SUPPLEMENTAL METHODS

Data sources

The effect of aneuploidy on gene expression was analyzed using DNA, RNA and protein data generated through the Cancer Cell Line Encyclopedia and available in processed form from the DepMap project (Nusinow et al. 2020; Nusinow and Gygi 2020; Broad DepMap 2020, 2021; Barretina et al. 2012; Ghandi et al. 2019).

Protein expression data was retrieved from the DepMap "Proteomics" dataset (downloaded August 3rd, 2020). The protein data had been normalized, $\log_2$ transformed, and mean expression was set at 0 (Nusinow et al. 2020; Nusinow and Gygi 2020). Multi-sample mass spectrometry has been shown to cause systemic underestimation of protein fold changes, a phenomenon called ratio compression (Savitski et al. 2013). However, the CCLE proteomics dataset was generated using triple-stage mass spectrometry (MS3), which includes an additional round of peptide ion fragmentation and isolation. The MS3 technique has been demonstrated to almost completely eliminate ratio compression compared to standard tandem mass spectrometry (Ting et al. 2011).

RNA expression data was retrieved from the DepMap "Expression" (Public 20Q4) dataset (Broad DepMap 2020; Ghandi et al. 2019). The RNA data was $\log_2$ transformed with a pseudocount of 1.

"*Cell Line Sample Info"* was downloaded from DepMap (May 21st 2020) and used to link various data types (Broad DepMap 2020).

Relative gene copy number data was retrieved from DepMap "CCLE gene cn" (Public 21Q1) dataset on April 29, 2021 (Broad DepMap 2021). The data was $\log_2$ transformed with a pseudocount of 1.

Chromosome arm copy number data were retrieved from Cohen-Sharir, Y et al. 2021 (Cohen-Sharir et al. 2021).

Oncogene and tumor suppressor gene lists were acquired from Bailey et al. (Bailey et al. 2018) based on a comprehensive analysis of TCGA data.

Aggregated reproducibility rank gene information was acquired from Upadhya & Ryan (Upadhya and Ryan 2021).

Potential buffering factors explored in Figure 5 were retrieved from multiple sources.

Several potential buffering factors were retrieved from MobiBD (Version: 4.0 - Release: 2020_09(Piovesan et al. 2021)). MobiBD data includes: *Protein intrinsic disorder* (prediction-disorder- mobidb_lite), *low complexity score* (prediction-low_complexity-merge), *homology score* (homology-domain -merge), *loops in protein score* (prediction-lip -anchor), *protein polyampholyte score* (prediction-polyampholyte-mobidb_lite_sub) and *protein polarity* (prediction-polar- mobidb_lite_sub).

Translation rate, transcription rate, protein length, mRNA length, mRNA abundance and protein abundance data was retrieved from Hausser et al. 2019 (Hausser et al. 2019). *Protein half life* data was from Methieson et al. 2018 (Mathieson et al. 2018), the mean protein half life was used. *mRNA decay rates* were retrieved from Yang et al. 2003 (Yang et al. 2003).

UTR data (5' and 3') was retrieved from the UCSC human genome browser (hg38 GRCh38). The difference between transcription start sites and coding region start sites was found for 5' UTR, and the difference between coding sequence end and transcription sequence end was found for 3' UTR regions. Only manually curated ("NM") mRNA genes were analyzed.

Datasets for protein *acetylation, methylation, phosphorylation, ubiquitination, sumoylation and gene regulatory sites* were downloaded from phosphosite plus (Hornbeck et al. 2015) (last updated on April 19, 2021). The number of regulatory and/or modification sites were found per human gene. Genes not found in the dataset were set to zero modification sites.

The *neutral variance for RNA and protein expression* was calculated from DepMap protein and RNA expression data. Cell lines without a gain or loss (neutral ploidy) of the chromosome arm a gene was located on were used to calculate the neutral variance for that gene. The coefficient of variance was taken from gene expression levels and $\log_2$ transformed.

The *mutation counts (all)* was calculated per gene using the DepMap CCLE mutations public dataset (21Q2, Broad 2021(Broad DepMap 2021)). The mutation data was filtered for the cell lines in our expression difference dataset, and the number of all mutations was counted per gene. For the *mutation counts (nonsense and missense)*, as well as the *mutation counts (frame shifts)* only counted mutations of the indicated variants.

Non-exponential decay delta scores were taken from McShane et al. 2016 (McShane et al. 2016).

The *aggregation score* per protein was extracted from Ciryam et al. 2013 (Ciryam et al. 2013).

Protein complex (CORUM) score was extracted from the comprehensive resource of mammalian protein complexes (CORUM) “complete complex” dataset, version 3.0 (last updated 2018)(Giurgiu et al. 2019). All human protein complexes were extracted, and the number of times a protein appeared in the dataset was calculated. Proteins not found in protein complexes were given a score of zero.

The number of *protein-protein interactions* per protein was extracted from the Human Integrated Protein-Protein Interaction rEference (HIPPIE), version 2.2 (last updated February 2019)(Alanis-Lobato et al. 2017). Interactions with confidence values >0.6 were used, and the sum of interactions per protein was calculated. Proteins not found to interact with other proteins were given a score of zero.

Dependency scores were extracted from the DepMap Achilles gene dependency (21Q2(Broad DepMap 2021)) dataset. We took the mean gene dependency from across all reported cell lines.

Ovarian tumor proteomics data used in this publication were generated by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH)(Edwards et al. 2015). *Ovarian tumor transcriptome data* was acquired from Broad institute GDAC(Broad Institute TCGA Genome Data Analysis Center 2016). Additionally, *ovarian tumor arm call/aneuploidy data* was obtained from Taylor et al. (Taylor et al. 2018).

Stable aneuploidy cell line data were retrieved from Stingele et al. (Stingele et al. 2012). Protein and RNA difference upon aneuploidy was calculated by taking the $\log_2$ fold change between aneuploidy clones and their near-euploid cell line controls. For protein data, the mean difference was taken for all four cell lines with a gain for Chromosome 5: RPE-1 trisomy 12 and 5, HCT-116 tetraploidy 5, HCT-116 H2B-GFP tetraploidy 5 and HCT-116 H2B-GFP trisomy 5. For RNA data, mean RNA difference was found from RPE-1 trisomy 12 and 5 and HCT-116 tetraploidy 5, the only two Chromosome 5 gain transcriptomes available. Mean gene expression difference was plotted against DepMap protein difference data.

The *reverse phase protein array (RPPA) values and antibody references* were acquired from the DepMap RPPA datasets (Ghandi et al. 2019): “CCLE_RPPA_20181003” and “CCLE_RPPA_Ab_info_20181226”.

Down syndrome aneuploidy data was retrieved from Liu et al. (Liu et al. 2017) and Letourneau et al. (Letourneau et al. 2014). Protein expression difference was calculated

by taking the $\log_2$ fold change in quantile-normalized protein expression from 11 Down syndrome fibroblast lines and matched controls. The difference in RNA expression was reported as the $\log_2$ fold change in RNA expression between a pair of monozygotic twins discordant for Chromosome 21.

Yeast aneuploidy gene expression difference was retrieved from Dephoure et al. (Dephoure et al. 2014). Protein and RNA expression differences were reported as $\log_2$ ratios for genes located on duplicated chromosomes.

Pseudogene analysis was done on the main RNA expression dataset (see above). Coding genes and pseudogenes were identified using BioMart "gene type" classification (Smedley et al. 2015), with all pseudogene types merged. *Non-coding RNA expression data* was retrieved from the Expression Atlas E-MTAB-2770 (Barretina et al. 2012) as this dataset had more non-coding RNA expression; data was quantile normalized, $\log_2$ transformed, and the difference upon aneuploidy was found as above. The list of genes per non-coding RNA category was retrieved from the HGNC non-coding RNA gene group (Wright and Bruford 2011).

Dataset filtering (continued)

We generated two datasets measuring *gene expression differences upon chromosome gain and loss within cells with either low or high levels of cellular aneuploidy.* We initially included cell lines with proteomics data, RNA expression data, and chromosome arm copy number data. Next, we isolated the quartile of the cell lines with the lowest cellular aneuploidy score (from 0 to 1773), and the quartile of cell lines with the highest cellular aneuploidy score (from 3309 to 6985). We only included a gene in our difference analysis if we had RNA and protein expression data for that gene from at least 3 cell lines per chromosome arm category, as described in the main methods section (Supplemental Table S11).

We generated *datasets measuring gene expression differences in near-diploid and near-triploid cell lines separately*. Our dataset of human cancer cell lines consisted of 50% (185/367) near-diploid cell lines and 37% (137/367) near-triploid cell lines, with the remaining 13% spread across other ploidies. We therefore analyzed the near-diploid and near-triploid cell lines separately, and we excluded the other ploidies due to the low number of cell lines available.

We generated *datasets controlling for one of several factors,* these datasets measured gene expression difference upon chromosome gain or loss. The "No mutations" dataset removed mutated expression data for genes per cell line as listed in the DepMap CCLE mutations public dataset (21Q2, Broad 2021(Broad DepMap 2021)); all mutations were

removed. The “no flipped genes” dataset removed expression data for genes whose gene copy number (DepMap “CCLE gene cn” (Public 21Q1)(Broad DepMap 2021)) increased (relative copy number >1.1, 10.38% of genes) upon chromosome arm loss, or whose relative gene copy number decreased (relative copy number <0.9, 2.4% of genes) upon chromosome arm gain. The “High RNA only” dataset removed all genes whose mean RNA expression was in the bottom 20% of the RNA expression dataset (between 0.00 and 0.19). The “high reproducibility genes” dataset took all the genes in the previously published “aggregated reproducibility rank” from Upadhya & Ryan(Upadhya and Ryan 2021) and removed the 20% least reproducible genes (aggregated reproducibility rank below 0.346). Genes without a reproducibility rank were excluded. Finally, the NSCLC dataset took data from only non-small cell lung cancer cell lines (64 cell lines). The NSCLC dataset was not combined with the merged-control dataset as it reduced the number of genes available for analysis to below one thousand. All datasets only contained data from cells with RNA, protein, and gene copy number data available, and only contained genes with 10 or more data points per chromosome arm category. Filtering out genes with less than 10 data points per category was done after removing undesired data points.

Assessing the conservation of aneuploidy-associated dosage compensation

Gene expression datasets from stable human aneuploidy cell lines, Down syndrome fibroblast lines, and aneuploid yeast strains were acquired and processed as described above. The yeast orthologs of human genes were identified using g:Profiler(Raudvere et al. 2019). For this analysis, only one-to-one orthologs were considered. That is, we only included a yeast gene if it had a single human ortholog and if that human gene only had a single yeast ortholog. Dosage compensation across these various aneuploidy conditions was compared using Pearson correlation coefficients.

Identifying genomic features that correlate with protein buffering

ROC plots and area under the curve analysis were performed using R package “pROC” version 1.17.0.1 (Robin et al. 2011) and the various genetic, biochemical, and biophysical datasets listed above. A gene’s classification as “buffering” was used as the true positive “sensitivity” fraction. Significance was calculated by performing 10,000 random permutations and bootstrapping significance for each factor.

The potential buffering factors we analyzed are not all independent of one another; conversely, several factors are closely linked. For example, the number of protein-protein interactions and number of protein complexes a protein is integrated in are closely linked. The 30 genetic, biochemical and biophysical datasets we selected for our analysis were chosen based on availability and relevance. Our factor analysis was limited by the number of robust (>1500 data points) datasets of human data publicly available for scientific use. We sought out factors we knew to be involved with

transcription, translation, post-translational modifications, protein structure, protein stability, and protein half life (Buccitelli and Selbach 2020). We were not able to identify a dataset specifying all genes affected by various forms of transcriptional regulation.

REFERENCES

Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. 2017. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res* **45**: D408–D414.

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, et al. 2018. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**: 371-385.e18.

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, et al. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**: 603–607.

Broad DepMap. 2020. DepMap 20Q4 Public. https://figshare.com/articles/dataset/DepMap_20Q4_Public/13237076/2 (Accessed June 3, 2021).

Broad DepMap. 2021. DepMap 21Q2 Public. 14112785609 Bytes. https://figshare.com/articles/dataset/DepMap_21Q2_Public/14541774/2 (Accessed June 6, 2021).

Broad Institute TCGA Genome Data Analysis Center. 2016. Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run. http://gdac.broadinstitute.org/runs/stddata__2016_01_28 (Accessed November 30, 2021).

Buccitelli C, Selbach M. 2020. mRNAs, proteins and the emerging principles of gene expression control. *Nat Rev Genet* **21**: 630–644.

Ciryam P, Tartaglia GG, Morimoto RI, Dobson CM, Vendruscolo M. 2013. Widespread aggregation and neurodegenerative diseases are associated with supersaturated proteins. *Cell Rep* **5**: 781–790.

Cohen-Sharir Y, McFarland JM, Abdusamad M, Marquis C, Bernhard SV, Kazachkova M, Tang H, Ippolito MR, Laue K, Zerbib J, et al. 2021. Aneuploidy renders cancer cells vulnerable to mitotic checkpoint inhibition. *Nature* **590**: 486–491.

Dephoure N, Hwang S, O'Sullivan C, Dodgson SE, Gygi SP, Amon A, Torres EM. 2014. Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. *eLife* **3**: e03023.

Edwards NJ, Oberti M, Thangudu RR, Cai S, McGarvey PB, Jacob S, Madhavan S, Ketchum KA. 2015. The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J Proteome Res* **14**: 2707–2713.

Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER, Barretina J, Gelfand ET, Bielski CM, Li H, et al. 2019. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**: 503–508.

Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Ruepp A. 2019. CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res* **47**: D559–D563.

Hausser J, Mayo A, Keren L, Alon U. 2019. Central dogma rates and the trade-off between precision and economy in gene expression. *Nat Commun* **10**: 68.

Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. 2015. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* **43**: D512–D520.

Letourneau A, Santoni FA, Bonilla X, Sailani MR, Gonzalez D, Kind J, Chevalier C, Thurman R, Sandstrom RS, Hibaoui Y, et al. 2014. Domains of genome-wide gene expression dysregulation in Down's syndrome. *Nature* **508**: 345–350.

Liu Y, Borel C, Li L, Müller T, Williams EG, Germain P-L, Buljan M, Sajic T, Boersema PJ, Shao W, et al. 2017. Systematic proteome and proteostasis profiling in human Trisomy 21

fibroblast cells. *Nat Commun* **8**: 1212.
Mathieson T, Franken H, Kosinski J, Kurzawa N, Zinn N, Sweetman G, Poeckel D, Ratnu VS, Schramm M, Becher I, et al. 2018. Systematic analysis of protein turnover in primary cells. *Nat Commun* **9**: 689.
McShane E, Sin C, Zauber H, Wells JN, Donnelly N, Wang X, Hou J, Chen W, Storchova Z, Marsh JA, et al. 2016. Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. *Cell* **167**: 803-815.e21.
Nusinow DP, Gygi SP. 2020. *A Guide to the Quantitative Proteomic Profiles of the Cancer Cell Line Encyclopedia*. Systems Biology http://biorxiv.org/lookup/doi/10.1101/2020.02.03.932384 (Accessed June 3, 2021).
Nusinow DP, Szpyt J, Ghandi M, Rose CM, McDonald ER, Kalocsay M, Jané-Valbuena J, Gelfand E, Schweppe DK, Jedrychowski M, et al. 2020. Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell* **180**: 387-402.e16.
Piovesan D, Necci M, Escobedo N, Monzon AM, Hatos A, Mičetić I, Quaglia F, Paladin L, Ramasamy P, Dosztányi Z, et al. 2021. MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res* **49**: D361–D367.
Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. 2019. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* **47**: W191–W198.
Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**: 77.
Savitski MM, Mathieson T, Zinn N, Sweetman G, Doce C, Becher I, Pachl F, Kuster B, Bantscheff M. 2013. Measuring and Managing Ratio Compression for Accurate iTRAQ/TMT Quantification. *J Proteome Res* **12**: 3586–3598.
Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, Arnaiz O, Awedh MH, Baldock R, Barbiera G, et al. 2015. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res* **43**: W589–W598.
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.
Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, Schumacher SE, Wang C, Hu H, Liu J, et al. 2018. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**: 676-689.e3.
Ting L, Rad R, Gygi SP, Haas W. 2011. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat Methods* **8**: 937–940.
Upadhya SR, Ryan CJ. 2021. *Experimental reproducibility limits the correlation between mRNA and protein abundances in tumour proteomic profiles*. Systems Biology http://biorxiv.org/lookup/doi/10.1101/2021.09.22.461108 (Accessed November 30, 2021).
Wright MW, Bruford EA. 2011. Naming “junk”: Human non-protein coding RNA (ncRNA) gene nomenclature. *Hum Genomics* **5**: 90.
Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, Darnell JE. 2003. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res* **13**: 1863–1872.