Integrating RAS Status into Prognostic Signatures for Adenocarcinomas of the Lung

Maud HW. Starmans1,2, Melanie Pintilie3, Michelle Chan-Seng-Yue1, Nathalie C. Moon1, Syed Haider1,4, Francis Nguyen1, Suzanne K. Lau3,5, Ni Liu3, Arek Kasprzyk1,6,7, Bradley G. Wouters2,3,5,8, Sandy D. Der3, Frances A. Shepherd3, Igor Jurisica3,5,9, Linda Z. Penn3,5, Ming-Sound Tsao3,5,10, Philippe Lambin2, and Paul C. Boutros1,3,5,11

Abstract

Purpose: While the dysregulation of specific pathways in cancer influences both treatment response and outcome, few current prognostic markers explicitly consider differential pathway activation. Here we explore this concept, focusing on K-Ras mutations in lung adenocarcinoma (present in 25%–35% of patients).

Experimental Design: The effect of K-Ras mutation status on prognostic accuracy of existing signatures was evaluated in 404 patients. Genes associated with K-Ras mutation status were identified and used to create a RAS pathway activation classifier to provide a more accurate measure of RAS pathway status. Next, 8 million random signatures were evaluated to assess differences in prognosing patients with or without RAS activation. Finally, a prognostic signature was created to target patients with RAS pathway activation.

Results: We first show that K-Ras status influences the accuracy of existing prognostic signatures, which are effective in K-Ras-wild-type patients but fail in patients with K-Ras mutations. Next, we show that it is fundamentally more difficult to predict the outcome of patients with RAS activation (RASmt) than that of those without (RASwt). More importantly, we demonstrate that different signatures are prognostic in RASwt and RASmt. Finally, to exploit this discovery, we create separate prognostic signatures for RASwt and RASmt patients and show that combining them significantly improves predictions of patient outcome.

Conclusions: We present a nested model for integrated genomic and transcriptomic data. This model is general and is not limited to lung adenocarcinomas but can be expanded to other tumor types and oncogenes. Clin Cancer Res; 21(6): 1–10. © 2014 AACR.

Introduction

Lung cancer has the highest mortality rate of all malignancies. Its predominant histologic type, non–small cell lung cancer (NSCLC), accounts for about 85% of cases (1). Standard care for NSCLC is based primarily on pathologic staging, with most early-stage patients receiving surgical resection (2). Despite this intervention, 30% to 60% of patients with stage IB to IIIA NSCLC relapse and die within 5 years of diagnosis (3).

For stage II–IIIA patients, the benefit of adjuvant chemotherapy for NSCLC has been shown (4). For stage IB patients, however, there is no overall effect (5), although some subgroups do derive benefit (6). It is a major clinical goal to personalize administration of adjuvant chemotherapy by identifying those stage I patients with more aggressive disease that might benefit, and subgroups of later stage patients who might neither need nor derive benefit from additional treatment.

To achieve this goal, several studies have used transcriptome profiling on surgically excised tumor samples. Biomarkers identified in these studies usually have been derived in an unbiased manner, without exploiting prior knowledge of the underlying tumor biology (7–11). One exception was a study by Bild and colleagues, who developed “pathway signatures” for delivery of targeted therapies (12), but the strength of these data is unclear as other work by this group is in question (13). With the advent of rapid and cost-effective genome sequencing (14), it is becoming possible to integrate transcriptomic, genomic, and pathway information into multimodal biomarkers.

The best way to derive clinically relevant conclusions from these diverse datasets remains unresolved. One initial step would be to create prognostic biomarkers that include both transcriptomic data and the mutation status of genes in pathways essential to tumor development (15). In several cancers, specific patient subgroups share activation of specific oncogenes (e.g., HER2/neu...
Translational Relevance

Many groups have described and validated transcriptome-based biomarkers that predict survival of patients with non-small cell lung cancer, but these have elevated error rates that hinder clinical translation. We sought to identify the origins of this phenomenon and identified the influence of RAS mutations as a critical variable. The accuracy of existing prognostic signatures is directly influenced by K-Ras status. Using unbiased computational approaches, we show that it is more difficult to predict prognosis of patients with RAS pathway activation, and different transcriptome-based biomarkers are prognostic in patients with or without RAS pathway activation. The use of nested classification schemes that separately stratify patients with different RAS activation status are indicated and greatly improve prediction of patient outcome. This work highlights the need to directly incorporate genetic heterogeneity into biomarker discovery studies.

RAS mutation status–associated genes

Two datasets (8, 34) that reported RAS mutation status and provided transcriptomic data were used to identify genes associated with RAS mutation status (191 patients). Using these, a linear model was fit to each gene to compare RASmt and RASwt patients:

\[ Y_i = A_{i,0} + A_{i,1} \times (\text{RAS}_{\text{mutation}} \text{status}) + A_{i,2} \times (\text{dataset}) \]

where

- \( Y_i \) = normalized signal intensity for gene \( i \),
- \( A_{i,0} \) = baseline expression of the \( i \)th gene,
- \( A_{i,1} \) = coefficient for RAS status for the \( i \)th gene,
- \( \text{RAS}_{\text{mutation}} \text{status} \) = an indicator where 1 indicates RASmt and 0 indicates RASwt,
- \( A_{i,2} \) = coefficient for the effect of the dataset parameter on the \( i \)th gene,
- dataset = an indicator where 0 indicates the Beer dataset (8) and 1 the Botling dataset (34).

After applying an FDR correction for multiple testing (38), genes associated with RAS mutation status were identified (FDR < 10%).

RAS activation status classifier

Genes associated with RAS mutation status were used to train a signature for predicting RAS pathway activation, where RASmt indicates an activation of the pathway and RASwt not. A random forest (39) of 20,000 trees was built based on expression of the 14 genes with FDR < 10% across the 2 training datasets (the same 2 datasets used for feature selection; refs. 8, 34) using the RandomForest package (v4.6-6). The random forest model is available upon request. Performance of the RAS status classifier was assessed using out-of-bag (OOB) error estimates. A fully independent validation cohort was created by merging 3 other datasets (12, 32, 35) which reported RAS mutation status for all or a subset of patients (226 validation patients).

As a secondary validation, a permutation study was performed to assess classifier performance relative to the null distribution (40). A series of 290,000 random sets of \( n \) genes (where \( n = 2, 3, 4, \ldots, 30; 10,000 \) random sets per size) were generated and individually used to build a random forest classifier for RAS status prediction. The accuracy of our RAS status classifier was compared with this empirical estimate of the null distribution.

Furthermore, performance of the RAS status classifier was compared with performance of a previous published RAS pathway dependency signature (41). This signature consisted of 147 genes that were up- or downregulated as signaling through the RAS pathway increased. In total, 142 of 147 genes could be mapped to Entrez Gene IDs, and only one of these was found in our RAS mutation status prediction signature. For each patient, a signature score was calculated as described by Loboda and colleagues (41). First, data were mean normalized and transformed to \( \log_{10} \) space. Next, a score was calculated by taking the mean intensity of the “up” genes and subtracting the mean scaling and housekeeping gene normalization (to the geometric mean of ACTB, BAT1, B2M, and TBP levels) was performed, as done previously (9, 10). We maximized statistical power by focusing only on the 4,858 genes present in all datasets except where stated otherwise.

Materials and Methods

Data preprocessing

All analyses were performed in R statistical environment (v2.15.2). Public mRNA abundance datasets derived from primary NSCLC adenocarcinomas were used; only datasets with publicly available raw expression data (Supplementary Table S1) and patient-level annotation were used (7, 8, 12, 32–35). All datasets used Affymetrix microarrays and were preprocessed using the RMA algorithm (affy package v1.30.0; ref. 36) combined with updated ProbeSet annotations (cdf v14.1.0; ref. 37). Genes were matched across datasets based on Entrez Gene ID.

Starmans et al.
intensity of the "down" genes. Finally, a signature score of zero was used as threshold to categorize samples.

**GLMNet permutation analyses**

The null distribution of prognostic performance was assessed by selecting random sets of genes and fitting elastic net regularized Cox proportional hazard models with the glmnet package (v1.8; ref. 42) in the R statistical environment (v2.15.2). Patients with survival data from all 10 datasets in this study were used in the permutation study [we treated the 4 sub-datasets in the Director’s Challenge separately (32, 43); (7, 8, 12, 32, 35). The RAS activation status predictor was applied to all 10 datasets and predicted RAS status was used to distinguish RAS$^{wt}$ and RAS$^{mt}$ patients. Two separate permutation studies were performed, each with distinct goals.

The first permutation study tested whether RAS status influences signature performance. To do so, we used 3 datasets [the Director’s Challenge MSKCC dataset (ref. 32), the Botling dataset (ref. 34), and the Fouret dataset (ref. 35)] to train the signatures (testing data: 193 RAS$^{mt}$ and 107 RAS$^{wt}$ patients). Validation was then performed on the RAS$^{mt}$ patients ($n = 253$) and RAS$^{wt}$ patients ($n = 382$) from the remaining 7 datasets.

The second study was aimed at assessing whether RAS$^{wt}$ and RAS$^{mt}$ patients are fundamentally different in predicting prognosis. It used the same datasets for training and validation. However, in this case, the numbers of RAS$^{wt}$ and RAS$^{mt}$ patients in training and validation were matched. In each permutation, 100 RAS$^{wt}$ and 100 RAS$^{mt}$ patients were randomly selected from the 3 training datasets. The signatures were then validated in all RAS$^{mt}$ patients ($n = 253$) and 253 randomly selected RAS$^{mt}$ patients from the 7 validation datasets. This has the effect of balancing power in the training and validation cohorts.

In each study, we tested 100,000 random signatures per gene set size for sizes ranging from 5 to 100 genes in steps of 5 genes, yielding 2,000,000 total signatures. For each random signature, glmnet (with the elastic-net mixing parameter $\alpha = 0.1$) was run on the training cohort. The regularization parameter was chosen such that all coefficients were non-zero. For each patient in the test cohort, a score was calculated by fitting the model, and the median risk score from the training data was used to split test patients into 2 groups. Prognostic performance of the random signature was evaluated by unadjusted Cox proportional hazards modeling, followed by the Wald test in all, RAS$^{wt}$, and RAS$^{mt}$ patients of the validation cohort (survival package v2.36-14).

**RAS-dependent prognostic signature identification by glmnet**

General, RAS$^{wt}$, and RAS$^{mt}$ prognostic signatures were created using an elastic-net regularized Cox proportional hazard model (42). Training was done using either 107 RAS$^{mt}$ patients, 193 RAS$^{wt}$ patients, or both patient sets combined of the 3 training datasets from the glmnet permutation study described above. The resulting test cohort contains all remaining patients (382 RAS$^{mt}$ and 260 RAS$^{wt}$).

An elastic-net regularized Cox proportional hazard model was fit to the training dataset using the elastic-net mixing parameter; $\alpha = 0.1$ and 10-fold cross validation. The value for the regularization parameter (lambda) that maximized cross-validation performance measured by partial likelihood was selected. Next, the median risk score was determined by re-running the identified model in the training cohort. The signature was then applied to the test cohort by fitting the elastic-net regularized Cox proportional hazard model to each patient to generate a risk score. Test cohort patients were then split into predicted low- and high-risk groups based on the median risk score from the training dataset. Performance for the 3 signatures was evaluated for both RAS$^{mt}$ patients and RAS$^{wt}$ patients using Cox proportional HR modeling, followed by the Wald test. The 3 glmnet models are available upon request.

**Visualization software**

All plotting was performed in R statistical environment (v2.15.2). The packages e1071 (v1.6), lattice (v0.19-28),
latticeExtra (v0.20-6), hexbin (v1.26.0), cluster (v1.14.2), and VennDiagram (v1.3.0; ref. 43) were used for data processing and graphical representation.

**Results**

**Prognostic influence of RAS mutations**

The significance of RAS mutation status in NSCLC remains controversial, with conflicting reports (23) and no known relationship to the many prognostic mRNA signatures that have been reported (7, 11). We therefore examined the association between RAS mutation and patient outcome in mRNA abundance datasets with known RAS mutation status (8, 12, 32, 34, 35). We focused on adenocarcinomas, which are the NSCLC subtype with the highest RAS mutation frequency, and stratified patients into RAS-wildtype (RASwt) and RAS-mutated (RASmt) groups. RAS mutation status was not associated with 5-year overall survival (Supplementary Fig. S1, HR = 1.11, P = 0.54; Wald test, stage-adjusted; n = 404), confirming previous reports (19, 26).

Next, we evaluated the relationship between RAS mutations and the performance of 2 published (9, 10) and independently validated (43) prognostic signatures based on the mRNA abundances of 3 and 6 genes, respectively (Fig. 1A). Both signatures stratified RASwt patients into subgroups with distinct risks (Fig. 2A, Supplementary Fig. S2A) but failed to stratify RASmt patients into subgroups with differential survival (Fig. 2B, Supplementary Fig. S2B). Taken together, these data suggest that RAS status itself is not prognostic but rather identifies patient subsets that require separate prognostic signatures.

**Predicting RAS pathway activation**

Next we investigated transcriptional changes associated with RAS mutation status. Linear modeling was applied to 2 datasets to

Figure 2.

RAS status influences prognostic performance. Kaplan–Meier survival curves for RASwt (A) and RASmt (B) patients, stratified according to the 3-gene signature. C, heatmap of clustered genes (rows) showing differential expression between clustered RASwt and RASmt patients (columns). The color bar on top displays patient and data characteristics (RAS mutation status and dataset of origin). Clustering profiles are independent of dataset (ARI, adjusted Rand index = −3.89 x 10^{-5}), indicating no batch effects are present. In contrast, clustering is strongly associated with RAS status (ARI = 0.322).
identify genes whose mRNA abundance was associated with RAS mutation status (8, 34). At an FDR of 10%, 14 genes were associated with RAS status, including cyclin D1 (CCND1) and ras homolog family member A (RHOA; Fig. 2C, Supplementary Table S2).

A signature was generated from these 14 genes by training a random forest classifier (39) to predict RAS status. A random forest classifier is generated by growing a large number of decision trees, each trained with a randomly selected subset of patients and genes. Each tree in the random forest votes on the RAS status and RAS status is predicted from the number of votes. Approximately a third of patients are omitted from each tree, called the OOB data, and provide an unbiased estimate of classifier performance (39).

Our random forest predictions of RAS status yielded OOB accuracies of 79.1%, with misclassifications equally divided between false-positives and false-negatives (Fig. 3A). Approximately a third of patients are omitted from each tree, called the OOB data, and provide an unbiased estimate of classifier performance (39).

Our random forest predictions of RAS status yielded OOB accuracies of 79.1%, with misclassifications equally divided between false-positives and false-negatives (Fig. 3A). The number of votes received by RASwt patients was significantly higher than for RASmut patients (Fig. 3B, P = 1.08 x 10^-16, Student t test). When segregating the patients into the different error classes (true negatives, false negatives, true positives, and false positives), there is a clear difference in the fractions of votes (Fig. 3C), potentially suggesting that false-positive and false-negative patients may have RAS pathway activation through other means, although this hypothesis is challenging to test experimentally.

While our accuracy of 79.1% was promising, we wondered if it could be improved. Before undertaking an extensive machine learning analysis, we first sought to determine whether our 14-gene RAS classifier had reached a global maximum, at least based on the set of 4,858 genes used for discovery. We created 290,000 random signatures, each comprising 2 to 30 genes. Each signature was trained using a random forest, as above, and its OOB accuracy was calculated. This approach gives an empirical estimate of the null distribution (10). The 79.1% classification accuracy of our 14-gene RAS classifier was superior to all of gene sets tested and in fact the accuracy of random signatures never exceeded 75%. This provides very strong confidence that our 14-gene RAS classifier is at or near a global optimum (P = 3.45 x 10^-31; Fig. 3D, dashed line is 14-gene classifier).

To further validate our 14-gene RAS classifier, we studied 226 patients from independent datasets not used in model training (Fig. 1B) (12, 32, 35). RAS mutation status was correctly predicted in 75.7%, 67.1%, and 71.1% of samples in these datasets (Fig. 3E–G), validating our classifier. Performance of the 14-gene RAS classifier was then compared with a published RAS pathway dependency signature generated from cell line data and reported to have about 60% accuracy in lung tumors (41). In each dataset, our 14-gene classifier outperformed the published signature by margins ranging from 9.9% to 31.8% (Supplementary Table S3).

We next sought to ensure that performance of the 14-gene RAS classifier was not limited by focus on a subset of the transcriptome. We analyzed the platform with the highest number of genes (U133 Plus 2.0 arrays). First, 2 RAS classifiers were trained in the Botling dataset (34) and tested in the Bild and Fouret datasets (12, 35) by taking either all genes on these platform (n = 19,070).
or the 4,858 genes studied above. Accuracies in the independent validation cohorts were unchanged (Supplementary Fig. S3A–S3F). Next, an empirical estimate of the null distribution was made by creating 290,000 random signatures as described previously, but considering all 19,070 genes in the combined Botling, Bild, and Fouret datasets. Again our 14-gene classifier was superior to all of gene sets tested (Supplementary Fig. S3G). These data confirm that performance of the 14-gene RAS classifier was not limited by looking at a subset of the transcriptome.

Finally, to demonstrate the use of this classifier, we applied it to public mRNA abundance datasets where RAS mutation status was not reported (n = 549, Fig. 1B). Predicted RAS status in this large, well-powered cohort (power of 0.90 to detect an HR of 1.46) replicated the lack of association with 5-year survival (Supplementary Fig. S4; HR = 1.17, P = 0.22; Wald test, stage-adjusted). Furthermore, predicted RAS status confirmed that prognostic signatures performed better in the RASwt patients (Supplementary Fig. S5 and Table S4). Thus, our RAS signature predictions have similar clinical correlates to actual RAS mutation status.

It is easier to predict survival of RASwt patients

Next, we sought to generalize our observation that RAS status confounds mRNA-based prognostic marker performance (Fig. 2, Supplementary Figs. S2 and S3) by again assessing the null distribution of the overall biomarker space. We used 3 datasets (n = 300 patients) for training and the remaining 7 (n = 635 patients) for testing/validation (Fig. 1C). We generated 2,000,000 gene sets, ranging in size from 5 to 100 genes, and trained/tested each for their prognostic capability separately in all patients, RASmt patients, and RASwt patients (as classified with the RAS activation status predictor) using an elastic-net regularized Cox proportional hazard model (42).

The distribution of HRs is right-shifted in RASwt patients relative to RASmt patients (Fig. 4A, log2-transformed for visualization). This shows that it is easier to predict prognosis for RASwt patients with mRNA signatures, independent of the specific genes used. Similarly, P values are smaller in the RASwt patient group than in the RASmt patients (Fig. 4B). Both these observations are independent of signature size, as shown by elevated $P_{\text{RASmt}}$ to $P_{\text{RASwt}}$ ratio (Fig. 4C; all boxplots are above the dashed line). Many more signatures reach significance in the RASwt patients (34.9%), compared with the RASmt cohort (13.4%). Interestingly, 5.9% of signatures were significant in both groups ($P < 1.0 \times 10^{-20}$, hypergeometric test).

To demonstrate that this effect is independent of the numbers of RASwt and RASmt patients used for training and testing, a second permutation study was performed where the number of RASwt and RASmt patients was balanced by subsampling (see Materials and Methods). Training datasets comprised 100 RASwt and 100 RASmt patients; validation datasets comprised 253 RASwt and 253 RASmt patients. Even after controlling for differential sample size, prognostic signatures for RASwt patients continue to have larger HRs (Fig. 4D) and lower $P$ values (Fig. 4E and F) than those for RASmt patients. The number of statistically significant signatures remains larger in RASwt patients (21.0%) than in RASmt patients (13.4%), with the overlap remaining larger than expected by chance (3.7%; $P < 1.0 \times 10^{-20}$, hypergeometric test). Thus, RASwt patients appear to be fundamentally easier to prognose, even after controlling for their larger sample size.

To demonstrate that the observed differences were not an effect of the selected subset of the transcriptome, both permutation studies were repeated. However, now we focused on the datasets profiled on the U133A and U133 Plus2.0 arrays increasing the number of genes from 4,858 to 12,140. The training datasets remain the same. Results for both permutation studies are...
Nested classification improves prognostic performance

Our permutation studies show that generating prognostic signatures for RAS<sup>mt</sup> patients requires training on only a RAS<sup>mt</sup> patient cohort. Taken together with the differential accuracy of existing signatures on RAS<sup>mt</sup> patients, these data suggest a nested classification scheme where patients would be first stratified according to their RAS status (in this case with the RAS activation status predictor) and then prognosed with different transcriptomic biomarkers used for each subgroup. Because several existing markers accurately classify RAS<sup>mt</sup> patients, we focused on predicting prognosis for RAS<sup>mt</sup> individuals. We trained signatures with glmnet (44) on 193 RAS<sup>wt</sup> patients (RAS<sup>wt</sup>sig), 107 RAS<sup>mt</sup> patients (RAS<sup>mt</sup>sig), and the union of these 2 cohorts (AllSig). Independent validation was performed on 382 RAS<sup>wt</sup> and 253 RAS<sup>mt</sup> patients (Fig. 1C).

These 3 signatures showed modest gene-level overlap (Fig. 5A; Supplementary Table S5). Both the RAS<sup>wt</sup> sig (HR = 0.84; P = 0.39; Wald test, stage-adjusted; Fig. 5A, Supplementary Table S6) and AllSig (HR = 1.26; P = 0.24; Wald test, stage-adjusted; Supplementary Table S6) failed to prognose RAS<sup>mt</sup> patients. In contrast, only the RAS<sup>mt</sup> sig robustly predicted survival of RAS<sup>mt</sup> patients (HR = 1.61; P = 1.30 × 10<sup>−2</sup>; Wald test, stage-adjusted). Stage 1 patients show an identical trend: only the RAS<sup>mt</sup> sig successfully stratified RAS<sup>mt</sup> patients into groups with differential survival (HR = 1.86, P = 3.29 × 10<sup>−2</sup>; Wald test, Supplementary Table S7). Consequently, nested classification, applying the AllSig to the RAS<sup>mt</sup> patients and the RAS<sup>mt</sup>sig to RAS<sup>wt</sup> patients, resulted in improved prognostication in the complete cohort (Supplementary Table S8). This clearly shows the benefit of stratified, joint genomic–transcriptomic prognostic models.

We next wondered whether the differences in prognosing RAS<sup>wt</sup> and RAS<sup>mt</sup> patients could be attributed to heterogeneity of genetic alterations. We examined the TCGA lung adenocarcinoma sequencing data (45). While the overall mutation rate does not differ between RAS<sup>wt</sup> and RAS<sup>mt</sup> patients (Fig. 6, P = 0.96, t test), specific individual genes do. We focused on those genes significantly recurrently altered greater than chance alone in the TCGA data (Fig. 6) and show that RAS<sup>mt</sup> patients have significantly more alterations in RBM10 and STK11 and significantly fewer in EGRF, NF1, and TP53. Furthermore, TCGA-defined mRNA subtypes are associated with Ras status as well (Fig. 6, P = 0.045, Fisher exact test). Overall, these results highlight the multifaceted impact of Ras activation on the tumor genome and transcriptome and on patient outcome and clinical presentation.

Discussion

We hypothesized that RAS status might confound the performance of mRNA abundance–based biomarkers in NSCLC. Because the incidence of RAS mutations is highest in adenocarcinomas, we focused on this subgroup. We started by studying the transcriptional effects of RAS mutations and created a 14-gene RAS status signature. At least 5 of 14 genes in this signature are known to be regulated downstream of RAS (ARF1, CCND1, DISP6, RHOA, and TFF1). Further, several genes (CCND1,
Figure 6.
Mutation rate of specific genes is different for RAS\textsuperscript{wt} and RAS\textsuperscript{mt} patients. Differences between the number of mutations and significantly recurrently mutated genes from the TCGA lung adenocarcinoma data for RAS\textsuperscript{wt} and RAS\textsuperscript{mt} patients were examined. \(P\) values are from \(t\) test (to assess differences in number of mutations) and Fisher exact tests (to assess differences for specific genes or mRNA subtypes). RAS status was associated with TCGA mRNA subtypes (Fisher; \(P = 0.045\)).
Our 14-gene RAS activity correctly identified about 75% of patients. But while 75% prediction accuracy is far above random chance, one might expect that improved machine learning methods could be used to improve it. Surprisingly, then, a permutation study shows that the 14-gene RAS classifier is near-optimal. The natural explanation for this result is that the RAS pathway is being activated by methods outside of direct mutations of RAS. For example, mutations in the upstream signaling molecule EGFR are reported in approximately 5% to 10% of the NSCLC adenocarcinomas (47). On the other hand, some RAS mutations could have limited downstream effects and do not show typical pathway activation (41). Ihle and colleagues (48) have shown that different RAS mutations can have different downstream effects, with some having more pronounced effects than others. Furthermore, RAS mutation status was determined on the basis of mutations in KRAS codons 12 and 13 (8). However, 10% of the RAS mutations found in lung adenocarcinoma are not in KRAS (49), making it possible that some RAS mutations are missed. However, assessing the individual impact of each of these considerations is not possible given the current, limited datasets; as larger datasets with more extensive follow-up times than the current TCGA data become available this will be an important question to ask. Nevertheless, it may be of value to test the use of RAS pathway signatures in clinical contexts, rather than solely evaluating the clinical use of RAS mutations.

Next, we demonstrated that our finding that mRNA-based biomarkers are confounded by RAS status is fully general and not restricted to the 2 initial biomarkers that spurred this discovery (9, 10). We used a very large permutation study and tested 4,000,000 gene signatures to show that RASwt patients are fundamentally easier to classify than RASmt patients. Furthermore, we show that in general different signatures are prognostic in the 2 patient groups.

These data strongly suggest the use of a separate transcriptomic biomarker for RASmt patients. Therefore, we developed a RASmt dependent signature on 107 patients and validated it on 253 patients. This signature robustly classifies RASmt patients (HR = 1.61, \( P = 1.30 \times 10^{-7} \); Wald test, stage-adjusted) and clearly demonstrates that future development of transcriptional signatures can be improved by stratifying patients based on key, recurrent genetic aberrations.

This approach may be immediately suitable to other tumor types. For example, MLL2 is mutated in about 20% of squamous cell lung tumors (14). MLLs are H3 lysine methyltransferases which have an important role in transcription regulation (50). A biomarker specific for patients with squamous cell lung cancer with MLL2 mutations may improve prognostic performance in a way analogous to that used here. Overall, a better understanding of the biology of RAS mutations in NSCLC may allow development of improved biomarkers for this large and important patient subgroup.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors’ Contributions

Conception and design: M.H.W. Starmans, F.A. Shepherd, P.C. Boutros
Development of methodology: M.H.W. Starmans, N. Liu, P.C. Boutros
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): S.K. Lau, F.A. Shepherd, I. Jurisica, M.-S. Tsao
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): M.H.W. Starmans, N.C. Moon, B.G. Wouters, S.D. Dee, P.C. Boutros
Writing, review, and/or revision of the manuscript: M.H.W. Starmans, M. Pintilie, M. Chan-Seng-Yue, A. Kasprzyk, B.G. Wouters, F.A. Shepherd, I. Jurisica, L.Z. Penn, M.-S. Tsao, P. Lambin, P.C. Boutros
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): M.H.W. Starmans, S. Haider, F. Nguyen, S.K. Lau, F.A. Shepherd, L.Z. Penn
Study supervision: B.G. Wouters, P. Lambin, P.C. Boutros

Acknowledgments

The authors thank Dr. Tom John for critical reading, advice, and helpful suggestions during the preparation of the article and all members of the Boutros laboratory for technical support and insightful conversations. The results published here are in whole or part based upon data generated by The Cancer Genome Atlas project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov/.

Grant Support

This study was conducted with the support of the Ontario Institute for Cancer Research to P.C. Boutros and A. Kasprzyk through funding provided by the Government of Ontario. Furthermore, we acknowledge financial support from the CTMM framework (AIRFORCE project) and EU 7th framework program (AIRFORCE) to M.H.W. Starmans and P. Lambin and the Canadian Cancer Society Research Institute (grant #020527) to M.S. Tsao. P.C. Boutros was supported by a Terry Fox Research Institute New Investigator Award and a CIHR New Investigator Award.

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.

Received July 8, 2014; revised December 23, 2014; accepted December 30, 2014; published OnlineFirst January 21, 2015.

References


www.aacrjournals.org Clin Cancer Res; 21(6) March 15, 2015 OF9

Published OnlineFirst January 21, 2015; DOI: 10.1158/1078-0432.CCR-14-1749

Downloaded from clinicalcancerres.aacrjournals.org on April 14, 2017. © 2015 American Association for Cancer Research.


Integrating RAS Status into Prognostic Signatures for Adenocarcinomas of the Lung

Maud HW. Starmans, Melania Pintilie, Michelle Chan-Seng-Yue, et al.

Clin Cancer Res Published OnlineFirst January 21, 2015.

Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-14-1749

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2015/01/22/1078-0432.CCR-14-1749.DC1

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.