ORIGINAL ARTICLE

DNA copy number aberrations in endobronchial lesions: a validated predictor for cancer

Robert A A van Boerdonk,[1] Johannes M A Daniels,[2] Peter J F Snijders,[1] Katrien Grünberg,[1] Erik Thunnissen,[1] Mark A van de Wiel,[3] Bauke Ylstra,[1] Pieter E Postmus,[2] Chris J L M Meijer,[1] Gerrit A Meijer,[1] Egbert F Smit,[2] Thomas G Sutedja,[2] Daniëlle A M Heideman[1]

ABSTRACT

We recently identified a DNA copy number aberration (CNA)-based classifier, including changes at 3p26.3-p11.1, 3q26.2-29, and 6p25.3-24.3, as a risk predictor for cancer in individuals presenting with endobronchial squamous metaplasia. The current study was set out to validate the prediction accuracy of this classifier in an independent series of endobronchial squamous metaplastic and dysplastic lesions.

The study included 36 high-risk subjects who had endobronchial lesions of various histological grades that were identified and biopsied by autofluorescence bronchoscopy and were subjected to arrayCGH in a nested case–control design. Of the 36 patients, 12 had a carcinoma in situ or invasive carcinoma at the same site at follow-up (median 11 months, range 4–24), while 24 controls remained cancer free (78 months, range 21–142).

The previously defined CNA-based classifier demonstrated 92% (95% CI 77% to 98%) accuracy for cancer (in situ) prediction. All nine subjects with CNA-based classifier-positive endobronchial lesions at baseline experienced cancer outcome, whereas all 24 controls and 3 cases were classified as being low risk.

In conclusion, CNAs prove to be a highly accurate biomarker for assessing the progression risk of endobronchial squamous metaplastic and dysplastic lesions. This classifier could assist in selecting subjects with endobronchial lesions who might benefit from more aggressive therapeutic intervention or surveillance.

INTRODUCTION

Lung cancer is a major cause of cancer-related mortality worldwide.[1] This is largely due to the fact that the majority of patients are diagnosed with advanced stage disease when treatment options with curative intent are generally unfeasible. The ability to detect lung cancer at early or preinvasive stages together with adequate curative treatment could have a favourable impact on disease outcome. Early detection is currently available given sensitive clinical tools such as low-dose CT (LDCT)[2 3] and autofluorescence bronchoscopy (AFB).[4 5] With the advent of these imaging techniques, the challenge has become to differentiate true premalignant lesions with high-risk of progression to cancer from the many suspicious, though regressive lesions[6] to better target interventions and avoid overtreatment.

Key messages

What is the key question?
Can we validate our recently identified classifier based on specific DNA copy number aberrations (CNAs) as risk predictor for endobronchial cancer in individuals presenting with squamous metaplastic and dysplastic lesions?

What is the bottom line?
Our predefined CNA-based classifier, including changes at 3p26.3-p11.1, 3q26.2-29 and 6p25.3-24.3, was validated by demonstrating 92% (95% CI 77% to 98%) accuracy for cancer (in situ) prediction in an independent sample series.

Why read on?
The validated CNA-based classifier holds great clinical promise as an alternative or additive to histological grading and may provide guidance to clinicians on early interventional decisions for a generally equivocal set of endobronchial lesions.

In case of AFB, endobronchial sites showing abnormal autofluorescence can be biopsied for histopathological examination. Histological grading according to the WHO classification, however, does not provide an accurate estimate of cancer risk.[7–9] Whereas endobronchial lesions with squamous metaplasia and dysplasia are at risk of developing into lung cancer, with the highest risk being associated with the most severe grades,[7 9] these lesions behave erratically and only a minority truly progresses towards lung cancer. As such, the prediction of the progression risk of an individual endobronchial lesion by means of histological classification is imprecise and subject to considerable inter-observer variation.[10 11] Therefore, biomarkers are needed that can aid in assessing an individual's risk of subsequent cancer to tailor intervention strategies in case these lesions are detected.

In the search for such biomarkers, we recently analysed a series of endobronchial squamous metaplastic lesions with known clinico-pathological outcome (ie, location-matched endobronchial cancer in cases and cancer-free outcome in controls)

by means of arrayCGH.[12] This study discovered two distinct clusters at the molecular level with significant association to lesion outcome, that is, one cluster with lesions showing multiple DNA copy number aberrations (CNAs), including exclusively progressive lesions of cases, and one with lesions showing no or relatively few CNAs, including all non-progressive lesions of controls and one lesion of a case. In this discovery series of endobronchial squamous metaplastic lesions, a molecular classifier was built based on CNAs at 3p26.3-p11.1 (loss), 3q26.2-29 (gain) and 6p25.3-24.3 (loss) that predicted cancer diagnosed within 44 months with 97% accuracy.[12]

In the current study, we validate the prediction accuracy of the CNA-based classifier using an independent set of endobronchial lesions of various histological grades. A case–control arrayCGH profiling study was performed nested within a cohort of subjects at risk of lung cancer, but without evidence of prevalent carcinoma in situ (CIS) or invasive carcinoma at baseline bronchoscopy, who underwent AFB at regular time intervals. The series comprised endobronchial lesions of 12 patients who were identified with CIS or invasive carcinoma (≥CIS) at the initial site on follow-up bronchoscopy (further referred to as cases) and 24 controls who remained cancer free. Accuracy of the previously defined CNA-based classifier in predicting endobronchial cancer in this series at baseline was determined.

METHODS
Patient population
This study was nested within a cohort of 479 subjects at risk of lung cancer based on smoking habits (ie, more than 20 pack years), COPD, signs and symptoms, and/or a history of ear–nose–throat (ENT) or lung cancer, who visited the Department of Pulmonary Diseases, VU University Medical Center (Amsterdam, The Netherlands) for baseline AFB examination between October 1995 and November 2011. Clinical and histological characteristics of individual AFB-visualised lesions were carefully documented as subjects underwent AFB surveillance at regular time intervals (on average 3–6-month time intervals) with biopsy collection for histopathological examination from all sites showing abnormal autofluorescence at the respective or previous visits. For each site, paired formalin-fixed paraffin-embedded (FFPE) and frozen biopsies guided by AFB were obtained after informed consent from subjects. Study approval was obtained by the Institutional Review Board of VU University Medical Center. Follow-up data of subjects were available through the Dutch nationwide network and registry of histopathology and cytopathology (PALGA, Bunnik, The Netherlands).

A total of 42 subjects without evidence of prevalent ≥CIS were documented with location-matched CIS or invasive carcinoma at the initial site within a follow-up period exceeding 90 days. Of these, 12 case subjects (10 men, 2 women) were eligible for this study based on the following: confirmed absence of ≥CIS in the location-matched baseline endobronchial biopsy by comprehensive histopathological review of H&E-stained FFPE and corresponding cryo-sections (see below); availability of sufficient DNA quality and quantity following laser-capture microdissection of frozen biopsy for molecular analysis; and not included in the previous study.[12]

Twenty-four control subjects were chosen in such a way that relative to cases a similar distribution was ensured for those parameters that in theory might confound the analyses, that is, baseline routine histopathological diagnosis, age, gender, smoking habits, COPD and history of cancer, but who remained cancer free for a follow-up period that exceeded that of

respective cases. At the individual level, each case had two matched controls. For matching, continuous variables were categorised and controls were chosen to fall within the same category as their matching case for at least four of the six above-mentioned parameters. The number of controls was chosen to be twice that of the cases because a further increase in sample size was considered to render only modest improvement of power given the fixed number of the case group and the expected homogeneous nature of arrayCGH profiles in the control group.[12] Follow-up data, supplemented through PALGA, supported the cancer-free status of control subjects for a median of 78 months (range 21–142 months).

Histopathological evaluation
Two designated pathologists (ET, KG) who were blinded for the clinicopathological information and molecular findings performed an independent review of all baseline and follow-up biopsies using H&E-stained FFPE and cryo-preserved tissue sections. Histopathological assessment of FFPE tissue sections was performed in accordance with the WHO/International Association for the Study of Lung Cancer histological classification system of preinvasive and invasive squamous lesions of the bronchus.[13 14] Cryo-sections were categorised using the following composite grouping scheme: normal histology (N); low-grade disease (LGD), including squamous metaplasia (SqM), mild dysplasia (miD) and moderate dysplasia (moD); high-grade disease (HGD), including severe dysplasia (SD); and CIS or invasive carcinoma. Despite the limitations faced in accurate histological examination of H&E-stained frozen sections, this approach is reasonably accurate for assessing sample representativeness, and the research purpose of correlating histology with molecular analysis.[15] In case the grade of dysplasia could not be further defined, the term 'dysplasia-indefinite for grade' (D-ind) was used. Lesions were ultimately classified into three categories: LGD, HGD and indeterminate on the basis of the highest histomorphological abnormality scored on H&E-stained FFPE and frozen sections.

ArrayCGH analysis
A designated molecular biologist (RvB) performed sample preparation and genomic DNA isolation of frozen biopsy specimens as described before.[12] To specifically enrich for epithelial cells (≥70%), all samples were laser-capture microdissected using a Leica ASLMD microscope (Leica, Wetzlar, Germany) prior to genomic DNA extraction, as described before.[12] Subsequent linear whole genome amplification, Cy3-/Cy5-labelling (both from Enzo Life Sciences, Farmingdale, New York, USA), and hybridisation of laser-capture microdissected biopsy-derived DNAs onto the Agilent 4×44 K arrayCGH platform (Agilent Technologies, Santa Clara, California, USA) were performed as previously described.[12 16 17] Samples were co-hybridised against a Megapool Reference DNA (Kreatech Diagnostics, Amsterdam, The Netherlands), a homogeneous pool of DNA isolates from 100 healthy male or female individuals (XY mix and XX mix, respectively). Across arrayCGH comparisons were made to measure test signals.[17] Array slide scanning and image data acquisition were performed as described before (Agilent CGH_107_Sep09 protocol).[12] Oligonucleotides were mapped according to the human genome build NCBI 36 (March 2006). The entire dataset is available through the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/), under GSE accession number GSE45287. Accompanying data analyses were done as previously described,[12] including smoothing, segmentation, calling and regioning of the current dataset.

Measuring *PIK3CA* copy number by quantitative real-time PCR

PIK3CA gene copy number was determined by quantitative real-time PCR (qPCR) on a 7500 Fast Real-Time PCR System (Applied Biosystems, Nieuwerkerk a/d IJssel, The Netherlands) using non-amplified genomic DNA essentially as described before.[12] All qPCR reactions were carried out in duplicate and the threshold cycle numbers were averaged.

Data and statistical analysis

The study conformed to guidelines regarding Standards for the Reporting of Diagnostic accuracy studies (STARD). The molecular analyses were performed in a blinded fashion and results were correlated to clinicopathological data afterwards. Unsupervised hierarchical clustering analysis of arrayCGH data was performed to determine the association between genomic profiles and lesion outcome.[18] The association between clustering results and case/control status was determined by χ^2 testing. Cancer risk was predicted by using the previously defined statistical 'endobronchial cancer risk' model based on CNAs at 3p26.3-p11.1, 3q26.2-29 and 6p25.3-24.3.[12] When the predicted probability is larger than 0.5, classification of the sample is 'high risk' for endobronchial cancer. Prediction accuracy, sensitivity and specificity of the CNA-based classifier for endobronchial cancer (in situ) were calculated in this series. Given the skewed study population, the denominators for positive and negative predictive value calculations are essentially unknown and therefore these parameters were not calculated. Strength of the association between copy number classification based on arrayCGH and real-time PCR analyses was examined using McNemar testing. Two-sided p values below 0.05 were considered statistically significant. Statistical analyses were performed using IBM SPSS Statistics V.20.0 software package (New York, USA).

RESULTS

Baseline characteristics of study subjects and their endobronchial lesions

Table 1 shows demographics, histology and clinical variables for the individuals of the current study stratified for lesion outcome (ie, cases and controls). Detailed histological diagnoses at study entry and molecular findings are listed per lesion analysed in the online supplementary data file. The arrayCGH DNA copy number profiles of the endobronchial lesions of all controls (24/24; 100%) and 3 (25%) cases were quiescent with only few genomic aberrations (average percentage of altered features 0.2%, range 0.0–2.4%; cluster 1, figure 1) while the DNA copy number profiles of the remaining cases (9/12, 75%) displayed multiple genomic aberrations (average percentage of altered features 51.8%, range 27.6–78.2%; cluster 2, figure 1). A significant association between cluster assignment by unsupervised hierarchical clustering analysis and lesion outcome was found (p<0.001). To determine most common CNAs in our series, the frequency of gains and losses was plotted per genomic probe (figure 2). Highly frequent genomic aberrations (≥50% frequency) in cases included gains at 3q23-q29 and 19q13.12-q13.2 and losses at 3p26.3-p11.1, 5q11.1-q35.3, 8p23.3-p12 and 9p24.3-p21.1.

Validation of the risk classifier

We next applied the output from the model built in our previous study,[12] that is, a risk classifier based on specific CNAs at 3p26.3-p11.1 (loss), 3q26.2-q29 (gain) and 6p25.3-p24.3 (loss),

Table 1 Characteristics of study subjects and their lesions

Characteristic	Lesion outcome	
	Cancer (in situ) (case subjects)	Cancer free (control subjects)
Subjects, n	12	24
Gender		
Male	10	19
Female	2	5
Categorised baseline histology*		
LGD	3	11
HGD	7	10
Indeterminate	2	3
Age at baseline (years)		
Median	67	66
Range	62–81	46–75
Mean±SD	68±5	66±6
Smoking status		
Current	6	11
Former	6	13
Number of pack-years		
Median	40†	42
Range	30–51†	23–120
COPD status		
Yes	9	17
No	3	6
Unknown	–	1
COPD staging		
GOLD I—II	6	9
GOLD III—IV	3	8
History of ENT/lung cancer		
Yes	10	14
No	2	10
AFB follow-up time (months)		
Median	11	28‡
Range	4–24	10–80‡
Overall follow-up time (months)		
Median	32	78
Range	8–118	21–142

*Lesions were classified into three categories: LGD, HGD and indeterminate on the basis of the highest histomorphological abnormality scored on H&E-stained FFPE and frozen sections (see online supplementary table 1 for detailed information per lesion).
†Number of pack-years is missing for one subject.
‡AFB follow-up is missing for three subjects.
AFB, autofluorescence bronchoscopy; ENT, ear–nose–throat; GOLD, Global Initiative for Chronic Obstructive Lung Disease; HGD, high-grade disease; indeterminate, indefinite for histological grade; LGD, low-grade disease.

using a predicted probability cutoff of 0.5 to assign individuals to high or low risk groups, and tested whether the assignments are related to lesion outcome (table 2). All subjects with CNA classifier-positive endobronchial lesions at baseline experienced cancer outcome, whereas all 24 controls and the remaining 3 cases were classified as low-risk. No intermediate test results with a probability around the threshold level of 0.5 were observed in this series. The CNA classifier had a 92% (95% CI 77% to 98%) prediction accuracy, 75% (95% CI 46% to 92%) sensitivity and 100% (95% CI 84% to 100%) specificity for endobronchial carcinoma (in situ) in the current series (table 2).

Predominant value of chromosomal region 3q

Aiming to fine-tune the CNA-based classifier, logistic regression analysis with forward feature selection using both arrayCGH

Figure 1 Unsupervised hierarchical clustering analysis of 36 baseline endobronchial lesions revealed two clusters: cluster 1, in which the complete set of controls congregated, and cluster 2, in which the majority of cases were assigned. Genomic loci are marked by corresponding chromosome numbers (Y axis). Loci that were lost, gained or amplified are shown in red, green and white, respectively. Cases (black boxes) and controls (white boxes) are represented on the X axis.
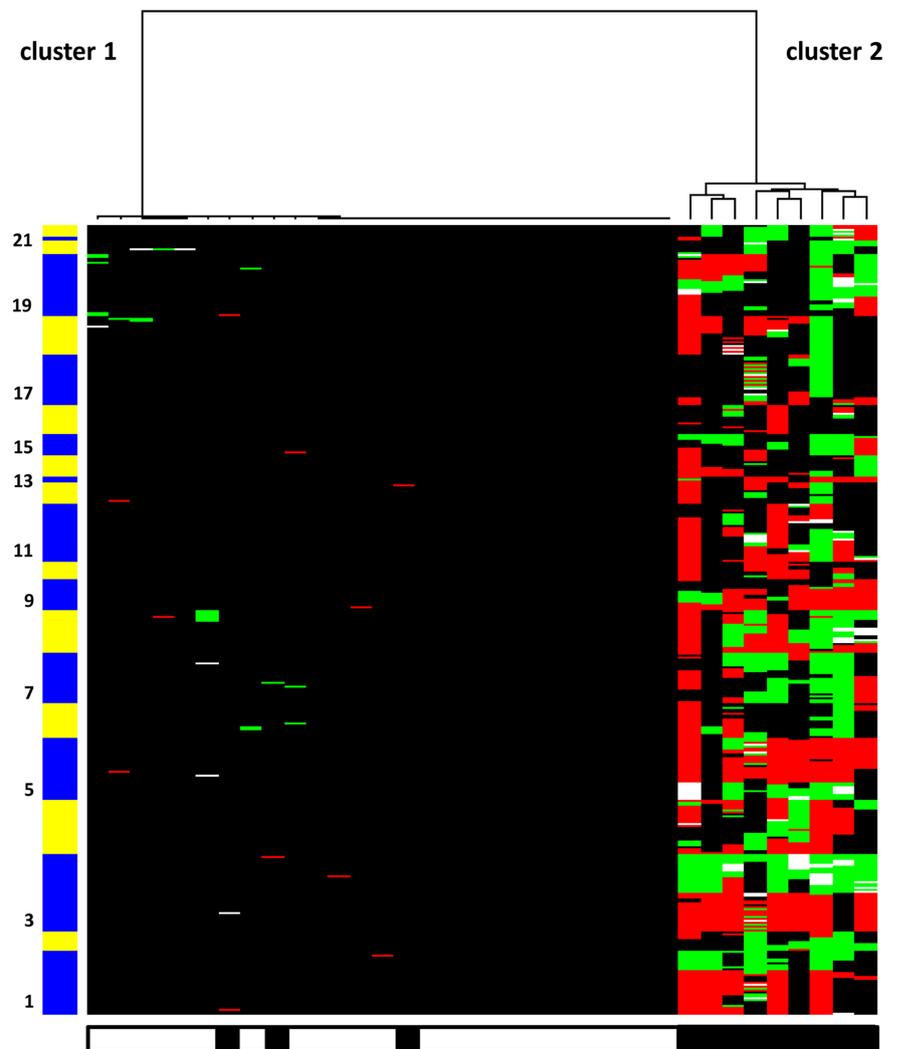
datasets combined (GSE23644 and GSE45287; total of 18 cases and 47 controls) was performed. This confirmed CNAs at 3p26.3-p11.1, 3q26.2-q29 and 6p25.3-p24.3 as parameters of the classifier, and could not reveal additional predictive DNA copy number feature(s) for refinement of the risk model. When we examined the relative contribution of the individual chromosomal regions to our risk model, the gain at 3q26.2-q29 contributed most. Gain at 3q26.2-q29 was present in virtually all lesions classified as high risk by the model (13/14, 93%). We further assessed the value of a gain at 3q26.2-q29 for risk prediction by independent qPCR analysis of the *PIK3CA* gene located within this chromosomal region (figure 3). A 95% concordance between arrayCGH and qPCR analyses was obtained, with one case sample revealing a 3q26.32 gain by arrayCGH while not by qPCR, and two control samples showing solely a gain by qPCR. The qPCR had an 85% (95% CI 74% to 92%) prediction accuracy, 61% (95% CI 39% to 80%) sensitivity and 94% (95% CI 82% to 98%) specificity for endobronchial carcinoma (in situ).

DISCUSSION

In the present work, we validated our previously defined CNA-based classifier[12] to assess an individual's risk for subsequent endobronchial cancer among subjects who present with endobronchial squamous metaplastic and dysplastic lesions. Our predefined classifier of CNAs at 3p26.3-p11.1 (loss), 3q26.2-29

(gain) and 6p25.3-24.3 (loss) demonstrated 92% (95% CI 77% to 98%) accuracy for predicting cancer outcome in this independent validation set. All subjects with CNA classifier-positive endobronchial lesions experienced cancer outcome. Of note, the classifier performed well in lesions graded as squamous metaplasia, as shown here and in the previous study,[12] and had predictive value along the full spectrum of preinvasive disease at the histological level while retaining 100% specificity. Thus, CNAs prove to be a highly accurate biomarker for assessing the progression risk of endobronchial squamous metaplastic and dysplastic lesions. The validated classifier proposed here holds great promise for improving the clinical management of subjects diagnosed with endobronchial squamous metaplastic or dysplastic lesions.

During recent years, numerous molecular studies have attempted to unravel whether specific (epi)genetic aberrations underlying lung carcinogenesis are detectable in the premalignant respiratory epithelium and would allow for assessing a person's risk of developing lung cancer.[19–27] Yet, molecular biomarkers that confer accurate risk stratification for subjects harbouring one or more endobronchial preinvasive lesions have not been validated. The lack of validated markers may be explained by the fact that the majority of discovery studies involved cross-sectional sample series, thereby not taking into account individual lesion's outcome.[20–24 26] Since only few endobronchial squamous metaplastic and dysplastic lesions eventually progress

Figure 2 Frequency plots of copy number aberrations (CNAs) are shown for (A) baseline endobronchial lesions of cases (n=12) and (B) baseline endobronchial lesions of controls (n=24). Percentages of gains (in green) and losses (in red) as determined by 4×44 K arrayCGH analysis are shown per spotted oligonucleotide on the positive and negative Y axis, respectively. Dashed lines within the figures indicate 50% frequency of gain or loss.
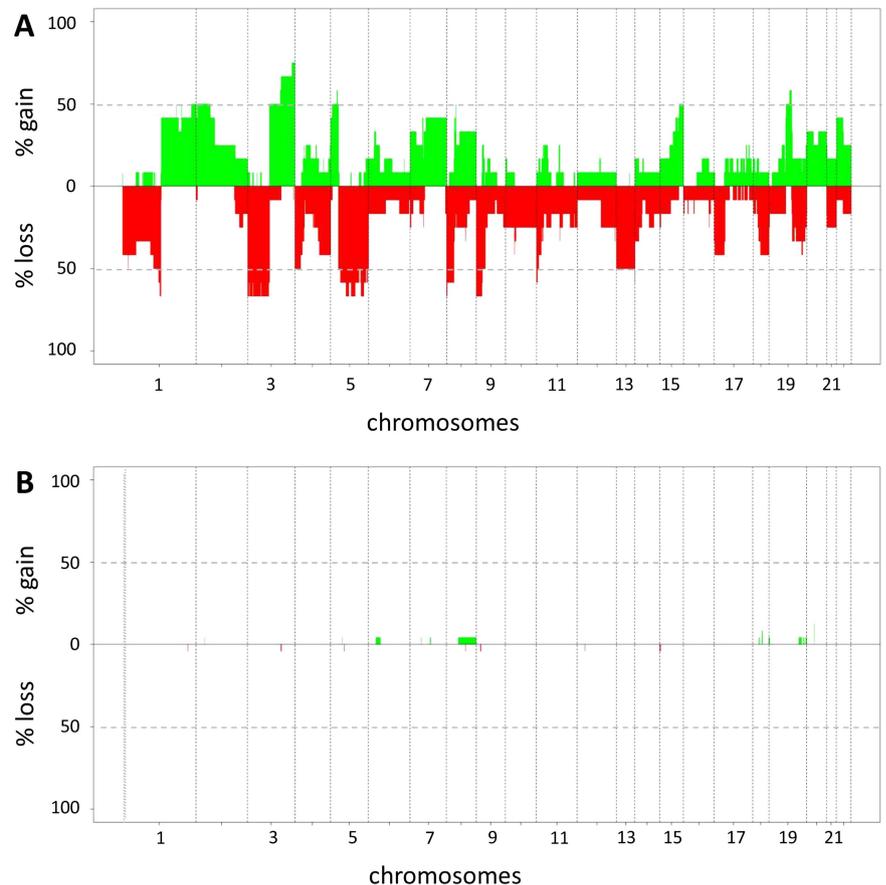
to carcinoma (in situ), longitudinal cohort studies are desired to identify molecular events that drive progression of these lesions into a malignant phenotype. However, collecting these longitudinal series requires an extensive period of close surveillance of large numbers of individuals at risk of lung cancer. This underscores the uniqueness of the series examined in this study.

The results presented here, together with our previous findings,[12] provide a consistent pattern of CNAs in endobronchial squamous metaplastic and dysplastic lesions that are predictive of endobronchial cancer. This is in agreement with the few longitudinal studies that previously reported amplifications of 3q[19 25] and regional losses at chromosomes 3p,[19 27] 5q and 9p.[27] Though these studies mainly defined markers within a very limited sample set or by low-resolution analyses, they

further confirm the robustness of CNAs as classifiers. The almost uniform finding of chromosome 3q26.2-q29 gain/amplification in our case series pinpoints the key importance of this chromosomal aberration in squamous lung carcinogenesis,[19 25 28] as was also recognised in other tumour types.[29 30] Putative target oncogenes within this region are PIK3CA, SOX2 and TP63.[31] Interestingly, the two regions that added up to our classifier (ie, losses at 3p26.3-p11.1 and 6p25.3-24.3) were also (partly) identified as areas of significant CNAs in a large cohort of squamous cell lung cancers analysed in The Cancer Genome Atlas project, and include putative target tumour suppressor genes such as FOXP1 and FHIT.[32]

Regarding the clinical performance of the classifier, the specificity of 100% (95% CI 84% to 100%) is proficient; allowing early interventional decisions in subjects with endobronchial lesions who were classified as high risk by the presence of CNAs. Intermediate test results with a probability around the threshold level of 0.5 were infrequently observed (ie, not in this series and once in the previous study[12]), supporting the robustness and utility of the CNA classifier. However, the sensitivity of the CNA classifier could be the subject of improvement. In its current format, the classifier lacks sufficient sensitivity for dismissal of subjects with endobronchial lesions that were classified as low risk by the absence of CNAs from further surveillance. The three cases of the current series that were inaccurately classified included case 3 with at maximum the clinical endpoint diagnosis of CIS, case 9 with inadequate quality measures for accurate histological diagnosis of the cryo-biopsy obtained 16 months prior to cancer diagnosis, and case 8 with an indefinite (ie, no consensus) diagnosis in the cryo-biopsy 4 months prior to cancer diagnosis. For the latter case, raw data processing suggested a gained chromosomal segment at 3q, however

Table 2 Risk classifier

Outcome risk classifier*	Lesion outcome			
	Cancer (in situ)	Cancer free	Total	
Low risk	3	24	27	Prediction accuracy: 92% (95% CI 77% to 98%)
High risk	9	–	9	Sensitivity: 75% (95% CI 46% to 92%)
Total	12	24	36	Specificity: 100% (95% CI 84% to 100%)

*Risk was predicted by using the previously defined model based on copy number aberrations (CNAs) at 3p26.3-p11.1, 3q26.2-29 and 6p25.3–24.3.[12] When the predicted probability is larger than 0.5, classification of the sample is 'high risk' for endobronchial cancer (in situ). Prediction accuracy, sensitivity and specificity of the CNA classifier for endobronchial cancer (in situ) were calculated in this series.
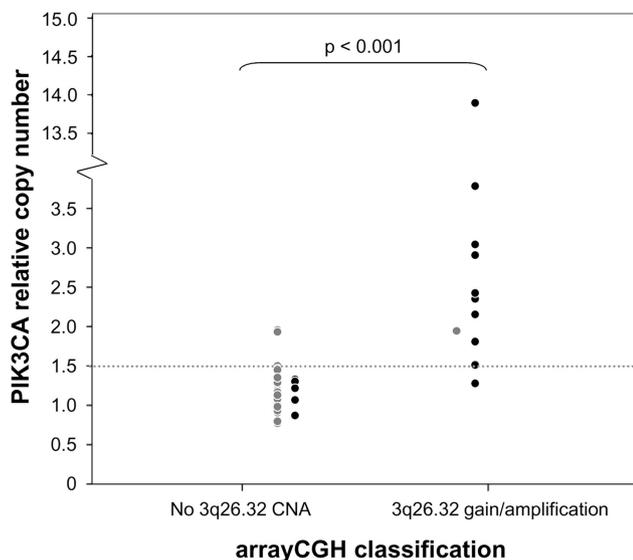
Figure 3 PIK3CA quantitative real-time PCR. Dot plot of the PIK3CA gene copy number ratios per haploid genome as determined by quantitative PCR (qPCR) (Y axis) in relation to the copy number status of the 3q26.32 locus as determined by arrayCGH analysis (X axis) is shown for all lesions of the current study (12 cases, 24 controls) and previous study[12] (6 cases, 23 controls). Cases are indicated by black circles and controls by grey circles. A significant difference in PIK3CA copy number ratios was found between arrayCGH-classified groups (ie, 'no 3q26.32 CNA' or '3q26.32 gain/amplification') (p<0.001). Dashed gray horizontal line indicates cut-off value (ie, 1.5 gene copies per haploid genome) for defining gained or amplified PIK3CA gene copy numbers. Copy number classifications were highly similar between arrayCGH and real-time PCR methodology (p=1.00). Prediction accuracy of qPCR assay for endobronchial cancer (in situ) using a cut-off of 1.5 gene copies per haploid genome is 85% (55/65; 95% CI 74% to 92%), sensitivity 61% (11/18; 95% CI 39% to 80%) and specificity 94% (44/47; 95% CI 82% to 98%).

values did not reach the threshold level for being finalised as a gain by the calling algorithm used. Overall, these case data may point to sampling bias, limitation of biopsy quality and/or histopathological examination,[33] or inaccuracy of the classifier at early onset of (pre)malignant disease as critical factors in the clinical sensitivity of the classifier. Particularly, the contribution of 3q gain to the classifier offers possibilities for a simple diagnostic method, for example qPCR assay, to be applied in biomaterials such as sputum[34] or bronchial brushings,[35] however at the cost of some diagnostic accuracy as shown in this study.

The strengths of this study include the validation of the CNA-based classifier as an objective, molecular determinant for lung cancer risk assessment in an independent sample set and the uniqueness of the longitudinal sample series. A limitation of this study may be the number of lesions included. Although we adjusted for potential confounders, some influence of demographic or clinical parameters on lesion outcome cannot be completely ruled out. Furthermore, the prediction accuracy, sensitivity and specificity figures of the CNA-based classifier may be imprecise due to the small numbers. The latter illustrates the relative infrequency of central airway lesions and underscores the uniqueness of the cohort described in this study.

In summary, our data validate that the presence of specific CNAs in endobronchial squamous metaplastic and dysplastic lesions predicts endobronchial cancer. The validated classifier holds great promise, as alternative or additive to histological grading, for stratification of subjects with endobronchial lesions

for risk of subsequent cancer. The classifier may have important practical implications providing guidance to clinicians on early interventional decisions for a generally equivocal set of endobronchial lesions, thereby allowing one to tailor treatment choices and avoid unnecessary follow-up procedures. The classifier may further be of great value as a surrogate endpoint reflecting cancer development or risk in therapeutic and/or chemopreventive trials. A challenge for future research is to assess the effect of earlier intervention of endobronchial squamous metaplastic and dysplastic lesions as guided by the classifier onto lung cancer-specific mortality and the quality of life.

REFERENCES

1 Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2009. *CA Cancer J Clin* 2009;59:225–49.
2 Aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011;365:395–409.
3 Gohagan JK, Marcus PM, Fagerstrom RM, et al. Final results of the Lung Screening Study, a randomized feasibility study of spiral CT versus chest X-ray screening for lung cancer. *Lung Cancer* 2005;47:9–15.
4 Chen W, Gao X, Tian Q, et al. A comparison of autofluorescence bronchoscopy and white light bronchoscopy in detection of lung cancer and preneoplastic lesions: a meta-analysis. *Lung Cancer* 2011;73:183–8.
5 Lam S, Kennedy T, Unger M, et al. Localization of bronchial intraepithelial neoplastic lesions by fluorescence bronchoscopy. *Chest* 1998;113:696–702.
6 Vansteenkiste J, Dooms C, Mascaux C, et al. Screening and early detection of lung cancer. *Ann Oncol* 2012;23(Suppl 10):x320–7.
7 Breuer RH, Pasic A, Smit EF, et al. The natural course of preneoplastic lesions in bronchial epithelium. *Clin Cancer Res* 2005;11(2 Pt 1):537–43.
8 Ishizumi T, McWilliams A, MacAulay C, et al. Natural history of bronchial preinvasive lesions. *Cancer Metastasis Rev* 2010;29:5–14.
9 Jeremy GP, Banerjee AK, Read CA, et al. Surveillance for the detection of early lung cancer in patients with bronchial dysplasia. *Thorax* 2007;62:43–50.
10 Nicholson AG, Perry LJ, Cury PM, et al. Reproducibility of the WHO/IASLC grading system for pre-invasive squamous lesions of the bronchus: a study of inter-observer and intra-observer variation. *Histopathology* 2001;38:202–8.
11 Venmans BJ, van der Linden HC, Elbers HR, et al. Observer variability in histopathologic reporting of bronchial biopsy specimens: influence on the results of autofluorescence bronchoscopy in detection of preinvasive bronchial neoplasia. *J Bronchol* 2000;7:210–4.
12 van Boerdonk RA, Sutedja TG, Snijders PJ, et al. DNA copy number alterations in endobronchial squamous metaplastic lesions predict lung cancer. *Am J Respir Crit Care Med* 2011;184:948–56.
13 Travis WD, Colby TV, Corrin B. *World Health Organization classification of lung and pleural tumors*. 3rd edn. Berlin: Springer, 1999.
14 Travis WD, Brambilla E, Muller-Hermelink KM. *Pathology and genetics: tumours of the lung, pleura, thymus and heart*. Lyon: IARC, 2004.
15 Flake GP, Rivera MP, Funkhouser WK, et al. Detection of pre-invasive lung cancer: technical aspects of the LIFE project. *Toxicol Pathol* 2007;35:65–74.
16 Buffart TE, Tijssen M, Krugers T, et al. DNA quality assessment for array CGH by isothermal whole genome amplification. *Cell Oncol* 2007;29:351–9.
17 Buffart TE, Israeli D, Tijssen M, et al. Across array comparative genomic hybridization: a strategy to reduce reference channel hybridizations. *Genes Chromosomes Cancer* 2008;47:994–1004.
18 Van Wieringen WN, van de Wiel MA, Ylstra B. Weighted clustering of called array CGH data. *Biostatistics* 2008;9:484–500.

19 Foster NA, Banerjee AK, Xian J, et al. Somatic genetic changes accompanying lung tumor development. *Genes Chromosomes Cancer* 2005;44:65–75.

20 Jeanmart M, Lantuejoul S, Fievet F, et al. Value of immunohistochemical markers in preinvasive bronchial lesions in risk assessment of lung cancer. *Clin Cancer Res* 2003;9:2195–203.

21 Jonsson S, Varella-Garcia M, Miller YE, et al. Chromosomal aneusomy in bronchial high-grade lesions is associated with invasive lung cancer. *Am J Respir Crit Care Med* 2008;177:342–7.

22 Ma J, Gao M, Lu Y, et al. Gain of 1q25-32, 12q23-24.3, and 17q12-22 facilitates tumorigenesis and progression of human squamous cell lung cancer. *J Pathol* 2006;210:205–13.

23 Massion PP, Taflan PM, Shyr Y, et al. Early involvement of the phosphatidylinositol 3-kinase/Akt pathway in lung cancer progression. *Am J Respir Crit Care Med* 2004;170:1088–94.

24 Massion PP, Zou Y, Uner H, et al. Recurrent genomic gains in preinvasive lesions as a biomarker of risk for lung cancer. *PLoS One* 2009;4:e5611.

25 McCaughan F, Pole JC, Bankier AT, et al. Progressive 3q amplification consistently targets SOX2 in preinvasive squamous lung cancer. *Am J Respir Crit Care Med* 2010;182:83–91.

26 Pelosi G, Del Curto B, Trubia M, et al. 3q26 Amplification and polysomy of chromosome 3 in squamous cell lesions of the lung: a fluorescence in situ hybridization study. *Clin Cancer Res* 2007;13:1995–2004.

27 Salaun M, Sesboue R, Moreno-Swirc S, et al. Molecular predictive factors for progression of high-grade preinvasive bronchial lesions. *Am J Respir Crit Care Med* 2008;177:880–6.

28 McCaughan F, Pipinikas CP, Janes SM, et al. Genomic evidence of pre-invasive clonal expansion, dispersal and progression in bronchial dysplasia. *J Pathol* 2011;224:153–9.

29 Bierkens M, Wilting SM, Van Wieringen WN, et al. Chromosomal profiles of high-grade cervical intraepithelial neoplasia relate to duration of preceding high-risk human papillomavirus infection. *Int J Cancer* 2012;131:E579–85.

30 Singh B, Stoffel A, Gogineni S, et al. Amplification of the 3q26.3 locus is associated with progression to invasive cancer and is a negative prognostic factor in head and neck squamous cell carcinomas. *Am J Pathol* 2002;161:365–71.

31 Heist RS, Sequist LV, Engelman JA. Genetic changes in squamous cell lung cancer: a review. *J Thorac Oncol* 2012;7:924–33.

32 Hammerman PS, Hayes DN, Wilkerson MD, et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012;489:519–25.

33 Lam S, MacAulay C, LeRiche JC, et al. Detection and localization of early lung cancer by fluorescence bronchoscopy. *Cancer* 2000;89(11 Suppl):2468–73.

34 Gomperts BN, Spira A, Elashoff DE, et al. Lung cancer biomarkers: FISHing in the sputum for risk assessment and early detection. *Cancer Prev Res (Phila)* 2010;3:420–3.

35 Gustafson AM, Soldi R, Anderlind C, et al. Airway PI3K pathway activation is an early and reversible event in lung cancer development. *Sci Transl Med* 2010;2:26ra25.

SUPPORTING INFORMATION

DNA COPY NUMBER ABERRATIONS IN ENDOBRONCHIAL LESIONS: A VALIDATED PREDICTOR FOR CANCER by *Robert A.A. van Boerdonk; Johannes M.A. Daniels; Peter J.F. Snijders; Katrien Grünberg; Erik Thunnissen; Mark A. van de Wiel; Bauke Ylstra; Pieter E. Postmus; Chris J.L.M. Meijer; Gerrit A. Meijer; Egbert F. Smit; Thomas G. Sutedja; Daniëlle A.M. Heideman.*

SUPPLEMENTARY TABLE 1. DETAILS OF BASELINE ENDOBRONCHIAL LESIONS: HISTOLOGY AND CORRESPONDING MOLECULAR FEATURES

Case subjects	Study ID	Routine histology diagnosis FFPE biopsy	Review histology diagnosis FFPE biopsy	Review histology diagnosis frozen biopsy	Aberrant genomic features (%)	CNA-based classifier
	1	N	ntd	LGD	27.6	pos
	2	SqM	miD	LGD	60.6	pos
	3	SqM	SqM	LGD	0.0	neg
	4	SqM	D-ind	no consensus (HGD / LGD)	78.2	pos
	5	moD	miD	no consensus (HGD / LGD)	71.0	pos
	6	SD	no consensus (ntd / SD)	HGD	58.2	pos
	7	SD	D-ind	D-ind	56.9	pos
	8	SD	moD	no consensus (D-ind / N)	0.2	neg
	9	SD	D-ind	ntd*	0.1	neg
	10	SD	SD	HGD	31.7	pos
	11	SD	ntd	D-ind	44.3	pos
	12	D-ind	SD	HGD	38.0	pos

Control subjects	Study ID	Routine histology diagnosis FFPE biopsy	Review histology diagnosis FFPE biopsy	histology diagnosis frozen biopsy	Aberrant genomic features (%)	CNA-based classifier
	1	SqM	SqM	LGD	0.0	neg
	2	SqM	no consensus (miD / SqM)	LGD	0.4	neg
	3	SqM	SqM	LGD	2.4	neg
	4	SqM	SqM	LGD	0.0	neg
	5	miD	miD	LGD	0.0	neg
	6	miD	miD	LGD	0.1	neg
	7	miD	SqM	LGD	0.0	neg
	8	miD	miD	LGD	0.0	neg
	9	moD	miD	LGD	0.0	neg
	10	moD	D-ind	D-ind	0.0	neg
	11	moD	ntd	LGD	0.0	neg
	12	moD	SqM	LGD	0.0	neg
	13	SD	SqM	LGD	0.0	neg
	14	SD	miD	LGD	0.0	neg
	15	SD	no consensus (moD / SqM)	N	0.1	neg
	16	SD	no consensus (moD / SqM)	LGD	0.0	neg
	17	SD	SD	HGD	0.0	neg
	18	SD	SqM	LGD	0.0	neg
	19	SD	miD	N	0.1	neg
	20	SD	ntd	N	0.0	neg
	21	SD	SqM	LGD	0.0	neg
	22	SD	D-ind	N	0.8	neg
	23	D-ind	ntd	LGD	0.1	neg
	24	N	ntd	D-ind	0.1	neg

LEGENDS FOR SUPPLEMENTARY TABLES

Histology review diagnoses are reported as the consensus diagnosis, or 'no consensus' (*i.e.*, no agreement between the two expert pathologists) with individual assignment indicated between brackets. Abbreviations used: D-ind, dysplasia-indefinite for grade; HGD, high-grade disease; LGD, low-grade disease; miD, mild dysplasia; moD, moderate dysplasia; N, normal histology; neg, negative (CNAs not present); ntd, not able to determine (*i.e.*, specimen of inadequate quality measures); pos, positive (CNAs present); SD, severe dysplasia; SqM, squamous metaplasia. *) absence of ≥CIS confirmed.