bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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.

Candidate targets of copy number deletion events across 17 cancer types

Qingyao Huang[a,b], Michael Baudis[a,b,*]

[a] *Department of Molecular Life Science, Winterthurerstrasse 190, Zurich, 8057, Zurich, Switzerland*
[b] *Swiss Institute of Bioinformatics, Winterthurerstrasse 190, Zurich, 8057, Zurich, Switzerland*

Abstract

Genome variation is the direct cause of cancer and driver of its clonal evolution. While the impact of many point mutations can be evaluated through their modification of individual genomic elements, even single copy number aberrations (CNAs) may encompass hundreds of genes and therefore pose challenges to untangle potentially complex functional effects. However, consistent, recurring and disease-specific patterns in the genome-wide CNA landscape imply that particular CNA may promote cancer type-specific characteristics. Discerning essential cancer-promoting alterations from the inherent co-dependency in CNA would improve the understanding of mechanisms of CNA and provide new insights into cancer biology and potential therapeutic targets. Here we implement a model using segment breakpoints to discover non-random gene coverage by copy number deletion (CND). With a diverse set of cancer types from multiple resources, this model identified common and cancer type-specific oncogenes and tumor suppressor genes as well as cancer-promoting functional pathways. Confirmed by differential expression analysis of data from corresponding cancer types, the results show that for most cancer types, despite dissimilarity of their CND landscapes, similar canonical pathways are affected. In 25 analyses of 17 cancer types, we have identified 20-170 significant genes by copy deletion, including RB1, PTEN and CDKN2A as the most significantly deleted genes among all cancer types. We have also shown a shared dependence on core pathways for cancer progression in different cancers as well as cancer-type separation by genome-wide significance scores. While this work provides a reference for gene specific significance in many cancers, it chiefly contributes a general framework to derive genome-wide significance and molecular insights in CND profiles with a potential for the analysis of rare cancer types as well as non-coding regions.

Keywords: Copy Number Aberration, Somatic Variation, Cancer Genomics

1. Introduction

Cancer genomes are characterized by a wide range of mutations in comparison to the unaltered germline genome. These "somatic" mutations emerge during an individuals life

*To whom correspondence may be addressed.
Email address: michael.baudis@mls.uzh.ch (Michael Baudis)

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

time and may accumulate sufficiently to lead to malignant transformation and tumorigenesis. Oncogenic mutations can impact the regulation and level of gene expression as well as the completeness and properties of gene products. While deviation from the physiological *status quo* typically impair cell viability, two features inherent to malignant transformation, genome instability and high replication rate, frequently promote the generation of a large pool of somatic genome alterations. This pool potentiates the selection of the sporadic cases where the mutated genome promotes a growth advantage and protection from apoptotic mechanisms. However, since most variations do not confer a strong growth advantage[1], the detection of the few key cancer-promoting variations hidden in a complex mutational landscape constitutes a major challenge in cancer genome research.

Depending on the scale of somatic variations, they can be grouped into specific sequence alterations ("point" mutations - e.g. SNVs, small INDELs) and structural variations, including copy number aberrations (CNAs). CNA change the dosage of the covered genetic elements in the affected segment and also may disrupt the local genomic context, e.g. by affecting regulatory elements.

Especially since the advent of next-generation sequencing technologies, point mutations have been thoroughly studied for their functional impact, for their effect as gain of function (GOF) mutations for oncogenes which promote proliferation or inhibit regulatory mechanisms; and loss of function (LOF) mutations for tumor suppressor genes (TSG) with negative impact on cell cycle control and other cellular surveillance functions. Here, the principal modes of action of oncogenes and TSG can be related to differing mutational characteristics, also known as an empirical "20/20 rule" [1]: 20% mutations within a gene locate at recurrent positions to define an oncogene, whereas in case of TSG, 20% mutations inactivate it at various locations.

Similarly to the functional attributions for point mutations, CNA can be separated into amplifications (GOF) and deletions (LOF). While deletion of a fraction of gene results in LOF due to truncated or untranscribed gene product, GOF requires amplification of the coding region and potentially regulatory part of a gene. Taken together, the mechanism and impact for amplification and deletion are dissimilar. In this study, we particularly focus on the copy number deletion (CND) patterns.

Whereas point mutations target one particular genetic element at the specific location, single CNDs potentially affect hundreds of genetic elements with a subsequent co-dependency of the affected genes. In addition, overall CNA involvement is highly correlated with the disease stage[2, 3, 4, 5], indicating an accumulation of unrepaired replication defects instead of a predominant selection of driver events. These factors present additional layers of complexity to distinguish the significant genes within large segmental CNA. Yet, CNAs manifest as genome-wide landscapes with frequently recurring features within related cancer types [6] (Figure 1). This observation implies that particular CNA patterns may be specifically tolerated and/or contain elements which provide selective advantage during malignant transformation and disease progression.

Earlier research by our group [7, 8] and others [9, 10] has described amplification and deletion hotspots among multiple cancer types. However, a systematic multi-cancer analysis for CNA-exerted susceptibility discovery is needed to compare between cancer types, extract relevant changes and delineate their functional impact. In recent years, work from several data curation projects and international research consortia has led to an improved availabil-

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

ity for generally compatible, genome-wide CNA profiling data associated information and thereby enabled the development and benchmarking of such an integrative approach. Particularly, the Progenetix CNA database has gathered 115k samples across 788 cancer types from published studies and cohorts, including the CNA data from 11k patients of 182 cancer types from the Cancer Genome Atlas (TCGA) Project[11, 12, 13].

In the last decade, GISTIC has been widely used to assess the significance of individual genomic regions in CNA data sets from individual genomic platforms [9, 14]. It uses a semi-parametric permutation to calculate a score for each probe based on both amplitude and frequency and identifies regions significant for amplification and deletion. However, beyond the probe-level and region-level significance discovery, it does not offer a statistical test for per-gene significance which would allow cross-study comparisons.

Here we describe an approach to evaluate per-gene significance in CND which utilizes the non-random features in gene disruption. We benchmark the gene sets identified as significant with known cancer driver gene sets and reduced gene expression in respective cancer types. Additionally, we explore the biological impact with pathway analysis and cancer type clustering.

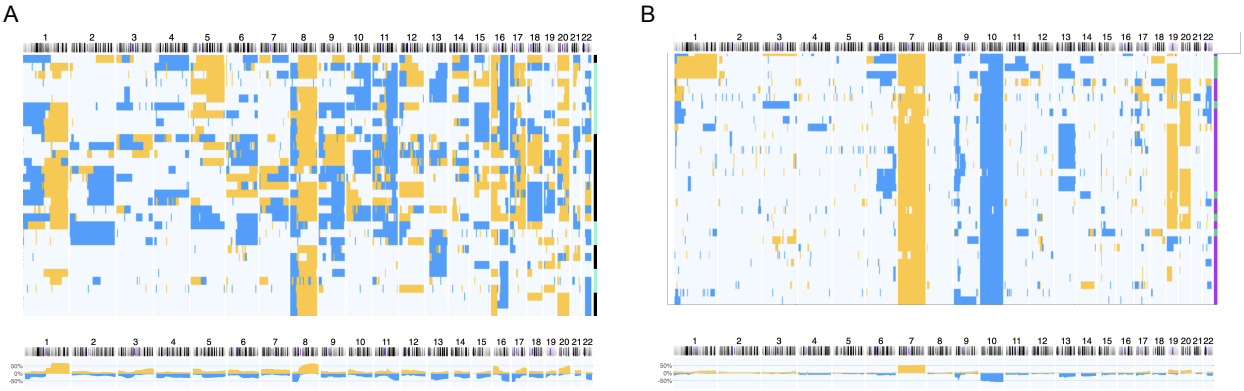

Figure 1: **Genome-wide copy number landscape across multiple samples.** Top panel stacks around 30 individual samples' CNA profile and the lower panel indicates the aggregated CNA landscape by percentage of samples exhibiting gain (+) and loss (-) across the genome. A) Ductal Breast Carcinoma, NCIT:C4017; B) Glioblastoma, NCIT:C3058

2. Results

2.1. Significance across multiple cancer types

We used data from three independent data sources depending on sample availability: 13 cancer types from the arrayMap resource, which represents a subset of the Progenetix database with available probe-specific genomic array data [15]; 12 cancer types from the TCGA project processed on genome-wide SNP6 arrays; as well as 4 cancer types from TCGA-GENIE project derived from whole exome sequencing (WES) experiments. Among these, 9 cancer types were represented by more than one source allowing comparison and benchmarking for source or technology related biases (Figure C.8).

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

We assessed the breakpoint density in the gene-dense and gene-poor regions within each analysis. While genome-wide SNP array derived datasets from TCGA and arrayMap sources show similar density, the WES-derived data from cBioPortal are biased against gene-poor regions, which causes inflation of gene significance level, making it not comparable with the array-derived data where probes are equally distributed across the whole genome (Figure C.9). Therefore, the WES-based results are not used to derive a gene set by a significance cutoff.

In the 17 cancer types, the number of significant genes in each analysis ranged from 23 to 168 (Table A.2). We used the three cancer-driving gene sets as "gold standard" to test enrichment of significant genes. With few exceptions, the significant genes from most analysis were enriched in all three driver sets (Table A.1). Many cancer-driving genes were shared by a few analyses. RB1 was found significant in 15 out of 25 analyses, followed by PTEN, CDKN2A, PTPRD, SMAD2, NRG1, JAK2, FHIT, DLC1, SMAD4, MAP2K4, RET, LRP1B, BRCA2, EPHA7, MLLT3, KANSL1, CTNNB1, APC, FGFR1, NCOR1, FLCN. Their functions spanned PI3K/Akt pathway, Cell cycle regulation, Wnt signaling and chromatin histone modification pathways. The cross-study significant genes showed local clusters of significance, particularly in chromosome 8p, 9p and 17p (Figure 2), while fewer others e.g. RB1, FHIT and PDE4D, appeared as singletons.

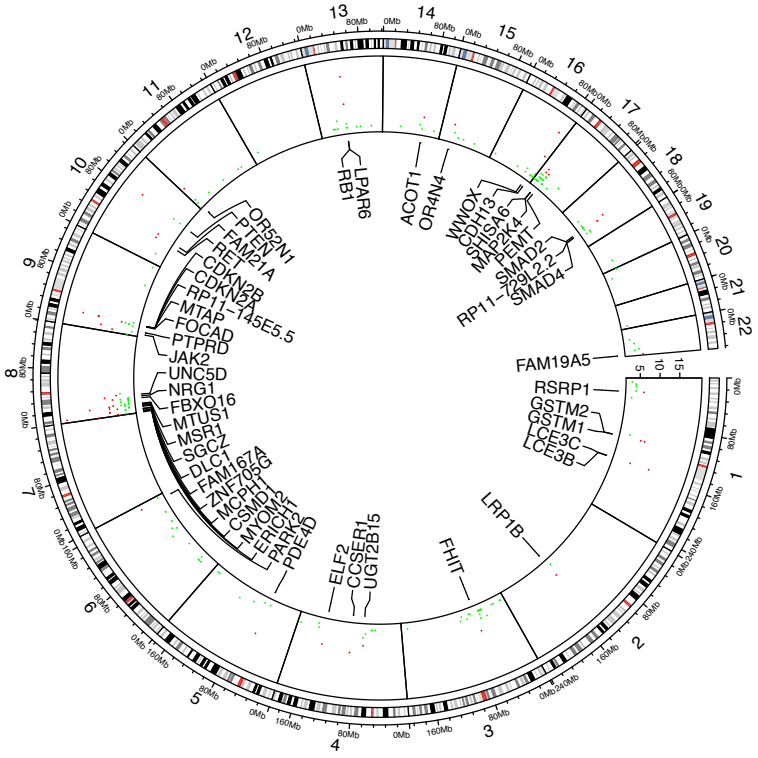

Figure 2: **Significant genes shared among performed analyses.** 49 genes found significant in more than 5 analyses are marked red and labeled accordingly.

2.2. Differential expression against normal tissue

Since cancer CNA has been previously found to be strongly correlated with gene expression, for a majority of genes, we expect that genes significantly covered by CND should have

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

reduced gene expression compared to the matched normal samples [16]. We tested this hypothesis with the RNA sequencing data of paired cancer and normal samples from TCGA. For 9 cancer types with available data, we determined a validation list of genes with significant mRNA expression reduction in tumor samples as compared to normal samples. We expect that the reduced dosage at DNA level may result in a cascade effect on downstream effectors with potentially magnified impact, and that other somatic variations unrelated to copy number alterations may additionally influence a wide range of expression values. In turn, the significant genes should demonstrate an evident reduction in expression but may not be among the most down-regulated by log fold change. Thus, we tested the over-representation of significant genes for CND with the respective cancer types. Here, all the tests were significant (Fisher exact p values range from 1.6×10^{-6} to 5×10^{-21}), corroborating the LOF at the expression level.

2.3. Cancer type clustering and pathway analysis

Using a FDR cutoff is useful to discover a set of significant genes within an analysis. However, a genome-wide evaluation with all significance scores can help to assess similarities between cancer types by CND. We used uniform manifold approximation and projection (UMAP) to reduce the 19k gene significance scores to two dimensions and indicated identical cancer types from different data sources with the same color (Figure 3). All matched cancer types are represented in a neighborhood, except for prostate adenocarcinoma. The separation between clusters is not pronounced, likely due to shared significance and small group size (2-3 studies per group).

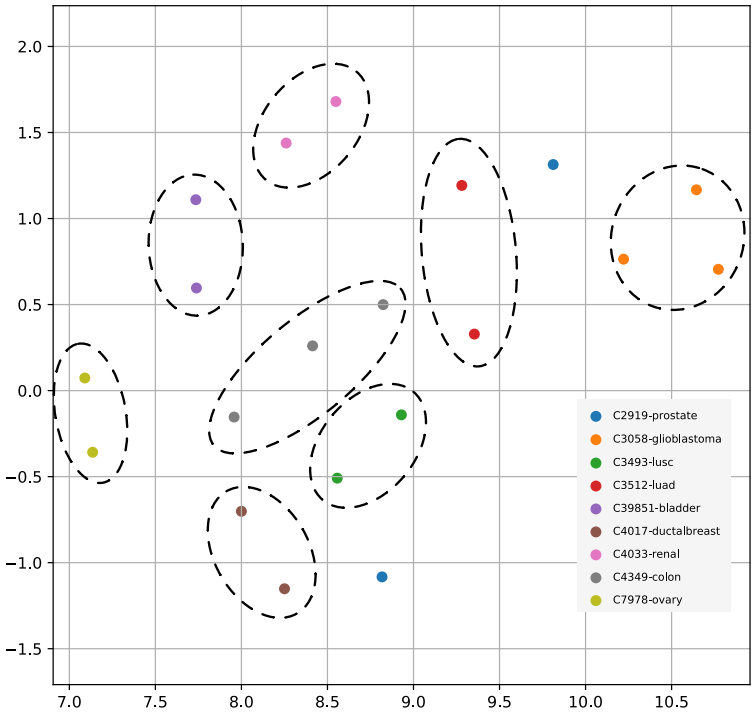

Figure 3: **UMAP projection of genome-wide significance scores among 9 paired cancer types from multiple data sources**

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

We also evaluated the dependency of the analyzed cancer types on different functional pathways. The clustering of pathways by significant genes showed a universal enrichment in the pathways related to cancer (Figure B.6; zoomed area of 48 pathways). Among these were "TGF-beta signaling", "stem cell", "viral infection", receptor signaling pathways (related to growth, apoptosis), "senescence", "energy metabolism" and "cell adhesion". With paired cancer types derived from different data source, two cancer types out of nine, ovary serous cystadenocarcinoma (C7978) and glioblastoma (C3058) were clustered together. Furthermore, a list of 29 canonical pathways from Supplementary Table 5 of [1] were used as hallmarks in cancer development. For these pathways, clustering was performed without standardization to compare the influence among cancer types as well as between multiple resources (Figure B.7). It was conspicuous that none of the analyses was enriched in mismatch repair pathway(MMR). MMR is common for familial cancers, including hereditary non-polyposis colorectal cancer (HNPCC) or Lynch Syndrome[17]. Microsatellite instability (MSI) caused by MMR defects is mutually exclusive with the chromosomal instability related to CNA [18], which is confirmed by the result. Three pathways, "TGF-beta signaling", "Cell cycle" and "p53 signaling", were enriched in a majority of cancer types. Ras signaling pathway was enriched in prostate adenocarcinoma from Array Map dataset and GnRH signaling pathway was enriched in ovary serous carcinoma from TCGA dataset.

We listed the top 10 pathways by enrichment significance in 25 analyses (Table B.3). "Pathways in cancer" appears in 11 analyses. Different types of viral infection appear in 12 analyses. Different types of drug metabolism appear in 9 analyses. "Metabolism of xenobiotics by cytochrome P450", "Cell cycle" and "Cellular senescence" appear in 7 analyses. "Chemical carcinogenesis" appears in 6 analyses.

3. Discussion

We have proposed a data-driven metric featured on frequency of gene disruption by breakage to discern non-random CNDs on gene level and generated a list of genes significantly affected by CND as well as a genome-wide significance score. From 29 independent runs on 18 cancer types, we have benchmarked the cancer-driving effects of the resulting genes through enrichment analysis for three independent "driver" gene sets and found significant enrichment all three sets. In addition, the genes show significant enrichment for cancer-related pathways and reduction in mRNA expression. Using the genome-wide significance scores we have clustered the analyses from multiple independent data resources and showed moderate separation between cancer types.

Genes with significant scores across multiple analyses include established tumor suppressor genes that control cell cycle and regulate proliferation and programmed cell death [19]. RB1 is implicated in multiple processes, including cell cycle, stress responses and apoptosis[20]. FOXO and PTEN are key regulators in phosphatidylinositol 3–kinase (PI3K) pathway which reacts to growth signals[21, 22]. CDKN2A(p16), CDKN2B(p15) at 9p21 controls S-phase checkpoint and their deletions have been reported in multiple cancers [23, 24], SMAD2/4 are involved in TGF-beta signaling pathway [25, 26]. The list also includes genes with sporadic or ambiguous oncogenetic attribution. FGFR1 has been reported to have both oncogenic and tumor suppressive potential [27] and the tyrosine kinase RET has long been established as a classic proto-oncogene but was found to act as a TSG in colorectal

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

carcinomas [28, 29]. As many high-level regulators have alternative cancer-promoting roles depending on the cellular context, the evidence of their selective deletion in a multi-cancer analysis provides additional support for their relevance in oncogenetic processes. Our results also point out genes with emerging cancer-related roles outside of classical cancer pathways such as GSTM1/GSTT1 in xenobiotic metabolism [30], ELF2 as an ETS transcription factor regulating various biological pathways[31] or DLC1 as a Rho-GTPase activating protein regulating *RhoA* pathway in hepatocellular carcinoma [32]. In addition, we have identified large genes residing in common fragile sites to be significantly affected by deletions and contributing to cancer development, including CSMD1, WWOX and FHIT [33].

Through our pathway analysis, we have observed prevalent enrichment in cancer-related pathways as well as hallmark mechanisms for cancer progression. In summary, the TGF-beta signaling pathway, cell cycle regulation and p53 signaling pathways emerged as the most frequently affected among all cancer types. Prostate, ovary and ductal breast adenocarcinoma samples were enriched for a majority of hallmark pathways, confirming their prominent dependence on CNA compared to discrete mutational events, as previously established[34].

The cancer type cluster separation by genome-wide significance scores is visible but not pronounced. The genomic CNA patterns as in Figure C.8 can be trained to separate the cancer types much better with image analysis techniques, but the simplified signatures are not readily interpretable for biological functions. The reduction to gene-based significance has certainly lost sizeable information, which may be related to the vast amount of non-coding areas unexplored here. Indeed, as an example in the local context of CDKN2A/B, a long non-coding RNA ANRIL is responsible for the transcriptional regulation, miRNA interaction, which modulates proliferation, senescence, motility and inflammation[35]. Additionally, heterogeneity within the sample set, non-CNA driven cancer samples as well as shared significance of core CNA genes may have overshadowed the less impactful cancer type-specific genes for the cancer type clustering.

In this article, we have developed a method to extract non-random significance from copy deletion in cancer. Specific features in the gene deletion or disruption mechanism is exploited to emphasize the implication in biological function. On the other hand, to discern non-random copy gain significance, such a framework can be adapted. Namely, cancer-promoting genes are expected to show under-represented disruption by either gain or loss copy endpoints as well as over-representation of endpoints in close proximity to gene start and end. In summary, we provide a general framework for integrative analysis on copy number deletion. It has confirmed well-known tumor suppressor genes as well as identified genes with incomplete characterization of their mode of action, suggesting the value in novel discovery and promoting further research into less studied genes. With the growing collection in high-quality CNA data, this method can be expanded to rare cancers which will potentiate discovery of novel cancer susceptibilities and dependencies and complement the overall understanding of malignancy development. Confirmed by the functional characterization of the known coding genes, this tool can be extended to the non-coding area and provide a better overview of the CNA functional landscape.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

4. Experimental procedures

4.1. Data availability

CNA data has been accessed from three different sources, which had been integrated into the Progenetix database:

- arrayMap, as a subset of Progenetix, contains CNA data derived from an in-house processing pipeline of Affymetrix 250K Nsp/Sty assays and Genome-wide SNP6 array platforms. This subset includes prostate adenocarcinoma (C2919), medulloblastoma (C3222), lung adenocarcinoma (C3512), ductal breast carcinoma (C4017), clear cell renal cell carcinoma (C4033), ovary serous cystadenocarcinoma (C7978), plasmacytoma (C9349), glioblastoma (C3058), lung squamous cell carcinoma (C3493), gastric adendocarcinoma, esophageal adenocarcinoma (C4025), colon adenocarcinoma (C4349), diffuse Large B-Cell Lymphoma, NOS (C80280).

- TCGA provides pre-processed CNA segment data derived from Genome-wide SNP6 arrays. The data used here includes prostate adenocarcinoma (C2919), hepatocellular carcinoma (C3099), lung adenocarcinoma (C3512), ductal breast carcinoma (C4017), thyroid gland papillary carcinoma (C4035), endometrial endometrioid adenocarcinoma (C6287), glioblastoma (C3058), lung squamous cell carcinoma (C3493), bladder urothelial carcinoma (C39851), clear cell renal cell carcinoma (C4033), colon adenocarcinoma (C4349), ovary serous cystadenocarcinoma (C7978).

- TCGA-GENIE project, accessed from cBioPortal provides pre-processed CNA segment data derived from WES experiments. This includes: glioblastoma (C3058), bladder urothelial carcinoma (C39851), colon adenocarcinoma (C4349), pancreas (C8294).

For differential expression analysis, transcriptomics data in raw HT-Seq counts was accessed for respective TCGA projects from GDC Data Portal. Paired RNAseq data was available for 11 cancer types: prostate adenocarcinoma (C2919), hepatocellular carcinoma (C3099), lung adenocarcinoma (C3512), ductal breast carcinoma (C4017), thyroid gland papillary carcinoma (C4035), endometrial endometrioid adenocarcinoma (C6287), lung squamous cell carcinoma (C3493), bladder urothelial carcinoma (C39851), clear cell renal cell carcinoma (C4033), colon adenocarcinoma (C4349), ovary serous cystadenocarcinoma (C7978).

4.2. Method description

The model design is based on the observed non-random features discovered in CND which can be harnessed to characterize its underlying mechanism. On the one hand, we demonstrate that CNDs are minimized in length for gene-rich regions. In all analyzed cancer types, we show the decay of average gene counts as the CND size increases (Figure 4). A genome-wide average of gene counts per 100kb is about 1.5. For CND segments shorter than 5Mb, this number amounts to 10-25, while it decays below genome average as segment length increases to 10-20Mb and stabilizes at 0.5-0.8 depending on the cancer type. This indicates that CND is targeted towards specific genes and long un-targeted CND in gene-rich regions is unfavored and eliminated.

CND does not require a full coverage of the gene element to disrupt the transcription or integrity of gene product. We show that they tend to recur and locate within the range of

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

cancer-related genes. We have tested the recurrence with CND segment endpoints in expert curated driver genes [36] against the non-driver genes. The localization of breakpoints in driver gene sets is highly over-represented among all but one datasets (Fisher exact p value in the range of 1.17×10^{-9} to $<2.225 \times 10^{-328}$, with exception of medulloblastoma at 0.07).

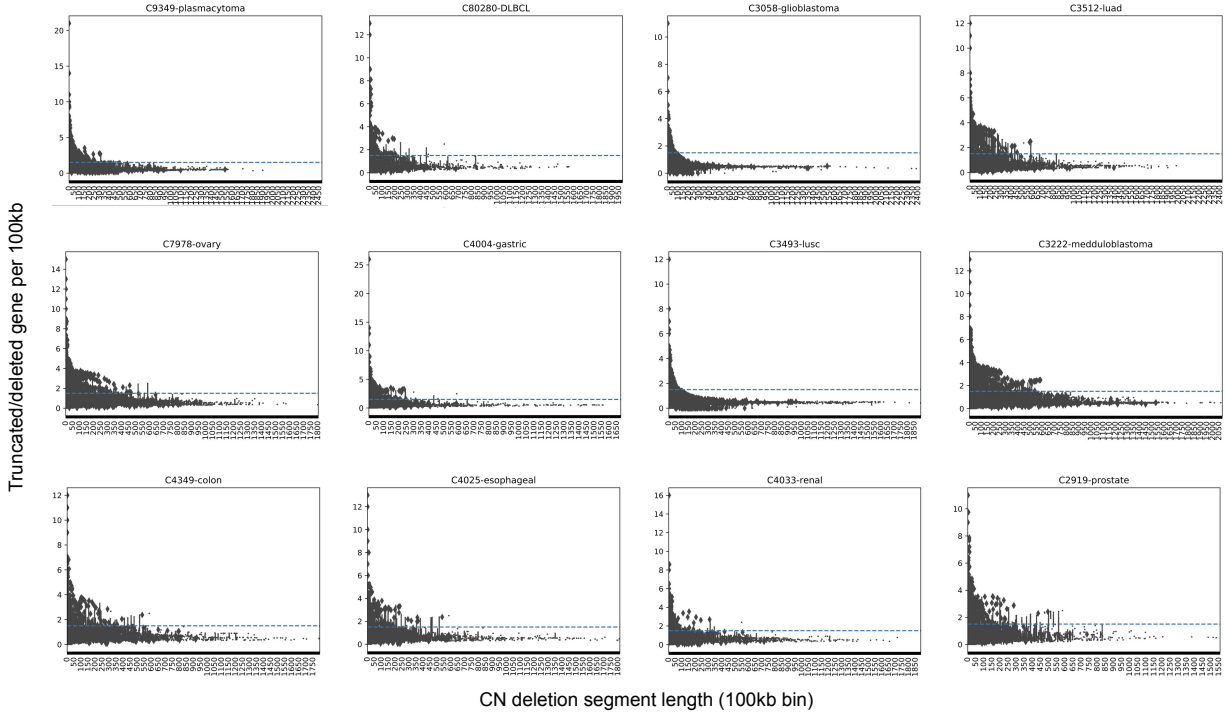

Figure 4: **Gene density in relation to copy deletion segment length** Gene density decreases as deletion segment (binned by 100kb) increases in all analyses. Horizontal dash line shows the genome-wide average of gene density.

4.3. General model design

Based on these initial observations, we then designed a model to capture these non-random features in CNDs as illustrated in Figure 5. In each analyzed cancer type, we aggregated all CND segments and created new intervals in a "reference genome track" for every new endpoint. We calculated a gene score for all genes based on abundance as well as penalizing the number of endpoints locating within the gene. Then we generated a background distribution with chromosome-wide CN reshuffling to derive gene-wise significance (gene score).

The gene score was defined to reward the high number of sample recurrence and penalize the length of segments and genes:

$$Score_g = \sum_{i=1}^{N} \left(\frac{S_i}{L_i + L_g}\right) \tag{1}$$

For each gene g, a score is defined by summing up all overlapping deletion segments i, the division of sample count S_i and the sum of fragment length L_i and gene length L_g.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

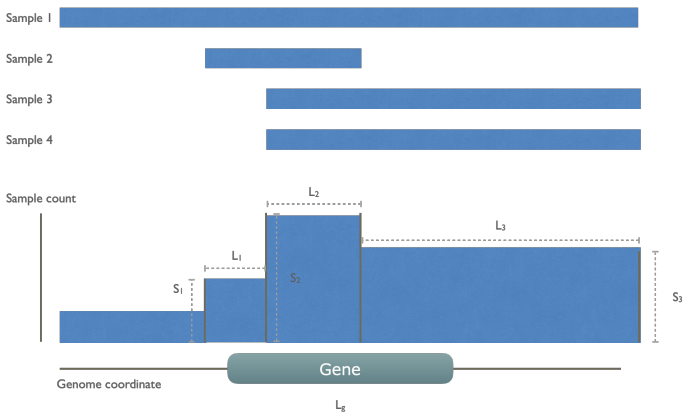

Figure 5: **A graphical illustration of gene score calculation** In this example, there are four cancer samples, all of which have a whole or partial deletion on the indicated gene. These CN segments are collapsed to a collective track, leaving four collective segments, but only three of them overlap with gene. The gene score sums over these three segments with count of involved sample S_i divided by sum of segment length L_i and the common gene length L_g.

The positions of collective CNA segments are shuffled within the same chromosome. The gene scores are calculated for each shuffling, to generate a background score distribution for each gene. The gene score on the real data is compared with the background distribution to calculate a empirical p value to denote the gene's significance.

This way of shuffling preserves the genomic context and gene neighborhood structure. Consider two genes A and B in close proximity. If they are within the same CND segment (co-segregation, equal disease relevance), the randomization would give them same background sample count. The gene-length difference is also reflected in the background rate and they will have the same significance score. In contrast, if gene A is more disease-relevant, there would be more CND segments overlapping with gene A, resulting in smaller collective segments for gene A score calculation. With the similar background rate due to location proximity, gene A will hence a higher score compared to gene B.

In addition, the method accounts for the variation in overall CNA content. Specifically, if a set of samples mostly consists of profiles with an overall low amount of CNA, the probability of CNA at each given region remains and will not affect the outcome of the analysis. Finally, it is robust for the noise introduced by baseline setting and segmentation errors from individual samples.

On the other hand, this method requires a dataset selection that compromises between sample size and cancer type specificity (for a representative and comparable CND profile), as a mixture of cancer subtypes with highly variable CNA landscape introduces breakpoint bias, and conversely a small sample size and few breakpoints result in low statistical power. Also, it is expected that in cases of chromothripsis-like events (CTLP; focal, extreme hyper-segmentation) [37, 38] the significance score of local genes can be increased and the significance of other genes on the same chromosome can be reduced due to the rise in the overall CNV rate. Such samples should be pre-filtered and investigated carefully on a case-by-case level.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

4.4. Differential expression analysis

We used R-package EdgeR for differential expression analysis between tumor and normal groups [39]. For each cancer type, gene-wise counts from paired tumor — normal samples were used. A TMM normalization was performed to calibrate for the library size (total counts) per sample. Negative binomial model was used to estimate the common and gene-wise dispersion parameters. A gene-wise general linear model was fit by the paired design and the differentially expressed genes were determined by likelihood ratio test.

4.5. Pathway analysis

All 326 KEGG pathways were used to determine the enrichment in each pathway with a Fisher exact test with the contingency table of significant genes - genes with a significance score on one axis and genes in/not in pathway on the other axis. For clustering analysis including all pathways, log10 of Fisher exact p values were calculated and standardized to 0-1 scale, i.e. 0 with lowest p. Hierarchical clustering with Euclidean distance and average linkage method was performed on both cancer types and pathways. For the clustering analysis of the 29 canonical cancer pathways, the original Fisher exact p value was used. Hierarchical clustering with Euclidean distance and average linkage method was performed on pathways only.

5. Acknowledgement

We would like to thank the scientific input from members of the Zurich Seminars in Bioinformatics as well as the Theoretical Cytogenetics and Oncogenomics group at University of Zurich for continuous work on data collection and curation for the Progenetix database.

6. Author contributions

QH conceived the project, performed analysis and wrote the manuscript. MB provided the CNA data assembly, gave insights about data analysis and clinical cancer biology and edited the manuscript.

7. Declaration of interests

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

References

[1] B. Vogelstein, N. Papadopoulos, V. Velculescu, S. Zhou, L. Diaz, K. Kinzler, Cancer genome landscapes., Science 339 (6127) (2013) 1546â–1558.

[2] M. Gerstung, C. Jolly, I. Leshchiner, S. C. Dentro, S. Gonzalez, D. Rosebrock, T. J. Mitchell, Y. Rubanova, P. Anur, K. Yu, et al., The evolutionary history of 2,658 cancers, Nature 578 (7793) (2020) 122–128.

[3] A. H. Shain, I. Yeh, I. Kovalyshyn, A. Sriharan, E. Talevich, A. Gagnon, R. Dummer, J. North, L. Pincus, B. Ruben, et al., The genetic evolution of melanoma from precursor lesions, New England Journal of Medicine 373 (20) (2015) 1926–1936.

[4] D. Tamborero, C. Rubio-Perez, F. Muiños, R. Sabarinathan, J. M. Piulats, A. Muntasell, R. Dienstmann, N. Lopez-Bigas, A. Gonzalez-Perez, A pan-cancer landscape of interactions between solid tumors and infiltrating immune cell populations, Clinical Cancer Research 24 (15) (2018) 3717–3728.

[5] H. Hieronymus, N. Schultz, A. Gopalan, B. S. Carver, M. T. Chang, Y. Xiao, A. Heguy, K. Huberman, M. Bernstein, M. Assel, R. Murali, A. Vickers, P. T. Scardino, C. Sander, V. Reuter, B. S. Taylor, C. L. Sawyers, Copy number alteration burden predicts prostate cancer relapse, Proceedings of the National Academy of Sciences 111 (30) (2014) 11139–11144. doi:10.1073/pnas.1411446111.

[6] P. C. Cordo, M. Baudis, Copy number variant heterogeneity among cancer types reflects inconsistent concordance with diagnostic classifications, bioRxiv (2021).

[7] M. Baudis, Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal cgh data, BMC cancer 7 (1) (2007) 1–15.

[8] N. Kumar, H. Cai, C. Von Mering, M. Baudis, Specific genomic regions are differentially affected by copy number alterations across distinct cancer types, in aggregated cytogenetic data (2012).

[9] R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson, M. Urashima, et al., The landscape of somatic copy-number alteration across human cancers, Nature 463 (7283) (2010) 899–905.

[10] C. Aouiche, B. Chen, X. Shang, Predicting stage-specific recurrent aberrations from somatic copy number dataset, Frontiers in Genetics 11 (2020) 160.

[11] Cancer Genome Atlas Research Network, J. Weinstein, E. Collisson, G. Mills, K. Shaw, B. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. Stuart, The cancer genome atlas pan-cancer analysis project., Nat Genet 45 (10) (2013) 1113–1120.

[12] National Cancer Institute, The Cancer Genome Atlas Program, `https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga`, accessed: 2021-05-20 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

[13] Q. Huang, P. Carrio-Cordo, B. Gao, R. Paloots, M. Baudis, The progenetix oncogenomic resource in 2021, Database : the journal of biological databases and curation 2021 (July 2021). doi:10.1093/database/baab043.
URL https://europepmc.org/articles/PMC8285936

[14] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, G. Getz, Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers, Genome biology 12 (4) (2011) 1–14.

[15] H. Cai, S. Gupta, P. Rath, N. Ai, M. Baudis, arraymap 2014: an updated cancer genome resource, Nucleic acids research 43 (D1) (2015) D825–D830.

[16] X. Shao, N. Lv, J. Liao, J. Long, R. Xue, N. Ai, D. Xu, X. Fan, Copy number variation is highly correlated with differential gene expression: a pan-cancer study, BMC medical genetics 20 (1) (2019) 1–14.

[17] P. Hsieh, K. Yamane, Dna mismatch repair: molecular mechanism, cancer, and ageing, Mechanisms of ageing and development 129 (7-8) (2008) 391–407.

[18] K. Søreide, Molecular testing for microsatellite instability and dna mismatch repair defects in hereditary and sporadic colorectal cancers–ready for prime time?, Tumor Biology 28 (5) (2007) 290–300.

[19] C. J. Marshall, Tumor suppressor genes, Cell 64 (2) (1991) 313–326.

[20] B. N. Chau, J. Y. Wang, Coordinated regulation of life and death by rb, Nature Reviews Cancer 3 (2) (2003) 130–138.

[21] L. Salmena, A. Carracedo, P. P. Pandolfi, Tenets of pten tumor suppression, Cell 133 (3) (2008) 403–414.

[22] N. Chalhoub, S. J. Baker, Pten and the pi3-kinase pathway in cancer, Annual Review of Pathology: Mechanisms of Disease 4 (2009) 127–150.

[23] J. Tsihlias, L. Kapusta, J. Slingerland, The prognostic significance of altered cyclin-dependent kinase inhibitors in human cancer, Annual review of medicine 50 (1) (1999) 401–423.

[24] W. D. Foulkes, T. Y. Flanders, P. M. Pollock, N. K. Hayward, The cdkn2a (p16) gene and human cancer, Molecular medicine 3 (1) (1997) 5–20.

[25] A. Nakao, T. Imamura, S. Souchelnytskyi, M. Kawabata, A. Ishisaki, E. Oeda, K. Tamaki, J.-i. Hanai, C.-H. Heldin, K. Miyazono, et al., Tgf-β receptor-mediated signalling through smad2, smad3 and smad4, The EMBO journal 16 (17) (1997) 5353–5362.

[26] S. A. Hahn, M. Schutte, A. S. Hoque, C. A. Moskaluk, L. T. Da Costa, E. Rozenblum, C. L. Weinstein, A. Fischer, C. J. Yeo, R. H. Hruban, et al., Dpc4, a candidate tumor suppressor gene at human chromosome 18q21. 1, Science 271 (5247) (1996) 350–353.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

[27] M. Katoh, H. Nakagama, Fgf receptors: cancer biology and therapeutics, Medicinal research reviews 34 (2) (2014) 280–300.

[28] C. Eng, Ret proto-oncogene in the development of human cancer, Journal of Clinical Oncology 17 (1) (1999) 380–380.

[29] Y. Luo, K. D. Tsuchiya, D. Il Park, R. Fausel, S. Kanngurn, P. Welcsh, S. Dzieciatkowski, J. Wang, W. M. Grady, Ret is a potential tumor suppressor gene in colorectal cancer, Oncogene 32 (16) (2013) 2037–2047.

[30] G. Ginsberg, S. Smolenski, P. Neafsey, D. Hattis, K. Walker, K. Z. Guyton, D. O. Johns, B. Sonawane, The influence of genetic polymorphisms on population variability in six xenobiotic-metabolizing enzymes, Journal of Toxicology and Environmental Health, Part B 12 (5-6) (2009) 307–333.

[31] A. Seth, D. K. Watson, Ets transcription factors and their emerging roles in human cancer, European journal of cancer 41 (16) (2005) 2462–2478.

[32] W. Xue, A. Krasnitz, R. Lucito, R. Sordella, L. VanAelst, C. Cordon-Cardo, S. Singer, F. Kuehnel, M. Wigler, S. Powers, et al., Dlc1 is a chromosome 8p tumor suppressor whose loss promotes hepatocellular carcinoma, Genes & development 22 (11) (2008) 1439–1444.

[33] D. I. Smith, Y. Zhu, S. McAvoy, R. Kuhn, Common fragile sites, extremely large genes, neural development and cancer, Cancer letters 232 (1) (2006) 48–57.

[34] G. Ciriello, M. L. Miller, B. A. Aksoy, Y. Senbabaoglu, N. Schultz, C. Sander, Emerging landscape of oncogenic signatures across human cancers, Nature genetics 45 (10) (2013) 1127–1133.

[35] Y. Kong, C.-H. Hsieh, L. C. Alonso, Anril: a lncrna at the cdkn2a/b locus with roles in cancer and metabolic disease, Frontiers in endocrinology 9 (2018) 405.

[36] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, S. A. Forbes, The cosmic cancer gene census: describing genetic dysfunction across all human cancers, Nature Reviews Cancer 18 (11) (2018) 696–705.

[37] H. Cai, N. Kumar, H. C. Bagheri, C. von Mering, M. D. Robinson, M. Baudis, Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens, BMC genomics 15 (1) (2014) 1–13.

[38] I. Cortés-Ciriano, J. J.-K. Lee, R. Xi, D. Jain, Y. L. Jung, L. Yang, D. Gordenin, L. J. Klimczak, C.-Z. Zhang, D. S. Pellman, et al., Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing, Nature genetics 52 (3) (2020) 331–341.

[39] M. D. Robinson, D. J. McCarthy, G. K. Smyth, edger: a bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (1) (2010) 139–140.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

[40] M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendl, J. Kim, B. Reardon, et al., Comprehensive characterization of cancer driver genes and mutations, Cell 173 (2) (2018) 371–385.

[41] F. Dietlein, D. Weghorn, A. Taylor-Weiner, A. Richters, B. Reardon, D. Liu, E. S. Lander, E. M. Van Allen, S. R. Sunyaev, Identification of cancer driver genes based on nucleotide context, Nature genetics 52 (2) (2020) 208–218.

Appendix A. Significance across cancer types

Table A.2: Significant genes in analyses

Source	NCIt	No.genes	Genes
arrayMap	C4017 (ductalbreast)	76	DGCR2,KANK1,AMY1A,AGPAT5,MSR1,SPG7,PARP4,CDRT4,COX10,CTNNB1,RFWD3, PRSS45,TUSC3,CUL4A,PCID2,SLC5A10,MAP2K4,CCDC25,TRPC6,BRCA2,CES1,WWOX, GSTM1,RHD,APOBEC3A,DLGAP2,SLC35F2,RBL2,SFMBT1,CSMD1,CDRT1,GSTM2, KMT2A,ZNF778,UGT2B15,USP17L8,TVP23C-CDRT4,OR4N4,FGFR1,ARHGAP44,UGT2B28, ZAR1L,TRIM16,MCPH1,RSRP1,CCDC134,USP6,DNAH9,PRSS50,JAK2,CWF19L2,ERICH1, NRG1,MYOM2,BLK,PTEN,ZNF705G,ARL17B,NUP50,ATXN3,FAM21A,ADAMTS18, SGCZ,UNC5D,SLC25A37,PSG9,KANSL1,SLC25A18,PI4KA,CDH13,FLT3,HYDIN, PARK2,RB1,HLA-C,CDH15
arrayMap	C3512 (luad)	47	AMY1A,LCE3C,MSR1,GRIN3A,GPC6,LINGO2,LIPK,ZNF385D,FAM46A,EPHA3, NKAIN2,NCAM2,XKR6,BRCA2,ZNF705B,GSTM1,SIRPB1,APOBEC3A,RBMS3,UFL1,RFX3, CSMD1,GSTM2,PCDH15,PTPRD,NEGR1,ROBO2,MTAP,GRIK2,JAK2,ERICH1,OR52N1, ZNF705G,SMAD2,SGCZ,PCDH17,ARPP21,PSG9,LCE3B,FOCAD,PCDH9,OR5B3,RETN, CNTN6,EYS,TUSC3,EPHA7
TCGA	C39851 (bladder)	53	HEXB,HSCB,CREBBP,LRP1B,CNTLN,CDKN2B,KIAA1456,SH3PXD2B,PHYKPL,FMN1, RASD2,ERBB4,LRRC4C,WWOX,IRF2,RNASEK,KCTD9,COL23A1,CSMD1,BIN3,PTPRD, NEFH,PTPN13,MTAP,PACRG,KCNU1,RAD51B,OR7G2,IKZF2,MCC,STC1,ERCC8,TENM2, HNRNPAB,JMJD8,PTEN,MPPED1,UNC5D,SGCZ,SLC25A37,ATAD1,SPAG16,BTNL8, FAM19A5,CDKN2A,WRAP53,FOCAD,RB1,CCSER1,INPP4B,LONRF1,PDE4D,PEMT
arrayMap	C9349 (plasmacytoma)	40	TPP2,KRTAP9-7,C16orf95,PRDM7,PARP4,TGDS,KIAA0513,SPATA13,FAM46C,KLF12, NOD2,ZNF705B,GSTM1,LPAR6,WWOX,SIRPB1,CSMD1,GSTM2,HEATR4,CYSLTR2,CEP128, ERICH1,FAM155A,ZNF705G,TMCO3,SHCBP1,CRYL1,PLCG2,ADAMTS18,ACOT1, LCE3B,KANSL1,PCDH9,RB1,RCBTB2,FLT3,IGLL5,HYDIN,TRAF3,USP12
arrayMap	C3058 (glioblastoma)	35	TACC2,CELF2,CDKN2B,ASCC1,ACER2,FFAR3,DHRS4L2,KIAA1598,C10orf111, MALRD1,FAM208B,CPXM2,MTAP,OR4N4,OR4X1,SLC24A2,MLLT3,SLC16A12,CTNNA3, RET,ACTA2,SH2D4B,OR52N1,IFNB1,FAM21A,ACOT1,AKR1E2,LCE3B,ATAD1,CDKN2A, FOCAD,ELAVL2,ATRNL1,HEATR4,ZRANB1
arrayMap	C80280 (DLBCL)	30	LGALS9C,CFHR3,USP17L7,AMY1A,LCE3C,EPHA7,TAS2R30,LPA,CDKN2B,TBPL1, TAS2R14,PRH1,CFHR2,PRR4,FAM106A,GSTM2,TBC1D32,MTAP,DEFA1,UGT2B15,OR4N4, UGT2B28,MRC1,TMPRSS11E,FAM21A,LCE3B,CDKN2A,EYS,IGLL5,RB1
TCGA	C4017 (ductalbreast)	85	MTHFD1L,ZNF395,TXN2,RERE,KIRREL3,SHISA6,CDC42BPB,ADAM18,WWOX,DSCAML1, AKTIP,ANKRD11,EXD2,KCNU1,KCNJ12,JAK1,DUSP4,NRG1,PINX1,KAZN,SGCZ, WDTC1,CDKN2A,EXTL3,OSGIN1,RB1,PEMT,ATP8A2,COX10,STX12,LPAR6,BIN3, SYNGR1,WSCD1,MYO1C,ST3GAL4,TRIM16,TRPV1,DNAH2,RAI1,FAM167A,BANP,VPS37A, GTF2E2,MTUS1,ARHGEF10,KIAA1456,MAP2K4,HS3ST3A1,MYOCD,SLC47A1,DLC1, PEBP4,CSMD1,ARHGAP44,AJAP1,DNAH9,ZDHHC2,FBXO16,LRRC75A,UNC5D,SERPINF1, RARA,BTNL8,RCBTB2,NCOR1,MT1A,PDE4D,RBPMS,PFKFB4,NUGGC,TTI2,IRF2, TDGF1,EFNA5,PPP2R2A,PTPRD,MTAP,RAD51B,MFHAS1,MCPH1,PUM1,PTEN,SLC7A2, EPN2
arrayMap	C3222 (medulloblastoma)	26	LCE3C,OR51A4,MTUS1,CCDC25,PTCH1,WWOX,RHD,MMP26,SIRPB1,DDHD2,RBL2, EVPLL,CSMD1,MRC1,TRIM5,FBXO16,ERICH1,OR52N1,PTEN,ZNF705G,FAM21A,LCE3B, OR51A2,CDH13,ATRNL1,HYDIN
TCGA	C3058 (glioblastoma)	59	ABCC2,CAMK1D,PTPLAD2,GSTO2,ADAM8,IFNA1,CDKN2B,ACER2,CAMTA1,COL13A1, LINGO2,CHAT,CTBP2,TEK,CDH23,UNC5B,ANKRD22,NPAS3,LPAR6,C10orf105, RSU1,KLHL9,MALRD1,ADARB2,ZNF365,ADAMTSL1,PAOX,DMRTA1,MTAP,MOB3B,IFT74, RNLS,PARD3,SLC24A2,MLLT3,STAMBPL1,RET,ACTA2,PFKFB3,PTEN,PFKP,ATAD1, PLAA,IFNA16,FOCAD,CDKN2A,PAPSS2,LIPA,MAT1A,PRKG1,ELAVL2,RCBTB2,PLG, DUSP13,LRRC20,HIST1H4B,RB1,JAKMIP3,GRID1
TCGA	C6287 (endometrium)	56	MMP15,DYNC1LI2,LRP1B,ATP10A,ZCCHC14,C16orf95,JPH3,PRR5,VAT1L,CALB2, C10orf35,EP300,NPAS3,COL4A1,MYO15A,WWOX,BPIFC,ANKS1B,DNHD1,CSMD1, TSPO,TRIM32,PGR,PTPRD,EHD4,GAN,GABBR2,SLC12A3,ST5,MARCH1,PLA2G3,ASTN2, ZFHX3,PTEN,COL15A1,PLCG2,TMEM132E,CKLF,CNTNAP4,KAZN,SGCZ,FSIP1, FAM167A,FAM19A5,PEX11G,MMP2,RB1,CCSER1,DAPK1,CDH13,BANP,POLR2M,GPC5, VAC14,PDE4D,ELF2

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

Table A.2: Significant genes in analyses

Source	NCIt	No.genes	Genes
TCGA	C2919 (prostate)	151	SCARA3,DCTN6,DIAPH3,SCARA5,CALB2,ZNF292,KLF12,CHD1,LYRM2,ENOX1,LIPJ, WWOX,CNOT7,MAN1A1,PMFBP1,RIMS1,LEPROTL1,ALOX15B,GRIN2B,WDR7,PM20D2, SLC18A1,THSD7B,NRG1,EMC8,PDGFRL,FUT9,SIM1,SGCZ,PCDH17,ASAH1,SORBS3, ITM2B,RB1,ALOX12B,RABEP1,PPAP2A,HS3ST5,SLC35F1,LCP1,CCDC70,CEP162, CHMP7,CCDC102B,PANK1,NEDD4L,C16orf46,ANKRD22,LPAR6,CHRNB1,MTSS1L, CDKN1B,IL34,NDRG4,PKD1L3,LIN28B,CAB39L,WSCD1,GATA4,MTMR7,CCAR2,LOXL2, OR1A1,FTO,RFPL4B,PLCG2,ATAD1,CMTM4,FAM167A,GABRR2,FREM2,CSGALNACT1, BANP,PIWIL2,MEI4,VPS37A,MSR1,PGM3,MTUS1,ZC3H13,ARHGEF10,ASCC3,SNX3, PROSER1,WDR59,NT5E,TRIM35,BNIP3L,DLC1,FRK,ETV6,PKD1L2,CSMD1,PEBP4, WDR16,RNLS,TOX3,TNFRSF10C,SMAD4,GRIK2,ZDHHC2,RCBTB1,FGF20,FBXO16, CYB5R4,IKBKB,CDH13,EPHA7,PDE4D,NECAB2,LONRF1,GSE1,LRP1B,ATF7IP,HTRA4, VAT1L,FNDC3A,SARAF,MTMR9,NUGGC,HACE1,TMPRSS2,PSD3,CDYL2,ERICH6B,STAR, MINPP1,PKIB,SPRYD7,TERF2IP,LRCH1,MFHAS1,PELP1,APC,C6orf163,ERG, ATP1B2,NCKAP5,ZFHX3,PTEN,CDH8,SLC7A2,STX8,PDS5B,SLC25A37,FXR2,PAPSS2, WBP4,APOLD1,TUSC3,VAC14
TCGA	C3493 (lusc)	58	FAM107A,VGLL4,ATG7,LRP1B,CADM2,DCANP1,MAML1,ARL15,CDKN2B,FOXP1, TBC1D9B,TIFAB,FRMD4B,IFNW1,EPHA3,KLF12,STK24,WWOX,LPAR6,ROBO1,DLC1, FARP1,CSMD1,PTPRD,AKR1C1,RPS12,ROBO2,RASA1,RAB3C,MTAP,MARCH1, AADACL3,AKR1C2,FHIT,MLLT3,FAM19A1,OXCT2,ZNF197,TENM2,NOTCH2,PTEN, ANKRD32,UNC5D,CACNA2D3,SBF2,ATAD1,BTNL8,FOCAD,CDKN2A,GBE1,ERC2,CCSER1, ELAVL2,ITM2B,PPARG,RB1,PDE4D,USP22
TCGA	C4035 (thyroid)	19	PDXP,TBC1D22A,RASD2,PHF21B,KCNJ11,HBB,SCUBE1,PITPNB,APOL4,SYNGR1, MICAL3,TMC1,TPST2,SULT4A1,PRR5-ARHGAP8,FAM19A5,TCF20,TAB1,SH3BP1
TCGA	C7978 (ovary)	139	SCARA3,TIAM2,SKAP1,OR51E2,NF2,SMPD3,NDEL1,SHISA6,YWHAE,WWP2,LRSAM1, WWOX,ELAVL1,ESR1,ANKRD11,PMFBP1,BTBD2,E2F4,CPD,CDK12,BRCA1,TNRC18, KCNQ1,MAP3K4,NRG1,NLRP1,MYOM2,CHEK2,SGCZ,MON1B,POLDIP3,MYH10,EPB41L2, TCF20,ITM2B,RB1,PEMT,HEXB,UHRF1,ZNF559-ZNF177,ADAP2,SLC5A4,NIPSNAP1, NUFIP2,DPP9,LPAR6,SAFB2,OMG,GNRH1,WDR18,SLC24A5,IL34,NF1,EPHX2, GNA11,TENM3,LATS1,PNPLA6,FAM118A,MCC,DNAH2,ATP6V1B2,QKI,CCSER1,SUPT6H, SLC9A5,AP1B1,USP22,IFNA4,OR51F1,PDXP,CDH1,PACSIN2,ACSL1,LSAMP, TBC1D22A,ACLY,RANBP3,MAP2K4,EP300,ZSWIM6,DBF4B,THSD4,PLIN3,ZFP90,DLC1, MIEF1,CSMD1,TTC38,PHF12,KDM4B,OR1A2,DYRK1A,DNAH9,FBXO16,SLC25A41, NAE1,RAB11FIP4,MAST4,CDH13,RCBTB2,DEPDC1B,PARK2,PDE4D,ELF2,NR3C2,AP2B1, MARVELD2,LRP1B,LRBA,EVI2B,CHRNA2,STAT3,SPON1,SPATA13,NCOA7,IRF2, KCTD9,TADA2A,ANKS1B,PPP2R2A,CDYL2,RNF135,DCLK2,CES4A,RAD51B,MAP3K1, MCPH1,CDH8,PTEN,SHCBP1,STX8,IQGAP2,RNF111,FAM19A5,FOXN1,FLCN,VAC14, SGSM2
arrayMap	C4349 (colon)	52	LGALS9C,GPR42,RBFOX1,PIK3R6,DSG3,FOXN3,VPS53,SLC5A10,CCDC25,TCF4, GOLGA8A,FLRT2,TAF4B,GSTM1,ZNF595,ZNF519,ZBTB7C,APOBEC3A,SIRPB1,CCDC11, MPPE1,CSMD1,GSTM2,KCTD1,XKR3,TSR1,UGT2B15,SERPINB5,UGT2B28,FHIT, LSM10,RSRP1,SMAD4,C18orf25,CCDC178,FBXO16,ERICH1,OR52N1,MYO5B,ELAC1, ATXN3,DOK6,SMAD2,PSTPIP2,ACOT1,MACROD2,KANSL1,CCSER1,TRAPPC8,CROCC, PARK2,PDE4D
TCGA	C4033 (renal)	118	ADAMTS9,CDCP1,SLC25A26,RPP14,CACNA1D,CCR2,SUCLG2,CAV3,FBLN2,RBMS3, IP6K2,OXNAD1,ZFYVE20,UBE2E1,NEGR1,IL17RD,ACAA1,UBP1,SLC6A1,SLC6A20, NRG1,CNTN4,ZCWPW2,ITGA9,NUP210,ERC2,ITPR1,FBXL2,METTL6,RAD18,KIF9, COLQ,ATG7,CADM2,ADCK1,ITIH1,CDKN2B,CTNNB1,MKRN2,FRMD4B,NKTR,ULK4, LMCD1,PLCL2,IL17RB,CCDC66,DOCK3,RFTN1,FLNB,EDEM1,FGD5,PRKAR2A,C3orf67, DUSP7,PHF7,PSMD6,TBC1D5,STAC,WDR82,BCL11B,PTPRG,DNAH1,FAM208A,SEC13, PBRM1,NRXN3,RPS6KA2,LRIG1,ZNF385D,SYNPR,DAZL,EPHA3,RARB,LTF,VIPR1, CADPS,CSMD1,FYCO1,TGM4,FHIT,FAM19A1,KAT2B,NEK10,CAND2,MAGI1,CACNA2D3, ATP2B2,SUMF1,TRANK1,SCN11A,FAM3D,OSBPL10,PDZRN3,SRGAP3,GADL1,NPAS3, WNT7A,TDGF1,ROBO1,ARHGEF3,SLC6A11,THRB,SFMBT1,DHX30,PTPRD,SLC22A14, FAM19A4,CLEC3B,ROBO2,SMIM4,GRM7,GRIP2,PXK,NCKIPSD,SETD2,GBE1,SNRK, IQSEC1
arrayMap	C4004 (gastric)	86	KANK1,ATP10A,KCNIP4,WNK1,WWOX,MEOX2,UGT2A2,HLA-DRA,APOBEC3A,PGR, ARAP2,DMRT1,TDRP,RAF1,ERICH1,MYOM2,PINX1,SGCZ,OR51A2,LCE3B,CDH7,LCE3C, CFHR1,CTNNB1,KDM4C,CDK6,ATP8B1,UGT2B17,GSTM1,WT1,DOCK8,GATB,GSTM2, PDXDC1,SIGLEC14,NCAPG2,RIF1,OR4N4,UGT2B28,TERT,JAK2,MGAM,SOX7,KANSL1, CCSER1,EGFR,ZRSR1,FIP1L1,ZNF385D,CSMD1,KIT,UGT2B15,FHIT,TEC,SMAD4, ZNF705G,ARL17B,SMAD2,UGT2A1,PI4KA,GMDS,PARK2,PDE4D,SLC1A1,HLA-C,HAL, ELF2,GAB1,HEATR2,KLHL14,TMPRSS2,PPP2R3B,SFMBT1,QRFPR,PTPRD,HRG, SLC2A9,APC,ERG,ETS1,GLIS3,PDGFRA,SLC25A37,PSG9,ACOT1,CTDSPL
TCGA	C3512 (luad)	109	DCC,ATP10A,MTHFD1L,SHISA6,CACNA1D,FMN1,USP43,WWOX,TSPAN16,ESD,NRG1, MYOM2,SGCZ,STK11,PRR5-ARHGAP8,FOCAD,CDKN2A,ITPR1,TMEM40,C15orf27, RB1,PEMT,RPSAP58,CNTLN,ATP8A2,COX10,CDKN2B,FOXP1,GLP2R,SCML4,NKAIN2, TFDP1,OR5AU1,RORA,LMCD1,SLC22A2,EMC7,RGCC,ADAMTSL1,RYR3,WSCD1,GJA3, GATA4,RCL1,ZNF675,SMYD4,LOXL2,BLK,ALOXE3,RFX2,FAM167A,NFIB,CSGALNACT1, PREP,MSR1,GALK2,MTUS1,RPS6KA2,PIK3R5,DHRS7C,ARHGEF10,DEFA6,MYO15A, SLC43A2,MYOCD,FBN3,DLC1,VIPR1,ETV6,CSMD1,ARID1B,KDM4B,MOB3B,ZDHHC14, MLLT3,DNAH9,MYO5B,UNC5D,SERPINF1,AVEN,PARK2,PDE4D,C16orf95,OSBPL10, GPC6,SARAF,NUGGC,SPATA13,CPB2,PSD3,AIM1,PTPRD,NTN1,MTAP,SCAMP4, SLC24A2,GRM7,MFHAS1,MCPH1,ACSBG2,SMARCA4,C8orf74,STX8,FAM19A5,OR1B1, PEX11G,FLCN,MYH1,ARHGAP8
TCGA	C3099 (liver)	126	SCARA3,ZNF395,UBXN8,SHISA6,SCARA5,ENOX1,WWOX,IPCEF1,ZMAT4,UST,LRAT, FGL1,KCNAB2,ERICH1,NRG1,MYOM2,PDGFRL,GAS7,KAZN,SGCZ,ASAH1,SMTNL2, CDKN2A,CCDC158,ZNF594,RB1,PEMT,RABEP1,SRRM1,ALB,MYH13,CDKN2B,CMIP, ATPAF2,SCML4,LPAR6,DLGAP2,TMEM229B,RGCC,ELAC2,PKD1L3,WSCD1,ABCA1, TENM3,CCAR2,PAPLN,NEIL2,TRPV1,DNAH2,BLK,MTOR,TNKS,RAI1,ANXA3,FAM167A, CCSER1,TRPV3,TOM1L2,CSGALNACT1,ENPP6,CENPV,ACSL1,MSR1,XPO7,MTUS1, IGSF21,PIK3R6,PIK3R5,RPS6KA2,CAMTA1,SCIMP,ARHGEF10,NRXN3,SOX5,HS3ST3A1, PPP2CB,SLC13A5,BNIP3L,DLC1,PPP1R14C,CSMD1,SH2D5,CYSLTR2,WDR16, TNFRSF10B,TNFRSF10C,GRIK2,TNK1,RCBTB1,WDR17,ALDH3A2,PTK2B,UNC5D,CYP4V2, CDH13,NCOR1,SORBS2,LONRF1,ELF2,LRP1B,LRBA,GPC6,SARAF,CCDC122,C8orf86, IRF2,SERPINA3,WDR27,PIGL,PSD3,SHPRH,CDCA2,MTAP,MFHAS1,MCPH1,LRCH1, LRRC48,PELP1,AFP,PTEN,SLC7A2,CNGB1,STX8,ANGPT2,ADORA2B,VAC14

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

Table A.2: Significant genes in analyses

Source	NCIt	No.genes	Genes
arrayMap	C4033 (renal)	22	GALNT15,IFNK,PRSS45,EPHA3,RARB,DOCK3,THRB,CSMD1,SFMBT1,ROBO2,GFRA2, RSRP1,PRSS50,SCAP,IRAK2,LCE3B,TRANK1,PPP4R2,PRAMEF12,CNTN6,RAD18,ELF2
arrayMap	C3493 (lusc)	163	KANK1,VGLL4,ATP10A,ACOT2,ANXA6,OR52N5,SHISA6,CUL4A,ZNF705B,WWOX, PCDHA7,CAV3,TATDN2,PCDHA13,FANCD2OS,RYBP,USP4,DDHD2,EVPLL,TSEN2, CACNA1B,ATG10,SMARCA2,HNRNPCL1,PCDHA11,ALG1L2,ERICH1,NRG1,SGCZ,OR51A2, LCE3B,CDKN2A,FOCAD,HELQ,RAD18,RB1,PCDHA5,TACC2,MTMR14,CADM2,HAVCR1, CTNNB1,FOXP1,KDM4C,SLC27A6,FRMD4B,FAM153A,BRCA2,TFDP1,ZNF717,GSTM1, LPAR6,EXD3,DOCK8,ERI1,FLNB,DOCK3,GATB,GSTM2,SHQ1,OR4N4,ZNF8,UGT2B28, MYH11,RSRP1,GSR,JAK2,BDP1,RBM6,OR52N1,PSG4,FAM21A,PCDH9,DEFA3,HEATR4, TRIM52,NAIP,PTPLAD2,MSR1,ERVV-1,MTUS1,PCDHA6,PCDHA9,EPHA3,N4BP2L2, DBN1,RARB,PCDHA2,MMP26,GHRL,BTNL3,CSMD1,CCDC36,UGT2B15,RNLS,FHIT, FAM153B,MLLT3,TRIM5,TEC,PCDHA1,RCBTB1,FBXO16,PCDHA3,ZNF705G,CAND2, MAGI1,SMAD2,BTNL8,SUMF1,PCDHA8,OR2A42,CDH13,FUT10,ZDHHC3,ELF2,PDE4D, HLA-C,C21orf59,AMY1A,USP17L7,SGCD,AP1S3,C3orf62,LRBA,RBFOX1,IFNA10, PRSS45,ACER2,PALLD,PCDHA4,CCDC25,DYNC1LI1,GOLGA8A,ATE1,ZNF595,GLRX, SFMBT1,RHOBTB3,DEFA1B,PPP2R2A,PTPRD,DDX58,DHX30,GYPB,SLC22A14,MTAP, OCIAD2,ZAR1L,SLC24A2,GRM7,MRC1,NAT1,PRSS50,CTSB,RET,TENM2,PTEN,GLIS3, PSG9,ACOT1,GPX3,PCDHA10
TCGA	C4349 (colon)	77	RNMT,DCC,SHISA6,WWOX,CELF4,GUCA2A,WDR7,CEP104,NRG1,ASB2,MYOM2,GAS7, KAZN,PPM1F,SGCZ,PRKG1,NPC1,MYH10,DTNA,RIT2,PEMT,RABEP1,MYH13,MKRN3, COX10,XKR6,ASXL3,WSCD1,TMX3,GPM6A,CCDC178,DNAH2,FAM167A,CCSER1, KIAA1328,CNDP1,MSR1,PIK3R6,DEFA6,MAP2K4,KCNIP1,DLC1,CSMD1,ALDH3A1, ZNF407,FHIT,AJAP1,SMAD4,MYO5B,LRRC75A,SMAD2,EDDM3A,SERPINF1,GMDS,PARK2, PDE4D,GABRG3,PITPNM3,CCBE1,LDLRAD4,RBFOX1,SLC47A2,DLGAP1,AGBL4,ZBTB7C, PSD3,SLC14A2,PACRG,FHOD3,MCPH1,APC,PTEN,SLC7A2,STX8,MACROD2,FAM19A5, B2M
arrayMap	C4025 (esophageal)	24	LCE3C,ETS2,WWOX,PPP2R3B,BTNL3,PTPRD,MTAP,UGT2B15,OR4N4,FHIT,DUSP22, MRC1,RSRP1,JAK2,PRB1,MGAM,ERICH1,OR52N1,ZNF705G,FAM21A,SMAD2,LCE3B, USP17L2,C21orf59
arrayMap	C7978 (ovary)	54	LGALS9C,KANK1,KRTAP9-7,OR51A4,DDX24,KRTAP9-6,MTUS1,RFPL4AL1,TEX9, OR52N5,EFCAB6,KIAA1671,SLC5A10,CCDC25,PML,EP300,GOLGA8A,ABL1,PTCH1, ZNF705B,WWOX,MMP26,SPP1,RBL2,PKD1L2,CSMD1,NF1,UGT2B15,USP17L8,ATG10, ZAR1L,TSC1,PRR23D2,TRIM5,TEC,APC,FBXO16,ERICH1,OR52N1,RAB40C,TBCK, ERBB2,ZNF705G,SGCZ,PSG9,OR51A2,ACOT1,LMNB1,KANSL1,LCE3B,CDH13,PARK2, PDE4D,ELF2
arrayMap	C2919 (prostate)	49	CCND2,USP17L7,GSTT2B,LCE3C,MSR1,THAP7,LIPK,BRCA2,ABL1,GSTM1,TMPRSS2, LPAR6,KCTD9,SIRPB1,APOBEC3A,DLC1,DDHD2,MRAP2,CDCA2,CSMD1,GSTM2, PRR23D1,EMR1,FGFR1,UCK1,MCPH1,TRIM5,LOXL2,ERICH1,OR52N1,DHX38,PTEN, FBXL17,FGFR2,SLC7A2,SNAP91,SLC25A37,LCE3B,GUCY2C,IKBKB,TMEM201, SGCZ,PI4KA,CDH13,EYS,BANP,TUSC3,RB1,ELF2

Appendix B. Clustering of enriched pathways

Appendix C. Technical information of analyses and benchmarking sets

Three independent driver sets include Bailey set [40] consisting of 299 genes from tumor exome analysis with experimental validation, the Dietlein set [41], including 461 genes from nucleotide context as well as the CGC set [36], including 724 genes from expert curation across multiple cancer types. The three gene sets share 144 consensus genes and the number of genes private to one set ranges from 73 to 481 (Figure C.10).

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

Table A.1: The number of significant genes belonging to the three cancer driving gene sets and the Fisher exact test p value for each

source	Cancer Type	No. Significant	No. in Bailey	Fisher p	No. in Dietlein	Fisher p	No. in CGC	Fisher p
TCGA	C2919-prostate	153	8	3.54E-03	15	7.38E-06	17	1.02E-04
arrayMap	C2919-prostate	49	5	9.82E-04	5	6.19E-03	8	4.42E-04
TCGA	C3058-glioblastoma	59	4	1.42E-02	4	5.5E-02	6	2.44E-02
arrayMap	C3058-glioblastoma	35	2	1.02E-01	2	2.02E-01	3	1.41E-01
TCGA	C3099-liver	127	8	1.03E-03	9	4.12E-03	12	3.58E-03
arrayMap	C3222-medulloblastoma	27	2	6.41E-02	2	1.33E-01	2	2.64E-01
TCGA	C3493-lusc	59	7	4.19E-05	6	3.08E-03	12	2.02E-06
arrayMap	C3493-lusc	168	10	3.32E-04	12	7.69E-04	16	6.06E-04
TCGA	C3512-luad	111	4	9.93E-02	6	5.4E-02	14	9.03E-05
arrayMap	C3512-luad	48	5	8.55E-04	3	1.04E-01	7	1.87E-03
TCGA	C39851-bladder	53	7	2.06E-05	6	1.78E-03	9	1.64E-04
arrayMap	C4004-gastric	86	13	1.06E-09	7	4.58E-03	18	2.98E-09
TCGA	C4017-ductalbreast	86	8	7.14E-05	6	1.85E-02	10	1.63E-03
arrayMap	C4017-ductalbreast	77	11	3.33E-08	9	8.99E-05	12	2.71E-05
arrayMap	C4025-esophageal	25	3	6.91E-03	0	1E+00	4	1.38E-02
TCGA	C4033-renal	118	5	4.17E-02	8	9.06E-03	13	6.73E-04
arrayMap	C4033-renal	23	0	1E+00	1	4.4E-01	1	6E-01
TCGA	C4035-thyroid	18	0	1E+00	0	1E+00	0	1E+00
TCGA	C4349-colon	78	6	1.6E-03	8	6.73E-04	11	1.97E-04
arrayMap	C4349-colon	54	4	9.6E-03	1	1E+00	3	4.53E-01
TCGA	C6287-endometrium	57	6	2.89E-04	4	4.98E-02	7	5.81E-03
TCGA	C7978-ovary	143	17	2.17E-10	16	5.44E-07	21	1.5E-07
arrayMap	C7978-ovary	54	7	2.08E-05	5	9.27E-03	8	8.6E-04
arrayMap	C80280-DLBCL	30	2	7.79E-02	4	5.35E-03	2	3.08E-01
arrayMap	C9349-plasmacytoma	41	5	4.5E-04	5	2.98E-03	3	2.01E-01

Table B.3: Top 10 pathways enriched by the significant genes of 25 analyses

arrayMap	C2919 (prostate)	Pathways in cancer,Prostate cancer,Drug metabolism - other enzymes,Hepatocellular carcinoma,Glutathione metabolism,PI3K-Akt signaling pathway, Fluid shear stress and atherosclerosis,Breast cancer,Central carbon metabolism in cancer,MicroRNAs in cancer
arrayMap	C3058 (glioblastoma)	Pathways in cancer,Cell cycle,Gastric cancer,Cushing's syndrome,Cellular senescence,Biosynthesis of unsaturated fatty acids,Olfactory transduction, Fatty acid elongation,Viral carcinogenesis,Thyroid cancer
arrayMap	C3222 (medulloblastoma)	FoxO signaling pathway,Cellular senescence,Olfactory transduction,Hedgehog signaling pathway,Human papillomavirus infection,Endometrial cancer, Basal cell carcinoma,Central carbon metabolism in cancer,PI3K-Akt signaling pathway,p53 signaling pathway
arrayMap	C3493 (lusc)	Hepatocellular carcinoma,Drug metabolism - other enzymes,Chemical carcinogenesis,Pathways in cancer,Glutathione metabolism, Biosynthesis of unsaturated fatty acids,Fatty acid elongation,Drug metabolism - cytochrome P450,Non-small cell lung cancer, Metabolism of xenobiotics by cytochrome P450
arrayMap	C3512 (luad)	Glutathione metabolism,Hepatocellular carcinoma,Axon guidance,Pathways in cancer,Drug metabolism - cytochrome P450, Metabolism of xenobiotics by cytochrome P450,Platinum drug resistance,Pancreatic cancer,Drug metabolism - other enzymes, Chemical carcinogenesis
arrayMap	C4004 (gastric)	Hepatocellular carcinoma,Drug metabolism - cytochrome P450,Metabolism of xenobiotics by cytochrome P450,Drug metabolism - other enzymes, Chemical carcinogenesis,Ascorbate and aldarate metabolism,Pathways in cancer,Gastric cancer,Pentose and glucuronate interconversions, Porphyrin and chlorophyll metabolism
arrayMap	C4017 (ductalbreast)	Drug metabolism - other enzymes,Drug metabolism - cytochrome P450,Metabolism of xenobiotics by cytochrome P450,Chemical carcinogenesis, Pathways in cancer,Breast cancer,Hepatocellular carcinoma,Porphyrin and chlorophyll metabolism,Prostate cancer, Kaposi's sarcoma-associated herpesvirus infection
arrayMap	C4025 (esophageal)	AGE-RAGE signaling pathway in diabetic complications,Th17 cell differentiation,Signaling pathways regulating pluripotency of stem cells,Tuberculosis, Ascorbate and aldarate metabolism,Galactose metabolism,Starch and sucrose metabolism,Endocytosis,Pentose and glucuronate interconversions, HTLV-I infection
arrayMap	C4033 (renal)	Axon guidance,Mucin type O-glycan biosynthesis,Non-small cell lung cancer,RIG-I-like receptor signaling pathway,Small cell lung cancer, Thyroid hormone signaling pathwayNeurotrophin signaling pathway,Gastric cancer,Jak-STAT signaling pathway,Tuberculosis
arrayMap	C4349 (colon)	Drug metabolism - cytochrome P450,Metabolism of xenobiotics by cytochrome P450,Drug metabolism - other enzymes,Chemical carcinogenesis, Hepatocellular carcinoma,Ascorbate and aldarate metabolism,Pentose and glucuronate interconversions,Porphyrin and chlorophyll metabolism, Apelin signaling pathway,Glutathione metabolism
arrayMap	C7978 (ovary)	MicroRNAs in cancer,Human papillomavirus infection,Pathways in cancer,Cell cycle,Endometrial cancer,Basal cell carcinoma,Adherens junction, cAMP signaling pathway,PI3K-Akt signaling pathway,EGFR tyrosine kinase inhibitor resistance
arrayMap	C80280 (DLBCL)	Drug metabolism - cytochrome P450,Metabolism of xenobiotics by cytochrome P450,Viral carcinogenesis,Drug metabolism - other enzymes, Chemical carcinogenesis,HTLV-I infection,Ascorbate and aldarate metabolism,Cell cycle,Pentose and glucuronate interconversions,Bladder cancer
arrayMap	C9349 (plasmacytoma)	Pathways in cancer,Hepatocellular carcinoma,Glutathione metabolism,Kaposi's sarcoma-associated herpesvirus infection,Epstein-Barr virus infection, Drug metabolism - cytochrome P450,Non-small cell lung cancer,Glioma,Metabolism of xenobiotics by cytochrome P450,Platinum drug resistance
TCGA	C2919 (prostate)	Prostate cancer,Inositol phosphate metabolism,Chronic myeloid leukemia,Hepatitis B,Small cell lung cancer,Gastric cancer,MicroRNAs in cancer, Hepatocellular carcinoma,Endocytosis,Glioma
TCGA	C3058 (glioblastoma)	Pathways in cancer,Hepatitis B,Central carbon metabolism in cancer,Glioma,Platinum drug resistance,Melanoma,Cellular senescence, Chronic myeloid leukemia,Hepatocellular carcinoma,Small cell lung cancer
TCGA	C3099 (liver)	Glioma,Phospholipase D signaling pathway,Endocrine resistance,Cellular senescence,Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate, Kaposi's sarcoma-associated herpesvirus infection,PI3K-Akt signaling pathway,p53 signaling pathway,Melanoma,Pancreatic cancer
TCGA	C3493 (lusc)	Small cell lung cancer,Axon guidance,Non-small cell lung cancer,Glioma,MicroRNAs in cancer,Pathways in cancer,Melanoma, Cellular senescence,Endocrine resistance,Cell cycle
TCGA	C3512 (luad)	Cellular senescence,Cushing's syndrome,Cell cycle,Oxytocin signaling pathway,cGMP-PKG signaling pathway,Hepatocellular carcinoma,Tight junction, Circadian entrainment,Bladder cancer,Cholinergic synapse
TCGA	C39851 (bladder)	Cell cycle,Glioma,Cellular senescence,Melanoma,Viral carcinogenesis,Small cell lung cancer,Prostate cancer,Bladder cancer,HTLV-I infection, FoxO signaling pathway
TCGA	C4017 (ductalbreast)	Glioma,Hepatitis B,Melanoma,Pancreatic cancer,Glycosaminoglycan biosynthesis - heparan sulfate / heparin,EGFR tyrosine kinase,inhibitor resistance, Endocrine resistance,Viral carcinogenesis,Epstein-Barr virus infection,Bladder cancer
TCGA	C4033 (renal)	Arrhythmogenic right ventricular cardiomyopathy (ARVC),Cushing's syndrome,Proteoglycans in cancer,Gastric cancer,Hypertrophic cardiomyopathy (HCM), Dilated cardiomyopathy (DCM),GABAergic synapse,Small cell lung cancer,Aldosterone synthesis and secretion,Axon guidance
TCGA	C4035 (thyroid)	Vitamin B6 metabolism,African trypanosomiasis,Type II diabetes mellitus,Malaria,Leishmaniasis,Insulin secretion, NF-kappa B signaling pathway,Toll-like receptor signaling pathway,TNF signaling pathway,Toxoplasmosis
TCGA	C4349 (colon)	Colorectal cancer,Histidine metabolism,Hepatocellular carcinoma,beta-Alanine metabolism,FoxO signaling pathway,Apelin signaling pathway, Signaling pathways regulating pluripotency of stem cells,HTLV-I infection,Hepatitis B,Wnt signaling pathway
TCGA	C6287 (endometrium)	Bladder cancer,Glioma,Pathways in cancer,Small cell lung cancer,Prostate cancer,AGE-RAGE signaling pathway in diabetic complications, HTLV-I infection,Estrogen signaling pathway,Hepatitis B,Breast cancer
TCGA	C7978 (ovary)	Hepatitis B,GnRH signaling pathway,Cell cycle,EGFR tyrosine kinase inhibitor resistance,HTLV-I infection,Hippo signaling pathway, Tight junction,Kaposi's sarcoma-associated herpesvirus infection,Viral carcinogenesis,Epstein-Barr virus infection

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

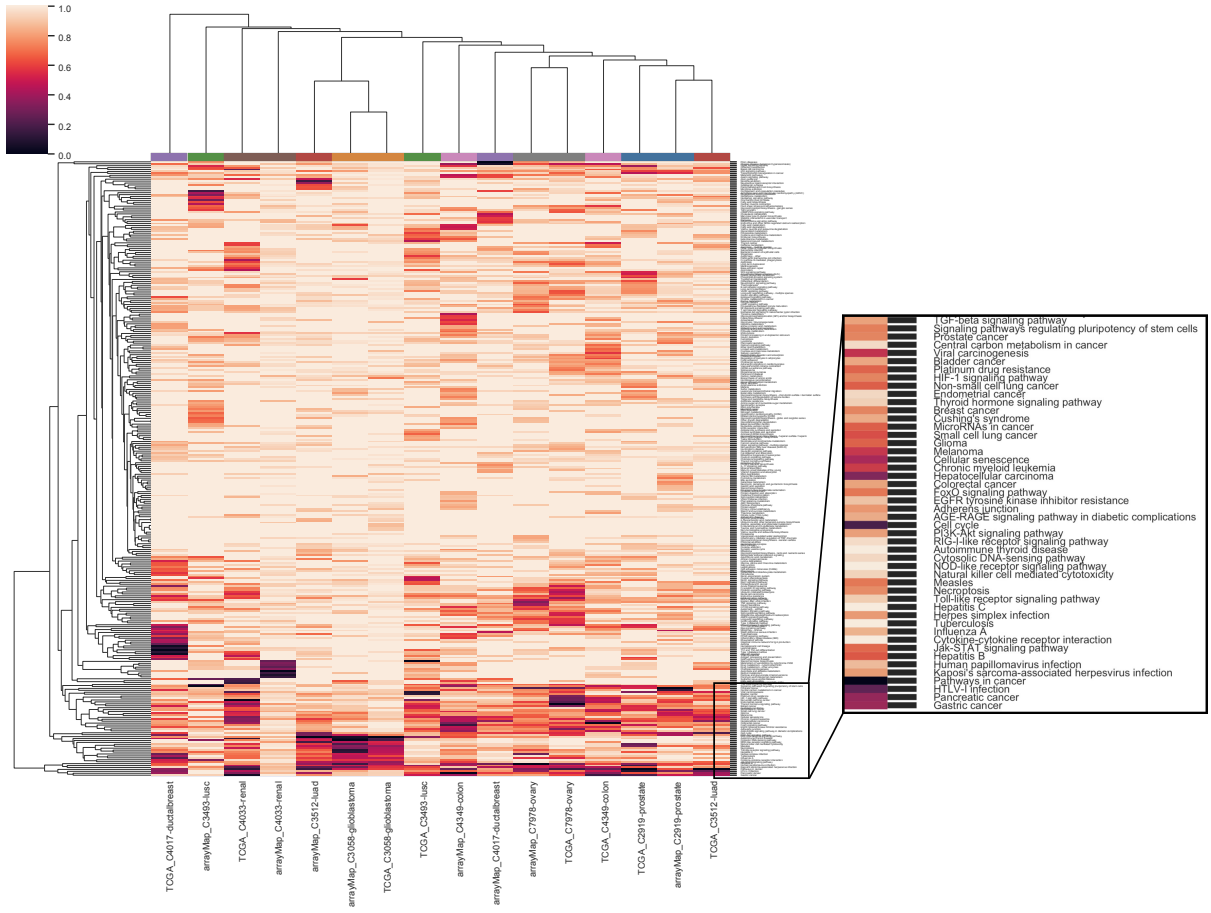

Figure B.6: **Clustering of genome-wide significance value of paired 9 cancer types with all 326 KEGG pathways**

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

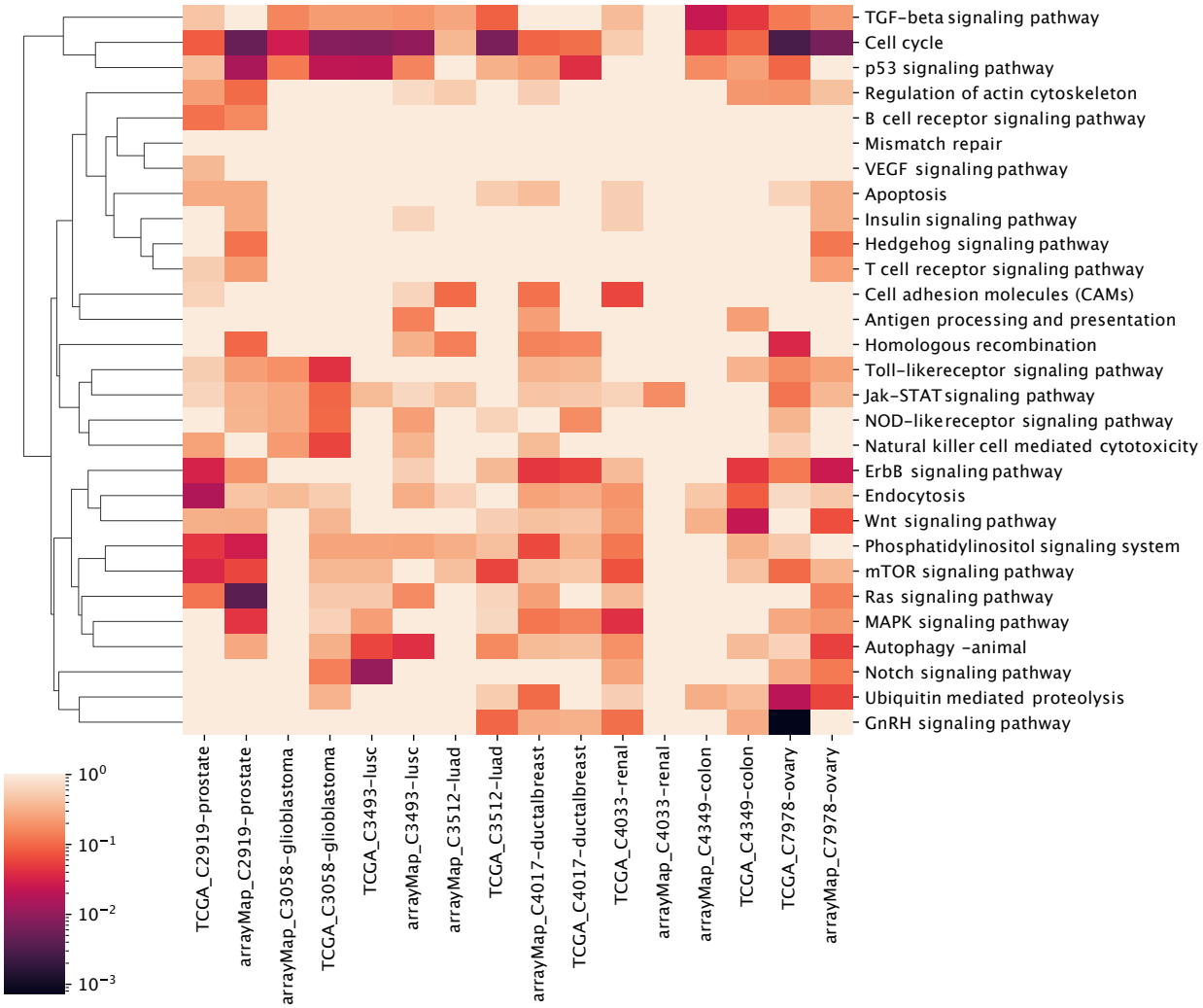

Figure B.7: **Clustering of genome-wide significance value of paired 9 cancer types with 29 cancer hallmark pathways**

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

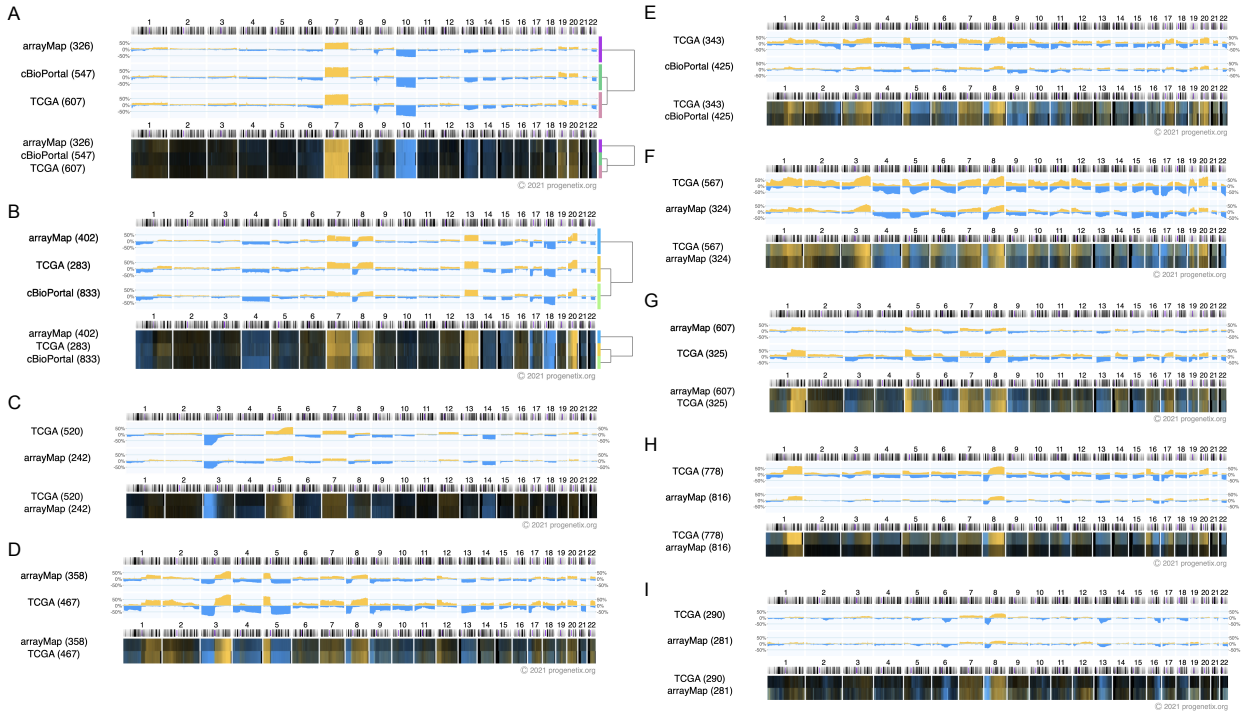

Figure C.8: **Genome-wide copy number landscape of nine cross-comparison cancer types** A) glioblastoma (C3058); B) colon adenocarcinoma (C4349); C) clear cell renal cell carcinoma (C4033); D) lung squamous carcinoma (C3493); E) bladder urothelial carcinoma (C39851); F) ovary serous cystadenocarcinoma (C7978); G) lung adenocarcinoma (C3512); H) ductal breast carcinoma (C4017); I) prostate adenocarcinoma (C2919).

bioRxiv preprint doi: https://doi.org/10.1101/2022.06.29.498080; this version posted June 30, 2022. 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.

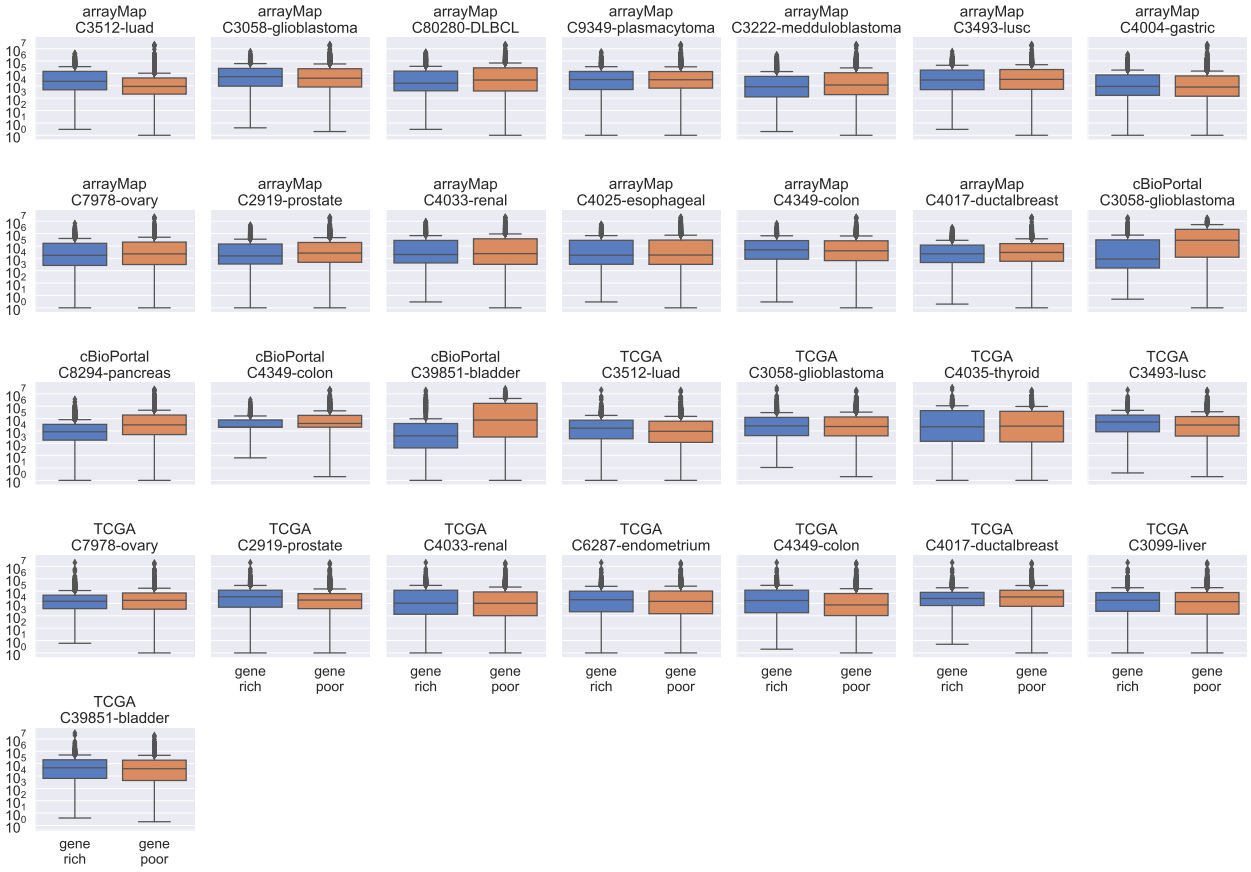

Figure C.9: **Number of segment breakpoints in gene-rich and poor regions by analysis** Most analyses have a nearly equal breakpoint density regardless of gene density but the four WES-derived (cBioPortal) analyses show clear segment sparsity in gene-poor regions.

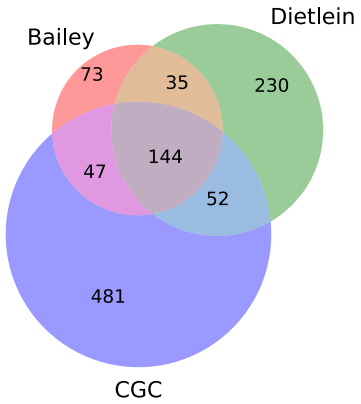

Figure C.10: **Overlap among the three cancer driving gene sets: Bailey, Dietlein and CGC**