Metagenomic Assay for Identification of Microbial Pathogens in Tumor Tissues

Don A. Baldwin, Michael Feldman, James C. Alwine, Erle S. Robertson

ABSTRACT Screening for thousands of viruses and other pathogenic microorganisms, including bacteria, fungi, and parasites, in human tumor tissues will provide a better understanding of the contributory role of the microbiome in the predisposition for, causes of, and therapeutic responses to the associated cancer. Metagenomic assays designed to perform these tasks will have to include rapid and economical processing of large numbers of samples, supported by straightforward data analysis pipeline and flexible sample preparation options for multiple input tissue types from individual patients, mammal's, or environmental samples. To meet these requirements, the PathoChip platform was developed by targeting viral, prokaryotic, and eukaryotic genomes with multiple DNA probes in a microarray format that can be combined with a variety of upstream sample preparation protocols and downstream data analysis. PathoChip screening of DNA plus RNA from formalin-fixed, paraffin-embedded tumor tissues demonstrated the utility of this platform, and the detection of oncogenic viruses was validated using independent PCR and sequencing methods. These studies demonstrate the use of the PathoChip technology combined with PCR and deep sequencing as a valuable strategy for detecting the presence of pathogens in human cancers and other diseases.

IMPORTANCE This work describes the design and testing of a PathoChip array containing probes with the ability to detect all known publicly available virus sequences as well as hundreds of pathogenic bacteria, fungi, parasites, and helminths. PathoChip provides wide coverage of microbial pathogens in an economical format. PathoChip screening of DNA plus RNA from formalin-fixed, paraffin-embedded tumor tissues demonstrated the utility of this platform, and the detection of oncogenic viruses was validated using independent PCR and sequencing methods. These studies demonstrate that the PathoChip technology is a valuable strategy for detecting the presence of pathogens in human cancers and other diseases.
have difficulty in discriminating between strains or reporting the presence of genomic variants or pathogenicity factors. Deep sequencing of the total DNA from a sample can certainly identify bacterial, viral, and other microbiome members (3, 17, 18) but with a severe penalty in efficiency. Even if the field attains the as-yet-unrealized goal of a cost of $1,000 per genome, total DNA sequencing will be an expensive method for screening hundreds or thousands of experimental and control samples to detect associations of pathogens with a particular disease. Depending on the specimen sampled, the data may overwhelmingly be from host human sequences, creating an unnecessarily large search space for locating pathogen signatures and resulting in the majority of sequence reads being discarded.

DNA microarrays have been used for metagenomics. The Lawrence Berkeley Lab/Affymetrix PhyloChip is based on rRNA sequences (19). An academically developed Virochip has probes for 1,500-plus viruses (20, 38-44) and has successfully detected viruses in pathology samples. The Virochip platform is limited to viruses and assays RNA that is reverse transcribed to cDNA for PCR amplification (20, 38-44). The Glomics GeoChip 4.0 focuses on RNA expression by bacteria in the human microbiome (21) and covers bacteriophages but no other viruses, nor any eukaryotic microorganisms. PathGen Dx has launched a PathChip kit that features an Affymetrix microarray for all known viruses and a broad selection of bacteria (22) but no eukaryotic pathogens.

These and other array-based tools illustrate the demand for methods to quickly and economically screen sets of samples for broad microbial content, including species beyond bacteria (23). This report describes development of the PathoChip platform containing 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 to independent regions of the target genome are used to improve an opportunity for detection. Furthermore, while the PathoChip content was developed from sequences to known targets, the ability to discover new strains or organisms is provided by the inclusion of probes to sequences that are conserved within and between viral families. To this end, a previously unknown virus with homology to a conserved sequence may produce a corresponding hybridization signal from such a probe, if not to a complete probe set. A supporting workflow is described for profiling large collections of tumor samples typically available as formalin-fixed, paraffin-embedded (FFPE) tissue in biobanks and includes simultaneous detection of DNA and RNA to expand the range of targets available for hybridization.

RESULTS

Microarray design. The PathoChip design goals were to cover all public NCBI viral genomes and genomic sequences from a broad selection of microorganisms (bacteria, fungi, and parasites) that are pathogenic to humans, using multiple probes to independent target sites in the genome of each species (Fig. 1A). The resulting collection of pathogen sequences was assembled into a metagenome containing 58 chromosomes of 448.9 million bp and 5,206 accessions for over 4,200 viruses, bacteria, fungi, and parasites. Agilent custom probe design algorithms built for comparative genomic hybridization applications were used to identify 5.5 million probes from the metagenome. Over 3 million of these probes were predicted to have low risk of cross-hybridization with a human genome sequence. Importantly, 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 regions of sequence conservation between at least two or more viruses was synthesized on PathoChip v2b arrays (Fig. 1B).

An enhanced feature of the PathoChip v2b was the inclusion of 2,085 probes tiled throughout the lengths of 22 accessions for agents known to be tightly associated with human cancers.

Pilot assays using Agilent reference human DNA showed median probe intensities of over 750 fluorescence units for probes to human sequences, around 17 fluorescence units for nonhuman specific probes on PathoChip v2a, and 120 fluorescence units for nonhuman conserved probes on PathoChip v2b (experiment 1, Table 1). These assays identified 6,560 probes with fluorescence values of >150 that would apparently be able to hybridize to human DNA and were therefore removed from consideration for generation of the PathoChip v3 design, which combined the unique and conserved probe sets (Fig. 1B).

Pilot experiment 1 indicated that the presence or absence of Cot-1 DNA made no difference in probe performance (Tables 1 and 2). Therefore, this reagent was omitted from subsequent assays. Interestingly, very high hybridization intensities were noted for probes to Epstein-Barr virus (EBV; human herpesvirus 4) from this control human 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. EBV detection in tumor screening assays that used this reagent was therefore possible if the signals were not normalized to EBV probe signals to the xhh Cy5 channel. Future assays for our screens utilized virus-free reference human DNA as the cross-hybridization control after validation through a number of stringent steps for detection of other known viruses.

Assay response to positive controls. The limited amounts of tissue available in most tumor archives or obtained from clinical procedures, such as fine needle aspirates and other biopsy specimens, require that metagenomic screening protocols include efficient strategies for nucleic acid extractions combined with an amplification step which allows for genome- or transcriptome-wide representation of microbial agents present in the sample. These methods must additionally be compatible with the degraded DNA and RNA typically produced by formalin tissue fixation. A draft workflow was designed to address these technical hurdles (Fig. 1C), and pilot experiments were conducted to test the amplification and detection of a number of positive-control viruses.

Three whole-genome amplification (WGA) methods that use phi29 polymerase rolling-circle amplification (24) (GenomiPhi), universal primer multiplex PCR (25) (GenomePlex and TransPlex), or single-primer Isothermal amplification (26) (Ovation WGA) were tested for their ability to detect a single bacteriophage genome spiked into a background of human DNA (15 ng). phiX174 DNA at copy numbers 1, 100, and 1000 relative to a single-copy human genomic locus was easily detected by PathoChip probes after any of the amplification reactions (Fig. 2A). To test detection of human DNA and RNA viruses, DNA from cell lines containing adenovirus type 5 or RNA containing respiratory syncytial virus was amplified by the GenomePlex DNA and TransPlex RNA methods (experiments 2 and 3, Table 2). The cell line DNA and RNA samples were then mixed and simultaneously amplified by TransPlex (experiment 4, Table 2). Probes for both viruses produced strong and specific
TABLE 1 Pilot PathoChip assays with human and nonhuman probes

| Expt | PathoChip Test (Cy3) | xhh cross-hybridization control (Cy5) | Amplification | Fluorescence value<sup>a</sup> |
|------|----------------------|---------------------------------------|---------------|---------------------------------|
|      |                      |                                       | Human probes  | Nonhuman probes                 |
|      |                      |                                       | Median Cy3    | Median Cy5                      |
|      |                      |                                       | Median Cy3    | Median Cy5                      |
| 1    | v2a                  | Human gDNA,* no Cot-1                 | Human gDNA, no Cot-1 | None | 794 | 785 | 18 | 17 |
| 1    | v2b                  | Human gDNA, no Cot-1                 | Human gDNA, no Cot-1 | None | 726 | 741 | 119 | 124 |
| 1    | v2a                  | Human gDNA + Cot-1                   | Human gDNA + Cot-1 | None | 758 | 794 | 17 | 17 |
| 1    | v2b                  | Human gDNA + Cot-1                   | Human gDNA + Cot-1 | None | 758 | 791 | 121 | 128 |
| 2    | v2a                  | Adenovirus type 5 + host gDNA        | Human gDNA    | GenomePlex WGA                  | 284 | 784 | 8 | 18 |
| 3    | v2a                  | Respiratory syncytial virus + host DNA | Human gDNA | TransPlex WTA                  | 448 | 825 | 10 | 16 |
| 4    | v2a                  | Adenovirus type 5 + respiratory syncytial virus + host gDNA | Human gDNA | TransPlex WTA                  | 371 | 832 | 7 | 15 |

<sup>a</sup> gDNA, genomic DNA.

<sup>b</sup> Relative fluorescence units.

Detection signals. This indicated that the TransPlex reverse transcription worked robustly in the presence of genomic DNA, and genomic DNA and cDNA targets were coamplified.

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, adenovirus type 5 was detected by PathoChip probes at all copy numbers while all polyomavirus probes produced detectable signal above background, detecting at least 100 genome copies (Fig. 2B).

The genomic sequence of human cytomegalovirus (CMV) strain AD169, a laboratory-adapted strain, differs at several locations from the NCBI reference CMV sequence, a clinical strain. This includes a large deletion. DNA from a cell line infected with CMV AD169 was amplified and hybridized to PathoChip v3, which includes a set of saturation tiling probes for the CMV reference genome. While most probes produced high signals, probes located at sites that are polymorphic or deleted in CMV AD169 had significantly reduced fluorescence signals, clearly delineating the polymorphisms or deletions in the laboratory strain (Fig. 2C).

**Assay performance with tumor tissue samples.** The AllPrep DNA/RNA FFPE kit provides efficient extraction of genomic DNA and total RNA from the same FFPE specimen, so this kit was tested for its ability, importantly, to extract nucleic acids from fungal cells and Gram-negative or Gram-positive bacteria, which are likely to be the most difficult microbial agents in the samples. DNA and RNA from Saccharomyces cerevisiae, Bacillus cereus, and Escherichia coli cultures were efficiently recovered using the kit (data not shown). This provided a preliminary indication that nucleic acids from eukaryotic and prokaryotic pathogens can be extracted and detected from the PathoChip tumor extraction procedure (Fig. 1C).

The screen was performed on an initial set of eight oropharyngeal squamous cell carcinoma (OSCC)/head and neck carcinoma samples from FFPE tissue specimens. Human p16 overexpression from the CDKN2A gene is correlated with oncogenic human papillomavirus (HPV) infection in OSCC, and p16 immunohistochemistry is used as a prognostic molecular biomarker in clinical pathology laboratories, with high sensitivity but poor specificity for HPV (27). The OSCC samples included five p16-positive tumors and three p16-negative tumors as determined by pathology at the Hospital of the University of Pennsylvania. From our PathoChip screen, four of the five p16(+) tumors produced high detection signals across the 68 PathoChip probes for HPV16, and the fifth tumor showed signals for a small subset of HPV16 probes. The remaining tumors were negative for PathoChip HPV16 detection (Fig. 3). Despite good hybridization to other p16(+) samples, three of the HPV probes for tumor 2025 and two probes for tumor 2028 had no detectable signal This is suggestive of an HPV strain variation, which was similar to the results from polymorphic sites in the CMV positive-control experiment.

**Development of PathoChip analysis strategies using OSCC tumor screening data.** Oncogenic viruses may undergo significant genomic rearrangements or deletions in host tumors. Furthermore, viral strains can be widely polymorphic, and 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 (Fig. 4). Accession signal (AccSig), the average of all probes for an accession adjusted for human DNA cross-hybridization, was calculated to screen for detection by a majority of probes in an accession’s set. MAT (model-based analysis of tiling arrays) scores (28) from a sliding window of probes were calculated to detect local areas of high signal regardless of accession boundaries. t tests with multiple testing correction were employed at the individual probe level to identify probes with signal consistently higher than back-

---

**FIG 1** PathoChip design and tumor screening workflow. (A) Sequence accessions for all viruses and selected human-pathogenic microorganisms were retrieved from the NCBI DNA sequence databases and concatenated to form a metagenome. Wherever possible, regions of target sequence unique to the accession (a) and (c) were used to select multiple 60-nt probes (1, 2, and 4 to 6) for microarray synthesis, and probes to target regions that share similar sequences in at least two viral accessions (b) were also identified. Probes to prokaryotic and eukaryotic pathogen genomes may map to intergenic, gene, or rRNA sequences or a mixture of target types, depending on the availability of sequence data. (B) Parallel and iterative design processes were used to assemble the PathoChip probe collection that covers unique and conserved target regions, supplemented with high-resolution probe tiling for known cancer-associated microorganisms. (C) The PathoChip tumor screening protocol simultaneously assays DNA and RNA from small amounts of tissue recovered from formalin-fixed, paraffin-embedded (FFPE) tumor specimens.
FIG 2  PathoChip assay performance assessed using positive-control DNA. (A) Whole-genome amplification kits that feature three different enzymatic processes were compared in their abilities to detect phiX174 bacteriophage genomic DNA spiked into human DNA. 1× DNA was equivalent to the molarity for a single-copy locus in the human genome. Green bars are the median Cy3 signal for the 14 phiX174 probes hybridized to test samples, and red bars show the median Cy5 signal from control samples (human DNA only). Error bars indicate standard deviations across probes. (B) 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). (C) Human cytomegalovirus (CMV) DNA was hybridized to a PathoChip containing 299 probes for saturation tiling across the reference CMV genome (NCBI accession NC_006273). The DNA was from CMV AD169, a strain that differs from the reference sequence at several locations, and was spiked into a background of human DNA for cohybridization with reference human DNA only (xhh). Red numerals indicate example probes for positive detection (1), low signal due to sequence polymorphisms (2, 3, and 4), and missing signal due to deletion in AD169 (5 and 6).
ground across the population of tumors, and an outlier analysis was conducted for probes with high signal but only in one or a few 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 (Fig. 5A; see also Table S1 in the supplemental material), 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. In Fig. 5B, metagenome regions with a MAT score of more than 3,000 were compiled for each sample, and the individual probes within each region were ordered by map position in a plot of probe signals. This analysis detected a number of other organisms, including pathogenic oral bacteria, although the signals were lower than those of the HPV

---

**FIG 3** HPV16 detection in naturally infected tumor samples. Eight oral squamous cell carcinoma samples were assayed on PathoChips that include 68 probes for human papillomavirus 16 (HPV16). Clinical pathology results for p16 overexpression, a diagnostic marker correlated with oncogenic HPV infection, are indicated in the color key for each tumor’s hybridization profile.
The pools were mixed with a panel of biotinylated DNA probes and were combined into six capture pools as indicated in Fig. 6A. Signals in probe sets identified above. Importantly, amplicons were not observed from HPV16-negative samples produced the appropriate amplicons from the high- and moderate-signal regions in tumors positive for HPV16 detection. PCR of genomic DNA from representative samples (pool 1, Fig. 6A and C). More HPV16 templates, up to 73 in pool 3, were observed in the libraries from HPV16-positive tumors. Interestingly, there were regions of high reads which showed high intensities in the majority of the tumors analyzed. The E4 region which was prevalent in the majority of the tumors (Fig. 6A to C) may provide a window into the transcription profiles of HPV16 in these particular type of OSCCs. The oncogenes E6 and E7 also showed a high number of signals in the tumors, but not as dramatic and surprising as the E4 open reading frame (ORF) region. However, this may be an interesting discovery which suggests a higher number of transcripts for the E4 ORFs in these tumors than was previously thought. Notably, the E1 region also showed greater signals in the tiled probes across the HPV16 genome. Predominantly, E4 seems to be the most prominent signal for the OSCC tumors and suggests a greater involvement of E4 in maintenance of the tumor by HPV than previously indicated.

**DISCUSSION**

The ability of a highly multiplexed, metagenomic assay to detect small nonhuman genomes in an overwhelming background of human sequences will be affected by several factors, including nucleic acid extraction and recovery, target size and copy number, participation in amplification reactions if used, and specific probe performance. The last three factors likely contributed to the differences in assay performance, in which 10,000 copies of JC or BK polyomavirus (5-kb genomes in a 4-kb double-stranded circular DNA plasmid vector) were detected with probe intensity ranges of 61 to 4,889 (JC virus, 42 probes) and 4 to 442 (BK virus, 9 probes). In contrast, adenovirus type 5 (36-kb genome, double-stranded linear nonintegrated DNA) was detected over an intensity range of 0 to 73 in pool 3, were observed in the libraries from HPV16-positive tumors. Interestingly, there were regions of high reads which showed high intensities in the majority of the tumors analyzed. The E4 region which was prevalent in the majority of the tumors (Fig. 6A to C) may provide a window into the transcription profiles of HPV16 in these particular type of OSCCs. The oncogenes E6 and E7 also showed a high number of signals in the tumors, but not as dramatic and surprising as the E4 open reading frame (ORF) region. However, this may be an interesting discovery which suggests a higher number of transcripts for the E4 ORFs in these tumors than was previously thought. Notably, the E1 region also showed greater signals in the tiled probes across the HPV16 genome. Predominantly, E4 seems to be the most prominent signal for the OSCC tumors and suggests a greater involvement of E4 in maintenance of the tumor by HPV than previously indicated.

**FIG 4** Model data illustrating three analysis strategies. 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.
342 to 65,325 using 63 probes and 100 target genome copies; differences in genome size and conformation may affect participation in whole-genome amplification reactions. Furthermore, probes clearly have different hybridization affinities despite sharing the same bioinformatic design criteria. PathoChip assay sensitivity will therefore vary across accessions and between protocol options, but the inclusion of multiple probes per accession and integration of candidates from different levels of data analysis provide avenues for optimizing the chances of detecting a pathogen in the screening projects. The HPV16 genome, for example, is a

**FIG 5** Accession average and sliding window 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). (B) Signals (tumor minus xhh) for each probe in metagenome regions with high sliding window scores are shown as a heat map of probes (y axis rows) hybridized to the tumor samples in panel A.
generally matched AccSig results but provided better ability to the strongest detection candidates. The sliding window analysis rather uncomplicated means to summarize the data and collect
iment.
lines of investigation as a follow-up to the initial screening exper-
potential agents having oncogenic activities and so lead to new
multiple host-genome integrated forms (35), but probe signal differ-
complicated by the ability of HPV16 to exist in episomal and mul-
primary transcript (34). Detailed interpretation of these data is
is known to be a highly abundant RNA splicing product from the
observed for E6/E7 sequences, and for the E4 region of E2, which
target preparation. Among the early genes, strong signals were
early gene signals is likely due to the RNA portion of the sample
probes to intergenic HPV16 sequences. Therefore, the boost in
HPV16 in 34 of 48 OSCC tumors individually tested by PathoChip
screening, a 71% occurrence rate somewhat higher than the esti-
ated 63% rate previously reported (29, 30) but not unreasonable
given the rapidly increasing prevalence of papillomaviruses in
oropharyngeal cancers (31). The results of the assay were highly concordant with the molecular pathology reports for p16 overex-
pression and in some cases suggested that an HPV strain other
than HPV16 may be responsible. HPV16 detections by PathoChip
assays were confirmed by PCR using primers that are independent of
PathoChip probes and by recovery and sequencing of HPV16
regions located outside those targeted by capture probes on the
HPV genome.
The ability of the PathoChip to combine saturation probe sets
and RNA and DNA detection enhances the screening for known
oncogenic pathogens. If sufficient target copies are present in a
sample, inferences regarding genomic structural variation and
RNA expression levels may be possible. For HPV16, probes map-
ing to early gene transcripts produced more overall signal than
those for late genes in OSCC samples; this is consistent with stud-
ies of actively infected cells and the oncogenic effects of E6/E7
expression (32, 33). Moreover, the transcription of HPV at the
level of detail that we are probing here is greater than previously
investigated and suggests a potential role for other transcripts,
including the E4 ORF and related antigen, in maintenance of the
transformed state in the tumors. Signals from probes to late gene
sequences were similar to or somewhat higher than those of
probes to intergenic HPV16 sequences. Therefore, the boost in
early gene signals is likely due to the RNA portion of the sample
target preparation. Among the early genes, strong signals were
observed for E6/E7 sequences, and for the E4 region of E2, which
is known to be a highly abundant RNA splicing product from the
primary transcript (34). Detailed interpretation of these data is
complicated by the ability of HPV16 to exist in episomal and mul-
tiple host-genome integrated forms (35), but probe signal differ-
cences within and between tumors will provide the identification of
potential agents having oncogenic activities and so lead to new
lines of investigation as a follow-up to the initial screening exper-
iment.
Averaging probe signals by accession provided a rapid and
rather uncomplicated means to summarize the data and collect
the strongest detection candidates. The sliding window analysis
generally matched AccSig results but provided better ability to
distinguish variants within the set of samples and offers the poten-
tial to detect candidates represented by only a portion of an acces-
sion. As used here, the MAT algorithm did require more labor
because it was applied to each sample in separate operations, but
this could be addressed by future automated scripting. Analyses at
the individual probe level helped to explain how candidates arose
in the AccSig and MAT results and are likely the only way in which
previously unknown pathogens with some sequence homology to
a conserved (or specific) probe can be detected. Thus, a PathoChip
screening project can generate a list of candidates prioritized by
the magnitude of detection, detection via multiple analysis strat-
gegies, and the rate of detection across the sample population.
Combining these results with annotations for the virus or patho-
genic microorganism such as host range, tissue specificity, or
prevalence in the general population will assist in determination
of which agents deserve further attention. This approach is likely
to provide a signature of a particular cancer or disease with agents
with various degrees of contribution. A window into the natural
conditions for commensal and pathogenic organisms will greatly
enhance our ability to diagnose and treat cancer and other possi-
bles diseases not yet linked to specific agents.
The PathoChip screening assay described here supported faster
laboratory and data analysis turnarounds than those for deep se-
quencing of whole-genome tumor DNA or RNA and coverage of
viral and eukaryotic genomes not assayed by 16S rRNA ap-
proaches. The Agilent SurePrint platform is relatively economical
compared to other microarray formats and is flexible for quick
production of customized probe subsets or updated metagenome
compilations, as well as being compatible with a variety of up-
stream sample preparation strategies. The PathoChip meta-
genomic assay allows for a comprehensive assessment of the fre-
cuency of coinfection by multiple organisms and their correlation
with driving oncogenic events. These events can lead to prolifera-
tion early in the infection process as well as over an extensive
period during which the contributions by these agents may vary or
have specific effects on the host cell important for disease devel-
oment. The data analysis workflows will test for statistically sig-
nificant interactions between these infectious agents. Critically, as
these studies unfold, the PathoChip data in combination with
patient genotyping, RNA profiling, and clinical data may be used
to search for genetic or environmental predispositions that influ-
ence the host-pathogen interactions important for initiation and
maintenance of the cancer phenotype.

**MATERIALS AND METHODS**

**Microarray design.** National Center for Biotechnology Information (NCBI) databases for genome, gene, and nucleotide accessions were qu-
eried (http://www.ncbi.nlm.nih.gov/pubmed) for all taxonomy virus an-
notations 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, Cat-
gegory A, B, and C Priority Pathogens). The resulting accessions were as-
sembled into a nonredundant concatenation with 100-N nucleotide
separators between accesses. This metagenome was divided into 58
“chromosomes” each around 5 to 10 million nucleotides (nt) in length
and submitted to Agilent Technologies (Santa Clara, CA) as a custom
design project. Probe sequences, at a maximum of 60 nt with nonhybrid-
izing spacers, were selected using the Agilent array comparative genomic
hybridization (aCGH) design algorithms and then filtered for low likeli-
hood of cross-hybridization to human genomic sequences.

Independently, low-complexity regions in the metagenome were

### TABLE 3 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 & test    | 64            | 11    |       | 4     |
| Pass outlier test| 65          | 66    |       | 9     |
| Conserved probes| 1            | 1     |       | 2     |
| Pass & test    | 1             | 1     |       | 0     |
| Pass outlier test| 1           | 0     |       | 2     |
FIG 6 Confirmation of HPV16 detection. (A) The heat map indicates test minus xhh signals for every HPV16 probe (columns) from PathoChip assays of the OSCC samples (rows). Row numbers are indicated for samples that are examples of no HPV16 signal (2021 and 2023), hybridization to nearly all probes (2022 and 2024), or hybridization to a smaller subset of probes (2032, 2035, 2053, and 2061). Probe locations are indicated relative to the transcript map for early (E) and late (L) genes, and black arrows show the positions of forward (f) and reverse (r) PCR primers. Probe names in boxes correspond to the oligomers used for capture bead enrichment and deep sequencing (cap-seq) of samples that were pooled as marked by the right axis bars. The histogram shows the sum of cap-seq

(Continued)
were moved ahead for cDNA conversion. In most of the samples, the fragment sizes were acceptable and recovery of most samples was relatively good, allowing for further processing. As expected, formalin-exposed RNA was partially degraded. However, the probes were filtered for minimum interprobe spacing of 100 bp and distribution that roughly covers the full length of each accession while limiting the number of probes to 10 to 20 per accession. The number of probes was not restricted for known oncogenic viral agents, creating a saturation tiling set covering these accessions. Entire genome sequences were covered to the extent possible with all available Agilent-designed probes. The microarray was supplemented with an additional number of predesigned aCGH probes for 660 genes and 602 intergenic regions from the human genome and Saccharomyces cerevisiae. Probes and accession annotations are available in the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/).

Sample preparation. Purified phiX174 virion DNA was purchased from New England Biolabs (N3023S; Ipswich, MA, USA), total DNA from human MRC-5 cells infected with cytomegalovirus (human herpesvirus 5 strain AD169) was ATCC VR-338D, total DNA from human A549 cells infected with adenovirus type 5 (HAdV-5 strain Adenoid 75) was ATCC VR-5D, and total RNA from human HEp-2 cells infected with respiratory syncytial virus (HRSV strain Long) was ATCC VR-26D, all purchased from ATCC (Manassas, VA, USA). Plasmid minipreps were prepared from pBR322 subclones carrying JC or BK polyomavirus genomes (J.C. Alwine, University of Pennsylvania, Philadelphia, PA) and from pUC19 carrying the human papillomavirus 16 (HPV16) genome (obtained from Peter Howley, Harvard Medical School, Boston, MA).

Oropharyngeal squamous cell carcinoma tumor samples were obtained from the Abramson Cancer Center’s Tumor Tissue and Biospecimen Bank (https://somapps.med.upenn.edu/pbr/portal/tumor). All samples were reviewed by our resident pathologist for case history and confirmed for tumor type and demarcation of the cancer cells. If significant adjacent normal tissue was present, sections were mounted on non-charged glass slides for dissection of tumor tissue using a template slide. Specimens containing mostly cancer cells were sectioned with a hematoxylin-and-eosin (H&E)-stained section with the cancer region clearly demarcated. Specimens containing mainly cancer cells and accessions annotations are available in the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/).

Whole-genome amplifications (WGAs) of genomic DNA and/or cDNA from random-primed, reverse-transcribed total RNA were performed with the Illustra GenomiPhi v2 kit (GE Healthcare Bio-Sciences, Pittsburgh, PA, USA), the Ovation WGA system (NuGEN, San Carlos, CA, USA), and GenomePlex or TransPlex kits (WGA2 and WTA2; Sigma-Aldrich, St. Louis, MO) using manufacturer-recommended protocols and input amounts. Amplification products were purified with the QiAquick PCR purification kit (Qiagen), and 2 μg was used for Cy3 dye labeling by the SureTag labeling kit (Agilent). Cy5 dye labeling was performed on 2 μg of human reference DNA from the Agilent SureTag kit, without prior WGA (experiment 1, Table 1) or after WGA (other experiments), as a control to report probe cross-hybridization to human DNA. Labeled DNA was purified with SureTag kit spin columns, and specific activities were calculated for use in hybridization reactions.

Microarray production and processing. SurePrint glass slide microarrays (Agilent Technologies Inc.) were manufactured with 60-nt DNA oligomers synthesized in 60,000 features on eight replicate arrays per slide. PathoChip v2a and v2b contained 60,000 probes to unique target regions and conserved plus saturation target regions, respectively. PathoChip v3 contained 37,704 probes to unique targets and 23,627 probes to conserved targets or to saturate known oncogenic and pathogenic viral agents.

Labeled samples were hybridized to microarrays as described in the Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis protocol (version 7.2, G4410-90010). Master mixes containing aCGH blocking agent, HI-RPM hybridization buffer, and Cot-1 DNA (pilot assays only) were added to a mixture of the entire labeled test sample and the xhh DNA control sample, denatured, and hybridized to arrays under 8-chamber gasket slides at 65°C with 20-rpm rotation for 40 h in an Agilent hybridization oven. Arrays were processed using wash procedure A and scanned on an Agilent SureScan G4900DA microarray scanner.

Microarray data analysis. Scanned microarray images were analyzed using 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. The intensity distribution and channel balance were not used for quality control because they are expected to have little or no signal, except for the control human probes.

Feature intensities for Cy3 and Cy5 channels were imported into the Partek Genomics Suite (Partek Inc., St. Louis, MO, USA). The average intensity for human intergenic control probes was calculated for cohybridized test and xhh DNA samples, and a scale factor was determined which would make the Cy5 xhh DNA average equal to the Cy3 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) (28) as implemented in Partek was used for sliding window analysis of probe signals (Cy3 minus Cy5) for each tumor sample. MAT parameters were P value cutoff of 0.99, window of 5,000 bp, minimum number of positive probes of 5, and dis-
card 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 xh 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.

HPV16 PCR and capture sequencing. PCR amplification reaction mixtures for HPV16 detection contained 100 ng of tumor DNA and primer fl (5′-AAGGGAAGACCGGGGTATG)-2′ (5′-AGGATGACCA CGCACATGGGG), r1 (5′-TGGGTGTGCAATTAGTGTC), r2 (5′-TGGGTTGCTCCTCATACTC), r3 (5′-GTGGGTTGCTAGCCTTTCT GT), or r4 (5′-TGCGAACGAGAAGCTACA). DNA was denatured at 94°C for 2 min, followed by 30 cycles of 94°C for 30 s, 57°C for 60 s, and 65°C for 60 s.

Magnetic bead capture was used to create libraries of targeted sequences for deep sequencing. Selected PathoChip probes with high signals for candidate organisms were synthesized as 5′-biotinylated DNA oligomers (Integrated DNA Technologies, Coralville, IA, USA), mixed as a 36-probe panel, including six probes for HPV16 (Fig. 6A), and hybridized to pools of tumor targets. Targets were captured by pooling the TransPlex products used for PathoChip screening (100 tumors over six pools) and then adding a probe panel aliquot containing 2.5 pmol of each probe to 150 ng of each target pool in 100-μl reaction mixtures with 1× TMAC buffer (3 M tetramethylammonium chloride, 0.1% Sarkosyl, 50 mM Tris-HCl, 4 mM EDTA, pH 8.0). The reaction mixtures were denatured at 100°C for 10 min followed by hybridization at 60°C for 3 h. M-280 streptavidin Dynabeads (Life Technologies, Carlsbad, CA, USA) (1,500 μg) were then added with continuous mixing at room temperature for 3 h, followed by three washes of the magnetically captured bead-probe-target complexes with 1 ml 2× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate) and three washes with 1 ml 0.1× SSC. Captured single-stranded target DNA was eluted in 50 μl Tris-EDTA (TE) at 100°C for 10 min.

The six capture eluates (1 μl) were reamplified by GenomePlex reactions (WGA3; Sigma-Aldrich), purified by Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, IN, USA) using the manufacturer’s protocol, and assessed for yield by Qubit double-stranded DNA (dsDNA) assays (Life Technologies, Inc.) and for size distribution by agarose gel electrophoresis. Sequencing libraries were prepared using the Nextera XT DNA sample preparation kit (Illumina, San Diego, CA, USA), with dual indexing and bead library normalization according to the manufacturer’s protocols. After Qubit quantitation, libraries were submitted to the Washington University Genome Technology Access Center (St. Louis, MO) for quantitative PCR (qPCR) quality control measurements, library pooling, and sequencing on one flow cell of an Illumina MiSeq instrument with paired-end 250-nt reads. Approximately 400,000 reads from the six OSCC libraries generated were aligned to the PathoChip metagenome or the human genome using the Bowtie2 aligner (36) in sensitive-local mode. Reads mapping to HPV16 with MapQ scores of 20 or better were identified using Integrative Genomics Viewer 2.2.25 (37).

Institutional oversight. The research described does not involve animals. Tumors from human subjects were collected with informed consent for research use and were received as deidentified samples. This study was approved by the University of Pennsylvania institutional review board.
Lee DJ. 2014. Microbial dysbiosis is associated with human breast cancer. PLoS One 9:e83744. http://dx.doi.org/10.1371/journal.pone.0083744.

17. Cox MJ, Cookson WO, Moffatt MF. 2013. Sequencing the human microbiome in health and disease. Hum. Mol. Genet. 22:R88–R94. http://dx.doi.org/10.1093/hmg/ddt398.

18. Ma Y, Madupu R, Karaouz E, Nossa CW, Yang L, Yoosheh S, Yachiminski PS, Brodie EL, Nelson KE, Pei Z. 2014. Human papillomavirus community in healthy persons, defined by metagenomics analysis of HMP (human microbiome project) shotgun sequencing datasets. J. Virol. 88: 4786–4797. http://dx.doi.org/10.1128/JVI.00093-14.

19. Brodie EL, Desantis TZ, Jolley DC, Back SM, Larsen JT, Andersen GL, Hazen TC, Richardson PM, Herman DJ, Tokunaga TK, Wan JM, Firestone MK. 2006. Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. Appl. Environ. Microbiol. 72: 6288–6298. http://dx.doi.org/10.1128/AEM.00246-06.

20. Chen EC, Miller SA, DeRisi JL, Chiu CY. 2011. Using a pan-viral microarray assay (Virochip) to screen clinical samples for viral pathogens. J. Virol. Exp. 50:2536. http://dx.doi.org/10.1379/jvi.2010.052536.

21. Tu Q, Yu H, He Z, Deng Y, Wu L, Van Nostard JD, Zhou A, Vood-eckers J, Qin Y, Hemme CL, Shi Z, Xue K, Yuan T, Wang A, Zhou J. 2014. GeoChip 4: a functional gene array-based high throughput environmental technology for microbial community analysis. Mol. Ecol. Resour. 14:914–928. http://dx.doi.org/10.1111/j.1755-0998.12239.

22. Wong CW, Heng CL, Wan Yee I, Soh SW, Kartasasmita CB, Simoes EA, Hibberd ML, Sung WK, Miller LD. 2007. Optimization and clinical validation of a pathogen detection microarray. Genome Biol. 8:R93. http://dx.doi.org/10.1186/gb-2007-8-5-93.

23. Norman JM, Handley SA, Virgin HW. 2014. Kingdom-agnostic metagenomics and the importance of complete characterization of enteric microbial communities. Gastroenterology 146:1459–1469. http://dx.doi.org/10.1053/j.gastro.2014.02.001.

24. Sato M, Ohtsuka M, Ohmi Y. 2014. Usefulness of repeated GenomiPhi, a phi29 DNA polymerase-based rolling circle amplification kit, for generation of large amounts of plasmid DNA. Biomol. Eng. 22:129–132. http://dx.doi.org/10.1016/j.bioeng.2005.05.001.

25. Arneson N, Hughes S, Houlston R, Done S. 2008. GenomePlex whole-genome amplification. CSH Protoc. 2008:4920. http://dx.doi.org/10.1101/pdb.prot4920.

26. Hirsch D, Camps J, Varma S, Gaiser T, Patterson BK, Ganem D, DeRisi JL. 2006. Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. Appl. Environ. Microbiol. 72: 6288–6298. http://dx.doi.org/10.1128/AEM.00246-06.

27. Arneson N, Hughes S, Houlston R, Done S. 2008. GenomePlex whole-genome amplification. CSH Protoc. 2008:4920. http://dx.doi.org/10.1101/pdb.prot4920.

28. Robinson M, Schache A, Sloan P, Thavaraj S. 2012. HPV specific testing: a requirement for oropharyngeal squamous cell carcinoma patients. Head Neck Pathol. 6:583–590. http://dx.doi.org/10.1007/s12105-012-0370-7.

29. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, Liu R, Arneson N, Hughes S, Houlston R, Done S. 2007. GenomePlex whole-genome amplification. CSH Protoc. 2008:4920. http://dx.doi.org/10.1101/pdb.prot4920.

30. Lee DJ, Lee Y, Pfeiffer RM, Hernandez BY, Xiao W, Kim E, Jiang B, Goodman MT, Sibug-Saber M, Cozen W, Liu L, Lynch CF, Chen EC, Miller SA, DeRisi JL. 2011. Human papillomavirus and rising oropharyngeal cancer incidence in the United States. J. Clin. Oncol. 29:4294–4301. http://dx.doi.org/10.1200/JCO.2011.36.4596.

31. Cuming SA, Cheun-Im T, Milligan SG, Graham SV. 2008. Human papillomavirus type 16 late gene expression is regulated by cellular RNA processing factors in response to epithelial differentiation. Biochem. Soc. Trans. 36:522–524. http://dx.doi.org/10.1042/BST0360522.

32. Thierry F. 2009. Transcriptional regulation of the papillomavirus onco- genes by cellular and viral transcription factors in cervical carcinoma. Virology 384:375–379. http://dx.doi.org/10.1016/j.virology.2008.11.014.

33. Doorbar J. 2013. The E4 protein: structure, function and patterns of expression. Virology 445:80–98. http://dx.doi.org/10.1016/j.virology.2013.07.008.

34. Xu B, Chotewutmonti S, Wolf S, Klos U, Schmitz M, Durst M, Schwarz E. 2013. Multiplex identification of human papillomavirus 16 DNA integration sites in cervical carcinomas. PLoS One 8:e66693. http://dx.doi.org/10.1371/journal.pone.0066693.

35. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25. http://dx.doi.org/10.1186/gb-2009-10-3-r25.

36. Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. 14:178–192. http://dx.doi.org/10.1093/bib/bbs017.

37. Wang D, Cos coy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, DeRisi JL. 2002. Microarray-based detection and genotyping of viral pathogens. Proc. Natl. Acad. Sci. U. S. A. 99:15687–15692. http://dx.doi.org/10.1073/pnas.242579699.

38. Chiu CY, Rouskin S, Koshy A, Urisman A, Fischer K, Yagi S, Schnurr D, Eckburg PB, Tompkins LS, Blackburn BG, Merker JD, Patterson BK, Ganem D, DeRisi JL. 2006. Microarray detection of human parainfluen- zavirus 4 infection associated with respiratory failure in an immunocom- petent adult. Clin. Infect. Dis. 43:e71–e76. http://dx.doi.org/10.1086/507896.

39. Urisman A, Molinaro RJ, Fischer N, Plummer SJ, Casey G, Klein EA, Malathi K, Magi-Galluzzi C, Tubbs RR, Ganem D, Silverman RH, DeRisi JL. 2006. Identification of a novel gammaretrovirus in prostate tumors of patients homozygous for R462Q RNASEL variant. PLoS Pathog. 2:e25. http://dx.doi.org/10.1371/journal.ppat.0020025.

40. Chiu CY, Alizadeh AA, Rouskin S, Merker JD, Yeh E, Yagi S, Schnurr D, Patterson BK, Ganem D, DeRisi JL. 2007. Diagnosis of a critical respiratory illness caused by human metapneumovirus by use of a pan-viral microarray. J. Clin. Microbiol. 45:2340–2343. http://dx.doi.org/10.1128/JCM.00364-07.

41. Kistler A, Avila PC, Rouskin S, Wang D, Ward T, Yagi S, Schnurr D, Ganem D, DeRisi JL. 2007. Pan-viral screening of respira- tory tract infections in adults with and without asthma reveals unexpected human coronavirus and human rhinovirus diversity. J. Infect. Dis. 196:815–827. http://dx.doi.org/10.1086/520816.

42. Chiu CY, Brenner AL, Kanada K, Kwok T, Fischer KF, Runckel C, Louie JK, Glaser CA, Yagi S, Schnurr DP, Haggerty TD, Parsonnet J, Ganem D, DeRisi JL. 2006. Identification of cardiovascular and respiratory tract infections in adults with and without asthma. J. Infect. Dis. 196:815–827. http://dx.doi.org/10.1086/520816.

43. Kistler AL, Gancz A, Clubb S, Skews-Cox P, Fischer K, Sorber K, Chiu CY, Lublin A, Mechan S, Farnoushi Y, Brenner A, Wenz CC, Karlene SB, Ganem D, DeRisi JL. 2008. Recovery of divergent avian bornaviruses from cases of proventricular dilatation disease; identification of a candidate etiologic agent. Virol. J. 5:108. http://dx.doi.org/10.1186/1743-422X-5-88.