Genomics 99 (2012) 18–24

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

A computational index derived from whole-genome copy number analysis is a novel tool for prognosis in early stage lung squamous cell carcinoma

Ornella Belvedere [a,1], Stefano Berri [a,1], Rebecca Chalkley [a], Caroline Conway [a], Fabio Barbone [b], Federica Pisa [b], Kenneth MacLennan [a,c], Catherine Daly [a], Melissa Alsop [a], Joanne Morgan [a], Jessica Menis [a], Peter Tcherveniakov [a,d], Kostas Papagiannopoulos [a,d], Pamela Rabbitts [a], Henry M. Wood [a,*]

[a] Leeds Institute of Molecular Medicine, University of Leeds, Leeds, LS9 7TF, UK
[b] Division of Hygiene and Epidemiology, Department of Medical and Biological Sciences, University of Udine and University Hospital of Udine, 33100 Udine, Italy
[c] Department of Histopathology, St. James's University Hospital, Leeds, LS9 7TF, UK
[d] Department of Thoracic Surgery, St. James's University Hospital, Leeds, LS9 7TF, UK

ARTICLE INFO

Article history:
Received 15 August 2011
Accepted 19 October 2011
Available online 25 October 2011

Keywords:
Lung cancer
Copy number
Survival
Next-generation sequencing

ABSTRACT

Squamous cell carcinoma of the lung is remarkable for the extent to which the same chromosomal abnormalities are detected in individual tumours. We have used next generation sequencing at low coverage to produce high resolution copy number karyograms of a series of 89 non-small cell lung tumours specifically of the squamous cell subtype. Because this methodology is able to create karyograms from formalin-fixed paraffin-embedded material, we were able to use archival stored samples for which survival data were available and correlate frequently occurring copy number changes with disease outcome. No single region of genomic change showed significant correlation with survival. However, adopting a whole-genome approach, we devised an algorithm that relates to total genomic damage, specifically the relative ratios of copy number states across the genome. This algorithm generated a novel index, which is an independent prognostic indicator in early stage squamous cell carcinoma of the lung.

1. Introduction

Lung cancer genomes have been extensively studied benefitting from the large number of patient samples and the relative ease with which cell lines can be created [1,2]. Non-small cell lung cancer (NSCLC), although comprising three distinct major histological subtypes, adenocarcinoma (ADC), squamous cell carcinoma (SCC) and large cell carcinoma (LCC) has historically been viewed as a single group for the purposes of clinical management. Recently, this approach has been challenged and increasingly it is acknowledged that different histotypes should be treated as different diseases [3].

SCC appears to be the most homogenous of the subtypes, specifically with regard to commonality of genomic features. Non-squamous cancers are now frequently treated with pemetrexed; similarly, patients whose tumours harbour mutation of the epidermal growth factor receptor, most of which are adenocarcinomas, benefit from treatment with tyrosine kinase inhibitors. To date, however, no genomic-based therapy has been approved for SCC. Besides amplification of distal 3q, identified in almost all studies of SCC genomes, two recent large studies have shown that amplification of 8p12 is a common feature of SCC. At this time, as happens frequently in the history of identification of the driver gene within an amplicon, two different genes have been proposed as candidates; namely, BRF2 by Lockwood et al. [4] and FGFR1 by Weiss et al. [5]. Ultimately both may prove to have a role.

Because genomic changes initiate and drive cancer development, cancer genomics, besides suggesting potential drug targets has the potential to identify markers of prognosis and predictors of response to treatment. Most studies have adopted a candidate marker or candidate pathway approach, deciding *a priori* the marker(s) to be investigated [6]. Additionally, several gene signatures based on gene expression profiling, have been proposed to predict survival or response to treatment in NSCLC [7–9]. Two recent studies have reported the relationship between genomic changes and disease outcome in NSCLC. Kim et al. have identified several chromosomal regions as negative independent prognostic factors [10] and Huang et al. have discovered SNPs that may be prognostic for overall survival [11]. Neither of these studies differentiated between lung tumour histological subtypes in their analysis.

While the vast majority of copy number studies have examined cancer genomes in a *locus* by *locus* manner, one study by Hicks et al. [12] correlated survival in patients with breast cancer not to individual *loci* but to a pan-genomic index that measured the type and extent of genomic damage.

* Corresponding author at: Leeds Institute of Molecular Medicine, Section of Experimental Therapeutics, Wellcome Trust Brenner Building, St James's University Hospital, Leeds, LS9 7TF, UK. Fax: +44 113 3438601.
E-mail address: h.m.wood@leeds.ac.uk (H.M. Wood).
[1] These Authors contributed equally to this work.

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.10.006

We recently showed how massively parallel sequencing at very low genomic coverage can be used to study copy number from small amounts of DNA extracted from formalin-fixed paraffin-embedded (FFPE) blocks [13]. In the study reported here, we focused exclusively on early stage lung SCC to investigate the relationship between survival and chromosomal regions of gain and loss. Our objective was to determine whether patients with good or bad outcomes had distinctive genomic characteristics. In addition to conventional correlation of significant *loci* with disease outcome, we adopted a whole genome approach that negates the need to 'call' regions of gain and loss, which can be difficult in tumours of unknown ploidy, tumour cell content and clonal architecture. To this end, we devised a new computational algorithm that relates to total genomic damage, specifically the relative ratios of copy number levels across the genome. We also investigated whether this algorithm could generate a score with potential prognostic value for patients with early stage lung SCC.

2. Materials and methods

2.1. Patients and samples

Eighty-nine patients with early stage SCC of the lung who underwent surgery at the Department of Thoracic Surgery in Leeds, UK, between 1994 and 2003 were included in this study. Eligibility criteria were stage I–IIIA lung SCC, adequate surgery with curative intent, microscopically negative (R0) resection margins, survival ≥4 weeks after surgery, no pre-operative radiotherapy or chemotherapy, no other cancer in the previous 5 years. Clinical and outcome data for these patients were retrieved from the original patient records and the Yorkshire Cancer Registry. Patient demographic and clinical characteristics are summarised in Table 1. Formalin-fixed paraffin-embedded (FFPE) blocks of these 89 SCC were obtained in a pseudo-anonymised form from the Pathology archive of Leeds Teaching Hospitals. Staging work-up procedures for patients included chest and abdomen computed tomography (CT) scan. A database with relevant demographic, clinical and outcome information including donor's age at diagnosis, gender, histological diagnosis, stage of disease, type of surgical procedure, peri-operative chemotherapy and/or radiotherapy, was available. Ethics approval was obtained from the Leeds (East) Ethics Committee and from the Leeds Teaching Hospitals NHS Trust Research and Development Department. A further 29 samples were collected from eligible patients but did not yield DNA of sufficient quality to prepare DNA sequencing libraries.

Tumour size and nodal status were obtained from the original pathology reports; nodal stations were classified according to Naruke's map [14]. The tumours were staged according to the International Union Against Cancers tumour-node-metastasis (TNM) classification [15]. Histological subtype and grade were defined according to the World Health Organisation classification [16].

Table 1
Demographical and clinical characteristics of patients (n = 89).

Parameter	n	%
Age at surgery, years		
Median	68.2	
Range	39.2–84.5	
Gender		
Male	63	70.8
Female	26	29.2
Stage		
I	44	49.4
II	35	39.3
IIIA	10	11.2
Grade		
G1	2	2.2
G2	46	51.7
G3	37	41.6
Gx	4	4.5
Type of surgery		
Lobectomy/bilobectomy	65	73.0
Pneumonectomy	24	27.0
Post-operative radiotherapy		
Yes	27	30.3
No	62	69.7
Potential follow-up time (months)		
Median	103.2	
Range	50.0–170.1	
Survival (months)		
Median	28.7	
Range	1.1–152.2	

2.2. DNA preparation

Tumour genomic DNA was prepared from macrodissected FFPE tissue using a commercially available kit. Briefly, 4 μm-thick sections were cut from each FFPE tumour tissue block and stained with haematoxylin and eosin (H&E); the most representative tumour areas in each slide were marked using a fine-tipped permanent marker. An independent pathologist, blind to the patient identity and diagnosis, reviewed all the marked H&E slides in order to (i) confirm the diagnosis and histology reported in the original pathology report; (ii) evaluate the percentage of tumour cells in the marked area, corresponding to the macrodissected tissue used for DNA extraction. All histological diagnoses were confirmed, the average tumour cell content in macrodissected tissue was 74%. Seven further consecutive 10 μm-thick sections were cut from each block, heated on a hot plate at 65 °C for 3 min, and then rehydrated by immersion in xylene for 15 min, 100% ethanol for 3 min, 90% ethanol for 3 min, 70% ethanol for 3 min and ddH_2O. Sections were immediately macrodissected using sterile disposable scalpels to harvest the tumour tissue; the corresponding H&E stained, marked slide was used as a guide. All the macrodissected tissue from each case was placed in a separate sterile centrifuge tube containing 180 μl of Buffer ATL (Qiagen Hilden, Germany) and labelled with the unique patient study ID. DNA extraction was performed using the QIAamp DNA Mini Kit according to the manufacturer instructions (Qiagen).

The quality and quantity of genomic DNA was determined by UV spectroscopy using the ND 8000 Nanodrop spectrophotometer (NanoDrop Technologies, Thermo Fisher Scientific, Wilmington, DE) and the Quant-iT dsDNA BR Assay Kit (Invitrogen, Life Technologies Corporation, Carlsbad, CA).

2.3. Next-generation sequencing for copy number analysis

DNA libraries were prepared and sequenced using methods previously described [13]. Libraries were prepared for sequencing with a unique 6 bp adapter ligated to enable multiplexing. Twenty samples were pooled per lane on an Illumina GAII sequencer for 76 cycles of single end sequencing resulting in 70 bp of genomic sequence and 6 bp of adapter. Files were split according to adapter sequence and the remaining 70 bp aligned to the human genome (USCS hg19) using the Burrows–Wheeler Alignment tool (BWA) [17]. Only reads with the highest BWA mapping score of 37 were used.

Copy number was calculated by splitting the genome into windows averaging 300 tumour reads per window. A normal control sample was constructed from a pool of 20 normal British individuals downloaded from the 1000 genomes project [18]. The ratio for number of tumour and normal reads in each window was calculated, once read numbers had been adjusted for local GC content; breakpoints were called using the Circular Binary Segmentation algorithm available in the Bioconductor package DNAcopy [19]. Initially the read numbers were normalised so that the median number of reads per window was considered the normal.

Association of tumour copy number changes in specified genomic regions with tumour grade, stage and patient outcome was examined in two ways. First, gains and losses were called as those passing a log2 ratio threshold of ±0.25 and regions with a high proportion of gain and loss were visually inspected. A subsequent, statistical analysis was then undertaken using the Bioconductor package KC-SMARTR [20,21] which can detect significantly altered regions and compare two groups of samples.

To quantify and compare global patterns of copy number change we developed three novel mathematical measures, namely G-stat, H-stat and the combined GH index, as follows. First, the copy number data were smoothed using the smoothseg package [22]; a density plot was produced from the smoothed output and local peaks identified. The G-stat and H-stat were then calculated from the relative heights and positions of the peaks. Specifically, the 'G-stat' was calculated as being the proportion of the smoothed genome that was at a lower copy number state than the biggest peak. The 'H-stat' was calculated as the relative heights of the two biggest peaks, the second biggest divided by the biggest, as follows:

We divided both the test and normal genome in N non-overlapping windows of equal size. For each window i = (1, …, N) we calculated the ratio

$$r_i = \frac{t_i}{c_i}$$

where t_i and c_i are, respectively, the number of reads from test and control mapping to the window *i*. Using functions from the package CNAnorm, [22] we smoothed the signal and obtained N smoothed values s_i. Next we calculated the density function of s

$$d = \text{density}(\text{s}).$$

We labelled the local maxima (peaks) of the density function d(s) with M pairs of the kind

$$(q, p) = (s_i, d(s_i)).$$

Such peaks were defined by the condition

$$d(s_i) > d(s_{i-1}) \wedge d(s_i) > d(s_{i+1}).$$

We sorted such pairs from the maximum peak, corresponding to $s = q_1$, to the lowest one with $s = q_M$:

$$p_k = d(q_k) > p_{k+1} = d(q_{k+1}), \forall k \in \{1, \ldots, M-1\}.$$

Finally, we calculated the combined GH index = G × (1 − H) where

$$\text{G} = \left\{ \frac{S_i | S_i < q_1\}}{\{S_i\}} \right.$$

where #A is the number of elements of the set A, and

$$\text{H} = \frac{P_2}{P_1}.$$

GH value can be calculated from low coverage whole genome sequence data by downloading the CNAnorm package from and http://www.precancer.leeds.ac.uk/cnanorm or Bioconductor running it using code available at http://www.precancer.leeds.ac.uk/gh-index. This package is designed for sequence data but could be adapted for use with aCGH data. Alternatively, other packages that produce a density plot of smoothed data [23] could be adapted.

Additionally, we calculated the F-stat according to the method described by Hicks [12], which was developed to measure total genomic damage in breast cancer.

2.4. Survival analysis

Survival data were available for all 89 patients. Median potential follow-up was 103 months (range, 50 to 170 months).

Univariate survival analyses were performed by Kaplan–Meier estimates [24], and survival curves were compared using a log-rank test [25]. Patients still alive at the time of the final analysis were censored at the date of last contact. Overall survival was defined as the time interval between the date of surgery and death from any cause. We considered p values of <0.05 to be statistically significant. To assess whether the associations observed in the univariate analysis persisted after simultaneous adjustment, we performed a multivariable analysis using several models built according to standard methods [26]. We considered the following clinical-pathological variables: age at time of surgery (<65 vs. ≥65 years), gender, pTNM stage, tumour grading, surgical procedure (lobectomy vs. pneumonectomy), post-operative radiotherapy (yes vs. no vs. unknown). Genomic variables were G-stat, H-stat and the combined GH score. The software used for these analyses was Statistical Analysis Software (SAS, Cary, NC).

2.5. Unsupervised hierarchical clustering analysis

Unsupervised hierarchical clustering was performed by first producing a list of genomic regions with no breakpoints in any of the samples and then measuring the copy number ratios for all samples for all of these regions. These ratios were then analysed using the 'heatmap' function in R with row dendograms suppressed but otherwise default settings to produce heatmaps and dendograms to look for co-occurring patterns of gain and loss and clusters of similar samples.

3. Results

3.1. Sequence data and karyograms

DNA sequence was obtained from 89 early stage lung SCC cases. Sequences have been stored at the European Nucleotide Archive accession number ERP000834. Mean read number was 1,030,660 per sample, ranging from 200,000 to 3,000,000. Using 300 reads per window for copy number analysis, this equates to a resolution of approximately 900 kb. The number of breakpoints per sample ranged from 4 to 205.

Karyograms showing regions of gain and loss along the whole genome were generated for each case. Karyograms exhibited several different types of copy number patterns, in terms of both the proportion of the genomes involved and the complexity of the damage. This ranged from whole chromosome gain and loss to very small but highly amplified regions. The analysis of copy number aberrations across the entire data set revealed several features previously seen in lung SCC [27].

3.2. Relationship between frequently gained/lost regions and patient outcome

Two complementary methods were used to look for association between genomic regions of gain or loss and patient outcome, namely survival. Firstly, Kaplan–Meier analyses [24] were made of survival in cases incorporating all regions that were identified using the KC-SMARTR algorithm [20,21] as showing gain or loss in a significant number of samples (Table 2). No individual region was significantly associated with survival, although gains in 20p did show a non-significant trend towards better outcome (median survival 42 months vs 22 months; HR 0.6292, 95% CI 0.3872–1.022, uncorrected p-value = 0.0614). Secondly, the patients were split into two

Table 2
Association of commonly gained and lost regions with survival.

Chromosome arm	Position Mb	Copy number change gain or loss	n (%)	Median survival (months)		p
				With feature	Without feature	
3p	0–89	Loss	51 (57)	28	25	0.224
3q	93.6–197.7	Gain	84 (94)	27	13	0.180
4p	0–47	Loss	24 (27)	39	24	0.586
5p	0–46	Gain	60 (67)	27	24	0.617
5q	50.3–180.6	Loss	27 (30)	24	28	0.601
7p	0–49	Gain	50 (56)	23	28	0.677
7q	61.7–117	Gain	47 (53)	32	25	0.605
8p	0–32.8	Loss	30 (34)	30	25	0.291
8q	46.8–147.1	Gain	50 (56)	24	35	0.542
9p	0–44.1	Loss	27 (30)	23	28	0.948
12p	0–34.6	Gain	43 (48)	23	28	0.878
13q	19.1–114.9	Loss	32 (36)	27	25	0.506
17q	21.1–78.9	Gain	39 (44)	25	26	0.758
19q	27.9–28.9	Gain	53 (60)	32	19	0.221
20p	0–26	Gain	36 (40)	42	22	0.068
20q	29.4–39.1	Gain	39 (44)	26	25	0.944
22q	16.9–46.6	Gain	43 (48)	35	22	0.157

or more groups by recognised clinical and pathological features (survival time, tumour stage, tumour grade, etc.) to look for gained/lost regions associated with those features. Visual and statistical comparison between patient groups showed no regions significantly associated with any clinical features. Indeed, irrespective of which clinical parameter was used to subdivide the patients, the frequencies of gains and losses across the genome appeared almost indistinguishable between subsets (results not shown). This was corroborated using the KC-SMARTR algorithm, which showed that no regions were significantly different for any comparison made.

An alternative approach used unsupervised clustering analysis to produce heatmaps and dendograms. Again, this method did not separate the patients into any distinguishable subgroups. By contrast, and to indicate the efficacy of these analytical procedures, we compared the lung SCC samples presented here to an outgroup of head and neck SCC karyograms we had created for a separate study. Both the *locus* by *locus* approach of KC-SMARTR and the clustering analysis were able to separate the lung SCC samples from the head and neck samples (Supplementary Fig. 1), showing that the methods used were able to distinguish between different groups of tumours, but that this set of lung SCC samples did not contain identifiable subgroups.

3.3. Development of an algorithm to define relative chromosomal gain and loss

In view of our inability to identify copy number changes at individual genomic *loci* associated with clinical outcome, we decided to apply the approach of Hicks et al., and look for global patterns of copy number variation [12]. This approach classified largely diploid breast tumour genomes by patterns of damage named 'simplex' (few aberrations, mostly involving whole chromosome arms), 'sawtooth' (many aberrations spread throughout the genome) and 'firestorm' (like simplex, with local regions of complex damage), and generated an algorithm for calculating an index of genomic damage, named F-stat, which was associated with survival. The tumours from our series did not easily fit into the Hicks method of classification, mostly being in a continuous spectrum of genomic damage somewhere between the simplex and sawtooth patterns and with almost none being definitely in one group or the other. The F-stat was, however, calculated for each sample; there was no correlation with survival or any other clinical parameter (results not shown).

While investigating the gains and losses of individual regions, as well as the F-stat, it became apparent that the traditional approach of regarding the median copy number ratio as 'normal' may not always be the most appropriate. This approach by definition assumes that each has precisely the same amount of gain and loss. In addition, for several of our samples, there was relatively little of the genome near the median value; almost none of the genome could, therefore, be considered 'normal'. The alternative of looking for absolute copy number was not feasible either, since tumour cell population heterogeneity, contamination with normal cells and extensive aneuploidy meant that for many samples it was impossible to tell which regions were diploid or 'normal' and to accurately quantify any gains and losses. We decided instead to consider the total spread of copy number states throughout the entire population of cells in each tumour. Density plots were drawn for the copy number distributions of each

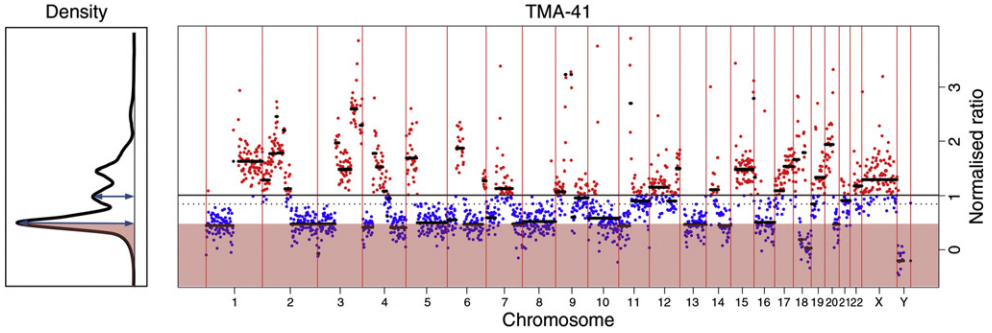

Fig. 1. How the G and H stats are calculated. For each genome, a density plot is drawn and peaks are identified. To calculate the G-stat, the proportion of the genome with copy number less than the highest peak is measured. In this example it corresponds to the number of points in the red box as fraction of the total. To measure the H-stat, the ratio of the heights of the two highest peaks (second over first) is calculated. For TMA-41 G = 0.30, H = 0.35. GH = G(1 − H) = 0.195.

sample, with the relative heights of each peak representing the proportion of the genome at that copy number state (Fig. 1). The peak nearest to the median was noted. For 55 of the 89 samples, the largest peak was the one nearest to the median, so the traditional method of treating the median as 'normal' would be a good approximation. For the remaining 34 samples however, the highest peak was not the one nearest the median, but was at a lower copy number state. We postulated that for these samples, the highest peak should be considered the normal, and observed that these genomes were characterised by considerable regions of gain, but very little loss.

We decided to improve on this simple classification because issues such as tumour heterogeneity or a relatively minor extra aberration might suddenly change which peak was nearest to the median and reclassify the tumour. The factors affecting the classification were the amount of genomic loss relative to each peak and the relative height of the peaks. To better quantify and compare global patterns of copy number change, we developed three novel mathematical measures, namely G-stat, H-stat and the combined GH index. The G-stat is a measure of genomic loss. It assumes that the highest peak represents the normal and is a simple calculation of the proportion of the genome which is less than this value; genomes with large amounts of loss will have a high G-stat. The H-stat is a measure of homogeneity and complexity of genome damage; it is calculated as the relative heights of the two highest peaks; a genome with one major peak and a number of minor peaks will have a low H-stat, while one with two or more peaks of equal height will have a high H-stat. The G and H stats both influenced the previous classification of biggest and median peaks, but were completely independent of each other (Pearson correlation of −0.047). The GH index was then calculated as $G \times (1 - H)$. An example of the derivation of the GH index for an individual SCC case is shown in Fig. 1. For all comparisons, the samples were split into two groups according to whether they were above or below the median value for the index being measured.

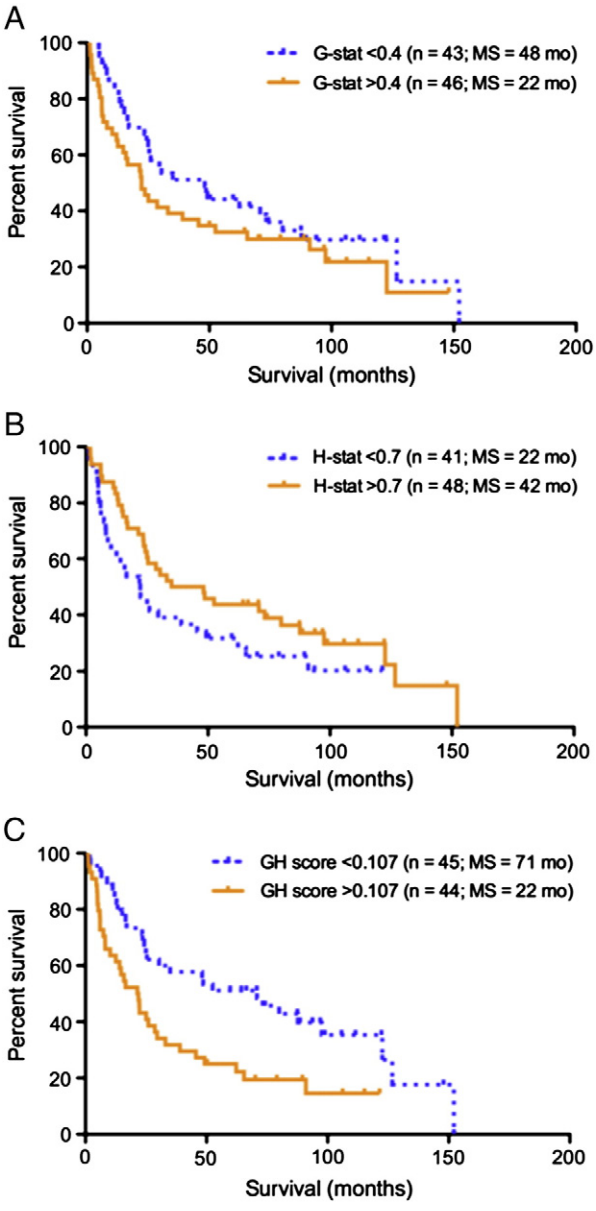

Fig. 2. Kaplan–Meier curves showing the association with survival of G-stat, H-stat and combined GH-stat. (MS = median survival).

3.4. Survival analysis

Overall survival rates at 1, 3 and 5 years were 75%, 44%, and 35%, respectively; median survival was 25.4 months (67 events observed).

In univariate analysis, a low G-stat did not significantly associate with better survival ($p = 0.18$, Fig. 2A) while a high H-stat showed a trend towards better survival ($p = 0.09$, Fig. 2B). Patients with a low combined GH index had significantly better survival than those with a higher value ($p = 0.003$, Fig. 2C). Neither G-stat nor H-stat showed a significant correlation with stage or grade, showing that they were separate, independent measures. Among the other variables, age<65 years and lobectomy were associated with better prognosis ($p = 0.04$ and $p = 0.02$). Univariate analysis results are summarised in Table 3.

Multivariate analysis confirmed GH score as an independent prognostic indicator (hazard ratio (HR) = 0.56, 95% CI 0.315–0.980, $p = 0.04$). Age<65 years (HR = 0.18, 95% CI 0.07–0.47, $p = 0.0004$) and stage I (HR = 0.29, 95% CI 0.11–0.76, $p = 0.01$) were significantly associated with better outcome. Multivariate analysis results are shown in Table 4.

4. Discussion

In this study, we have looked at copy number changes in a cohort of early stage, homogeneously treated, lung SCC. To provide an

Table 3
Univariate analysis.

Variable	HR	95% CI	p
Age (<65 vs. ≥65 years)	0.58	0.34–0.99	0.0438
Gender (male vs. female)	1.25	0.73–2.16	0.4192
Stage			
Stage I vs Stage III	0.46	0.21–1.00	0.0937
Stage II vs. Stage III	0.68	0.31–1.50	
Tumour grade (G1–G2 vs. G3–G4)	0.87	0.52–1.43	0.5753
Type of surgery (lobectomy vs. pneumonectomy)	0.54	0.32–0.91	0.0195
Post-op RT (yes vs. no)	0.97	0.57–1.66	0.9217
G-stat (<0.4 vs. ≥0.4)	0.72	0.44–1.17	0.1807
H-stat (<0.7 vs. ≥0.7)	1.53	0.93–2.51	0.0932
Combined GH-stat (<0.107 vs. ≥0.107)	0.47	0.29–0.78	0.0029

Table 4
Multivariate analysis.

Variable	HR	95% CI	p
GH score (<0.107)	0.56	0.32–0.98	0.0424
Age (<65)	0.18	0.07–0.47	0.0004
Stage			
Stage II	0.43	0.21–1.00	0.0706
Stage I	0.29	0.11–0.76	0.0112
Type of surgery (lobectomy)	0.70	0.38–1.30	0.2597
Tumour grade (G1–G2)	0.61	0.34–1.11	0.1079

evaluation of overall genomic damage in individual tumours, we have devised a novel computational index, the GH index, which is based on whole-genome copy number analysis. In an exploratory analysis we have shown the potential prognostic value of the GH index in lung SCC. Specifically, in our series, patients with a low GH score survived significantly longer (median survival 71 months vs. 22 months, $p = 0.003$); importantly, the GH score was confirmed as an independent prognostic factor in multivariate analysis ($p = 0.042$). To our knowledge, this is the first report of a pan-genomic damage index that correlates with survival in lung cancer.

Better stratification of patients with early stage NSCLC is required to optimise treatment and thereby improve outcomes. To date, disease stage remains the best prognostic indicator; recent progress in proteomics and genomics however holds the promise for novel molecular markers with prognostic value, complementary to and even more powerful than disease stage.

Previous studies in NSCLC [4–11,28], have sought to link alterations of particular genomic regions to clinical features such as histology, tumour grade, disease stage, response to treatment, or outcome. Despite some of these studies being apparently successful, major limitations include small sample size, heterogeneity of the study population and the lack of validated and universally accepted criteria for calling loss and gain. Importantly, these limitations also make any meta-analysis impossible. Indeed, it may be challenging to acquire sufficient numbers of clinically similar samples to make significant observations. We have sought to overcome this by identifying a study population as clinically homogenous as possible, i.e. patients with early stage lung SCC radically treated with surgery at a single centre. Despite this, it is possible that the samples might actually represent more than one currently unrecognised subtype and that this is the reason that in our series no single region of copy number change was found to associate with any clinical feature, including survival.

Even so, it is possible that no single region is associated with any individual clinical feature, no matter how many patients might be evaluated. As such, looking at the pan-genomic patterns of alteration offers an alternative approach that may be valuable for some tumour types. A pan-genomic approach has recently identified patterns of damage associated with survival in breast tumours [12]. The Hicks et al. 'F-stat' did not, however, associate with survival in our series, perhaps because the F-stat might be a tumour-type specific score. Similarly, the GH score we have developed might only be a prognostic indicator in lung SCC.

There are several major, but often overlooked, problems with more traditional approaches of copy number analysis in cancer research. First, the definition of "normal" copy number level, and by implication gain and loss, is not ideal. Conventionally, normal is defined at, or nearby the median copy number. This is an acceptable way of looking for germline alterations, where only a few portions of the genome deviate from an otherwise diploid normal genome, or looking for somatic copy number changes in tumour types with limited copy number variations. This definition of normality may not, however, be suitable when complex tumours are studied. Using the median as normal assumes the genome as having exactly equal proportions of gain and loss, even if that is not the case. The second problem relates to inter- and intra-sample tumour heterogeneity, as not only the percentage of tumour cells varies between tumour samples, but also different tumour cell subclones might be present within the individual tumour specimen. If a sample has 50% tumour content, then a gain of one copy will only be half the amplitude of a similar gain in a 100% tumour sample. The result would be the same if the tumour is heterogeneous and only half the cells exhibit a gain. These problems make the interpretation of copy number changes in complex tumours quite challenging. To draw a diagram showing frequency of gain and loss across many samples requires a threshold to be decided. If this threshold is too strict, then regions where only some of the cells are altered may be missed. If the threshold is too lenient, then noise from the data may be mis-called as gain or loss. It is also difficult to distinguish between regions where some of the cells have alterations and regions where every cell is altered. Similarly, a gain of two copies in all cells in a tetraploid tumour will look identical to a gain of one copy in all cells in a diploid tumour and a gain of three copies in half the cells of a triploid tumour. We are not proposing a solution to these issues here. There have been some bioinformatic attempts to remedy these problems [23,29,30] and most of the recent advances have been reviewed [31], but these techniques are not always used when clinical datasets are presented, making meta-analysis of different studies difficult.

This is an exploratory study on a small, but relatively homogeneous population. Exactly what the G and H values represent, and how they correlate with survival, needs further study. G is a measure of genomic loss, and more loss appears to indicate worse prognosis. It is possible that genomic loss exposes more tumour suppressor genes to haplo-insufficiency, as recently described [32]. Alternatively loss might be a less disruptive force than gain, so the cells are more normal, so more capable of surviving in situ and less visible to diagnostic screening. H is a measure of homogeneity and complexity of genomic damage. Homogenous tumours appear to have better prognosis than more complex heterogeneous ones. This could be because a heterogeneous tumour is one which has greater clonal variety and is thus more able to withstand changes in environment caused by systemic treatment or relocation to a distant metastatic site.

We believe that the pan-genomic way of thinking of copy number changes we have adopted in this study with the development of the GH score is potentially of great value, especially in tumour types where looking at individual gains and losses has not to date been associated with any specific clinical subgroup. Not being restricted by some arbitrary definition of what is 'normal', this approach allows the freedom to consider issues other than simple gain and loss, and address pan-genomic patterns of alteration. Our method of calculating a GH value (or other similar methods) may prove useful in many different tumour types, and as such we present these algorithms for validation in similar data sets. We are aware that genomic regions of gain and loss can be important, so we would propose that the two methods be used in a complementary fashion.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2011.10.006.

Acknowledgments

We are grateful to Professor Phil Quirke for providing access to the lung cancer series samples and clinical data.

This work was supported by Yorkshire Cancer Research (grant L341PG to PHR), by a Marie Curie Intra-European Fellowship from the European Commission within the 7th Framework Programme (to OB), by the Leeds Teaching Hospitals Charitable Foundation and the Betty Woolsey Bequest for Thoracic Research.

References

[1] A.F. Gazdar, L. Girard, W.W. Lockwood, W.L. Lam, J.D. Minna, Lung cancer cell lines as tools for biomedical discovery and research, J. Natl. Cancer Inst. 102 (2010) 1310–1321.
[2] F. Mitelman, Catalogue of chromosome aberrations in cancer, Cytogenet. Cell Genet. 36 (1983) 1–515.
[3] A.F. Gazdar, Should we continue to use the term non-small-cell lung cancer? Ann. Oncol. 21 (2010) vii225-vii229.
[4] W.W. Lockwood, R. Chari, B.P. Coe, K.L. Thu, C. Garnis, C.A. Malloff, J. Campbell, A.C. Williams, D. Hwang, C.Q. Zhu, T.P. Buys, J. Yee, J.C. English, C. Macaulay, M.S. Tsao, A.F. Gazdar, J.D. Minna, S. Lam, W.L. Lam, Integrative genomic analyses identify BRF2 as a novel lineage-specific oncogene in lung squamous cell carcinoma, PLoS Med. 7 (2010) e1000315.
[5] J. Weiss, M.L. Sos, D. Seidel, M. Peifer, T. Zander, J.M. Heuckmann, R.T. Ullrich, R. Menon, S. Maier, A. Soltermann, H. Moch, P. Wagener, F. Fischer, S. Heynck, M. Koker, J. Schottle, F. Leenders, F. Gabler, I. Dabow, S. Querings, L.C. Heukamp, H. Balke-Want, S. Ansen, D. Rauh, I. Baessmann, J. Altmuller, Z. Wainer, M. Conron, G. Wright, P. Russell, B. Solomon, E. Brambilla, C. Brambilla, P. Lorimier, S. Sollberg, O.T. Brustugun, W. Engel-Riedel, C. Ludwig, I. Petersen, J. Sanger, J. Clement, H.

Groen, W. Timens, H. Sietsma, E. Thunnissen, E. Smit, D. Heideman, F. Cappuzzo, C. Ligorio, S. Damiani, M. Hallek, R. Beroukhim, W. Pao, B. Klebl, M. Baumann, R. Buettner, K. Ernestus, E. Stoelben, J. Wolf, P. Nurnberg, S. Perner, R.K. Thomas, Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer, Sci. Transl. Med. 2 (2010) 62ra93.

[6] L.E. Coate, T. John, M.S. Tsao, F.A. Shepherd, Molecular predictive and prognostic markers in non-small-cell lung cancer, Lancet Oncol. 10 (2009) 1001–1010.

[7] Y. Lu, W. Lemon, P.Y. Liu, Y. Yi, C. Morrison, P. Yang, Z. Sun, J. Szoke, W.L. Gerald, M. Watson, R. Govindan, M. You, A gene expression signature predicts survival of patients with stage I non-small cell lung cancer, PLoS Med. 3 (2006) e467.

[8] H.Y. Chen, S.L. Yu, C.H. Chen, G.C. Chang, C.Y. Chen, A. Yuan, C.L. Cheng, C.H. Wang, H.J. Terng, S.F. Kao, W.K. Chan, H.N. Li, C.C. Liu, S. Singh, W.J. Chen, J.J. Chen, P.C. Yang, A five-gene signature and clinical outcome in non-small-cell lung cancer, N. Eng. J. Med. 356 (2007) 11–20.

[9] Z. Sun, D.A. Wigle, P. Yang, Non-overlapping and non-cell-type-specific gene expression signatures predict lung cancer survival, J. Clin. Oncol. : Off J. Am. Soc. Clin. Oncol. 26 (2008) 877–883.

[10] T.M. Kim, S.H. Yim, J.S. Lee, M.S. Kwon, J.W. Ryu, H.M. Kang, H. Fiegler, N.P. Carter, Y.J. Chung, Genome-wide screening of genomic alterations and their clinicopathologic implications in non-small cell lung cancers, Clin. Cancer Res. Off J Am Assoc. Cancer Res. 11 (2005) 8235–8242.

[11] Y.T. Huang, R.S. Heist, L.R. Chirieac, X. Lin, V. Skaug, S. Zienolddiny, A. Haugen, M.C. Wu, Z. Wang, L. Su, K. Asomaning, D.C. Christiani, Genome-wide analysis of survival in early-stage non-small-cell lung cancer, J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 27 (2009) 2660–2667.

[12] J. Hicks, A. Krasnitz, B. Lakshmi, N.E. Navin, M. Riggs, E. Leibu, D. Esposito, J. Alexander, J. Troge, V. Grubor, S. Yoon, M. Wigler, K. Ye, A.L. Borresen-Dale, B. Naume, E. Schlicting, L. Norton, T. Hagerstrom, L. Skoog, G. Auer, S. Maner, P. Lundin, A. Zetterberg, Novel patterns of genome rearrangement and their association with survival in breast cancer, Genome Res. 16 (2006) 1465–1479.

[13] H.M. Wood, O. Belvedere, C. Conway, C. Daly, R. Chalkley, M. Bickerdike, C. McKinley, P. Egan, L. Ross, B. Hayward, J. Morgan, L. Davidson, K. MacLennan, T.K. Ong, K. Papagiannopoulos, I. Cook, D.J. Adams, G.R. Taylor, P. Rabbitts, Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens, Nucleic Acids Res. 38 (2010).

[14] T. Naruke, K. Suemasu, S. Ishikawa, Lymph node mapping and curability at various levels of metastasis in resected lung cancer, J. Thorac. Cardiovasc. Surg. 76 (1978) 832–839.

[15] L.H. Sobin, I.D. Fleming, TNM classification of malignant tumors, fifth edition (1997). Union Internationale Contre le Cancer and the American Joint Committee on Cancer, Cancer 80 (1997) 1803–1804.

[16] W.D. Travis, L.H. Sobin, Histological Typing of Lung and Pleural Tumours, Springer-Verlag, Berlin; Paris, 1999.

[17] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (2009) 60.

[18] R.M. Durbin, G.R. Abecasis, D.L. Altshuler, A. Auton, L.D. Brooks, R.A. Gibbs, M.E. Hurles, G.A. McVean, A map of human genome variation from population-scale sequencing, Nature 467 (2010) 1061–1073.

[19] E.S. Venkatraman, A.B. Olshen, A faster circular binary segmentation algorithm for the analysis of array CGH data, Bioinformatics 23 (2007) 657–663.

[20] C. Klijn, H. Holstege, E. Schut, M. Reinders, J. de Ridder, J. Jonkers, L. Wessels, Candidate cancer gene discovery using KC-smart: a novel method for statistical multi-experiment ACGH data analysis, Cell. Oncol. 29 (2007) 121.

[21] J.J. de Ronde, C. Klijn, A. Velds, H. Holstege, M.J. Reinders, J. Jonkers, L.F. Wessels, KC-SMARTR: an R package for detection of statistically significant aberrations in multi-experiment aCGH data, BMC Res notes 3 (2010) 298.

[22] A. Gusnanto, H.M. Wood, Y. Pawitan, P. Rabbitts, S. Berri, Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next generation sequence data, Bioinformatics (in press), doi: 10.1093/bioinformatics/btr593.

[23] H.I. Chen, F.H. Hsu, Y. Jiang, M.H. Tsai, P.C. Yang, P.S. Meltzer, E.Y. Chuang, Y. Chen, A probe-density-based analysis method for array CGH data: simulation, normalization and centralization, Bioinformatics 24 (2008) 1749–1756.

[24] E.L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, J. Am. Stat. Assoc. 53 (1958) 457–481.

[25] R.G. Miller, G. Gong, A. MuÒoz, Survival Analysis, Wiley, New York, 1981.

[26] S. Selvin, Statistical Analysis of Epidemiologic Data, Oxford University Press, New York; Oxford, 2004.

[27] B.R. Balsara, G. Sonoda, S. du Manoir, J.M. Siegfried, E. Gabrielson, J.R. Testa, Comparative genomic hybridization analysis detects frequent, often high-level, overrepresentation of DNA sequences at 3q, 5p, 7p, and 8q in human non-small cell lung carcinomas, Cancer Res. 57 (1997) 2116–2120.

[28] P.S. Hammerman, M.L. Sos, A.H. Ramos, C. Xu, A. Dutt, W. Zhou, L.E. Brace, B.A. Woods, W. Lin, J. Zhang, X. Deng, S.M. Lim, S. Heynck, M. Peifer, J.R. Simard, M.S. Lawrence, R.C. Onofrio, H.B. Salvesen, D. Seidel, T. Zander, J.M. Heuckmann, A. Soltermann, H. Moch, M. Koker, F. Leenders, F. Gabler, S. Querings, S. Ansén, E. Brambilla, C. Brambilla, P. Lorimier, O.T. Brustugun, Ö. Helland, I. Petersen, J.H. Clement, H. Groen, W. Timens, H. Sietsma, E. Stoelben, J.r. Wolf, D.G. Beer, M.S. Tsao, M. Hanna, C. Hatton, M.J. Eck, P.A. Janne, B.E. Johnson, W. Winckler, H. Greulich, A.J. Bass, J. Cho, D. Rauh, N.S. Gray, K.-K. Wong, E.B. Haura, R.K. Thomas, M. Meyerson, Mutations in the DDR2 kinase gene identify a novel therapeutic target in squamous cell lung cancer, Cancer Discov. 1(1) (2011), doi:10.1158/2159-8274.CD-11-0005.

[29] M.A. van de Wiel, K.I. Kim, S.J. Vosse, W.N. van Wieringen, S.M. Wilting, B. Ylstra, CGHcall: calling aberrations for array CGH tumor profiles, Bioinformatics 23 (2007) 4.

[30] B.P. van Houte, T.W. Binsl, H. Hettling, W. Pirovano, J. Heringa, CGHnormaliter: an iterative strategy to enhance normalization of array CGH data with imbalanced aberrations, BMC Genomics 10 (2009) 401.

[31] M.A. van de Wiel, F. Picard, W.N. van Wieringen, B. Ylstra, Preprocessing and downstream analysis of microarray DNA copy number profiles, Brief. Bioinform. 12 (2011) 10–21.

[32] A.H. Berger, P.P. Pandolfi, Haplo-insufficiency: a driving force in cancer, J. Pathol. 223 (2011) 138–147.