

ARTICLE

Received 5 May 2014 | Accepted 3 Oct 2014 | Published 28 Nov 2014

DOI: 10.1038/ncomms6469

# Integrated Omic analysis of lung cancer reveals metabolism proteome signatures with prognostic impact

Lei Li<sup>1</sup>, Yuhong Wei<sup>1</sup>, Christine To<sup>2</sup>, Chang-Qi Zhu<sup>2</sup>, Jiefei Tong<sup>1</sup>, Nhu-An Pham<sup>2</sup>, Paul Taylor<sup>1</sup>, Vladimir Ignatchenko<sup>2</sup>, Alex Ignatchenko<sup>2</sup>, Wen Zhang<sup>1,3</sup>, Dennis Wang<sup>2</sup>, Naoki Yanagawa<sup>2</sup>, Ming Li<sup>2</sup>, Melania Pintilie<sup>2,4</sup>, Geoffrey Liu<sup>2,5</sup>, Lakshmi Muthuswamy<sup>5,6</sup>, Frances A. Shepherd<sup>2,7</sup>, Ming Sound Tsao<sup>2,5,8</sup>, Thomas Kislinger<sup>2,5</sup> & Michael F. Moran<sup>1,2,3</sup>

Cancer results from processes prone to selective pressure and dysregulation acting along the sequence-to-phenotype continuum DNA → RNA → protein → disease. However, the extent to which cancer is a manifestation of the proteome is unknown. Here we present an integrated omic map representing non-small cell lung carcinoma. Dysregulated proteins not previously implicated as cancer drivers are encoded throughout the genome including, but not limited to, regions of recurrent DNA amplification/deletion. Clustering reveals signatures composed of metabolism proteins particularly highly recapitulated between patient-matched primary and xenograft tumours. Interrogation of The Cancer Genome Atlas reveals cohorts of patients with lung and other cancers that have DNA alterations in genes encoding the signatures, and this was accompanied by differences in survival. The recognition of genome and proteome alterations as related products of selective pressure driving the disease phenotype may be a general approach to uncover and group together cryptic, polygenic disease drivers.

<sup>1</sup>Department of Molecular Structure and Function, The Hospital For Sick Children, 686 Bay Street, Toronto, Ontario, Canada M5G0A4. <sup>2</sup>Princess Margaret Cancer Centre, 101 College Street, Toronto, Ontario, Canada M5G 1L7. <sup>3</sup>Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada M5G 0A4. <sup>4</sup>The Dalla Lana School of Public Health Sciences, University of Toronto, Toronto, Ontario, Canada M5T 3M7. <sup>5</sup>Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada M5G 1L7. <sup>6</sup>Ontario Institute for Cancer Research, 661 University Avenue, Toronto, Ontario, Canada M5G 0A3. <sup>7</sup>Department of Medicine, University of Toronto, Toronto, Ontario, Canada M5G 1L7. <sup>8</sup>Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada M5G 1L7. Correspondence and requests for materials should be addressed to M.F.M. (email: m.moran@utoronto.ca).

Lung adenocarcinoma (ADC) and squamous cell carcinoma (SCC) are the two major subtypes of lung cancer, and combined account for >900,000 deaths per year worldwide. Characterization of tumour genomes has revealed large numbers of tumour-associated DNA sequence alterations, including genes frequently mutated in particular cancers including lung ADC and SCC<sup>1–4</sup>. Indeed, tumorigenesis is driven by combinatorial changes in tumour suppressor and oncogene (that is, driver) gene sequence and dosage<sup>5</sup>. However, the roles of large numbers of tumour-associated DNA sequence alterations that occur at only low frequency and may act in concert remain to be determined.

The cancer phenotype is ultimately a product of a set of processes, each prone to selective pressure and dysregulation in cancer, acting at each stage of the sequence continuum DNA → RNA → protein<sup>6</sup>. Selective pressure is accepted as a driving force behind cancer-associated remodelling of the genome and epigenome, but has not been systematically linked to the proteome. Indeed, the extent to which the cancer phenotype is a product of the proteome is not known. The ability of early-stage non-small cell lung carcinoma (NSCLC)<sup>7</sup> and breast cancer<sup>8</sup> primary tumour explants to establish primary tumour-derived xenograft (PDX) models is prognostic of poor outcome. Hence, tumour engraftment may select for critical, aggressive aspects of the cancer phenotype linked to disease progression. This is consistent with reports that PDX models harbour genetic aberrations closer to metastatic disease than corresponding primary tumours in breast and pancreas cancers, and melanoma<sup>9–12</sup>. However, the extent to which proteome remodelling in primary tumours is recapitulated in PDX models and links to disease progression is not known.

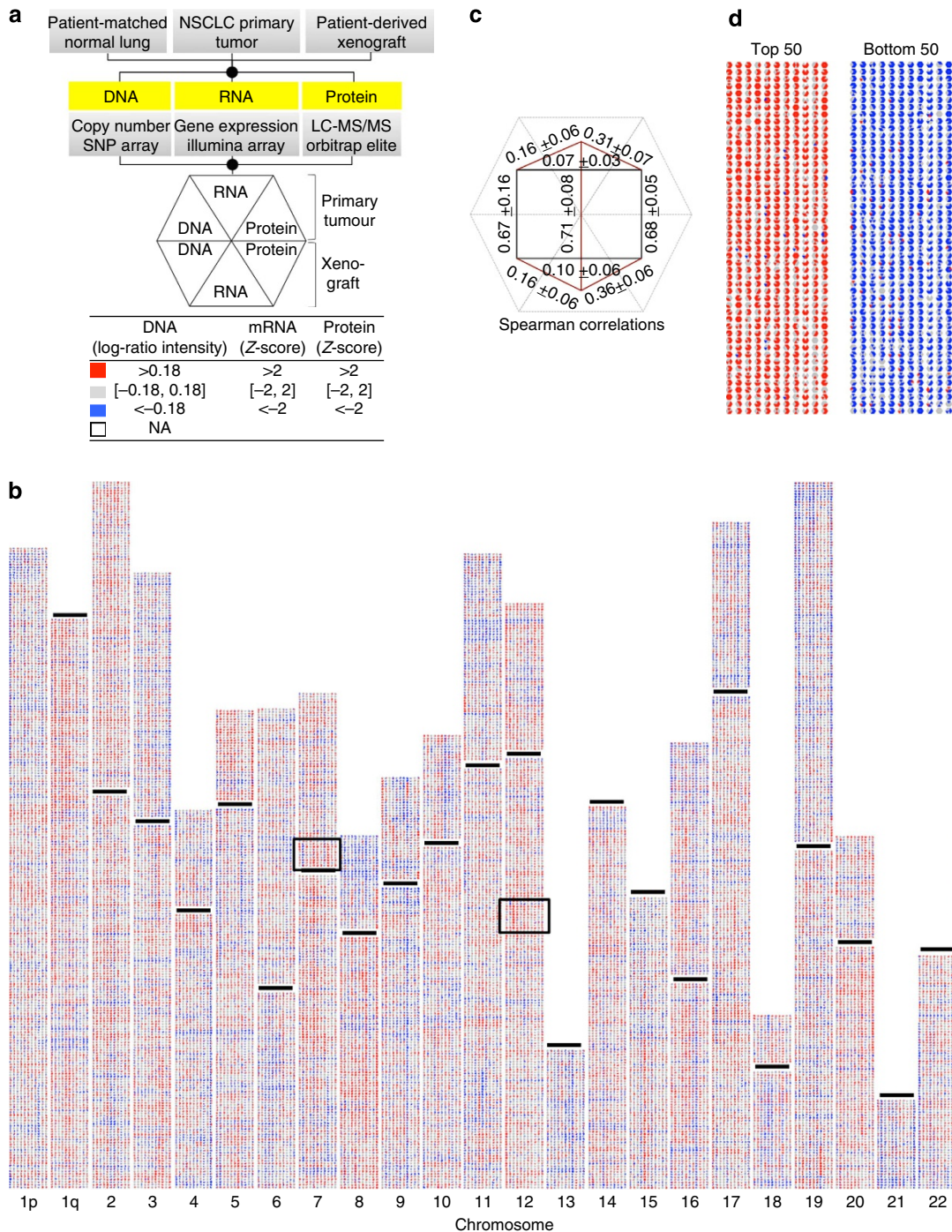
In order to uncover molecular features linked to NSCLC clinical parameters, we integrate DNA, RNA and proteomics data sets spanning the tissue spectrum from normal lung to patient-matched primary and PDX tumours. The data are integrated in the form of a genetic array and reveal genetically linked, NSCLC-associated changes in the proteome that are conserved between cognate primary and PDX tumours. We report that proteins not previously implicated as cancer drivers are encoded throughout the genome including, but not limited to, regions of recurrent DNA amplification/deletion in NSCLC. We hypothesized that polygenic cancer genes that function in concert may have remained cryptic if they respond to selective pressure through proteome remodelling, and consequently are infrequently mutated or subject to copy number changes. We further posited that since PDX engraftment is prognostic of overall survival in NSCLC<sup>7</sup>, then proteome remodelling that is linked to overall survival would be highly recapitulated in PDX models. Unsupervised clustering reveals signatures composed of sets of proteins involved in metabolism that are especially highly recapitulated between primary and PDX tumours, but which differ between the major NSCLC subtypes ADC and SCC. Interrogation of The Cancer Genome Atlas (TCGA) reveals sizeable cohorts of patients with DNA alterations in genes encoding the metabolism proteome signatures, and this is accompanied by differences in survival. Signatures with prognostic impact discriminate lung ADC and SCC, and in some instances are associated with non-lung cancers. Serine hydroxymethyltransferase 2 (SHMT2), a key enzyme in serine/glycine and folate-dependent, one-carbon metabolism, is upregulated in the proteomes of NSCLC primary and PDX tumours, and is implicated as a driver of chromosome 12q14.1 amplification, which is recurrent in NSCLC. SHMT2, along with other enzymes implicated as anti-folate targets, is also part of a metabolism signature associated with poor outcome in lung ADC. The interrogation of cancer genomes and proteomes for alterations that are related products of selective pressures

driving the cancer phenotype may be a general approach to uncover and group together cryptic, polygenic cancer drivers, which might represent new anticancer therapeutic targets.

## Results

**An integrated omic map of NSCLC.** A set of 33 samples consisting of 11 each primary NSCLC tumour, patient-matched normal lung and PDX<sup>7</sup> were analysed for DNA, RNA and protein as outlined in Fig. 1a. Tumours scored >70% for cellularity, and histological and immunohistochemical marker (p63/CK5/TF1) evaluation established 7 as ADC, and 4 as SCC (Supplementary Table 1 and Supplementary Fig. 1). Ultrahigh-resolution mass spectrometry<sup>13,14</sup> was used to identify and quantify 4,030 protein groups (Supplementary Data 1), and with Venn analysis showing 72% overlap in protein expression across the three sample types (Supplementary Fig. 1). This level of lung proteome coverage is comparable to the 3,621 proteins measured with pooled NSCLC samples and non-patient-matched controls in Kikuchi *et al.*<sup>15</sup> For each quantified protein, corresponding gene DNA copy number, mRNA expression level and protein amount were calculated relative to matched normal lung, and the data were incorporated into a symmetrically divided hexagon, with colour-encoded values for primary tumour in the top quadrants and with corresponding values for the cognate PDX entered in the bottom sections, as depicted in Fig. 1a. By this arrangement, the top and bottom halves of the hexagon represent mirror images of the DNA, RNA and protein measures for the primary tumours and cognate xenografts, respectively. In order to facilitate the recognition and visualization of genetically linked trends across tumours and grafts, an omic map was assembled in which 44,330 hexagons (4,030 genes × 11 patients) were assembled into an array by ordering horizontally according to NSCLC patient (from left-to-right 7 ADC and 4 SCC), and as a linear genetic map in the vertical direction (Fig. 1b and Supplementary Data 2).

Protein expression from chromosomes was highly positively correlated with chromosome gene density ( $r_s > 0.9$ ) and was not generally related to cytogenetic size, indicating no gross differential regulation of the proteome at the level of chromosomes (Supplementary Fig. 1). Spearman correlation analysis of the integrated data sets indicated that each of the DNA, RNA and protein measures was highly correlated between primary and PDX tumours (Spearman  $r_s > 0.6$ ), whereas RNA → protein translations were correlated to a moderate level ( $r_s \leq 0.4$ ), and DNA → RNA ( $r_s < 0.2$ ) and DNA → protein ( $r_s < 0.1$ ) to a much lesser extent (Fig. 1c). These data confirm that generally protein abundance is not a simple function of DNA copy and mRNA levels<sup>16,17</sup>, and therefore largely unpredictable based on these indirect measures. However, the omic map reveals genes whose dysregulation at the level of DNA, RNA and protein, relative to normal lung, is highly correlated between primary and PDX tumours. For example, a gene predominantly upregulated at the DNA, mRNA and/or protein level in primary tumours, and grafts would give rise to a horizontal red feature across the encoding chromosome in the array. Conversely, a broadly downregulated gene would give rise to a horizontal blue feature. Such horizontal features are dispersed throughout the array (and extracted from the map in Fig. 1d and Supplementary Data 2), and in some instances intersecting with vertical features representing regions of linked gene amplification or loss (Fig. 1b). We suggest that proteins consistently up- or downregulated across tumours, recapitulated in grafts and whose genes map into regions associated with focal amplification or deletion, respectively, may reflect proteome remodelling in response to selective pressure for an aspect of the tumour phenotype that is manifest at the protein level.

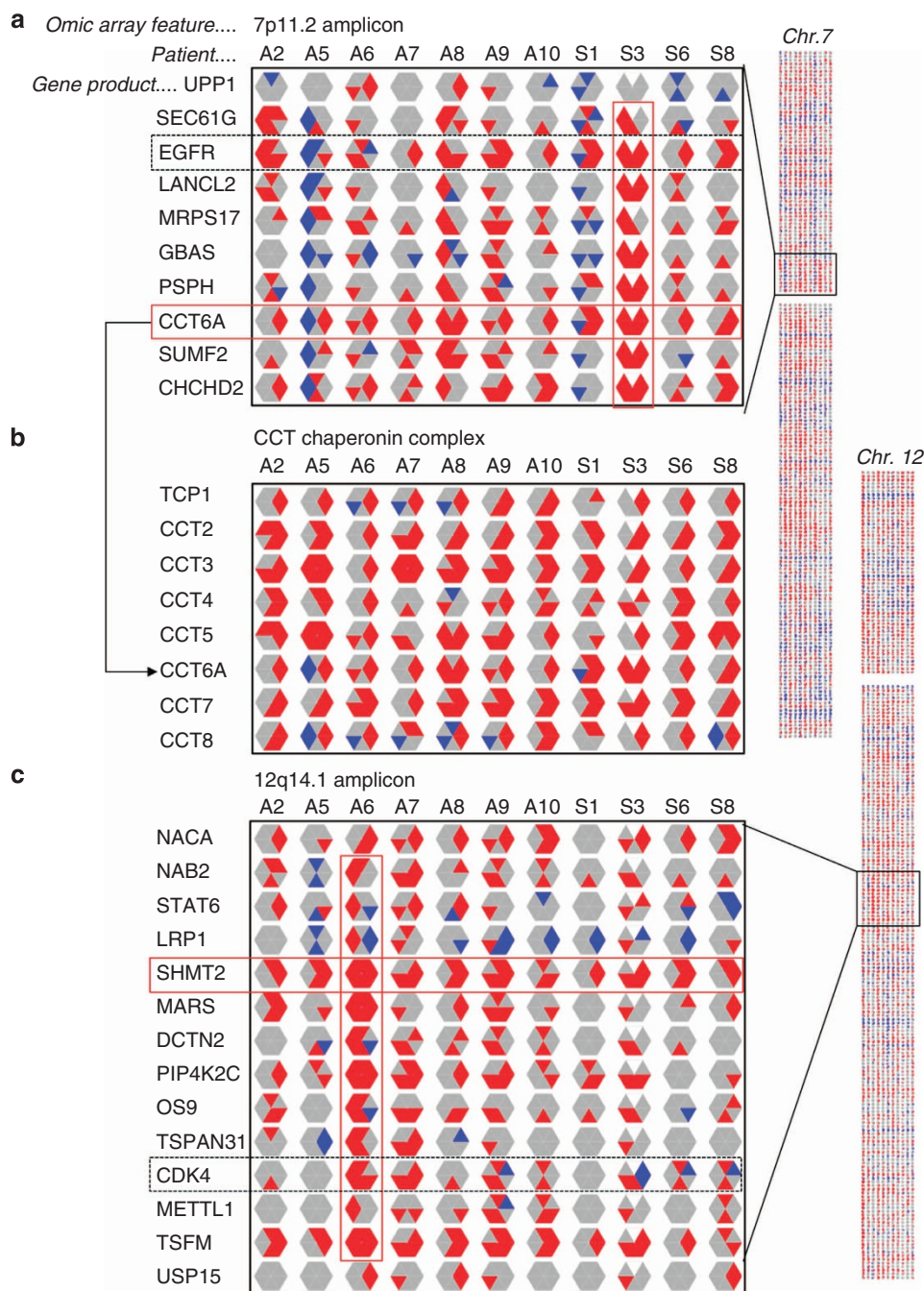


**Figure 1 | An integrated omic platform and DNA → RNA → protein array for the characterization of NSCLC.** (a) Normal lung, patient-matched primary tumours and PDXs were quantified for DNA copy number, mRNA and comprehensive protein. Data for each primary tumour and PDX, relative to matched normal lung, were colour-encoded as indicated, and integrated into six-sided polygons in a symmetrical manner, with primary tumour data in the top three quadrants, and data from the cognate PDX in the mirror-image bottom sections. By this approach,  $> 4 \times 10^5$  data points were integrated into 44,459 polygons. (b) The polygons were assembled into a two-dimensional genetic array: As a linear genetic map in the vertical direction, organized by chromosome and with bars indicating centromeres, and by patient in the horizontal direction. Increases in DNA, RNA and protein relative to normal lung are shown in red and decreases in blue. Highly concordant changes in DNA, RNA and protein are therefore evident as streaks of red or blue. Such features in the vertical direction represent genetically linked changes affecting adjacent genes/gene products and typically driven by regions of DNA amplification or loss. In some instances, these are intersected by streaks of the same trend (that is, up/red or down/blue) in the horizontal direction. These reflect correlated changes in one or more of the linked genes across the set of individual patient tumours and recapitulated in PDXs. Two examples of such intersecting features near 7p11 and 12q14 are shown boxed (and further expanded in Fig. 2). See also Supplementary Data 2 for higher resolution images of individual chromosome arrays further annotated with gene name, chromosome number and nucleotide position. (c) Spearman correlations were determined between DNA, RNA and protein, and between respective primary tumour and PDXs as indicated. (d) The 50 genes most highly upregulated (Top 50, red) or downregulated (Bottom 50, blue) and concordant across the DNA → RNA → protein sequence continuum in their dysregulation compared with normal lung were excised from the integrated array/map. SNP, single nucleotide polymorphism; NA, data not available. See also Supplementary Data 2.



Two examples of such intersecting features, indicated by boxes in Fig. 1b, are located near 7p11.2 and 12q14.1. The chromosome 7 amplicon is in tumour SCC3 and contains the gene encoding the NSCLC drug target, epidermal growth factor receptor (EGFR) (Fig. 2a), whereas the chromosome 12 amplicon is in ADC6 and

contains the oncogene *CDK4* (Fig. 2c). Both of these amplicons are recurrent in NSCLC<sup>4</sup>. Tumour SCC3 was observed to have a cytogenetically defined amplification of *EGFR* (Supplementary Table 1). Consistent with this, in both the primary and PDX tumour, the nine adjacent genes were co-amplified giving rise to a



**Figure 2 | Analysis of the omic array for evidence of proteome selection. (a)** An expanded segment of the array from the short arm of chromosome (chr.) 7 (7p11.2) as indicated, showing the detected genetically linked components of an EGFR amplicon. Note that mRNA was not available for primary tumour sample SCC3 (S3). The boxed vertical and horizontal red features reflect increases relative to normal lung: The vertical red feature results from the amplification of the indicated genes in patient SCC3. The horizontal features reflect increases associated with the gene *CCT6A* across the 22 tumours and grafts (solid red box) and *EGFR* (dashed black box). **(b)** Each of the eight components of the CCT chaperonin complex (including *CCT6A*) were extracted from the array and show *in toto* at the protein level a twofold increase in tumours and 2.3-fold increase in grafts (Table 1). **(c)** An expanded segment of the array from the long arm of chromosome 12 (12q14.1), which includes a vertical red feature (boxed in red) associated with an amplicon spanning the indicated 12 genes in patient sample ADC6 (A6). Horizontal boxes highlight *SHMT2* (solid red box), which was increased relative to normal lung in most of the primary and xenograft tumours, and *CDK4* (dashed black box), which is a known proto-oncogene product within the amplicon. The quantification of amplicon proteins is summarized in Table 1. See also Supplementary Fig. 2 and Supplementary Data 2.

vertical red feature in the array (see the vertical box in Fig. 2a). These genes were also amplified in ADC8 as part of a non-focal amplification. As a proto-oncogene product, the EGFR is an obvious driver of 7p11.2 amplification and tumour growth. Strikingly, every PDX showed elevated EGFR protein relative to normal lung or the cognate primary tumour (Fig. 2a and Table 1). This is consistent with increased EGFR translation associated

with tumour hypoxia<sup>18</sup> and/or impaired EGFR downregulation that occurs in various cancers<sup>19</sup>.

Examination across the array for protein expression from the genes co-amplified with EGFR in SCC3 suggests additional drivers in CCT6A and CHCHD2, which were increased in both primary and PDX tumours (Fig. 2a and Table 1). Every primary tumour and xenograft had elevated CCT6A protein relative to normal lung (Fig. 2a). The CCT6A protein is part of the CCT (cytosolic chaperonin containing t-complex polypeptide 1) chaperonin complex that comprises two rings, each containing eight paralogous subunits<sup>20,21</sup>. The complex plays a central role in proteostasis<sup>22</sup>, and is involved in the folding of actin and tubulin<sup>23</sup>, and proteins associated with cell proliferation and cancer including B and E type cyclins, p21 RAS and the von Hippel-Lindau (VHL) tumour suppressor<sup>24–26</sup>. Genomic studies have implicated individual CCT complex genes in cancer, but the data are perplexing, given the obligate hetero-oligomeric structure of the CCT protein complex<sup>20,21</sup>. For example, CCT3 alone was identified as part of a prognostic NSCLC gene expression signature<sup>27</sup>, whereas CCT2 and CCT5 transcript levels are related to colorectal cancer progression.<sup>28</sup>

Inspection of data from TCGA via cBio Cancer Genomics Portal<sup>29</sup> indicates that various CCT subunit genes are subject to copy number gains in many cancers (Supplementary Fig. 2). For example, >40% of lung ADC, the highest among measured cancer types, contain a DNA alteration (that is, gene amplification, or, rarely mutation or gene copy number loss) in at least one of the eight CCT subunit genes, as do >30% of lung SCC. In nine other cancer types, the incidence of CCT gene alteration exceeds 20%, and they present largely in a mutually exclusive manner (Supplementary Fig. 2). However, if equivalent subunit stoichiometry is required in order to generate functional CCT protein complexes, an increase in only one subunit would be futile. One explanation for these observations would be if CCT subunit stoichiometry is maintained at the protein level. In order to test this and to determine whether the CCT complex or just the CCT6A subunit was altered, data for all of the CCT complex subunits were extracted from the array (Fig. 2b). This revealed that indeed all eight CCT subunits were significantly elevated at the protein level in primary and PDX tumours compared with normal lung (Table 1), and this is consistent with the conclusion that CCT chaperonin activity is broadly activated in NSCLC.

Coiled-coil-helix-coiled-coil-helix domain containing 2 (CHCHD2) was co-amplified with EGFR in SCC3 (Fig. 2a). Indeed, there is a strong tendency for their co-amplification in ~7% of NSCLC (Supplementary Fig. 2). Our analysis suggests that their upregulation at the protein level is more frequent, with 8 of 11 cases showing elevated CHCHD2 in primary and/or PDX tumours (Fig. 2a), and elevated on average eightfold in primary and sevenfold in PDX tumours (Table 1). CHCHD2 promotes cell migration and regulates oxidative phosphorylation both at the protein level in mitochondria and transcriptionally<sup>30–33</sup>. Therefore, elevated expression of CHCHD2 may contribute to increased mitochondrial function and cell motility in NSCLC<sup>34</sup>. These data suggest that amplification at 7p11.2 may to some extent be driven by selective pressure to increase protein expression of CCT6A and CHCHD2, in addition to EGFR.

CDK4 resides within the 12q14.1 amplicon identified in ADC6 (Fig. 2c) and is associated with recurrent amplification in lung cancer<sup>35</sup>. However, with the exception of ADC6 and PDX-ADC5, CDK4 protein expression was not generally elevated in the NSCLC samples, and was in fact decreased relative to normal lung in four of the primary tumours (Fig. 2c). By contrast, proteins encoded by SHMT2 and TSMF were each increased in 21 of the 22 tumour and xenograft samples (Fig. 2c and Table 1).

**Table 1 | Genetically or functionally linked dysregulated proteins in NSCLC.**

Gene	Chr. no.	Protein expression* (Fold, relative to normal lung)			
		Primary tumour	P	PDX	P
<i>Chr. 7 amplicon</i>					
SEC61G	7	1.59	0.17	0.82	0.62
<b>EGFR</b>	<b>7</b>	<b>4.99</b>	<b>0.01</b>	<b>31.2</b>	<b>0.00</b>
LANCL2	7	1.84	0.31	2.02	0.27
MRPS17	7	2.12	0.01	1.43	0.32
GBAS	7	1.16	0.66	0.99	0.96
PSPH	7	1.27	0.72	1.56	0.43
<b>CCT6A</b>	<b>7</b>	<b>2.28</b>	<b>0.00</b>	<b>2.78</b>	<b>0.00</b>
SUMF2	7	1.49	0.22	1.19	0.54
<b>CHCHD2</b>	<b>7</b>	<b>8.35</b>	<b>0.00</b>	<b>7.04</b>	<b>0.00</b>
<i>CCT chaperonin complex</i>					
<b>TCP1</b>	<b>6</b>	<b>2.09</b>	<b>0.00</b>	<b>2.28</b>	<b>0.00</b>
<b>CCT2</b>	<b>12</b>	<b>2.19</b>	<b>0.00</b>	<b>2.55</b>	<b>0.00</b>
<b>CCT3</b>	<b>1</b>	<b>2.25</b>	<b>0.00</b>	<b>2.38</b>	<b>0.00</b>
<b>CCT4</b>	<b>2</b>	<b>1.66</b>	<b>0.01</b>	<b>1.82</b>	<b>0.01</b>
<b>CCT5</b>	<b>5</b>	<b>1.92</b>	<b>0.00</b>	<b>2.35</b>	<b>0.00</b>
<b>CCT6A</b>	<b>7</b>	<b>2.28</b>	<b>0.00</b>	<b>2.78</b>	<b>0.00</b>
<b>CCT7</b>	<b>2</b>	<b>1.82</b>	<b>0.00</b>	<b>2.19</b>	<b>0.00</b>
<b>CCT8</b>	<b>21</b>	<b>2.18</b>	<b>0.00</b>	<b>2.12</b>	<b>0.00</b>
<i>Chr. 12 amplicon</i>					
NAB2	12	1.91	0.10	1.82	0.14
STAT6	12	1.68	0.29	1.24	0.69
LRP1	12	0.61	0.04	0.63	0.07
<b>SHMT2</b>	<b>12</b>	<b>13.93</b>	<b>0.00</b>	<b>22.09</b>	<b>0.00</b>
<b>MARS</b>	<b>12</b>	<b>2.21</b>	<b>0.00</b>	<b>2.40</b>	<b>0.00</b>
DCTN2	12	1.20	0.34	1.04	0.87
<b>PIP4K2C</b>	<b>12</b>	<b>5.06</b>	<b>0.00</b>	<b>10.41</b>	<b>0.00</b>
OS9	12	1.26	0.39	1.40	0.50
SPAN31	12	1.03	0.95	1.01	0.98
CDK4	12	0.69	0.42	0.89	0.82
METTL1	12	0.80	0.34	1.40	0.27
<b>TSMF</b>	<b>12</b>	<b>9.84</b>	<b>0.00</b>	<b>18.51</b>	<b>0.00</b>
<i>Ser/Gly/One-carbon</i>					
PHGDH	1	3.62	0.03	3.79	0.08
<b>PSAT1</b>	<b>9</b>	<b>6.94</b>	<b>0.00</b>	<b>12.63</b>	<b>0.00</b>
GLS	2	0.96	0.91	1.25	0.58
<b>ASNS</b>	<b>7</b>	<b>4.31</b>	<b>0.01</b>	<b>5.47</b>	<b>0.00</b>
PSPH	7	1.27	0.72	1.56	0.43
CBS	21	1.67	0.19	2.55	0.02
PTDSS1	8	1.63	0.11	1.48	0.23
<b>SHMT2</b>	<b>12</b>	<b>13.93</b>	<b>0.00</b>	<b>22.09</b>	<b>0.00</b>
SHMT1	17	1.67	0.25	1.95	0.06
<b>MTHFD1</b>	<b>14</b>	<b>2.90</b>	<b>0.00</b>	<b>2.98</b>	<b>0.00</b>
<b>MTHFD1L</b>	<b>6</b>	<b>5.14</b>	<b>0.00</b>	<b>7.59</b>	<b>0.00</b>
<b>MTHFD2</b>	<b>2</b>	<b>3.96</b>	<b>0.00</b>	<b>5.48</b>	<b>0.00</b>
<b>MARS</b>	<b>12</b>	<b>2.21</b>	<b>0.00</b>	<b>2.40</b>	<b>0.00</b>
<b>GCSH</b>	<b>16</b>	<b>5.89</b>	<b>0.00</b>	<b>9.30</b>	<b>0.00</b>
GLDC	9	1.28	0.41	1.60	0.18

Chr., chromosome; NSCLC, non-small cell lung carcinoma; PDX, primary tumour-derived xenograft.

\*Fold changes >1.5 and P<0.05 in primary and PDX tumours are in bold

SHMT2 was upregulated at the protein level 14-fold in primary tumours and 22-fold in PDX tumours (Table 1). It is functionally redundant with SHMT1 (ref. 36), which was not significantly differentially expressed in tumours relative to normal lung (Table 1). SHMT2 catalyses folate-dependent conversion of serine to glycine as part of the one-carbon pathway<sup>37</sup>, and therefore plays a key role in the biosynthesis of nucleotides and S-adenosylmethionine required for the synthesis and methylation of DNA. Interestingly, MARS-encoded methionyl-tRNA synthetase, which is genetically linked to SHMT2 (Fig. 2), was upregulated more than twofold in primary and PDX tumours (Table 1). A similar level of ectopic MARS overexpression was sufficient to stimulate epithelial cell proliferation<sup>38</sup>. Recent reports indicate that tumours, or tumour-initiating cells<sup>39</sup>, become 'addicted' to dysregulated serine biosynthesis and glycine cleavage pathways<sup>39,40</sup>. Additional enzymes involved in serine/glycine/one-carbon metabolism were upregulated in the primary and PDX tumours (Table 1 and Supplementary Fig. 2). These proteins include PSAT1, which is elevated in colon tumours and tumorigenic in related cell lines<sup>41</sup>, and PHGDH, which is upregulated in several cancers including melanoma<sup>42</sup>, and is required for anaplerotic metabolism when upregulated in estrogen receptor (ER)-negative breast cancer models<sup>40</sup>. Although PHGDH protein was elevated in both ADC and SCC subtypes compared with normal lung (Table 1), it was significantly higher in SCC than ADC primary tumours (15-fold,  $P=0.02$ ), consistent with recent analyses of NSCLC xenografts<sup>43,44</sup>. The cytosolic folate-dependent dehydrogenase, MTHFD1, and especially its mitochondrial equivalents, MTHFD1L and MTHFD2, were upregulated in primary and PDX tumours (Table 1). This upregulation at the protein level, along with that of SHMT2, is consistent with their elevated mRNA expression in various cancer types<sup>45</sup>. SHMT2 was implicated as a driver required for survival in a series of cancer cell lines<sup>46</sup>, and was shown to be upregulated and required for maintenance of redox balance and tumour cell survival in response to hypoxia<sup>47</sup>. Finally, in apparent contrast to its metabolism function in mitochondria, SHMT2 was recently identified as an adaptor subunit of a cytosolic BRISC (BRCC36 isopeptidase complex) deubiquitinase<sup>48</sup>. These data, and additional results described below, suggest that amplification at 12q14.1 may to some extent be driven by selective pressure to increase the expression of SHMT2.

TSMF is a mitochondrial translation elongation factor required for the production of respiratory proteins encoded by the mitochondrial genome. However, of the three mitochondrial genome encoded proteins detected (ATP8, COX1 and COX2), none was upregulated in the PDX models, and only COX1 was upregulated in primary tumours (fourfold;  $P=0.03$ ). Therefore, the functional significance of TSMF upregulation and, in particular, any effect on the 10 undetected proteins expressed from the mitochondria genome remain to be determined.

#### Analysis of tumour proteome recapitulation in PDX models.

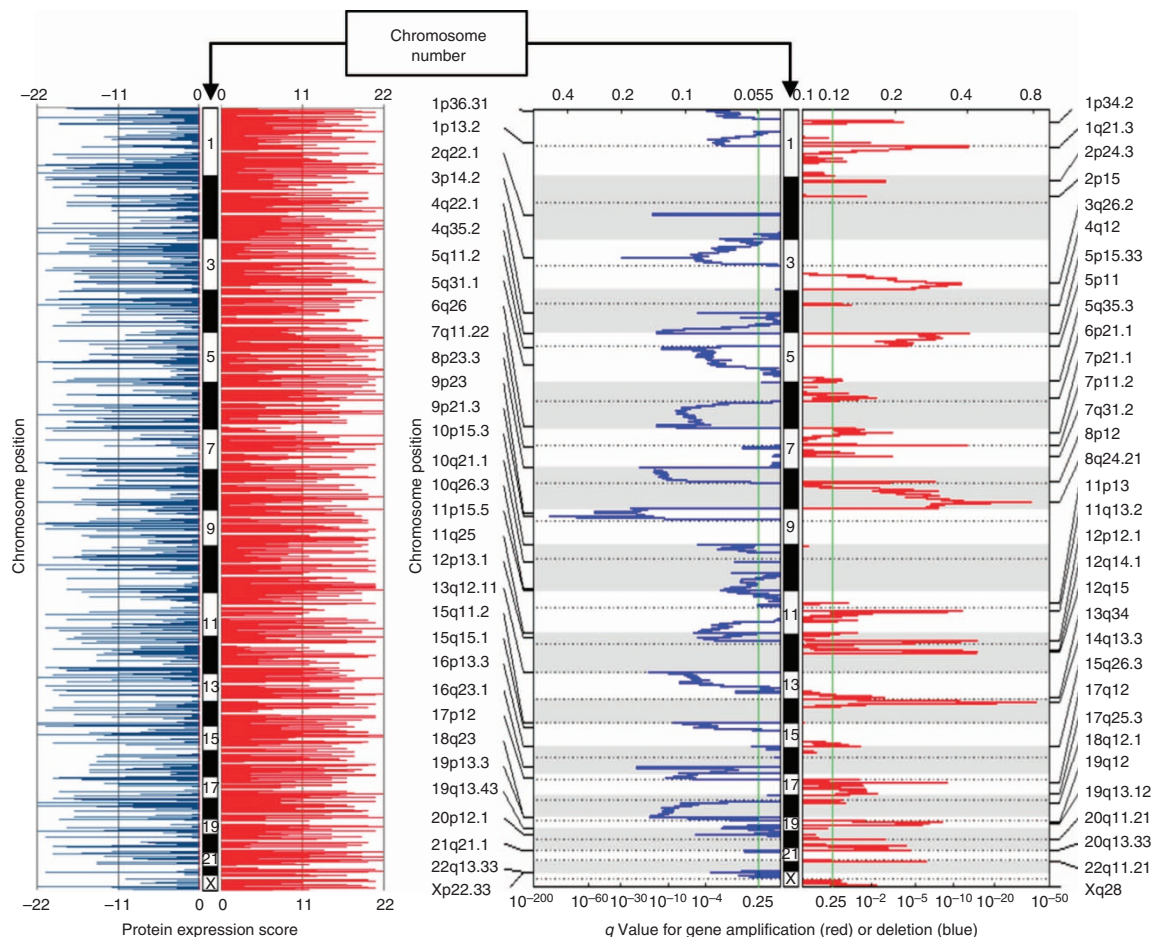
The analyses of the 7p11.2 and 12q14.1 amplicons are examples that illustrate how the integrated omic map establishes an approach to explore how cancer genomes and transcriptomes impact the proteome. However, a limitation of the omic map visualization approach is that trends in the proteome may be less apparent if non-concordant with DNA and RNA measures. Indeed, analysis of the omic map revealed vast numbers of highly dysregulated proteins encoded broadly across the genome, and which are not limited to corresponding regions of recurrent genome amplification or deletion in NSCLC (Fig. 3 and

Supplementary Fig. 3). This is consistent with the Spearman analysis (Fig. 1c), and suggests that protein expression-based positive and negative drivers of the cancer phenotype have remained cryptic at the genome and transcriptome level.

The histology, and DNA and RNA profiles of the PDX models closely mirrors that of cognate primary tumours (Fig. 1c and Supplementary Fig. 1)<sup>7</sup>. Moreover, early-stage NSCLC tumours that engraft are biologically more aggressive and appear representative of cancers with a higher propensity to relapse after surgery<sup>7</sup>. Therefore, we compared the proteomes of primary and PDX tumours with the rationale that proteins important to the cancer phenotype would be particularly highly recapitulated in the PDX models. Unsupervised hierarchical clustering of the normal, primary tumour and PDX samples according to protein abundance fully resolved normal lung from primary and PDX tumours, and showed that none of the normal lung samples was matched to its cognate primary tumour or PDX (Fig. 4a). Hence, normal lung proteomes are distinctly different from tumour proteomes. This comprehensive proteome analysis indicated that only five primary and cognate PDX tumours were most similar to each other (Fig. 4a). We questioned whether proteins most highly differentially expressed in tumours relative to normal lung would be more similar between primary and PDX tumours. Surprisingly, when the basis of comparison was a subset of the proteome comprising the 359 most highly differentially expressed proteins (that is, fold change  $>5$ ,  $P<0.05$ ), only six primary PDX pairs matched up, including ADC6, and the same five that matched by using the comprehensive proteome data set (Fig. 4b).

The Kyoto Encyclopedia of Genes and Genomes ([www.genome.jp/kegg](http://www.genome.jp/kegg)) was used to analyse the measured proteomes, and the most prevalent annotation assigned to the proteins detected in the 33 samples was metabolism. When the proteomes were searched against a curated set of 2,752 metabolic enzymes and transporters<sup>40</sup>, a total of 838 non-redundant metabolism proteins were identified (Supplementary Data 1), including 200 dysregulated in primary and PDX tumours relative to normal lung (|fold change|  $>2$ ,  $P<0.05$ ; Supplementary Fig. 4). Changes in metabolism underlie the cancer phenotype, and inhibitors targeting metabolic enzymes are in clinical development as anticancer therapeutics<sup>49</sup>. Indeed, several specific enzyme isoforms involved in central carbon metabolism were dysregulated (Supplementary Fig. 5). However, a comprehensive characterization of metabolism proteins in primary cancer tissue has not been made. When the tissue samples were subjected to unsupervised clustering according to the metabolism proteome, that is, the 838 quantified metabolism proteins, normal lung remained as a distinctive group, and, surprisingly, a substantial increase in the pairing of cognate PDX and primary tumours was observed with nine matches out of a possible 11. Furthermore, among the paired primary and PDX tumours, the histological subtypes ADC and SCC were resolved from each other (Fig. 4c). This high degree of correct primary tumour-to-PDX pairing could not be replicated randomly ( $P<1.0\times 10^{-5}$ , permutation test). When clustering was conducted according to proteins annotated for macromolecular synthesis and stability, or cell/extracellular interactions, which, similar to metabolism, are protein classes substantially represented in the proteome data set and dysregulated in the primary and PDX tumours (Table 2), the clustering of normal lung samples was again retained, but correct primary tumour-PDX pairing occurred in only three cases. Hence the metabolism-proteome represents a subset of the proteome that is exceptionally similar between tumours and their grafts including the recapitulation of histological distinctions, but significantly different from normal lung. This is consistent with the notion that the transformed metabolic state constitutes a core, driving component of the tumour phenotype<sup>6</sup>.





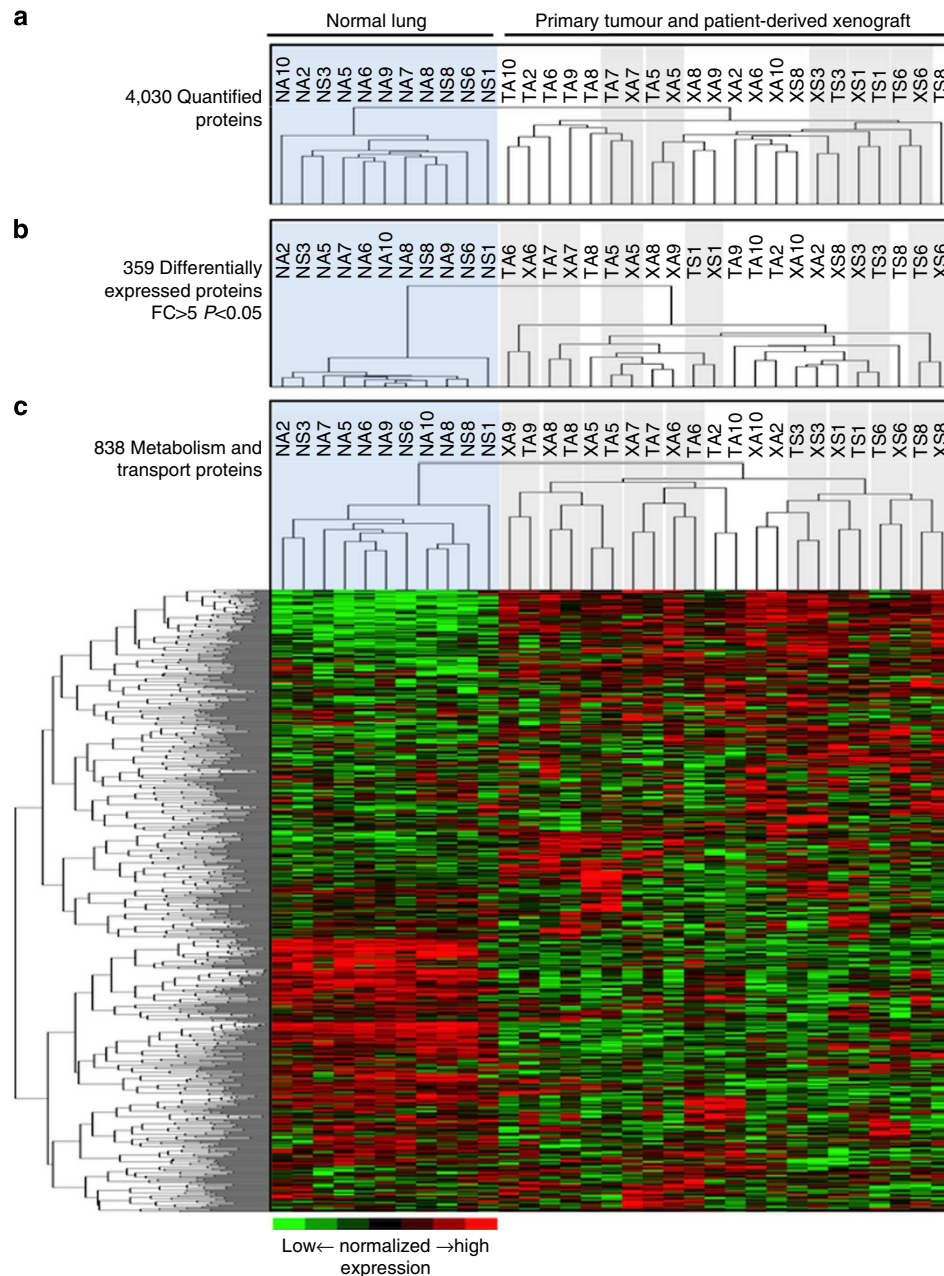
**Figure 3 | Genome-wide protein differential expression compared with DNA amplification/deletion in NSCLC.** Protein expression was determined by mass spectrometry for each of 11 primary tumours and corresponding PDX samples and scored +1 if upregulated (red, right) or -1 if downregulated (blue, left) relative to normal lung, using the same criteria used to generate the omic map (Fig. 1b). For each quantified protein, the sum of the 22 individual sample scores represents the differential expression score, which was plotted as a linear genetic map (left panel) and aligned with an analogous GISTIC analysis of DNA amplification (right panel, red) and deletion (right panel, blue)<sup>67</sup> based on the analysis of >300 NSCLC, sourced from TCGA and by using Tumorscape ([www.broadinstitute.org/tumorscape](http://www.broadinstitute.org/tumorscape))<sup>68</sup>. Genomic Identification of Significant Targets in Cancer (GISTIC)<sup>67</sup> score (top) and FDRs (q values, bottom; vertical green line is 0.25 cutoff for significance) for each alteration (x axis) are plotted at each genome position (y axis; chromosome numbers indicated); dotted lines indicate the centromeres. Amplifications (red lines) and deletions (blue lines) are shown. See also Supplementary Fig. 3.

**Metabolism-proteome signatures and clinical significance.** We reasoned that if the protein clusters (Fig. 4c) reflect concerted selective pressure to maintain the cancer phenotype during engraftment, then there should exist DNA alterations in the encoding genes that define patient subgroups with shared phenotypic features. To test this, genomics and clinical data from 1,900 patients spanning seven different cancers (Table 3) were retrieved through TCGA and cBioPortal<sup>29</sup>. For each cancer type, every node of the metabolism-proteome dendrogram (Fig. 4c) was iteratively searched for DNA alterations (that is, amplification, deletion and mutation) associated with the genes encoding the clustered proteins. This revealed clusters that were altered in >10% of patients and correlated with overall survival in NSCLC and other cancers (Table 3 and Supplementary Data 3). Clusters with prognostic impact were selected from the dendrogram analysis based on their local minimal log-rank *P* values (see Methods section). An internal validation algorithm was applied in which for each cluster,  $10^4$  mock clusters comprising the same number of proteins from the 838 detected metabolism proteins were randomly generated and were similarly tested. Only clusters for which the log-rank *P* value was in the top fifth percentile of the corresponding mock clusters were accepted.

Table 3 presents a subset of 18 clusters, ranging in size from 3 to 87 proteins and comprising 295 non-redundant proteins, in which overall survival was significantly different between NSCLC patient groups with or without DNA alterations in the set of genes encoding the clustered proteins (log-rank *P* value < 0.05). Specification sheets for each of the 18 clusters, including frequency and type of DNA alteration by gene and cancer type, and Kaplan–Meier survival curves are assembled in Supplementary Data 4.

The gene-based clusters were exclusive for lung tumour histology, with 12 associated with lung SCC and another six lung ADC. Eleven of the SCC clusters were associated with better overall survival, whereas all but one of the ADC clusters was associated with worse overall survival (Table 3). Therefore, the segregation of tumours according to histology, which was driven by the clustering of metabolism proteins, was largely reiterated when the metabolism proteome clusters were extrapolated to encoding genes and clinical outcomes.

In some instances, individual clusters were statistically significantly correlated with overall survival in cancers other than NSCLC (Table 3, and Supplementary Data 3 and 4). For example, cluster C10, associated with better outcome in lung



**Figure 4 | Distinctive metabolism protein profiles and altered metabolism pathways in NSCLC.** Dendrograms produced by unsupervised hierarchical clustering (Euclidean distance with complete linkage) of proteins quantified in 33 samples consisting of 11 each patient-matched normal lung, primary tumour and patient-derived xenograft (PDX). **(a)** In all, 4,030 quantified proteins. **(b)** Proteins (359) identified as highly dysregulated relative to normal lung ( $|\text{fold change}| > 5$ ,  $P < 0.05$ ). **(c)** Proteins (838) annotated for metabolism according to the Kyoto Encyclopedia of Genes and Genomes and Possemato *et al*<sup>40</sup>. Aligned with the dendrogram of patient samples is the corresponding heat map showing expression values for the 838 metabolism proteins organized by unsupervised hierarchical clustering by using Euclidian distance and with complete linkage. FC, fold change. See also Supplementary Figs 4 and 5.

SCC, also defined a subset of patients with head and neck SCC, but was associated with worse outcome. Cluster C2, comprising 43 genes/gene products including the cancer driver *IDH2*, was associated with better outcome in both lung SCC and glioblastoma multiforma, but worse outcome in acute myeloid leukaemia. Cluster C6, which was associated with better outcome in lung SCC, was altered in 40% of breast invasive carcinoma and was associated with worse outcome. Three clusters, which were associated with worse (C9) or better (C11) outcomes in lung SCC or worse outcome in lung ADC (C14), were associated with better outcomes in ovarian serous cystadenocarcinoma (Table 3 and Supplementary Data 4). Hence, metabolism proteome clusters,

when extrapolated to encoding genes constitute signatures conserved across different cancer types.

The clusters in some instances comprises differentially expressed proteins (Table 3; downregulated are in *italics* and upregulated are in **boldface**). Of the 295 cluster proteins, 65 were differentially expressed and highly correlated between primary and PDX tumours ( $R^2 = 0.86$ ). A Venn analysis and outline of metabolic pathways defined by the clusters and differentially expressed proteins are represented in Supplementary Fig. 6. Two clusters (C10 and C15) comprise entirely of proteins upregulated in NSCLC primary and PDX tumours relative to normal lung, and coincidentally had the greatest statistical significance between



**Table 2 | KEGG pathway analysis and dysregulated proteins in NSCLC.**

Process*	Identifier†	Identified proteins	Dysregulated proteins‡
<i>Metabolism</i>			
Arginine and proline metabolism	hsa00330	37	15
Fructose and mannose metabolism	hsa00051	21	10
Glycolysis/gluconeogenesis	hsa00010	42	17
Pentose phosphate pathway	hsa00030	20	7
Pyruvate metabolism	hsa00620	27	8
Propanoate metabolism	hsa00640	22	7
Butanoate metabolism	hsa00650	22	8
Amino sugar and nucleotide sugar metabolism	hsa00520	35	12
Valine, leucine and isoleucine degradation	hsa00280	37	11
Glutathione metabolism	hsa00480	37	11
Cysteine and methionine metabolism	hsa00270	22	6
Fatty acid metabolism	hsa00071	32	8
Citrate cycle (TCA cycle)	hsa00020	25	3
Oxidative phosphorylation	hsa00190	72	9
<i>Macromolecular synthesis and stability</i>			
Ribosome	hsa03010	77	59
Spliceosome	hsa03040	104	78
DNA replication	hsa03030	22	15
RNA degradation	hsa03018	35	21
Aminoacyl-tRNA biosynthesis	hsa00970	28	17
Proteasome	hsa03050	38	16
Regulation of actin cytoskeleton	hsa04810	97	32
Lysosome	hsa04142	61	14
<i>Cell/extracellular interactions</i>			
Focal adhesion	hsa04510	107	46
ECM-receptor interaction	hsa04512	51	33
Tight junction	hsa04530	67	19
Adherens junction	hsa04520	40	11

FDR, false discovery rates; KEGG, The Kyoto Encyclopedia of Genes and Genomes; NSCLC, non-small cell lung carcinoma.  
 \*DAVID enrichment analysis; Benjamin-adjusted  $P$  value < 0.01; human genome as background and 4,030 quantified proteins as the gene list.  
 †The KEGG Pathway classification code.  
 ‡Relative to patient-matched normal lung (FDR < 0.05, [fold change] > 2).

patient groups, with Kaplan–Meier log-rank  $P$  values of 0.004 and 0.0003, respectively (Fig. 5). C10 is defined by seven proteins, including the cancer-associated M2 isoform of the glycolytic enzyme pyruvate kinase (PKM2)<sup>50</sup>, and three enzymes involved in nucleotide biosynthesis (Table 3). Based on inspection of isoform-specific peptides, PKM2 was confirmed as elevated in tumours ( $\log_2$  (tumour/normal) = 1.7, false discovery rate (FDR) < 0.001), whereas PKM1 was not ( $\log_2$  (tumour/normal) = 0.06, FDR > 0.9; Supplementary Data 5). C15 has features of a folate signature, since three of its proteins (PAICS, PPAT and GART; Table 3) were previously identified as methotrexate-binding, candidate anti-folate targets<sup>51</sup>, whereas SHMT2 was recently identified as a target of the next-generation anti-folate, pemetrexed<sup>52</sup>, and showing co-expression with other folate enzymes<sup>53</sup>. Folate-dependent metabolism, involving one-carbon

units provided by SHMT activity as noted above, is required for the replication and methylation of DNA. Inhibition of folate-dependent one-carbon metabolism by methotrexate has been an anticancer therapeutic for > 60 years<sup>54</sup>, but has toxic side effects<sup>55</sup>. Pemetrexed is indicated for the treatment of non-squamous NSCLC, but the molecular basis for differential efficacy is not fully understood<sup>56</sup>. Coincidentally, the folate signature C15 was associated with worse outcome in lung ADC, but not SCC. Other enzymes in C15 are annotated for nucleotide binding or metabolism, and/or mitochondrion localization. Functional interactions and concerted contributions to the cancer phenotype remain to be determined for the metabolism proteome clusters.

In conclusion, the integration of comprehensive data sets spanning the sequence-to-phenotype continuum DNA → RNA → proteome → disease uncovered proteome alterations and molecular signatures linking cancer metabolism and overall survival in lung and other cancers, which have not been predicted by genomics and transcriptomics alone. Genes encoding the metabolism proteome in NSCLC are subject to selective pressures that manifest (1) proteome remodelling conserved between primary and PDX tumours, and (2) DNA alterations with prognostic impact in NSCLC and other cancers. We suggest that the linking of proteome remodelling with DNA alterations, both products of selective pressure, may be a general approach to uncover and group together cryptic, polygenic cancer driver genes, which individually display only low-frequency mutation and/or copy number variation<sup>57</sup>. The analysis of proteome recapitulation as demonstrated herein suggests that the NSCLC PDX system might have utility for the development of anti-metabolism therapeutics.

## Methods

The protocols for use of surgically resected NSCLC samples for omic and PDX studies have been approved by the University Health Network Research Ethics Board and the Animal Care Committee.

**Proteome analysis.** Fresh tissue samples were placed in cryovials, were rapidly frozen with liquid nitrogen and were stored at  $-80^{\circ}\text{C}$ . Upon retrieval from the University Health Network Biobank, aliquots of tissue ( $\sim 50$  mg) were mixed with lysis buffer (1 ml buffer per 10 mg tissue; 20 mM HEPES, pH 8.0, 9 M urea, 1 mM sodium orthovanadate, 2.5 mM sodium pyrophosphate and 1 mM beta-glycerophosphate) and sonicated for 1 min, followed by centrifugation (20,000g) for 20 min (ref. 43) The concentration of clarified lysates was typically  $\sim 2$  mg ml<sup>-1</sup> (protein). Equal amounts of the lysate were reduced with dithiothreitol, alkylated with iodoacetamide, further diluted with (HEPES buffer) and then digested with trypsin overnight at  $23^{\circ}\text{C}$ . Peptides were then passed through a C-18 column and were eluted with 50% acetonitrile. The eluted peptides were dried under vacuum and stored at  $-80^{\circ}\text{C}$ . The digested peptides were loaded onto a  $75\ \mu\text{m}$  inside diameter  $\times 50$  cm ( $2\ \mu\text{m}$  C18) analytical column (EASY-Spray, Thermo Fisher Scientific, Odense Denmark). The peptides were eluted over 2 h at  $250\ \text{nl min}^{-1}$  with a 0–35% acetonitrile gradient in 0.1% formic acid by using an EASY nLC 1000 nano-chromatography system (Thermo Fisher Scientific). The eluted peptides were introduced by nano-electrospray into a LTQ Velos-Orbitrap Elite hybrid mass spectrometer (Thermo Fisher, Bremen, Germany) operated in a data-dependent mode. Mass spectra were acquired at 240,000 full-width at half-maximum resolution in the FTMS Orbitrap (with a target value of  $5 \times 10^5$  ions) and tandem MS (MS/MS) was carried out in the linear ion trap. Ten MS/MS scans were obtained per MS cycle using a target of  $1 \times 10^4$  ions and a maximum injection time of 50 ms. All ions passing the monoisotopic precursor selection filter were fragmented.

Raw MS files were analysed by MaxQuant<sup>14</sup> version 1.3.0.5. MS/MS spectra were searched by the Andromeda search engine<sup>58</sup> against the decoy Human SwissProt database (version 2013.1) containing forward and reverse sequences. In addition, the database included 248 common contaminants. MaxQuant analysis included an initial search with a precursor mass tolerance of 20 p.p.m., the results of which were used for mass recalibration. In the main Andromeda search precursor mass and fragment mass had an initial mass tolerance of 6 and 20 p.p.m., respectively. The search included variable modifications of methionine oxidation and N-terminal acetylation, and fixed modification of carbamidomethyl cysteine. Minimal peptide length was set to seven amino acids and a maximum of two miscleavages was allowed. The FDR was set to 0.01 for peptide and protein identifications. In the case of identified peptides that are all shared between two proteins, these are combined and are reported as one protein group. Proteins

**Table 3 | DNA alterations and prognostic impact associated with metabolism proteome clusters in various cancers.**

MPCI	Cancer type	Log-rank P value*	HR	Patients with DNA alterations <sup>†</sup>	Cluster size <sup>‡</sup>	DE proteins <sup>§</sup>	Cluster proteins (gene names) <sup>  </sup>
<b>C1</b>	<b>LSCC</b>	0.0055	0.50	63% 111	56	59%	<i>BLVRB, BPGM, CA1, CA2, CAT, CP, HAGH, LYZ, PONI, PRDX2, SLC4A1, SULT1A1, TFLCAT1, SLC27A3, HSDL2, ACOT2, PCBD1, PTGDS, SOD1, HPGD, ACAA2, CBR1, FN3KRP, PPAP2B, SLC25A11, ADH1B, AOC3, MAOB, PGM5, SOD3, ALDH2, DPYSL2, AK1, AQP1, CLIC6, CYB5A, CYBRD1, CYP27A1, ENTPD1, MARC2, SULT1A3, ABHD14B, ALDH3B1, ATP2B4, CLIC4, CYB5R1, CYB5R3, KCTD12, LIPA, LPCAT3, MPO, NPC2, PPT1, PTGIS, SLC34A2</i>
<b>C2</b>	<b>LSCC</b>	0.0053	0.51	53% 94	43	21%	<b>ATP13A1, GYS1, MBOAT7, PAPSS1, PRPSAP2</b> , AMPD2, LYPLA1, PGM3, SLC2A3, UGGT1, GALK1, KCNAB2, MOGS, TCIRG1, ACOX1, COX1, IAH1, MAT2B, MPI, <b>NSDHL, PFKP, RDH11, SLC25A13</b> , <i>IDH2, ADA, ALDH1B1, ALDOC, ECI2, GAA, GALK2, GUSB, H6PD, NAGA, NDUFA6, NDUFS1, NDUFV1, NUDT16, OXCT1, PCCA, PGK1, PLCG1, SHMT1, UGP2</i>
	<b>GBM</b>	0.0003	0.33	35% 31			
	<b>AML</b>	0.0406	1.64	18% 28			
<b>C3</b>	<b>LSCC</b>	0.0475	0.61	42% 73	6	0%	COASY, FAH, MGST1, NCEH1, OGDH, SLC44A1
<b>C4</b>	<b>LSCC</b>	0.0068	0.43	24% 43	6	0%	AGPAT9, CTH, GK, HGD, UGT1A6, SLC12A7
<b>C5</b>	<b>LSCC</b>	0.0439	0.54	24% 42	9	0%	ATP1B1, CYP2S1, GALNS, GAMT, MOCS3, NAPRT1, NOS2, NPR3, SLC12A2
<b>C6</b>	<b>LSCC</b>	0.0070	0.40	22% 40	11	0%	ACBD5, ACLY, ACOX3, CERS2, DHCR24, GPX8, LSS, MAN2A1, RRM2B, SLC35B2, ST3GAL1
	<b>BRCA</b>	0.0049	1.84	40% 293			
<b>C7</b>	<b>LSCC</b>	0.0208	0.46	21% 37	14	14%	<b>FDXR, IDH3G, SDHB</b> , CHID1, GNS, ITPR3, NDUFA13, NDUFA5, NDUFB9, NDUFS2, NDUFS3, NDUFS8, STT3B, NDUFS6
<b>C8</b>	<b>LSCC</b>	0.0138	0.36	13% 23	7	0%	DHRS2, DHRS9, FMO3, GFPT2, HSD17B13, PLCB3, RHCG
<b>C9</b>	<b>LSCC</b>	0.0166	2.13	13% 22	6	0%	AMPD3, ATP11A, CDA, CRAT, SDSL, TSTA3
	<b>OSC</b>	0.00116	0.67	41% 126			
<b>C10</b>	<b>LSCC</b>	0.0039	0.25	11% 20	7	100%	<b>ADSS, ATP2A2, CTPS1, IMPDH2, PKM2, PTGES3, SGPL1</b>
	<b>HNSC</b>	0.0103	2.47	11% 18			
<b>C11</b>	<b>LSCC</b>	0.0134	0.26	11% 19	6	0%	GSTK1, HIBADH, PPCS, SCCPDH, SLC25A24, SUCLG2
	<b>OSC</b>	0.0077	0.64	31% 95			
<b>C12</b>	<b>LSCC</b>	0.0293	0.34	10% 18	7	0%	CS, DCK, GPD2, <b>IDH1</b> , NDUFB7, NIT1, PPA1
<b>C13</b>	<b>LADC</b>	0.0089	2.76	73% 149	87	0%	GGT5, NDUFB4, PLCG2, TBXAS1, ACO2, BDH1, CA12, NDUFA10, NDUFA8, NDUFA9, ABCD1, ACADM, AGPAT1, AHCYL1, AKR1A1, ALDH16A1, ALDH3A2, ALOX15, APIP, ATP6VOA1, ATP6VOD1, ATP6VID, ATP6VIE1, CLC, CNDP2, COLGALT1, COX2, COX4I1, COX5A, COX5B, COX6A1, COX6B1, <b>COX6C</b> , COX7A2, CRYM, CRYZ, CYBA, DBT, DERA, EPX, FDX1, FOLR2, GALM, GBET, GCDH, GLA, GLOD4, GLUL, GNPAT, GPX1, GSTO1, GUK1, GYG1, HADH, HADHA, HADHB, HIBCH, IL4I1, IMPA1, LAP3, ME1, ME2, MICAL1, NAGK, NDUFA12, NDUFA4, NDUFA7, NDUFB10, NDUFB11, NDUFB5, NNMT, NPL, PAFAH1B2, PFKL, PGLS, PPA2, PPOX, PRDX1, PTGS1, RDH14, RENBP, SCP2, SDHA, SOD2, TAP1, TAP2, TYMP
<b>C14</b>	<b>LADC</b>	0.0175	2.06	13% 26	4	0%	GGT5, NDUFB4, PLCG2, TBXAS1
	<b>OSC</b>	0.0113	0.6	20% 62			
<b>C15</b>	<b>LADC</b>	0.0003	2.76	22% 45	13	100%	<b>NAT10, NME2, OAT, PPAT, SHMT2, GART, PAICS, SRM, UMPS, QARS, ABCE1, ABCF2, ACOT7</b>
<b>C16</b>	<b>LADC</b>	0.0059	2.06	32% 65	14	0%	FDFIT1, GCLC, IMPA2, NPC1, PSPH, ABCC5, ADI1, ALDH3A1, CD38, CYB5R2, ELOVL5, GPX3, GSTA1, GSTA4
C17	<b>LADC</b>	0.0354	2.01	12% 25	3	0%	AKR1B10, DLD, G6PD
C18	<b>LADC</b>	0.0084	0.28	16% 33	10	30%	ACOT2, <i>PCBD1, PTGDS</i> , HPGD, SOD1, ACAA2, CBR1, FN3KRP, PPAP2B, SLC25A11

AML, acute myeloid leukemia; BRCA, breast invasive carcinoma; DE, differentially expressed; GBM, glioblastoma multiforma; HNSC, head and neck squamous cell carcinoma; HR, hazard ratio; KIRP, kidney renal papillary cell carcinoma; LADC, lung adenocarcinoma; LSCC, lung squamous cell carcinoma; MPCI, metabolism-proteome cluster index; OSC, ovarian serous cystadenocarcinoma; TCGA, The Cancer Genome Atlas.

\*Cancer patient survival and DNA alteration data obtained from TCGA through the cBioPortal.

<sup>†</sup>Percent of cases (>10%) with DNA Alteration (mutation, deletion, amplification) and absolute number of affected cases.

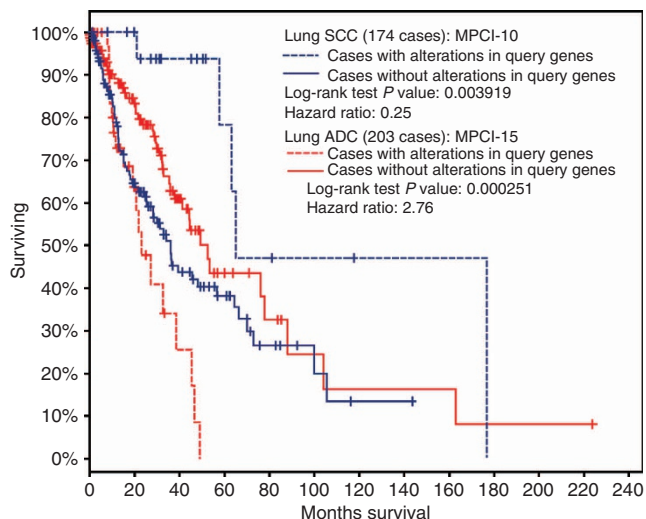
<sup>‡</sup>Number of proteins (>1).

<sup>§</sup>Percentage of differentially expressed proteins relative to normal lung; fold change >2, P<0.05.

<sup>||</sup>DE proteins are shown in bold (upregulated) or italics (downregulated). Cancer genes from the Catalogue of Somatic Mutations in Cancer database are underlined.

annotated as 'only identified by site', 'reverse' and 'contaminant' were filtered out by using Perseus tools within the MaxQuant environment. For comparison between samples, we used label-free quantification (LFQ) with a minimum of two ratio counts to determine the normalized protein intensity and retained the proteins quantified in at least two samples. Proteins without gene names (for

example, isoforms) were removed, as well as immunoglobulins, human leukocyte antigens and blood proteins. Finally, 4,030 proteins were quantified according to log<sub>2</sub> LFQ intensity. Zero protein values were filled with intensities from the lower part of a normal distribution (imputation width = 0.3 and shift = 1.8) by using Perseus. Complete identified and quantified protein tables are provided in



**Figure 5 | NSCLC subtype-specific prognostic impact of metabolism proteome clusters C10 and C15.** The dendrogram of metabolism proteins (Fig. 4c) was systematically tested for association of DNA alterations in the encoding genes. Kaplan-Meier plots relate survival of patients with (dotted line) and without (straight line) alterations. DNA alterations in genes encoding cluster C10 were found in 11% of lung SCC and associated with better overall survival (blue). DNA alterations associated with cluster C15 were found in 22% of lung ADC, and were associated with worse outcome (red). The total number of patients is shown in parenthesis. Patient data were obtained through cBioPortal. See also Supplementary Data 3 and 4. MPCl, metabolism-proteome cluster index, as in Table 3.

Supplementary Data 1 and 5. We note that proteome coverage, approaching 10,000 identifications, has been achieved in the analysis of colon cancer materials by using filter aided sample preparation (FASP) coupled with extensive off-line peptide fractionation before liquid chromatography tandem MS<sup>59–61</sup>.

Mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (<http://www.proteomexchange.org>) via the PRIDE partner repository<sup>62</sup> with the data set identifier PXD000853.

**Genome and transcriptome analysis.** Illumina platforms were used for DNA copy number (Human Omni1 SNP) and mRNA analyses (whole-genome DASL HT)<sup>63–65</sup>. The Illumina GenomeStudio (version 20120.1) genotyping module (version 1.6.3) was used to calculate B Allele Frequency (BAF) values and the Log R ratio (LRR) for each probe, and copy number alterations were inferred from these values by using the R-Gada package<sup>65</sup>. Microarray data were analyzed and normalized by using the LUMI package of R. If a gene was measured using multiple probe sets in the microarray, the one with largest variance across the samples was selected to present the expression of this gene<sup>66</sup>.

**Integrated Omics polygon array.** The array of polygons (hexagons) was generated by using an algorithm and perl script termed *Visygon* for visualization of polygons, wherein the 'GD' graphics library was used to draw graphics. The integrated DNA/RNA/protein data were arranged as a matrix sorted by gene location chromosome and by tumour type, that is, primary tumour followed by PDX, and sorted by histology subtype. The matrix was pre-processed by comparison of tumour data with matched normal lung and by application of thresholds that were used to define differential values: 1, upregulation; -1, downregulation; and 0, no change. After reading the matrix, the script was used to generate hexagon arrays and with the six quadrants of each hexagon coloured according to the values: 1, red; 1, blue; 0, grey; no data, white. The omic array was generated for individual chromosomes in gene order according to nucleotide number, and with a space added to denote centromere location separating p and q arms for clarity.

**Statistical analysis.** Statistical analysis was conducted by using R (<http://www.R-project.org>). A paired moderated *t*-test in the Limma package was applied for testing of differences in LFQ protein and gene intensities in clinical samples with the threshold value FDR (that is, adjusted *P* value) of 0.05 and fold change cutoff as 2.

**Testing proteome clusters for prognostic impact.** Clinical annotation of patient cohorts was retrieved from TCGA through cBioPortal for Cancer Genomics<sup>29</sup> ([www.cbioportal.org](http://www.cbioportal.org)). We included patients that were screened for DNA alterations and excluded patients with missing values. For each tumour type, we

iteratively searched with clusters of genes, defined by each node of the clustering dendrogram, for patients bearing a DNA alteration in one or more of the genes. The search progressed through each node of the dendrogram from the entire set of 838 down to individual proteins/genes (Supplementary Data 3). The clinical signals (log-rank *P* value for overall survival) for the cohorts of patients with and without DNA alterations, for which we arbitrarily set a lower limit at 10%, were compared and a likelihood ratio test was performed. Nodes with a global or local minimal likelihood ratio test *P* value < 0.05 were initially selected for further analysis. Local minimum was defined as a node wherein contiguous nodes (that is, having more or fewer proteins) had relatively higher *P* values. The statistical significance of initially selected groups was assessed by using an internal validation algorithm to evaluate the non-random association between clusters of a given number of proteins and clinical outcome. For each cluster,  $1 \times 10^4$  mock clusters of the same number of proteins from the 838 detected metabolism proteins were randomly generated and their log-rank *P* values were calculated for each of the indicated cancer types. Only clusters for which the log-rank *P* value was in the top fifth percentile (0.05) of the corresponding mock clusters were accepted.

## References

- Parmigiani, G. *et al.* Design and analysis issues in genome-wide somatic mutation studies of cancer. *Genomics* **93**, 17–21 (2009).
- Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
- Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
- TCGA. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
- Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat. Med.* **10**, 789–799 (2004).
- Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
- John, T. *et al.* The ability to form primary tumor xenografts is predictive of increased risk of disease recurrence in early-stage non-small cell lung cancer. *Clin. Cancer Res.* **17**, 134–141 (2011).
- DeRose, Y. S. *et al.* Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nat. Med.* **17**, 1514–1520 (2011).
- Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).
- Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
- Nemati, F. *et al.* Establishment and characterization of a panel of human uveal melanoma xenografts derived from primary and/or metastatic tumors. *Clin. Cancer Res.* **16**, 2352–2362 (2010).
- Tentler, J. J. *et al.* Patient-derived tumour xenografts as models for oncology drug development. *Nat. Rev. Clin. Oncol.* **9**, 338–350 (2012).
- Michalski, A. *et al.* Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol. Cell. Proteomics* **11**, O111 013698 (2012).
- Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
- Kikuchi, T. *et al.* In-depth proteomic analysis of nonsmall cell lung cancer to discover molecular targets and candidate biomarkers. *Mol. Cell. Proteomics* **11**, 916–932 (2012).
- Chen, G. *et al.* Discordant protein and mRNA expression in lung adenocarcinomas. *Mol. Cell. Proteomics* **1**, 304–313 (2002).
- Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).
- Franovic, A. *et al.* Translational up-regulation of the EGFR by tumor hypoxia provides a nonmutational explanation for its overexpression in human cancer. *Proc. Natl Acad. Sci. USA* **104**, 13092–13097 (2007).
- Mosesson, Y., Mills, G. B. & Yarden, Y. Derailed endocytosis: an emerging feature of cancer. *Nat. Rev. Cancer* **8**, 835–850 (2008).
- Kalisman, N., Adams, C. M. & Levitt, M. Subunit order of eukaryotic TRiC/CCT chaperonin by cross-linking, mass spectrometry, and combinatorial homology modeling. *Proc. Natl Acad. Sci. USA* **109**, 2884–2889 (2012).
- Leitner, A. *et al.* The molecular architecture of the eukaryotic chaperonin TRiC/CCT. *Structure* **20**, 814–825 (2012).
- Hartl, F. U., Bracher, A. & Hayer-Hartl, M. Molecular chaperones in protein folding and proteostasis. *Nature* **475**, 324–332 (2011).
- Sternlicht, H. *et al.* The t-complex polypeptide 1 complex is a chaperonin for tubulin and actin in vivo. *Proc. Natl Acad. Sci. USA* **90**, 9422–9426 (1993).
- Hansen, W. J. *et al.* Diverse effects of mutations in exon II of the von Hippel-Lindau (VHL) tumor suppressor gene on the interaction of pVHL with the cytosolic chaperonin and pVHL-dependent ubiquitin ligase activity. *Mol. Cell. Biol.* **22**, 1947–1960 (2002).



25. Melki, R., Batelier, G., Soulie, S. & Williams, Jr. R. C. Cytoplasmic chaperonin containing TCP-1: structural and functional characterization. *Biochemistry* **36**, 5817–5826 (1997).
26. Won, K. A., Schumacher, R. J., Farr, G. W., Horwich, A. L. & Reed, S. I. Maturation of human cyclin E requires the function of eukaryotic chaperonin CCT. *Mol. Cell. Biol.* **18**, 7584–7589 (1998).
27. Boutros, P. C. *et al.* Prognostic gene signatures for non-small-cell lung cancer. *Proc. Natl Acad. Sci. USA* **106**, 2824–2828 (2009).
28. Coghlin, C. *et al.* Characterization and over-expression of chaperonin t-complex proteins in colorectal cancer. *J. Pathol.* **210**, 351–357 (2006).
29. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
30. Seo, M., Lee, W. H. & Suk, K. Identification of novel cell migration-promoting genes by a functional genetic screen. *FASEB J.* **24**, 464–478 (2010).
31. Baughman, J. M. *et al.* A computational screen for regulators of oxidative phosphorylation implicates SLIRP in mitochondrial RNA homeostasis. *PLoS Genet.* **5**, e1000590 (2009).
32. Aras, S. *et al.* Oxygen-dependent expression of cytochrome c oxidase subunit 4-2 gene expression is mediated by transcription factors RBPJ, CXXC5 and CHCHD2. *Nucleic Acids Res.* **41**, 2255–2266 (2013).
33. Wei, Y. *et al.* CHCHD2 is co-amplified with EGFR in non-small cell lung carcinoma and regulates mitochondrial function and cell migration. (submitted), (2014).
34. Parker, J. J. *et al.* Gefitinib selectively inhibits tumor cell migration in EGFR-amplified human glioblastoma. *Neuro Oncol.* **15**, 1048–1057 (2013).
35. Wikman, H. *et al.* CDK4 is a probable target gene in a novel amplicon at 12q13.3-q14.1 in lung cancer. *Genes Chromosomes Cancer* **42**, 193–199 (2005).
36. Anderson, D. D. & Stover, P. J. SHMT1 and SHMT2 are functionally redundant in nuclear de novo thymidylate biosynthesis. *PLoS ONE* **4**, e5839 (2009).
37. Anderson, D. D., Woeller, C. F., Chiang, E. P., Shane, B. & Stover, P. J. Serine hydroxymethyltransferase anchors de novo thymidylate synthesis pathway to nuclear lamina for DNA synthesis. *J. Biol. Chem.* **287**, 7051–7062 (2012).
38. Pavon-Eternod, M., Gomes, S., Rosner, M. R. & Pan, T. Overexpression of initiator methionine tRNA leads to global reprogramming of tRNA expression and increased proliferation in human epithelial cells. *RNA* **19**, 461–466 (2013).
39. Zhang, W. C. *et al.* Glycine decarboxylase activity drives non-small cell lung cancer tumor-initiating cells and tumorigenesis. *Cell* **148**, 259–272 (2012).
40. Possemato, R. *et al.* Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. *Nature* **476**, 346–350 (2011).
41. Vie, N. *et al.* Overexpression of phosphoserine aminotransferase PSAT1 stimulates cell growth and increases chemoresistance of colon cancer cells. *Mol. Cancer* **7**, 14 (2008).
42. Locasale, J. W. *et al.* Phosphoglycerate dehydrogenase diverts glycolytic flux and contributes to oncogenesis. *Nat. Genet.* **43**, 869–874 (2011).
43. Wei, Y. *et al.* Primary tumor xenografts of human lung adeno and squamous cell carcinoma express distinct proteomic signatures. *J. Proteome Res.* **10**, 161–174 (2011).
44. Zhang, W. *et al.* Proteomic profiles of human lung adeno and squamous cell carcinoma using super-SILAC and label-free quantification approaches. *Proteomics* **14**, 795–803 (2014).
45. Nilsson, R. *et al.* Metabolic enzyme expression highlights a key role for MTHFD2 and the mitochondrial folate pathway in cancer. *Nat. Commun.* **5**, 3128 (2014).
46. Lee, G. Y. *et al.* Comparative oncogenomics identifies PSMB4 and SHMT2 as potential cancer driver genes. *Cancer Res.* **74**, 3114–3126 (2014).
47. Ye, J. *et al.* Serine catabolism regulates mitochondrial redox control during hypoxia. *Cancer Discov.* doi:10.1158/2159-8290.CD-14-0250 (2014).
48. Zheng, H. *et al.* A BRISC-SHMT complex deubiquitinates IFNAR1 and regulates interferon responses. *Cell Rep.* **5**, 180–193 (2013).
49. Vander Heiden, M. G. *et al.* Metabolic pathway alterations that support cell proliferation. *Cold Spring Harb. Symp. Quant. Biol.* **76**, 325–334 (2011).
50. Christofk, H. R. *et al.* The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature* **452**, 230–233 (2008).
51. Toledo-Sherman, L. *et al.* New targets for an old drug: A chemical proteomics approach to unraveling the molecular mechanism of action of methotrexate. *Clin. Proteomics* **1**, 45–67 (2004).
52. Daidone, F. *et al.* In silico and in vitro validation of serine hydroxymethyltransferase as a chemotherapeutic target of the antifolate drug pemetrexed. *Eur. J. Med. Chem.* **46**, 1616–1621 (2011).
53. Hebbbring, S. J. *et al.* Serine hydroxymethyltransferase 1 and 2: gene sequence variation and functional genomic characterization. *J. Neurochem.* **120**, 881–890 (2012).
54. Meyer, L. M., Miller, F. R., Rowen, M. J., Bock, G. & Rutzky, J. Treatment of acute leukemia with amethopterin (4-amino, 10-methyl pteroyl glutamic acid). *Acta Haematol.* **4**, 157–167 (1950).
55. Abolmaali, S. S., Tamaddon, A. M. & Dinarvand, R. A review of therapeutic challenges and achievements of methotrexate delivery systems for treatment of cancer and rheumatoid arthritis. *Cancer Chemother. Pharmacol.* **71**, 1115–1130 (2013).
56. Scagliotti, G. *et al.* Treatment-by-histology interaction analyses in three phase III trials show superiority of pemetrexed in nonsquamous non-small cell lung cancer. *J. Thorac. Oncol.* **6**, 64–70 (2011).
57. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
58. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
59. Wisniewski, J. R. *et al.* Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol. Syst. Biol.* **8**, 611 (2012).
60. Wisniewski, J. R., Ostasiewicz, P. & Mann, M. High recovery FASP applied to the proteomic analysis of microdissected formalin fixed paraffin embedded cancer tissues retrieves known colon cancer markers. *J. Proteome Res.* **10**, 3040–3049 (2011).
61. Zougman, A. *et al.* Retinoic acid-induced protein 3: identification and characterisation of a novel prognostic colon cancer biomarker. *Eur. J. Cancer* **49**, 531–539 (2013).
62. Vizcaino, J. A. *et al.* The PRoteomics IDentifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41**, D1063–D1069 (2013).
63. Diskin, S. J. *et al.* Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* **36**, e126 (2008).
64. Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547–1548 (2008).
65. Pique-Regi, R., Caceres, A. & Gonzalez, J. R. R-Gada: a fast and flexible pipeline for copy number analysis in association studies. *BMC bioinformatics* **11**, 380 (2010).
66. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
67. Beroukhi, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. USA* **104**, 20007–20012 (2007).
68. Beroukhi, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).

## Acknowledgements

This research was supported in part by the Canada Research Chairs Program (T.K. and M.F.M.), Ontario Research Fund-Research Excellence Grant (RE-03-020), Canadian Institutes for Health Research (CIHR) (MOP-102536), Canadian Cancer Society Research Institute and Ontario Ministry of Long Term Health. M.S.T. is the M. Oasim Choksi Chair in Lung Cancer Translational Research. F.A.S. is the Scott Taylor Chair in Lung Cancer Research. G.L. is the Alan B. Brown Chair of Molecular Genomics and Cancer Care Ontario Chair in Experiment Therapeutics and Population Studies. Y.W. was supported by Canadian Institutes of Health Research Fellowship and N.Y. was supported by the Terry Fox Foundation Strategic Training Initiative in Health Research (STIHR) grant in Molecular Pathology of Cancer at CIHR (STP-53912). We are most grateful to The Princess Margaret Cancer Foundation and TCGA Research Network, and their generous specimen donors and the Computational Biology Centre at Memorial Sloan-Kettering Cancer Center for providing access to their comprehensive data sets and tools for their analysis and visualization.

## Author contributions

M.S.T., F.A.S., G.L., T.K. and M.F.M. devised the experiments; Y.W., N.-A.P., J.T., C.T., L.M., N.Y. and P.T. collected and organized analytical data; M.F.M., L.L., Y.W., V.I., A.I., P.T., C.T., L.M., M.P., C.-Q.Z., W.Z., M.S.T. and T.K. completed data processing and analysis; and M.F.M. wrote the paper.

## Additional information

**Accession code:** Gene expression data have been deposited in the Gene Expression Omnibus (GEO) under the accession code GSE62113.

**Supplementary Information** accompanies this paper at <http://www.nature.com/naturecommunications>

**Competing financial interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**How to cite this article:** Li, L. *et al.* Integrated Omic analysis of lung cancer reveals metabolism proteome signatures with prognostic impact. *Nat. Commun.* 5:5469 doi: 10.1038/ncomms6469 (2014).