ARTICLE

doi:10.1038/nature13438

Proteogenomic characterization of human colon and rectal cancer

Bing Zhang[1,2], Jing Wang[1], Xiaojing Wang[1], Jing Zhu[1], Qi Liu[1], Zhiao Shi[3,4], Matthew C. Chambers[1], Lisa J. Zimmerman[5,6], Kent F. Shaddox[6], Sangtae Kim[7], Sherri R. Davies[8], Sean Wang[9], Pei Wang[10], Christopher R. Kinsinger[11], Robert C. Rivers[11], Henry Rodriguez[11], R. Reid Townsend[8], Matthew J. C. Ellis[8], Steven A. Carr[12], David L. Tabb[1], Robert J. Coffey[13], Robbert J. C. Slebos[2,6], Daniel C. Liebler[5,6] & the NCI CPTAC*

Extensive genomic characterization of human cancers presents the problem of inference from genomic abnormalities to cancer phenotypes. To address this problem, we analysed proteomes of colon and rectal tumours characterized previously by The Cancer Genome Atlas (TCGA) and perform integrated proteogenomic analyses. Somatic variants displayed reduced protein abundance compared to germline variants. Messenger RNA transcript abundance did not reliably predict protein abundance differences between tumours. Proteomics identified five proteomic subtypes in the TCGA cohort, two of which overlapped with the TCGA 'microsatellite instability/CpG island methylation phenotype' transcriptomic subtype, but had distinct mutation, methylation and protein expression patterns associated with different clinical outcomes. Although copy number alterations showed strong *cis*- and *trans*-effects on mRNA abundance, relatively few of these extend to the protein level. Thus, proteomics data enabled prioritization of candidate driver genes. The chromosome 20q amplicon was associated with the largest global changes at both mRNA and protein levels; proteomics data highlighted potential 20q candidates, including *HNF4A* (hepatocyte nuclear factor 4, alpha), *TOMM34* (translocase of outer mitochondrial membrane 34) and *SRC* (SRC proto-oncogene, non-receptor tyrosine kinase). Integrated proteogenomic analysis provides functional context to interpret genomic abnormalities and affords a new paradigm for understanding cancer biology.

TCGA has characterized the genomic features of human cancers[1–6] and this has presented a new challenge of explaining how genomic alterations drive cancers[7]. As proteins link genotypes to phenotypes, the Clinical Proteomic Tumour Analysis Consortium (CPTAC) is performing proteomic analyses of TCGA tumour specimens for selected cancer types. Here we present the first integrated proteogenomic characterization of human cancer with an analysis of the TCGA colorectal cancer (CRC) specimens[6].

The TCGA study affirmed well-established genomic features of CRC and described three transcriptional subtypes, 17 chromosomal regions of significant focal amplification and 28 regions of significant focal deletion, and linked genomic features of CRC to critical signalling pathways. The drivers underlying copy number alterations (CNAs) and transcriptional subtypes are largely unknown, and an integrative analysis of both genomic and proteomic data may provide a more comprehensive understanding of the information flow from DNA to protein to phenotype.

Peptide and protein identification

We performed liquid chromatography–tandem mass spectrometry (LC-MS/MS)-based shotgun proteomic analyses on 95 TCGA tumour samples (Extended Data Fig. 1 and Methods), the clinical and pathological characteristics and TCGA data sets for which are summarized in Supplementary Table 1. Benchmark quality control samples from one basal and one luminal human breast tumour xenograft were analysed in alternating order after every five CRC samples (Methods).

We identified a total of 124,823 distinct peptides among the 95 samples, corresponding to 6,299,756 spectra in an assembly of 7,526 protein groups with a protein-level false discovery rate (FDR) of 2.64% (Methods and Extended Data Fig. 2). To facilitate integration between genomic and proteomic data, a gene-level assembly of the peptides identified 7,211 genes.

A fundamental question in proteogenomics is which protein coding alterations are expressed at the protein level. As standard database search approaches cannot identify variant peptides from MS/MS data, we also performed database searches with customized sequence databases from matched RNA sequencing (RNA-seq) data for individual samples[8,9] (Methods and Extended Data Fig. 3).

We identified 796 single amino acid variants (SAAVs) across all 86 tumours for which matched RNA-seq data were available (Fig. 1a, b and Supplementary Tables 2 and 3), among which 64 corresponded to somatic variants reported by TCGA and 101 were reported in the COSMIC database (that is, COSMIC-supported variants). Of the remaining SAAVs, 526 were listed in the Single Nucleotide Polymorphism database (dbSNP) (that is, dbSNP-supported variants) and are likely to be germline variants. The 162 previously unreported SAAVs might be explained by novel somatic or germline variants, RNA editing, or, in some cases, false discovery.

The identified somatic variants were clearly enriched in the hypermutated samples, whereas the germline variants showed no association with hypermutation (Fig. 1a). Although 58% of the germline variants occurred in two or more samples, almost all somatic variants occurred

[1]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. [2]Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. [3]Advanced Computing Center for Research and Education, Vanderbilt University, Nashville, Tennessee 37232, USA. [4]Department of Electrical Engineering and Computer Science, Vanderbilt University, Tennessee 37232, USA. [5]Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. [6]Jim Ayers Institute for Precancer Detection and Diagnosis, Vanderbilt-Ingram Cancer Center, Nashville, Tennessee 37232, USA. [7]Directorate of Fundamental and Computational Sciences, Pacific Northwest National Laboratory, Richland, Washington 99352, USA. [8]Department of Internal Medicine, Washington University School of Medicine, St. Louis, Missouri 63110, USA. [9]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2-B500, Seattle, Washington 98109, USA. [10]Department of Genetics and Genomic Sciences, Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1498, New York, New York 10029, USA. [11]Office of Cancer Clinical Proteomics Research, National Cancer Institute, Bethesda, Maryland 20892, USA. [12]Broad Institute of MIT and Harvard, Cambridge, Maryland 02142, USA. [13]Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA.
*Lists of participants and their affiliations appear at the end of the paper.

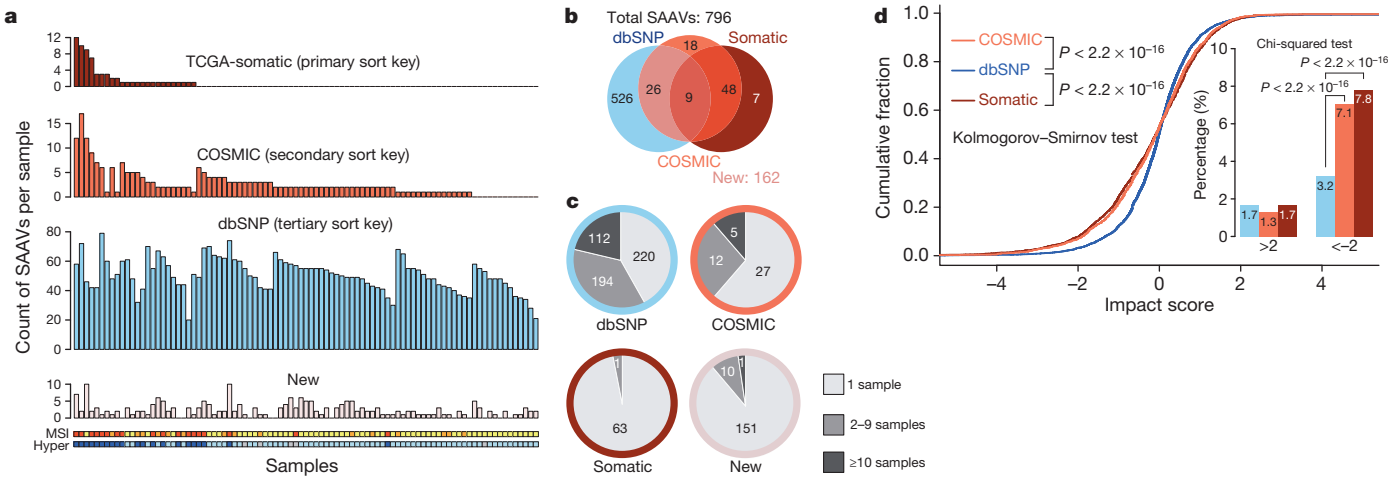


Figure 1 | Summary of detected single amino acid variants (SAAVs) and the impact of single nucleotide variants (SNVs) on protein abundance. **a**, The number of different types of SAAVs (TCGA-reported somatic variants, COSMIC-supported variants, dbSNP-supported variants and new variants) in individual tumour samples. The samples are ordered by the number of detected somatic variants, then COSMIC-supported variants, and then dbSNP-supported variants. The MSI and hypermutation (Hyper) status are labelled below the bar charts for each sample (red, MSI-high; orange, MSI-low; yellow, microsatellite stable; blue, hypermutated; light blue, non-hypermutated; grey, no data). The number of somatic variants and COSMIC-supported variants were significantly higher in MSI-high and hypermutated tumours, whereas the other two types of SAAVs were randomly distributed across the data set. **b**, The total numbers for different types of SAAVs and their overlapping relations. All 796 detected SAAVs were annotated based on previous reports in dbSNP (left circle), COSMIC (middle circle) or TCGA-reported somatic variants (right circle), and their overlapping relations are shown in the Venn diagram. There are 162 SAAVs that have not been reported previously in these databases (new). **c**, Distribution of the frequency of occurrence for different types of SAAVs. Border colours of the pie charts correspond to different SAAV types using the same colour scheme as in **a**. Whereas 58% of dbSNP-supported variants occurred in two or more samples, almost all somatic variants occurred in only one sample each. **d**, SNVs detected in RNA-seq data were separated into three categories (dbSNP-supported, COSMIC-supported and TCGA-somatic). The impact of individual SNVs on protein abundance was calculated (see Methods) and the impact scores for different categories of SNVs were plotted as cumulative fraction curves with two-sided *P* values from the Kolmogorov–Smirnov test labelled. The percentage of SNVs with an absolute impact score greater than 2 was also plotted as an inset, with *P* values from the Chi-squared test. Sample size for the dbSNP-supported, COSMIC-supported and TCGA-somatic variants were 1,2184, 7,492 and 3,302, respectively.

in only one sample (Fig. 1c). The low identification rate for somatic variants may reflect relatively low sequence coverage in shotgun proteomics; however, somatic variants also might negatively impact protein abundance, possibly by reducing translational efficiency or protein stability[10]. Using the protein abundance quantification method described below and detailed in the Methods, we found that somatic variants exerted a significantly stronger negative impact on protein abundance than did dbSNP-supported variants ($P < 2.2 \times 10^{-16}$, Kolmogorov–Smirnov test; Fig. 1d and Methods). The percentage of variants with an impact score of less than −2 was doubled for somatic variants compared to dbSNP-supported variants ($P < 2.2 \times 10^{-16}$, Chi-squared test; Fig. 1d).

Cancer-related variant proteins may serve as candidate protein biomarkers or therapeutic targets. The 108 somatic or COSMIC-supported protein variants mapped to 105 genes, including known cancer genes in the Cancer Gene Census database such as *KRAS*, *CTNNB1*, *SF3B1*, *ALDH2* and *FH*. The list also included 14 targets of FDA-approved drugs or drugs in clinical trials[4], such as *ALDH2*, *HSD17B4*, *PARP1*, *P4HB*, *TST*, *GAK*, *SLC25A24* and *SUPT16H*. A subset of variant peptide sequences, including *KRAS*(Gly12Asp) were verified by targeted analyses of tumour lysates spiked with synthetic, isotope-labelled peptide standards (Methods). One example is shown in Extended Data Fig. 4.

Quantification of protein abundance

To quantify protein abundance, we used spectral counts, which are the total number of MS/MS spectra acquired for peptides from a given protein[11] (Methods and Supplementary Table 4). Analysis of data from benchmark quality control samples demonstrated platform reproducibility throughout the analyses and enabled evaluation of data normalization methods (Extended Data Fig. 5a, b). Based on the minimal spectral count requirement established using the quality control data set (Extended Data Fig. 5c), 3,899 genes with a protein-level FDR of 0.43% were used to compare relative protein abundance across tumour samples.

mRNA versus protein abundance

The matched proteomic and RNA-seq measurements from the TCGA CRC tumours enabled the first global analysis of transcript–protein relationships in a large human tumour cohort (Methods). First, we compared the steady state mRNA and protein abundance for each gene within individual samples (Methods and Extended Data Fig. 6a). All samples showed significant positive mRNA–protein correlation (multiple-test adjusted $P < 0.01$, Spearman's correlation coefficient) and the average correlation between steady state mRNA and protein abundance in individual samples was 0.47 (Fig. 2a), which is comparable to previous reports in multi-cellular organisms[12].

Next, we examined the concordance between mRNA and protein variation of individual genes across the 87 tumours for which 3,764 genes had both mRNA and protein measurements suitable for relative abundance comparison (Methods). Although 89% of the genes showed a positive mRNA–protein correlation, only 32% had statistically significant correlations (Fig. 2b). The average Spearman's correlation between mRNA and protein variation was 0.23, which was comparable to reported values for yeast, mouse and human cell lines[13–15].

To test whether the concordance between protein and mRNA variation is related to the biological function of the gene product, we performed KEGG (Kyoto Encyclopedia of Genes and Genomes) enrichment analysis (Methods and Supplementary Table 5). Genes involved in several metabolic processes showed concordant mRNA and protein variation, whereas other gene classes showed low or even negative concordance in mRNA and protein variation (Fig. 2c). We also found that genes with stable mRNA and stable protein tend to have higher mRNA–protein correlation than those with unstable mRNA and unstable protein ($P = 5.27 \times 10^{-6}$, two-sided Wilcoxon rank-sum test, Methods, Extended Data Fig. 6b). Thus, mRNA measurements are poor predictors of protein abundance variations and both biological functions of the gene products and mRNA and protein stability may govern mRNA–protein correlation.

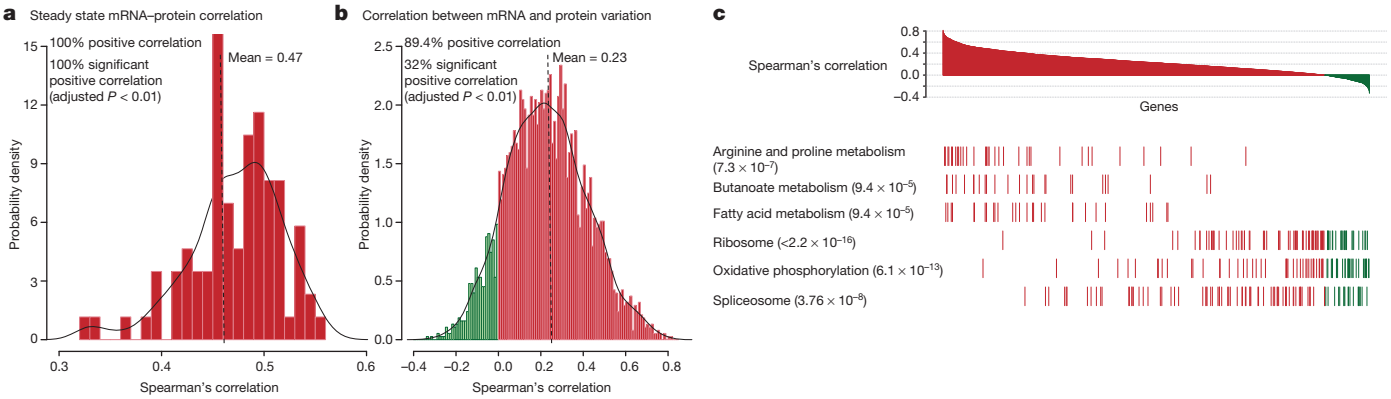


Figure 2 | Correlations between mRNA and protein abundance in TCGA tumours. **a**, Steady state mRNA and protein abundance were positively correlated in all 86 samples (multiple-test adjusted $P < 0.01$) with a mean Spearman's correlation coefficient of 0.47. **b**, mRNA and protein variation were positively correlated for most (89.4%) mRNA-protein pairs across the 87 samples, but only 32% showed significant correlation (multiple-test adjusted $P < 0.01$), with a mean Spearman's correlation coefficient of 0.23. **c**, mRNA and protein levels displayed dramatically different correlation for genes involved in different biological processes. Genes encoding intermediary metabolism functions showed high mRNA–protein correlations, whereas genes involved in oxidative phosphorylation, RNA splicing and ribosome components showed low or negative correlations. Multiple-test adjusted two-sided *P* values from the Kolmogorov–Smirnov test were provided in the parentheses following the KEGG pathway names. Red and green in the figures indicate positive and negative correlations, respectively.

Impact of copy number alterations

The study by TCGA identified 17 regions of significant focal amplification and 28 regions of significant focal deletion. We examined the impact of CNAs on mRNA and protein abundance, including both *cis*-effects on the abundance of genes in the same loci and *trans*-effects on the abundance of genes at other loci in the genome (Methods).

For all 23,125 genes with a CNA measurement in the TCGA data set, we calculated Spearman's correlation with mRNA and protein abundance, respectively for the 3,764 genes with both mRNA and protein measurements (Methods). Examination of the matrix visualizing significant CNA–mRNA correlations (multiple-test adjusted $P < 0.01$) revealed strong positive correlations along the diagonal (Fig. 3a), suggesting strong *cis*-effects of CNAs on mRNA abundance. Most of the diagonal signals corresponded to previously reported arm-level changes[6]. In contrast, the diagonal pattern was much weaker for CNA–protein correlations (Fig. 3b).

To investigate further the *cis*-effects of CNAs, we separated all genes with CNA, mRNA and protein measurements into those in focal amplification regions, focal deletion regions and non-focal regions (that is, chromosomal regions without focal amplification or deletion). As shown in Extended Data Fig. 7, CNA–mRNA correlations were significantly higher than CNA–protein correlations for genes in all three groups ($P < 1.0 \times 10^{-10}$, Kolmogorov–Smirnov test). Moreover, genes in the focal amplification regions showed significantly higher CNA–mRNA and CNA–protein correlations than genes in the non-focal regions ($P = 4.4 \times 10^{-4}$ and 0.02, respectively, Kolmogorov–Smirnov test). However, the same trend was not observed for genes in the focal deletion regions. Therefore, focal amplifications have the strongest *cis*-effects on both mRNA and protein abundance, suggesting that selection for high protein abundance may drive CNA in regions of focal amplification. However, many CNA-driven mRNA level increases do not translate into increased abundance of the corresponding proteins.

Figure 3a, b also revealed multiple *trans*-acting CNA hot spots, defined as chromosomal loci whose alteration is significantly associated with abundance changes of many transcripts or proteins at other loci. Chromosomes 20q, 18, 16, 13 and 7 contained the five strongest hot spots driving global mRNA abundance variation. These hot spots also were strongest at the protein level. Most hot-spot-related transcript changes did not propagate to the protein level, presumably reflecting buffering of protein abundance by post-transcriptional regulation[16,17]. Notably, many hot-spot-associated protein-level alterations occurred in the absence of corresponding mRNA alterations, suggesting that the same *trans*-acting hot spot may exert independent effects at both the transcriptome and proteome levels.

The 20q amplification was associated with the largest global changes in both mRNA and protein levels in this univariate analysis. The same conclusion was reached with a regularized multivariate regression analysis method, remMap[18] (Methods and Supplementary Tables 6–9). These

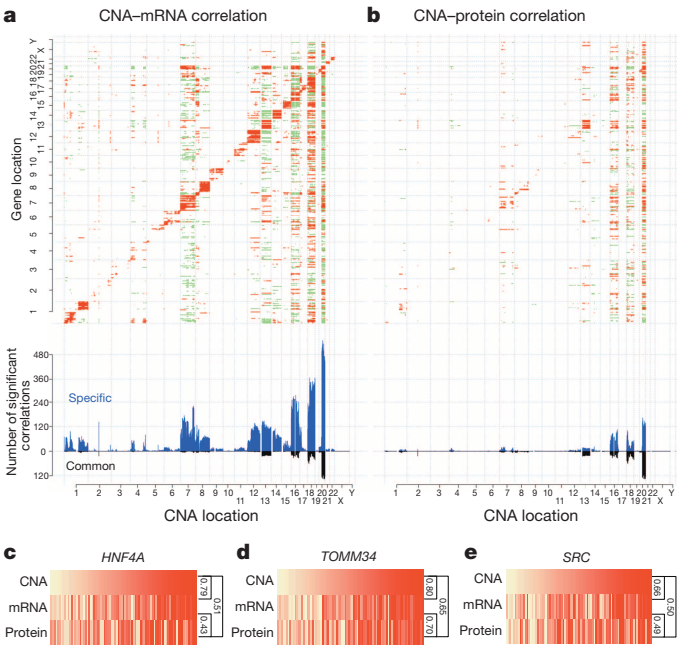


Figure 3 | Effects of copy number alterations on mRNA and protein abundance. **a**, **b**, The top panels show copy-number-abundance correlation matrices for mRNA abundance (**a**) and protein abundance (**b**) with significant positive and negative correlations (multiple-test adjusted $P < 0.01$, Spearman's correlation coefficient) indicated by red and green, respectively, and genes ordered by chromosomal location on both *x* and *y* axes. The bottom panels show the frequency of mRNAs and proteins associated with a particular copy number alteration, where blue and black bars represent associations specific to mRNA and protein or common to both mRNA and protein, respectively. **c**–**e**, *HNF4A* (**c**), *TOMM34* (**d**) and *SRC* (**e**) showed significant CNA–mRNA, mRNA–protein, and CNA–protein correlations (Spearman's correlation coefficient). The colour grade from light yellow to red indicates relatively low-level (yellow) to high-level (red) of copy number, mRNA abundance, or relative protein abundance among the 85 samples, which were ordered by copy number data.

data highlight the importance of 20q amplification in CRC, which has not been well documented in previous studies. Among the 79 genes in the 20q region with quantifiable protein measurements, 67 (85%) showed significant CNA–mRNA correlation, but only 40 (51%) showed significant CNA–protein correlation (multiple-test adjusted $P < 0.01$, Spearman's correlation coefficient, Supplementary Table 10).

As significant CNA–protein correlations identify amplified sequences that translate to high protein abundance, proteomic measurements can help prioritize genes in amplified regions for further examination. Of particular interest among the 40 genes is *HNF4A* (Fig. 3c), a candidate driver gene nominated by TCGA for the 20q13.12 focal amplification peak[6]. HNF4α is a transcription factor with a key role in normal gastrointestinal development[19] and is increasingly being linked to CRC[20]. However, there are contradictory reports on whether HNF4α acts as an oncogene or a tumour suppressor gene in CRC[20]. Upon reanalysis of the *HNF4A* short hairpin RNA (shRNA) knockdown data for CRC cell lines from the Achilles project[21], we found that the dependency of CRC cells on HNF4α correlated significantly with the amplification level of *HNF4A* (Methods and Extended Data Fig. 8), which may partially explain the contradictory roles reported for HNF4α in CRC. Other interesting candidates included *TOMM34* (Fig. 3d), which is overexpressed frequently in CRC tumours and is involved in the growth of CRC cells[22], and *SRC* (Fig. 3e), which encodes a non-receptor tyrosine kinase implicated in several human cancers including CRC[23].

Proteomic subtypes of CRC

The TCGA study reported three transcriptomic subtypes of CRC, designated 'microsatellite instability/CpG island methylator phenotype' (MSI/CIMP), 'invasive', and 'chromosomal instability' (CIN). Given the limited correlation between mRNA and protein levels, we asked whether CRC subtypes can be better represented with proteomics data. Using the consensus clustering[24] method (Methods and Extended Data Fig. 9), we identified five major proteomic subtypes in this tumour cohort, with 15, 9, 25, 11 and 19 cases in subtypes A to E, respectively (Fig. 4a).

We tested the association between the subtype classification and established genomic and epigenomic features of CRC using Fisher's exact test (Fig. 4b and Supplementary Table 11). Almost all hypermutated and MSI-high tumours were included in subtypes B and C, as well as tumours with *POLE* and *BRAF* mutations. However, statistically significant association with these features was only observed for subtype B (multiple-test adjusted $P < 0.05$). Moreover, subtype B was significantly associated with the TCGA CIMP -high methylation subtype, whereas subtype C was significantly associated with a non-CIMP subtype (cluster 4). Another unique feature of subtype B was the lack of *TP53* mutations and chromosome 18q loss. These results clearly established the association between proteomic subtype B and MSI-high and CIMP, but suggest that subtype C may have different biological underpinnings.

The remaining three subtypes were associated with CIN, another well-accepted genetic property of CRC. In particular, subtype E was significantly associated with both *TP53* mutations and 18q loss, genomic features frequently associated with CIN tumours[25]. Interestingly, subtype E was also associated with *HNF4A* amplification and relatively higher abundance of HNF4α protein (Fig. 4c). HNF4α abundance was significantly higher in subtype E tumours compared to normal colon samples (multiple-test adjusted $P = 1.09 \times 10^{-6}$, two-sided Wilcoxon rank-sum test); however, significant upregulation of HNF4α was not observed for other subtypes (Methods). This result, together with our reanalysis of shRNA knockdown data from the Achilles project (Extended Data Fig. 8), suggests that HNF4α dependency may be particularly associated with the subset of tumours or cells with *HNF4A* amplification.

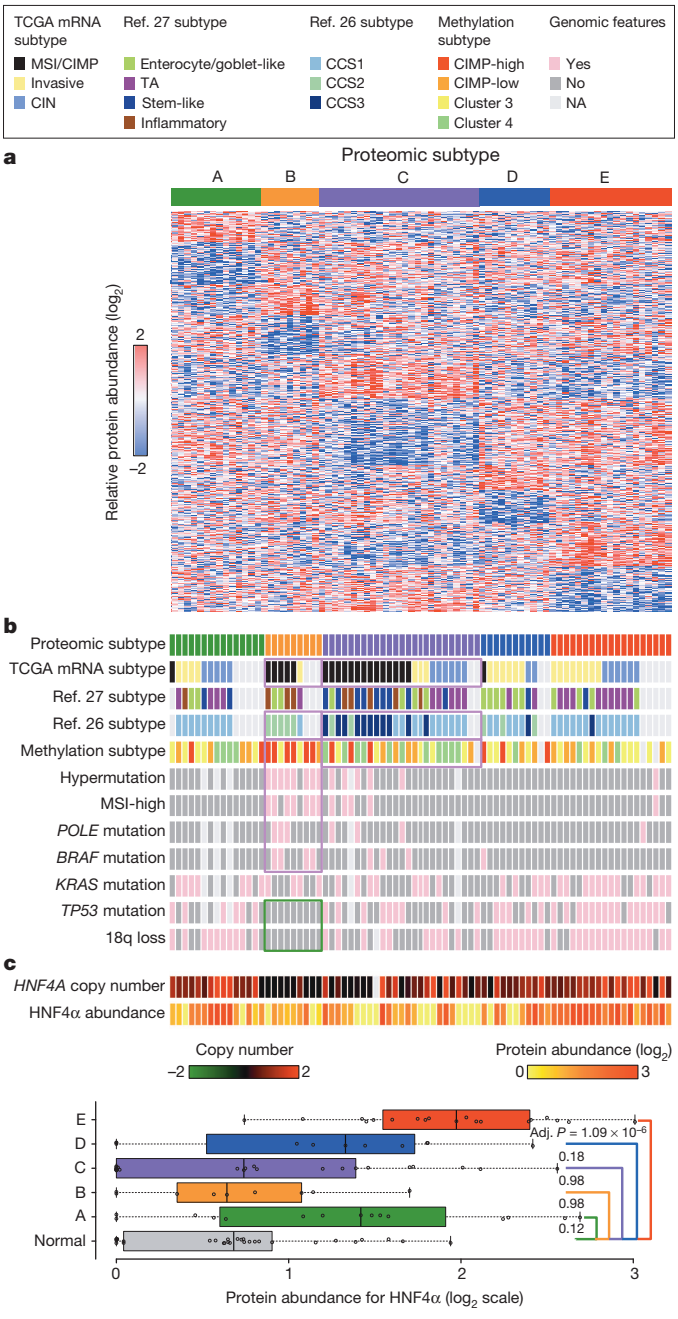


Figure 4 | Proteomic subtypes of colon and rectal cancers, associated genomic features, and relative abundance of HNF4α. **a**, Identification of five proteomic subtypes. Tumours are displayed as columns, grouped by proteomic subtypes as indicated by different colours. Proteins used for the subtype classification are displayed as rows. The heat map presents relative abundance of the proteins (logarithmic scale in base 2) in the 90-tumour cohort. **b**, Association of proteomic subtypes with major colorectal-cancer-associated genomic alterations and previously published transcriptomic and methylation subtypes. Subtypes that are significantly overlapped with a transcriptomic or methylation subtype are highlighted by pink boxes. Both proteomic subtypes B and C showed significant overlap with the TCGA MSI/CIMP subtype. In addition, they showed significant overlap with the CCS2 and CCS3 subtypes in the ref. 26 classification, respectively. Proteomic subtype B significantly overlapped with the TCGA CIMP-high methylation subtype, whereas subtype C significantly overlapped with a non-methylation subtype (TCGA cluster 4 methylation subtype). Subtypes overrepresented with a specific genomic alteration are also highlighted by pink boxes. The green box highlights the absence of *TP53* mutations and 18q loss in subtype B. **c**, The top panel shows *HNF4A* copy number and relative abundance of HNF4α protein in the five subtypes; the bottom panel compares relative abundance of HNF4α in the five subtypes to that in normal colon samples, respectively, and the adjusted *P* values are based on the two-sided Wilcoxon rank-sum test followed by multiple-test adjustment.

We also examined the association between the subtype classification and clinical features and found only that stage II tumours were significantly enriched in subtype C (multiple-test adjusted $P < 0.05$; Supplementary Table 11). Supervised statistical analyses at the individual protein level for 13 clinical and genomic features also identified few, if any, significant protein effects of these features, except for hypermutation status, MSI status and 18q loss (Supplementary Table 12), suggesting that the proteomic subtypes identified by the unsupervised clustering analysis captured the major proteome variations across the tumours.

Next, we compared the proteomic subtype classification with the TCGA transcriptional subtype classification for the 62 samples that had both subtype labels. Proteomic subtypes B and C both showed significant association with the TCGA subtype MSI/CIMP (Fig. 4b and Supplementary Table 11); however, they differ considerably at genomic, epigenomic and proteomic levels (Fig. 4a, b). We also examined alternative classifications of the TCGA samples based on two recently published transcriptomic subtype classifiers[26,27]. Proteomic subtype C, but not subtype B, showed enriched overlap with the 'stem-like' subtype described in ref. 27 and the colon cancer subtype 3 (CCS3) subtype described in ref. 26. Interestingly, tumours with stem-like and CCS3 classifications both have poor prognosis, which suggests that proteome subtype C also may be associated with poor prognosis. Therefore, the ability to distinguish subtype B from C through proteomics data are important, because MSI-high tumours typically have better prognosis[25].

Signatures for proteomic subtypes

To better understand the biology underlying the proteomic subtypes, we identified protein signatures for each subtype by supervised comparison of protein abundance in that subtype against all others; we also required signature proteins for a subtype to be significantly different in abundance compared to normal colon samples from 30 individuals analysed on the same proteome analysis platform (Methods and Supplementary Tables 13 and 14). As shown in Extended Data Fig. 10a, all CRC subtypes displayed more than 2,000 (>60%) significant protein abundance differences compared to normal colon. Although a full validation of the proteomic subtypes and protein signatures for the subtypes will require proteomic profiling data from an independent tumour cohort, a low cross-validation error rate of 3.8% demonstrated good generalizability of the subtypes and their signature proteins (Methods).

We performed Gene Ontology enrichment analysis for the subtype signatures using WebGestalt[28] (Methods and Supplementary Table 15). Genes involved in 'response to wounding' were significantly enriched in the up-signature of subtype C (multiple-test adjusted $P < 2.2 \times 10^{-16}$, Fisher's exact test). The wound-response gene signature is a powerful predictor of poor clinical outcome in patients with early stage breast cancers[29]. This result further links our subtype C to poor prognosis.

To understand better the functional networks underlying this subtype with potential clinical importance, we uploaded the up and down signatures of subtype C to NetGestalt[30] for enriched protein–protein interaction network module analysis. Four network modules were enriched with genes in the up signature for subtype C, whereas two modules were enriched with genes in the down signature (multiple-test adjusted $P < 0.01$, Fisher's exact test; Extended Data Fig. 10b). Notably, the down-signature-enriched module (III) included the E-cadherin (CDH1)-β-catenin (CTNNB1)-α-catenin (CTNNA1) complex (Extended Data Fig. 10c, e). E-cadherin, the most under-expressed protein in the subnetwork, suppresses invasion in lobular breast carcinoma[31] and is a switch for the epithelial-to-mesenchymal transition (EMT), which is associated with poor prognosis in colon cancer[32]. Other components of the module were desmosomal proteins (PKP2, JUP and DSG2) and cytokeratins (KRT18, KRT6A and KRT8). Reduction in both desmosome formation and cytokeratin expression is associated with EMT[33]. Moreover, proteins in the most significantly upregulated network module (Extended Data Fig. 10d, f) included collagens (COL1A1 and COL3A1) and extracellular matrix glycoproteins (FN1, BGN, FBN1 and FBN2) that also are markers of EMT[34,35]. These data strengthen the association of subtype C with poor prognosis and relate it to EMT activation.

Discussion

Our proteomic characterization of the genomically annotated TCGA colon tumours illustrates the power of integrated proteogenomic analysis. The data demonstrate that protein abundance cannot be reliably predicted from DNA- or RNA-level measurements. mRNA and protein levels were modestly correlated, as earlier cell and animal model studies suggested[36], but over two-thirds of these correlations were not statistically significant in the TCGA tumour set. Although most CNAs in CRC drive mRNA abundance changes, relatively few translated to consistent changes in protein abundance.

Genomic and proteomic technologies provide reinforcing data. RNA-seq data facilitate the discovery of variant proteins, which could serve as possible biomarker candidates or therapeutic targets. Combined mRNA and protein profiling data can identify potentially relevant genes in amplified chromosomal regions. This approach, which revealed the importance of chromosome 20q amplification and provided new insights into the role of HNF4α in CRC, can be broadly extended to understand roles of CNAs in other cancers. Proteomics identified CRC subtypes similar to those detectable by transcriptome profiles, but further captured features not detectable in transcript profiles. The separation of the TCGA MSI/CIMP subtype into distinct proteotypes illustrates the unique potential of proteomics-based subtyping. After validation in independent cohorts, protein subtype signatures could be directly translated into laboratory tests for tumour classification. Integrated proteogenomic analysis, as demonstrated in this study, will enable new advances in cancer biology, diagnostics and therapeutics.

METHODS SUMMARY

All tumour samples for the current study were obtained through the TCGA Biospecimen Core Resource (BCR) as described previously[6]. No other selection criteria other than availability were applied for this study. Patient-derived xenograft tumours from established basal and luminal B breast cancer intrinsic subtypes[37,38] were raised subcutaneously in 8-week-old NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ mice (Jackson Laboratories, Bar Harbour, Maine) as described previously[39,40]. Normal colon biopsies were obtained from screening colonoscopies performed between July 2006 and October 2010 under Vanderbilt University Institutional Review Board (IRB) approval no. 061096.

Tissue proteins were extracted and tryptic peptide digests were analysed by multidimensional liquid chromatography-tandem mass spectrometry. Xenograft quality control samples were run after every five colorectal tumour samples. Raw data were processed for peptide identification by database and spectral library searching and identified peptides were assembled as proteins and mapped to gene identifiers for proteogenomic comparisons. Quantitative proteomic comparisons were based on spectral count data. Detailed descriptions of the samples, LC-MS/MS analysis, and data analysis methods can be found in the Methods. All of the primary mass spectrometry data on TCGA tumour samples are deposited at the CPTAC Data Coordinating Center as raw and mzML files and complete protein assembly data sets for public access (https://cptac-data-portal.georgetown.edu).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 September 2013; accepted 2 May 2014.
Published online 20 July 2014.

1. The Cancer Genome Atlas Research Network Integrated genomic characterization of endometrial carcinoma. *Nature* **497,** 67–73 (2013).
2. The Cancer Genome Atlas Research Network Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).
3. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* 474**,** 609–615 (2011); erratum **490,** 292 (2012).
4. The Cancer Genome Atlas Research Network Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012); corrigendum **491,** 288 (2012).
5. The Cancer Genome Atlas Research Network Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

6. The Cancer Genome Atlas Research Network Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
7. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).
8. Wang, X. & Zhang, B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **29,** 3235–3237 (2013).
9. Wang, X. *et al.* Protein identification using customized protein sequence databases derived from RNA-Seq data. *J. Proteome Res.* **11,** 1009–1017 (2012).
10. Kim, W. K. *et al.* Identification and selective degradation of neopeptide-containing truncated mutant proteins in the tumors with high microsatellite instability. *Clin. Cancer Res.* **19,** 3369–3382 (2013).
11. Liu, H., Sadygov, R. G. & Yates, J. R. 3rd A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76,** 4193–4201 (2004).
12. de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels. *Mol. Biosyst.* **5,** 1512–1526 (2009).
13. Foss, E. J. *et al.* Genetic variation shapes protein networks mainly through non-transcriptional mechanisms. *PLoS Biol.* **9,** e1001144 (2011).
14. Ghazalpour, A. *et al.* Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet.* **7,** e1001393 (2011).
15. Gry, M. *et al.* Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* **10,** 365 (2009).
16. Foss, E. J. *et al.* Genetic basis of proteome variation in yeast. *Nature Genet.* **39,** 1369–1375 (2007).
17. Fu, J. *et al.* System-wide molecular evidence for phenotypic buffering in *Arabidopsis. Nature Genet.* **41,** 166–167 (2009).
18. Peng, J. *et al.* Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals Applied Statistics* **4,** 53–77 (2010).
19. Garrison, W. D. *et al.* Hepatocyte nuclear factor 4α is essential for embryonic development of the mouse colon. *Gastroenterology* **130,** 19.e1–19.e (2006).
20. Chellappa, K., Robertson, G. R. & Sladek, F. M. HNF4α: a new biomarker in colon cancer? *Biomark. Med.* **6,** 297–300 (2012).
21. Cheung, H. W. *et al.* Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc. Natl Acad. Sci. USA* **108,** 12372–12377 (2011).
22. Shimokawa, T. *et al.* Identification of TOMM34, which shows elevated expression in the majority of human colon cancers, as a novel drug target. *Int. J. Oncol.* **29,** 381–386 (2006).
23. Irby, R. B. *et al.* Activating *SRC* mutation in a subset of advanced human colon cancers. *Nature Genet.* **21,** 187–190 (1999).
24. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. R. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52,** 91–118 (2003).
25. Fearon, E. R. Molecular genetics of colorectal cancer. *Annu. Rev. Pathol.* **6,** 479–507 (2011).
26. De Sousa E Melo, F. *et al.* Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nature Med.* **19,** 614–618 (2013).
27. Sadanandam, A. *et al.* A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature Med.* **19,** 619–625 (2013).
28. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.* **33,** W741–W748 (2005).
29. Chang, H. Y. *et al.* Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl Acad. Sci. USA* **102,** 3738–3743 (2005).
30. Shi, Z., Wang, J. & Zhang, B. NetGestalt: integrating multidimensional omics data over biological networks. *Nature Methods* **10,** 597–598 (2013).
31. Polyak, K. & Weinberg, R. A. Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nature Rev. Cancer* **9,** 265–273 (2009).
32. Loboda, A. *et al.* EMT is the dominant program in human colon cancer. *BMC Med. Genomics* **4,** 9 (2011).
33. Geiger, T., Sabanay, H., Kravchenko-Balasha, N., Geiger, B. & Levitzki, A. Anomalous features of EMT during keratinocyte transformation. *PLoS One* **3,** e1547 (2008).
34. Kiemer, A. K., Takeuchi, K. & Quinlan, M. P. Identification of genes involved in epithelial-mesenchymal transition and tumor progression. *Oncogene* **20,** 6679–6688 (2001).
35. Zeisberg, M. & Neilson, E. G. Biomarkers for epithelial-mesenchymal transitions. *J. Clin. Invest.* **119,** 1429–1437 (2009).
36. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Rev. Genet.* **13,** 227–232 (2012).
37. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (2009).
38. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).
39. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999–1005 (2010).
40. Li, S. *et al.* Endocrine-therapy-resistant *ESR1* variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep.* **4,** 1116–1130 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by National Cancer Institute (NCI) CPTAC awards U24CA159988, U24CA160035, and U24CA160034; by NCI SPORE award P50CA095103 and NCI Cancer Center Support Grant P30CA068485; by National Institutes of Health grant GM088822; and by contract 13XS029 from Leidos Biomedical Research, Inc. Genomics data for this study were generated by The Cancer Genome Atlas pilot project established by the NCI and the National Human Genome Research Institute. Information about TCGA and the investigators and institutions comprising the TCGA research network can be found at http://cancergenome.nih.gov/.

Author Contributions B.Z., R.J.C.S., D.L.T., L.J.Z. and D.C.L. designed the proteomic analysis experiments, data analysis workflow, and proteomic–genomic data comparisons. K.F.S., L.J.Z., R.J.C.S. and D.C.L. directed and performed proteomic analysis of colon tumour and quality control samples. J.W., X.W., J.Z., Q.L., Z.S., P.W., S.W., R.J.C.S. and B.Z. performed proteomic-genomic data analyses. M.C.C., S.K., R.J.C.S. and D.L.T. performed analyses of mass spectrometry data and adapted algorithms and software for data analysis. S.R.D., R.R.T. and M.J.C.E. developed and prepared breast xenografts used as quality control samples. S.A.C., K.F.S. and D.C.L. designed strategy for quality control analyses. R.J.C.S., C.R.K, R.C.R. and H.R. coordinated acquisition, distribution and quality control evaluation of TCGA tumor samples. B.Z., J.W., R.J.C.S., R.J.C. and D.C.L. interpreted data in context of colon cancer biology. B.Z., R.J.C.S. and D.C.L. wrote the manuscript.

Author Information All of the primary mass spectrometry data on TCGA tumour samples are deposited at the CPTAC Data Coordinating Center as raw and mzML files for public access (https://cptac-data-portal.georgetown.edu). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.L. (daniel.liebler@vanderbilt.edu).

National Cancer Institute Clinical Proteomics Tumor Analysis Consortium (NCI CPTAC)

Steven A. Carr[1], Michael A. Gillette[1], Karl R. Klauser[1], Eric Kuhn[1], D. R. Mani[1], Philipp Mertins[1], Karen A. Ketchum[2], Amanda G. Paulovich[3], Jeffrey R. Whiteaker[3], Nathan J. Edwards[4], Peter B. McGarvey[4], Subha Madhavan[5], Pei Wang[6], Daniel Chan[7], Akhilesh Pandey[7], Ie-Ming Shih[7], Hui Zhang[7], Zhen Zhang[7], Heng Zhu[8], Gordon A. Whiteley[9], Steven J. Skates[10], Forest M. White[11], Douglas A. Levine[12], Emily S. Boja[13], Christopher R. Kinsinger[13], Tara Hiltke[13], Mehdi Mesri[13], Robert C. Rivers[13], Henry Rodriguez[13], Kenna M. Shaw[13], Stephen E. Stein[14], David Fenyo[15], Tao Liu[16], Jason E. McDermott[16], Samuel H. Payne[16], Karin D. Rodland[16], Richard D. Smith[16], Paul Rudnick[17], Michael Snyder[18], Yingming Zhao[19], Xian Chen[20], David F. Ransohoff[20], Andrew N. Hoofnagle[21], Daniel C. Liebler[22], Melinda E. Sanders[22], Zhiao Shi[22], Robbert J. C. Slebos[22], David L. Tabb[22], Bing Zhang[22], Lisa J. Zimmerman[22], Yue Wang[23], Sherri R. Davies[24], Li Ding[24], Matthew J. C. Ellis[24] & R. Reid Townsend[24]

[1]The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University Cambridge, Massachusetts 02142, USA. [2]Enterprise Science and Computing, Inc., 155 Gibbs St, Suite 420, Rockville, Maryland 20850, USA. [3]Clinical Research Division, Fred Hutchinson Cancer Research Center, 1100 Eastlake Avenue East, Seattle, Washington 98109, USA. [4]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, 3900 Reservoir Rd NW, Washington, DC 20057, USA. [5]Innovation Center for Biomedical Informatics, Georgetown University Medical Center, 2115 Wisconsin Ave NW, Suite 110, Washington, DC 20057, USA. [6]Icahn Institute and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, Hess CSM Building, Room S8-102, 1470 Madison Avenue, New York, New York 10029, USA. [7]Department of Pathology, The Johns Hopkins University, 600 North Wolfe Street, Baltimore, Maryland 21287, USA. [8]Department of Pharmacology and Molecular Science, the Johns Hopkins University, 733 N. Broadway, Baltimore, Maryland 21287, USA. [9]Antibody Characterization Laboratory, Advanced Technology Program, Leidos, Inc., 1050 Boyles Street, Frederick, Maryland 21701, USA. [10]Biostatistics Center, Massachusetts General Hospital Cancer Center, 55 Fruit Street, Boston, Massachusetts 02114, USA. [11]Department of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. [12]Gynecology Service/Department of Surgery, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, New York 10065, USA. [13]Office of Cancer Clinical Proteomics Research, National Cancer Institute, 31 Center Drive, MS 2580 Bethesda, Maryland 20892, USA. [14]Biomolecular Measurement Division, Material Measurement Laboratory, National Institute of Standards and Technology, 100 Bureau Drive, M/S 8300, Gaithersburg, Maryland 20899, USA. [15]Department of Biochemistry and Molecular Pharmacology, Smilow Research Building, Room 201, 522 First Avenue, New York University Langone Medical Center, New York, New York 10016, USA. [16]Biological Sciences Division, Pacific Northwest National Laboratory, 902 Battelle Boulevard, Richland, Washington 99352, USA. [17]Spectragen-Informatics, Rockville, Maryland 20850, USA. [18]Department of Genetics, Stanford University, Stanford, California 94305, USA. [19]The Ben May Department for Cancer Research, University of Chicago, 929 East 57th Street, W421 Chicago, Illinois 60637, USA. [20]University of North Carolina at Chapel Hill, 130 Mason Farm Road, Chapel Hill, North Carolina 27599, USA. [21]Department of Lab Medicine, University of Washington, Campus Box 357110, Seattle, Washington 98195, USA. [22]Vanderbilt University School of Medicine, 1161 21st Avenue South, Nashville, Tennessee 37232, USA. [23]Bradley Department of Electrical and Computer Engineering, Virginia Tech, 900 N. Glebe Road, Arlington, Virginia 22203, USA. [24]Department of Medicine, Washington University in St. Louis, 660 S. Euclid Avenue, St. Louis, Missouri 063110, USA.

METHODS

TCGA tumour samples. All samples for the current study were obtained through the TCGA Biospecimen Core Resource (BCR) as described previously[6]. Of the 276 samples described in the TCGA study, 95 specimens from 90 patients were available for the current study (Supplementary Table 1). No other selection criteria other than availability were applied for this study. Specimens for proteomic study were sectioned sequentially from the tumour blocks used for genomic studies, hence the 'bottom' slides from the genomic studies are representative of the material used for proteomics. The slides are available from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/). All samples contained at least 60% tumour nuclei, as described earlier[6]. Samples were shipped from the TCGA BCR in dry ice and kept frozen at −80 °C until processing. All TCGA colorectal tissue samples were washed before digestion to eliminate any residual optimal cutting temperature compound (OCT) . The tissue was placed in a 1.5-ml micro-tube and washed with 1 ml 70% ethanol in water for 30 s with vortexing. The supernatant was then discarded and the tissue washed with 1 ml of 100% H_2O for 30 s with vortexing and again the supernatant was discarded. One millilitre of 70% ethanol in water was added to the tissue sample and incubated for 5 min at 23 °C following by centrifugation at 20,000*g* for 2 min at 20 °C. The supernatant was removed and this wash step was repeated. Next, 1 ml of 85% ethanol in water was added to the tissue and incubated for 5 min at room temperature followed by centrifugation at 20,000*g* for 2 min at 20 °C. The supernatant was removed the wash was repeated. For the final wash, the tissue was washed in 1 ml 100% ethanol and incubated for 5 min at room temperature and centrifuged at 20,000*g* for 2 min at 20 °C. The supernatant was removed and the wash was repeated.

Breast tumour xenograft tumour samples. Patient-derived xenograft tumours from established basal (WHIM2) and luminal B (WHIM16) breast cancer intrinsic subtypes[37,38] were raised subcutaneously in 8-week-old NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/*SzJ* mice (Jackson Laboratories, Bar Harbour, Maine) as described previously[39,40]. These tumours have significantly different gene expression and proteomic signatures[40] that are related to their intrinsic biology and endocrine signalling. Tumours from each animal were collected by surgical excision at approximately 1.5 cm^3 with minimal ischaemia time by immediate immersion in a liquid nitrogen bath. The tumour tissues were then placed in pre-cooled tubes on dry ice and stored at −80 °C. A tissue 'pool' of cryopulverized tumours was prepared in order to generate sufficient material that could be reliably shared and analysed between multiple laboratories.

In brief, tumour pieces were transferred into pre-cooled Covaris Tissue-Tube 1 Extra (TT01xt) bags (Covaris no. 520007) and processed in a Covaris CP02 Cryoprep device using different impact settings according to the total tumour tissue weight: <250 mg = 3; 250–350 mg = 4; 350–440 mg = 5; 440–550 mg = 6. Tissue powder was transferred to an aluminium weighing dish (VWR no. 1131-436) on dry ice and the tissue was thoroughly mixed with a metal spatula pre-cooled in liquid nitrogen. The tissue powder was then partitioned (~100-mg aliquots) into precooled cryovials (Corning no. 430487). All procedures were carried out on dry ice to maintain tissue in a powdered, frozen state.

Protein extraction and peptide fractionation. *Protein extraction and digestion of tissue specimens.* Following OCT removal, tissue specimens were placed in 1.5-ml micro-centrifuge tubes and re-suspended in 100 µl of trifluoroethanol (TFE) and 100 µl of 100 mM ammonium bicarbonate, pH 8.0. If additional buffer was required, equal volumes of TFE and 100 mM ammonium bicarbonate pH 8.0, were added accordingly. In addition, powderized xenograft tumour tissue representing the comparison and reference (CompRef) samples for luminal (WHIM16) and basal (WHIM2) breast cancer subtypes were analysed with each set of 10 TCGA colorectal tissue samples. Samples were sonicated using a Fisher Scientific Sonic Dismembrator Model 100 at a setting of 20 W for 20 s followed by 30 s incubation on ice. This sonication step was repeated twice and samples were placed on ice between sonications. The resulting homogenate was heated with shaking at 1000 r.p.m. for 1 h at 60 °C followed by a second series of sonication steps, as described above. A protein measurement then was obtained for each sample using the BCA Protein Assay (ThermoFisher Pierce, Rockford, Illinois) using the manufacturer's protocol.

An aliquot equivalent to 200 µg was removed and reduced with tris(2-carboxyethyl)phosphine (TCEP, 20 mM) and dithiothreitol (DTT, 50 mM) at 60 °C for 30 min followed by alkylation with iodoacetamide (IAM, 100 mM) in the dark at room temperature for 20 min. The lysate was diluted with the appropriate volume of 50 mM ammonium bicarbonate, pH 8.0, to reduce the TFE concentration to 10%, trypsin was added at a ratio of 1:50 (w:w) and digested overnight at 37 °C. The digested mixture was frozen at−80 °C and lyophilized to dryness. The lyophilized samples were re-suspended in 350 µl of high-pressure liquid chromatography (HPLC)-grade water and vortexed vigorously for 1 min and desalted using an Oasis HLB 96-well µElution plate (30 µm, 5 mg, Waters Corp., Milford, Massachusetts), which was pre-washed with 500 µl of acetonitrile and equilibrated with 750 µl of HPLC-grade water. The flow-through was discarded and the plates were washed with 500 µl of HPLC-grade water and the peptides eluted with 80% acetonitrile and the eluates were evaporated to dryness *in vacuo*. Samples were stored in the freezer until further analysis.

Peptide fractionation by basic reverse-phase liquid chromatography. Samples were reconstituted in 300 µl of solvent A (1.0 M triethylamine bicarbonate (TEAB), pH 7.5). The reconstituted sample was then diluted with an additional 100 µl of solvent A and the entire 400 µl of solution was injected into the basic reverse-phase liquid chromatography (basic RPLC) column. Tryptic peptides were fractionated using high-pH RPLC separation with an XBridge BEH C18, 250 mm × 4.6 mm analytical column (130A, 5 µm particle size) equipped with a XBridge BEH C18 Sentry guard cartridge at a flow rate of 0.5 ml min^{-1}. The solvents were 10 mM TEAB, pH 7.5, in water as mobile phase A and 100% acetonitrile as mobile phase B. Sample fractionation was accomplished using the following multistep linear gradient: from 0 to 5% B in 10 min, from 5 to 35% B in 60 min, from 35–70% in 15 min and held at 70% B for an additional 10 min before returning to initial conditions. A total of 60 fractions were collected over the 105-min gradient and concatenated into 15 fractions by combining fractions 1, 16, 31, 46; 2, 17, 32, 47; and so on up to fractions 15, 30, 45 and 60. The samples were evaporated to dryness in a Speed-Vac sample concentrator and stored at −80 °C until LC-MS/MS analysis.

LC-MS/MS analysis. Resulting peptide fractions were resuspended in 50 µl of water containing 2% acetonitrile and 0.1% formic acid and analysed using a Thermo LTQ Orbitrap Velos ion trap mass spectrometer equipped with an Eksigent NanoLC 2D pump and AS-1 autosampler. A 2-µl injection volume of the peptide solution was separated on a packed capillary tip (Polymicro Technologies, 100 µm × 11 cm) containing Jupiter C18 resin (5 µm, 300 Å, Phenomenex) using an in-line solid-phase extraction column (100 µm × 6 cm) packed with the same C18 resin using a frit generated with liquid silicate Kasil[41]. Mobile phase A consisted of 0.1% formic acid and Mobile phase B consisted of 0.1% formic acid in acetonitrile. A 95-min gradient was preceded by a 15-min washing period (100% A) at a flow rate of 1.5 µl min^{-1} to remove residual salt. Following the wash, the mobile phase was programmed to 25% B by 50 min, followed by an increase to 90% B by 65 min and held for 9 min before returning to the initial conditions. A full MS scan was collected for peptides from 400–2,000 *m/z* on the Orbitrap at a resolution of 60,000 followed by eight data-dependent MS/MS scans from lowest to highest signal intensity on the linear trap quadrupole (LTQ). Centroided MS/MS scans were acquired using an isolation width of 3 *m/z*, an activation time of 30 ms, an activation *q* of 0.250 and 35% normalized collision energy. One microscan with a max ion time of 100 ms and 1,000 ms was used for each MS/MS and full MS scan, respectively. MS/MS spectra were collected using a dynamic exclusion of 60 s with a repeat of 1 and repeat duration of 1.

All TCGA samples were analysed on the same Thermo LTQ Orbitrap Velos instrument, with sample analysis beginning on 26 July 2012 and concluding on 17 February 2013. Benchmark quality control samples from one basal and one luminal human breast tumour xenograft were analysed in alternating order after every five CRC samples. Specifically, five TCGA samples were run on the instrument, followed by a luminal CompRef sample, and then another five TCGA samples followed by a basal CompRef sample. Bovine serum albumin (BSA) tryptic digest standards were analysed before and after every ten TCGA samples and were used to monitor instrument sensitivity, BSA standard sequence coverage and chromatographic performance to determine acceptance or rejection of the acquired data.

LC-MS/MS data analysis. *Peptide identification.* Basic protein identification used the RefSeq human protein sequence database, release version 54, and both database and peptide library search strategies. Two bovine trypsin sequences and one porcine trypsin sequence were appended to these 34,586 sequences. The 14 June 2011 National Institute of Standards and Technology (NIST) human spectral library for ion traps (617,000 spectra, counting paired decoys) was indexed against this sequence database. Thermo RAW files were converted to mzML peaklists by the ScanSifter algorithm[42] or by ProteoWizard msConvert[43]. The ScanSifter files were employed by Pepitome 1.0.42 (ref. 44) for spectral library search and MyriMatch 2.1.87 (ref. 45) for database search, whereas the msConvert files were used by MS-GF+ v9176 (ref. 46). Pepitome and MyriMatch used precursor tolerances of 10 p.p.m., while MS-GF+ used a 20-p.p.m. window; all three algorithms allowed fragments to vary by up to 0.5 m/z, and both database search engines considered semi-tryptic peptides equally with fully tryptic peptides, allowed for isotopic error in precursor ion selection, conducted on-the-fly peptide sequence reversal, and applied static +57 modifications to cysteines and dynamic +16 oxidations to methionines. MS-GF+ considered acetylation for protein amino termini, whereas MyriMatch added pyroglutamine modifications to the N termini of peptides starting with Gln residues. Pepitome considered any modification variants and trypsin specificities that were included in the spectral library.

Protein assembly. Spectral identification files from each of the search engines (pepXML for Myrimatch and Pepitome, mzIdent for MS-GF+) were converted to IDPicker 3 index files (idpXML) using IDPicker 3 (ref. 47). The resulting 4,275 idpXML files (95 samples × 15 fractions × 3 search engines) were used for a final protein assembly using IDPicker 3. For initial protein assembly, peptide identification stringency was set at a maximum of 1% reversed peptide matches, that is, 2% peptide-to-spectrum matches (PSM) FDR and a minimum of 2 unique peptides to identify a given protein within the full data set. Because the majority of false identifications occur with low frequency, the number of proteins identified by reversed peptides (false-positive protein identifications) expands exponentially with the size of the data set. In this case, a 2% PSM FDR resulted in an unacceptably high 32% FDR at the protein level. To maximize both the number of proteins identified as well as the number of spectra observed for each protein, we adopted a procedure similar to that outlined in ref. 48. To optimize the number of proteins identified we applied a very stringent filter at 0.1% PSM FDR and a minimum of 2 distinct peptides identified for each protein. This filter resulted in the identification of a total of 94,442 distinct peptides among the 95 samples representing 7,526 protein groups with a protein-level FDR of 2.64%. To rescue high quality PSMs that were excluded by the stringent PSM FDR threshold, we relaxed the PSM FDR threshold to 1% for the confidently identified proteins. This process increased the number of distinct peptides identified to 124,823, corresponding to a total number of 6,299,756 spectra; the rescued PSMs were of high quality, with a median PSM FDR of less than 0.2% (Extended Data Fig. 2), indicating the maintained integrity of the data set. To facilitate the integration between genomic and proteomic data, we further assembled the peptides at the gene level. This assembly resulted in 7,211 gene groups with a gene-level FDR of 2.7%. Genes identified from each sample ranged from 3,372 to 5,456, with a median gene count of 4,656 for the 95 samples. 1,530 genes (21%) were found in all 95 samples, 4,628 genes (64%) in more than half of the samples, and only 10 genes (0.1%) in just a single sample.

Variant peptide identification. To identify variant peptides, we used a customized protein sequence database approach[9], in which we derived customized protein sequence databases from matched RNA-seq data and then performed database searches using the customized databases for individual samples.

To identify SNVs and indels from RNA-seq data, BAM files for 86 of the 90 patients were downloaded from CGHub in February 2013. Although 87 samples were analysed by RNA-seq, we were unable to obtain the bam file for one of the 87 samples. Tophat (v2.0.7) was used to re-align reads to human reference genome (hg19) in a spliced mode using default parameters, allowing a maximum of 10 hits per read. The resulting BAM files were indexed using samtools (0.1.18, http://samtools.sourceforge.net/). We used a custom-written script to summarize the reads mapping information (Extended Data Fig. 3a) and calculate the exon coverage. As shown in Extended Data Fig. 3b, 76% of exons were covered by RNA-seq reads, and 64% had an average coverage greater than 1.

Putative SNVs and short indels were called one library at a time using *mpileup* from samtools and *varFilter*. Putative SNVs were further filtered based on the following criteria: SNP quality ≥ 20; mapping quality ≥ 20; and read depth ≥ 3 and then recorded in a variant call format (VCF) file. For short INDEL candidates, gapped reads ≥ 3 were required.

For customized database construction and variant peptide identification we used an R package *customProDB*[8] (http://bioconductor.org/packages/2.13/bioc/html/customProDB.html) to annotate variations predicted from RNA-seq, including mapping to dbSNP135 and COSMIC64 databases. For each sample, *customProDB* generates a protein FASTA database by appending proteins with nonsynonymous protein coding SNVs and aberrant proteins to the end of the standard RefSeq human protein sequence database. Owing to the low coverage of this RNA-seq data set, we did not remove the low abundance transcripts from the standard RefSeq database. Peptide identification was performed for each sample separately using corresponding customized FASTA database and MyriMatch 2.1.87 (ref. 45). Search settings were identical to those described for Myrimatch above. IDPicker 3 was used for protein assembly as described earlier, except that the data set was filtered at 1% PSM FDR and a minimum of 5 spectra identified per protein. The full data set consisted of 8,352 protein groups with 1.8% protein FDR. Identified SAAVs were further annotated for existence in the somatic variant list published by TCGA[6] (that is, TCGA-somatic variants), existence in the COSMIC64 database (that is, COSMIC-supported variants), and existence in the dbSNP135 database (that is, dbSNP-supported variants). To identify TCGA-somatic variants, we downloaded the MAF (mutation annotation format) files from the Firehose website (http://gdac.broadinstitute.org, version 20130523). As the coordinates in MAF files were based on hg18, liftOver (http://hgdownload.cse.ucsc.edu/admin/exe/) from UCSC was used to convert genome coordinates to hg19. Support of these somatic mutations by RNA-seq data are shown in Extended Data Fig. 3c. All identified variant peptides as well as SAAVs and their annotations can be found in Supplementary Tables 2 and 3.

Protein quantification. We used spectral count, or the total number of MS/MS spectra taken on peptides from a given protein in a given LC/LC−MS/MS analysis as the basis for protein quantification. Spectral count is linearly correlated with the protein abundance over a large dynamic range[11]. This simple but practical quantification method has found broad application in detecting differential or correlated protein expression[49–53], and multiple groups have concluded that spectral counting achieves similar accuracy to more complex methods such as the intensity-based techniques[53–55]. Previously, we have confirmed proteomic changes detected from spectral count data by targeted proteomics with multiple reaction monitoring (MRM) in different data sets[56,57].

For the quality control sample data set, spectral count data were summarized at the protein group level, where a protein group is defined as the set of proteins that are indiscernible on the basis of the observed peptides. For each group, a random protein was selected to represent the group. The final spectral count table has data for 7,440 proteins and 20 samples, with 10 basal samples and 10 luminal samples.

For the TCGA tumour data set, to facilitate the integration between genomic and proteomic data, spectral count data were summarized at the gene group level, where a gene group is defined as the set of genes that are indiscernible on the basis of the observed peptides. To ensure reproducibility, for each gene, the longest protein was selected for the calculation of protein length. For each gene group, the gene with the shortest protein length was selected to represent the group following the Occam's razor principle. The final spectral count table has data for 7,211 genes and 95 samples (5 tumours have duplicated samples; Supplementary Table 4). For analysis that required only one sample from the duplicates, the sample with a larger total spectral count was selected.

For platform evaluation, we used the quality control data set to evaluate the technical variability in both protein identification and quantification based on data generated from the replicates. For the basal breast carcinoma xenografts, the numbers of identified proteins among the 10 replicates ranged from 4,771 to 6,190 and the coefficient of variance was 8%; for the luminal breast carcinoma xenografts, the numbers ranged from 4,639 to 5,842 and the coefficient of variance was 7%.

For quantitative analysis, spectral count data from the quality control data set were subjected to quantile normalization[58] using the *normalizeQuantiles* function in the *limma* package in Bioconductor, followed by log base 2 transformation. Pairwise Spearman's correlation coefficient was calculated for all sample pairs and the results were plotted in R. Samples from the same group showed a high similarity (Spearman's correlations of 0.85 and 0.88 for W16 and W2, respectively), whereas samples from different groups are clearly different (Spearman's correlation, 0.68) (Extended Data Fig. 5a). Consistent with this, in the TCGA tumour data set (Supplementary Table 4), the five duplicate pairs showed an average Spearman's correlation coefficient of 0.81.

In addition to the quantile normalization, other normalization methods commonly used in shotgun proteomics data analysis include global normalization[59] and the NSAF[60] (normalized spectral abundance factor). The quantile normalization makes the distributions of spectral count data from individual samples comparable to each other. The global normalization method makes the total numbers of identified spectra comparable across all samples. The NSAF method normalizes for both protein length and the total number of identified spectra from a sample. We performed all three types of normalization for the quality control data set and log transformation was applied for all normalized data. Among the three normalization methods, NSAF has a clear advantage for comparing abundance level across proteins because it is the only one that considers protein length. However, it is not clear which method is the best for comparing relative abundance of individual proteins among different samples. We compared the three methods on the basis of the intraclass correlation coefficient (ICC) analysis. For each normalization method, we used the *icc* function in the R package *irr* to calculate ICCs for individual proteins in the quality control data set. For each protein, the ICC score estimates the correlation between replicated measurements within a group, or the ratio of the between-group variance to the total variance observed for the protein. Our assumption is that better normalization methods should produce higher ICC scores. The analysis was done for the top 1,000, 500 or 100 proteins with the largest variance. The cumulative fraction curves for the ICC scores were plotted in R (Extended Data Fig. 5b). Each curve shows the cumulative fraction of all proteins in a given normalization method that had an ICC score less than or equal to a given value. The results suggested that quantile normalization generated slightly more consistent quantification than total spectral count normalization, and both methods clearly out-performed the NSAF normalization. Therefore, quantile normalized spectral count data were used in all analyses comparing relative abundance of individual proteins among different samples.

Spectral-count-based quantification is not accurate enough for comparing relative expression of low-abundance proteins. To decide a minimal number of

spectral counts to be required for reliable relative comparison, we used the ICC analysis to compare the ICC scores for proteins with different abundance levels. Specifically, we sorted all proteins in the quality control data set based on their total spectral counts and then divided the proteins into 10 bins with equal numbers of proteins. For each bin, we used the *icc* function to calculate ICCs for individual proteins in the bin. The analysis was done for the top 300, 200 or 100 proteins, with the largest variance in each bin. The cumulative fraction curves for the ICC scores were plotted in R (Extended Data Fig. 5c). Protein bins with spectral counts less than 1.4 showed clearly lower ICC scores, whereas the ICC score curves started to converge when the average spectral count was greater than 1.4. These data suggest that an average minimum spectral count of 1.4 per sample is required for reliable comparison of relative protein abundance. Applying this cutoff to the TCGA tumour data set identified 4,122 protein groups with a protein-level FDR of 0.47%. The 3,899 corresponding genes were used to compare relative protein abundance across tumour samples.

Impact of SNVs on protein expression. All nonsynonymous protein coding SNVs detected from the RNA-seq data were annotated for their existence in dbSNP135, COSMIC64, and previously reported somatic mutations[6]. For each SNV in one of the above three categories, if the SNV-containing gene is included in the 3,899 genes with sufficient spectral count data for quantitative comparison across samples, we quantified the impact of the SNV on protein expression by comparing protein expression level of the gene in the SNV-containing sample with those in samples without the SNV using the formula:

$$\text{Impact score} = (\text{EXP} - \text{MEDIAN}_{\text{non-SNV}})/\text{MAD}_{\text{non-SNV}}$$

where EXP is the expression level of the protein in the sample containing the SNV, and $\text{MEDIAN}_{\text{non-SNV}}$ and $\text{MAD}_{\text{non-SNV}}$ are the median and MAD (median absolute deviation) of the expression levels of the protein in all samples without the SNV.

The cumulative fraction curves for the impact scores of SNVs in the three categories were plotted in R (Fig. 1d). Differences between the distributions were evaluated using the Kolmogorov–Smirnov test. A Chi-squared test was also used to directly compare the percentage of high impact variations (absolute impact score greater than 2) for SNVs in different categories (Fig. 1d).

Parallel reaction monitoring analysis. We selected three distinct SAAVs detected in four TCGA samples by the shotgun analysis for further validation by targeted proteomic analyses using parallel reaction monitoring (PRM)[61], including *KRAS*(Gly12Asp) in TCGA-AA-3818 and TCGA-AG-A00Y, *ANXA11*(Ile278Val) in TCGA-AF-3400, *SRSF9*(Tyr35Phe) in TCGA-AA-A01P. These samples were prepared by the same method as used for shotgun analyses, including basic RPLC fractionation. The samples were re-suspended in 50 μl water containing 2% acetonitrile and 0.1% formic acid containing 12.5 fmol μl^{-1} each of the $^{13}C^{15}N$-isotopically labelled standards for the target variant peptides. All fractions from each sample were analysed separately.

All peptide separations were performed using an Easy nLC-1000 pump and autosampler system (Thermo Fisher Scientific). For each analysis, 2 μl of each sample was injected onto an in-line solid-phase extraction column (100 μm × 6 cm) packed with ReproSil-Pur C18 AQ 3 μm resin (Dr. Maisch GmbH) and a frit generated with liquid silicate Kasil 1 and washed with 100% Solvent A (0.1% formic acid) at a flow rate of 2 μl min^{-1}. After a total wash volume of 7 μl, the precolumn was placed in-line with a PicoFrit capillary column (New Objective, 11 cm × 75 μm) packed with the same resin. The peptides were separated using a linear gradient of 2% to 35% Solvent B (0.1% formic acid in acetonitrile) at a flow rate of 300 nl min^{-1} over 40 min, followed by an increase to 90% B over 4 min and held at 90% B for 6 min before returning to initial conditions of 2% B.

PRM analyses were performed on a Q-Exactive mass spectrometer (Thermo Fisher Scientific). For ionization, 1,800 V was applied and a 250-°C capillary temperature was used. All basic RPLC fractions from each sample were analysed using an acquisition method that combined a full scan selected ion monitoring (SIM) event followed by 14 PRM scans as triggered by an unscheduled inclusion list containing the target precursor ions representing variant peptides. The SIM scan event was collected using a m/z 380–1,500 mass selection, an Orbitrap resolution of 17,500 (at m/z 200), target automatic gain control (AGC) value of 3×10^6 and a maximum injection time of 30 ms. The PRM scan events used an Orbitrap resolution of 17,500, an AGC value of 1×10^6 and maximum fill time of 80 ms with an isolation width of 2 m/z. Fragmentation was performed with a normalized collision energy of 27 and MS/MS scan were acquired with a starting mass of m/z 150.

All PRM data analysis was performed using Skyline software[62]. Validation was achieved by comparing the fragment ion ratios and retention times of the endogenous variant peptide to that of the isotopically labelled standard. In addition, the MS/MS spectra of the endogenous, unlabelled and isotope-labelled standards acquired during the PRM analyses were compared to the original MS/MS spectra collected for the same peptides during the shotgun analyses. We successfully confirmed all selected mutations and one example is shown in Extended Data Fig. 4.

Evaluating mRNA–protein correlation. *mRNA data.* We downloaded the TCGA CRC RNA-seq data from the TCGA portal, which was from Illumina HiSeq 2000 RNA Sequencing Version 2 analysis and normalized by the RSEM algorithm[63]. This included RSEM measurements for 87 samples with matched proteomic data. It has been shown that RSEM is preferred over the popular FPKM[64] (fragments per kilobase (of exon) per million fragments mapped) measure for comparing gene expression across samples. However, our analysis suggested that the RSEM measure is highly correlated with gene length (Extended Data Fig. 6a) and thus not appropriate for comparing expression of different genes within a sample. Therefore, we used cufflinks[64] to generate FPKM measures. As mentioned above, we were unable to obtain the BAM file for one of the 87 samples, thus only 86 samples had FPKM data. Unlike RSEM, FPKM is independent of gene length (Extended Data Fig. 6a)

Number of overlapping genes in the proteomics and RNA-seq data sets. The proteomics data set for TCGA tumours included a total of 7,211 genes, among which 7,176 genes were included in the RNA-seq data set. The analysis for correlation between steady state mRNA and protein abundance (see below) (Fig. 2a) did not involve the comparison of relative protein abundance for the same gene across samples. For this analysis, all 7,176 genes were included. The analysis for correlation between mRNA and protein variation (Fig. 2b) involved relative protein abundance comparison for the same gene across samples. Our study of the quality control data set suggests that a minimal average spectral count of 1.4 is required for a reliable relative protein abundance comparison. Among the 7,211 genes in the proteomics data set, only 3,899 met this requirement. Moreover, among the 3,899 genes, only 3,764 were included in the RNA-seq data set. Therefore, the 3,764 genes were used in the analysis for correlation between mRNA and protein variation.

Correlation between steady state mRNA and protein abundance. To compare the steady state mRNA and protein abundance within individual samples, all mRNA and protein measurements within a sample have to be comparable. Thus, we used the FPKM and NSAF values to estimate mRNA and protein abundance, respectively. For each of the 86 samples, we calculated the Spearman correlation coefficient between FPKM and NSAF measurements for the 7,176 genes. Next, *P* values corresponding to the coefficients were computed and adjusted by the Benjamini–Hochberg procedure[65]. Significant calls were made based on an adjusted *P* value cutoff of 0.01.

Correlation between mRNA and protein variation. To compare mRNA and protein variations across samples, we focused on the 3,764 genes with both RSEM measurement in RNA-seq data and a minimal spectral count of 1.4 per sample in the proteomics data. The quantile normalized proteomics data were used for this analysis according to the quality control databased method comparison results (Extended Data Fig. 5b). We first calculated the Spearman correlation coefficient between RSEM and quantile normalized measures for each of 3,764 genes. Then, *P* values corresponding to the coefficients were computed and adjusted by the Benjamini–Hochberg procedure. Significant calls were made based on an adjusted *P* value cutoff of 0.01.

KEGG enrichment analysis. Based on the Spearman correlation coefficients between RSEM and quantile normalized measurements of the 3,764 genes, we performed KEGG enrichment analysis using the Kolmogorov–Smirnov test. Then, *P* values were adjusted by the Benjamini–Hochberg procedure and significant calls were made based on an adjusted *P* value cutoff of 0.05.

mRNA–protein correlation versus stability of the molecules. A recent study in a mouse fibroblast cell line suggests a poor correlation between mRNA and protein half-lives[66]. To investigate the relationship between mRNA–protein correlation and the stability of the molecules, we downloaded mRNA and protein half-life data from the mouse study. Only common genes in both our study and the mouse study were included in the analysis. Following the criteria used in the original publication[66], we defined the top third mRNAs and proteins with the highest half-lives as stable mRNAs and proteins and the bottom third with the lowest half-lives as unstable mRNAs and proteins. Accordingly, we separated human genes into four categories based on the mRNA and protein half-lives of their mouse orthologues: stable mRNA–stable protein; stable mRNA–unstable protein, unstable mRNA–stable protein, and unstable mRNA–unstable protein. Distribution of mRNA–protein correlations for genes in each category was plotted in Extended Data Fig. 6b. Correlation difference among the four categories were evaluated based on the Kruskal–Wallis non-parametric ANOVA test, and the difference between the stable mRNA–stable protein, and unstable mRNA–unstable protein groups were calculated based on a two-sided Wilcoxon rank-sum test.

Impact of copy number alterations on gene and protein abundance. *Univariate analysis.* The TCGA CRC gene level CNA data were downloaded from the output of GISTIC2 (ref. 67) in Firehose. The data set was generated on the Affymetrix Genome-Wide Human SNP Array 6.0 array and contained 23,125 genes and 575

samples. The matched CNA, proteomics and RNA-seq measurements from 85 samples were used to study the impact of CNA on gene and protein expression. First, for each of the 23,125 genes in the CNA data, we calculated the Spearman correlation coefficient between CNA measures and mRNA and protein abundance for the 3,764 genes with both RSEM measurements in RNA-seq and a minimal spectral count of 1.4 per sample in proteomics. Then, *P* values corresponding to the coefficient were corrected using the Benjamini–Hochberg procedure. Significant CNA–mRNA and CNA–protein correlations were identified based on an adjusted *P* value cutoff of 0.01.

Multivariate analysis. In addition to the univariate analysis, we also employed a recently developed statistical tool—regularized multivariate regression for master predictors (remMap)[18]—to jointly model CNAs and mRNA and protein abundance.

For this analysis, the level-three segmented DNA copy numbers profiles were downloaded from Firehose (http://gdac.broadinstitute.org, version 20130809). To align the segment data from different samples, we first broke the genome using the union of the break points detected in all tumour samples and filtered the small regions with less than 10 megabase pairs. This resulted in 7,219 regions. Then for each region of each sample, we recorded its copy number based on the inferred DNA copy number of the corresponding segment in the sample, with tail values truncated at ± 1.5. Owing to the high spatial correlation in DNA copy number profiles, we further condensed these 7,219 regions into 1,586 CNA intervals, which have consistent DNA copy number pattern across all samples, by applying fixed order clustering (FOC)[68]. The copy number for each interval in each sample was then calculated as the mean of the copy number of all regions within the interval. In the end, we normalized the copy number of each CNA interval across all samples to have mean 0 and standard deviation 1.

RNA-seq data and proteomics data were the same as those used in the univariate analysis. For the RNA-seq data, we excluded genes that show interquartile range (IQR) less than the 75% quantile, which resulted in 941 genes. Then we normalized the abundance of each gene across all samples to have mean 0 and standard deviation 1. For the proteomics data, we excluded proteins that show IQR less than the 75% quantile, which resulted in 941 proteins. Then we normalized the abundance of each gene across all samples to have mean 0 and standard deviation 1.

The data matrices of CNV intervals, mRNA abundance and protein abundance obtained as described above were used to fit the remMap model[18]. Specifically, mRNA (or protein) abundances were treated as responses and CNA data were treated as predictors. Non-zero coefficients in the multivariate regression model suggest regulation relationships between the corresponding CNA and mRNA (or protein). The tuning parameters in remMap were selected based on two-dimensional grid research using fivefold cross validation error scores. The optimal tuning parameter was $(l_1, l_2) = (6, 90)$ for regression models between mRNA and CNA; and $(l_1, l_2) = (10, 60)$ for models between protein and CNA. Regulations between a CNA interval and an mRNA (or protein) were declared if the corresponding coefficients in the regression models from all five models inferred in the cross validation were non-zero.

The results are summarized in Supplementary Tables 6–9. Five CNA intervals in cytoband regions 11p15.5, 18p12.32–1811.21, 18q21.2–18q23, 20q11.21–20q11.23 and 20q11.23–20q13.33 were detected to be *trans* hubs for mRNA abundance. And four of these regions on chromosomes 18 and 20 were also detected to be *trans* hubs for protein abundance. Specifically, the two CNA intervals on chromosome 20q had the highest number of *trans* regulations for both mRNA and protein abundance (Supplementary Tables 6 and 7).

Reanalysis of HNF4A *shRNA knockdown data.* HNF4α is a transcription factor with a key role in normal gastrointestinal development[19] and is increasingly being linked to liver and colon cancer[20]. The observations that *HNF4A* is located in an amplification peak and shows significant CNV–mRNA, CNV–protein and mRNA–protein correlations (Fig. 3c) support its oncogenic role. Consistent with this view, a recent study on 102 human cancer cell lines in the Achilles project found that shRNA knockdown of *HNF4A* has a relatively stronger negative impact on proliferation and viability of colon cancer cells compared to other cancer cell types[21]. However, there are contradictory reports on whether HNF4α acts as an oncogene[69,70] or a tumour suppressor gene[71,72] in CRC.

It has been suggested that HNF4α isoforms driven by different promoters may have different roles in colon cancer: promotor 1 (P1) HNF4α acts as a tumour suppressor, whereas the exact role of P2 HNF4α remains to be determined[20,72]. The P1- and P2-driven isoforms differ by only 16–29 amino acids in their N-terminal domain[20], and our shotgun proteomics data were not able to distinguish these two types of isoforms (Extended Data Fig. 8a). However, one of the shRNAs used in the Achilles project specifically targets P1 HNF4α (Extended Data Fig. 8a). Therefore, we reanalysed data from the Achilles project to compare the effect of different *HNF4A*-targeting shRNAs on the proliferation of colon cancer cell lines. We also took into consideration the *HNF4A* amplification status of the cell lines.

Specifically, we downloaded the shRNA knockdown data from the Achilles project website (http://www.broadinstitute.org/software/cprg/?q=node/10). The study had five *HNF4A* shRNAs: TRCN0000019189, TRCN0000019190, TRCN 0000019191, TRCN0000019192 and TRCN0000019193. A consistency score (C score) for each shRNA that represents the confidence that its observed phenotypic effects are the result of on-target gene suppression is provided on the website based on the ATARiS algorithm, with a higher value representing higher confidence. Based on the C score, the website suggests that only three shRNAs are acceptable: TRCN0000019189, TRCN0000019191 and TRCN0000019193. Therefore, we only included the three shRNAs in our analysis. Amplification data for the cell lines were downloaded from the Cancer Cell Line Encyclopedia (CCLE) project[73]. There were 18 common CRC cell lines in the two projects. For each shRNA, we calculated the Spearman's correlation coefficient between its effect score on the 18 cell lines and the log base 2 transformed copy number values for *HNF4A*.

As shown in Extended Data Fig. 8b–d, shRNAs simultaneously targeting both P1 and P2 isoforms (TRCN0000019189 and TRCN TRCN0000019191) showed a primarily negative impact on cell proliferation, whereas the P1-specific shRNA TRCN0000019193 showed mixed impacts. Interestingly, a stronger negative impact was associated with increased copy number, both for all shRNAs ($P = 0.01$, Spearman's correlation *P* values for individual shRNAs summarized by the Fisher's combined probability test) and for the P1-specific shRNA ($P = 0.04$, Spearman's correlation). These data suggest that the role of HNF4α is not only isoform-specific, but also depends on the status of *HNF4A* amplification. It is possible that the oncogenic role of HNF4α is primarily related to tumours with *HNF4A* amplification. Consistently, compared to normal colon samples, significant upregulation of HNF4α was only observed in a specific CRC subtype (Fig. 4d).

Proteomic subtype identification and characterization. *Proteomic subtype identification.* The normalized protein expression data set of 90 CRC samples was filtered to identify 1,263 proteins that were expressed (that is, with non-zero values) in at least 95% samples and also variably expressed among the 90 samples with a MAD value greater than 0.5. Based on the 1,263 selected proteins, we performed consensus clustering[24] implemented in GenePattern[74]. In consensus clustering, perturbations of the original data are simulated by resampling techniques. Clustering algorithm is applied to each of the perturbed data sets and the consensus among the multiple runs is assessed and summarized in a consensus matrix. Visual inspection of the consensus matrixes, and of the corresponding summary statistics (for example, area under the curve) can help determine the optimal number of clusters as described in the original publication[74]. The parameters used were set as follows: clustering algorithm = hierarchical clustering; clustering metrics = (1–Pearson correlation) distance and average linkage; *n* resamplings = 1,000; proportion of samples and proteins used in each resampling = 80%; k tested = from 2 to 8.

According to the consensus matrices and the empirical cumulative distribution function (CDF) plots shown in Extended Data Fig. 9a, b, $k = 7$ led to a clean consensus matrix and no obvious increase in clustering stability was observed going from $k = 7$ to $k = 8$. Thus, the 90 CRC samples were divided into seven clusters. As it is difficult to interpret the biological meaning of small clusters, only five clusters with more than five samples were kept. Thus, five subtypes were used for the following analysis.

Core sample identification. Following ref. 75, we defined the 'core samples' for each subtype as those with higher similarity to their own class than to any other classes and identified 79 core samples, as indicated by positive silhouette width scores[76] (Extended Data Fig. 9c).

Association with transcriptomic subtypes, methylation subtypes and genomic features. To associate the proteomic subtypes with genomic, epigenomic and clinical features of CRC, we downloaded the somatic mutation matrix (gene by sample) from the cBio cancer genomics portal[77] and obtained other sample information, including mRNA subtype, methylation subtype, hyper-mutation information, MSI information, stage information, cancer type, histological type and tumour site from the Supplementary Information table of TCGA CRC paper[6]. To obtain 18q loss information, we downloaded the CNV data from the Firehose website (http://gdac.broadinstitute.org, version 20130116). The association between each proteomic subtype and each feature was determined using a two-sided Fisher's exact test. The *P* values were corrected for multiple testing using the Benjamini–Hochberg procedure[65]. The results are summarized in Supplementary Table 11.

Proteomics data set for normal colon samples. Normal colon epithelium biopsies were obtained from screening colonoscopies performed between July 2006 and October 2010 under Vanderbilt University Institutional Review Board (IRB) approval no. 061096. During colonoscopy, multiple pinch biopsies were obtained from both ascending and descending colon and immediately frozen in liquid nitrogen. Biopsies obtained from 30 subjects with completely normal findings during colonoscopy were included in the study. A total of 60 specimens, one from ascending and one from descending colon, were used for proteomic analysis. Samples were processed identically to the colorectal tumour specimens as described

above and subjected to MS/MS analysis as described. The data were searched using the three search engines as outlined above. A protein assembly was made for the comparison of the TCGA tumours with the normal samples using IDPicker 3 as described in earlier at 0.2% PSM FDR and a minimum of 2 (distinct) spectra required per protein. The resulting assembly consisted of 6,044 confidently identified protein groups with a protein FDR of 4.1%. This data set was merged with the TCGA tumour data set using IDPicker 3, resulting in a combined tumour-normal proteomic data set of 7,548 protein groups with a protein FDR of 2.7% identified with a grand total of 9,028,208 filtered spectra. These proteins mapped to 7,244 unique genes. The data set were then subjected to quantile normalization, as described in the section on the protein quantification, followed by log transformation. Data from normal ascending and descending colon of the same individual were averaged, resulting in a final data set with 30 normal samples and 90 tumour samples. Only 3,718 quantifiable genes with a minimal average spectral count of 1.4 were included for downstream quantitative comparison (Supplementary Table 13). The median effect size in the data set for all 3,718 quantifiable genes is 0.96 as calculated by Cohen's d. According to power analysis, with sample sizes 30 and 90 and an effect size of 0.96, the power for detecting a difference at the significance level of 0.01 is 0.97 using a two-sided *t*-test.

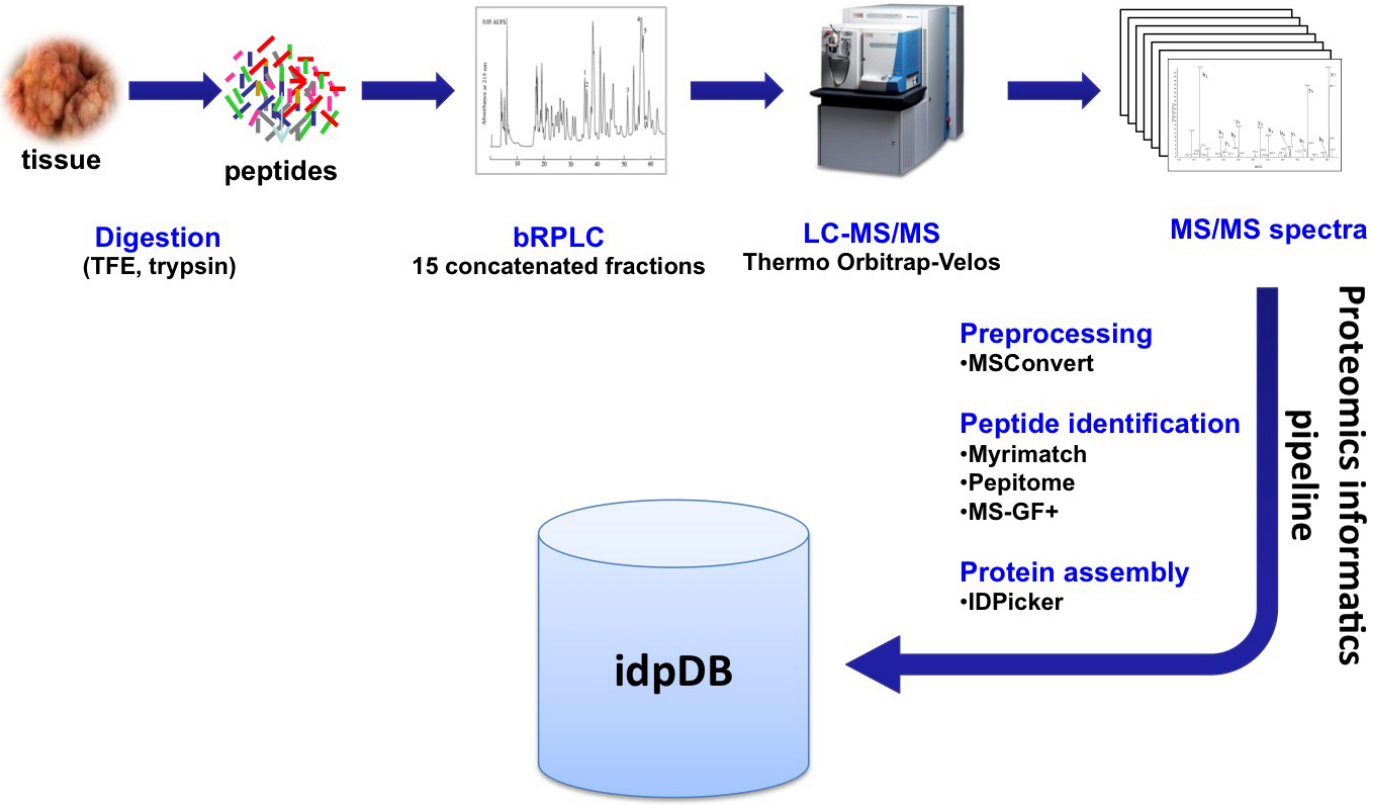
Subtype signature identification and Gene Ontology and network analysis. To identify protein signatures for individual proteomic subtypes, we compared protein expression in each subtype against all remaining subtypes. We also required signature proteins for a subtype to be significantly differently expressed in the subtype compared to normal colon samples. The analysis was based on data in Supplementary Table 13. A two-sided Wilcoxon rank-sum test was used for differential expression analysis. The *P* values were corrected for multiple testing using the Benjamini–Hochberg procedure[65] and the statistical significance was determined based on a corrected *P* value of less than 0.05. Signature proteins are listed in Supplementary Table 14.

To evaluate the generalizability of the proteomic subtypes and their signature proteins, we performed a leave-one-out cross validation. Specifically, one of the 79 samples was set aside and the remaining 78 samples and the normal samples were used to identify protein signatures and train a nearest shrunken centroid classifier for the proteomic subtypes using the R package *pamr*[78]. The trained classifier was then applied to the set-aside sample. This was repeated 79 times for all tumour samples and the cross-validation error rate was calculated. We obtained a low error rate of 3.8%, suggesting good generalizability of the proteomic subtypes and their signature proteins.

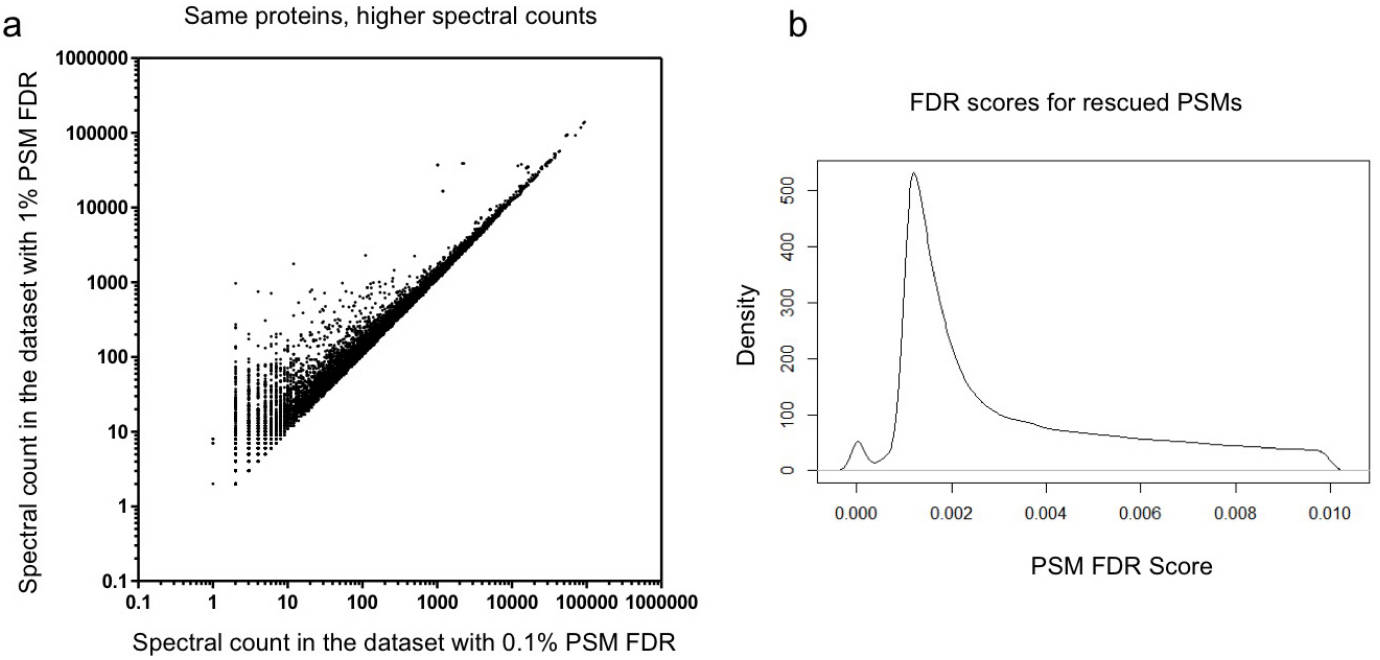
Gene Ontology enrichment analysis was carried out in WebGestalt[28,79] using the Fisher's exact test with an adjusted *P* value cutoff of 0.05, and enriched Gene Ontology terms were further analysed using the Gene-Ontology-function algorithm[80] to generate a parsimonious list of enriched terms (Supplementary Table 15).

Network analysis was performed in NetGestalt[30] using the iRef protein–protein interaction network[81] as a reference network. Using the NetSAM[30] algorithm, NetGestalt derived a linear order of all genes in the iRef network according to the hierarchical organization of the network and identified network modules at different hierarchical levels. As a result, expression data and gene lists can be co-visualized in the system as tracks and enriched network modules can be identified and visualized. Enriched network modules were identified based on the Fisher's exact test as implemented in NetGestalt. The Fisher's exact test *P* values were corrected for multiple testing using the Benjamini–Hochberg procedure[65] and the statistical significance was determined based on an adjusted *P* value of less than 0.01.

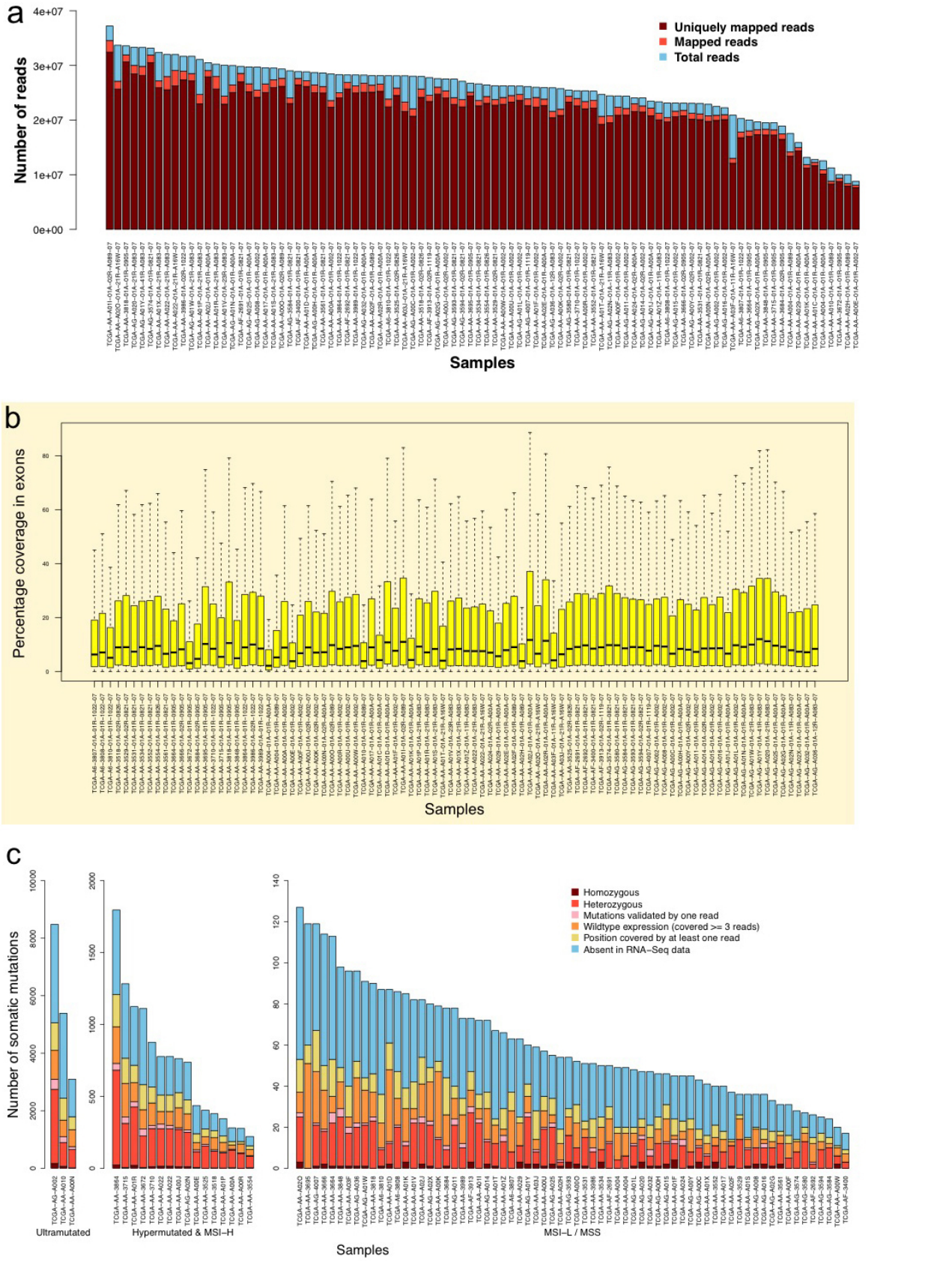
41. Licklider, L. J., Thoreen, C. C., Peng, J. & Gygi, S. P. Automation of nanoscale microcapillary liquid chromatography-tandem mass spectrometry with a vented column. *Anal. Chem.* **74,** 3076–3083 (2002).
42. Ma, Z. Q. *et al.* Supporting tool suite for production proteomics. *Bioinformatics* **27,** 3214–3215 (2011).
43. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnol.* **30,** 918–920 (2012).
44. Dasari, S. *et al.* Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment. *J. Proteome Res.* **11,** 1686–1695 (2012).
45. Tabb, D. L., Fernando, C. G. & Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **6,** 654–661 (2007).
46. Kim, S. *et al.* The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Mol. Cell. Proteomics* **9,** 2840–2852 (2010).
47. Ma, Z. Q. *et al.* IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.* **8,** 3872–3881 (2009).
48. Zhou, J. Y. *et al.* Improved LC-MS/MS spectral counting statistics by recovering low-scoring spectra matched to confidently identified peptide sequences. *J. Proteome Res.* **9,** 5698–5704 (2010).
49. Halvey, P. J., Zhang, B., Coffey, R., Liebler, D. C. & Slebos, R. J. Proteomic consequences of a single gene mutation in a colorectal cancer model. *J. Proteome Res.* **11,** 1184–1195 (2012).
50. Kislinger, T. *et al.* Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* **125,** 173–186 (2006).
51. Zhang, B. *et al.* Detecting differential and correlated protein expression in label-free shotgun proteomics. *J. Proteome Res.* **5,** 2909–2918 (2006).
52. Li, M. *et al.* Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling. *J. Proteome Res.* **9,** 4295–4305 (2010).
53. Ning, K., Fermin, D. & Nesvizhskii, A. I. Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. *J. Proteome Res.* **11,** 2261–2271 (2012).
54. Zybailov, B., Coleman, M. K., Florens, L. & Washburn, M. P. Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal. Chem.* **77,** 6218–6224 (2005).
55. Old, W. M. *et al.* Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteom.* **4,** 1487–1502 (2005).
56. Halvey, P. J. *et al.* Proteogenomic analysis reveals unanticipated adaptations of colorectal tumor cells to deficiencies in DNA mismatch repair. *Cancer Res.* **74,** 387–397 (2014).
57. Slebos, R. J. *et al.* Proteomic analysis of oropharyngeal carcinomas reveals novel HPV-associated biological pathways. *Int. J. Cancer* **132,** 568–579 (2013).
58. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19,** 185–193 (2003).
59. Liu, Q. *et al.* Integrative omics analysis reveals the importance and scope of translational repression in microRNA-mediated regulation. *Mol. Cell. Proteomics* **12,** 1900–1911 (2013).
60. Zybailov, B. *et al.* Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J. Proteome Res.* **5,** 2339–2347 (2006).
61. Gallien, S. *et al.* Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. *Mol. Cell. Proteom.* **11,** 1709–1723 (2012).
62. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26,** 966–968 (2010).
63. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).
64. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7,** 562–578 (2012).
65. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57,** 289–300 (1995).
66. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).
67. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12,** R41 (2011).
68. Wang, P. Statistical Methods for CGH Array Analysis. (VDM Verlag, 2010).
69. Darsigny, M. *et al.* Hepatocyte nuclear factor-4alpha promotes gut neoplasia in mice and protects against the production of reactive oxygen species. *Cancer Res.* **70,** 9423–9433 (2010).
70. Schwartz, B. *et al.* Inhibition of colorectal cancer by targeting hepatocyte nuclear factor-4α. *Int. J. Cancer* **124,** 1081–1089 (2009).
71. Saandi, T. *et al.* Regulation of the tumor suppressor homeogene Cdx2 by HNF4α in intestinal cancer. *Oncogene* **32,** 3782–3788 (2013).
72. Chellappa, K. *et al.* Src tyrosine kinase phosphorylation of nuclear receptor HNF4α correlates with isoform-specific loss of HNF4α in human colon cancer. *Proc. Natl Acad. Sci. USA* **109,** 2302–2307 (2012).
73. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
74. Reich, M. *et al.* GenePattern 2.0. *Nature Genet.* **38,** 500–501 (2006).
75. Verhaak, R. G. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17,** 98–110 (2010).
76. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20,** 53–65 (1987).
77. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2,** 401–404 (2012).
78. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA* **99,** 6567–6572 (2002).
79. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res.* **41,** W77–W83 W77–83 (2013).
80. Wang, J. *et al.* GO-function: deriving biologically relevant functions from statistically significant functions. *Brief. Bioinform.* **13,** 216–227 (2012).
81. Turner, B. *et al.* iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford)* **2010,** baq023 (2010).



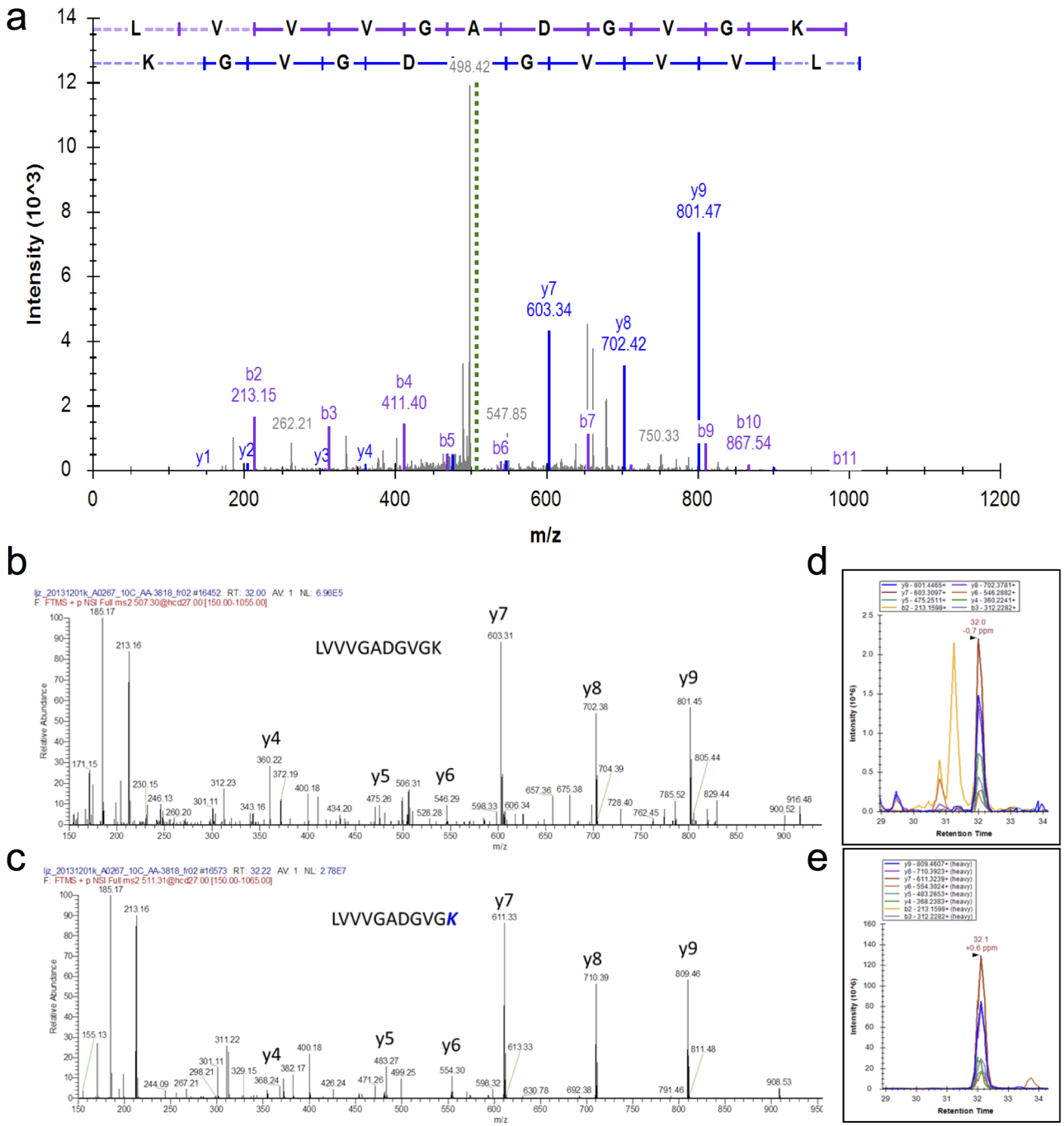
Extended Data Figure 1 | Mass-spectrometry-based proteomics workflow. Protein was extracted from frozen tumour tissue and used to generate tryptic digests. The resulting tryptic peptides were fractionated using off-line basic reverse-phase (high-pressure) liquid chromatography (basic RPLC). Collected fractions were pooled and used for reverse-phase HPLC in line with a Thermo Orbitrap-Velos MS instrument. Raw data were processed by MSConvert and then used for database and spectral library searching using three different search engines (Myrimatch, Pepitome and MS-GF+). Identified peptides were assembled using IDPicker 3 with selected filters as described in the methods. IDPicker 3 stores its protein assemblies for a specified set of filters in the idpDB format. These SQLite databases associate spectra with peptides, peptides with proteins, and LC-MS/MS experiments with a hierarchy of experiments.



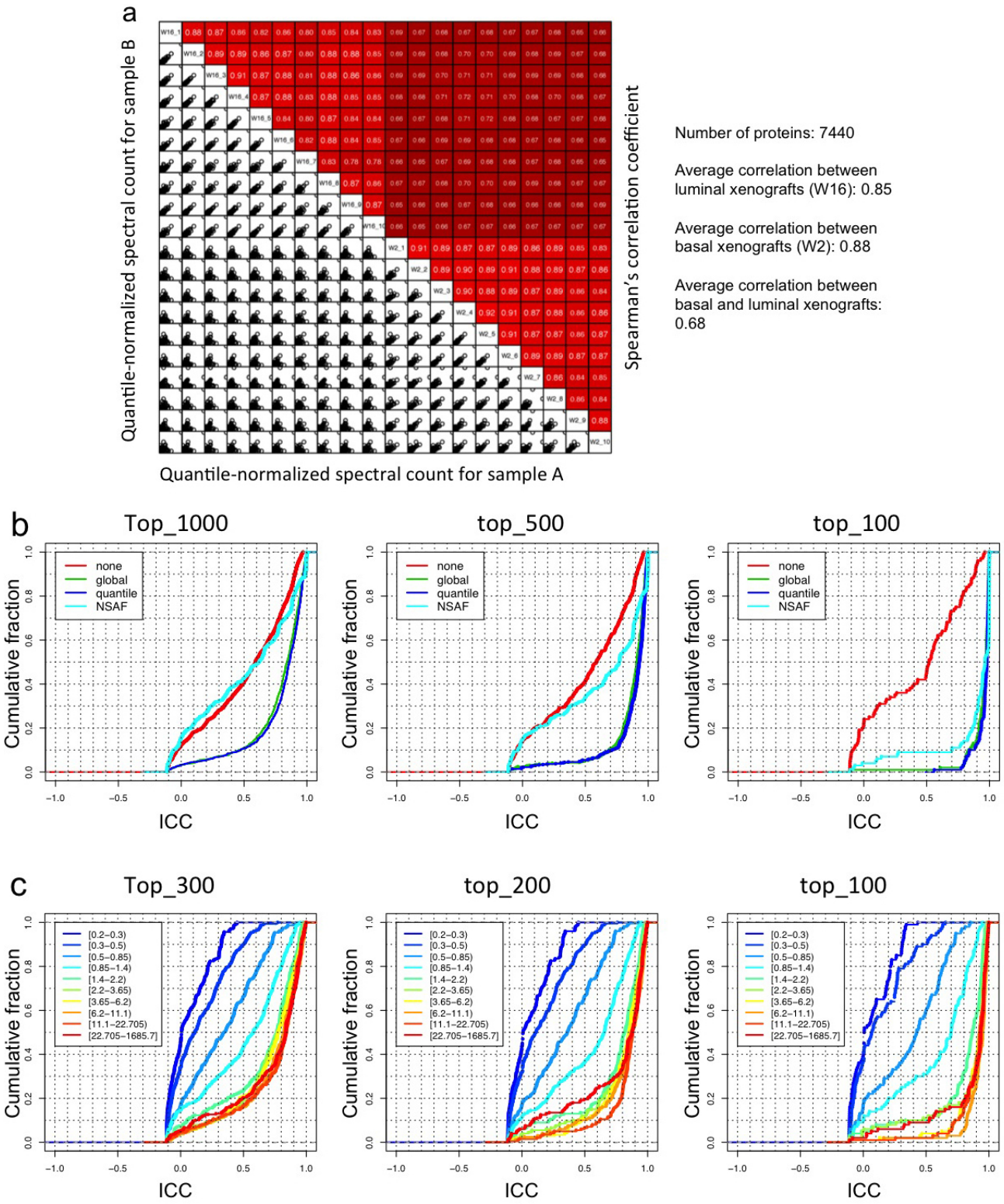
Extended Data Figure 2 | Relaxing the false discovery rate of peptide-spectrum match for high-confident proteins increases spectral counts. To increase spectral counts and improve statistical comparisons, we first created a protein assembly that maximized the number of proteins identified (at 0.1% peptide-spectrum match false discovery rate (PSM FDR)) and then relaxed the PSM FDR to 1% exclusively for the set of confidently identified proteins. This strategy led to increased spectral counts from 4,896,831 to 6,299,756, a 29% increase. **a**, Spectral count plot of all 7,526 confidently identified proteins demonstrates the increase in the absolute number of spectra identified for each protein, but no decrease for any of the proteins. Each dot in the figure represents one of the 7,526 proteins; *x* axis and *y* axis represent the spectral counts obtained in the data sets with 0.1% and 1% PSM FDR, respectively, both plotted on a log scale. **b**, Density plot showing the distribution of PSM FDR scores for all rescued PSMs. Rescued PSMs are of high quality with a median PSM FDR score of less than 0.2%, indicating the maintained integrity of the data set.



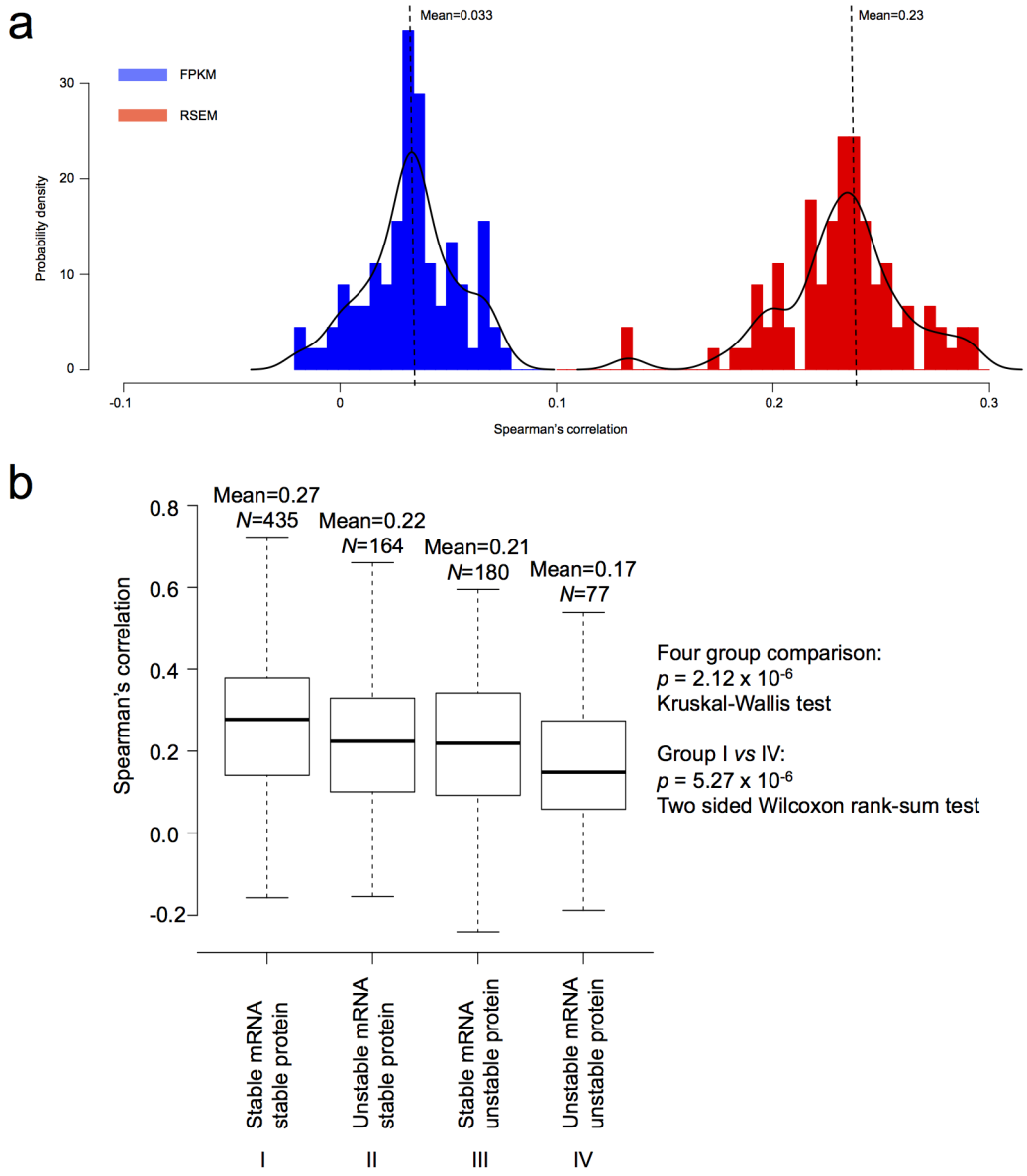
Extended Data Figure 3 | Read mapping, exon coverage and missense somatic variants in RNA-seq data. **a**, Summary of total RNA-Seq read counts and mapping results for individual samples. **b**, Distribution of percentage sequence coverage in exons for individual samples. Among all 228,157 exons, 76% were expressed, but only 64% had an average coverage greater than 1. Exons with no coverage were not included in the box plots. **c**, Number of missense somatic variants detected by RNA-seq in individual samples. Approximately 54% of the mutation positions were covered by RNA-seq reads and only 43% were covered by three or more reads.



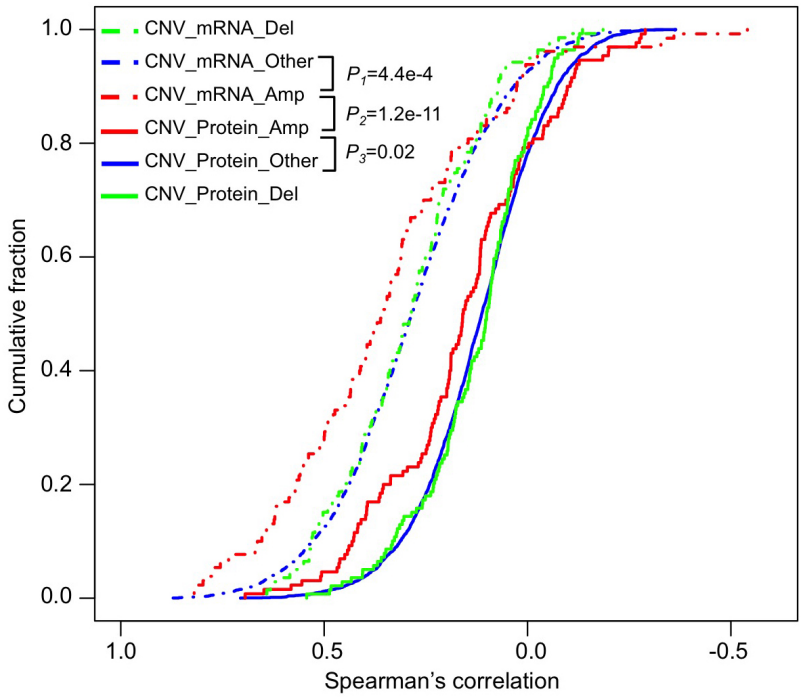
Extended Data Figure 4 | Parallel-reaction-monitoring validation results. Single amino acid variants (SAAVs) identified in the TCGA shotgun data set were validated using parallel-reaction-monitoring (PRM) analyses. Three distinct SAAVs in four TCGA samples were selected for validation. The TCGA samples were freshly prepared in the same manner as the original samples analysed by shotgun proteomics. Each sample was spiked with 12.5 fmol μl^{-1} of a mixture of all isotopically labelled peptides. Using an inclusion list containing the precursor *m*/*z* values representing both unlabelled (endogenous) and labelled peptides, each fraction was analysed by PRM for the variant peptides. This figure shows the PRM data for the variant sequence LVVVGADGVGK (*KRAS*(Gly12Asp) in TCGA-AA-3818. **a**, The MS/MS spectrum identified in the initial shotgun analyses. **b**, The annotated MS/MS spectrum of the unlabelled endogenous variant peptide in the PRM analysis. **c**, The annotated MS/MS spectrum of the spiked, labelled peptide in the PRM analysis. **d**, The chromatographic trace showing the overlapping transitions and retention time of the endogenous variant peptide. **e**, The chromatographic trace showing the overlapping transitions and retention time of the labelled variant peptide.



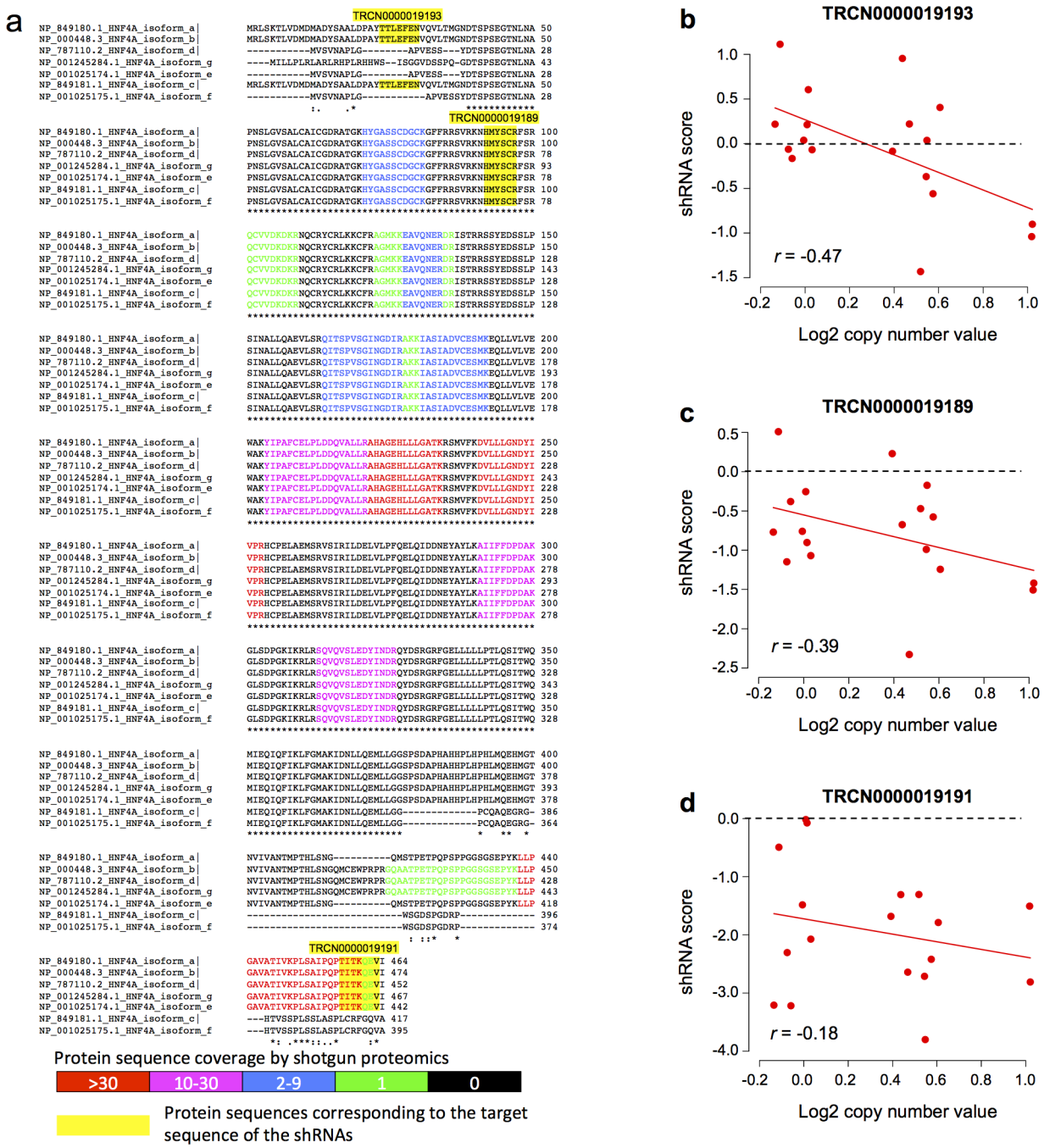
Extended Data Figure 5 | Platform evaluation and analysis method selection using quality control samples. **a**, The lower-left half (uncoloured) depicts pairwise scatter plots of the samples, with *x* and *y* axes representing quantile-normalized spectral counts for samples in corresponding columns and rows, respectively. The upper-right half (red) depicts pairwise Spearman's correlation coefficients for the same comparisons. **b**, For each normalization method (none, global, quantile and NSAF), we calculated the intraclass correlation coefficients (ICCs) for individual proteins in the quality control data set. The analysis was done for the top 1,000, 500 or 100 proteins with the largest variance and the cumulative fraction curves were plotted. In most scenarios, quantile normalization generated slightly higher ICC scores than global normalization, and both methods clearly outperformed the NSAF normalization. **c**, We sorted all proteins in the quality control data set based on their total spectral counts and then divided the proteins into 10 bins with equal number of proteins. Average spectral count ranges for each bin are shown in the brackets in the legend box. For each bin, we calculated the ICCs for individual proteins in the bin. The analysis was done for the top 300, 200 or 100 proteins with the largest variance in each bin. The cumulative fraction curves were plotted. Protein bins with spectral counts less than 1.4 showed clearly lower ICC scores, whereas the ICC score curves started to converge when the average spectral count was greater than 1.4.



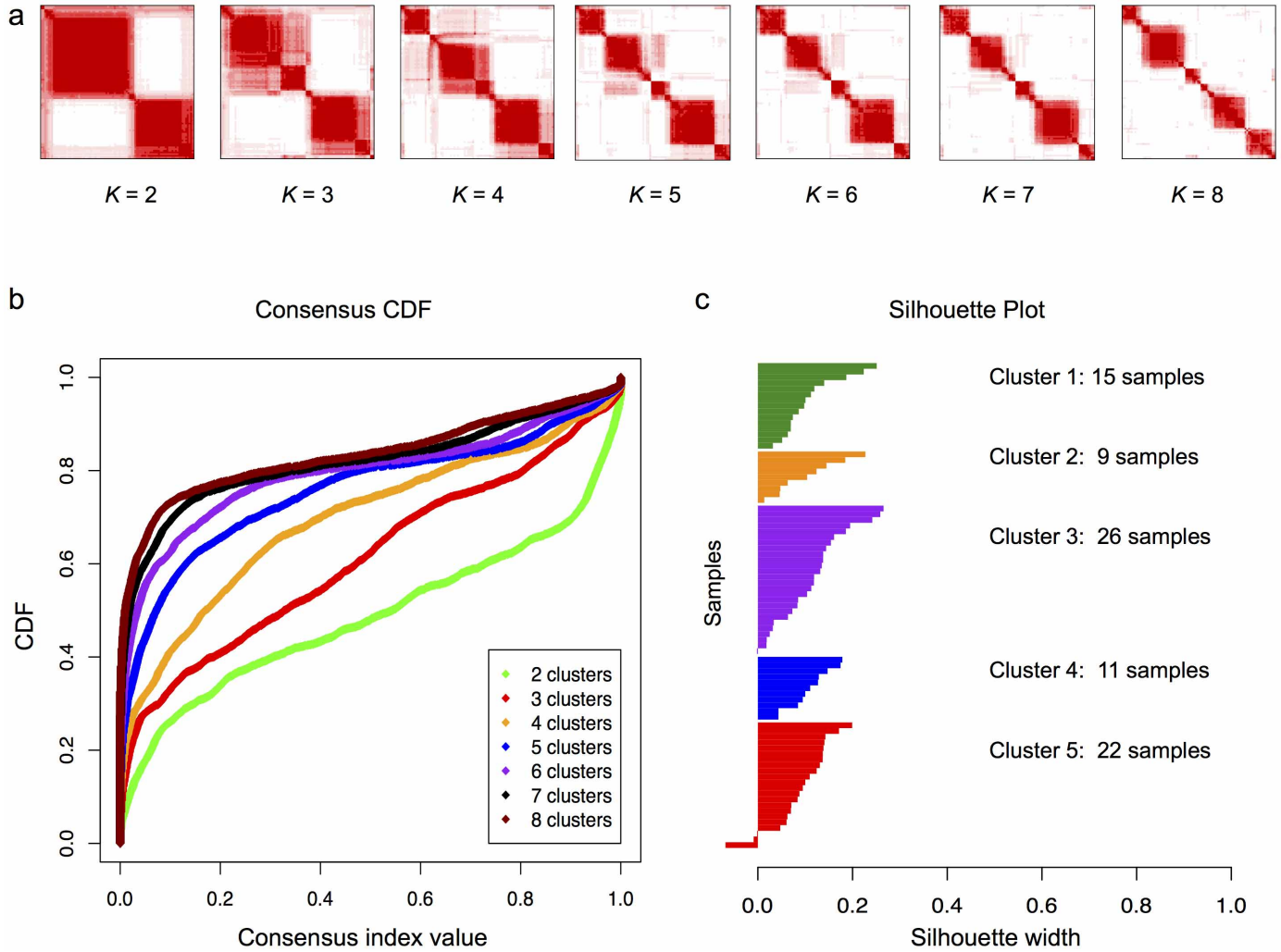
Extended Data Figure 6 | Extended data for mRNA–protein correlation analysis. **a**, Evaluation of the length bias in different RNA-Seq-based gene abundance estimation methods. The plot shows the distribution of correlation between gene length and estimated transcript abundance based on FPKM (fragments per kilobase of exon per million fragments mapped, blue curve) and RSEM (RNA-seq expectation maximization, red curve), respectively. FPKM measure is independent of gene length, whereas the RSEM measure strongly correlates with gene length. **b**, Relationship between mRNA–protein correlation and the stability of the molecules. Human genes were separated into four categories based on the mRNA and protein half-lives of their mouse orthologues: stable mRNA–stable protein; stable mRNA–unstable protein, unstable mRNA–stable protein, and unstable mRNA–unstable protein. Distribution of mRNA–protein correlations for genes in each category was plotted in the box plots. Genes with stable mRNA and stable protein showed relatively higher mRNA–protein correlation whereas those with unstable mRNA and unstable protein showed relatively lower mRNA–protein correlation. Only common genes in both our study and the mouse study were included in the analysis. The total number of genes in each category (*N*) is labelled in the figure. The *P* value indicating correlation difference among the four categories was calculated based on the Kruskal–Wallis non-parametric analysis of variance (ANOVA) test. The *P* value indicating correlation difference between the stable mRNA–stable protein group and the unstable mRNA–unstable protein group was calculated based on the two-sided Wilcoxon rank-sum test.



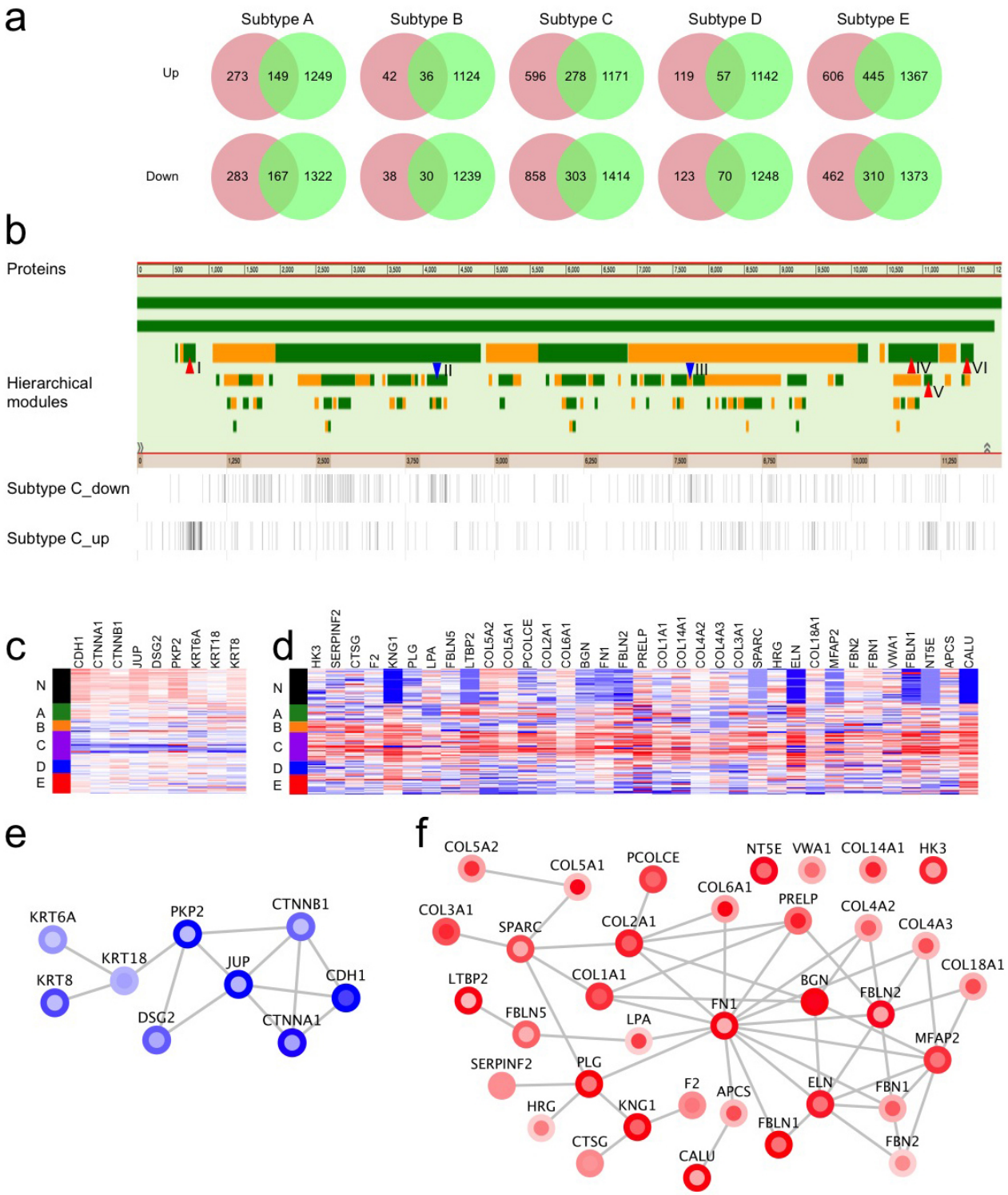
Extended Data Figure 7 | mRNA and protein-level *cis*-effect of copy number alterations in focal amplification, focal deletion and non-focal regions. The figure plots cumulative fraction curves of copy number alteration (CNA)–mRNA (dashed lines) and CNA–protein (solid lines) expression correlations for genes in the focal amplification regions (red), focal deletion regions (green), and non-focal regions (blue), respectively. Focal alteration regions were defined in the TCGA study. Any chromosomal regions outside the focal amplification and deletion regions were considered as non-focal regions. CNA–mRNA correlations were significantly higher than CNA–protein correlations for genes in any of the three groups. Moreover, genes in the focal amplification regions showed the highest level of CNA–mRNA and CNA–protein correlations among the three groups of genes. *P* values were based on the two-sided Kolmogorov–Smirnov test.



Extended Data Figure 8 | HNF4α isoforms and the effect of *HNF4A* shRNA on the proliferation of colon cancer cells. **a**, Multiple sequence alignment of the HNF4α isoforms, with peptides detected by shotgun proteomics and sequences corresponding to the shRNA target sequences highlighted. Different colours of the letters indicate different levels of sequence coverage in the shotgun proteomics study, as indicated by the colour scale bar. Yellow boxes highlight sequences corresponding to the shRNA target sequences. TRCN0000019193 specifically targets P1 promoter-driven isoforms, whereas the other two target both types of isoforms. **b–d**, The P1-HNF4α-specific shRNA showed mixed impacts (**b**), whereas shRNAs simultaneously targeting both P1 and P2 HNF4α showed a primarily negative impact on cell proliferation (**c**, **d**). Moreover, a stronger negative impact was associated with increased copy number, both for the P1- HNF4α specific shRNA ($P = 0.04$, Spearman's correlation (r)) and for all shRNAs ($P = 0.01$, Spearman's correlation P values for individual shRNAs summarized by the Fisher's combined probability test).



Extended Data Figure 9 | Consensus matrices, the empirical cumulative distribution function plot and core sample identification. **a**, Consensus matrices of the 90 CRC samples for $k = 2$ to $k = 8$. The consensus matrices show the robustness of the discovered clusters to sampling variability (resampling 80% samples) for cluster numbers $k = 2$ to 8. In each consensus matrix, both the rows and the columns were indexed with the same sample order and samples belonging to the same cluster frequently are adjacent to each other. For each pair of samples, a consensus index, which is the percentage of times they belong to the same cluster during 1,000 runs of the clustering algorithm based on resampling was calculated. The consensus index for each pair of samples was represented by colour gradient from white (0%) to red (100%) in the consensus matrix. **b**, Cumulative distribution function (CDF) plots corresponding to the consensus matrices for $k = 2$ to $k = 8$. This plot shows the cumulative distribution of the entries of the consensus matrices within the 0–1 range. Skew towards 0 and 1 indicates good clustering. As k increases, the area under the CDF is hypothesized to increase markedly until k reaches the k_{true}. In this case, 7 was considered as k_{true} because the change of the area under the CDF was close to zero when k increased from 7 to 8. **c**, Silhouette plot for core sample identification. For each sample (y axis), the silhouette width (x axis) compares its similarity to its assigned class and to any other classes. Samples with higher similarity to their assigned class than to any other classes will get positive silhouette width score and be selected as core samples.



Extended Data Figure 10 | Network analysis of the subtype signature proteins. **a**, The number of signature proteins for each subtype. For a given subtype, the red circle represents proteins that were different in abundance between the subtype and all other subtypes, the green circle represents proteins that were different in abundance between the subtype and normal colon tissues. The intersection between red and green circles contains the signature proteins for each subtype. **b**, Visualizing subtype-C-signature proteins in NetGestalt. Proteins in the iRef protein–protein interaction network are placed in a linear order together with the hierarchical modular organization of the network. Alternating bar colours (green and orange) are used to distinguish neighbouring modules. Proteins in the up and down signatures of subtype C were visualized as two separate tracks below the network modules, where each bar represents a protein. These proteins are not randomly distributed in the network. Highlighted by red or blue arrows are four Network modules (I, IV, V, VI) significantly enriched with up-signature proteins and two modules (II and III) significantly enriched with down-signature proteins (adjusted p value <0.01). **c**, **d**, Heat maps depicting relative abundance of down- and up-signature proteins of subtype C in modules III and I, respectively. Tumours are displayed as rows, grouped by normal controls (N) and proteomic subtypes (A–E) as indicated by different side bar colours. Proteins are displayed as columns. **e**, **f**, Network diagrams depicting the interaction of down- and up-signature proteins of subtype C in modules III and I, respectively. Node and node-border colours represent relatively higher or lower abundance in the subtype compared to other subtypes and normal colon tissues, respectively. Red and blue in the heat maps and the network diagrams represent relatively higher or lower abundance, respectively.