bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Chromosomal rearrangements are commonly post-transcriptionally attenuated in cancer

Emanuel Gonçalves [1], Athanassios Fraguais [2], Luz Garcia-Alonso [1], Thorsten Cramer [2], Julio Saez-Rodriguez [1#], Pedro Beltrao [1,2#]

[#] co-last authors: pedrobeltrao@ebi.ac.uk; saezrodriguez@gmail.com

[1] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge CB10 1SD, UK

[2] RWTH Aachen University, Faculty of Medicine, Joint Research Centre for Computational Biomedicine, Aachen 52057, Germany

Running title: Chromosomal rearrangement attenuation in cancer

Keywords: Cancer; Gene dosage; Proteomics; Copy-number variation; Protein complexes

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Chromosomal rearrangements, despite being detrimental to normal organismal fitness, are ubiquitous in cancer and often act as driver events. Copy number variations (CNVs) in the autosomal chromosomes are not known to have specific balancing mechanists but cancer cells can thrive with large dosage imbalances. Since the effect of CNVs on the cellular proteome of tumours is poorly understood, we analyzed recently generated proteogenomic data-sets on 282 patient samples of solid tumours, and investigated the implication of CNVs in the proteome of these cells. We found that the impact of CNVs is post-transcriptionally attenuated in 23-33% of the proteins with a strong enrichment for protein complex subunits. Interaction partners have a high co-regulation of their abundances and some complex subunits act as rate-limiting steps of complex assembly, indirectly controlling the abundance of other complex members. We identified 48 such regulatory interactions providing insight into assembly pathways of protein complexes involved in relevant biological processes in cancer. Lastly, we found that a gene-signature of the proteome attenuation is associated with increased resistance to chaperone and proteasome inhibitors. This study presents novel insights into the widespread importance of post-transcriptional mechanisms in cancer that allow cells to cope with their altered genomic profile.

Introduction

Cancer development is driven by the acquisition of somatic genetic variation that includes point mutations, copy number variations (CNVs) and larger chromosome rearrangements or duplications (i.e aneuploidy) (Beroukhim et al., 2010). These events can result in a fitness advantage and cancer progression but they are most often detrimental to cellular fitness. While

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

somatic gene amplification of key oncogenes such as MYCN, AKT2, ERBB2 and others (Santarius et al., 2010) can drive cancer development, germline CNVs are rare and are under negative selection (Itsara et al., 2009). Gene amplifications and other CNVs are thought to be detrimental due to changes in gene-expression that cause an imbalance to the cell. In females, one of the two X chromosomes is inactivated by a specialized RNA based silencing mechanism (Avner and Heard, 2001; Lyon, 1961) but such mechanism does not exist for gene dosage imbalances in the autosomal chromosomes. Protein and mRNA abundance measurements in models of aneuploidy in yeast (Dephoure et al., 2014) and human cells (Stingele et al., 2012) have shown that most autosomal gene duplications are propagated to the protein level, with the notable exception of protein complex subunits (Dephoure et al., 2014; Stingele et al., 2012). The lack of attenuation of changes in expression resulting from duplicated chromosomes causes global stress responses that include cell-cycle and metabolic defects and proteotoxic stress among others (Tang and Amon, 2013). While somatic CNVs are known to be drivers of cancer development and that aneuploidy is a common feature of tumor cells, the impact of gene dosage changes on the proteome of cancer cells have yet to be studied.

Comprehensive characterisation of hundreds of tumour samples at the genomic and transcript level have been instrumental for the study of tumour heterogeneity across patients and to assess the implications of genomic alterations to cancer development (Cancer Genome Atlas Research Network et al., 2013). Functional annotation of genomic events have been limited to the transcript level until very recently, when the first broad proteomics studies with deep coverage of the proteome were made available (Mertins et al., 2016; Zhang et al., 2014, 2016).

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In this study, we investigated the implications of CNVs on the proteomes of tumours by taking advantage of the comprehensive data-sets made available by the TCGA and CPTAC consortiums comprising copy-number, transcript and protein measurements for hundreds of tumours (Cancer Genome Atlas Network, 2012a, 2012b; Cancer Genome Atlas Research Network, 2011; Mertins et al., 2016; Zhang et al., 2014, 2016). This data revealed that CNVs are often propagated to the protein level although we observed that post-transcriptional mechanisms attenuate this impact in 23-33% of the measured proteins. Protein complexes were notably attenuated, likely due to the degradation of free subunits, resulting in strong protein abundance coregulation across samples. Not all complex subunits are attenuated, with some acting as potential rate-limiting factors for complex assembly, and we identified 48 novel and known regulatory interactions, whereby the abundance of one of the subunits can modulate the abundance of other complex members. In addition, ranking the samples by their potential to attenuate gene dosage effects identifies putative mechanisms involved in autosomal gene dosage compensation. Finally, a gene expression signature of attenuation potential was found to be associated with drugs targeting chaperones, the proteasome and the E3 ligase MDM2. Using 282 tumour samples we revealed the widespread importance of post-transcriptional mechanisms to ameliorate the impact of CNVs in cancer cells.

Results

Tumour pan-cancer proteomics reveals attenuation of copy-number alterations in protein complex subunits

In order to study the implication of gene dosage changes on the proteome of cancer cells we compiled and standardized existing data-sets made available by the TCGA and CPTAC

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

consortiums, comprising 3 different cancer types: breast (BRCA) (Cancer Genome Atlas Network, 2012b; Mertins et al., 2016), high-grade serous ovarian (HGSC) (Cancer Genome Atlas Research Network, 2011; Zhang et al., 2016) and colon and rectal (COREAD) (Cancer Genome Atlas Network, 2012a; Zhang et al., 2016) (Figure 1A). These data-sets provide molecular characterisation of gene CNVs, gene expression and protein abundance of solid tumour samples of 282 patients for which clinical information is also available (Figure 1A, Supp. Table 1).

Current methods can reliably measure the complete expressed transcriptome in a cell but measuring the total proteome is still a challenge with current techniques only providing partial snapshots (Nagaraj et al., 2011). Thus, we quantified the fraction of expressed transcripts measured in the proteomics experiments in each tumour sample (Figure 1B) (see Methods). COREAD samples displayed the lowest average coverage of the expressed transcriptome (22.3%) compared to the coverage measured for the HGSC (42.0%) and BRCA (56.1%) samples. The proteomics experiments were not conducted using the same methodologies, and therefore it is crucial to take in consideration the potential confounding effects. In particular, the COREAD (Zhang et al., 2014) quantifications were done with a label free approach while the HGSC and BRCA were quantified using isobaric labelling (Mertins et al., 2016; Zhang et al., 2016). To ensure comparable measurements among data-sets we removed confounding and systematic effects from the proteomics and transcriptomics, by regressing-out batch effects associated with experimental technologies used, patient gender and age and tumour type (see Methods). The associations between these possible confounding factors and the principal components were completely removed after correction (Figures S1 and S2).

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Having assembled this compendium of data-sets we then set out to understand the implication of CNVs events in the expression of the proteome (Figure 1C). We calculated for each gene the agreement between the CNVs and transcriptomics and proteomics using the pearson correlation coefficient (Figure 1D). Transcript abundance is on average well correlated with gene CNVs changes (median pearson's r=0.43) and this contrasts with the significant decrease (Welch's t-test p-value < 1e-4) of agreement of CNVs with protein abundance (median pearson's r=0.20) (Figure 1D, Supp Table 2). As transcription is intermediate between the copy-number alterations and protein abundance, it sets the maximum possible agreement between both. We can then define as attenuated genes those that have a lower agreement between CNVs and protein abundance than expected by their CNV to gene-expression correlation (see Methods). In these samples we found 1,496 - 2,119 genes that are significantly attenuated by this definition, corresponding to 23-33% of all genes with available measurements (6,418). This result shows that a significant fraction of the proteome undergoes gene dosage balancing. Additionally, this group of attenuated genes highlight the complexity of the regulation of protein abundance, hinting at regulatory constraints that control protein translation or degradation rates.

To understand the biological processes that are affected by this attenuation we performed an unbiased enrichment analysis using GO terms (Ashburner et al., 2000; Subramanian et al., 2005; The Gene Ontology Consortium, 2015) (Figure 2A) (see Methods). The enrichment analysis revealed that proteins involved in complexes and modules of functionally interacting proteins displayed a significant agreement at the transcript with the copy-number measurements but this agreement is generally lost at the protein level (Figure 2B). This attenuation of protein complex subunits recapitulates previous findings in models of aneuploidy

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in yeast (Dephoure et al., 2014) and human cell lines (Stingele et al., 2012) showing that the underlying mechanism generalizes from the aneuploidy models to the hundreds of patient tumour samples studied here. In order to validate the generality of the set of attenuated genes we confirmed that these are also recapitulated in an independent proteomic panel of triple negative breast cancer and ovarian cancer cell lines (Coscia et al., 2016; Lawrence et al., 2015) (Figure 2C). This attenuation mechanism has been shown, in yeast aneuploid strains, to be mostly due to control of protein abundance by degradation (Dephoure et al., 2014). For protein complexes in particular, this result fits with a model where subunits are degraded when free from the complex (Abovich et al., 1985). To test if degradation plays a role in the attenuation observed in human cells, we used publically available data on changes in protein ubiquitylation after proteasome inhibition as markers of degradation (Kim et al., 2011). We observed that genes defined as attenuated in our study show a faster increase in ubiquitylation after proteasome inhibition than other genes (Figure 2D), suggesting that degradation plays a key a role in the attenuation of these genes. Recently some studies have shown, in degradation time-series experiments, that many protein complex subunits have degradation profiles that are best fit by a two-state model, suggesting that the degradation rate of these proteins changes, presumably when free or when assembled into the complex (McShane et al., 2016). These results suggest that the abundance of protein subunits of large stable protein complexes are under active control to maintain their co-regulation, possibly to guarantee the stability and formation of the associations or prevent the accumulation of free subunits that might be prone to aggregate.

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Proteomic correlation analysis uncovers strong co-regulation of protein complexes

Considering that protein complexes seem to be coregulated to maintain their abundance, likely via the degradation of free subunits, we systematically explored this by performing protein-protein correlation analysis using the proteomics measurements (Figure 3A). We performed all possible pair-wise correlation of protein abundance for all the 6,434 proteins measured in at least 50% of the samples across the 3 different tumour types (see Methods). Consistently, proteins within the same complexes display coordinated changes of abundance across samples (Figure 3A). Then, we assessed if this co-regulation effect is ubiquitous in a curated set of human protein complexes from the CORUM database (Ruepp et al., 2010). Pairs of proteins present together in a complex display a degree of co-regulation (mean pearson's r=0.25) that is significantly higher than what is observed for random pairs (mean pearson's r=0). We also assessed if this co-regulation was visible at the transcript level, and while there is a significant increase over random associations (mean pearson's r=0.15) this correlation is significantly lower than the one seen at the protein level (Figure 3B). Protein pairs that have functional interactions but are not complex subunits show a lower degree of abundance correlation (pearson's r=0.15) that is also closer to the observed at the transcript level (pearson's r=0.11) (Figure 3B).

In light of this agreement between functionally related proteins we examined the capacity of protein-protein correlation profiles to predict known protein-protein interactions (Figure 3C) (see Methods). We found that direct and indirect functional interactions could be well identified with proteomics (AROC 0.86 and 0.75, respectively), and significantly worst predicted with

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

transcriptomics (AROC 0.69 and 0.67, respectively) (Figure 3C). This finding goes in line with a recent work that showed that proteins within similar biological processes or pathways display better agreement at the protein than at the transcript level (Wang et al., 2016). We noticed that protein interactions derived from signalling networks displayed in general poor agreement at the protein and transcript abundance levels (AROC 0.55 and 0.54) (Figure 3C). Furthermore, metabolic enzymes involved in the same metabolic pathways displayed some degree of agreement at the protein and transcript level (AROC 0.65 and 0.62) (Figure 3C).

Proteogenomics analysis identifies subunits that control the protein abundance levels of other members of the complex

Previous studies conducted in aneuploidy models have indicated that, while protein complex members tend to be attenuated, not all subunits are. It has been hypothesized that these non-attenuated subunits could act as scaffolding proteins or rate-limiting for the assembly of the complex (Dephoure et al., 2014). However, these studies were conducted on a small number of yeast strains or cell lines (Dephoure et al., 2014; Stingele et al., 2012). Given the large number of tumour samples analyzed here we reasoned that we could more readily identify such subunits that can act as drivers of complex assembly**.** To study this we assessed if CNVs of a given protein within the complex could explain the changes in abundance of other subunits once we discount their transcriptional changes (see Methods). In other words, if the presence or absence of certain proteins of the complex could be associated with the protein degradation rate of other members. This was performed systematically for all identifiable protein-pairs within protein complexes using linear regression models where the CNVs of a protein (Px) was used to estimate the protein abundance variation of the corresponding paired protein (Py) (Figure 4A) (see Methods). To consider the degradation levels of the protein, the transcript measurements

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

were regressed-out from the protein abundance (Figure 4A) (see Methods). This allowed us to consider the variability arising only from post-transcriptional regulatory events and, importantly, to discard possible confounding regulatory events occurring at the genomic and transcript level, such as close genomic localization. Out of the 58,627 possible directed protein interactions 64 were found to be significantly associated (FDR < 5%) (Figure 4A, Supp Table 3) (see Methods). To ensure that the association was not only visible at the genomic level but also at the transcript level, the same associations were performed using transcriptomics data instead of CNVs. As expected since that transcript abundance is a closer measurement to the protein abundance, we found a substantial increase of significant associations, 2,846 (FDR < 5%) (Figure S3, Supp Table 3). Also, 75% (48) of the associations found at the genomic level were found to be significant at the transcript level (Figure 4A, Supp Table 3).

This analysis identified putative regulators of the assembly of protein complexes that are potentially rate-limiting for the assembly of the full complex. We found, for example, an association between the copy-number of COG3 and the protein variability of COG2 (pearson's r=0.39, p-value 9.90e-12) (Figure 4B). COG3 is also significantly associated with COG4 (Figure 4B) increasing the possibility that COG3 is a regulator of the assembly of the COG complex. These findings are corroborated by an existing study where COG3 knockdown leads to a decrease abundance of COG2 and COG4 (Zolov and Lupashin, 2005). Besides hinting at known regulatory interactions our analysis also points to possibly novel interactions within the COG complex with COG6 being significantly associated with COG2 (Figure 4A, C). Additional positive regulatory interactions were found for subunits of the eukaryotic Initiation Factor 3 (eIF3), Transcription Factor IIH (TFIIH), Adaptor Related Protein Complex 3 (AP3), among

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

others (Supp Table 3), providing with information on the putative assembly pathways of these complexes.

The number of significant negative associations was remarkably lower than the number of positive associations (Figure 4A, Figure S3 C). SMARCA2 copy-number alterations were significantly negatively associated with the degradation of SMARCA4 (Figure 4A) and this was also visible at the protein level (Figure 4B). Negative associations are likely to represent mutually exclusive events of the protein complex regulation, thus when one protein is present the other will not be necessary for the complex formation and may undergo degradation. Indeed, current evidence in the literature seems to suggest that SMARCA2 and SMARCA4 are mutually exclusive elements of the SWI/SNF complex (Karnezis et al., 2016). The lower number of negative associations suggests that these types of events are less frequent.

Besides providing us with clues on the assembly pathways of protein complexes, these associations provide concrete examples of trans protein quantitative loci associations (pQTLs) (Chick et al., 2016), illustrating how the genomic variation can have consequences through protein interaction partners.

Protein complex amplifications associated with high attenuation potential

Having assessed the attenuation of CNVs effects in the proteome we set out to quantify the extent of this regulation in each tumour sample. We reasoned that by stratifying the samples by their capacity to attenuate the CNV changes, we could identify the underlying attenuation mechanisms. Similarly to the protein analysis (Figure 1D) we performed a correlation analysis between the CNVs and transcriptomics and proteomics for each sample (Figure 5A), instead of

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

each protein. Furthermore, recurring to a gaussian mixture model we classified 50 samples (18%) as those having a general strong attenuation effect (see Methods). Such tumour samples have a higher number of genes with strong attenuation, suggesting either an overall increase in degradation or decrease in translation rates in these samples. We then searched for complexes and complex subunits that are more likely to be amplified in the tumours with stronger attenuation and could therefore contribute to the attenuation potential (see Methods). Tumours with strong attenuation effects displayed a significant enrichment of gene amplifications in several complex subunits including genes involved in the endoplasmic reticulum-associated degradation (ERAD) pathway (DERL1 and VIMP), cell polarity (SCRIB, LLGL2 and VANGL2), GPI-anchor biosynthesis (PIGT and PIGU) and RNA interference (AGO2) (Figure 5B and 5C). SCRIB protein complexes have been previously reported to play an important role in cancer progression in breast cancer and their inhibition has been linked to a decrease in cell migration (Anastas et al., 2012). The proteasome system is important for the regulation of focal adhesions in migrating cells (Teckchandani and Cooper, 2016) and inhibition of the proteasome inhibits migration and invasion in breast cancer cells (Xie et al., 2009). However, it is not clear how the overexpression of these cell polarity factors would result in an increase in attenuation potential. Interestingly and consistently with the increased protein attenuation profile of these tumours, we observe amplifications of the ERAD components DERL1 and VIMP, that are part of an endoplasmatic reticulum (ER) complex which is responsible for the retrotranslocation of misfolded proteins to the cytosol for proteasomal degradation (Lilley and Ploegh, 2004; Ye et al., 2004). The association between increased attenuation and amplification of AGO2 could be explained by its role in repressing the initiation of mRNA translation (Kiriakidou et al., 2007).

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Gene expression profile of protein attenuation is associated with specific drug responses

Since the tumours with strong attenuation of CNVs effects displayed particular characteristics, we defined a gene expression signature by systematically correlating each gene with the attenuation potential (see Methods). This signature provides a putative ranking of the agreement between gene expression and the protein attenuation profile of the samples. Next, we explored the capacity of this signature to identify particular cellular states that can be informative for drug response. Samples with a strong correlation with the signature would be predicted to have higher attenuation and could, for example, display a higher proteasomal capacity. Thus, we considered an independent cell line panel for which gene expression and drug response is available (Iorio et al., 2016b), and ranked the cell lines according to their predicted protein attenuation potential (see Methods). Then we assessed the association between this predicted attenuation potential and drug response measurements for 265 compounds (see Methods) (Figure 5D and 5E). Among the top predicted compounds are a proteasome (Bortezomib) and chaperone inhibitors (AUY922, 17-AAG, Elesclomol, CCT018159) which displayed a significant (FDR < 5%) positive association, suggesting that a stronger predicted attenuation potential is associated with increased resistance to proteasome/chaperone inhibitors (Supplementary Table 4). This unbiased search also revealed significantly positive associations of Nutlin-3a and JNJ-26854165 and the proteome attenuation profile (pearson's r 0.20 and 0.16, respectively). Interestingly, both compounds target the oncoprotein E3 ligase murine double minute 2 (MDM2) which, in p53 wild type tumours, suppresses the activity of p53 by ubiquitination and thereby is a potential therapeutical target (Shangary and Wang, 2008). This analysis suggests that the gene expression signature for the

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

proteome attenuation may be associated with an increased capacity of the protein quality control machinery and an increased resistance to drugs that target this system.

Discussion

Although chromosomal rearrangements is an ubiquitous feature of cancer cells, still little is known about their impact on the cellular phenotype, in particular on the regulation of the proteome. Recent proteogenomics studies released by the CPTAC consortium (Cancer Genome Atlas Research Network et al., 2013; Mertins et al., 2016; Zhang et al., 2014, 2016) presented for the first time a broad compendium of proteogenomics measurements across tumours, enabling a systematic analysis of this question.

In this study, we observed that while CNVs have a good agreement with transcript measurements, 23-33% of the proteins undergo post-transcriptional regulation which attenuates the impact of CNVs (Figure 1 C, D). This information alone is very relevant for the identification of causal genes within amplified genome regions. Since amplifications of the attenuated genes are not observed at the protein level, these are less likely to be drivers of cancer progression and similarly less likely to explain changes in drug associations. Notably, this attenuation was more pronounced in protein subunits and complexes, in agreement with previous observations (Dephoure et al., 2014; Stingele et al., 2012). This is likely explained by the fact that the stoichiometry of complexes need to be preserved and that proteins over-represented compared to other members of the complex are likely degraded due to increased instability (McShane et al., 2016). Furthermore, we observed that proteins with stronger attenuation are more quickly ubiquitinated (Kim et al., 2011) (Figure 2D) suggesting that the attenuation may be mostly driven by changes in degradation instead of translation rates.

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Protein complexes subunits abundance co-regulation was a general feature and this agreement was not so strikingly visible at the transcript abundance, recapitulating recent findings with the same data-set (Wang et al., 2016). Within these subunits we identified 48 putative-regulators, that may act as rate-limiting factors for complex assembly, capable of regulate the abundance and assembly of other complex subunits (Figure 4A). This systematic analysis recapitulated known, COG3 - COG2 and COG3 - COG4, and novel, COG6 - COG2, regulatory interactions within the conserved oligomeric Golgi (COG) complex, which is located in the golgi apparatus and involved in protein sorting and glycosylation (Smith & Lupashin, 2008). COG3 associations were validated in an independent study where knocking down COG3 leads to the depletion of the interacting proteins (Zolov and Lupashin, 2005). Other putative-regulatory interactions were proposed within complexes with important roles in cancer, such as EIF3A regulation of several members of the eIF3 complex, which plays an important role in translation initiation and regulation of protein synthesis (des Georges et al., 2015; Zhang et al., 2007). Also, a regulatory interaction between RPA2-RPA3 was reported, these proteins are part of the replication protein A (RPA) complex which is involved in the response to DNA damage and thereby is involved in the control of DNA repair mechanisms and the activation of DNA damage checkpoints (Zou et al., 2006). Interestingly, potential mutual exclusivity associations were present in much lower numbers. The most compelling negative association was SMARCA2 - SMARCA4, which was supported by current literature where the two are reported to be mutual exclusive ATPases within the SWI/SNF complex, and dual deficiency of these proteins induces differentiation from a normal cell to high-grade tumour (Karnezis et al., 2016). These results provide examples and putative mechanistic explanations for how variation in copy number of gene expression of a

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

protein can have trans effects in the abundance of interacting proteins, as seen in protein quantitative trait loci analyses (Chick et al., 2016).

Tumour samples with strong attenuation of CNVs effects in protein abundance displayed a significant enrichment for amplifications of several protein complexes involved in the response to misfolded proteins in the ER, cell polarity, trafficking and gene repression. The amplification of genes associated with cell polarity in cells with increased attenuation potential would suggest that increased cell migration might result in an increased proteasomal function or decreased translation rates. We derived a gene expression signature to characterise the attenuation effect and used this to associate with drug responses in an independent large panel of approximately 1,000 cell lines and 265 compounds (Iorio et al., 2016b) (Figure 5D). Proteome attenuation was associated with increased resistance to proteasome and chaperone inhibitors, suggesting that tumours where attenuation is more pronounced are more resistant to perturbations in the chaperone/proteasome system. Interestingly, the two compounds in the screen targeting MDM2 were among the top associated with the gene expression signature, suggesting that tumours with high predicted attenuation potential may have a high proteasome capacity and therefore be less sensitive to the inhibition of MDM2, that is the E3 ligase responsible for the degradation of p53 in TP53 wild-type tumours (Shangary and Wang, 2008).

In this study, we provide novel insights into how cancer cells regulate their proteome in the presence of abnormal chromosomal rearrangements, in particular of copy-number alterations. This study emphasises the importance of integrating multiple types of biological data to allow functional assessment of genomic alterations. While relevant insights were possible here, more tailored experimental work should be carried-out to understand the mechanisms of how are the

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

co-regulatory effects of complexes maintained. Mutations can have similar implications to copy-number alterations by affecting the correct synthesis of the protein or its transcript, but these are generally harder to interpret. We have uncovered a potential way in how cancer cells manage to cope with often dramatic chromosomal rearrangements (Thompson and Compton, 2011) and these can possibly provide insights into their functional implications and hopefully open novel therapeutic opportunities.

Methods and materials

Data compendium

Proteomics measurements at the protein level for the three tumour types analysed here were compiled from the the CPTAC data portal (Edwards et al., 2015) (accession date 2016/07/06) for the following publications: BRCA (Mertins et al., 2016), HGSC (Zhang et al., 2016) and COREAD (Zhang et al., 2014). Transcriptomics RNA-seq raw counts were acquired from (Rahman et al., 2015) (GSE62944) and processed using the Limma R package (Ritchie et al., 2015) with the voom transformation (Law et al., 2014). GISTIC (Mermel et al., 2011) thresholded copy-number variation measurements and clinical data were obtained from the http://firebrowse.org/ portal (accession date 2016/06/08).

Data processing and normalisation

Transcriptomics raw counts were downloaded from (Rahman et al., 2015) (GSE62944). To ensure that lowly expressed transcripts are removed, genes with average counts per million (CPM) across samples lower or equal to 1 were excluded. Data was normalised by the trimmed mean of M-values (TMM) method (Robinson and Oshlack, 2010) using edgeR (Robinson et al.,

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2010) R package. Finally, the log-CPM values derived from the voom (Law et al., 2014) function in Limma (Ritchie et al., 2015) package were extracted for this analysis.

Coverage of the proteomics samples was assessed using the jaccard index for each sample with matching transcriptomics. Transcriptomics and proteomics measurements were used at the gene symbol level annotation. For each sample it was only considered transcripts passing the expression threshold, defined above, and proteins with matching measurement. The jaccard index for each sample was calculated with the intersection over the union.

Considering that proteomics and transcriptomics principal component analysis (PCA) revealed associations with possible confounding factors, i.e. age, gender, tumour type and measurement technology, we regressed them out from the original data-sets using linear regression models (Figure S1). For each protein a multiple linear regression model was fitted with protein measurements across the tumour samples as the dependent variable and the confounding factors mentioned above as independent discrete variables, apart from the age which was represented with a continuous variable. Once the model was fitted the estimated weights were used to regress-out their impact in the protein measurement and thereby removing their effects (Figure S2). Due to the sparseness of mass-spectrometry measurements for the proteomics data-set we only considered proteins that were consistently measured in at least 50% of the samples, leaving a total of 6,734 proteins. The same procedure was performed in the transcriptomics measurements. Transcript and protein measurements were normalised and centred across the samples using a gaussian kernel density estimation function.

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Proteome attenuation analysis

Agreement between the copy-number variation and the transcriptomics and proteomics was calculated for each gene/protein across the tumour samples using pearson correlation coefficient. Enrichment of biological processes for proteins displaying an attenuation of the correlation at the protein level compared to the transcript level was performed using Gene Set Enrichment Analysis (GSEA) approach (Subramanian et al., 2005). For the enrichment the protein attenuation levels was used, which is calculated by the difference between the pearson coefficient of the transcript correlation (correlation between copy-number variation and transcript measurements) and the pearson coefficient of the protein correlation (correlation between copy-number and protein measurements). To ensure a normal distribution centered around zero for the GSEA enrichments a gaussian kernel density estimation function was used to normalise the protein attenuation distribution. Gene signatures of Gene Ontology (GO) (Ashburner et al., 2000; The Gene Ontology Consortium, 2015) terms for biological processes (BP) and cellular compartments (CC) were acquired from the MSigDB data-base (Subramanian et al., 2005). Gene signatures of post-translational modifications (PTMs) were also used and acquired from Uniprot data-base (The UniProt Consortium, 2015). The estimated enrichment scores were statistically assessed by performing 1,000 random permutations of the signatures and p-values were then adjusted using false-discovery rate (FDR).

Proteins were classified according to their copy-number attenuation effect using a gaussian mixture model with 2 mixture components. Proteins in the group with larger mean attenuation were considered highly attenuated. More stringent classification of the attenuation effect was performed by only considering proteins with an attenuation score higher than 0.3.

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

For samples the attenuation potential was estimated similarly as for proteins but instead correlations were calculated across the proteins measured in the sample. Samples were then classified as before with a gaussian mixture model with 2 mixture components. Enrichment analysis of amplifications in protein complexes in tumour samples with high protein attenuation potential was performed using SLAPenrich (Iorio et al., 2016a).

A gene expression signature of the sample attenuation potential was calculated by systematically correlating the samples attenuation potential with every gene-expression profile in the transcriptomics data-set.

Pairwise correlation analysis

Correlations between protein pairs, or genes, across samples were calculated using pearson correlation coefficient. Only proteins that were also measured at the transcript level were considered, i.e. 6,434. The systematic analysis of all unique pairwise correlations generated a total of 41,389,922 correlation coefficients both at the protein and gene level.

Protein sets of known protein complexes were acquired from the CORUM data-base (Ruepp et al., 2010, 2008). A protein-protein interaction list of the complexes was assembled by considering that two proteins interact if they are present within the same complex at least once, this generated a total of 67,927 interactions. More indirect but functional associations were also considered by using the STRING data-base (Franceschini et al., 2013). For STRING only interactions with the highest confidence score (900) were used performing a total of 214,815 interactions. 9,273 protein interactions within signalling pathways were assembled from

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

kinase/phosphatase-substrate interactions reported in SIGNOR data-base (Perfetto et al., 2016). Metabolic enzyme interactions associated with metabolic pathways were extracted from KEGG pathways (Kanehisa et al., 2016) reported in MSigDB (Subramanian et al., 2005). Two enzymes were considered to be interacting if they were present in the same metabolic pathway, making a total of 121,134 interactions. Enrichment of the different types of protein-protein interactions, i.e. complexes, functional, signalling and metabolic, were estimated using receiving operating characteristic (ROC) curves and by calculating the area under the ROC curve (AROC). True-positive sets of protein interactions were defined as the ones reported in the different resources used. Due to the strong unbalance between the number of true positives and false positives the ROC curves were calculated using 5 different and randomised sets within the false positive group. The variability of the AROC score is represented by error bars.

Proteogenomics analysis to identify protein complex regulators

The identification of protein complex regulators only focused on protein-protein interactions reported in the CORUM data-base (Ruepp et al., 2010, 2008) with a protein-protein interaction list assembled as described before.

For each protein-protein interaction reported within a complex, its association was tested using two linear regression models. Given a pair Protein Y ~ Protein X (Py ~ Px), a first linear model is used to regress-out the transcript variability from the protein measurement of Py. The dependent variable of the model is the proteomics measurements of Py and the independent variable is the transcriptomics measurements (Ty) (Eq. 1):

Eq. 1: $Py = \beta \, . \, Ty + \psi$

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The model is fit with an intercept (for simplicity omitted from Eq.1) and noise term, ψ. After fitting the estimated weight (β), the residuals of Py, (Py') are calculated as (Eq. 2):

$$\text{Eq. 2: } Py' = Py - \beta \, . \, Ty$$

Py' represents the variability measured due to post-transcriptional and post-translational regulation. Then a second linear model is performed to calculate the association between Py' and the CNV of Px, (Px) (Eq: 3):

$$\text{Eq. 3: } Py' = \beta \, . \, Px + \psi$$

Statistical significance is estimated by calculating an F statistic over an F-distribution, p-values are then adjusted using FDR correction. A total of 58,627 tests are performed. The same analysis is performed using transcriptomics measurements instead of the copy-number of Px, generating a total of 57,462 tests. Associations estimated with the copy-number variation that are significant with the transcriptomics are highlighted with a red border (Figure 3A).

Cell lines drug response analysis

Gene expression measurements (E-MTAB-3610) acquired with Affymetrix Human Genome U219 array for approximately 1,000 cell lines was used in this analysis (Iorio et al., 2016b). Drug response measurements were obtained as the area under the curve (AUC) for 265 compounds (Iorio et al., 2016b). Cell lines proteome attenuation potential was calculated by performing pearson correlation between their transcriptomics profile and the proteome attenuation potential signature derived from tumours. Cell line correlations with the signature were then used as a feature in single linear regression models to systematically predict the response of each compound in the screen. Statistical assessment of the association was calculated using a Wald test and multiple hypothesis correction with FDR.

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Code availability

All the computational analysis were performed in Python version 2.7.10, apart from the transcriptomics RNA-seq processing which as done in R version 3.3.1 with Limma package version 3.28.21 and edgeR 3.14.0, and are available under GNU General Public License V3 as a GitHub project in the following url https://github.com/saezlab/protein_attenuation. Plotting was done using Python modules Matplotlib version 1.4.3 (Hunter, 2007) and Seaborn version 0.7.0 (Waskom et al., 2014). Generalised linear models were built using Python module Sklearn version 0.17.1 (Pedregosa et al., 2011). Data analysis and structuring was carried out using Python module Pandas version 0.18.1 (McKinney and Others, 2010).

Acknowledgments

We thank M. Schubert for help integrating the copy-number variation data.

References

Abovich, N., Gritz, L., Tung, L., Rosbash, M., 1985. Effect of RP51 gene dosage alterations on ribosome synthesis in Saccharomyces cerevisiae. Mol. Cell. Biol. 5, 3429–3435.

Anastas, J.N., Biechele, T.L., Robitaille, M., Muster, J., Allison, K.H., Angers, S., Moon, R.T., 2012. A protein complex of SCRIB, NOS1AP and VANGL1 regulates cell polarity and migration, and is associated with breast cancer progression. Oncogene 31, 3696–3708.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Michael Cherry, J., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29.

Avner, P., Heard, E., 2001. X-chromosome inactivation: counting, choice and initiation. Nat. Rev. Genet. 2, 59–67.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., Mc Henry, K.T., Pinchback, R.M., Ligon, A.H., Cho, Y.-J., Haery, L., Greulich, H., Reich, M., Winckler, W., Lawrence, M.S., Weir, B.A., Tanaka, K.E., Chiang, D.Y., Bass, A.J., Loo, A., Hoffman, C., Prensner, J., Liefeld, T., Gao, Q., Yecies, D., Signoretti, S., Maher, E., Kaye, F.J., Sasaki, H., Tepper, J.E., Fletcher, J.A., Tabernero, J., Baselga, J., Tsao, M.-S., Demichelis, F., Rubin, M.A., Janne, P.A., Daly,

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

M.J., Nucera, C., Levine, R.L., Ebert, B.L., Gabriel, S., Rustgi, A.K., Antonescu, C.R., Ladanyi, M., Letai, A., Garraway, L.A., Loda, M., Beer, D.G., True, L.D., Okamoto, A., Pomeroy, S.L., Singer, S., Golub, T.R., Lander, E.S., Getz, G., Sellers, W.R., Meyerson, M., 2010. The landscape of somatic copy-number alteration across human cancers. Nature 463, 899–905.

Cancer Genome Atlas Network, 2012a. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330–337.

Cancer Genome Atlas Network, 2012b. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70.

Cancer Genome Atlas Research Network, 2011. Integrated genomic analyses of ovarian carcinoma. Nature 474, 609–615.

Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., 2013. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 45, 1113–1120.

Chick, J.M., Munger, S.C., Simecek, P., Huttlin, E.L., Choi, K., Gatti, D.M., Raghupathy, N., Svenson, K.L., Churchill, G.A., Gygi, S.P., 2016. Defining the consequences of genetic variation on a proteome-wide scale. Nature 534, 500–505.

Coscia, F., Watters, K.M., Curtis, M., Eckert, M.A., Chiang, C.Y., Tyanova, S., Montag, A., Lastra, R.R., Lengyel, E., Mann, M., 2016. Integrative proteomic profiling of ovarian cancer cell lines reveals precursor cell associated proteins and functional status. Nat. Commun. 7, 12645.

Dephoure, N., Hwang, S., O'Sullivan, C., Dodgson, S.E., Gygi, S.P., Amon, A., Torres, E.M., 2014. Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. Elife 3, e03023.

des Georges, A., Dhote, V., Kuhn, L., Hellen, C.U.T., Pestova, T.V., Frank, J., Hashem, Y., 2015. Structure of mammalian eIF3 in the context of the 43S preinitiation complex. Nature 525, 491–495.

Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan, S., Ketchum, K.A., 2015. The CPTAC Data Portal: A Resource for Cancer Proteomics Research. J. Proteome Res. 14, 2707–2713.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., Jensen, L.J., 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 41, D808–15.

Hunter, J.D., 2007. Matplotlib: A 2D Graphics Environment. Computing in Science Engineering 9, 90–95.

Iorio, F., Garcia-Alonso, L., Brammeld, J., Martincorena, I., Wille, D.R., McDermott, U., Saez-Rodriguez, J., 2016a. Pathway-based dissection of the genomic heterogeneity of cancer hallmarks with SLAPenrich. bioRxiv. doi:10.1101/077701

Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., Cokelaer, T., Greninger, P., van Dyk, E., Chang, H., de Silva, H., Heyn, H., Deng, X., Egan, R.K., Liu, Q., Mironenko, T., Mitropoulos, X., Richardson, L., Wang, J., Zhang, T., Moran, S., Sayols, S., Soleimani, M., Tamborero, D., Lopez-Bigas, N., Ross-Macdonald, P., Esteller, M., Gray, N.S., Haber, D.A., Stratton, M.R., Benes, C.H., Wessels, L.F.A., Saez-Rodriguez, J., McDermott, U., Garnett, M.J., 2016b. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 166, 740–754.

Itsara, A., Cooper, G.M., Baker, C., Girirajan, S., Li, J., Absher, D., Krauss, R.M., Myers, R.M., Ridker, P.M., Chasman, D.I., Mefford, H., Ying, P., Nickerson, D.A., Eichler, E.E., 2009.

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Population analysis of large copy number variants and hotspots of human genetic disease. Am. J. Hum. Genet. 84, 148–161.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457–62.

Karnezis, A.N., Wang, Y., Ramos, P., Hendricks, W.P., Oliva, E., D'Angelo, E., Prat, J., Nucci, M.R., Nielsen, T.O., Chow, C., Leung, S., Kommoss, F., Kommoss, S., Silva, A., Ronnett, B.M., Rabban, J.T., Bowtell, D.D., Weissman, B.E., Trent, J.M., Gilks, C.B., Huntsman, D.G., 2016. Dual loss of the SWI/SNF complex ATPases SMARCA4/BRG1 and SMARCA2/BRM is highly sensitive and specific for small cell carcinoma of the ovary, hypercalcaemic type. J. Pathol. 238, 389–400.

Kim, W., Bennett, E.J., Huttlin, E.L., Guo, A., Li, J., Possemato, A., Sowa, M.E., Rad, R., Rush, J., Comb, M.J., Harper, J.W., Gygi, S.P., 2011. Systematic and quantitative assessment of the ubiquitin-modified proteome. Mol. Cell 44, 325–340.

Kiriakidou, M., Tan, G.S., Lamprinaki, S., De Planell-Saguer, M., Nelson, P.T., Mourelatos, Z., 2007. An mRNA m7G cap binding-like motif within human Ago2 represses translation. Cell 129, 1141–1151.

Law, C.W., Chen, Y., Shi, W., Smyth, G.K., 2014. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 15, R29.

Lawrence, R.T., Perez, E.M., Hernández, D., Miller, C.P., Haas, K.M., Irie, H.Y., Lee, S.-I., Blau, C.A., Villén, J., 2015. The proteomic landscape of triple-negative breast cancer. Cell Rep. 11, 630–644.

Lilley, B.N., Ploegh, H.L., 2004. A membrane protein required for dislocation of misfolded proteins from the ER. Nature 429, 834–840.

Lyon, M.F., 1961. Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature 190, 372–373.

McKinney, W., Others, 2010. Data structures for statistical computing in python, in: Proceedings of the 9th Python in Science Conference. pp. 51–56.

McShane, E., Sin, C., Zauber, H., Wells, J.N., Donnelly, N., Wang, X., Hou, J., Chen, W., Storchova, Z., Marsh, J.A., Valleriani, A., Selbach, M., 2016. Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. Cell. doi:10.1016/j.cell.2016.09.015

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., Getz, G., 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 12, R41.

Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., Kawaler, E., Mundt, F., Krug, K., Tu, Z., Lei, J.T., Gatza, M.L., Wilkerson, M., Perou, C.M., Yellapantula, V., Huang, K.-L., Lin, C., McLellan, M.D., Yan, P., Davies, S.R., Townsend, R.R., Skates, S.J., Wang, J., Zhang, B., Kinsinger, C.R., Mesri, M., Rodriguez, H., Ding, L., Paulovich, A.G., Fenyö, D., Ellis, M.J., Carr, S.A., NCI CPTAC, 2016. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 534, 55–62.

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., Mann, M., 2011. Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Syst. Biol. 7, 548.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Perfetto, L., Briganti, L., Calderone, A., Perpetuini, A.C., Iannuccelli, M., Langone, F., Licata, L.,

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Marinkovic, M., Mattioni, A., Pavlidou, T., Peluso, D., Petrilli, L.L., Pirrò, S., Posca, D., Santonico, E., Silvestri, A., Spada, F., Castagnoli, L., Cesareni, G., 2016. SIGNOR: a database of causal relationships between biological entities. Nucleic Acids Res. 44, D548–54.

Rahman, M., Jackson, L.K., Johnson, W.E., Li, D.Y., Bild, A.H., Piccolo, S.R., 2015. Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. Bioinformatics 31, 3666–3672.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.

Robinson, M.D., Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11, R25.

Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O.N., Stümpflen, V., Mewes, H.W., 2008. CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res. 36, D646–50.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., Mewes, H.-W., 2010. CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res. 38, D497–501.

Santarius, T., Shipley, J., Brewer, D., Stratton, M.R., Cooper, C.S., 2010. A census of amplified and overexpressed human cancer genes. Nat. Rev. Cancer 10, 59–64.

Shangary, S., Wang, S., 2008. Targeting the MDM2-p53 interaction for cancer therapy. Clin. Cancer Res. 14, 5318–5324.

Stingele, S., Stoehr, G., Peplowska, K., Cox, J., Mann, M., Storchova, Z., 2012. Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. Mol. Syst. Biol. 8, 608.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550.

Tang, Y.-C., Amon, A., 2013. Gene copy-number alterations: a cost-benefit analysis. Cell 152, 394–405.

Teckchandani, A., Cooper, J.A., 2016. The ubiquitin-proteasome system regulates focal adhesions at the leading edge of migrating cells. Elife 5. doi:10.7554/eLife.17440

The Gene Ontology Consortium, 2015. Gene Ontology Consortium: going forward. Nucleic Acids Res. 43, D1049–D1056.

The UniProt Consortium, 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43, D204–D212.

Thompson, S.L., Compton, D.A., 2011. Chromosomes and cancer cells. Chromosome Res. 19, 433–444.

Wang, J., Ma, Z., Carr, S.A., Mertins, P., Zhang, H., Zhang, Z., Chan, D.W., Ellis, M.J.C., Townsend, R.R., Smith, R.D., McDermott, J.E., Chen, X., Paulovich, A.G., Boja, E.S., Mesri, M., Kinsinger, C.R., Rodriguez, H., Rodland, K.D., Liebler, D.C., Zhang, B., 2016. Proteome profiling outperforms transcriptome profiling for co-expression based gene function prediction. Mol. Cell. Proteomics. doi:10.1074/mcp.M116.060301

Waskom, M., Botvinnik, O., Hobson, P., Cole, J.B., Halchenko, Y., Hoyer, S., Miles, A.,

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Augspurger, T., Yarkoni, T., Megies, T., Coelho, L.P., Wehner, D., cynddl, Ziegler, E., diego, Zaytsev, Y.V., Hoppe, T., Seabold, S., Cloud, P., Koskinen, M., Meyer, K., Qalieh, A., Allan, D., 2014. seaborn: v0.5.0 (November 2014). ZENODO. doi:10.5281/zenodo.12710

Xie, Y., Wolff, D.W., Wei, T., Wang, B., Deng, C., Kirui, J.K., Jiang, H., Qin, J., Abel, P.W., Tu, Y., 2009. Breast cancer migration and invasion depend on proteasome degradation of regulator of G-protein signaling 4. Cancer Res. 69, 5743–5751.

Ye, Y., Shibata, Y., Yun, C., Ron, D., Rapoport, T.A., 2004. A membrane protein complex mediates retro-translocation from the ER lumen into the cytosol. Nature 429, 841–847.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., Davies, S.R., Wang, S., Wang, P., Kinsinger, C.R., Rivers, R.C., Rodriguez, H., Townsend, R.R., Ellis, M.J.C., Carr, S.A., Tabb, D.L., Coffey, R.J., Slebos, R.J.C., Liebler, D.C., NCI CPTAC, 2014. Proteogenomic characterization of human colon and rectal cancer. Nature 513, 382–387.

Zhang, H., Liu, T., Zhang, Z., Payne, S.H., Zhang, B., McDermott, J.E., Zhou, J.-Y., Petyuk, V.A., Chen, L., Ray, D., Sun, S., Yang, F., Chen, L., Wang, J., Shah, P., Cha, S.W., Aiyetan, P., Woo, S., Tian, Y., Gritsenko, M.A., Clauss, T.R., Choi, C., Monroe, M.E., Thomas, S., Nie, S., Wu, C., Moore, R.J., Yu, K.-H., Tabb, D.L., Fenyö, D., Bafna, V., Wang, Y., Rodriguez, H., Boja, E.S., Hiltke, T., Rivers, R.C., Sokoll, L., Zhu, H., Shih, I.-M., Cope, L., Pandey, A., Zhang, B., Snyder, M.P., Levine, D.A., Smith, R.D., Chan, D.W., Rodland, K.D., CPTAC Investigators, 2016. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. Cell 166, 755–765.

Zhang, L., Pan, X., Hershey, J.W.B., 2007. Individual overexpression of five subunits of human translation initiation factor eIF3 promotes malignant transformation of immortal fibroblast cells. J. Biol. Chem. 282, 5790–5800.

Zolov, S.N., Lupashin, V.V., 2005. Cog3p depletion blocks vesicle-mediated Golgi retrograde trafficking in HeLa cells. J. Cell Biol. 168, 747–759.

Zou, Y., Liu, Y., Wu, X., Shell, S.M., 2006. Functions of human replication protein A (RPA): from DNA replication to DNA damage and stress responses. J. Cell. Physiol. 208, 267–273.

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure legends

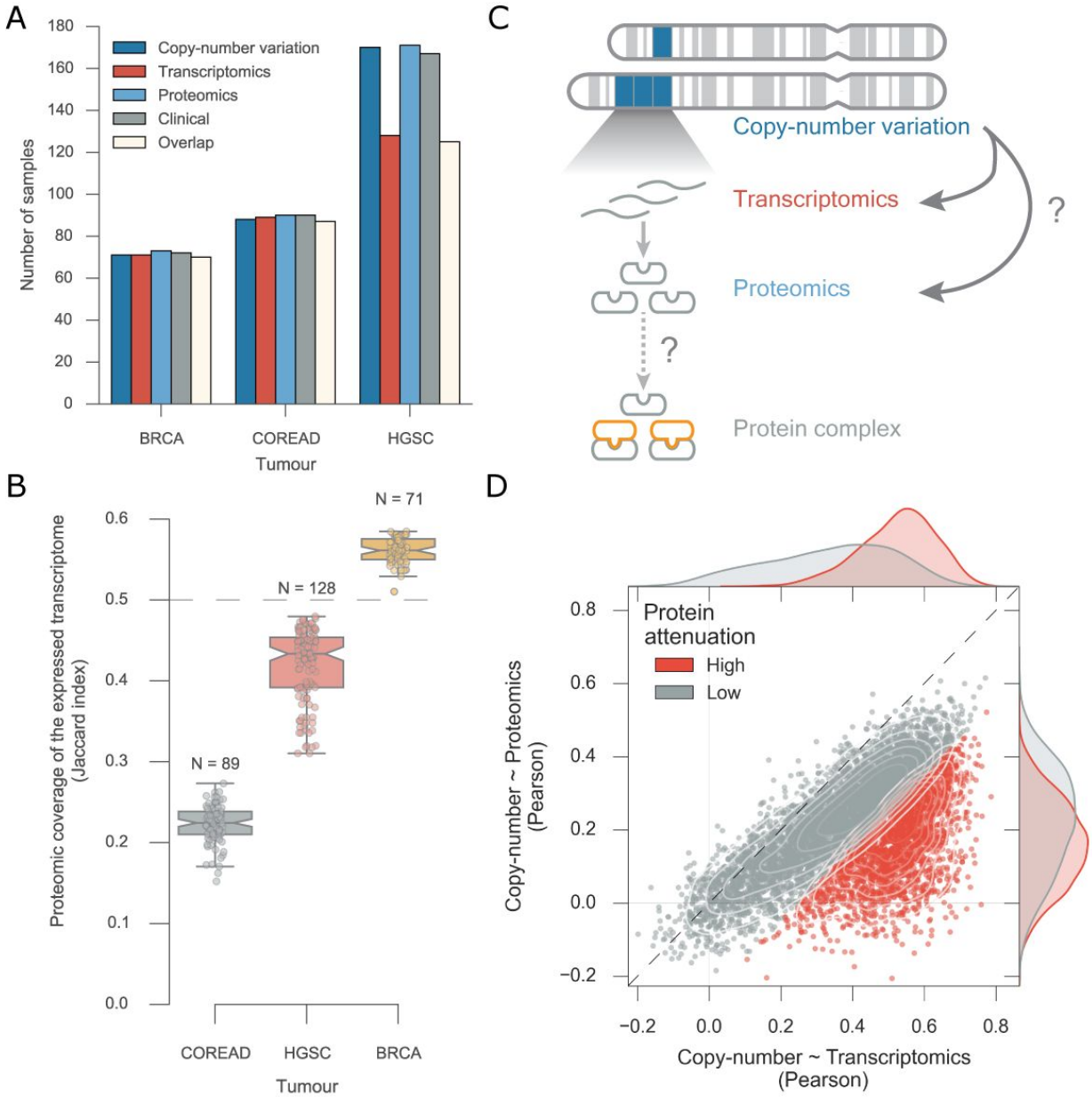


Figure 1. Pan-cancer effects of copy-number variation implications on transcript and protein abundances. A) Overview of the number of samples used in this study overlapping with the proteomics measurements for each tumour type. B) Coverage of the proteomics measurements

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of the expressed transcripts in each sample, for each tumour type. C) Diagram depicting the implication of copy-number alterations along the central dogma of biology. D) Scatter plot depicting the pearson correlation between copy-number variation and transcriptomics in the x axis and copy-number variation and proteomics in the y axis. A gaussian mixture model was used to identify proteins with high attenuation levels and these are colored in red.

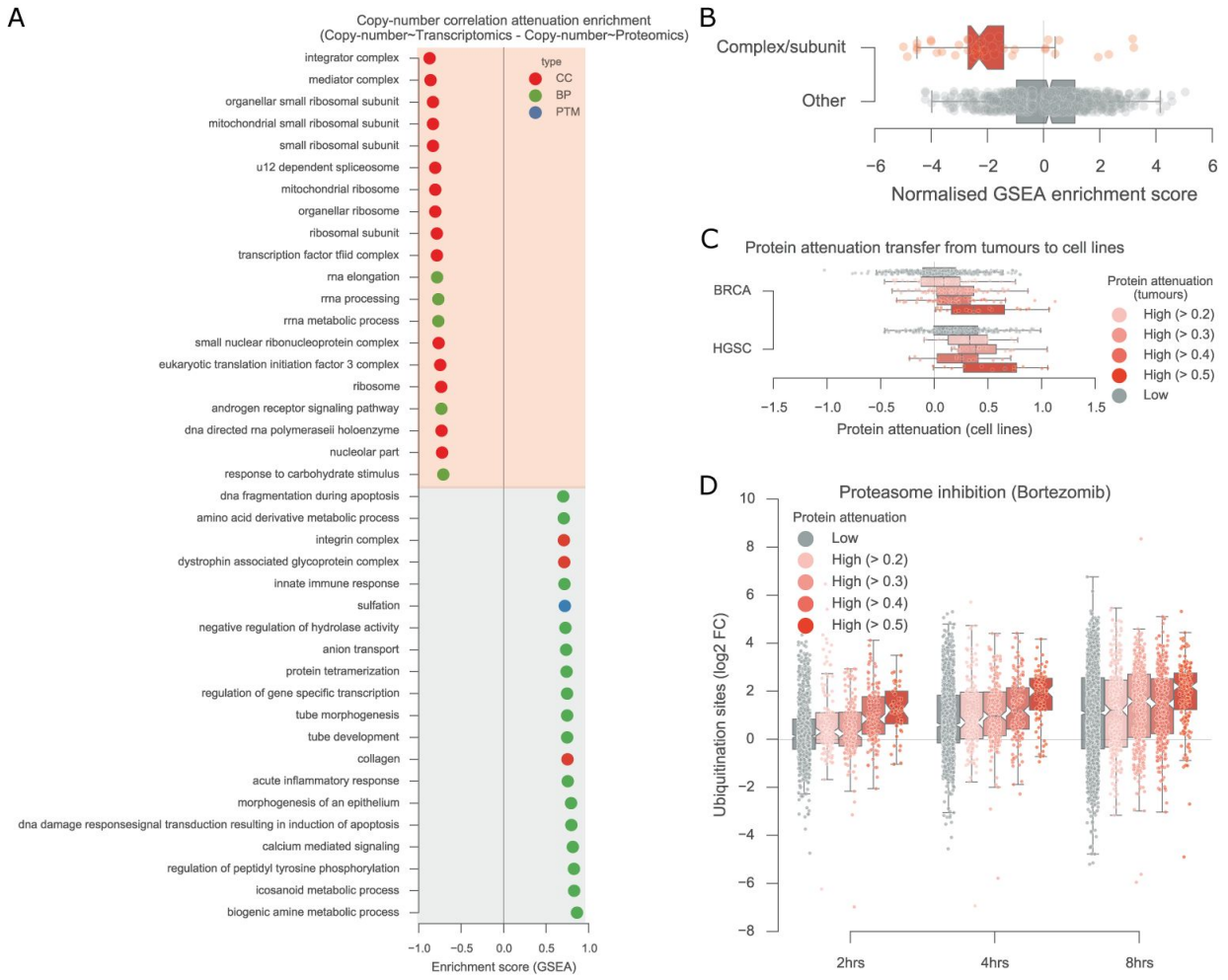


Figure 2. Enrichment analysis of the proteins undergoing copy-number attenuation. A) Enrichment analysis of the correlation differences between copy-number variation and transcriptomics and copy-number variation and proteomics. Protein subsets used represent

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

biological processes (BP; green), cellular components (CC; red) and post-translational modifications (PTM; blue). B) The distribution of the enrichment scores for terms referring to protein complexes or subunits are represented in red and all the rest in gray. C) Proteins classified according to their attenuation profile in tumours are mapped against their attenuation in breast and ovarian cancer cell lines. D) Ubiquitination sites fold-changes (log2) over time after proteasome inhibition with bortezomib discretized according the protein attenuation level in tumours.

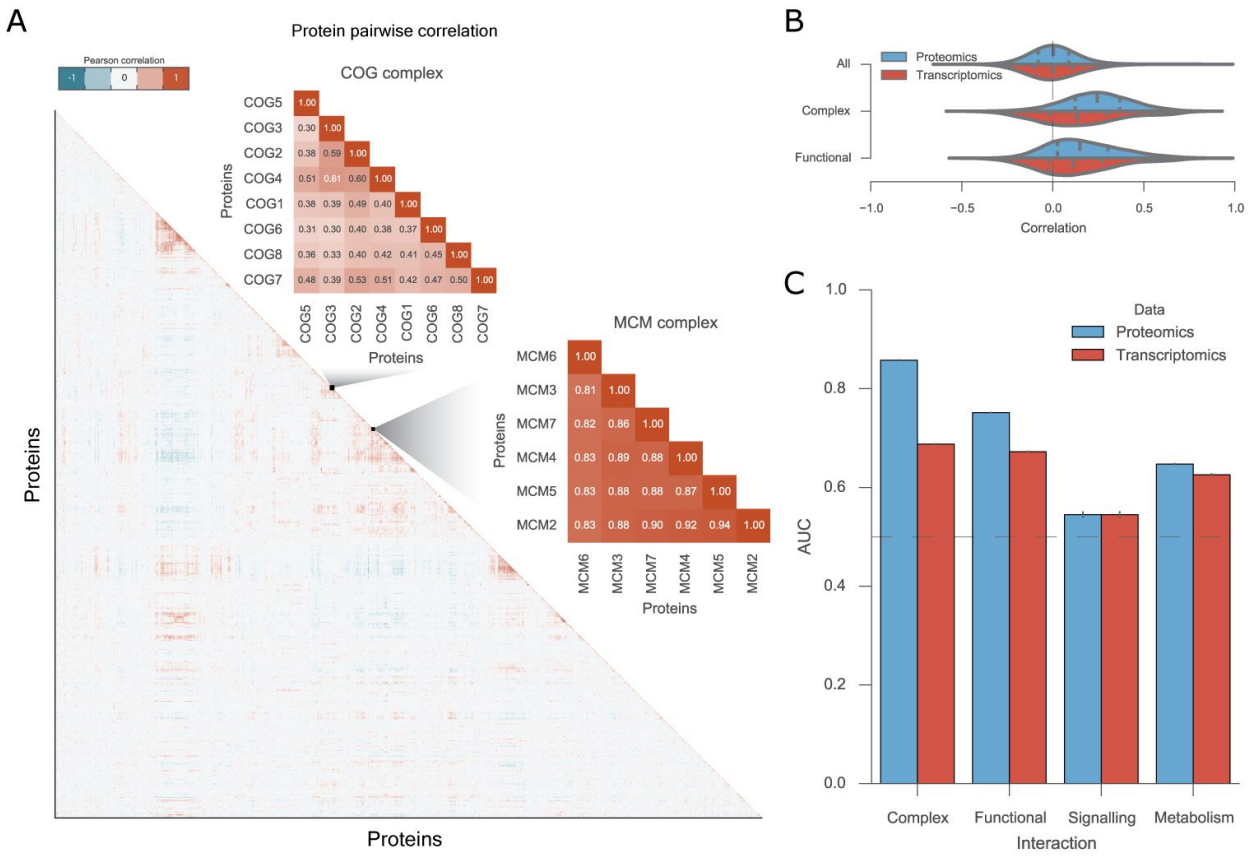


Figure 3. Copy-number variation attenuation for protein complex subunits results in strong co-regulation of abundances across samples. A) Protein-protein correlation matrix using pearson correlation coefficient and representative cases of top correlated members of protein

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

complexes and subunits. B) Distribution of all protein-protein correlation at the protein level (Proteomics), and transcript level (Transcriptomics). Protein interactions within complexes are represented by the Complex label and protein functional, but not necessarily direct, interactions are represented by the Functional label. C) Enrichment analysis by the means of the area under the receiving operating characteristic curves (AROC) using pairwise correlation coefficients, for both proteomics and transcriptomics measurements.

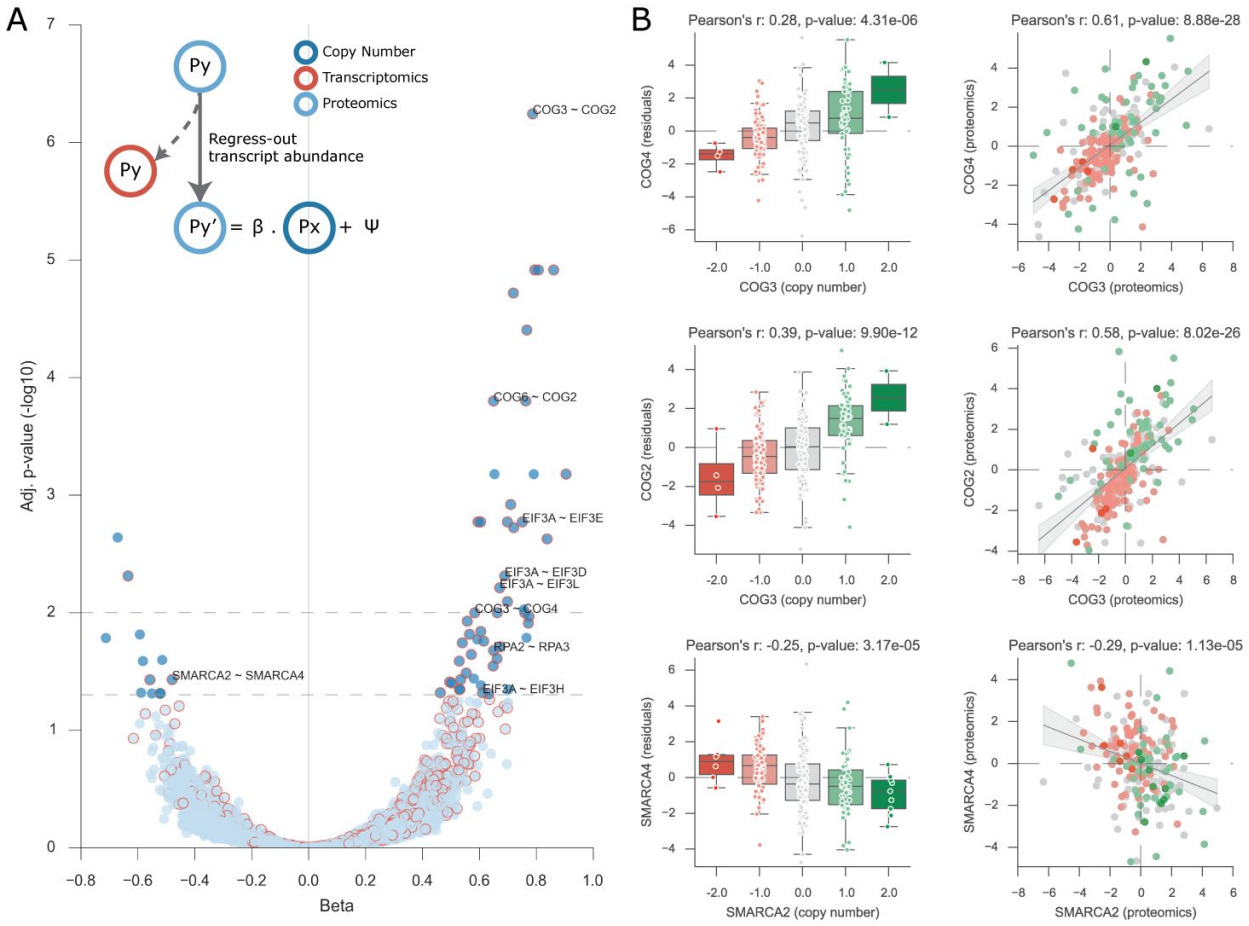


Figure 4. Protein complex regulators. A) Volcano plot displaying the effect size and adjusted p-value of all the tested regulatory interactions. Associations were performed using the

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

copy-number variation of the putative regulatory protein, Px, and the protein residuals of the regulated protein, Py. Significant associations found with the transcript measurements of Px are denoted with a red border. B) Representative significant associations. Boxplots show the agreement between the copy-number variation of Px and the residuals of the regulated Py. Scatter plot show the agreement between the protein pairs at the original proteomics measurements.

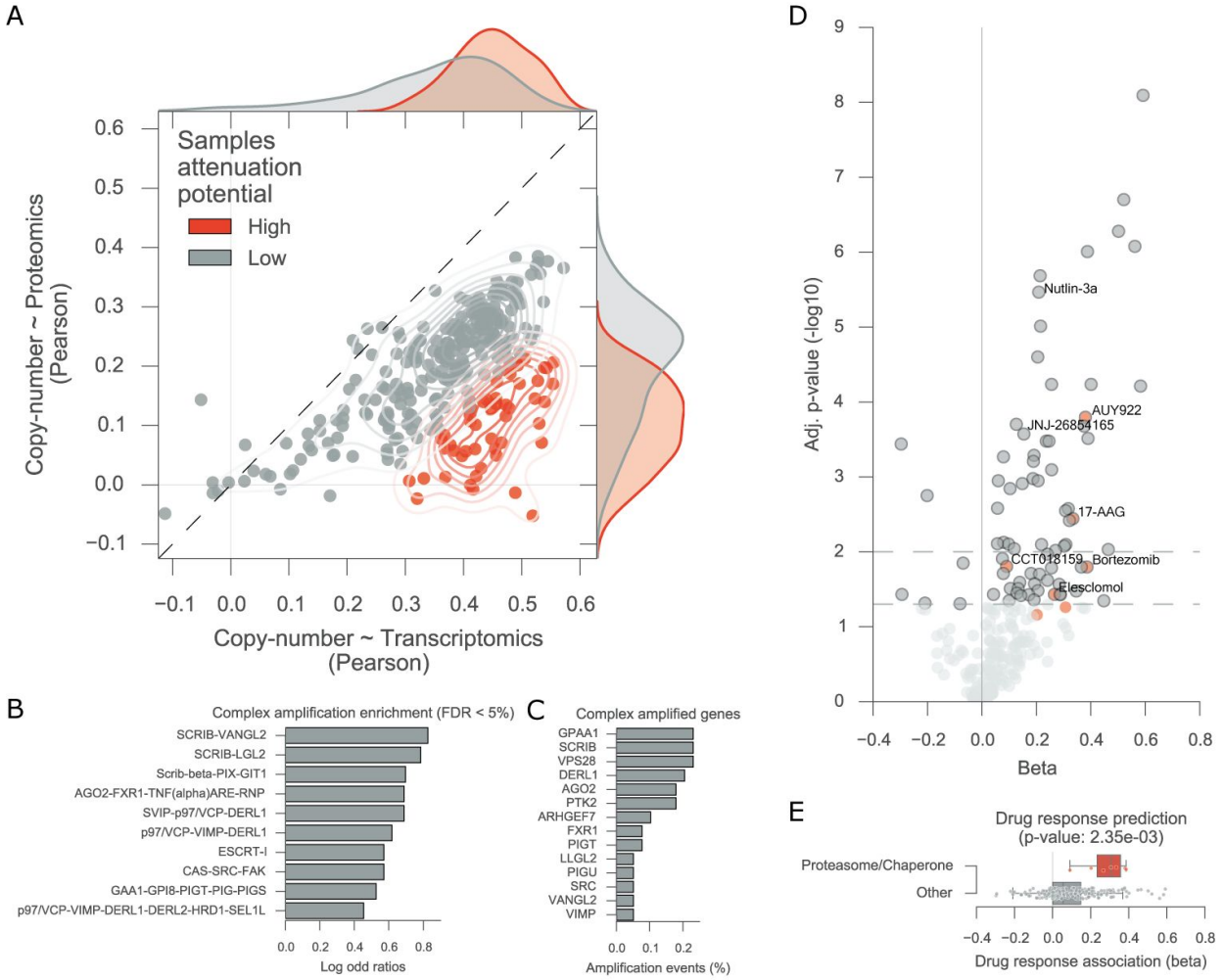


Figure 5. Tumour attenuation potential predicts attenuation mechanisms and is associated with chaperone/proteasome drug resistance. A) Tumour samples correlation of the copy-number

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

changes and the transcript (x axis) and protein (y axis) measurements. Samples classified with high attenuation potential, in red, display the stronger attenuation of the copy-number variation on the proteins. B) Protein complexes significantly enriched (FDR < 5%) on the samples with high protein attenuation. C) Top strongly amplified genes within the significantly enriched complexes. D) Drug response associations performed in a large cell line panel using the cell lines putative attenuation potential as the predictive feature. Significant associations (FDR < 5%) of chaperone inhibitors are labelled and marked in red. Ubiquitin-protein ligase MDM2 inhibitors are labelled. E) Boxplots representing the distributions of the drug associations effect sizes of all the proteasome and chaperones inhibitors in the drug panel.

Tables and their legends

Supplementary Table 1. Overview of the biological measurements available for the TCGA and CPTAC samples.

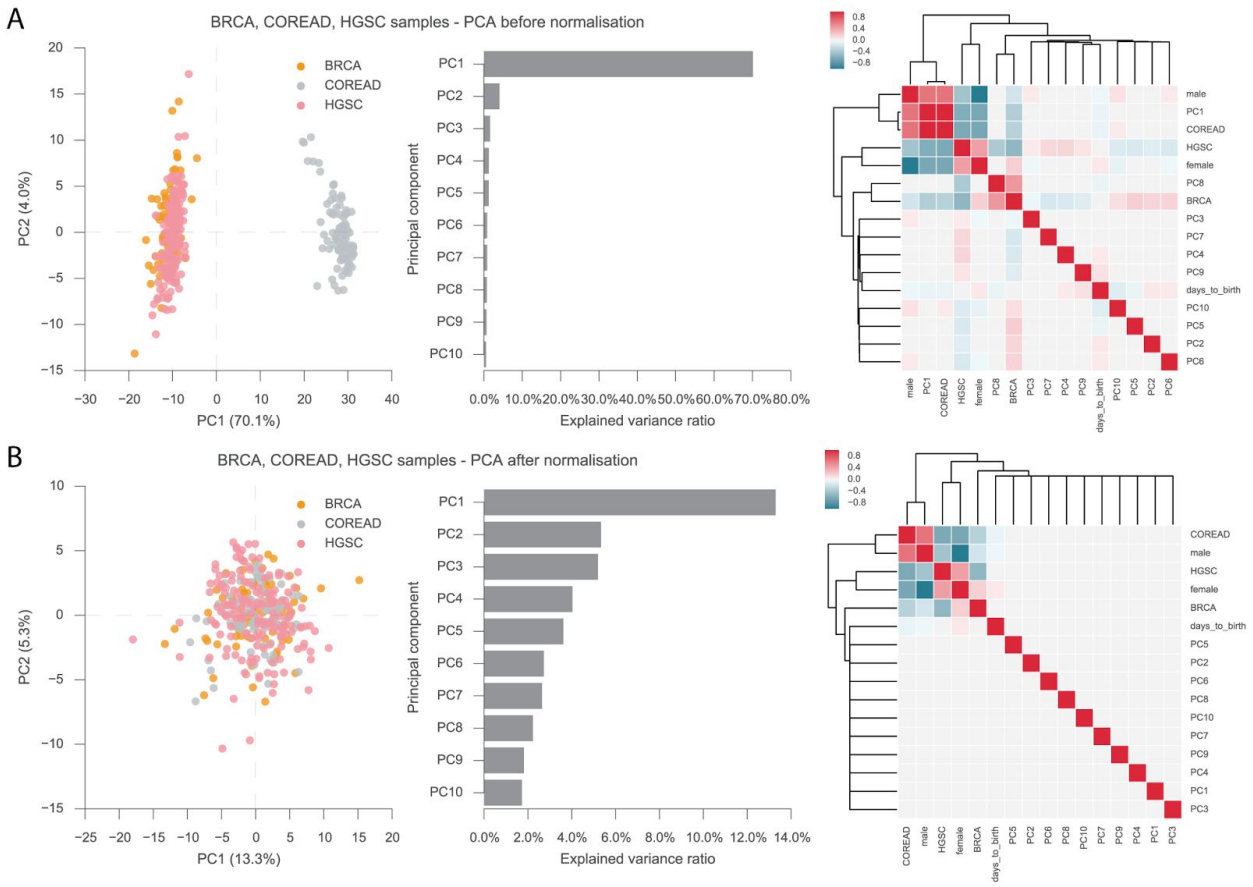
Supplementary Table 2. Pearson correlation coefficients for all genes overlapping between copy-number variation, transcriptomics and proteomics data-sets.

Supplementary Table 3. Protein complex regulatory associations estimated between the putative regulatory protein, Px, and the regulated protein, Py, using the copy-number variation and transcriptomics measurements.

Supplementary Table 4. Tumour protein attenuation potential gene-expression signature and drug response analysis results.

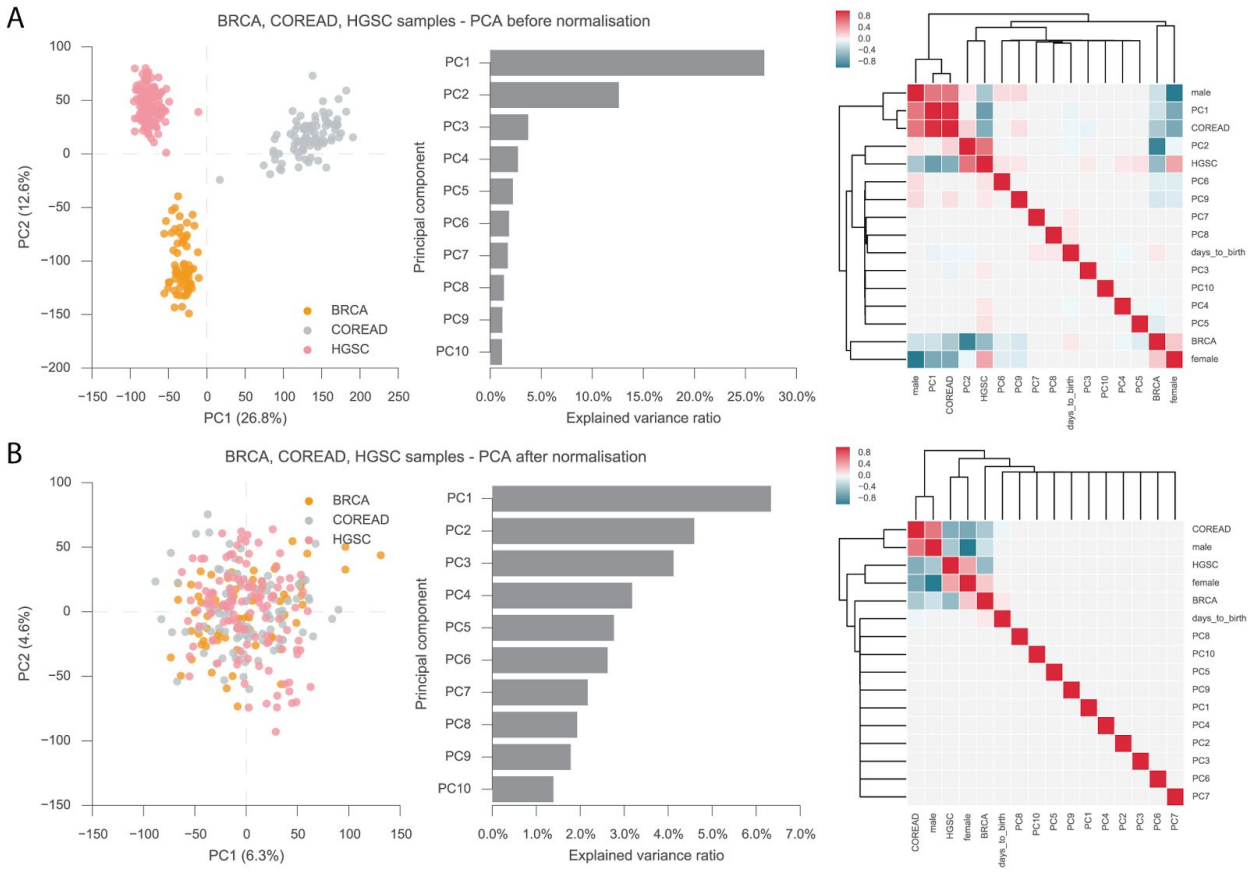
bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Expanded view figure legends



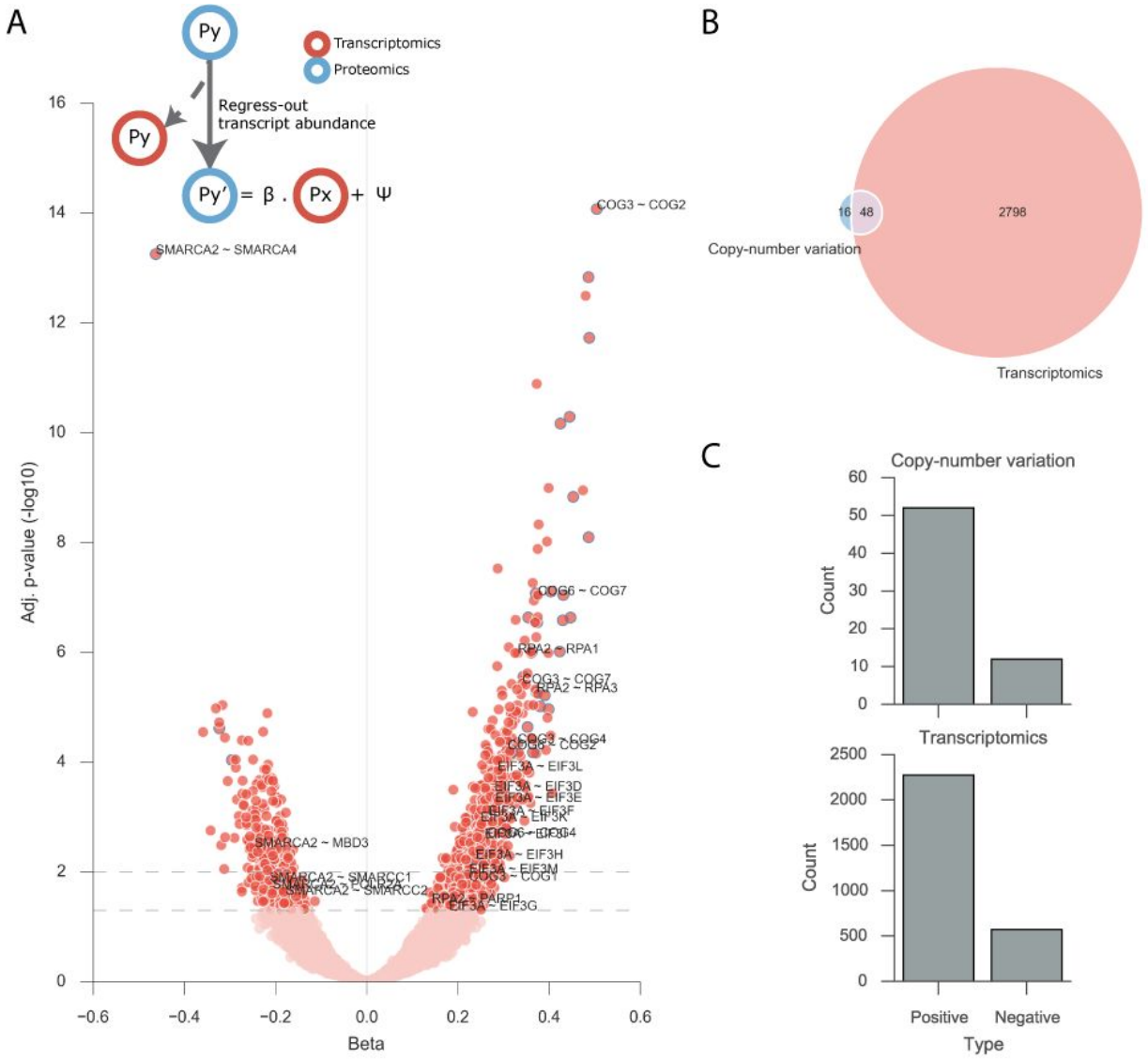
Supplementary Figure 1. Proteomics data-sets PCA analysis using proteins consistently measured across all the samples and pearson correlation coefficient between the first 10 principal components and the possible confounding factors, i.e. age, tumour type and gender. A) Analysis performed on the published proteomics data-sets. B) Analysis performed on the proteomics data-set where the confounding factors were regressed-out.

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure 2. Transcriptomics data-sets PCA analysis and pearson correlation coefficient between the first 10 principal components and the possible confounding factors, i.e. age, tumour type and gender. A) Analysis performed on the voom transformed transcriptomics measurements. B) Analysis performed on the transcriptomics data-set where the confounding factors were regressed-out.

bioRxiv preprint doi: https://doi.org/10.1101/093369; this version posted December 12, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure 3. Protein complexes regulatory interactions identified using transcriptomics of the putative regulatory protein, Px. A) Volcano plot representing the effect size on the x axis and FDR adjusted p-value on the y axis. Diagram representing the linear model used to perform the associations. B) Overlap between the significant regulatory associations found using the copy-number variation and transcriptomics of the Px proteins. C) Number of significant associations with a Positive or Negative effect size.