Prediction of Clinical Outcome in Multiple Lung Cancer Cohorts by Integrative Genomics: Implications for Chemotherapy Selection

Philippe Broët,1,2 Sophie Camilleri-Broët,1,2,3 Shenli Zhang,4 Marco Alifano,1,2 Dhinoth Bangarusamy,5 Maxime Battistella,1 Yonghui Wu,6 Marianne Tuefferd,1 Jean-François Régnard,2,3 Elaine Lim,4 Patrick Tan,5,6 and Lance D. Miller7

1JE2492, Faculty of Medicine Paris-Sud, Bicêtre, France; 2Assistance Publique Hôpitaux de Paris; 3Faculty of Medicine Paris Descartes, Paris, France; 4Yong Loo Lin School of Medicine, National University of Singapore; 5Genome Institute of Singapore; 6Duke-NUS Graduate Medical School, Singapore, Republic of Singapore; and 7Department of Cancer Biology, Wake Forest University School of Medicine, Winston-Salem, North Carolina

Abstract

The role of adjuvant chemotherapy in patients with stage IB non–small-cell lung cancer (NSCLC) is controversial. Identifying patient subgroups with the greatest risk of relapse and, consequently, most likely to benefit from adjuvant treatment thus remains an important clinical challenge. Here, we hypothesized that recurrent patterns of genomic amplifications and deletions in lung tumors could be integrated with gene expression information to establish a robust predictor of clinical outcome in stage IB NSCLC. Using high-resolution microarrays, we generated tandem DNA copy number and gene expression profiles for 85 stage IB lung adenocarcinomas/large cell carcinomas. We identified specific copy number alterations linked to relapse-free survival and selected genes within these regions exhibiting copy number–driven expression to construct a novel integrated signature (IS) capable of predicting clinical outcome in this series (P = 0.02). Importantly, the IS also significantly predicted clinical outcome in two other independent stage I NSCLC cohorts (P = 0.003 and P = 0.025), showing its robustness. In contrast, a more conventional molecular predictor based solely on gene expression, while capable of predicting outcome in the initial series, failed to significantly predict outcome in the two independent data sets. Our results suggest that recurrent copy number alterations, when combined with gene expression information, can be successfully used to create robust predictors of clinical outcome in early-stage NSCLC. The utility of the IS in identifying early-stage NSCLC patients as candidates for adjuvant treatment should be further evaluated in a clinical trial. [Cancer Res 2009;69(3):1055–62]

Introduction

Non–small-cell lung carcinoma (NSCLC) is the most common cause of worldwide cancer mortality, with a global 5-year survival rate of 15%. For patients with early-stage disease, the survival rate after surgery is 40% to 55% (1, 2), raising the need to accurately identify subgroups who might benefit from adjuvant chemotherapy (3). The utility of adjuvant chemotherapy for the stage IB tumors, however, remains controversial. Preliminary results of the CALGB 9633 trial suggested a potential survival benefit for adjuvant chemotherapy in stage IB disease, but updated results from the same trial show no benefit in overall survival. Stage IB NSCLC thus represents an excellent opportunity for applying genomic strategies to stratify patients with low and high risks of recurrence, with adjuvant therapy being a treatment option for the high-risk category.

One major feature of NSCLCs is chromosomal instability, which can result in the amplification and deletion of either specific genomic regions or entire chromosomes. Regions exhibiting copy number alterations (CNA) can affect the expression of cis-localized tumor suppressor genes and oncoproteins. However, only a few reports have suggested a potential relationship between recurrent CNAs and NSCLC patient prognosis (4, 5). In addition, the architecture of CNAs is often complex (multiple “subalterations”) and not all genes within a CNA region will necessarily show altered gene expression (“copy number–driven” expression; refs. 6, 7). These observations suggest that a substantial proportion of genes within CNAs may be inconsequential for tumor behavior, and including such genes into a survival model may only add noise and reduce predictive accuracy.

To address these limitations, we developed an integrative strategy combining both genomic CNA and transcriptomic copy number–driven expression. We applied this strategy to a cohort of stage IB lung adenocarcinomas profiled using both high-resolution array-based comparative genomic hybridization (array-CGH) and gene expression platforms. We found that the integrated signature (IS) was an accurate predictor of relapse-free survival in the original cohort and also robustly predicted survival in two other independent cohorts.

Materials and Methods

Patients and tumor samples. A series of 85 consecutive patients who underwent surgery at the Hôtel-Dieu Hospital (AP-HP, France) between August 2000 and February 2004 for stage IB (pT2N0) primary adenocarcinoma or large cell lung carcinoma of peripheral location were included in the study. Patients with bronchioalveolar adenocarcinomas or large cell neuroendocrine carcinomas were excluded from the study, as well as those having received chemotherapy. Pathologic slides were reviewed without any information about the outcome (S.C.-B., M.B.). The clinical and pathologic parameters collected for analysis included age, sex, tobacco exposure, type of resection, laterality, necrosis, size of the tumor, histologic subtype (8), differentiation, vessel invasion, visceral pleura involvement, and TTF1 expression. The quality of frozen tissue was checked by touch preps on microscopic glass slide; only tissue samples with tumor content >50% were selected. This study was approved by institutional ethics committees.
Array-CGH and gene expression microarrays. DNA was extracted from frozen samples using the Nucleon DNA extraction kit (BACC2, Amersham Biosciences) according to the manufacturer's procedures. Briefly, frozen tumor sections were cut into small pieces and digested in proteinase K overnight at 42°C. Deproteinization was carried out in 5 mol/L sodium perchlorate followed by extraction in chloroform/alcohol isomylamylate. After centrifugation, the upper phase was precipitated in cold alcohol 100. DNA pellets were dried and resuspended in Tris-EDTA. For each tumor, 2 μg of tumor and reference genomic DNAs were directly labeled with Cy3-ICTP or Cy5-ICTP, respectively, and hybridized onto CGH microarrays containing 32,000 DOP-PCR amplified bacterial artificial chromosome (BAC) genomic clones providing tiling coverage of the human genome (spotted on two arrays). Hybridizations were done using a MAUI hybridization station, and after washing, the slides were scanned on a GenePix 4000B scanner, as described previously (9).

For total RNA extraction, frozen tumor samples were shattered in liquid nitrogen and homogenized in 1-ml TRIzol (Invitrogen). Extraction was done using a standard chloroform/isopropanol method. RNA quality was assessed on an Agilent Bioanalyzer before storage at −80°C. RNA from 74 of the 85 tumor samples was deemed of sufficient quality to enable reliable gene expression analysis. For measuring gene expression, the Human U133 Plus 2.0 oligonucleotide gene chips (Affymetrix) containing a total of 47,000 transcripts with 61,000 probe sets were used according to the manufacturer's protocol. Briefly, 5 μg of total RNA were used in the amplification reaction, and 20 μg of labeled cRNA were added to the hybridization.

The array-CGH data have been deposited in National Center for Biotechnology Information Gene Expression Omnibus and are accessible through GEO Series accession no. GSE10144.

Preprocessing of the array data. The array-CGH signal intensities were normalized using a two-channel microarray normalization procedure (10) implemented in Genedata Expressionist Pro software. BAC genomic clones mapping to sex chromosomes (X and Y) were not considered for the analysis. Inferences about the gain/loss/modal status of each BAC clone for each sample was obtained using the CGHmix classification procedure (11), which computes the posterior probabilities of a clone belonging to either of three defined genomic states. We assigned each clone to one of two modified copy number states (loss or gain) if its corresponding posterior probability was above a defined threshold value; otherwise the clone was assigned to the modal/unaltered copy state. This latter threshold value was selected to obtain a false discovery rate of 5% for each sample, where false discovery corresponded to a clone incorrectly defined as amplified or deleted. Clones with an absolute fluorescence intensity log ratio of >0.5 and a posterior probability of being a clone considered as an exclusively deleted recurrent CNA.

Defining patterns of CNA. To analyze the propensity of each genomic region (BAC clone) to be deleted or amplified across the series, we modeled the distribution of the number of observed deletions, modal/unmodified, and amplifications for all the genomic regions using a latent class model relying on a finite mixture of multinomial distributions (13). Here, we considered a latent class model with three (low, intermediate, high) levels for both amplification and deletion representing a total of nine (3²) chromosomal patterns. Each of these nine chromosomal patterns describes the joint propensity of a given genomic region for being deleted/unmodified/amplified. From our series, we estimated for each genomic region its posterior probabilities for each of the nine chromosomal patterns using Monte Carlo Markov chain techniques (14), implemented in Winbugs software (15). Then, a classification rule was applied, which assigned each genomic region to the chromosomal pattern to which it had the highest probability of belonging. From the nine chromosomal patterns, the one corresponding to the highest frequency for amplification and lowest for deletion was defined as an “exclusively amplified” recurrent CNA, and vice versa (“exclusively deleted” recurrent CNA).

Statistical analysis to identify copy number–driven genes. To identify copy number–driven genes, each probe set was assigned to the nearest mapped BAC clone. For each probe set, a classic linear regression model was applied where gene expression was the dependent variable and DNA copy number change was the explanatory variable (coded as −1, 0, 1 for loss, modal, and gain, respectively). From the resulting test statistics, we calculated the posterior probability of relationship between genomic and transcriptomic changes using the Gmix procedure (16), a fully Bayesian normal mixture model with an unknown number of components. A probe set was classified as a copy number–driven gene if its posterior probability of relationship between genomic and transcriptomic changes was >0.5, according to the Bayes rule.

Relapse-free survival: assessing prognostic effect of genomic and transcriptomic changes. Relapse-free survival (RFS) time was calculated from the date of the patients' surgery until disease-related death, disease recurrence (either local or distant), or last follow-up examination. To analyze the prognostic effect of either genomic or transcriptomic changes, we computed two sets of univariate score test statistics based on the semiparametric Cox proportional hazards model (17). Here, the null hypothesis corresponded to the absence of a relationship between the instantaneous hazard rate for relapse and either genomic (copy number) status or gene expression measurement. To increase statistical power, we also used information from our analysis of chromosomal patterns. Specifically, for a genomic clone considered as an exclusively amplified recurrent CNA, the few deleted samples for this clone were gathered with those having a modal genomic status. The converse was also done for a clone considered as an exclusively deleted recurrent CNA.

Using the Gmix procedure (16), the posterior probabilities of RFS being related to either the genomic status (genomic-survival posterior probabilities) or gene expression measurements (transcriptomic-survival posterior probabilities) were calculated.

Gene signature building procedure overview. We devised a novel gene selection strategy to construct a copy number–driven gene expression signature, termed integrated signature (IS) in the following text, to predict RFS (Fig. 1). In parallel, we also constructed a conventional transcriptomic signature (TS), with the aim of comparing the performance of the IS to that of a more conventionally derived expression signature not restricted to specific pathologic properties of the cancer. For both signatures, we considered a two-step procedure: (a) In the first step (feature selection), the genomic clones or genes were individually ranked based on either their genomic survival or transcriptomic survival posterior probabilities. For IS (as seen below), we also take into account the relationship between genomic and transcriptomic changes. From these results, gene subset selections were done. (b) In the second step (signature development), a linear combination of the genes belonging to the selected subsets was computed leading to a gene expression signature.

Feature selection. The major difference between the IS and TS feature selection steps is that the former (IS) incorporates genomic information. For the IS, we first selected genomic clones based on their genomic-survival posterior probabilities. Among the genes localized to those high-priority genomic areas, we then restricted our feature selection only to genes exhibiting copy number–driven expression. In the classic way, for the TS we selected the genes based on their transcriptomic survival posterior probabilities. In practice, we selected the clones/gens in a top-down manner, starting with a genomic/transcriptomic survival posterior probability of 99% and decreasing down to 75% with regular spacings (0.05 unit). This operation generated a series of nested gene/clone feature sets of different sizes depending on the chosen posterior probability threshold. This ranking approach is conceptually similar to previous reports (18, 19) but considers posterior probabilities rather than P values.

Signature development. The survival-associated gene expression signatures (IS, TS) were defined as linear combinations of the gene expression measurements of the selected genes weighted by their estimated Cox proportional hazards model regression coefficients (association between gene expression and RFS). More precisely, for feature gene sets (obtained in the feature selection step), the IS and TS signatures for each patient i were calculated as follows:
To quantify the amount by which the consensus weights differ from the optimally trained weights (defined as the weights derived from each independent data set), we computed the dispersion over the IS and TS gene sets by averaging the squared distance of the consensus weights from the independent data set, we computed the dispersion over the IS and TS gene sets by averaging the squared distance of the consensus weights from the

External validation of the consensus gene signatures. Because individual cross-validation runs can output distinct feature sets, we defined consensus feature sets for IS and TS comprising genes that were selected in at least two of five of the cross-validated gene sets obtained at their optimal posterior probability thresholds. Finally, the IS and TS consensus feature sets were reapplied to our series to determine consensus gene weighted scores for the final consensus IS and TS signatures.

The external validation or the transportability of the two consensus signatures (IS and TS) were tested on two independent publicly available microarray expression data sets, done on either Affymetrix U133 Plus 2.0 or U133A oligonucleotide arrays. The first data set (GEO accession no. GSE3141) from Duke University (23) included a subselection of 31 stage I lung adenocarcinomas. The second independent data set (GEO accession no. GSE4573) from Michigan University (19) included a subselection of 73 patients having stage I squamous cell lung carcinomas. For both data sets, the MA55-calculated signal intensities were normalized using quantile normalization.

To quantify the amount by which the consensus weights differ from the optimally trained weights (defined as the weights derived from each independent data set), we computed the dispersion over the IS and TS gene sets by averaging the squared distance of the consensus weights from the optimal ones.

Results

This study was based on a series of 85 lung cancer patients diagnosed with stage IB primary adenocarcinoma/large cell carcinoma (Table 1). Because the effect of comorbidity on survival

Figure 1. Flow chart of the lung cancer gene signature building process. Integrative signature (IS), A, CGHmix: allocation for each BAC to a copy state (gain/loss/modal); B, ranking of the genomic clones based on their genomic-survival posterior probabilities; C, selection of nested feature sets of clones depending on the chosen posterior probability threshold; D, identification of copy number–driven expression genes; E, selection of copy number–driven expression genes located in selected genomic-survival clones. Transcriptomic signature (TS), F, ranking of the genes based on their transcriptomic survival posterior probabilities. G, selection of nested feature sets of genes depending on the chosen posterior probability threshold. H, signature development. I. 5-fold internal cross-validation for different nested sets of genes for IS and TS. J, evaluation of the discriminating ability of IS and TS to separate high-risk from low-risk patients at different posterior probability thresholds; identification of the optimal threshold. K, external validation using independent data sets.

where \( \beta_j \) (resp. \( \beta^s_j \) for TS) was the transcriptomic Cox's regression coefficient for a gene \( j \) belonging to the feature sets \( \Omega \) for IS (resp. \( \Psi \)) and \( Z_{ij} \) (resp. \( Z^s_{ij} \)) was the gene expression measurement of a gene \( j \) for the patient \( i \) over \( \Omega \) (resp. \( \Psi \)).

These signatures can be viewed as a compound covariate predictor for survival data (20, 21). Using these signatures, we classified patients into low-risk or high-risk profile group using a cutoff value determined by the median of the estimated scores obtained through the cross-validation procedure described below.

Performance evaluation of the signature building processes. The discriminating ability of each signature building process (IS and TS) to separate high-risk from low-risk patients was evaluated at different posterior probability thresholds, leading to different feature gene set sizes. At each threshold, the entire process of feature gene selection, signature computation, and high/low-risk group allocation was assessed using a 5-fold cross-validation strategy. At the end of the cross-validation procedure, each patient had a predicted group membership and the log-rank score statistic was retained and regarded as the optimal threshold for that signature.

To establish if the differences between the two survival distributions (low/high risk) were statistically significant (i.e., the gene signature performance is better than chance), we randomly permuted the survival times (and associated censoring indicators) among the tumor samples, repeated the entire cross-validation procedure, and calculated a log-rank score statistic as described above. Then, we calculated the proportion of permutations having a log-rank statistic greater or equal to the real (unpermuted) data (20) and used it to detect a significant difference at the 5% level.

External validation of the consensus gene signatures. Because individual cross-validation runs can output distinct feature sets, we defined consensus feature sets for IS and TS comprising genes that were selected in at least two of five of the cross-validated gene sets obtained at their optimal posterior probability thresholds. Finally, the IS and TS consensus feature sets were reapplied to our series to determine consensus gene weighted scores for the final consensus IS and TS signatures.

The external validation or the transportability of the two consensus signatures (IS and TS) were tested on two independent publicly available microarray expression data sets, done on either Affymetrix U133 Plus 2.0 or U133A oligonucleotide arrays. The first data set (GEO accession no. GSE3141) from Duke University (23) included a subselection of 31 stage I lung adenocarcinomas. The second independent data set (GEO accession no. GSE4573) from Michigan University (19) included a subselection of 73 patients having stage I squamous cell lung carcinomas. For both data sets, the MA55-calculated signal intensities were normalized using quantile normalization.

To quantify the amount by which the consensus weights differ from the optimally trained weights (defined as the weights derived from each independent data set), we computed the dispersion over the IS and TS gene sets by averaging the squared distance of the consensus weights from the optimal ones.
after surgical resection of stage I NSCLC patients has been recognized (24), we focused on RFS as a clinical end point. The median follow-up was 46 months. At the time of analysis, 29 disease-related deaths or tumor relapses had occurred. For the entire cohort, the RFS rate was 79.3% (95% confidence interval, 70.8–88.9) at 24 months (Supplementary Fig. S1), similar to previous observations (25). No significant relationships between RFS and classic clinicopathologic variables (age, sex, histologic differentiation, pleural involvement, vascular invasion, and TTF1 expression) were found (Supplementary Table S1).

**Patterns of CNAs.** Using BAC array-CGH technology, we analyzed the frequencies of genomic amplification/deletion events in our series (Fig. 2A). The global copy number patterns observed in our series were similar with those of previous studies (4, 26–28), showing the well-known amplifications at 1q, 5p, 7, and 8q and deletions at 3p, 5q, 8p, 9p, and 13. In particular, we found that the most common genomic alteration in our series was a gain of chromosome 5p found in 56.5% of cases, a similar rate as that published by Weir and colleagues (27). On 5p, we detected two distinct amplification events centered on the hTERT and SKP2 genes (Supplementary Fig. S2), both of which have been functionally implicated in lung carcinogenesis. Additionally, we observed the previously described common segmental amplifications such as 8q24 (c-MYC), 11q13 (CCND1), and the more recently reported region 1q41, corresponding to the NKX2-1 (TTF1) gene (27). In our study, the majority of oncogenes and tumor suppressor genes known to be associated with quantitative genomic changes in NSCLC (Supplementary Table S2) were commonly found in close proximity to the central peaks of recurrent CNAs (Supplementary Fig. S2). For example, we observed a strong correlation between array-CGH and in situ hybridization for amplification of EGFR-1 and c-MYC genes (Supplementary data).

We next defined patterns of recurrent CNAs that reflect the propensity of each genomic region to be amplified or deleted. From this chromosomal patterns analysis, 14.4% and 20.9% of the clones were classified as “exclusively amplified” or “exclusively deleted” recurrent CNAs, respectively. The most frequent exclusively amplified CNAs were observed at chromosomes 1q, 5p, 6p, 7, 8q, and 20, whereas the most frequent exclusively deleted CNAs occurred at 3p, 5q, 6q, 8p, 13, 15, 16q, 17p, and 18q (Fig. 2B). The PIK3CA gene, located at 3q26.3 locus, has been reported to be amplified in squamous cell carcinoma (4, 28) and, as expected, was not identified as a recurrent CNA in our series. In a similar vein, we observed recurrent gains of 6p and recurrent losses of 13, both of which have been shown to occur in lung adenocarcinomas (5, 26).

**Copy number–driven genes.** Using a Bayesian normal mixture model approach (16), we quantified for each gene its posterior probability for having expression changes correlated with copy number changes. The distribution of the linear correlation-based statistics formed a normal-shaped curve shifted toward positive values (Supplementary Fig. S3). Although we observed several competing mixture models that provided a good fit to the data, the estimated component means of normal distributions for these mixture models were always single, consistent with the notion that amplifications are associated with increased expression, and deletions with loss of expression. Applying the Bayes allocation rule, 42% of the genes were classified as copy number driven, consistent with a global influence of DNA CNAs on gene expression in lung cancer. Similar observations have been reported for breast cancer (7). An example of a positive correlation validated at the DNA, mRNA, and protein levels is shown for CCND1 (Supplementary Fig. S3). In addition, we observed a positive relationship between amplification of NKX2-1 (TTF1, TTF1) and its expression at both transcript and protein levels. The mean transcript levels by microarray were 5.89 and 6.90 units for nonamplified and amplified NKX2-1, respectively (P = 0.02). Furthermore, all 16 cases of amplified NKX2-1 showed detectable expression of the protein, whereas protein was detected in only 40 of 65 (62%) cases deemed not amplified for NKX2-1 (P < 0.005).

**Prognostic effect of genomic/transcriptomic changes.** To examine the relationships between copy number changes and RFS, we computed score statistics based on Cox models (Supplementary Fig. S4A). At a false discovery rate threshold of 10%, the clones with the highest posterior probabilities of being correlated to the time to relapse were located in the following regions: 1p36, 7p12, 7q11, 7q31-33, 8q22, 11q12, 14q21, 16p11-13, 16q22-q24, 20q11, 21q21-22, and 22q11-12. Of note, a highly significant increased risk for relapse was found for the amplified region 7q31-33, known to contain the oncogene MYCN whose deletion may act to reduce cancer progression and thus explain the protective effect of this chromosomal loss (29).

The prognostic effect of global gene expression changes on RFS was also calculated. Unlike the survival score statistics for the BAC genomic clones, the gene expression statistics did not show a clear trend over the chromosomes (Supplementary Fig. S4B). For a

---

**Table 1. Patient clinicopathologic characteristics**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at diagnosis (y)</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>63</td>
</tr>
<tr>
<td>Range</td>
<td>42–84</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>63 (74)</td>
</tr>
<tr>
<td>Female</td>
<td>22 (26)</td>
</tr>
<tr>
<td>Tobacco (n = 78)</td>
<td></td>
</tr>
<tr>
<td>Smokers</td>
<td>73 (86)</td>
</tr>
<tr>
<td>Non smokers</td>
<td>5 (6)</td>
</tr>
<tr>
<td>Type of resection</td>
<td></td>
</tr>
<tr>
<td>Wedge-resection/segmentectomy</td>
<td>4 (5)</td>
</tr>
<tr>
<td>Lobectomy/bilobectomy</td>
<td>78 (92)</td>
</tr>
<tr>
<td>Pneumonectomy</td>
<td>3 (3)</td>
</tr>
<tr>
<td>Necrosis</td>
<td>54 (64)</td>
</tr>
<tr>
<td>Histology</td>
<td></td>
</tr>
<tr>
<td>Adenocarcinomas of mixed subtype</td>
<td>56 (66)</td>
</tr>
<tr>
<td>Other adenocarcinomas</td>
<td>9 (11)</td>
</tr>
<tr>
<td>Large cell carcinomas/other</td>
<td>20 (23)</td>
</tr>
<tr>
<td>Histologic differentiation</td>
<td></td>
</tr>
<tr>
<td>Well differentiated</td>
<td>42 (49)</td>
</tr>
<tr>
<td>Moderately differentiated</td>
<td>7 (8)</td>
</tr>
<tr>
<td>Poorly/ nondifferentiated</td>
<td>36 (43)</td>
</tr>
<tr>
<td>Other histologic parameters</td>
<td></td>
</tr>
<tr>
<td>Lymphatic invasion</td>
<td>44 (52)</td>
</tr>
<tr>
<td>Blood vessel invasion</td>
<td>53 (62)</td>
</tr>
<tr>
<td>Visceral pleura invasion (n = 84)</td>
<td>53 (63)</td>
</tr>
<tr>
<td>TTF1 expression (n = 84)</td>
<td>51 (61)</td>
</tr>
</tbody>
</table>
global 10% false discovery rate, the selected scores were exclusively positive, indicating that overexpression increases relapse risk, whereas underexpression decreases relapse risk. Among the selected genes, SRA1, GNA12, and NTSR1 have previously been implicated in cancer progression in breast, ovarian, and glial cell cancers (30–32). Other RFS-associated genes are related to immune cell function (SLAMF9, IFIH1, IL11, and CD2BP2) and oxidative stress response (MAPK11 and TXNRD2), perhaps indicating involvement of the microenvironment. It is also worth noting the selection of PTK9 (TWF1) and PTK9L (TWF2), coding for two recently described proteins of the twinfilin subfamily that modulate cell motility by inhibiting actin polymerization (33).

**Construction and internal validation of prognostic gene signatures.** Next we sought to build an “integrated” predictive model of RFS based solely on the expressed portions of the most clinically relevant cytogenetic abnormalities. For this purpose, we restricted our gene selection specifically to copy number–driven genes located within exclusively amplified or deleted recurrent CNAs, the latter having posterior probabilities of being associated with RFS above a defined statistical threshold (see Materials and Methods). We then constructed a compound covariate predictor, termed the integrated signature (IS), using an approach similar to that of Simon and colleagues (20). We performed 5-fold cross-validation to evaluate the two classifier-building processes (feature selection and signature construction) with respect to their discriminatory capabilities. To compare the IS with a more conventionally derived expression signature, we also constructed a transcriptomic signature (TS) using the same methods, with the exception of feature selection. For TS, we considered all genes irrespective of their copy number status and ranked them based solely on their expression correlations with RFS.

Both the IS and TS processes were able to select signatures that provided statistically significant discrimination between low-risk and high-risk patients. Nevertheless, the IS process showed higher and more stable discriminating power than the TS process when increasing or decreasing the feature selection threshold (posterior probability), which relates to the number of selected clones/gene across the different cross-validation runs.

Based on the cross-validation curves, we defined optimal threshold values (0.92 for IS and 0.88 for TS) that strike a balance between having a good discriminating ability and allowing for a minimum number of selected genes. Thus, the IS defined low-risk and high-risk groups with RFS rates at 24 months of 94.5% (95% confidence interval, 87.3–100.0) and 63.7% (95% confidence interval, 48.2–84.2), respectively (Fig. 3 A). Similarly, the TS defined low-risk and high-risk groups with RFS rates at 24 months of 87.1% (95% confidence interval, 76.1–99.7%) and 74.0% (95% confidence interval, 60.6–90.3%), respectively (Fig. 3B). By performing random permutations, we found that the survival differences between the low-risk and high-risk groups defined by the IS and TS were significantly better than expected by chance ($P = 0.02$ and $P = 0.05$, respectively).

Finally, we identified final consensus gene sets for the IS and TS comprising genes that were commonly selected in repeated
cross-validations. The consensus IS was composed of 171 probe sets representing 103 unique genes located on chromosomes 7, 16, 20 and 22 (Supplementary Table S3). The consensus TS was composed of 58 probe sets representing 43 unique genes scattered over the genome (Supplementary Table S4). Not surprisingly, these two signatures included completely different sets of genes (only one gene in common), suggesting that they may reflect different biological aspects of carcinogenesis.

**External validation of the consensus IS and TS signatures.** Next, we assessed the transportability of our consensus IS and TS in two independent lung cancer data sets. Importantly, we did not retrain the weights on the new data sets, but rather directly applied the original gene weights as derived from our series. In the Duke data set subselection (consisting of 31 stage I lung adenocarcinomas analyzed on the same platform U133Plus 2.0, ref. 23), the consensus IS showed a statistically significant difference in RFS between low-risk and high-risk patients ($P = 0.003$), whereas the TS did not (Fig. 4A and B). It is worth noting that varying the number of genes for the TS did not improve its internal or external prognostic performance.

Because the locations and frequencies of recurrent CNAs are highly similar between adenocarcinomas and squamous cell carcinomas (28), we then asked if the IS retained its prognostic significance when applied to squamous cell carcinomas as well. Specifically, we tested a series of 73 patients with stage I squamous cell carcinomas from a Michigan University study (19). Because the Michigan series was analyzed on the Affymetrix U133A microarray, only 93 of 171 probe sets for the IS and 27 of 58 for the TS could be applied in validation. Nevertheless, the consensus IS showed a statistically significant difference in RFS between low-risk and high-risk patients ($P = 0.025$), whereas the TS did not (Fig. 4C and D).

To investigate the disparity between IS and TS performance, we analyzed the squared distance between the original consensus...
weights and optimally trained ones derived from the Duke and Michigan series. The distances were markedly smaller for the IS (Duke: 1.19, Michigan: 0.58) compared with the TS (Duke: 3.06, Michigan: 1.67), indicating that on the whole, the genes comprising the IS are more reproducibly associated with patient outcome in the independent series than the genes of the TS, which explains, in part, the better transportability of the IS. Together, these findings show a robust prognostic performance of the IS in predicting outcome in stage I NSCLC.

Discussion

In this work, we combined genomic and gene expression information to derive a survival model rooted in recurrent CNAs associated with NSCLC. By restricting the model to genes exhibiting copy number-driven expression, we generated a generalizable (reproducible and transportable) predictor of outcome in a subgroup of early-stage lung cancer patients for which there is clearly a need for new prognostic factors. Specifically, the IS accurately distinguished patients with high risk and low risk of relapse in our initial series and was transportable to two independent stage I NSCLC series. These results clearly show that genome copy number information can be effectively used for generating prognostic models of lung cancer survival. Interestingly, we also found that a classically constructed prognostic signature, based solely on gene expression, failed to show significance in the independent cohorts. This may suggest that integrating copy number information with gene expression may provide added power in the generation of robust prognostic signatures. Although an exhaustive comparison about whether the integrated approach is truly superior to pure gene expression approaches is beyond the scope of this report, it clearly represents an avenue for the conduct of future studies.

It is perhaps worthwhile to juxtapose our study against the backdrop of other reports describing genomic approaches to discriminate patients with early-stage NSCLC. Bhattacharjee and colleagues (34) described three distinct and stable clusters of adenocarcinoma subclasses using mRNA expression. One subclass included most bronchioloalveolar carcinomas (BAC) and were stage I tumors. In contrast, another subclass expressed neuroendocrine markers and had a significantly poorer prognosis. In our series, the genes described in the neuroendocrine subclass (KLK11, DDC, ASCL1, CALCA, PCSK, and SPE) were not significantly related with survival. This was not surprising because we excluded cases with neuroendocrine differentiation and BAC histology. Recently, Potti and colleagues (35) combined gene expression information with Bayesian statistics to describe a multifactorial model for predicting clinical outcome in early-stage NSCLC. Chen and colleagues (36) also described a simpler five-gene classifier for the same purchase. Although promising, these previous studies are also not without limitations. First, most of the signatures have been largely inferred by treating NSCLC as a single disease type, whereas in reality NSCLCs comprise a diverse mix of distinct histologic subtypes that differ radically in their global gene expression profiles (37). Furthermore, there is mounting evidence that different histologic subtypes of NSCLC may in fact exhibit different optimal molecular signatures for survival (19). This failure to incorporate histologic subtype might reduce model robustness and predictive accuracy in the pure gene expression–based models. Indeed, we found that two published pure gene expression–based models, the 5- and 16-gene signatures from Chen and colleagues (36) and a 50-gene prognostic signature from Beer and colleagues (18) and Raponi and colleagues (19), were not able to significantly discriminate between low-risk and high-risk patients in our cohort (data not shown). In contrast, the survival-associated recurrent CNAs described in our report are known to be observed across multiple NSCLC subtypes, such as amplifications of chromosome 7 and deletion of 16q (28). The commonality of these CNAs may explain why our integrated predictor was also applicable to a squamous cell lung carcinoma cohort despite it being built on an initial cohort, which was a mixture of adenocarcinomas and large cell carcinomas.

Another limitation of the gene expression–based studies is that it can be difficult to infer if genes belonging to the prognostic signatures reflect transcription within cancer cells or in the tumor stroma consisting of various fibroblast, endothelial, or infiltrating immune cells. Given the variation of stromal tissue content within and across tumor specimens, how gene expression levels arising from these tumor subcompartments relate to prognosis may be difficult to predict and may explain why some gene expression signatures show limited performance when measured over different populations. In contrast, our IS is grounded on genomic regions exhibiting recurrent CNAs. Because such CNAs are likely to be present solely within tumor cells, the IS may present a more “tumor-centric” view of gene activity and thereby improve the transportability of a survival model.

Figure 5. RFS from high-risk group stage I and stage II patients. A, RFS curves for our series (blue) and the stage I adenocarcinoma patients from the Duke series (orange). B, high-risk (orange) and low-risk (dashed orange) patients according to the IS for stage I patients from the Duke series, with the RFS for stage II patients from the same series (green) shown superimposed.
From a clinical aspect, it is worth considering the potential effect of our study on the treatment of stage IB NSCLC patients—an important clinical population where treatment options are controversial. In a preliminary analysis, we found that in the Duke series (23), the clinical outcome of stage I patients classified as “high risk” and stage II patients were similar (Fig. 5). This observation raises the potential implication that stage IB patients classified as high risk by the IS should be treated with adjuvant chemotherapy similar to stage II patients because the benefit of adjuvant treatment has already been conclusively shown in the latter group. By extension, stage IB patients designated “low risk” by the IS might consider not undergoing adjuvant treatment. The utility of an IS as a chemotherapy indicator should definitely be further evaluated in the context of a prospective clinical trial.

In conclusion, we have described in this report an integrative genomic strategy combining information about recurrent CNAs with genes exhibiting copy number–dependent expression for the creation of survival models. We then showed the robustness and transportability of this IS for stratifying stage IB NSCLC patients. Our results conclusively show that genome abnormalities in copy number are likely to exert an influence in determining patient prognosis in NSCLC. Our study highlights the relevance of combining genomic information from multiple levels to address problems of high clinical priority.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

Received 4/1/2008; revised 10/15/2008; accepted 10/18/2008; published OnlineFirst 01/27/2009.

Grant support: This collaborative work was initiated when the two first authors were visiting the Genome Institute of Singapore (visiting program 2005) and is supported by a French-Singaporean MERLION program (#5-07-06, P. Broet and L.D. Miller). Sample preparation and validation were supported by funding from Association pour la Recherche sur le Cancer (#3905) and Groupement des Entreprises Francaises dans la Lutte contre le Cancer (S. Camilleri-Broet). The BAC array-CGH profiling work was supported by grants BS-0001 (P. Tan) and IN40 (E. Lim) from the Singapore Cancer Syndicate. The expression profiling work was supported by A*STAR, the Singapore Agency for Science, Technology and Research.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank E. Liu for encouraging this work, and V. Ducruet, F. Devez, and B. Marmey for technical help.

References

30. Vartilainen MK, Sarkkinen EM, Matilainen T, Salminen L, Lappalainen P. Mammals have two twoflinn iso
Prediction of Clinical Outcome in Multiple Lung Cancer Cohorts by Integrative Genomics: Implications for Chemotherapy Selection

Philippe Broët, Sophie Camilleri-Broët, Shenli Zhang, et al.

Cancer Res  Published OnlineFirst January 27, 2009.

Updated version
Access the most recent version of this article at:
doi:10.1158/0008-5472.CAN-08-1116

Supplementary Material
Access the most recent supplemental material at:
http://cancerres.aacrjournals.org/content/suppl/2009/01/26/0008-5472.CAN-08-1116.DC1