between at least two or more viruses was synthesized on PathoChip v2b arrays[6]. An enhanced feature of the PathoChip v2b was the inclusion of 2,085 probes tiled throughout the lengths of 22 accessions for pathogens known to be highly associated with human cancers. Pilot assays using Agilent reference human DNA showed median probe intensities of over 750 fluorescence units (RFU) for probes to human sequences, around 17 RFU for nonhuman specific probes on PathoChip v2a, and 120 RFU for nonhuman conserved probes on PathoChip v2b[1]. Overall, these assays identified 6,360 probes with fluorescence values >150 that would likely be able to hybridize to human DNA. As these would lead to high background, they were removed from consideration for generation of the PathoChip v3 design, which combined both unique and conserved probe sets.

Interestingly, high hybridization intensities were noted for probes to Epstein-Barr virus (EBV; human herpesvirus 4) from the Agilent female reference DNA. The manufacturer confirmed that cell lines used to prepare the male and female SureTag human reference DNA were infected with this virus to generate the cell lines. This can generate a false positive result for EBV if the signals are not normalized to EBV probe signals in the female reference Cy5 channel. We then switched our reference DNA to a virus-free reference human DNA from a B-cell origin. Following a number of stringent detection steps, this reference has exhibited no signal for human viruses.

Development of a Modified Labelling Workflow Unique to PathoChip Assay

In contrast to the standard CGH workflow that only processes genomic DNA (gDNA), the PathoChip assay uses both gDNA and total RNA. Many commercially available DNA and RNA purification kits can process FFPE samples; however, we have found that QIAGEN AllPrep DNA/RNA FFPE kit provides efficient extraction of gDNA and total RNA from the same FFPE specimen. This kit successfully extracted nucleic acids from fungal cells as well as both Gram-negative and Gram-positive bacteria, which are likely to be the most difficult microbial agents in the samples. This kit also efficiently recovered both gDNA and RNA from *Saccharomyces cerevisiae, Bacillus cereus, and Escherichia coli cultures* (data not shown).

Another consideration is that, due to the potential low range of copy numbers among different microbial agents, DNA and RNA must be amplified to increase copy numbers above the detection threshold. The PathoChip screening technology utilizes an amplification step via the TransPlex WTA-2 amplification kit, allowing the detection of microorganisms and viruses present at low genomic copy numbers. These features of the PathoChip array allow multiple tumor samples to be rapidly and sensitively screened for the presence of a wide variety of microbial agents.

In initial testing, human adenovirus type 5, JC polyomavirus, or BK polyomavirus DNA was added to a background of 15 ng of human DNA at absolute copy numbers ranging from 10,000 to 10 viral genomes. After TransPlex amplification, we detected adenovirus type 5 with PathoChip probes at all copy numbers while the polyomavirus probes detected signal above background as low as 100 genome copies (Figure 2A). DNA from cell lines containing adenovirus type 5 and RNA containing respiratory syncytial virus was mixed and simultaneously amplified by TransPlex WTA-2. Probes for both viruses produced strong and specific detection signals. This indicated that the TransPlex WTA-2 provided robust reverse transcription in the presence of genomic DNA and that gDNA and cDNA targets were coamplified[6].

Altogether, we have demonstrated that nucleic acids from eukaryotic and prokaryotic pathogens can be extracted and detected using the PathoChip tumor extraction procedure (based on a modified CGH workflow [Figure 2B]).

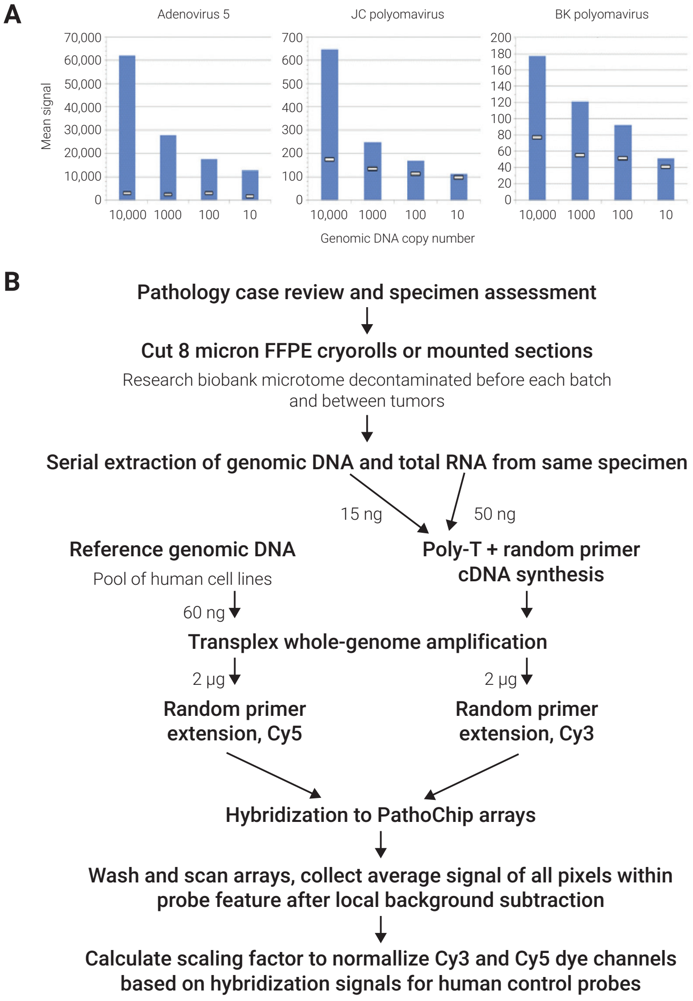


Figure 2. PathoChip workflow to detect control DNA pathogens. Detection responses for three viruses were measured over a dilution series from 10,000 to 10 genomic copies per sample. Genomic DNA for each virus was spiked into a reference amount of human DNA. Blue bars are the average Cy3 signals for all probes to the indicated viruses hybridized to test samples, and white lines indicate the probes' Cy5 average from control samples (human DNA only).

Development of A Data Analysis Strategy for PathoChip Array with OSCC Samples

Oncogenic viruses may undergo significant genomic rearrangements or deletions in host tumors. Furthermore, as viral strains can be widely polymorphic, detection of a new pathogen may rely on signal from a single probe. Several levels of data analysis are therefore needed to detect three main classes of "hits" that might be expected in a screening project (Figure 3). Accession signal (AccSig) is the average of all probes for an accession adjusted for human DNA cross-hybridization. AccSig was calculated to screen for general pathogen detection based on the majority of probes in an accession's set.

MAT (model-based analysis of tiling arrays) scores from a sliding window of probes were calculated to detect local areas of high signal regardless of accession boundaries. T-tests with multiple testing corrections were employed at the individual probe level to identify probes with signal consistently higher than background across the population of tumors. An outlier analysis was conducted for probes with high signal but only in one (or several) tumors from the screening population.

Data from a screening project of 100 OSCC tumors were used to evaluate these analysis methods. AccSig for HPV16 was consistent with p16 pathology reports[6] (Figure 4), with 80% of p16(+) tumors producing an AccSig value of more than 100. Of the eight p16(+) tumors with low or no HPV16 AccSig, four showed high signals for a subset of HPV16 probes or produced significant AccSig values for HPV26 or HPV92. The sliding window analysis recapitulated AccSig results and highlighted the differences between detection events for full or partial HPV16 genomes[6].

Analyses at the individual probe level also demonstrated utility for identifying candidates. Most HPV16 probes passed a t-test significance threshold for detection greater than background across the tumor population (Table 1). This is expected for a genome that is so common in OSCC. Many HPV16 probes also passed the outlier test indicating that, although the signals are consistently different from background, the population's range of intensities is wide (and therefore also contains outliers). In contrast, fewer HPV18 and HPV26 probes were significant by t-test, reflecting the much lower apparent occurrence of these genomes in this tumor population.

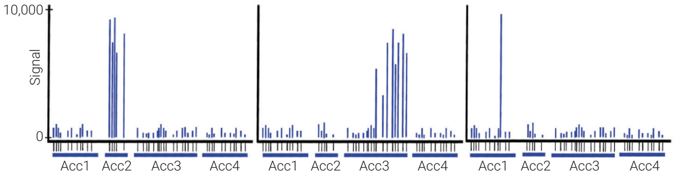


Figure 3. Model data illustrating three analysis strategies with PathoChip. Signals from individual probes (x axis) to four genome accessions (Acc) are plotted after hybridization to three hypothetical tumor samples. All probes for Acc2 show high signal in tumor 1 (left), so this candidate should be detectable by comparing the accession's all-probe averages from test samples with those of control samples. A subset of Acc3 probes show high signal in tumor 2 (middle), perhaps due to strain sequence differences or partial deletion of the genome, reducing the all-probe accession average and making detection more difficult. In this case, a sliding window analysis of local probe signals is not biased by accession annotation and may be more sensitive for candidate identification. A single probe for Acc1 has high signal in tumor 3 (right), so a third tier of analysis based solely on individual probe performance is needed to detect organisms not specifically targeted by the PathoChip but sharing sequence homology with one or a few probes.

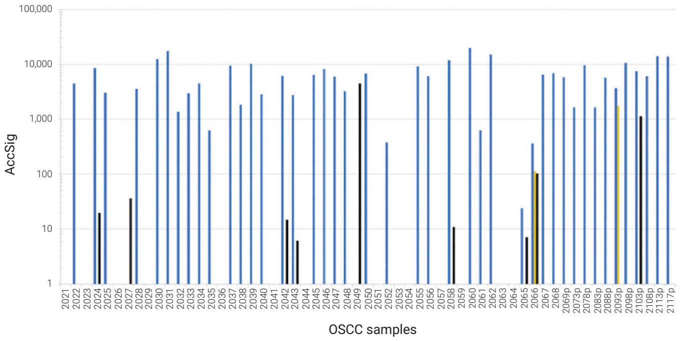


Figure 4. Accession average analyses for HPV in tumors. (A) The accession signals (AccSig) for HPV16 (blue), HPV18 (orange), and HPV26 (black) were calculated from PathoChip results for 100 oral squamous cell carcinoma (OSCC) samples, assayed individually (2021 to 2068) or in pools (2069p to 2117p).

Table 1. Individual probe analyses for human papillomavirus detection.

Probe	No. of probes		
	HPV16	HPV18	HPV26
Total probes	68	85	13
Specific probes	67	84	11
Pass *t* test	64	11	4
Pass outlier test	65	66	9
Conserved probes	1	1	2
Pass *t* test	1	1	0
Pass outlier test	1	0	2

However, the outlier analysis easily identified the relatively larger number of probes that produced HPV18 or HPV26 detections by AccSig or MAT score in a few positive samples. For these rarer candidates, some probes were significant by t-test because they produced lower but consistent signals over background throughout the population. This may be due to the copy number of genomes present and is not surprising. This also illustrates the need to examine probe-level hybridization intensities—not just to analyze algorithm output scores—when considering candidates for follow-up validation, regardless of the method used for their initial identification.

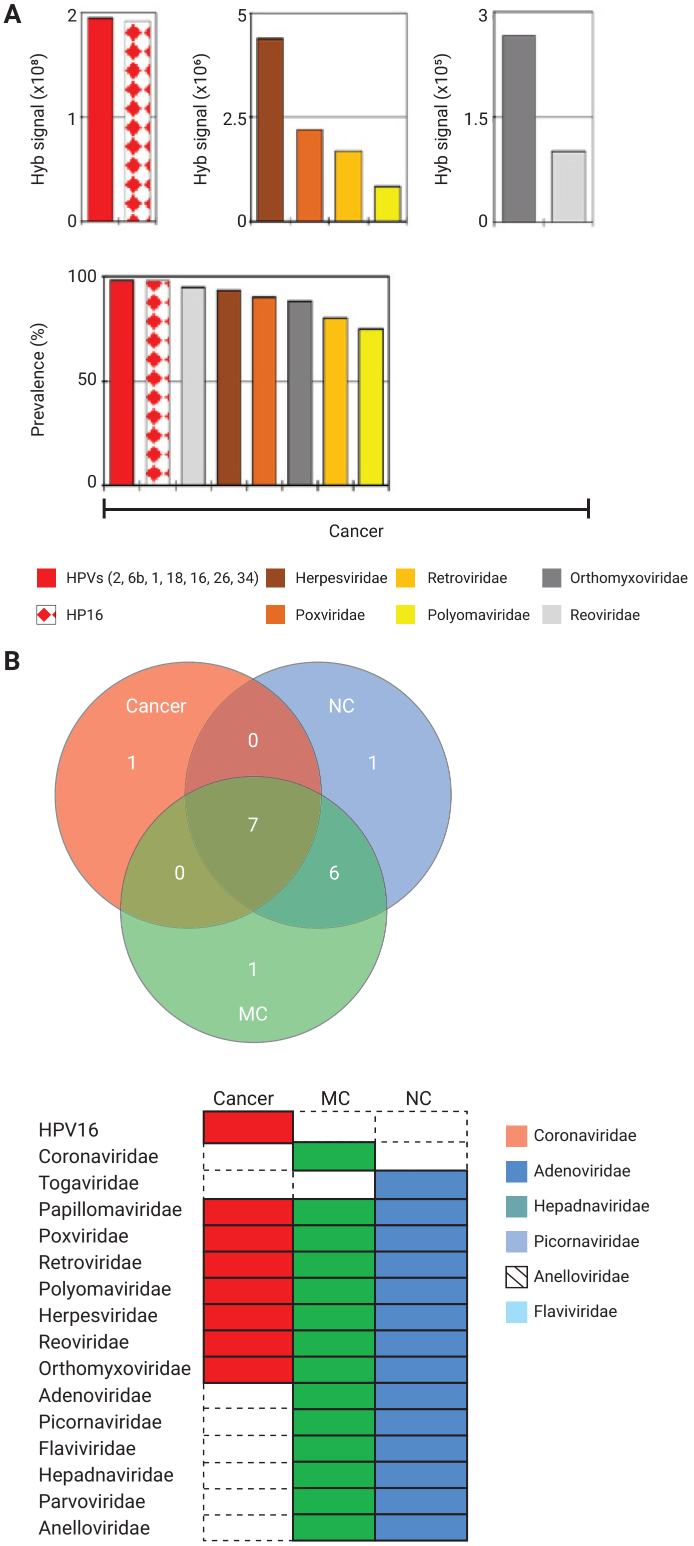
Using Pathochip Array to Detect Distinct Microbial Signatures Associated with Oropharyngeal and Oral Squamous Cell Carcinomas

In earlier data analysis[6], metagenome regions with a MAT score of more than 3,000 were compiled for each sample. The individual probes within each region were then ordered by map position in a plot of probe signals. This analysis detected a number of other organisms, including pathogenic oral bacteria, demonstrating potential involvement of a wide range of microbes in OSCC beyond HPV.

Using the PathoChip technology, we screened 100 FFPE pathologically defined Oral Cavity Squamous Cell Carcinoma (OCSCC) patient samples as well as 20 cancer adjacent normal controls (matched) and 20 oral tissue (uvula) from healthy individuals (non-matched controls). This screen assayed for distinct viral and microbial signatures associated with the tumor tissue[7]. Samples analyzed in this study were carcinomas taken from tongue, base of tongue, tonsil, floor of mouth, and cheek, but were predominantly oropharynx. Following our modified workflow, both DNA and RNA were extracted from the samples, subjected to whole genome and transcriptome amplification (referred to here as WGTA), labelled, and hybridized to the probes on the PathoChip.

Viral signatures associated with OCSCC

We identified RNA and DNA viruses associated with the cancer and control samples (Figure 5). Viral sequences belonging to *Papillomaviridae* showed the highest hybridization signal in the OCSCC samples screened, followed by that of *Herpesviridae*, *Poxviridae*, *Retroviridae*, and *Polyomaviridae* (Figure 5A). Viral signatures belonging to all of these families were seen to be >75% prevalent among the 100 OCSCC samples screened. Interestingly, *Papillomaviridae* was detected in 98% of the cases (Figure 5A). The hybridization signal for all papillomaviruses was much higher in the OCSCC samples compared to the matched and non-matched controls[9].



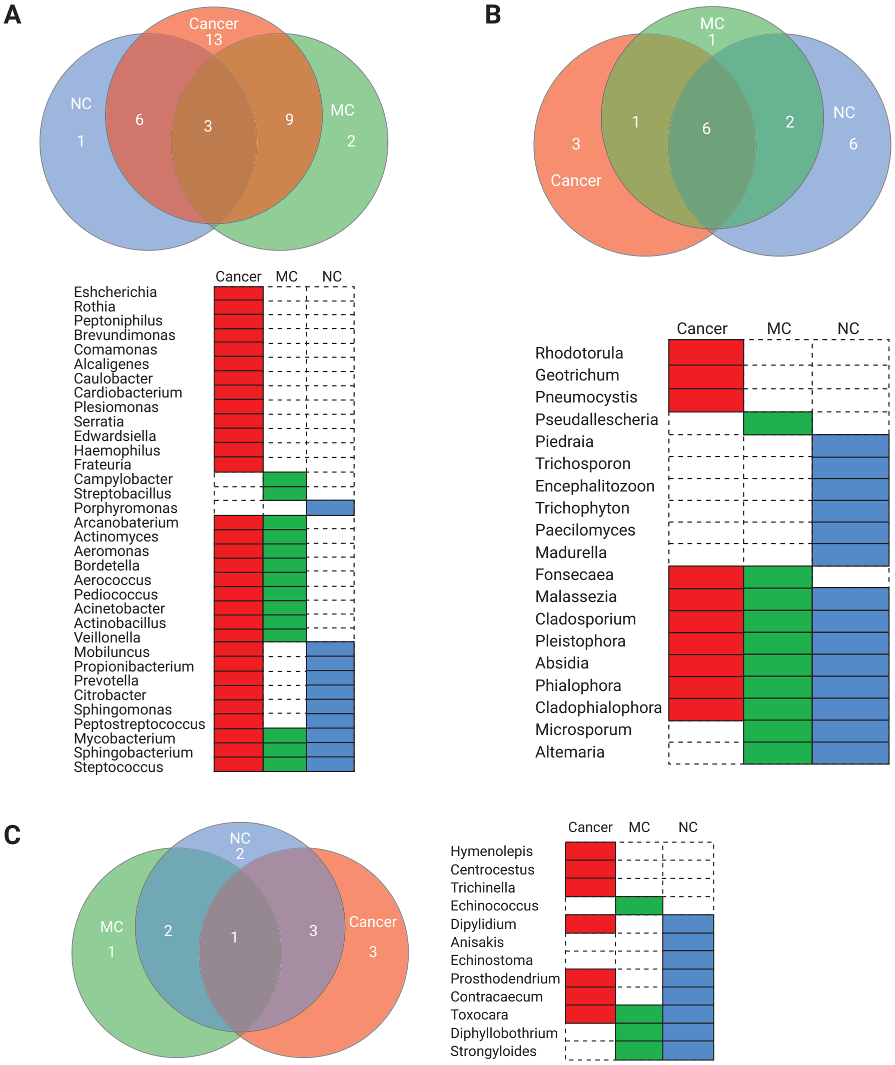
	Cancer	MC	NC
HPV16	🟥		
Coronaviridae		🟩	
Togaviridae			🟦
Papillomaviridae	🟥	🟩	🟦
Poxviridae	🟥	🟩	🟦
Retroviridae	🟥	🟩	🟦
Polyomaviridae	🟥	🟩	🟦
Herpesviridae	🟥	🟩	🟦
Reoviridae	🟥	🟩	🟦
Orthomyxoviridae	🟥	🟩	🟦
Adenoviridae		🟩	🟦
Picornaviridae		🟩	🟦
Flaviviridae		🟩	🟦
Hepadnaviridae		🟩	🟦
Parvoviridae		🟩	🟦
Anelloviridae		🟩	🟦

Figure 5. Viral signatures detected in oral cancer and control samples. (A) The viral signatures that are detected with hybridization signal (g–r > 30) by PathoChip screen of 100 oral cancer samples are shown and ranked according to decreasing hybridization signal (weighted score sum of all the probes per accession) and prevalence. (B) The association of different molecular signatures of viral families with cancer and controls, represented as a Venn diagram, and as colored bars.

Importantly, HPV16 was detected with both high hybridization signal and prevalence (98%) only in the OCSCC samples (Figure 5A, 5B). Signatures of *Reoviridae*, *Herpesviridae*, *Poxviridae*, *Orthomyxoviridae*, *Retroviridae* and *Polyomaviridae* were detected in OCSCC samples with high prevalence and at hybridization signals that were 2–3 logs higher than in controls (Figure 5A). Notably, viral signatures of *Coronoviridae*, *Picornaviridae*, *Adenoviridae*, *Anelloviridae*, *Hepadnaviridae* and *Flaviviridae* were significantly and specifically detected in the controls along with signatures of non-HPV16 papillomaviridae[7]. These data show that viral signature is significantly changed when compared specifically to the OCSCC tissue.

Bacterial signatures associated with OCSCC

Figure 6A shows the variety of bacterial signatures found in OCSCC, matched control, and non-matched control samples. The Venn diagram summarizes our findings, showing that bacterial signatures representing 13 genera are found to be associated with OCSCC samples but not with the matched or non-matched controls. These included 11 genera of *Proteobacteria*, 1 genus each of *Actinobacteria* and *Firmicutes*[7]. *Proteobacteria Brevundimonas* and *Actinobacteria Mobiluncus* were the most prevalent (98%): probes of Proteobacteria generas *Escherichia* and *Brevundimonas* were detected in 88 and 98% of cancer cases, respectively, with high hybridization signals. *Actinobacteria* probes detected in the OCSCC samples also had high hybridization signals with the highest being that of Arcanobacterium. As in the case of the viruses, the bacterial microbial signatures showed a significant divergence in the OCSCC when compared to the normal signatures and were more robust. Among matched or non-matched control samples, bacterial signatures of genera *Actinomyces*, *Mobiluncus*, and *Mycobacterium* were detected. Importantly, it should be noted that most of the bacterial signatures detected in the control samples are of the normal oral flora.



	Cancer	MC	NC
Eshcherichia	🟥		
Rothia	🟥		
Peptoniphilus	🟥		
Brevundimonas	🟥		
Comamonas	🟥		
Alcaligenes	🟥		
Caulobacter	🟥		
Cardiobacterium	🟥		
Plesiomonas	🟥		
Serratia	🟥		
Edwardsiella	🟥		
Haemophilus	🟥		
Frateuria	🟥		
Campylobacter		🟩	
Streptobacillus		🟩	
Porphyromonas			🟦
Arcanobaterium	🟥	🟩	
Actinomyces	🟥	🟩	
Aeromonas	🟥	🟩	
Bordetella	🟥	🟩	
Aerococcus	🟥	🟩	
Pediococcus	🟥	🟩	
Acinetobacter	🟥	🟩	
Actinobacillus	🟥	🟩	
Veillonella	🟥	🟩	
Mobiluncus	🟥		🟦
Propionibacterium	🟥		🟦
Prevotella	🟥		🟦
Citrobacter	🟥		🟦
Sphingomonas	🟥		🟦
Peptostreptococcus	🟥		🟦
Mycobacterium	🟥	🟩	🟦
Sphingobacterium	🟥	🟩	🟦
Steptococcus	🟥	🟩	🟦

	Cancer	MC	NC
Rhodotorula	🟥		
Geotrichum	🟥		
Pneumocystis	🟥		
Pseudallescheria		🟩	
Piedraia			🟦
Trichosporon			🟦
Encephalitozoon			🟦
Trichophyton			🟦
Paecilomyces			🟦
Madurella			🟦
Fonsecaea	🟥	🟩	
Malassezia	🟥	🟩	🟦
Cladosporium	🟥	🟩	🟦
Pleistophora	🟥	🟩	🟦
Absidia	🟥	🟩	🟦
Phialophora	🟥	🟩	🟦
Cladophialophora	🟥	🟩	🟦
Microsporum		🟩	🟦
Altemaria		🟩	🟦

	Cancer	MC	NC
Hymenolepis	🟥		
Centrocestus	🟥		
Trichinella	🟥		
Echinococcus		🟩	
Dipylidium	🟥		🟦
Anisakis			🟦
Echinostoma			🟦
Prosthodendrium	🟥		🟦
Contracaecum	🟥		🟦
Toxocara	🟥	🟩	🟦
Diphyllobothrium		🟩	🟦
Strongyloides		🟩	🟦

Figure 6. Bacterial, fungal and parasitic signatures detected in oral cancer samples. Figure A, B and C shows the association of molecular signatures of different bacterial, fungal and parasitic genera with oral cancer and/or controls, represented as a Venn diagram, and as colored bars.

Fungal and parasitic signatures associated with OCSCC

The Venn diagram shows the shared and specific fungal signatures between OCSCC, matched and non-matched controls. Noteworthy are the three fungal signatures, *Rhodotorula*, *Geotrichum*, and P*neumocystis*, associated specifically with OCSCCs (Figure 6B). Molecular signatures of *Fonsecaea*, *Malassezia*, *Pleistophora*, *Rhodotorula*, *Cladophialophora*, and *Cladosporium* were detected in all the OCSCC 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[7]. We note that a significant change in the fungal biome of OCSCC was observed when compared to control oral samples. Meanwhile, the Venn diagram in Figure 6C 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 OCSCC. Signatures of *Echinococcus* were 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.

Using Pathochip to Detect Distinct Microbial Signatures Associated with Triple-Negative Breast Cancer (TNBC)

Breast cancer is one of the most prevalent cancers and is categorized based on presence or absence of certain hormone and growth receptors. The most aggressive form of breast cancer is triple-negative breast cancer (absence of estrogen, progesterone, and HER2 receptors) as it cannot be treated by endocrine therapy. Genetic, environmental, and lifestyle factors have been implicated in the progression of breast cancers, but several studies with breast cancer have shown an association with herpesviruses, polyomaviruses, papillomaviruses, and retroviruses[9].

A total of 100 TNBC samples, along with 17 matched and 20 non-matched controls, were screened using the PathoChip. All samples were derived from FFPE archival samples (see Methods). Of the 100 TNBC samples screened, 40 were screened individually, and 60 were screened in pools of five samples (10 ng each of RNA/DNA) per reaction. In total, 52 arrays were used to screen the 100 TNBC samples. Samples were pooled for the 17 matched and 20 non-matched controls so that four arrays were used for each set

A probe was considered positive when the PathoChip screen detected a higher Cy3 (g) signal than Cy5 (r) signal for a particular probe. Probes of a particular organism were considered associated with cancer samples when the detectable hybridization signal ($g - r > 30$) was found significantly higher in cancer samples compared to matched or non-matched control samples. Also, multiple detection methods were used to carefully evaluate positive probes. These methods (Table 2) include Accession outlier, Accession t-test, Specific probe outliers, Specific probe t-test, Conserved probe outlier, Conserved probe t-test, and Model based analysis for tiling arrays.

For reporting purposes, we list the names of specific viruses and microorganisms that were detected by specific probes on the PathoChip. However, we note that the detection by specific probes may suggest a closely related family member and not the specific organism. This is particularly relevant in cases where TNBC samples showed a range of hybridization signals, or no hybridization signals for some probes across the probe set for a specific virus or microorganism. It could also mean that genomic regions of these agents are deleted in that particular tumor or that a strain has exhibited variance.

Among the conserved probes, viral signatures belonging to *Herpesviridae*, *Retroviridae*, *Parapoxviridae*, *Polyomaviridae*, and *Papillomaviridae* families were detected[8]. For example, for the herpesviridae family, probes of Human Cytomegalovirus (HCMV), Human Herpesvirus 1 (HHV1; Herpes simplex type 1), Kaposi sarcoma herpes virus (KSHV), Epstein-Barr virus, or Human Herpesvirus 4 (EBV/HHV4) were significantly detected among 92%, 65%, 96 and 78% of the breast cancer samples, respectively. For the papillomavirus family, specific probes detected Human Papillomavirus (HPV) 6b, HPV18, HPV2 and HPV16 in 78.8%, 75%, 84.6%, and 78.8% of the breast cancer samples, respectively. Specific probes also detected signals for Hepatitis GB, C and B in 82.7%, 90.4%, and 86.5% of the cancer samples, respectively. Interestingly, not all the specific probes of these viral agents were detected[8]. This could be due to several possibilities, including similar organism with identical sequence for the probe region, fragments of an organism being present, or integrated fragments of organismal DNA.

The viral probes, when ranked according to percent prevalence (regardless of hybridization intensity), showed signatures of Hepadnaviruses and Flaviviruses (86.5%), followed by Parapoxviruses (83.3%), Herpesviruses (83.2%), Retroviruses (79.6%), and Papillomaviruses (79.3%). However, when ranked according to decreasing hybridization signal, Herpesvirus probes had the highest hybridization signal across the tumors, followed by high hybridization signal for the probes of *Parapoxviruses*, *Flaviviruses*, *Polyomaviruses*, *Retroviruses*, *Hepadnaviruses* and *Papillomaviruses* (Table 2). Similar data were also generated for bacterial, fungal, and parasitic agents [8].

Table 2. Hybridization signal (calculated as sum of hybridization signal of all the probes per accession) and prevalence of viral probes detected in 100 triple negative breast cancer samples. The methods that detected the candidates are mentioned; AO: Accession outlier, AT: Accession t-test, SO: Specific probe outliers, ST: Specific probe t-test CO: Conserved probe outlier, CT: Conserved probe t-test, MAT: Model based analysis for tiling arrays.

(a) Associated viral agent	Detection methods	Percent detected (%)	Probe sum/ accession
Human herpesvirus 5/HCMV	AO, SO, ST, CO, MAT	92	14332000
Human herpesvirus 8/KSHV	AO, SO, MAT	96	12119800
Simian virus 40	AO, SO, MAT	75	8113970
Hepatitis C virus genotype 1	SO, CO, MAT	90	7199330
Human T-lymphotropic virus 2	AO, SO, CO, MAT	88	7040500
Orf virus	CO, MAT	75	6422460
Pseudocowpox virus	AO, SO, CO, MAT	90	5037880
Human herpesvirus 4/EBV	AO, SO, CO, MAT	79	5024970
Bovine papular stomatitis virus	AO, SO, CO, MAT	85	4214040
Okra mosaic virus	AO, SO, CO, MAT	75	3435060
Human papillomavirus 2	SO, MAT	85	3361460
Human T-lymphotropic virus 1	AO, SO, CO, MAT	83	2745990
Hepatitis B virus	AO, SO, CO, MAT	87	2621640
Human herpesvirus 1	AO, SO, CO, MAT	65	2319570
Human papillomavirus type 16	SO, MAT	79	1651350
Moloney murine leukemia virus	SO, CO, MAT	58	1587600
Merkel cell polyomavirus	AO, SO, MAT	90	1551830
Mouse mammary tumor virus	AO, SO, MAT	79	1464980
Human paillomavirus type 6b	AO, SO, MAT	79	1271950
Human paillomavirus 18	SO, CT, MAT	75	1184610
JC polyomavirus	AO, CO, SO, MAT	77	755288
Hepatitis GB virus A	SO, MAT	83	749098
Fujinami sarcoma virus	SO, CO, MAT	90	691071

A summary of detected viral conserved and specific probes—as well as bacterial, fungal and parasitic probes in cancer samples—is shown in Table 3 & 4[8]. The viral, bacterial, fungal, and parasitic signatures detected in the triple-negative breast cancer samples were found to be significantly associated with the cancer samples ($p < 0.05$) compared to the non-matched and matched control samples analyzed[8].

Table 3. Number of viral probe signatures detected by screening triple negative breast cancer samples by the PathoChip.

	Retroviridae				Polyomaviridae				Herpesviridae			
	MMTV	MMLV	HTLV1	HTLV2	FSV	SV40	JC	MCPV	HCMV	EBV	KSHV	HHV1
Total probes	31	24	41	86	8	41	42	62	299	235	259	22
Specific	31	15	37	84	5	41	40	62	275	149	256	15
Outlier	1	4	24	43	4	25	12	27	139	67	132	7
t-test	30	0	0	0	0	0	0	0	1	0	0	0
Conserved	0	9	4	2	3	0	2	0	24	86	3	7
Outlier	0	2	3	2	2	0	1	0	15	2	3	3
t-test	0	0	0	0	0	0	0	0	0	0	0	0

	Papillomaviridae				Hepadnaviridae	Flaviviridae		Poxviridae		
	HPV16	HPV18	HPV6b	HPV2	HBV	HCV-1	HepGB virus A	BPSV	PCP	ORF
Total probes	68	85	91	92	49	121	14	109	105	111
Specific	67	84	90	92	47	119	14	12	12	13
Outlier	19	28	37	49	25	72	7	1	3	1
t-test	0	0	0	0	0	0	0	0	1	0
Conserved	1	1	1	0	2	2	0	97	93	98
Outlier	0	0	0	0	2	0	0	74	80	76
t-test	0	1	0	0	0	0	0	0	0	0

Table 4. Number of bacterial, fungal and parasitic probe signatures detected by screening triple negative breast cancer samples by the PathoChip.

Microbial signatures	Total no. of probes in the Chip	**Total no. of probes detected**	Detection in triple negative breast tumors	Type of agent
Members	**Specific**	**Specific**	**Percent positive**	**Organism**
Arcanobacterium	4	4	75	Bacteria
Brevundimonas	3	3	73	Bacteria
Sphingobacteria	5	5	67	Bacteria
Providencia	1	1	67	Bacteria
Prevotella	2	2	67	Bacteria
Brucella	10	10	65	Bacteria
Escherichia	13	10	64	Bacteria
Actinomyces	4	4	52	Bacteria
Mobiluncus	4	4	50	Bacteria
Propiniobacteria	2	2	50	Bacteria
Geobacillus	2	1	44	Bacteria
Rothia	3	3	40	Bacteria
Peptinophilus	2	2	39	Bacteria
Capnocytophaga	1	1	37	Bacteria

Microbial signatures	Total no. of probes in the Chip	**Total no. of probes detected**	Detection in triple negative breast tumors	Type of agent
Members	**Specific**	**Specific**	**Percent positive**	**Organism**
Pleistophora	8	8	98	Fungi
Piedra	6	6	90	Fungi
Foncecaea	3	3	89	Fungi
Phialophora	4	4	87	Fungi
Paecilomyces	4	4	69	Fungi

Trichuris	7	7	96	Parasite
Toxocara	1	1	62	Parasite
Leishmania	6	5	60	Parasite
Babesia	2	2	56	Parasite
Thelazia	1	1	40	Parasite
Paragonimus	3	2	15	Parasite

Conclusions

The ability of a highly multiplexed metagenomic assay to detect small nonhuman genomes in an overwhelming background of human sequences depends on several factors. These include nucleic acid extraction and recovery, target size and copy number, amplification efficiency (if WTA if used), and specific probe performance. Several modifications have been made during PathoChip development to enhance the sensitivity of the assay. Modifications comprise the inclusion of multiple probes per accession and integration of candidates from different levels of data analysis, allowing optimization of pathogen detection in screening projects. The ability of the PathoChip to combine saturation probe sets, RNA and DNA detection enhances screening for known oncogenic pathogens or other microbial agents. The PathoChip assay will thus allow for a comprehensive assessment of the frequency of coinfection by multiple organisms and their correlation with driving oncogenic or other pathogenic events.

A PathoChip screening project can generate a list of candidates prioritized by the magnitude of detection, detection via multiple analysis strategies including hierarchical clustering, and the rate of detection across the sample population. Combining these results with annotations for the virus or pathogenic microorganism such as host range, tissue specificity, or prevalence in the general population will assist in determining which agents deserve further attention. These results can be followed up via either PCR or capture probe-based hybridization with next-generation sequencing. This approach is likely to provide a promising microbial signature of a particular cancer or disease with agents with various degrees of contribution.

References

1. de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, Forman D, Plummer M. “Global burden of cancers attributable to infections in 2008: a review and synthetic analysis.” *Lancet Oncol.* 13:607–615 **(2012)**.
2. Brodie EL, Desantis TZ, Joyner DC, Baek SM, Larsen JT, Andersen GL, Hazen TC, Richardson PM, Herman DJ, Tokunaga TK, Wan JM, Firestone MK. “Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation.” *Appl. Environ. Microbiol.* 72:6288–6298, **(2006)**.
3. Wong CW, Heng CL, Wan Yee L, Soh SW, Kartasasmita CB, Simoes EA, Hibberd ML, Sung WK, Miller LD. “Optimization and clinical validation of a pathogen detection microarray.” *Genome Biol.* 8:R93, **(2007)**.
4. Chen EC, Miller SA, DeRisi JL, Chiu CY. “Using a pan-viral microarray assay (Virochip) to screen clinical samples for viral pathogens.” *J. Vis. Exp.* 50:2536 **(2011)**.
5. Tu Q, Yu H, He Z, Deng Y, Wu L, Van Nostard JD, Zhou A, Voodeckers J, Qin Y, Hemme CL, Shi Z, Xue K, Yaun T, Wang A, Zhou J. “GeoChip 4: a functional gene array-based high throughput environmental technology for microbial community analysis.” *Mol. Ecol. Resour.* 14:914–928, **(2014)**.
6. Baldwin, D. A., Feldman, M., Alwine, J. C. & Robertson, E. S. “Metagenomic assay for identification of microbial pathogens in tumor tissues.” *MBio 5.* e01714–01714, **(2014)**.
7. Banerjee S, Tian T, Wei Z, Peck KN, Shih N, Chalian AA, O'Malley BW, Weinstein GS, Feldman MD, Alwine J, Robertson ES. “Microbial Signatures Associated with Oropharyngeal and Oral Squamous Cell Carcinomas.” *Sci Rep.* 7(1):4036 **(2017)**.
8. Banerjee S, Wei Z, Tan F, Peck KN, Shih N, Feldman M, Rebbeck TR, Alwine JC, Robertson ES. “Distinct microbiological signatures associated with triple negative breast cancer.” *Sci Rep*. 5:15162, **(2015)**.
9. Shiovitz S, Korde LA. “Genetics of breast cancer: a topic in evolution.” *Ann Oncol.* 20, **(2015)**.

www.agilent.com

For Research Use Only. Not for use in diagnostic procedures.
PR7000-2665

This information is subject to change without notice.

Printed in the USA, December 2020
5994-1968EN

Agilent

Trusted Answers