FULL PAPER

BJC

British Journal of Cancer (2016) 115, 497–504 | doi: 10.1038/bjc.2016.213

Keywords: non-small cell lung cancer; NSCLC; carcinogenesis; virus; microarray; retrovirus; human papillomavirus; HPV

Molecular evidence of viral DNA in non-small cell lung cancer and non-neoplastic lung

Lary A Robinson*[,1,2], Crystal J Jaing[3], Christine Pierce Campbell[2,4], Anthony Magliocco[5], Yin Xiong[5], Genevra Magliocco[5], James B Thissen[3] and Scott Antonia[1]

[1]Department of Thoracic Oncology, Moffitt Cancer Center, Tampa, Florida 33612-9416, USA; [2]Center for Infection Research in Cancer (CIRC), Moffitt Cancer Center, Tampa, Florida 33612-9416, USA; [3]Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California 94559-9698, USA; [4]Department of Epidemiology, Moffitt Cancer Center, Tampa, Florida 33612-9416, USA and [5]Department of Pathology, Moffitt Cancer Center, Tampa, Florida 33612-9416, USA

Background: Although ~20% of human cancers are caused by microorganisms, only suspicion exists for a microbial cause of lung cancer. Potential infectious agents were investigated in non-small cell lung cancer (NSCLC) and non-neoplastic lung.

Methods: Seventy NSCLC tumours (33 squamous cell carcinomas, 17 adenocarcinomas, 10 adenocarcinomas with lepidic spread, and 10 oligometastases) and 10 non-neoplastic lung specimens were evaluated for molecular evidence of microorganisms. Tissues were subjected to the Lawrence Livermore Microbial Detection Array, an oncovirus panel of the International Agency for Research on Cancer, and human papillomavirus (HPV) genotyping. Associations were examined between microbial prevalence, clinical characteristics, and p16 and EGFR expression.

Results: Retroviral DNA was observed in 85% squamous cell carcinomas, 47% adenocarcinomas, and 10% adenocarcinomas with lepidic spread. Human papillomavirus DNA was found in 69% of squamous cell carcinomas with 30% containing high-risk HPV types. No significant viral DNA was detected in non-neoplastic lung. Patients with tumours containing viral DNA experienced improved long-term survival compared with patients with viral DNA-negative tumours.

Conclusions: Most squamous cell carcinomas and adenocarcinomas contained retroviral DNA and one-third of squamous cell carcinomas contained high-risk HPV DNA. Viral DNA was absent in non-neoplastic lung. Trial results encourage further study of the viral contribution to lung carcinogenesis.

Over the past three decades, research has linked a number of cancers to infectious agents, including viruses (e.g., Epstein–Barr virus, human papillomavirus (HPV), hepatitis B and C, human T-lymphotropic retrovirus, and polyomavirus), bacteria (*Helicobacter pylori*), and parasites (*Schistosoma haematobium* and *Clonorchis sinesis*; zur Hausen, 2011b). Perhaps the strongest documented causal relationship exists for HPV, which causes 99% of cervical cancers, 90% of anal cancers, 50% of penile, vulvar and vaginal cancers, and >70% of oropharyngeal cancers (Doorbar, 2006; Chaturvedi *et al*, 2011). An estimated, 20% of newly diagnosed cancer cases worldwide are caused by infectious agents, and it is believed that 'further links between specific infectious agents and human malignancies' will be indentified in the next few years (Sarid and Shou-Jiang, 2011; zur Hausen, 2011a).

In view of the strong link of HPV to epithelial cancers, the question has been raised whether this virus is involved in lung cancer. Over the past two decades, almost 90 published studies have searched for HPV in lung cancer with largely inconsistent results. Data from PCR-based studies suggest that the prevalence of HPV in squamous cell lung cancer is ~25%, with the lowest prevalence observed in the United States (15%) and Europe (17%), and the highest prevalence found in Asia (36%) (Klein *et al*, 2009). Differences in methodology and risk of contamination by HPV DNA may explain the inconsistent results observed in previously conducted studies.

Furthermore, few studies have investigated whether organisms other than HPV play a significant role in lung carcinogenesis, and this is likely related to prior methods employed to search for

*Correspondence: Dr LA Robinson; E-mail: lary.robinson@moffitt.org

Received 5 March 2016; revised 16 June 2016; accepted 20 June 2016; published online 14 July 2016

suspect microbes. To determine whether other microorganisms are associated with lung cancer, a generalised, non-targeted approach is needed for the detection of microbial DNA/RNA. Thus, this study was aimed at broadly screening archived frozen non-small cell lung cancer (NSCLC) specimens for suspect microorganisms using several methods, including a novel Lawrence Livermore Microbial Detection Array (LLMDA), an oncovirus panel developed by the International Agency for Research on Cancer (IARC), and HPV DNA genotyping. Additional molecular methods were employed to correlate with results, including p16 and epidermal growth factor receptor (EGFR) expression by immunohistochemistry.

MATERIALS AND METHODS

Patient and tumour selection. Patients who underwent surgical resection for NSCLC at Moffitt Cancer Center and consented to the Total Cancer Care protocol between 2000 and 2013 were eligible for this study. The institutional honest broker chose the tumours and control lung specimens randomly based on (1) having enough volume of frozen tissue to perform the studies, (2) preoperative radiographs showed no associated pneumonia or distal atelectasis, (3) gene expression data were already available for future studies, and (4) no patient had chemotherapy or radiotherapy before resection. The following patient data were collected: age, gender, prior history of cancer, smoking status, pack-years, survival (follow-up available in 56 out of 60 patients), and the date of tissue collection/storage. Approval for the use of archived tissue and patient information was obtained from the University of South Florida IRB, protocol no. MCC16765.

Snap-frozen tissue samples from a total of 70 NSCLC tumours and 10 non-neoplastic lung specimens were obtained for this study. Of the surgically resected stage I tumours, 30 were squamous cell carcinomas (SCCs), 10 adenocarcinomas, and 10 adenocarcinomas with lepidic spread (formerly bronchioloalveolar carcinomas). In addition, we selected 10 resected stage IV tumours (3 squamous cell and 7 adenocarcinoma) and their matched surgically resected distant oligometastases.

Tissue processing. Nucleic acids were extracted from snap-frozen tumour and non-neoplastic lung specimens (obtained at lung cancer resections). During tissue macrodissection, the selected frozen tissue section was mounted on the Cryostat and a 5 μm tissue section was cut and placed on a glass slide. The slides were stained with haematoxylin and eosin (H&E). The H&E slide was reviewed by a trained pathologist who marked areas of the desired tissue type. Using the H&E slide as a template, the desired tissue-type fragments were collected by scalpel from the OCT-embedded tissue on the chuck that represented the mirror image of the H&E slide.

DNA was extracted from harvested tissue using the DNeasy Blood and Tissue kit (Qiagen, Germantown, MD, USA; Kit: DNeasy Blood and Tissue Kit, 2014). Briefly, crushed tissue was homogenised for 60 s in 20 mg ml^{-1} lysozyme and 10 mg ml^{-1} RnaseA solution in buffer (20 mM TrisCl, pH 8.0, 2 mM sodium EDTA, and 1.2%Triton X-100) and glass beads using MiniBead-Beater (BioSpec Products, Bartlesville, OK, USA). The lysate was incubated at 37 °C for 30 min. Proteinase K was added to the tissue homogenate and incubated for 1–3 h at 56 °C. Subsequently, AL buffer and ethanol were added to the mixture, and were transferred on the silica column and spun. The column was washed twice with AW1 and AW2 buffers; DNA was eluted with AE buffer. Extracted DNA from tissue was frozen for analysis.

Lawrence Livermore Microbial Detection Array. The LLMDA was developed at the Lawrence Livermore National Laboratory (LLNL) and designed to detect all sequenced viral and bacterial families, with appropriate controls (Gardner *et al*, 2010; Victoria *et al*, 2010; Erlandsson *et al*, 2011). The 135 K format of the LLMDA (v.5) targets all vertebrate pathogens including 1856 viruses, 1398 bacteria, 125 archaea, 48 fungi, and 94 protozoa. Detailed methodology for the LLMDA assay has been previously published (Gardner *et al*, 2010). Oligonucleotide probes ($\approx$60 nt in length) were designed to detect all sequenced viral and bacterial families with a large number of probes per sequence (average 30 probes) to improve sensitivity and specificity in evaluating tissue nucleic acid samples for microorganisms (Gardner *et al*, 2010).

Probes were selected to avoid sequences with high levels of similarity to human, bacterial and viral sequences not in the target family. More conserved probes within a family were favoured to enable minimisation of the total number of probes needed to cover all existing genomes with high probe density per target. In the development of this microarray, PCR was used extensively to validate the results and verify that the statistical algorithm was accurate (Hewitson *et al*, 2014; Thissen *et al*, 2014; Jaing *et al*, 2015). The high-density oligo LLMDA and statistical analysis method has been extensively tested in numerous problems in viral and bacterial detection from pure or complex environmental or clinical samples. Recent published examples of this technology were used to identify a contaminating pig virus from a rotavirus vaccine and to detect various viral infections from human clinical samples (Hewitson *et al*, 2014), Epstein–Barr virus in human lymphoma tissues (Tellez *et al*, 2014), Kaposi's sarcoma-associated herpesvirus from human bladder cancer samples (Paradžik *et al*, 2014), bacterial pathogens from wounded soldiers (Be *et al*, 2014), and emerging viruses from human clinical infected samples (Rosenstierne *et al*, 2014).

The LLMDA was run on 50 NSCLC tumours (10 SCC, 10 adenocarcinoma, 10 adenocarcinomas with lepidic spread, 10 stage IV primary tumours, and 10 matched metastases) and 10 non-neoplastic lung tissues. Following extraction, nucleic acids were sent frozen to the LLNL. Nucleic acid samples were whole-genome amplified using a random primer PCR amplification protocol described previously in detail (Gardner *et al*, 2010).

Microarray hybridisation. For each sample, 1 μg of amplified product was fluorescently labelled using the Roche NimbleGen One-Color DNA Labeling Kit (#05223555001, Madison, WI, USA) according to the recommended protocols. The DNA and cDNA were purified after labelling, and hybridised using the NimbleGen Hybridization kit (Roche NimbleGen. Inc., Madison, WI, USA; Cat no. 05583683001) to the LLMDA according to the manufacturers' instructions. The microarrays were hybridised for 17 h and washed using the NimbleGen Wash Buffer kit (Roche NimbleGen. Inc.; #05584507001) according to the manufacturer's instructions. Microarrays were scanned on a MS-200 2 μm scanner from Roche Diagnostics (Pleasanton, CA, USA). The scanned tif image files were aligned using NimbleScan Version 2.4 software (Roche NimbleGen. Inc.) and pair text files were exported for analysis.

Data analysis. Data were analysed using automated LLMDA analysis algorithm-Composite Likelihood Maximization Method (Gardner *et al*, 2010). A 95% threshold was used in the data analysis to analyse only probes with signal intensity above 95% of random controls. The log likelihood for each possible target was estimated from the BLAST similarity scores of the array feature and target sequences, together with the feature sequence complexity and other covariates derived from the BLAST results. Non-specific probe binding or cross-hybridisation of probes was mitigated in the results by this statistical evaluation.

IARC oncovirus panel. A subset of NSCLC tumours (10 SCCss) was evaluated using an oncovirus screening panel developed at the IARC in Lyon, France. In this panel, the presence of 61 viral agents

(46 HPV types, 10 polyomaviruses, and 5 herpesviruses) was determined using type-specific multiplex genotyping assays, which combine multiplex PCR and bead-based Luminex technology (Luminex Corp., Austin, TX, USA; Corbex *et al*, 2014). Two primers for the amplification of β-globin were added to provide a positive control for the quality of the template DNA.

HPV PCR genotyping. All 70 NSCLC tumours (50 tumours evaluated by LLMDA and 20 additional SCCss) underwent HPV genotyping. The short-PCR-fragment technique employed in this assay amplifies the 65-bp fragment of the L1 open reading frame allowing detection based on reverse hybridisation. Specifically, DNA extracted from the frozen tissue underwent genotyping to detect HPV DNA using an AutoBlot 3000H processor (MedTec Biolab, Medtec, Inc., Hillsborough, NC, USA) and the INNO-LiPA Genotyping Extra Assay (Fujirebo Diagnostics, Inc., Malvern, PA, USA), which detects 28 HPV genotypes classified as high or low risk, depending on their association with carcinogenesis (Ragin *et al*, 2014).

Immunohistochemical studies. Formalin-fixed, paraffin-embedded (FFPE) tissue samples were obtained from the surgical pathology archives from the same patients from whom we extracted nucleic acids from their frozen, resected tumour. Of the 60 frozen primary tumour specimens studied, 5 FFPE blocks could not be located for the studies. Ten metastases from the primary tumours and the non-neoplastic lung specimens were not evaluated. These 55 tumours underwent immunohistochemical (IHC) evaluation for p16 and EGFR expression. Also, 26 documented positive and negative controls were utilised containing normal tissue, non-lung cancers, and known HPV-positive tissue. Sequencing EGFR for exon deletions was not performed.

Clinically, IHC staining of the tumour suppressor p16 is generally manually scored as positive or negative based on the subjective opinion of an experienced pathologist who makes the decision on the basis of the percentage of positive tumour cells. A positive result is felt to be indicative of the presence of HPV-positive tumour, especially oropharyngeal or cervical carcinoma, as p16 is generally considered to be a surrogate marker for high-risk HPV infection. Likewise, IHC staining for the mutant cell surface receptor EGFR was interpreted by manual scoring that, although subjective, has also been shown to be reproducible (Rüschoff *et al*, 2013).

For the current research study, we employed manual and machine scoring of p16 and EGFR to obtain the greatest possible accuracy. In contrast to the manual method indicating the percentage of positive staining tumour cells, machine scoring is based on the pixel intensity using the sum of positive pixels divided by the total number of pixels, creating the so-called H-score (Früh and Pless, 2012).

Tissue microarray construction. After collection of all of the corresponding FFPE blocks, two 1 mm diameter cores were taken from each lung tumour (55 primary tumours) and one core from each of 26 controls (6 non-lung cancers, 9 normal lung and tracheal mucosa, 5 normal organs, 3 normal laryngeal mucosa, and 3 HPV-containing tissues). The cores were then placed into the Beecher Manual Tissue Microarrayer TMA-1 (Beecher Instruments, Inc., Sun Prairie, WI, USA) recipient array block according to the manufacturer's protocol (Kononen *et al*, 1998).

p16 immunohistochemical staining. Standard microtome-sectioning techniques were used to make slides that were stained using Ventana Discovery XT automated system (Ventana Medical Systems, Tucson, AZ, USA) as per the manufacturer's protocol with proprietary reagents. Briefly, sections were de-paraffinized on the automated system with EZ Prep solution (Ventana). The heat-induced antigen retrieval method was used in Cell Conditioning 1 (CC1, Ventana). The mouse monoclonal antibody that reacts to p16 (#551154, Becton Dickinson, San Jose, CA, USA) was used at 1 : 100 concentration in Dako antibody diluent (Carpenteria, CA, USA) and incubated for 32 min. The Ventana OmniMap Anti-mouse Secondary Antibody was used for 16 min. The detection system is the Ventana ChromoMap kit and slides were then counterstained with haematoxylin. Sections were then dehydrated and cover-slipped per normal laboratory protocol for manual and machine scoring.

EGFR immunohistochemical staining. Standard microtome-sectioning techniques were used to make slides that were stained using Ventana Discovery XT automated system (Ventana Medical Systems) as per the manufacturer's protocol with proprietary reagents. Briefly, sections were de-paraffinized on the automated system with EZ Prep solution (Ventana). Enzymatic retrieval method was used in protease 1 for 4 min (Ventana). The mouse primary antibody that reacts to EGFR (#M3563, Dako) was used at 1 : 50 concentration in Dako antibody diluent and incubated for 60 min. The Ventana OmniMap Anti-mouse Secondary Antibody was used for 16 min. The detection system is the Ventana ChromoMap kit and slides were then counterstained with haematoxylin. Sections were then dehydrated and cover-slipped as per the normal laboratory protocol for manual and machine scoring.

Manual scoring. Stained sections of the tissue microarray (TMA) slides were carefully scored by an expert pathologist (A.M.). p16 and EGFR TMA slides were each manually scored on a scale from 0 to 3+ depending on the percentage of positive tumour cells: 0 (negative or occasional cells positive); 1+ (<10% cells positive); 2+ (10–49% cells positive); and 3+ (50–100% cells positive).

Machine scoring. The same slides were scanned by a digital optical microscope at ×20 (Aperio XL Brightfield Scanner, Leica Biosystems, Buffalo Grove, IL, USA). The TMA cores were then analysed by automated methods using Aperio's Positive Pixel Count v9 algorithm (Aperio Technologies, Inc., Vista, CA, USA). On the basis of the intensity of the pixel from the DAB staining, the pixel was placed into one of the four categories: negative, weak, moderate, and strong. The unit of measurement for this test is positivity, which shows the percentage of positive pixels or rather the sum of positive pixels divided by the total number of pixels. An H-score was calculated for each of the TMA cores using the automated raw data results found for positive pixel count. These values for both EGFR and p16 were used in the following formula: H-Score = 3 X (percentage of strong positive pixels) + 2 X (percentage of moderate positive pixels) + 1 X (percentage of weak positive pixels). Note the computerised H-score is different from manual H-score because the former measures pixel intensity rather than cell count. The computerised scores were then mapped onto the TMA map and were entered into the database for analysis.

Receiver operating characteristic analysis of machine scores. Receiver operating characteristic (ROC) curves for p16/EGFR were created with GraphPad (GraphPad Software, Inc., La Jolla, CA, USA) using 3 lung alveolar epithelial and 3 lung bronchial epithelial laboratory control samples *vs* the 55 lung cancer samples. The maximum sum of sensitivity and specificity (max (sensitivity + specificity)) was used as the threshold for p16/EGFR positive *vs* p16/EGFR negative.

Statistical evaluation

Microbial detection microarray data. Data from the pan-microbial detection array were analysed using the LLNL Composite Likelihood Maximization Method, described elsewhere (Gardner *et al*, 2010).

Immunohistochemical studies and clinical data correlation. Pearson correlation tests were performed on manual scores *vs* machine scores for p16 and EGFR, respectively, to decide how close the manual-scoring and machine-scoring techniques are to each other. χ^2-tests were performed on p16/EGFR positive *vs* negative for histology, tumour type, cigarette smoking status, pack-per-year, and prior cancer. Pearson correlation tests were also performed on the clinical data when p16 and EGFR were used as continuous variables. Kaplan–Meier method was used to examine survival distributions for 'with virus' *vs* 'without virus' and compared using the log-rank test.

RESULTS

LLMDA studies. Table 1 summarises the composite microarray results from the 70 tissue specimens tested using the LLMDA (Gardner *et al*, 2010). Virtually, all specimens were positive for endogenous human retrovirus, which at present are not thought to be pathogenic. When the frozen storage time of the tissue was compared with the presence of virus, we found that the two Delta-retroviruses BLV and STLV-6 are negatively correlated with tissue storage time (BLV Pearson Coefficient −0.29373643, *P*-value 0.028001; STLV-6 Pearson Coefficient −0.34021534, *P*-value 0.010301). That is, the longer the storage time, the less likely that BLV and STLV-6 are found in the tumour tissue, implying either DNA degradation over time or a change in the types of virus present in lung cancer over the years.

HPV type 57 was found in 60% of SCCs, but was absent in normal lung tissue and rare in other cell types. Exogenous Delta-retroviruses HTLV-2 and Bovine leukaemia virus were detected in 85% of SCCs, and Alpha or Delta retrovirus (Y53 sarcoma virus or STLV-2) were found in 47% of human invasive adenocarcinomas. None of these viruses were found in the 10 normal control lung specimens. Only 1 out of 10 (10%) adenocarcinomas with lepidic spread was positive for a retrovirus (Y53 sarcoma virus). As expected, all tumours and normal tissues were positive for endogenous retroviruses.

Oncovirus screening panel. When DNA extracted from the first 10 SCCs was sent to the IARC for type-specific multiplex genotyping assays, 3 out of 10 (30%) tumours were positive for high-risk (type 16 in two tumours) or low-risk HPV (type 44 in one tumour) and one tumour also was positive for Epstein–Barr virus and hepatitis B. Sixty-seven per cent of HPV tumours positive on the Oncovirus Panel were also positive for HPV on the microarray study. No retroviruses were included in this panel nor was HPV type 57 included. All specimens were positive for β-globulin that provided positive control indicating good quality of the template DNA.

HPV genotyping by PCR. Nine out of 30 (30%) SCCs were positive for HPV (types 16, 44, 51, or 52), 5 out of 17 (29.4%) adenocarcinomas were positive for HPV (types 16, 18, 39, or 68), 0 out of 10 adenocarcinomas with lepidic spread were positive for HPV, and 1 out of 10 (10%) of non-neoplastic lung tissue specimens was positive for HPV (type 16). All SCCs that were positive for HPV on the genotyping PCR assay were positive for HPV on the microarray study, and 50% SCCs were positive on all three assays (microarray, oncovirus panel, and genotyping PCR). Of note, each of the non-neoplastic lung tissue specimens was obtained at lung cancer surgery, but was not matched to the tumours studied.

Immunohistochemical studies

Manual vs machine scoring. When the manual scoring is compared with the machine H-score, the p16 manual and machine scores are highly correlated (r-square = 0.5775, *P*-value <0.0001) and the EGFR manual and machine scores are also highly correlated (r-square = 0.5417, *P*-value <0.0001). As machine scoring is likely more objective and reproducible, we will report only the continuous variable machine scores. On the basis of the ROC curves, we will also report some of the data, when appropriate, as binary variables (positive or negative) for p16 and EGFR.

p16 results. Using the ROC analysis of p16, the H-score value 43.4 was designated as the cutoff, such that a score greater than this value was considered positive. On the basis of this binary scoring (positive or negative), machine scoring, the p16 results for each cell type are shown in Table 2.

Table 2. Machine-scored binary (positive or negative) immunohistochemistry results

Tumour cell type	p16 positive (%)	EGFR positive (%)
Squamous cell carcinoma	23 out of 29 (88.5%)	21 (81%)
Of the HPV-positive tumours ($n=13$)	10 (77%)	6 (46%)
Adenocarcinoma ($n=17$)	15 (88%)	9 (55%)
Of the HPV-positive tumours ($n=5$)	4 (80%)	4 (80%)
Adenocarcinoma with lepidic spread ($n=9$)	5 (56%)	7 (78%)
No HPV-positive tumours ($n=0$)	0 (0%)	0 (0%)

Abbreviations: EGFR = epidermal growth factor receptor; HPV = human papillomavirus.

Table 1. Viral DNA found in tumour by microarray analysis

Lung cancer cell type	HPV (human papillomavirus, type 57)	HBV (hepatitis B virus)	HTLV-2 (human T-cell lymphotropic virus 2), a Delta-retrovirus	Bovine leukaemia virus (BLV, a Delta-retrovirus, similar to HTLV-1)	Y53 sarcoma virus (an Alpha-retrovirus)	STLV-1, 2, or 6 (simian T-cell lymphotropic viruses), all Delta retroviruses
Squamous cell ca. ($n=10$)	6 (60%)	9 (90%)	7 (70%)	8 (80%)	0	6 (60%)
Adenoca. ($n=10$)	1 (10%)	2 (20%)	0	0	6 (60%)	1 (10%)
Adenoca. with lepidic spread ($n=10$)	0	0	0	0	1 (10%)	0
Adenoca. stage IV ($n=7$)	0	0	0	0	1 (14%)	0
Squamous cell ca. stage IV ($n=3$)	0	0	0	0	1 (33%)	0
Normal lung ($n=10$)	0	0	0	0	0	0

Abbreviations: Adenoca. = adenocarcinoma; Ca. = carcinoma; HPV = human papillomavirus; HBV = hepatitis B virus; HTLV-2 = human T-cell lymphotropic virus-2; BLV = bovine leukaemia virus; STLV-1 = Simian T-cell lymphotrophic virus-1.

P16-positive and p16-negative tumours were equally likely to have viruses detected except that the p16 score positively correlated with the presence of high-risk HPV detected by PCR (Pearson coefficient P-value = 0.04949).

EGFR results. Using the ROC analysis of EGFR, the H-score value 0.7329 was designated as cutoff such that a score greater than this value was considered positive. On the basis of binary scoring (positive or negative), machine scoring, EGFR results for each cell type are shown in Table 2.

EGFR-positive tumours of all cell types compared to EGFR-negative tumours were more likely to have 'any virus detected' as well the hepatitis B, HPV-57, and the HTLV-2-positive tumours (Pearson coefficient P-values = 0.00676, 0.00032, 0.00002, and 0.00001, respectively). In addition, the continuous variable EGFR score positively correlated with the presence of any virus detected including hepatitis B, HPV 56, PLPPV-1, HTLV-2, and BLV (Pearson coefficient P-value = 0.00001–0.00676). However, there was no correlation of EGFR-positive tumours and the presence of high-risk HPV (P = 0.86161).

Virus species *vs* histology. Table 3 displays the correlations between virus species and histology as well as the correlations to continuous variable-scored p16 and EGFR. The most notable finding is SCCs are more likely to have HTLV-2 and BLV viruses than the other cell types (χ^2 P-value = 0.04379 and 0.02413). And invasive adenocarcinomas are more likely to have Y73SV virus (χ^2 P-value = 0.01079).

Correlations with patient characteristics. The median age is 77 (range 45–88) for the 50 stage I patients and 77 (range 67–82) for the 10 stage IV patients. Of stage I patients, there were 58% male and 42% female; for stage IV patients, there were 70% male and 30% female. There were 3 never-smokers (all adenocarcinoma patients), 13 current smokers, and 34 former smokers in the 50 stage I patients; there were 3 current and 7 former smokers (no never-smokers) in the 10 stage IV patients. In stage I patients, the median pack-year smoking history for current and former smokers combined is 50 (range 1–150) and for stage IV patients it is 60 (range 32–135). Table 4 displays the correlations between patient demographic characteristics and viral DNA found, and binary scoring of p16 and EGFR. Frozen tissue storage time correlations are included.

The Kaplan–Meyer (KM) survival curves in Figure 1 graphically display the comparison of patient survival based on whether they had a stage I lung cancer resection (estimated 5-year survival rate 72%) *vs* stage IV patients who had resection of a lung cancer and a metachronous resection of an oligometastasis (estimated 5-year survival rate 50%). In contrast to overall stage survival curves in Figure 1, there is longer patient survival after surgery regardless of stage if their tumour contained DNA from HPV or 'any virus detected,' or if the

Table 3. Histology, p16/EGFR, and virus correlations

Histology/statistical test	Virus	p16	EGFR
Squamous cell carcinoma/CS/tt	More likely presence of HTLV-2 and BLV (P = 0.04379 and 0.02413)	NC	Higher score than for adenoca. or adenoca. w/lep. (P = 0.00810 and 0.00187)
Adenocarcinoma/CS	More likely presence of Y73SV (P = 0.01079)	Higher score than adenoca. w/lep. (P = 0.01001)	NC
Adenocarcinoma with lepidic spread/tt	NC	Lower score than adenoca. (P = 0.04932)	NC
p16/PC	More likely presence of HPV (P = 0.04949)	—	—
EGFR/PC	More likely presence of 'any virus detected', HBV, HPV-57, PLPPV-1, HTLV-2, and BLV (P = 0.00676, 0.00032, 0.00002, 0.00619, 0.00001, and 0.00060)	—	—

Abbreviations: Adenoca = adenocarcinoma; Adenoca w/Lep = adenoca with lepidic features; BLV = bovine leukaemia virus; CS = χ^2-test; HBV = hepatitis B virus; HPV = human papillomavirus; HPV-57 = human papillomavirus-56; HTLV-2 = human T-lymphotropic virus-2; NC = no correlation; PLPPV-1 = lambda papillomavirus-1; PC = Pearson coefficient test; STLV-6 = Simian T-cell leukaemia virus-6; tt = *t*-test; Y73SV = Y73 sarcoma virus.

Table 4. Correlates with patient characteristics

Patient characteristic/statistical test	Virus	p16	EGFR
Age/PC	NC	NC	NC
Gender/CS	NC	NC	NC
Long-term survival/LH	Univar: longer survival if HPV positive (P = 0.00619). Multivar: longer survival if HPV positive or 'any virus detected' (P = 0.04075 and 0.01618). Shorter survival if Y73SV or HBV (P = 0.01335 and 0.00543).	Multivar: longer survival if p16 positive (P = 0.02391)	NC
Prior cancer/CS	STLV-6 more likely (P = 0.0444)	NC	NC
Smoking status/CS Never-smoker Former smoker Current smoker	Y73SV higher proportion in never-smokers (P = 0.02368).	Higher in current smokers (P = 0.01378)	NC
Pack-years/PC	STLV-6 more likely present in the greater number of pack-years (P = 0.03730).	NC	NC
Tissue storage time/PC	Longer the storage, the less likely to find BLV and STLV-6 (P = 0.028001 and 0.010301).	NC	NC

Abbreviations: BLV = bovine leukemia virus; CS = χ^2-test; HPV = human papillomavirus; HBV = hepatitis B virus; LH = log hazard ratio; Multivar = multivariable; NC = no correlation; PC = Pearson coefficient test; STLV-6 = Simian T-cell leukemia virus-6; Univar = univariable; Y73SV = Y73 sarcoma virus.

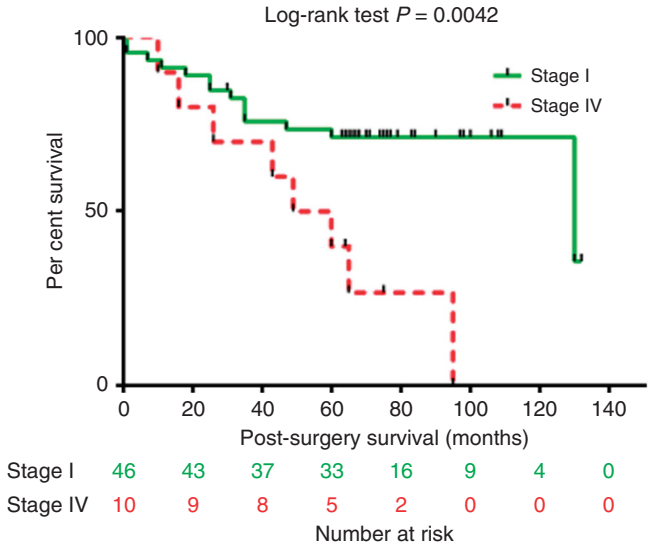

Figure 1. Months survival of NSCLC patients by stage.

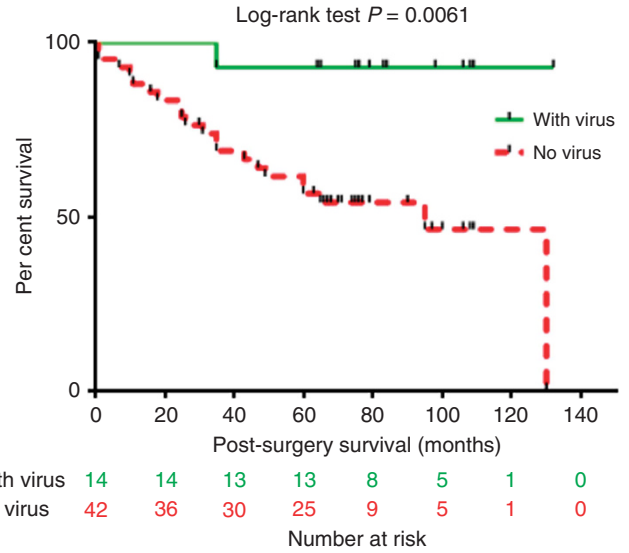

Figure 2. Months survival of all NSCLC patients by the presence or absence of high-risk HPV DNA in their tumour.

tumour was p16 positive (an independent predictor). The KM survival curves in Figure 2 graphically display the comparison of patient survival (all stages combined) based on the presence (estimated 5-year survival rate 93%) or absence (estimated 5-year survival rate 60%) of high-risk HPV DNA detected by PCR in each tumour.

DISCUSSION

Although sequencing with high-throughput methods is theoretically the optimal method to look for microorganisms, human DNA may make up >99% of reads, which makes it difficult to detect viruses that often are at low concentrations. Microarrays span the middle ground between PCR and sequencing, offering high probe density to detect diverse targets with lower costs and fast turnaround time. These reasons led to our choice of this technique in this study to screen for microorganisms in lung cancer, the most frequent cancer worldwide and the most common cause of cancer deaths annually (1.59 million deaths in 2012; World Health Organization, 2014).

Viruses in tumours. The pan-microbial microarray employed has a high probe density and provided unique results indicating a high incidence (69%) of DNA from all types of HPV in SCC, of which ~1/3 were high-risk HPV subtypes, which is a higher incidence than that found in prior Western studies. Interestingly, ~70% of SCCs were also positive for hepatitis B virus, which is a known oncovirus-causing liver cancer (Di Bisceglie, 2009) and may be involved in multiple myeloma (Li *et al*, 2016), but its relationship to lung malignancies is unclear.

However, most intriguing was the finding of a high incidence (85%) of several Delta retroviruses in lung SCC, in which one (Bovine leukaemia virus) also has recently been found with high incidence (59%) in 218 breast cancer specimens (Buehring *et al*, 2015). Usually transmitted through infected milk, Bovine leukaemia virus is present in 80–100% of raw milk, although it is likely rendered inactive with careful pasteurisation (Chung *et al*, 1986). A documented statistical link exists between the risk of breast cancer and development of a subsequent lung cancer (Prochazka *et al*, 2002). The Alpha-retrovirus Y53 sarcoma virus was found in ~1/2 of the invasive adenocarcinomas and this virus is similar to the oncogenic Rous sarcoma virus in fowl described ~100 years ago (Rous, 1911).

The retrovirus family is known to contain a number of oncoviruses including HTLV-1, a proven cause of human adult T-cell leukaemia (Bangham and Ratner, 2015). Almost all 26 cancers in animals with a known aetiology are caused by retroviruses, including the naturally occurring lung adenocarcinoma in sheep, which is identical in clinical presentation and histology to human lung adenocarcinoma. This tumour is known to be caused by the Jaagsiekte sheep Beta-retrovirus (JSRV; Palmarini and Fan, 2001). In addition, a recently published study using human lung cancer tissue arrays found human adenocarcinomas express an antigen that reacts with a JSRV Env-specific monoclonal antibody, and that exogenous JSRV-like *env* and *gag* sequences can be amplified from human lung cancer tissue arrays, all suggesting that a 'JSRV-like' virus may infect humans (Linnerth-Petrik *et al*, 2014).

High-risk HPV types, especially 16 and 18, infect squamous epithelia and appear to result in malignant transformation by expressing the oncoproteins E6 and E7, which dysregulate protein complexes, especially the tumour suppressor proteins p53 and pRB, which normally control cellular proliferation, differentiation, and apoptosis (McLaughlin-Drubin and Munger, 2008). HPV is an attractive oncovirus that may contribute to the genesis of at least some cases of SCC of the lung.

Most retroviruses with oncogenic capabilities have a long latency period and do not carry their own viral oncogenes. Rather, they induce tumours by integrating their provirus into the genome, and apparently by chance, the provirus integrates near a cellular proto-oncogene. Over 70 proto-oncogenes including c-myc and KRas have been known to be activated by the provirus (Oxnard *et al*, 2013). When the provirus is inserted upstream in the same transcriptional orientation near normal cellular proto-oncogenes, they activate their expression by proviral insertional mutagenesis (Butel, 2000; Maeda *et al*, 2008), where the promotor and enhancer elements of viral long terminal repeats can increase expression of the proto-oncogene. The finding of a high percentage of human SCCs and adenocarcinomas with retroviral DNA (none were found in normal lung specimens) provides attractive candidates to pursue for their carcinogenic potential.

Immunohistochemistry findings. The tumour suppressor protein p16 is a cyclin-dependent kinase inhibitor that has an important role in regulating the cell cycle, by slowing down the progression from G1 phase to S phase. This protein is being used clinically as a

prognostic biomarker for a number of cancers including oropharyngeal and cervical SCCs, where the presence of this biomarker is associated with a more favourable prognosis and is strongly associated with high-risk HPV infection (Oguejiofor *et al*, 2013).

Clinically, a p16-positive tumour is assumed to be a HPV-associated cancer. In the current study, the finding of p16 in a lung cancer specimen is significantly correlated with the presence of HPV virus (Table 3). High p16 expression may provide more convincing evidence that the virus is having a direct oncogenic effect on the cellular replication machinery and provides more evidence of a causative role.

Epidermal growth factor receptor exists on the cell surface and is activated by a number of specific ligands (including epidermal growth factor: coded on chromosome 4), leading to cell proliferation, inhibition of apoptosis, angiogenesis, and migration/adhesion/invasion (Oda *et al*, 2005). Mutations resulting in EGFR overexpression that lead to uncontrolled cell division are strongly linked to lung cancer (and other cancers) and are the therapeutic target of several effective pharmacologic inhibitors. In the current study, EGFR was overexpressed in 57% (8 out of 14) of HPV-positive tumours, 73% (11 out of 15) retrovirus-positive tumours, and 76% (16 out of 21) of tumours positive for any virus, strongly suggesting that upregulation of this cell surface receptor might have a role in viral carcinogenesis. It is conceivable that EGFR overexpression could attract virus to the tumour cell, perhaps defining a subgroup of cells more susceptible to viral invasion and leading to the false conclusion that viruses might contribute to lung tumorigenesis.

Correlations with patient characteristics. Of the various demographic factors, smoking status was significant in that tumours of never-smokers were more likely to contain Y73SV Alpha-retroviral DNA. The Delta-retrovirus STLV-6 was more likely to be found in tumours from smokers with a greater number of pack-years and from patients with a prior cancer. However, most intriguing was striking improvement in long-term patient survival in tumours containing retroviral DNA, HPV, or any virus (Figure 2), perhaps because these viruses stimulate a more vigorous immune response. However, being p16 positive is also an independent predictor of improved survival. Nevertheless, the viruses Y73SV or HBV, when present, appeared to decrease long-term patient survival.

Limitations of the study. Lung cancer specimens used were collected at surgery, snap-frozen and maintained at $-80\,^{\circ}C$ in the tissue bank from 2000 to 2013. We found that the longer the storage time of the frozen tumour in the tissue bank (some specimens were >10 years old), the less likely that BLV and STLV-6 were detected in the tumour tissue, suggesting that DNA degradation may have occurred over time. However, it is widely believed that long-term storage of frozen tissue will not result in significant degradation of nucleic acids (Baleriola *et al*, 2011).

Although PCR was used to validate construction of the pan-microbial array used in this study and verify that the statistical algorithm was accurate, this is the first time that this microarray technology has been used to evaluate lung cancer for the presence of microorganisms. Previously, the microarray technology has been used to identify EBV from human lymphoma (Tellez *et al*, 2014) and human herpesvirus 8 from bladder cancer samples (Paradžik *et al*, 2014). PCR was used to confirm the HPV findings in our study from the microarray; however, limited specimen quantity prohibited us from conducting PCR to confirm the various retrovirus findings. It is possible that the intriguing findings that HPV and several exogenous retroviral species were observed in a large proportion of archived lung cancers and not in normal lung may be explained by cross-hybridisation, as some retroviral species have high sequence similarity. The results in this discovery study will require additional confirmatory studies, which are currently in progress by the study investigators.

In addition, the frozen extracted nucleic acids were shipped overnight from Florida to California for the analysis periodically in batches over a 6-month period, and there is the potential that some specimens were degraded during transport and/or there were slight technical differences in the microarrays or technique that may have influenced the results.

Finally, the mere presence of viral DNA in lung tumours does not necessarily indicate causality, and conceivably the microorganism found might be a commensal or even be a contaminant of the tissue fixation or analysis technique. However, none of these suspicious viruses were found in non-neoplastic lung processed in the same manner.

An infectious agent, if truly present in tumour, may be involved in direct or indirect carcinogenesis (zur Hausen, 2011b). Specific, well-defined criteria such as that of zur Hausen (zur Hausen, 2011b) or Evans (Evans and Mueller, 1990) must be satisfied in future studies to confirm or disprove a causal role of an infectious agent in carcinogenesis. However, in this pilot study, we felt the first step was to discover what infectious agent(s) were present in the tumour tissue and then subsequent studies can be performed confirming functional activity of the virus and its possible contribution to malignant transformation.

CONCLUSIONS

Viral carcinogenesis is well documented in at least 20% of human cancers (zur Hausen, 2011b). Using a unique pan-microbial microarray and PCR, this pilot discovery study demonstrated the consistent presence of viral DNA from retroviruses, HPV, and several other viruses in a large percentage of NSCLC specimens, but not in non-neoplastic lung tissue. A number of intriguing immunohistochemical findings and clinical correlates accompanied the molecular findings. Although the mere physical presence of viral DNA in lung tumours does not prove causality, results encourage further study of the potential viral contribution to human lung oncogenesis.

ACKNOWLEDGEMENTS

This study was supported by the Paul Hoenle Foundation, Sarasota, Florida, USA.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Baleriola C, Johal H, Jacka B, Chaverot S, Bowden S, Lacey S, Rawlinson W (2011) Stability of hepatitis C virus, HIV, and hepatitis B virus nucleic acids in plasma samples after long-term storage at $-20\,^{\circ}C$ and $-70\,^{\circ}C$. *J Clin Microbiol* **49**: 3163–3167.

Bangham C, Ratner L (2015) How does HTLV-1 cause adult T-cell leukaemia/lymphoma (ATL)? *Curr Opin Virol* **14**: 93–100.

Be NAJ, Brown T, Gardner S, McLoughlin K, Forsberg J, Kirkup B, Chromy B, Luciw P, Elster E, Jaing C (2014) Microbial profiling of combat wound infection through detection microarray and next-generation sequencing. *J Clin Microbiol* **52**: 2583–2594.

Buehring G, Shen H, Jensen H, Jin D, Hudes M, Block G (2015) Exposure to bovine leukaemia virus is associated with breast cancer: a case-control study. *PLos One* **10**(9): e0134304.

Butel J (2000) Viral carcinogenesis: revelation of molecular mechanisms and etiology of human disease. *Carcinogenesis* **21**: 405–426.

Chaturvedi A, Engels E, Pfeiffer R, Hernandez B, Xiao W, Kim E, Jiang B, Goodman M, Sibug-Saber M, Cozen W, Liu L, Lynch C, Wentzensen N, Jordan R, Altekruse S, Anderson W, Rosenberg P, Gillison M (2011) Human papillomavirus and rising oropharyngeal cancer incidence in the United States. *J Clin Oncol* **29**: 4294–4301.

Chung Y, Prior H, Duffy P, Rogers R, MacKenzie A (1986) The effect of pasteurization on bovine leucosis virus-infected milk. *Aust Vet J* **63**: 379–380.

Corbex M, Bouzbid S, Traverse-Giehen A, Aouras H, McKay-Chpoin S, Carreira C, Lankar A, Tommasino M, Gheit T (2014) Prevalence of Papillomaviruses, Polyomaviruses, and Herpesviruses in triple-negative and inflammatory breast tumors from Algeria compared with other types of breast cancer tumors. *PLos One* **9**(12): E114559.

Di Bisceglie A (2009) Hepatitis B and hepatocellular carcinoma. *Hepatology* **49**(5 Suppl): S56–S60.

Doorbar J (2006) Molecular biology of human papillomavirus infection and cervical cancer. *Clin Sci* **110**: 525–541.

Erlandsson L, Rosenstierne MW, McLoughlin K, Jaing C, Fomsgaard A (2011) The microbial detection array combined with random Phi29-amplification used as a diagnostic tool for virus detection in clinical samples. *PLoS One* **6**(8): e22631.

Evans A, Mueller N (1990) Viruses and cancer. Causal associations. *Ann Epidemiol* **1**: 71–92.

Früh M, Pless M (2012) EGFR IHC score for selection of cetuximab treatment: Ready for clinical practice? *Transl Lung Cancer Res* **1**(2): 145–146.

Gardner SN, Jaing CJ, McLoughlin KS, Slezak TR (2010) A microbial detection array (MDA) for viral and bacterial detection. *BMC Genomics* **11**: 668.

Hewitson L, Thissen J, Gardner S, McLoughlin K, Glausser M, Jaing C (2014) Screening of viral pathogens from pediatric ileal tissue samples after vaccination. *Adv Virol* **2014**: 720585.

Jaing C, Thissen J, Gardner S, McLoughlin K, Hullinger P, Monday N, Niederwerder M, Rowland R (2015) Application of a pathogen microarray for the analysis of viruses and bacteria in clinical diagnostic samples from pigs. *J Vet Diagn Invest* **27**: 313–325.

Kit: DNeasy Blood and Tissue Kit (2014) Available at http://www.qiagen.com/products/catalog/sample-technologies/dna-sample-technologies/genomic-dna/dneasy-blood-and-tissue-kit.).

Klein F, Amin Kotb WFM, Petersen I (2009) Incidence of human papilloma virus in lung cancer. *Lung Cancer* **65**: 13–18.

Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schrami P, Leighton S, Torhorst J, Mihatsch M, Sauter G, Kallionimeni O (1998) Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* **4**(7): 844–847.

Li Y, Bai O, Liu C, Du Z, Wang X, Wang G, Li W (2016) Association between hepatitis B virus infection and risk of multple myeloma: a systemic review and meta-analysis. *Intern Med J* **46**(3): 307–314.

Linnerth-Petrik N, Walsh S, Bognaer P, Morrison C, Wooton S (2014) Jaagsiekte sheep retrovirus detected in human lung cancer tissue arrays. *BMC Res Notes* **7**: 160–171.

Maeda N, Fan H, Yoshikai Y (2008) Oncogenesis by retroviruses: old and new paradigms. *Rev Med Virol* **18**: 387–405.

McLaughlin-Drubin M, Munger K (2008) Viruses associated with human cancer. *Biochim Biophys Acta* **1782**: 127–150.

Oda K, Matsuoka Y, Funahashi A, Kitano H (2005) A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol Syst Biol* **1**(1): 0010.

Oguejiofor K, Hall J, Mani N, Douglas C, Slevin N, Homer J, Hall G, West C (2013) The prognostic significance of the biomarker p16 in oropharyngeal squamous cell carcinoma. *Clin Oncol (R Coll Radiol)* **25**(11): 630–638.

Oxnard G, Binder A, Janne P (2013) New targetable oncogenes in non-small cell lung cancer. *J Clin Oncol* **31**(8): 1097–1104.

Palmarini M, Fan H (2001) Retrovirus-induced ovine pulmonary adenocarcinoma, an animal model for lung cancer. *J Natl Cancer Inst* **93**: 1603–1614.

Paradžik M, Bučević-Popović V, Šitum M, Jaing C, Degoricija M, McLoughlin K, Ismail S, Punda-Polić V, Terzić J (2014) Association of Kaposi's sarcoma-associated herpesvirus (KSHV) with bladder cancer in Croatian patients. *Tumour Biol* **35**: 567–572.

Prochazka M, Granath F, Ekbom A, Shields P, Hall P (2002) Lung cancer risks in women with previous breast cancer. *Eur J Cancer* **38**(11): 1520–1525.

Ragin C, Obikoya-Malomo M, Kim S, Chen Z, Flores-Obando R, Gibbs D, Koriyama C, Aguayo F, Koshiol J, Caporaso N, Carpagnano G, Ciotti M, Dosaka-Akita H, Fukayama M, Goto A, Spandidos D, Gorgoulis V, Heideman D, van Boerdonk R, Hiroshima K, Iwakawa R, Kastrinakis N, Kinoshita I, Akiba S, Landi M, Eugene Liu H, Wang J, Mehra R, Khuri F, Lim W, Owonikoko T, Ramalingam S, Sarchianaki E, Syrjanen K, Tsao M, Sykes J, Hee S, Yokota J, Zaravinos A, Taioli E (2014) HPV-associated lung cancers: an international pooled analysis. *Carcinogenesis* **35**: 1267–1275.

Rosenstierne M, McLoughlin K, Olesen M, Papa A, Gardner S, Engler O, Plumet S, Mirazimi A, Weidmann M, Niedrig M, Fomsgaard A, Erlandsson L (2014) The microbial detection array for detection of emerging viruses in clinical samples–a useful panmicrobial diagnostic tool. *PLoS One* **9**: e100813.

Rous P (1911) A sarcoma of the fowl transmissible by an agent separable from the tumor cells. *J Exp Med* **13**: 397–411.

Rüschoff J, Kerr K, Grote H, Middel P, von Heydebreck A, Alves V, Baldus S, Büttner R, Carvalho L, Fink L, Jochum W, Lo A, López-Ríos F, Marx A, Molina T, Olszewski W, Rieker R, Volante M, Thunnissen E, Wrba F, Celik I, Störkel S (2013) Reproducibility of immunohistochemical scoring for epidermal growth factor receptor expression in non-small cell lung cancer: round robin test. *Arch Pathol Lab Med* **137**: 1255–1261.

Sarid R, Shou-Jiang G (2011) Viruses and human cancer: from detection to causality. *Cancer Lett* **305**: 218–227.

Tellez J, Jaing C, Wang J, Green R, Chen M (2014) Detection of Epstein-Barr virus (EBV) in human lymphoma tissue by a novel microbial detection array. *Biomark Res* **2**: 24.

Thissen J, McLoughlin K, Gardner S, Gu P, Mabery S, Slezak T, Jaing C (2014) Analysis of sensitivity and rapid hybridization of a multiplexed microbial detection microarray. *J Virol Methods* **201**: 73–78.

Victoria JG, Wang C, Jones MS, Jaing C, McLoughlin K, Gardner S, Delwart EL (2010) Viral nucleic acids in live-attenuated vaccines: Detection of minority variants and an adventitious virus. *J Virol* **84**: 6033–6040.

World Health Organization (2014) Cancer Fact Sheet. Available at http://www.who.int/mediacentre/factsheets/fs297/en/ (accessed on 27 July 2014).

zur Hausen H (2011a) Cancers with a possible infectious etiology. In *Infections Causing Human Cancer*, zur Hausen H (ed.), pp 485–503. Wiley-Blackwell: Weinheim, Germany.

zur Hausen H (2011b) Historical review. In *Infections Causing Human Cancer*, zur Hausen H (ed.), pp 17–26, 27–40. Wiley-Blackwell: Weinheim, Germany.