Pan-cancer Analysis of Trans-effects from Copy Number Amplifications.

Authors: Daniel Lewis, Chelsie Minor, Avery Bell

Abstract

Mutations are at the root of cancer. One type of mutation, copy number amplifications, can alter protein levels, leading to downstream effects that could cause or worsen cancer. Not much is known about the effects of CNVs across multiple types of cancer. In this study, *trans* effects of common CNVs across 7 different cancer types are evaluated using data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) dataset. We found that of the 307 genes that are significantly copied across cancer types, most duplication events fall on chromosome 8. An analysis of the *trans* effects of these genes showed that 62 effects were common among six of the seven cancers in the study, and that one of the proteins effected, YWHAZ, is related to the TP53 pathway. These findings point to chromosome 8 and YWHAZ as important avenues of study when considering the effects of copy number amplifications in multiple types of cancer.

Introduction

Cancer is a highly diverse set of diseases, with cancer type typically defined based on the anatomical site of origin, for example, colon cancer. Cancer type can greatly influence a patient's prognosis, with some cancer types, such as thyroid, having a high 5-year survival rate (98%, Howlander et al., 2019), and others, such as pancreatic, most often leading to early death (9%, Howlander et al., 2019). Statistics like these lead researchers to ask why certain cancers are more aggressive than others, why certain cancers evade immune response, and why certain cancers develop in specific tissues or metastasize throughout the body. The answers to these questions could make the difference between a poor prognosis and a cure.

Cancer is driven by mutations, alterations to a cell's normal genetic code that can have downstream effects on the cell's growth and maturation cycle. Not all mutations have the same severity: base substitutions, for example, do not affect the reading frame of the DNA, while insertions and deletions can affect all the base pairs downstream. One of the most serious kinds of mutation is copy number variation (CNV), which is the addition or subtraction of DNA fragments that are at least 1 kb or larger (Freeman et al., 2006). Changing an entire portion of chromosome can have devastating effects on the normal functioning of a cell, sometimes making it cancerous (Ding, Tsang, Ng, & Xue, 2014). CNVs have been discovered in many cancer types, including breast and ovarian cancer (Gonçalves et al., 2017).

CNVs directly affect gene dosage for the genes that have been altered, and they can have indirect effects, referred to as *trans* effects, by altering the abundance of other genes (Stranger et al., 2007). Previous research has found the *trans* effects of CNVs in clear cell renal carcinoma (Clark et al., 2019) and ovarian cancer (Zhang et al., 2016). These studies found direct relationships between CNVs and proteins commonly implicated in cancer, such as TP53 and CCNE1 (Zack et al., 2013).

Despite much research on CNVs within individual cancer types, there has been little done on CNVs and their *trans* effects across cancer types. This study will analyze the relationships between copy number amplification (CNVs in which genetic material has been duplicated) and protein expression across cancer types. A pan-cancer approach to investigating copy number effects will allow us to determine which copy number amplifications are common among different types of cancer and of those, which are significantly associated with changes in protein expression. Through finding chromosomes across cancer types with significant copy number amplifications and their *trans* effects, this study will shed more light on underlying mechanisms of cancer in different tissues, which can further the work of researchers determining why certain types of cancers so difficult to treat.

Materials and Methods

CPTAC Data

CNV data and proteomic data were obtained from the CPTAC dataset, a centralized data repository with proteomic, CNV and clinical data for eight cancer types (Edwards et al., 2015). We used seven of the cancer types in this study: glioblastoma (GBM), ovarian, head and neck squamous cell carcinoma (HNSCC), clear cell renal cell carcinoma (CCRCC), endometrial (ENDO), and lung adenocarcinoma (LUAD). The data table for each cancer type contains numeric values expressing CNV duplication and protein levels in cancerous tissue in different patients, normalized against noncancerous tissue samples from the same patients. Each column contained data for a gene or protein depending on the table. Each row contained data for each patient. The cptac Python package was used to provide access to the CPTAC data and to aid with calculations, see pypi.org/cptac.

Finding Genetic Outliers in Copy Number Variation

Using the CNV tables in the cptac data package for each of the seven cancer types, we determined the number of patients with significantly amplified values. For the purposes of the study, CNVs with duplication values above 0.2, indicating an amplification well above the normal value of 0, were considered significant and included in the count. With values representing the number of samples with significant amplification for each gene in each cancer type, we calculated which genes were outliers in their cancer type. Outliers were determined using the distribution of the gene counts and calculating 1.5x the 4th quartile of the count values (see Figure 1). Cutoff values for each cancer type are shown in Table 1. The code to calculate these cutoffs is located here: https://github.com/dlewis27/CNVpan-cancer/blob/master/findingCNVs/finding_genetic_outliers.ipynb

	ENDO	GBM	BRCA	HNSCC	LUAD	OVARIAN	CCRCC
# of Patients Above	8.5	10.0	41.0	19.5	14.5	42.5	10.0

Table 1. Cutoffs for genes to be considered in our analysis. Each number represents the minimum number of patients for whom a gene must show copy number variation (i.e., the gene abundance measures above 0.2 or below -0.2 for the patient). The cutoffs

for each cancer are different because certain cancers, such as ovarian cancer, have higher copy number variation across all genes, while others, such as endometrial cancers, generally have lower copy number variation.

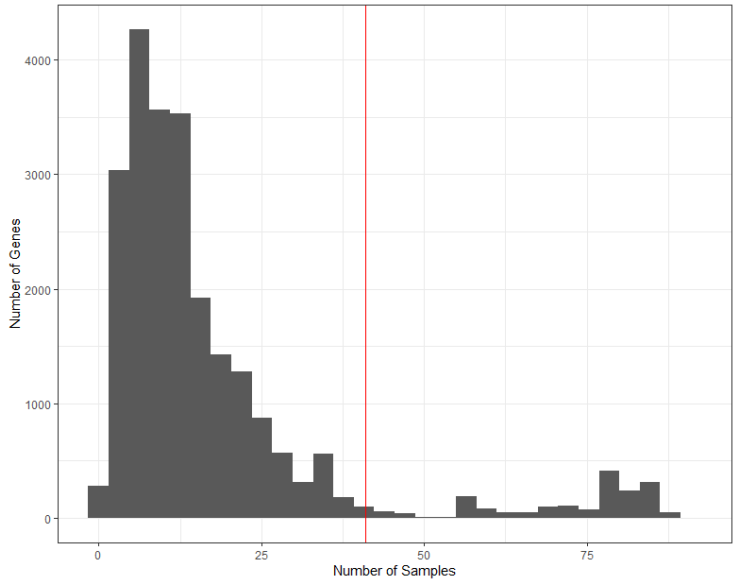


Figure 1. Distribution of Gene counts in Breast Cancer. Histogram showing the number of genes above the given threshold of 0.2 in Breast Cancer. In red is the calculated threshold for breast cancer (n=41.0). Any gene with >41.0 tumor samples above the threshold is considered to have a significant increase in copy number variation

After filtering to include genes that only met the patient cutoffs for each cancer type, the list of genes was further filtered to include only genes that were significant in multiple cancers. An intersection of the outlier genes for each cancer determined genes that appeared in multiple of the seven cancer types (see figure 2). Any gene found across 4 or more cancer types were kept for analysis which resulted in 307 genes of interest.

Trans **Effects**

The CNV data for the 307 genes was isolated in all cancer types. Linear regressions were performed for each of the 307 genes against all proteins using SciPy (Virtanen et al., 2020), a statistical python package, to find what proteins were affected by CNV events involving the 307 genes. For each gene, the corresponding column in the CNV table with data for that gene was used in a linear regression against every column in the protein table. This resulted in a list of CNV protein pairs representing the effects of the levels of CNV duplication of a gene on protein levels. The pairs were filtered using a Bonferroni-corrected p-value. The number of proteins considered was multiplied by the number of CNVs considered (number of tests) for each cancer type. The original p-value of 0.05 was divided by the number of tests to get the new p-value which resulted in a list of 806 CNV-protein pairs. The instructions to perform the linear regressions and calculate the p-value are located here: https://github.com/dlewis27/CNVpan-cancer/tree/master/transEffects/workflow.txt

Significant Pathways

Using these 806 CNV-protein pairs, we performed a simple intersect on all the cancers to find which CNV-protein pairs were significant (i.e., passed the p-value test) among all the cancers. None were found to be common between all 7 cancers, though a few were found when GBM was excluded. We also determined whether significant CNV protein pairs exhibited different linear regression slopes between cancers. The code to analyze patterns in the linear regressions is located here: https://github.com/dlewis27/CNVpan-cancer/blob/master/transEffects/Effects_Across_Cancers.ipynb

To determine the pathways these gene-protein pairs were involved in, we performed a gene set enrichment analysis (GSEA) (Subramanian et al., 2005) using the python gseapy package (https://pypi.org/project/gseapy/). This package compares a list of genes to known pathways, indicating what pathways the genes influence. The package includes access to several gene set libraries. Our data was run against 7 libraries: GO Biological Process 2018, GO Cellular Component 2018, GO Molecular Function 2018, NCI-Nature 2016, Reactome 2016, Wikipathways 2019 Human, and KEGG 2019 Human. The code to perform GSEA is located here: https://github.com/dlewis27/CNVpan-cancer/blob/master/pathways/GSEA.ipynb

Results

Copy Number Amplifications Across Cancer Types

To narrow down which CNVs may be significant in cancer, we determined which genes in each of the seven cancers were commonly duplicated in multiple patients (see Methods). From these findings it is apparent that the number of copy number amplifications vary across cancer type (see Figure 2). While some types of cancer such as HNSCC and GBM have many patients with a high number of copy number amplifications, other cancer types, such as ovarian cancer have relatively few copy number amplifications common across multiple patients.

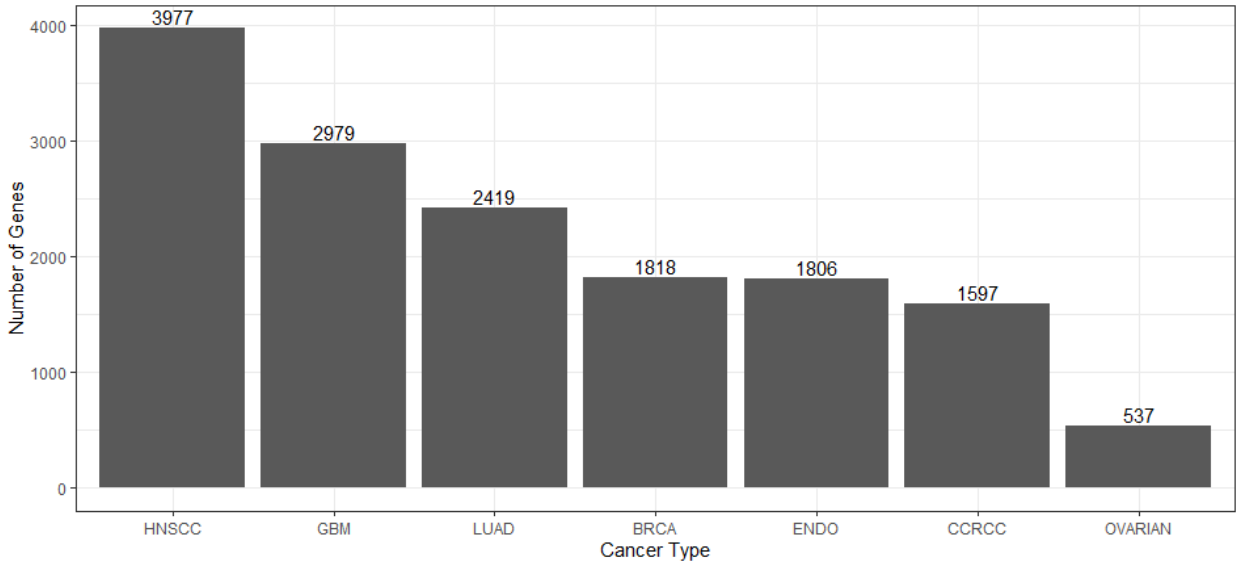


Figure 2. Number of Considered Genes by Cancer Type. The number of genes with many samples (more than the cutoff values in table1) that show amplifications above what we would normally suspect (>0.2). See Methods.

Previous studies have shown that cancer may be linked to chromosomal instability (Bakhoum & Compton, 2012), which leads to an increase in duplication and deletion events on certain chromosomes. We analyzed trends in copy number variation with respect to chromosome number and found certain chromosomes exhibit high copy number variation in different cancers. Figure 3 shows the chromosomes that are important contributors of copy number variation in each of the seven cancer types in this study.

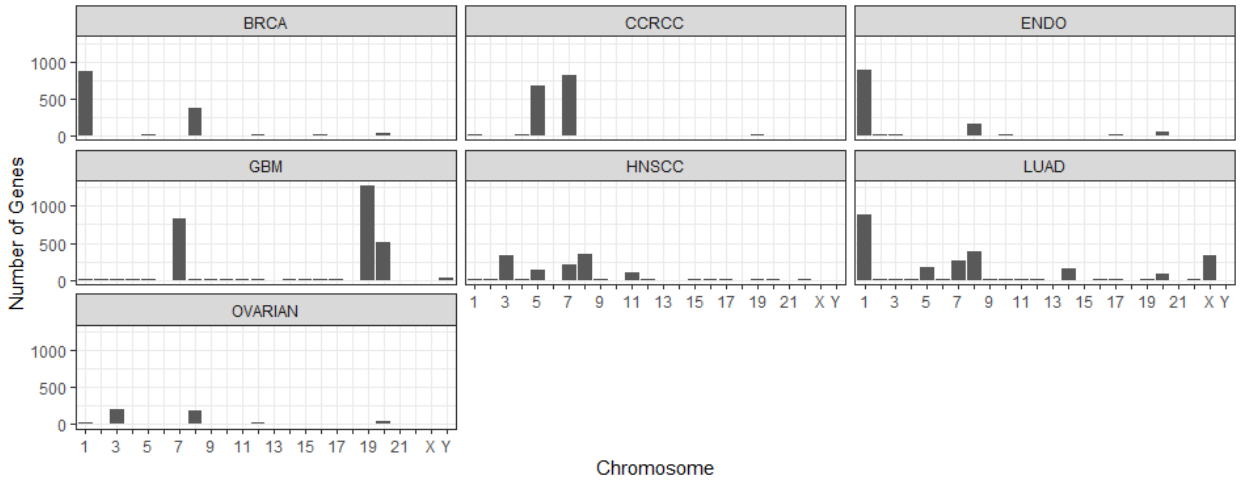


Figure 3. Chromosomal locations of Copy Number Amplification in 7 cancer types. Locations and counts of copy number variants differ across cancer types.

We also were interested in understanding how similar the copy number variants were between cancer types. Figure 4 shows the numbers of genes in CNVs that each pair of cancers have in common. LUAD, BRCA, and endometrial cancer share many of the same variants, while certain cancers, such as ovarian and clear cell renal cell carcinoma share few variants. This suggests that cancers with similar copy number variants may be analyzed together for greater insights.

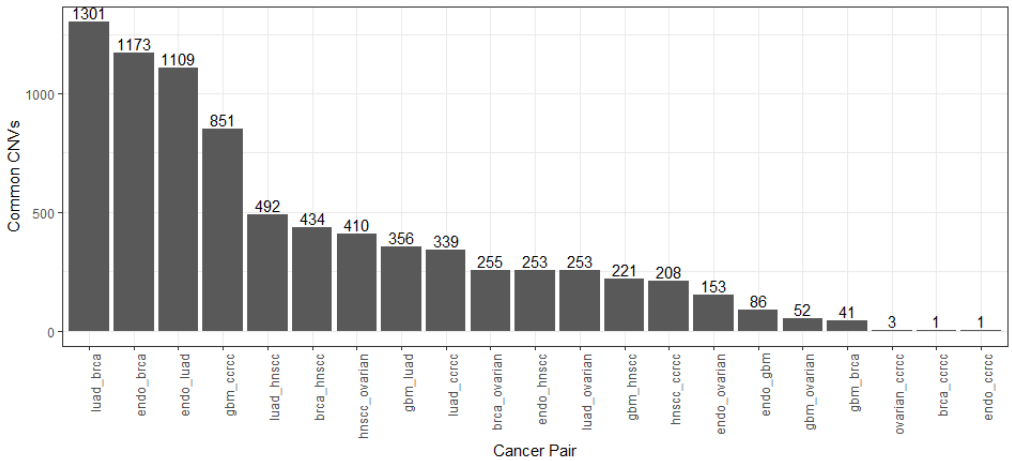


Figure 4. Similarities in copy number variants in cancer pairs. Along the x axis, there are pairs of cancers that we compared. Along the y axis, there is the number of significant gene amplifications that each pair shares in common.

In addition to understanding the impact of copy number variations in individual cancer types, we are interested in understanding copy number variants that are common across multiple cancers. After finding the genes that were significant in 4 or more cancer types, we obtained a list of 307 genes to be studied. We also discovered an interesting pattern that emerges on the chromosomes of these genes. While each cancer type seems to be characterized by copy number alterations on different chromosomes, when analyzed through a pan-cancer lens, copy number alterations tend to converge on chromosome 8 (see figure 5). Mapping these genes on the chromosomes, it becomes apparent that the genes are clustered on the q arm of chromosome 8 (see figure 6). The remainder of this paper will focus on this list 307 genes and their effects.

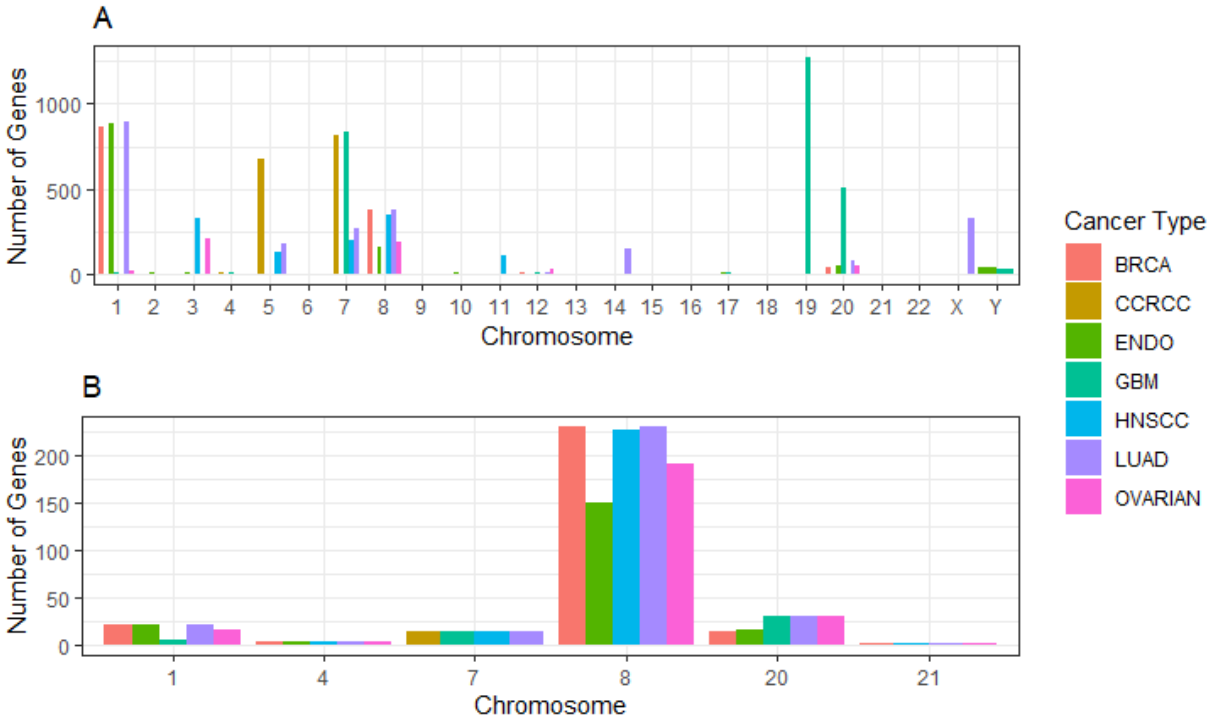


Figure 5. Importance of Chromosome 8 in Pan-Cancer Analysis. A) Location of significant genes before filtering for genes common across multiple cancer types. B) Location of significant genes after filtering for genes common across multiple cancer types. While chromosome 8 appears relatively insignificant among individual cancer types, when looking at genes common among many cancer types, chromosome 8 appears to have significant effect.

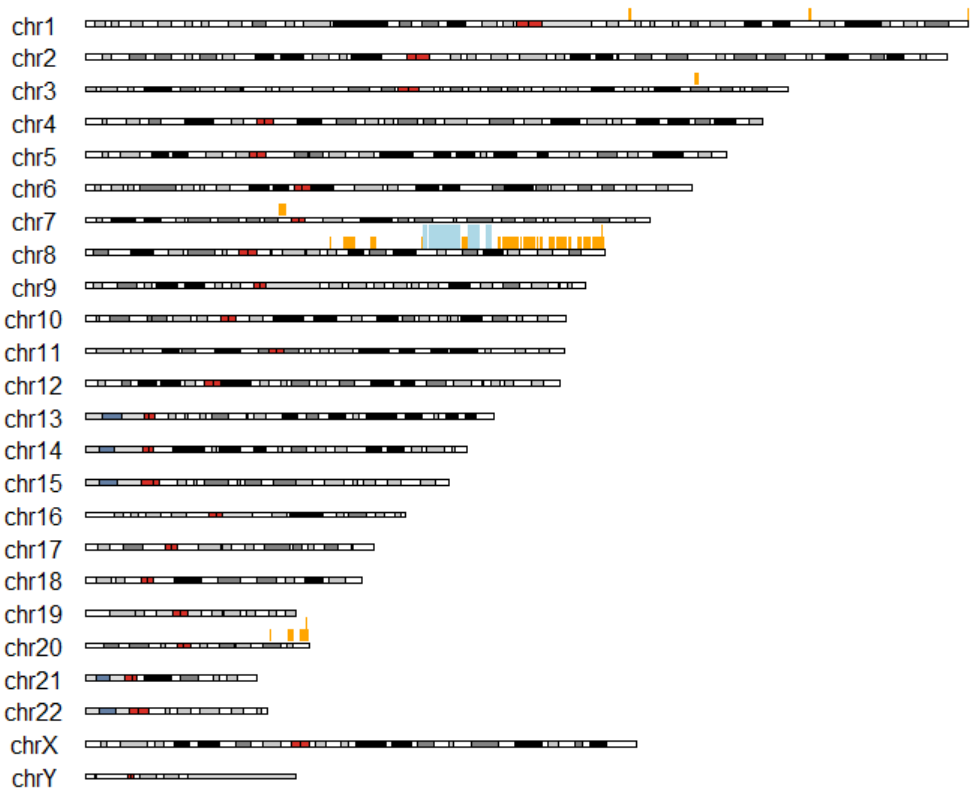


Figure 6. The location of the 307 genes are shown in orange. The 62 genes with significant effects on proteins across all cancer types except GBM are shown in light blue and are grouped in the same location on chromosome 8. This is indicative of a possible carcinogenic CNV event occurring on the q arm that may be found across various cancers.

Trans **Effects of Significant Copy Number Amplifications**

Linear regressions were performed with CNV data for the 307 genes against data for all proteins for each cancer. This gave CNV protein pairs with the effects of the CNV on each protein. A Bonferroni correction was applied to the CNV protein pairs of each cancer type ($p = .05$). This yielded 806 gene-protein pairs that showed significant correlation in the seven cancers. Certain cancers had more significantly correlated gene-protein pairs than others (see Supplementary Figure 1). There were some amplified genes that had both a positive and negative effect on different proteins within a cancer type, but no genes that had opposite effects on the same protein in different cancers (see Supplementary Data 1).

62 gene-protein pairs were significantly correlated in all cancer types except for GBM. Pairs from GBM were excluded because including it resulted in no common pairs between all cancers. These gene-protein pairs represent 62 unique genes that affect two unique proteins. The genes and proteins were positively correlated in the six cancers in which they are significant (see Figure 7). The cancers can be

grouped into high (0.5-0.7) and moderate (0.2-0.4) linear regression slopes for each protein.

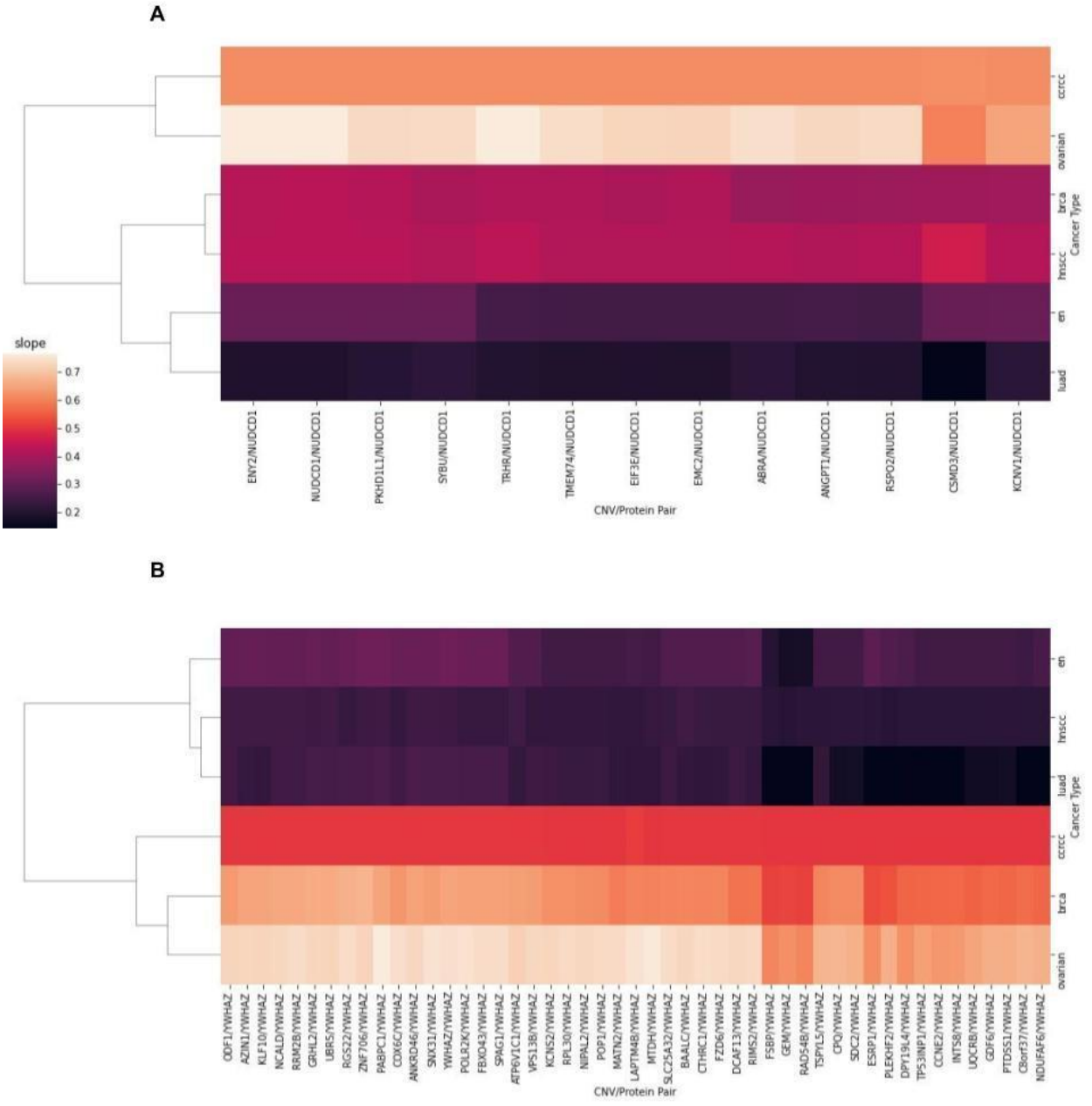


Figure 7. Effects of CNVs on two proteins related to cancer. Linear regressions were performed on every CNV/protein pair for every cancer. Significant gene/protein pairs were determined using a Bonferroni-corrected P-value. 62 significant gene/protein pairs were common among all the cancers except for GBM. All 62 involved one of two proteins: A) NUDCD1 or B) YWHAZ, which are both implicated in cancer (Han et al., 2018; Shi et al., 2019). Cancer types fell into two main groups: those for whom more copies of a gene lead to a greater increase in protein (high slope=yellow/orange) and those for whom more copies of a gene did not greatly increase the protein (low slope=purple/indigo). The branched lines on the left indicate additional similarities between cancers.

Pathways Affected Across Cancer Types

All 62 genes were located near each other on chromosome 8 and the 2 proteins are involved in similar functions, suggesting a copy number increase in this arm of chromosome 8 has a significant *trans* effect. The location of the 307 genes and 62 genes are shown in figure 6.

To further understand the role of this copy number increase involving 62 genes, a gene set enrichment analysis was performed. Selected results of the GSEA are shown in table 2. Pathways indicated involvement in protein transfer, TP53 regulation, and cell division.

Gene Set	Pathway	Overlap	P-value	Genes
Reactome	Transcriptional Regulation by TP53 Homo sapiens	6/348	0.00071	RRM2B, CCNE2, TP53INP1, COX6C, YWHAZ, POLR2K
NCI-Nature	Wnt signaling network Homo sapiens	2/28	0.00339	FZD6, CTHRC1
KEGG	Oocyte meiosis	3/125	0.00689	CCNE2, FBXO43, YWHAZ
Wikipathways	Endoderm Differentiation	3/141	0.00957	UBR5, PABPC1GRHL2
GO Molecular Function	hydrogen ion transmembrane transporter activity	2/51	0.01093	UQCRB, ATP6V1C1
GO Cellular Components	nucleolar part	3/153	0.01194	POP1, POLR2K, MTDH

Table 2. Gene Set Enrichment Analysis Pathways. Pathways involving the 62 genes were found through a gene set enrichment analysis. The gene list was compared against 6 different datasets. The results with the lowest p-value of each gene set are shown in the table above, ordered from smallest to largest p-value. Full results of the GSEA can be found at https://github.com/dlewis27/CNVpan-cancer/tree/master/pathways

One particularly interesting pathway is transcriptional regulation by TP53. TP53 is famously implicated in cancer. YWHAZ, one of the proteins significantly affected by the chromosome 8 copy number variation across cancers, is also involved in that pathway. Figure 8 shows the complete transcriptional regulation by TP53 pathway and highlights where YWHAZ is involved.

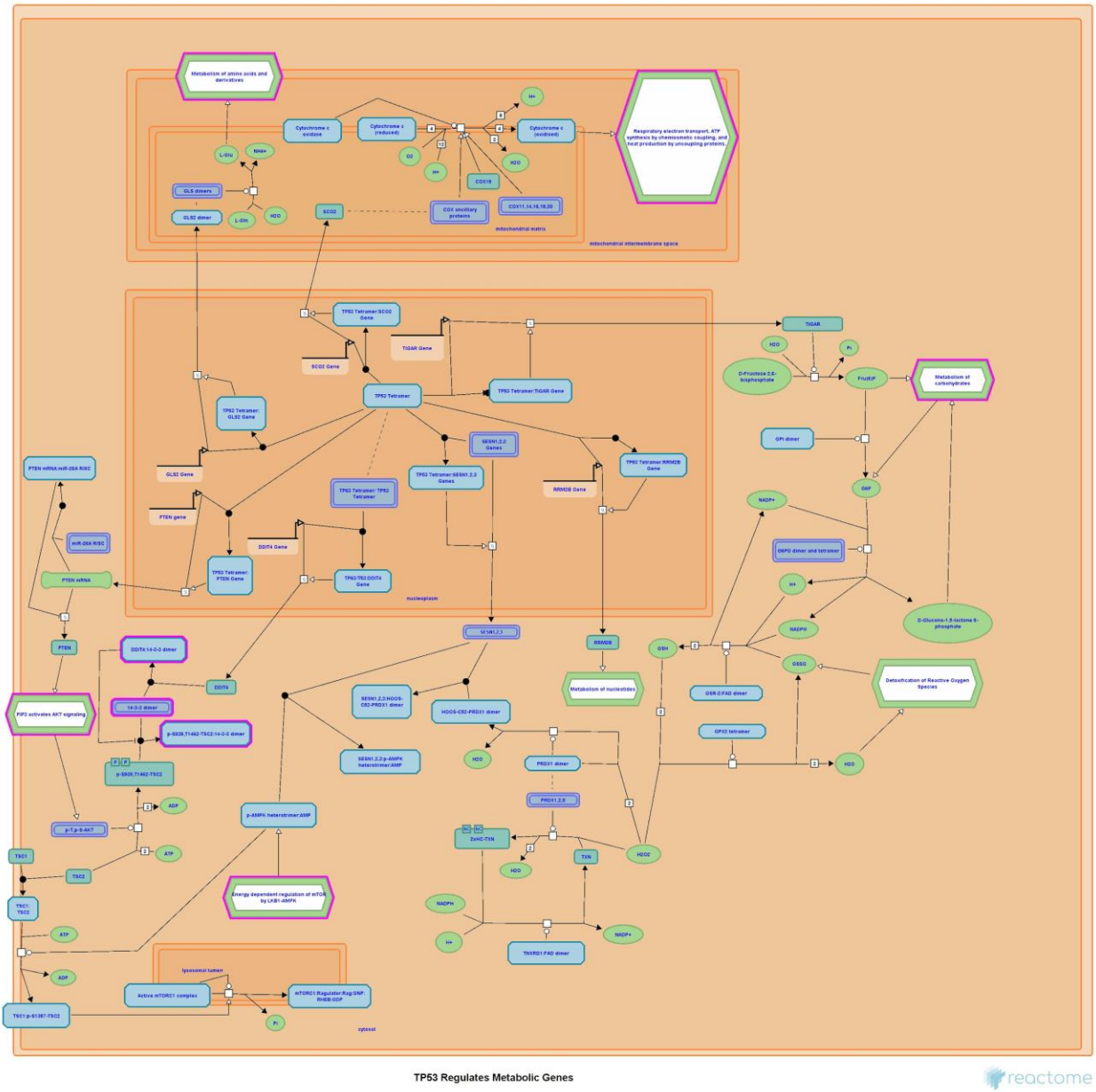
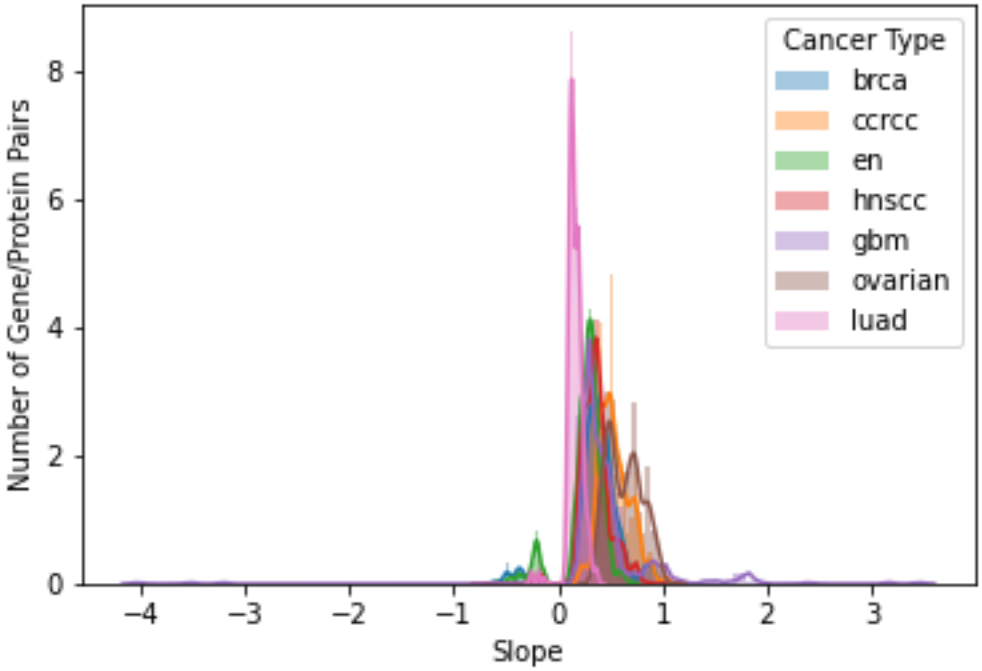


Figure 8. The Reactome Pathway "Transcriptional Regulation by TP53." The protein YWHAZ that was upregulated by copy number amplifications in six of seven cancers is involved in the steps highlighted in pink.

Conclusion

In our pan cancer analysis of copy number variation, a region on chromosome 8 (q22.1-q23.3) with several duplicated genes common across multiple cancer types was identified. This region of the chromosome has not yet been studied in relation to cancer, presenting a novel area for additional cancer research. A *trans* effect analysis of genes within the specified region of chromosome 8 identified two genes that seem to be influenced by the CNV event: NUCDC1 and YWHAZ, both of which have previous known connections to cancer. We additionally found that several of the 62 protein pairs play

critical roles in the TP53 pathway, including YWHAZ. Additional research extending our findings across other cancer types will further aid in providing a better understanding of the impact of copy number variations on protein expression and cancer. Further research on the role of YWHAZ in cancer may also lead to additional findings.



Supplementary Figure 1. The distribution of slopes from the 806 significant linear regressions (gene in the x axis against protein in the y axis). Most of the slopes are positive but some are negative. Certain cancers, such as BRCA (blue), endometrial (green), and LUAD (pink) showed both positive and negative relationships between gene and protein. Other cancers, such as GBM (purple) showed a high spread of slopes, some of which were well above the normal curve that the rest of the cancers followed.

Supplementary Data 1. To view the results of searching for gene/protein pairs that exhibited positive slope in one cancer and negative slope in another, please download and run this Python notebook: https://github.com/dlewis27/CNVpan-cancer/blob/master/transEffects/Effects_Across_Cancers.ipynb

References

Bakhoum, S. F., & Compton, D. A. (2012). Chromosomal instability and cancer: A complex relationship with therapeutic potential. *Journal of Clinical Investigation*, *122*(4), 1138–1143. https://doi.org/10.1172/JCI59954

Clark, D. J., Dhanasekaran, S. M., Petralia, F., Pan, J., Song, X., Hu, Y., … Zhang, H. (2019). Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. *Cell*, *179*(4), 964-983.e31. https://doi.org/10.1016/j.cell.2019.10.007

Ding, X., Tsang, S.-Y., Ng, S.-K., & Xue, H. (2014). Application of Machine Learning to Development of Copy Number Variation-based Prediction of Cancer Risk. *Genomics Insights*, *7*, GEI.S15002. https://doi.org/10.4137/GEI.S15002

Edwards, N. J., Oberti, M., Thangudu, R. R., Cai, S., McGarvey, P. B., Jacob, S., … Ketchum, K. A. (2015). The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *Journal of Proteome Research*, *14*(6), 2707–2713. https://doi.org/10.1021/pr501254j

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., … Lee, C. (2006). Copy number variation: New insights in genome diversity. *Genome Research*. https://doi.org/10.1101/gr.3677206

Gel, B., & Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics (Oxford, England)*, *33*(19), 3088–3090. https://doi.org/10.1093/bioinformatics/btx346

Gonçalves, E., Fragoulis, A., Garcia-Alonso, L., Cramer, T., Saez-Rodriguez, J., & Beltrao, P. (2017). Widespread Post-transcriptional Attenuation of Genomic Copy-Number Variation in Cancer. *Cell Systems*, *5*(4), 386-398.e4. https://doi.org/10.1016/j.cels.2017.08.013

Han, B., Zhang, Y.-Y., Xu, K., Bai, Y., Wan, L.-H., Miao, S.-K., … Zhou, L.-M. (2018). NUDCD1 promotes metastasis through inducing EMT and inhibiting apoptosis in colorectal cancer. American Journal of Cancer Research, 8(5), 810–823. Retrieved from https://pubmed.ncbi.nlm.nih.gov/29888104

Lehrer, S., Green, S., Ramanathan, L., Rosenzweig, K., & Labombardi, V. (2012). No consistent relationship of glioblastoma incidence and cytomegalovirus seropositivity in whites, blacks, and hispanics. *Anticancer Research*, *32*(3), 1113–1115.

Image for “Transcriptional Regulation by TP53". Reactome, 72, https://reactome.org/PathwayBrowser/#/R-HSA-5628897&SEL=R-HSA-206099&PATH=R-HSA-74160,R-HSA-73857,R-HSA-212436,R-HSA-3700989&FLG=P63104 Retrieved April 6, 2020.

Shi, J., Ye, J., Fei, H., Jiang, S.-H., Wu, Z.-Y., Chen, Y.-P., … Yang, X.-M. (2019). YWHAZ promotes ovarian cancer metastasis by modulating glycolysis. Oncology Reports, 41(2), 1101–1112. https://doi.org/10.3892/or.2018.6920

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazlsy, C., Thorne, N., … Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene phenotypes. *Science*, *315*(5813), 848–853. https://doi.org/10.1126/science.1136678

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43), 15545–15550.

https://doi.org/10.1073/PNAS.0506580102

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Yin, T., Cook, D., & Lawrence, M. (2012). ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biology*, *13*(8), R77. https://doi.org/10.1186/gb-2012-13-8-r77

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., … Beroukhim, R. (2013). Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, *45*(10), 1134–1140. https://doi.org/10.1038/ng.2760

Zhang, H., Liu, T., Zhang, Z., Payne, S. H., Zhang, B., McDermott, J. E., … Townsend, R. R. (2016). Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell*, *166*(3), 755–765. https://doi.org/10.1016/j.cell.2016.05.069