Accepted Manuscript

Proteogenomic Analysis of Surgically Resected Lung Adenocarcinoma


PII: S1556-0864(18)30784-6
DOI: 10.1016/j.jtho.2018.06.025
Reference: JTHO 1045

To appear in: Journal of Thoracic Oncology

Received Date: 24 January 2018
Revised Date: 12 June 2018
Accepted Date: 27 June 2018


This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
Title: Proteogenomic Analysis of Surgically Resected Lung Adenocarcinoma

Authors: Michael F. Sharpnack\textsuperscript{a}, Nilini Ranbaduge\textsuperscript{b,++}, Arunima Srivastava\textsuperscript{c,*}, Ferdinando Cerciello\textsuperscript{d}, Simona G. Codreanu\textsuperscript{e}, Daniel C. Liebler\textsuperscript{f}, Celine Mascaux\textsuperscript{g}, Wayne O. Miles\textsuperscript{h,**}, Robert Morris\textsuperscript{h}, Jason E. McDermott\textsuperscript{i}, James L. Sharpnack\textsuperscript{j}, Joseph Amann\textsuperscript{k}, Christopher A. Maher\textsuperscript{l}, Raghu Machiraju\textsuperscript{c}, Vicki H. Wysocki\textsuperscript{b}, Ramaswami Govindan\textsuperscript{l}, Parag Mallick\textsuperscript{m}, Kevin R. Coombes\textsuperscript{n}, Kun Huang\textsuperscript{a,**}, and David P. Carbone\textsuperscript{k,^}.

\textsuperscript{a}Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio
\textsuperscript{b}Department of Chemistry, The Ohio State University, Columbus, Ohio
\textsuperscript{c}Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio
\textsuperscript{d}University of Zurich, Zürich, Switzerland
\textsuperscript{e}Department of Chemistry, Vanderbilt University, Nashville, Tennessee
\textsuperscript{f}Department of Biochemistry, Vanderbilt University, Nashville, Tennessee
\textsuperscript{g}Department of Multidisciplinary Oncology and Therapeutic Innovations, Assistance Publique des Hôpitaux de Marseille, Aix-Marseille University, Marseille, France
\textsuperscript{h}Center for Regenerative Medicine, Massachusetts General Hospital, Boston, Massachusetts
\textsuperscript{i}Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA
\textsuperscript{j}Department of Statistics, University of California, Davis, California
\textsuperscript{k}Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio
\textsuperscript{l}Department of Medicine, Washington University in St. Louis, St. Louis, Missouri
\textsuperscript{m}Department of Radiology, Stanford University, Palo Alto, California
\textsuperscript{*}These authors contributed equally to this project.
\textsuperscript{^}Corresponding author, Phone: +1-614-685-4479, FAX: +1-614-366-1969, E-mail: david.carbone@osumc.edu
**Present addresses: KH, Department of Medicine, Indiana University School of Medicine, Indianapolis, Indiana: WOM, Department of Molecular Genetics, The Ohio State University, Columbus, Ohio: NR, Waters Corporation, Milford, Massachusetts

**Funding sources:** Financial support was received from the Vanderbilt lung cancer NCI SPORE grant (CA090949). MF is supported by a National Library of Medicine Fellowship (4T15LM011270).

**Conflicts of Interest:** DPC reports personal fees from Genentech/Roch, Bristol-Myers Squibb, and Astra Zeneca, outside the submitted work.

**ABSTRACT**

**Introduction:** Despite apparently complete surgical resection, approximately half of resected early stage lung cancer patients relapse and die of their disease. Adjuvant chemotherapy reduces this risk by only 5-8%. Thus, there is a need for better identifying who benefits from adjuvant therapy, the drivers of relapse and novel targets in this setting.

**Methods:** RNAseq and LC/LC MS proteomics data was generated from 51 surgically resected non-small cell lung tumors with known recurrence status.

**Results:** We present a rationale and framework for the incorporation of high-content RNA and protein measurements into integrative biomarkers and demonstrate the potential of this approach for predicting risk of recurrence in a group of lung adenocarcinomas. In addition, we characterize the relationship between mRNA and protein measurements in lung adenocarcinoma and show that it is outcome specific.
Conclusions: Our results suggest that mRNA and protein data possess independent biological and clinical importance, which can be leveraged to create higher-powered expression biomarkers.

Keywords: lung adenocarcinoma, non-small cell lung cancer, biomarkers, proteomics, proteogenomics

Introduction

Five-year survival of patients with surgically resected, early stage lung adenocarcinoma ranges from 50-70%¹, and adjuvant chemotherapy reduces this risk by only a small amount. An accurate prediction of the risk of tumor recurrence at the time of surgery could potentially spare patients the toxicity of adjuvant chemotherapy, and target other patients for increased therapy and surveillance. Many previous attempts have been made to predict recurrence and prognosticate outcomes after resection of lung adenocarcinomas; however, significant challenges to reproducibility and implementation have prevented the widespread use of these signatures in the clinic². Most of the previous classifiers have evaluated empirically selected protein expression by immunohistochemistry or RNA expression patterns based on microarrays. To date, there has been no effort to compare and integrate high content proteomic with transcriptomic approaches in carefully clinically annotated cases of this disease.

In this study, we present an integrative approach combining both transcriptomic and proteomic data. The central hypothesis of this study is that protein and mRNA measurements of lung adenocarcinoma tumors encompass independent information that can be leveraged to discover novel dysregulated genes and integrative clinical biomarkers.

There have been multiple proteogenomics studies in model systems, such as bacteria³, yeast⁴, and cell lines⁵. Initial studies in humans reiterated the poor correlation between mRNA and protein measurements⁶, highlighting the importance of regulation at
the post-transcriptional level. Recent studies in cell lines have proposed that a greater amount of protein variation can be explained by transcription than previously thought\textsuperscript{7}; however, a picture has emerged of bursts of mRNA transcription creating stable changes in protein expression in response to perturbation\textsuperscript{8}. In surgically resected tumor samples, the cell states vary from perturbed to steady state, implying that mRNA-protein correlation may vary as well. For example, Wei et al. showed that RNA-protein correlation differs between aging and young humans and rhesus macaques\textsuperscript{9}. The discovery that mRNA-protein correlation is a phenotype that can be correlated with biological and clinical outcomes necessitates further studies with matched mRNA and protein measurements. Large datasets of matched RNAseq and proteomics results were published by Zhang et al.\textsuperscript{10} in colorectal and Mertin et al.\textsuperscript{11} in breast cancer samples of convenience, however these study were not designed to explore an integrative clinical biomarker. Recently, Zhang et al.\textsuperscript{12} published a proteogenomic dataset from high-grade serous ovarian cancer tumors, which can be separated into early and late survivors, however there are not significant differences to mRNA and protein expression between the two groups.

Here, we investigate differential mRNA-protein correlation between recurrent and non-recurrent lung adenocarcinoma tumors. We then leverage this difference, in combination with differential mRNA and protein abundances, to predict lung adenocarcinoma recurrence with matched transcriptomic and proteomic data using a novel supervised classification algorithm.

**Materials and Methods**

*RNAseq Data collection and Preprocessing.*

RNA from tumor samples resected at Vanderbilt and MD Anderson was extracted from fresh frozen tissue with Qiagen RNeasy mini kit, converted to a poly-A selected cDNA library, and paired-end sequenced on Illumina HiSeq 2000. Raw fastq
files were filtered for adapters and low quality, and aligned to UCSC hg19 reference genome with TopHat2\textsuperscript{13} using default parameters. Read counts were generated with htseq-count\textsuperscript{14} using RefSeq gene definitions\textsuperscript{15}. RNA from tumor samples resected at WashU was extracted from fresh frozen tissue, converted to a poly-A selected cDNA library with NuGen v2 kit, and paired-end sequenced on Illumina HiSeq 2000. Raw fastq files were filtered for adapters and low quality, and aligned to UCSC hg19 reference genome with STAR 2-pass method\textsuperscript{16,17}. Read counts were generated with featureCounts\textsuperscript{18} using RefSeq gene definitions. Variants from both RNAseq datasets were extracted with samtools' mpileup\textsuperscript{19}.

\textit{The MD Anderson and Vanderbilt cohort tumor tissue preparation.}

Formalin fixed paraffin embedded tissues of tumor resections collected at Vanderbilt and MD Anderson and Washington University (WashU cohort) were used in protein extraction. The Vanderbilt and MD Anderson cohort tissue samples were deparaffinized using sub-x xylene (Surgipath, Richmond, IL) followed by rehydration in three ethanol washes as previously described (Sprung, 2009 #159). Samples were homogenized in lysis buffer containing trifluoroethanol (TFE) and 100 mM ammonium bicarbonate at pH 8.0 using Sonic Dismembrator model 100 (Fisher scientific, Pittsburgh, PA) at 20 W for 20 S with 30 S intervals. The sonication step was repeated twice, and the samples were stored on ice between sonications. The concentration of the proteins in each lysate was measured using BCA protein assay (Thermo Fisher Pierce, Rockford, Illinois) using the manufacturer's protocol. A total of 200 µg of lysate was reduced with 20 mM tris(2-carboxyethyl)phosphine (TCEP, Pierce, Rockford,IL) and 50 mM DTT (Sigma-Aldrich, St. Louis, MO) at 60 °C for 30 min followed by alkylation with 100 mM iodoacetamide (Sigma-Aldrich, St. Louis, MO) in dark for 20 min at room temperature. The concentration of TFE was reduced to 10% of the total volume by diluting in 50 mM ammonium bicarbonate. The samples were digested with trypsin
(Promega Corporation, Madison, WI) at a ratio of 1:50 (w:w) overnight at 37 °C followed by acidification with 0.5% TFA. Protein digests were frozen at -80 °C and lyophilized to dryness. The samples were re-suspended in HPLC-grade water with vortexing for 1 min and desalted using Oasis HLB 96-well µElution plate (30 µm, 5 mg, Waters Corporation, Milford, MA) as previously described (Zhang, 2014 #161).

WashU cohort tumor tissue preparation

The FFPE tumor tissues were deparaffinized in xylene followed by rehydration in ethanol as previously described (Scicchitano, 2009 #162). The tumor tissues were homogenized in a modified lysis buffer containing 0.2% RapiGest (Waters Corporation, Milford, MA) in 50 mM ammonium bicarbonate. The lysates were incubated at 105 °C for 30 min and stored on ice for 5 min. The samples were sonicated using Sonic Dismembrator model 100 (Fisher Scientific) at 20 W for 20 s with 30 s intervals. This sonication step was repeated twice, and the samples were incubated at 70 °C for 2 h. The protein concentration in each lysate was determined by BCA protein assay (Thermo Fisher Pierce, Rockford, Illinois) using the manufacturer’s protocol. A total of 100 µg of tissue proteins were reduced with 50 mM DTT at 60 °C for 30 min followed by alkylation with 100 mM iodoacetamide in dark at room temperature for 20 min. The samples were digested with sequencing grade trypsin (Promega Corporation, Madison, WI) at a ratio of 1:50 (w:w) and 0.01% ProteaseMax surfactant (Promega Corporation, Madison, WI) at 37 °C for 3 h. The samples were acidified with 0.5% TFA and centrifuged at 14000 g for 15 min. The supernatant was collected and evaporated to dryness in a Speed-Vac concentrator (Thermo Scientific). The samples were stored in -80 °C until LC/LC-MS/MS analysis.

The Vanderbilt and MD Anderson cohort peptide fractionation by off-line high pH reverse-phase chromatography.
The samples \((n = 44)\) were reconstituted in 400 \(\mu\text{L}\) of 1.0 M triethylammonium bicarbonate (TEAB) at pH 7.5 and injected into the chromatography system. Tryptic peptides were fractionated at high pH reverse-phase XBridge BEH C18 analytical column (250 mm x 4.6 mm, 130 Å, 5 µm) equipped with an XBridge BEH C18 centry guard cartridge. The separation was achieved at a flow rate of 0.5 \(\mu\text{L}/\text{min}\) in 10 mM TEAB and water at pH 7.5 (solvent A) and 100% acetonitrile (solvent B). A multi-step gradient with three linear gradients were used; from 0-5% B in 10 min, 5-35% B in 60 min, 35-60% B in 15 min and 70% B for 10 min before reaching the initial conditions. A total of 60 fractions were collected and recombined into 15 peptide fractions as previously described \cite{Zhang, 2014 #161}. The samples were evaporated to dryness in a Speed-Vac concentrator and stored in -80 ºC until LC MS/MS runs.

The Vanderbilt and MD Anderson cohort LC-MS/MS analysis.

The protein digests were reconstituted in 50 \(\mu\text{L}\) of 2% acetonitrile and 0.1% formic acid. An Eksigent NanoLC 2D pump with an AS1 auto-sampler reverse-phase LC system was used for peptide fractionation. A total of 8 \(\mu\text{g}\) were injected and separated using 0.1% formic acid (solvent A) and 0.1% formic acid in acetonitrile in a packed capillary tip (Polymicro Technologies) containing Jupiter C18 resin (Phenomenex, 5 µm, 300 Å) in-line with a solid phase extraction column (packed with the same resin). The gradient was programmed to desalt the samples on the column for 15 min at 100% A prior to separation at a flow rate of 1.5 \(\mu\text{L}/\text{min}\). The separation was achieved by changing mobile phase composition from 100% A to 25% B in 50 min, 25%-90% B in 65 min and held at 90% for extra 9 min. Peptides eluting the column were ionized at 1.45 kV and analyzed with a Thermo Velos Pro dual-pressure linear ion trap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) by data dependent acquisition. The top five MS/MS scans were acquired for every full MS scan for an m/z range from 400-2000. The method was used with an ion transfer tube temperature at
200 °C; S-lens RF 65%; dynamic exclusion with a repeat count 1 and repeat duration of 1 s for an exclusion list size of 50 mass-to-charges; CID with normalized collision energy of 30%, q= 0.25, and activation time of 10 ms; the minimum intensity threshold was set to 1000 counts.

*The WashU cohort LC/LC-MS/MS analysis.*

For the analysis of the WashU cohort (17 sample), liquid chromatography coupled to tandem mass spectrometry was performed using a Waters nanoacquity two-dimensional (2D) UHPLC system (Waters Corporation, Milford, MA) with two reverse-phases interfaced to a Thermo LTQ-Orbitrap Elite hybrid mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). A total of 8 µg of protein digest reconstituted in 100 mM ammonium formate was injected using Acquity UPLC autosampler (Waters Corporation, Milford, MA) and the peptides were fractionated online at high pH prior to analytical separation. The fractionation of peptides was achieved in the first reverse-phase column (Waters BEH C18, 130 Å, 1.7 µm, 300 µm, 100 mm) at pH 10.0 in buffer A1 (20 mM ammonium formate) by varying the amounts of solvent B1 (100% acetonitrile). The column was equilibrated at 3% B1 (v/v), which was increased to 4.7% (v/v) in 1 min eluting the first fraction of peptides and decreased back to 3% (v/v) B1 in the next 4 min. The column was held at 3% (v/v) B1 during separation at a steady flow rate of 2 µL/min. The solvent % B1 (v/v) was increased from 4.7%, 9.0%, 10.8%, 12.0%, 13.1%, 14.0%, 14.9%, 15.8%, 16.7%, 17.7%, 18.9%, 20.4%, 22.2%, 25.8% and to 65% over fifteen fractions. Each fraction eluted from the fractioning column was loaded onto a Waters symmetry C18 trap column (100Å, 5 µm, 180 µm x 20 mm) and desalted at a flow rate of 20 µL/min. The analytical separation was achieved in the second reverse-phase column (Waters HSS T3, C18, 100Å, 1.8 µm, 75 µm X 150 mm) at pH 2.4 which was equilibrated to initial conditions; 95% (v/v) A2 (water with 0.1 % formic acid) and 5% (v/v) B2 (acetonitrile with 0.1 % formic acid). The subsequent separation was achieved
by three linear gradients at 38 °C where, the % B2 was increased from 5%-9% in 3 min; 9%-30% over 44 min; 30%-40% over 5 min and 40%-85% over 5 min at a flow rate of 0.5 µL/min. The column was held at 5% (v/v) B2 from 65-70 min before reaching initial conditions. The 2D LC was coupled to LTQ-orbitrap Elite via a nanospray Flex ion source (Thermo Fisher Scientific, Bremen, Germany) containing a 30 µm inner-diameter stainless steel emitter (Thermo Fisher Scientific) with spray voltage between 1.7-1.8 kV. The orbitrap mass spectrometer was operated in data dependent acquisition mode, where the top fifteen MS/MS scans were acquired for every full MS-scan. The full MS-scan was acquired in the orbitrap MS-analyzer with resolution \( r = 120,000 \) at \( m/z \) 400 for every \( 10^7 \) charges acquired in the ion trap MS-analyzer. This acquisition was set to trigger MS/MS scans for the top fifteen most abundant \( m/z \) peaks after collision induced dissociation (CID) for an automated gain control (AGC) target value of 5000 charges. The method was programmed with an ion transfer tube temperature at 275 °C; S-lens RF 55%; dynamic exclusion with a repeat count 1 and repeat duration of 15 s for exclusion list size of 500 mass-to-charges; CID with normalized collision energy of 35%, q= 0.25 and activation time of 10 ms; the minimum intensity threshold was set to 6000 counts.

**Data processing and protein identification.**

For protein identification, Myrimatch version 2.1.111 was used with a customized RefSeq human database (version 54) and Peptitome version 1.0.42. The raw files generated in Xcaliber software (Thermo Fisher Scientific) for all fifteen fractions of each protein digest were used in the peptide identification. The MS/MS spectra were searched with fixed carbamidomethyl modification at cysteine, and variable acetylation at protein N-termini, oxidation of methionines and deamidation at asparagine and glutamine (only for WashU cohort data). A maximum of two missed cleavages were allowed for every fully tryptic peptide (proline rule applied) with a minimum peptide length of six amino
acids. The data were filtered in IdPicker software version 3.0.504. The proteins present in each sample were identified with a peptide false discovery rate (FDR) of 1% and a protein FDR of 4.45%. Protein groups were filtered to only include proteins with a minimum of two peptides and with spectra required per peptide. For proteogenomic analysis, protein groups identified in each sample were grouped based on the gene group and the respective number of spectral counts for each gene group per patient was recorded.

Normalization and Filtering.

Both proteomics and RNAseq datasets were normalized by dividing each patient column by the total number of counts in that column, and then multiplying by a million to get counts per million. We then filtered out any genes for which the median across more than half of the patients was zero. 3,960 genes were detected at RNA and protein levels in at least one sample in both the Vanderbilt/MD Anderson and WashU cohorts, and after filtering, 2,286 genes remained for downstream analysis. We only used features for which there were matching protein and RNA features from the same gene.

Differential gene expression and correlation.

We developed a novel method of differential gene expression by comparing the rank median expression of each group and dividing by the total number of genes to get a number between -1 and 1. This method is robust to outliers, simple, and non-parametric. All differential correlation was computed as the absolute value of the difference between RNA-protein spearman correlation values within each cohort. A cutoff for significance of 0.54 was used. We chose this cutoff by taking the value of correlation or anti-correlation necessary to achieve significance within a single cohort (Spearman $\rho > 0.27$, estimated $p < 0.05$) and multiplying by two, i.e. taking the minimum difference necessary between a significantly correlated and significantly anti-correlated RNA-protein pair.

Construction of the Integrative Biomarker.
To create an integrated biomarker of tumor recurrence, we employ a model selection approach. For each gene, we find a set of models, or functions, that relate the RNA measurements to the protein measurements in the non-recurrent and recurrent cohorts. Formally, we define the functions as follows.

\[ \text{Protein} \sim f_R(\text{RNA}) + N(0, \sigma_R^2) \]  
\[ \text{Protein} \sim f_{NR}(\text{RNA}) + N(0, \sigma_{NR}^2) \]  

Where \( f_R \) and \( f_{NR} \) are the recurrent and non-recurrent functions and \( N(0, \sigma_R^2) \) and \( N(0, \sigma_{NR}^2) \) are the normally distributed error terms of the models. After the models are generated on a training set, the likelihood that an expression measurement from a test sample came from a recurrent or non-recurrent patient is obtained by computing the probability density of the difference between the theoretical and test protein expression values for each model.

To learn the relationship between RNA and protein measurements for each gene, we use L1 trend filtering, which seeks to fit a piecewise linear function to the data. Trend filtering controls for over-fitting with a sparsity term, which is optimized using cross validation. We implemented trend filtering using the R package genlasso. Trend filtering seeks to optimize the following objective function.

\[ \frac{1}{2} \| \text{Protein} - f(\text{RNA}) \|^2 + \lambda \| D f(\text{RNA}) \|_1 \]  

Where \( \lambda \geq 0 \) is the regularization parameter, and \( D \) is the second-order difference matrix defined in Kim et al. Trend filtering enforces a piecewise linear regression model and the number of knots, or differing slope values, is determined by cross-validation within the training set to optimize the number of kinks given the noisiness of the data.

We compute the overall probability of a patient being recurrent or non-recurrent using Bayes’ theorem with an uninformative prior and independent genes.
Where \( P_{g,j}, R_{g,j} \) are the protein and RNA measurements for each gene in the signature for a given patient and \( M \) is the number of genes in the signature. To perform feature selection to find the final gene signature, we remove genes that are inaccurate on the training set based on the number of incorrectly predicted log odds ratios for each gene.

**Comparison of Individual vs. Integrative Biomarker Results.**

The integrative biomarker was benchmarked against a method using similar principles, except employing only protein or RNA data alone. The individual RNA and protein biomarkers were built using the same operation to build the individual RNA component of the integrative biomarker. First, samples were split into training or testing (here we use leave-one-out cross-validation). For each RNA and protein gene, measurements were first divided into recurrent or non-recurrent. Next, the validation sample’s RNA or protein values were compared to the distributions of the training recurrent and non-recurrent samples, and a log likelihood was generated that the validation sample came from either clinical group. Finally, the log likelihoods were combined into a single log likelihood using a naïve Bayesian classifier.

**Functional Analysis of Dys-regulated genes.**

The dys-regulated genes identified in this study were examined for enrichments of regulatory factors including RNA binding protein and microRNA binding sites. 5’ and 3’ UTR coordinates for all available transcripts were downloaded from the UCSC Table Browser for the human genome (hg38). UTR exon sequences were extracted for each
transcript using the R package BSgenome.Hsapiens.UCSC.hg38. Sequence motifs for 178 human RNA-binding Proteins (RBP) binding sites (101 RBPs) were collected from CISBP-RNA. Each UTR sequence (length L) was scanned for each motif (length M) using a single nucleotide sliding window providing L-M+1 scores. The maximum score for each transcript was selected as the motif representative score. The set of putative targets for each RBP motif across the whole genome were identified as the set of transcripts with representative scores >90% of the motifs' theoretical maximum. The set of targets were compare to the dys-regulated genes in order to identify the putative RBP dys-regulated targets. The background set of targets were identified as the targets associated with the global set of genes assayed (all genes for which RNA and protein data was available). A hypergeometric test was used to determine whether the dys-regulated genes were enriched as targets for each RBP motif.

MicroRNA data was collected from the TargetScan website. The human conserved microRNA family targets for were downloaded from the database (214 microRNA families). This provided a list of genomic coordinates for the microRNA binding sites. Using the UCSC liftover tool the original hg19 binding site coordinates were converted into the hg38 genomic coordinates. Overlap of these sites with transcribed regions provided the set of gene targets for each microRNA family. After identifying the set of dys-regulated microRNA targets a hypergeometric test analogous to the RBP analysis was used to calculate the putative enrichment for each of the microRNA targets in the dys-regulated gene set. RBP and microRNA motifs with a Benjamini-Hochberg corrected p-value < 0.25 were considered significantly enriched.

Results

Proteogenomic Analysis of Surgically Resected Non-Small Cell Lung Cancer

We collected fresh frozen and formalin-fixed paraffin embedded (FFPE) specimens from 61 patients, half selected for rapid recurrence after surgery, and half
selected for long-term (> 3 year) survival after surgical resection. 44 of these patients were recruited at Vanderbilt University and MD Anderson (tissue was processed at Vanderbilt for all samples), and 7 patients were recruited at Washington University in St. Louis (Supplementary Fig. 1). The patients were matched for recurrence and adjuvant chemotherapy status (Table 1). RNAseq was performed on the fresh frozen tissues and tandem liquid chromatography mass spectrometry (LC-MS) was performed on the FFPE tissues. In total, 5,482 and 6,581 protein groups were identified in the Vanderbilt and WashU cohorts, respectively. 5,284 and 5,253 of these proteins were matched by gene symbol to their corresponding mRNA in the Vanderbilt and WashU cohorts, respectively. A total of 6,577 genes were measured in at least one study, and 3,960 genes were quantified in both studies.

**RNA-protein Correlation Is Dependent on Tumor Recurrence Status In Lung Adenocarcinoma**

We observed high correlation of mRNA measurements across patients, as well as high correlation of protein measurements across patients, indicating that the data generated from each site are suitable to be combined for analysis (Supplementary Fig. 2). Median mRNA-protein spearman correlations were $\rho = 0.07$ in the WashU cohort (3004 genes compared, Fig. 1A) and $\rho = 0.17$ in the Vanderbilt cohort (4656 genes compared, Fig. 1B). These values are lower than those found in previous studies conducted in lung cancer (mean Spearman $\rho = 0.34^{24}$ and $\rho < 0.4^{25}$), colon and rectal cancer ($\rho = 0.47$), breast ($\rho = 0.39$), and ovarian cancers ($\rho = 0.45$).

Next, the pathway enrichments for low and high correlated genes were studied and found similar trends to those found in previous the cancer genome atlas studies. Interestingly, the mRNA splicing pathway, which is enriched for poor mRNA-protein correlation in colorectal, breast, and ovarian cancers, is enriched for high mRNA-protein
correlation in lung adenocarcinoma (Supplementary Table 1). Aberrant splicing has recently been implicated in lung adenocarcinoma, and may contribute to the overall low mRNA-protein correlation seen in this study\textsuperscript{26}.

Prior research has shown that the unexplained protein variability is not solely accounted for by technical noise, but also post-transcriptional regulation\textsuperscript{7}. As such, we sought to discover genes whose mRNA-protein correlation was dependent on the clinical outcome. mRNA and protein data were matched at the gene level and filtered by expression to obtain a set of 2,286 paired RNA and protein measurements per lung adenocarcinoma patient (N = 51, See Materials and Methods for details). Globally, there is a significant difference between the mRNA-protein correlation of all genes in the recurrent group and in the non-recurrent group (p value < 10\textsuperscript{-16} Wilcoxon Rank sum test, Fig. 1C). Overall, the genes we investigated were more highly correlated in the non-recurrent tumors (Fig. 1C).

Synergistic Detection of RNA and Protein Dysregulation

We investigated the gene-level differences in mRNA-protein correlation and abundances with spearman correlation (Fig. 2A) and a 2-dimensional differential expression method (Fig. 2B). We show that the mRNA-protein correlation of individual genes can vary greatly between recurrent and non-recurrent tumors (Fig. 2A). We hypothesized that mRNA-protein correlation itself may contain important information about the state of the cell. Poorly correlated mRNA and protein abundances may reflect post-transcriptional (splicing, microRNA, RNA localization, etc.) and post-translational (phosphorylation, ubiquitination, altered degradation, etc.) regulation. As such, differential correlation can be used to detect dysregulated genes in cancer, and necessitates the collection and analysis of large clinical cohorts with matched mRNA and protein data.
We found that genes can be differentially expressed independently at the mRNA and protein levels (Fig. 2B). Indeed, there is little overlap between genes that are differentially expressed at the mRNA and protein levels, including differential correlation (Fig. 2C, differential expression p-values are reported as uncorrected p-values produced by the R package, npSeq\textsuperscript{27}, see Online Methods). Were we to only use one data type, we would have found 66 differentially expressed proteins or 159 differentially expressed mRNA’s; however, the inclusion of both allows us to generate 325 hypotheses of dysregulated genes. The numbers of differentially expressed proteins and mRNA’s reported by npSeq are very low due to it’s stringency; however, we chose this non-parametric approach to minimize the chance of differential expression being driven by outliers. Outlier driven differential expression is not as useful in biomarker development, because it does not capture the behavior of an entire cohort. In addition, we observed high intragroup variability relative to intergroup variability.

We further investigated which genes were most differentially correlated. The most differentially correlated gene, Timm50, has highly correlated RNA-protein abundances among non-recurrent tumors but highly anticorrelated abundances among recurrent tumors (Fig. 3). Timm50 encodes the protein, Tim50, that is involved in the mitochondrial apoptosis pathway, is upregulated by mutant p53\textsuperscript{28}, and its loss induces apoptosis in breast cancer cells\textsuperscript{29}. Timm50 is weakly differentially expressed at the RNA level, and not differentially expressed at the protein level, such that its discovery as a dysregulated gene in our patient cohort requires the use of both data types.

To examine whether aberrant post-transcriptional regulation contributed to the poor RNA-protein spearman correlations in recurrent patients, we search for enriched RNA-binding proteins (RBPs) and micro-RNAs (miRNAs) motifs within these genes. This analysis of 178 RBP and 214 miRNA family motifs identified no significant enrichment
(FDR <0.25) for post-transcriptional motifs within this gene set \( \text{(Supplementary Table 2-4).} \)

**Integrating RNA and protein abundances for predicting tumor recurrence**

We next sought to leverage the RNA and protein data by developing a novel, comprehensive methodology to generate integrative expression biomarkers (Fig. 4). In brief, we separate patients into training and test cohorts, and then further separate the training cohort according to a binary clinical variable (Fig. 4A). In this study, the variable is recurrence status. For each gene, we perform regression using a recently developed machine learning technique, L1 trend filtering\(^{20}\), to find a piecewise-linear relationship between RNA and protein abundances in each cohort (Fig. 4A). Trend filtering produces a set of piecewise linear equations that seek to balance over and under-fitting of the model. For instance, if the relationship is highly non-linear with a high signal to noise ratio, then the model will have many knots that closely follow the data. In the case of a highly linear or low signal to noise ratio, then there will be no knots, and simple linear regression is performed. The test samples are then compared to the model, and an error is calculated that represents the difference between the model-predicted and test protein values, given the test RNA values (Fig. 4B). Errors are then calculated for each training sample and used to learn parameters for a normal distribution independently for each cohort. P-values for the test errors are extracted from recurrent and non-recurrent distributions and combined to generate a log odds ratio (LOR) for each gene-patient combination (Fig 4C). These LOR values are then summed for all genes included in the signature to generate a final LOR that a tumor will recur or not. For more details on how genes are included in the final signatures, see Online Methods.

Because our method considers protein as a function of RNA, a gene that has differential RNA expression in the absence of differential protein expression would not be considered as a useful biomarker. We remedy this situation by generating a separate
LOR that an RNA measurement was taken from recurrent or non-recurrent RNA abundance distributions (Fig. 4D). The accuracy of the LOR’s generated by the integrative or RNA-alone methods are compared on the training set for each gene, and using a simple objective function, the method decides whether to use each gene as an integrative or RNA biomarker.

Using a synthetic dataset, we show that our method is able to simultaneously utilize changes to protein concentrations, RNA concentrations, and RNA-protein correlations (Supplementary Fig. 3-5). Leave-one-out cross-validation results on our patient cohort are shown in Table 2. Our integrative method was able to correctly predict 36/51 (71%) patients’ recurrence status, including 20/26 (77%) non-recurrent patients and 16/25 (64%) recurrent patients (Supplementary Fig. 6). This is in contrast to results using protein and RNA expression separately, which collectively had an accuracy of ~50%. Interestingly, the majority of prediction errors using our integrative approach of non-recurrent patients (4/6, 67%) came from the misclassification of patients who received chemotherapy. This suggests that our method was able to find tumors that may have recurred without the intervention of adjuvant chemotherapy.

To find genes that best predict patient recurrence status, we include feature selection by evaluating each gene’s performance on the training cohort. The result is a signature generated by each cross-validation test (Supplementary Fig. 7). We evaluated the biological significance of each gene included in a majority of signatures: Sumo1, Pcbd1, Psmc5, Arc1n1, Ppa2, and Sri (For a full list of genes included in at least one signature, see Supplementary Table 5). Each of these genes was utilized as an integrative biomarker, not as an RNA biomarker. Sumo1 is covalently attached to target proteins in a process termed sumoylation. Sumoylation is involved in many cellular responses; most notably, sumoylation of DNA damage response proteins is necessary to repair DNA double-stranded breaks. Pcbd1 is a dimerization cofactor of Hnf1a,
which has been implicated in numerous cancers\textsuperscript{31,32}. \textit{Psmc5} has proteasomal functions, has been used as a biomarker of radiosensitivity in a lung cancer H460 cell line\textsuperscript{33}, and has been identified as a modifier of the \textit{Tgfb} transcriptional program\textsuperscript{34}. \textit{Arcn1} has been hypothesized to function in vesicle trafficking\textsuperscript{35}, and in one study, \textit{Arcn1} RNA expression was predictive of survival in surgically resected lung cancer\textsuperscript{36}. \textit{Ppa2} is a mitochondrial inorganic pyrophosphatase. \textit{Sri} has been shown to be involved in multidrug resistance in cancer\textsuperscript{37,38}, and protects against mitochondrial apoptosis\textsuperscript{39}. Our integrative biomarker method selected biologically relevant genes to predict lung adenocarcinoma recurrence.

\textbf{Discussion}

In this study, we present a novel comprehensive characterization of 51 lung adenocarcinoma tumors with matched RNA and protein abundance analysis. We further show that the combined analysis of RNA and protein abundances can be used to define candidate biomarkers of recurrence risk for surgically resected lung adenocarcinomas. Although several papers have used RNA data to inform the choice of protein biomarkers, our method is the first, to our knowledge, to integrate RNA and protein expression data into a single signature. In fact, our method can be more broadly implemented to perform supervised learning to predict a binary response variable using any two matched datasets.

There are several limitations of this study. First and foremost, RNA sequencing data was generated from fresh frozen tissues while proteomics data was generated from FFPE tissues. This is a possible explanation for the unusually low RNA-protein correlation. Second, although our integrative biomarker improved upon RNA or protein-based biomarkers for recurrence prediction in our dataset, the accuracy (71\%) is too low to be of clinical utility. This result is possibly due to the high intra-group variability observed in our data. Third, these patients do not have matched DNA sequencing data, so a comprehensive catalogue of driver mutations is lacking. Fourth, the majority of
recurrent tumors were male (72%) and the majority of non-recurrent tumors were female (69%). Fifth, our study was not designed with independent training and validation cohorts, so independent validation is necessary to reduce the potential for over-fitting explaining the observed results. A large, publicly available cohort of lung cancer patients with matched RNAseq, proteomics, and clinical data is sorely lacking to aid in the validation of studies such as the one presented here. Ideally, this validation set would be produced solely from fresh frozen tissue obtained from surgical resection and have sufficient numbers of patients with and without tumor progression. Sixth, protein extraction is difficult to perform and can produce variable results, particularly on heterogenous tumors. While our protein extraction methods for the Vanderbilt/MD Anderson and WashU cohorts were highly similar, they were not identical, and could be a source of technical error in our study.

Ultimately, independent validation is necessary to demonstrate robustness of our findings. A large

One interesting approach for a future study would be to find combinations of RNA’s and proteins that are predictive of a clinical or biological outcome that are not necessarily from the same gene. It might be that the expression of one protein as a function of an entirely different RNA, which is possibly non-coding, could be an excellent biomarker. This possibility highlights the fact that our method contextualizes the protein expression within the landscape of RNA expression.
List of Supplementary Material

Fig. S1. Schematic of study division of labor.

Fig. S2. Patient-to-patient correlation at RNA and protein levels.

Fig. S3. Simulated Protein differential expression.

Fig. S4. Simulated RNA differential expression.

Fig. S5. Simulated differential RNA-Protein correlation.

Fig. S6. Log Odds Ratios of Recurrence. Leave-One-Out Cross-Validation with Feature Selection

Fig. S7. Leave-One-Out Cross-Validation Biomarker Selection via Thresholding of Training Results.

Table S1. Pathway enrichment of high and low RNA-protein correlations (jto.sharpnack.supplementarytable2.xlsx)

Table S2. RBP and miRNA enrichment with assayed proteins as background. (jto.sharpnack.supplementarytable3.xlsx)

Table S3. RBP and miRNA enrichment with global protein background. (jto.sharpnack.supplementarytable4.xlsx)

Table S4. mRNA-protein combination genelists used in RBP and miRNA enrichment analysis. (jto.sharpnack.supplementarytable5.xlsx)

Table S5. Genes included in LOOCV signatures after feature selection

References


4417–4424 (2010).


14. Anders, S., Pyl, P. T. & Huber, W. HTSeq-A Python framework to work with high-


36. Tomida, S. *et al.* Gene expression-based, individualized outcome prediction for


**Acknowledgements:** **Funding:** Financial support was received from the Vanderbilt lung cancer NCI SPORE grant (CA090949). MF is supported by a National Library of Medicine Fellowship (4T15LM011270). **Author contributions:** MS, AS, FC, JS, PM, KC, & KH conceived of the method. MS & AS encoded the method in R and MS wrote the manuscript. NR & JA produced and processed MS/MS data for the WashU cohort. SC & DL produced MS/MS and RNAseq data for the Vanderbilt cohort. WM & RM ran the miRNA and RNA binding protein enrichment analyses. JM produced the functional enrichment. CM & RG provided the WashU samples. DC is the PI for the clinical trial initiated at Vanderbilt. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** RNAseq data will be available for download from dbGAP (accession number pending). Raw proteomics data will be able to be downloaded from ProteomeXchange (accession number pending). Processed data and biomarker code can be found at github.com/michaelsharpnack/proteogenomics. **None** of
the results presented in this paper have been published or are under consideration.

**Informed consent** was obtained after the nature and possible consequences of the study was explained.

**Figure Legends**

**Figure 1.** Gene-level mRNA-protein correlation in human lung adenocarcinoma. mRNA protein correlation in Vanderbilt (A) and WashU (B) datasets. (C) Histogram of mRNA-protein correlations within each cohort. The significance of the difference between recurrent and non-recurrent mRNA-protein correlations was determined by the wilcoxon rank sum test.

**Figure 2.** Synergistic discovery of differentially regulated genes using matched RNA and protein abundances. (A) RNA-protein correlations within recurrent and non-recurrent patient cohorts are shown in a scatterplot. Genes whose RNA-protein abundances are significantly correlated or anti-correlated (uncorrected p-value < 0.05) are shown in red. (B) RNA and protein differential expression is shown as the change in median rank abundance between non-recurrent and recurrent cohorts. Genes are differentially expressed at RNA and protein levels. (C) Overlap of genes differentially expressed at the protein and RNA levels, as well as genes that are differentially correlated. Zero genes displayed simultaneous differential expression at both levels and differential correlation. Please see Online Methods for more information about how differential expression and correlation was computed.

**Figure 3.** *Timm50* is differentially correlated between recurrent and non-recurrent tumors. (A) *Timm50* is weakly differentially expressed at the RNA level (p < 0.05), but
not differentially expressed at the protein level (B). Timm50 differential RNA-protein correlation between recurrent and non-recurrent tumors.

**Figure 4.** Overview of integrative RNA-protein biomarker discovery pipeline. (A) Patients are divided into their clinical groups, here we use binary recurrence status to group the patients. Regression is then performed using trendfiltering to find a relationship between RNA and protein abundances within each cohort. (B) This model is then used to test a separate test sample or samples. Given a test RNA abundance, the test error is calculated as the difference between the predicted protein value and the test protein value (arrows). (C) The test errors are then compared to the distributions of training errors in each cohort, and a log odds ratio is calculated (LOR1). (D) Since this integrative method does not detect differential RNA abundances in the absence of differential protein abundances, a second log odds ratio is calculated by comparing the test RNA abundances to the training RNA abundances in each cohort (LOR2).
### Table 1. Clinical patient attributes.

<table>
<thead>
<tr>
<th></th>
<th>Recurrent (N=25)</th>
<th>Non-Recurrent (N=26)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male Sex (%)</td>
<td>72</td>
<td>31</td>
</tr>
<tr>
<td>Adjuvant Therapy</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>No Adjuvant Therapy</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ia/b</td>
<td>8/8</td>
<td>6/14</td>
</tr>
<tr>
<td>IIa/b</td>
<td>3/4</td>
<td>2/1</td>
</tr>
<tr>
<td>IIIa/b</td>
<td>0/1</td>
<td>2/1</td>
</tr>
<tr>
<td>Collection Site</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vanderbilt</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>MD Anderson</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>WashU</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 2  Integrative Biomarker of Recurrence Leave-One-Out Cross Validation Performance.

<table>
<thead>
<tr>
<th>Datatypes Used</th>
<th>Feature Selection</th>
<th>Total (%)</th>
<th>Non Recurrent (%)</th>
<th>Recurrent (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA</td>
<td>All</td>
<td>24 (47)</td>
<td>12 (46)</td>
<td>12 (48)</td>
</tr>
<tr>
<td></td>
<td>FS</td>
<td>29 (57)</td>
<td>13 (50)</td>
<td>16 (64)</td>
</tr>
<tr>
<td>Protein</td>
<td>All</td>
<td>18 (35)</td>
<td>10 (38)</td>
<td>8 (32)</td>
</tr>
<tr>
<td></td>
<td>FS</td>
<td>24 (47)</td>
<td>12 (46)</td>
<td>12 (48)</td>
</tr>
<tr>
<td>RNA+ Protein</td>
<td>All</td>
<td>19 (37)</td>
<td>9 (35)</td>
<td>10 (40)</td>
</tr>
<tr>
<td></td>
<td>FS</td>
<td>25 (49)</td>
<td>11 (42)</td>
<td>14 (56)</td>
</tr>
<tr>
<td>Integrative</td>
<td>All</td>
<td>22 (43)</td>
<td>10 (38)</td>
<td>12 (48)</td>
</tr>
<tr>
<td></td>
<td>FS</td>
<td>36 (71)</td>
<td>20 (77)</td>
<td>16 (64)</td>
</tr>
</tbody>
</table>