frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 16 January 2023
DOI 10.3389/fgene.2022.1017657

OPEN ACCESS

EDITED BY
Anton A. Buzdin,
European Organisation for Research and Treatment of *Cancer*, Belgium

REVIEWED BY
Yupeng Cun,
Children's Hospital of Chongqing Medical University, China
Li (Charlie) Xia,
South China University of Technology, China

*CORRESPONDENCE
Michael Baudis,
michael.baudis@mls.uzh.ch

SPECIALTY SECTION
This article was submitted to *Cancer* Genetics and Oncogenomics, a section of the journal Frontiers in Genetics

RECEIVED 12 August 2022
ACCEPTED 28 December 2022
PUBLISHED 16 January 2023

CITATION
Huang Q and Baudis M (2023), Candidate targets of copy number deletion events across 17 cancer types.
Front. Genet. 13:1017657.
doi: 10.3389/fgene.2022.1017657

Candidate targets of copy number deletion events across 17 cancer types

Qingyao Huang[1,2] and Michael Baudis[1,2]*

[1]Department of Molecular Life Science, University of Zurich, Zurich, Switzerland, [2]Swiss Institute of Bioinformatics, Zurich, Switzerland

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 a single copy number aberration (CNA) 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 segmental 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 19 to 169 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

somatic variation, cancer genomics, meta-analysis in cancer, biomarker discovery, mutational signature, copy number aberration (CNA)

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 individual's life 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 state 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 (Vogelstein et al., 2013), 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 structure of somatic variations, they can be grouped into small-scale sequence alterations, including single nucleotide variations (SNVs), small insertions and deletions (INDELs), and structural variations, including copy number aberrations (CNAs). While the former affects isolated genetic elements, CNAs 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.

Point mutations have been reported and extensively studied for their functional impact. Affected genes can be evaluated regarding their relevance for oncogenesis through the general effect of their mutations. Briefly, cancer related genes are subdivided into two functional groups: oncogenes, of which gain-of-function (GOF) mutations promote proliferation or inhibit regulatory mechanisms and tumor suppressor genes (TSGs), of which loss of function (LOF) mutations confer a negative impact on cell cycle control and other cellular surveillance functions (Weinberg, 1994). The principal modes of action of oncogenes and TSG exhibit differing mutational characteristics. Namely, mutations for oncogenes tend to recur at the same locus; while mutations for TSGs scatter along the coding sequence (CDS) Accordingly, in Catalogue of Somatic Mutations in *Cancer* (COSMIC) database (Forbes et al., 2010), mutations are classified with a so-called "20/20 rule": to classify a gene as an oncogene, 20% of all the mutations within a gene's CDS recorded in database, need to reside at the same locus.; whereas to classify one as a TSG, 20% of recorded mutations need to be inactivating mutations but they mostly do not overlap in their location (Vogelstein et al., 2013).

In analogy to the diverging functional attributions of point mutations for oncogenes and TSGs, CNAs can be divided into amplifications and deletions. While deletion of a fraction of gene results in LOF due to truncated or untranscribed gene product, only when the entire CDS and potentially the regulatory regions outside CDS are amplified, a GOF arises.

On the mechanistic level, CN gains and losses emerge from erroneous recombination during DNA replication (Hastings et al., 2009) but present unique processes. Extrachromosomal oncogenes have been detected as copy number gain events (Turner et al., 2017; Decarvalho et al., 2018), while chromothripsis - chromosome shattering and rejoining of clustered segments - has been described as a phenomenon in cancer, which disrupts the genetic elements in the region and can result in CN deletions (Stephens et al., 2011). On the tumor evolution perspective, CN loss events tend to precede CN gain events, suggesting their different roles in oncogenesis (Watkins et al., 2020). 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 modeling the gene inactivation incorporating its unique feature of introducing segmental breakpoints within a gene's CDS.

Whereas point mutations target one particular genetic element at the specific location, a single CND potentially affects hundreds of genetic elements with a subsequent co-segregation of the affected genes. In addition, overall CNA involvement is highly correlated with the disease stage (Hieronymus et al., 2014; Shain et al., 2015; Tamborero et al., 2018; Gerstung et al., 2020), 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 (Cordo and Baudis, 2021) (Supplementary Appendix Figure S2). 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 has described amplification and deletion hotspots among multiple cancer types (Baudis, 2007; Beroukhim et al., 2010; Kumar et al., 2012; Aouiche et al., 2020). CNA-derived gene discovery can complement the knowledge of functional landscape during oncogenesis and pinpoint new genes previously unknown from point mutation analysis (Mullighan et al., 2007). In particular, an integrative multi-cancer analysis for CNA-exerted susceptibility discovery can increase the statistical power to extract disease-relevant genes and delineate their functional impact across cancer types. In recent years, work from several data curation projects and international research consortia has led to an improved availability for generally compatible, genome-wide CNA profiling data with associated information and thereby enabled the development and benchmarking of integrative approaches. Particularly, the Progenetix CNA database has gathered 115,357 samples across 788 cancer types from published studies and cohorts, including the CNA data from 11,090 patients of 182 cancer types from the *Cancer* Genome Atlas (TCGA) Project (Weinstein et al., 2013; National Cancer Institute, 2013; Huang et al., 2021).

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 (Beroukhim et al., 2010; Mermel et al., 2011). 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 gene-wise significance which would allow cross-study comparisons.

Here we describe an approach to evaluate gene-wise significance in CND which utilizes the non-random features in gene locus disruption. We verify our model by comparing the identified genes with known cancer driver gene sets as well as genes with reduced expression in the respective cancer types. Additionally, we corroborate the identified genes in terms of their biological impact with pathway analysis and cancer type clustering.

2 Results

We used data from three independent data sources depending on sample availability: 13 cancer types from the arrayMap collection, which represents a subset of the Progenetix database with available probe-specific genomic array data (Cai et al., 2015); 12 cancer types from the TCGA project processed on genome-wide SNP6 arrays (Weinstein et al., 2013); as well as four cancer types from cBioPortal database derived from whole exome sequencing (WES) experiments (Cerami et al., 2012) (Table 1). Among these, nine cancer types were represented by more than one source allowing comparison

TABLE 1 Analyzed cancer types across data sources.

	arrayMap	TCGA	cBioPortal
Bladder Urothelial Carcinoma (C39851)		y	y
Clear Cell Renal Cell Carcinoma (C4033)	y	y	
Colon Adenocarcinoma (C4349)	y	y	y
Diffuse Large B Cell Lymphoma NOS (C80280)	y		
Ductal Breast Carcinoma (C4017)	y	y	
Endometrial Endometrioid Adenocarcinoma (C6287)		y	
Esophageal Adenocarcinoma (C4025)	y		
Gastric Adendocarcinoma (C4004)	y		
Glioblastoma (C3058)	y	y	y
Hepatocellular Carcinoma (C3099)		y	
Lung Adenocarcinoma (C3512)	y	y	
Lung Squamous Cell Carcinoma (C3493)	y	y	
Medulloblastoma (C3222)	y		
Ovary Serous Cystadenocarcinoma (C7978)	y	y	
Pancreatic Adenocarcinoma (C8294)			y
Plasmacytoma (C9349)	y		
Prostate Adenocarcinoma (C2919)	y	y	
Thyroid Gland Papillary Carcinoma (C4035)		y	

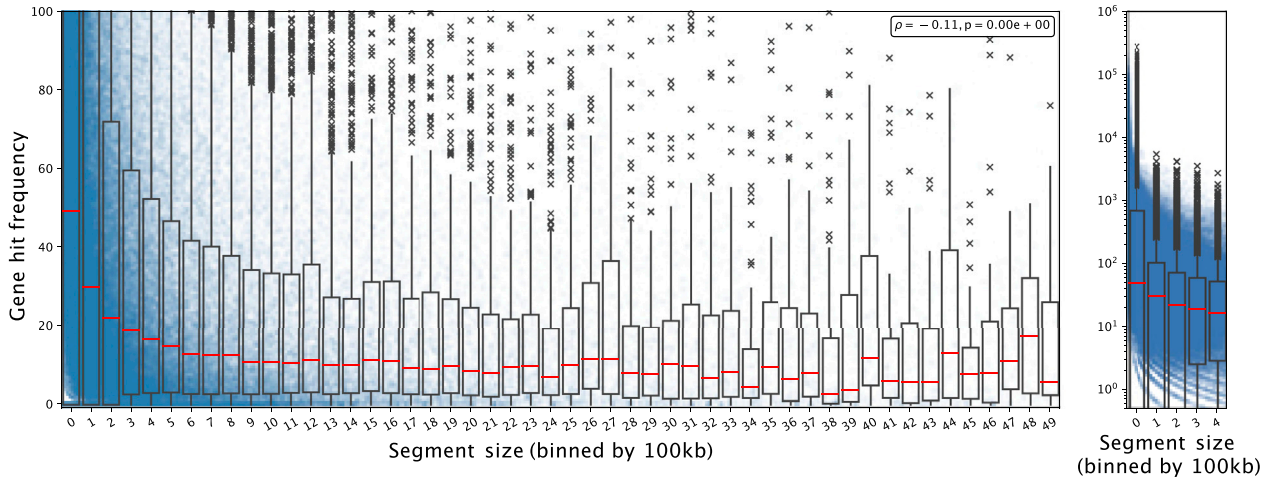

FIGURE 1
Gene hit frequency decreases as CND segment size increases Each dot shows an individual gene deletion event in a segment falling into the bin size range and the boxplot shows the summary statistics of the bin including its 25, 50 (red line; median), 75 percentiles. Left: The segment size is shown from 0 to 5MB; the gene hit frequency is limited up to 100 at a linear scale to show the decay with the bins. Right: for segment size between 0 and 500 kb with log scale on *y*-axis.

and benchmarking for source or technology related biases (Table 1). Their genome-wide CNA landscape differed among cancer types while remained comparable between data sources (Supplementary Appendix Figure S3).

2.1 Model design

We observed non-random features in CND which can be harnessed to characterize their underlying mechanisms. First, we

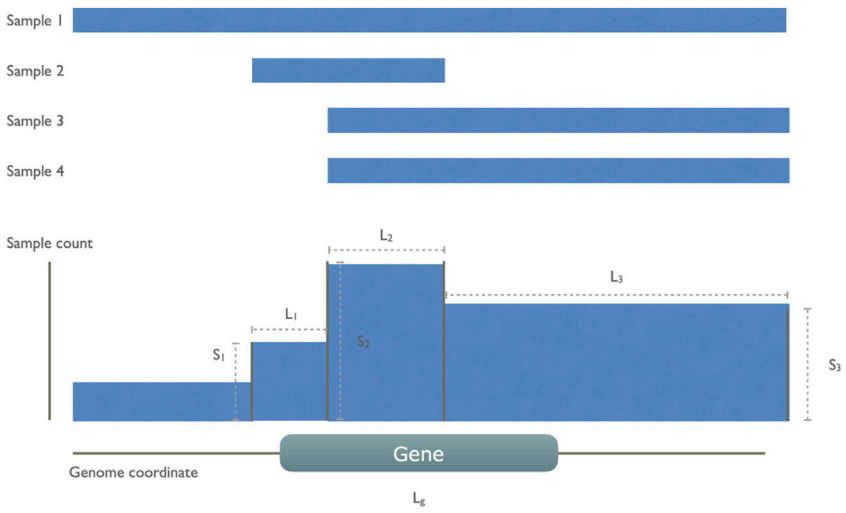

FIGURE 2
An illustration of gene score calculation The gene score was defined to reward the high number of sample recurrence and penalize the length of segments and genes. In this example, there are four cancer samples, all of which have a whole or partial deletion on the indicated gene *g*. These CN segments are collapsed to a collective track, leaving four collective segments of which only three of them overlap with the gene. The gene score sums over these three segments (*i*) with count of involved sample S_i divided by the sum of segment length L_i and the common gene length L_g.

noted that CND segment size tend to be reduced in gene-rich regions. We overlaid the CND segments from all the datasets included in the analysis one and computed the number of genes hit by the collective segment profile normalized by segment length in units of 100kb, multiplied by the number of samples containing the segment. For simplicity, we refer to this value as gene hit frequency (GHF) from here. GHF decayed with increasing CND size (Figure 1). In the segment sizes below 5 Mbp GHF decreased from 49.20 (bin 0, 0–100 kb) to 3.55 (bin 49, 4.9 Mbp - 5 Mbp) (Figure 1 Left). We noted that the individual GHF values in the first two bins' (0–200 kb) upper quantile were above the axis limit and the individual GHF values spanned five orders of magnitude. Subsequently, we plotted the GHF for the first five bins in log scale (Figure 1 Right). Visibile on both scales, the individual points formed curves of discrete gene hit number (1, 2, 3 . . .) divided by segment size (*x*-axis), while the zero-hit curve was not visible on the log-scale. We observed GHF with higher variability and higher median in the lower range of segment size and more consistently lower GHF as segment size increased, with a Spearman correlation coefficient at -0.11 and p-val 2.23×10^{-308} (python's minimal float value). Overall, the GHF decay indicated that CND favored targeting specific genes and long un-targeted CND in gene-rich regions were selected against.

Further, we showed that CND tended to recur and to locate within the genomic range of cancer-related genes. We tested the recurrence with CND segment endpoints in expert curated driver genes from COSMIC (Sondka et al., 2018) against the rest of the genome. The localization of breakpoints in driver gene sets is highly over-represented in all 29 analyses, with one-sided Fisher exact *p*-value in the range of 0.038 to $<2.23 \times 10^{-308}$. After correcting for gene length, breakpoints were still over-represented in the driver set in 27 analyses, with one-sided Fisher exact *p*-value in the range of 4.38×10^{-8} to $<2.23 \times 10^{-308}$. For Diffuse Large B Cell Lymphoma, NOS (C80280; dataset from arrayMap) the *p*-value equaled 0.51 and for medulloblastoma (C3222; arrayMap) it equaled 1.

Based on these initial observations, we then designed a model to capture these non-random CND features as illustrated in Figure 2. In each cancer type, we aggregated all CND segments and created new "collective segments" in a reference genome track. We calculated a gene score for all the genes giving weight to the sample size on the segments covering the gene while penalizing the length of segments to account for size-related unspecific deletion.

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

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

The positions of collective CND segments were shuffled within the same chromosome. The gene scores were calculated for each shuffling, to generate a background score distribution for each gene. The gene score on the real data was compared with the background distribution to calculate a empirical *p*-value to denote the gene's significance, which was subsequently adjusted with Benjamini-Yekutieli procedure. Significant genes from each analysis had the adjusted *p*-value below 0.05.

2.2 Significance across multiple cancer types

For the 29 datasets included in this study, we first 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 showed similar density, the WES-derived data from cBioPortal were biased against gene-poor regions, which causes inflation of gene significance level, making it not comparable with the array-derived data where probes are approximately equally distributed across the whole genome (Supplementary Appendix Figure S4). Therefore, the WES data-

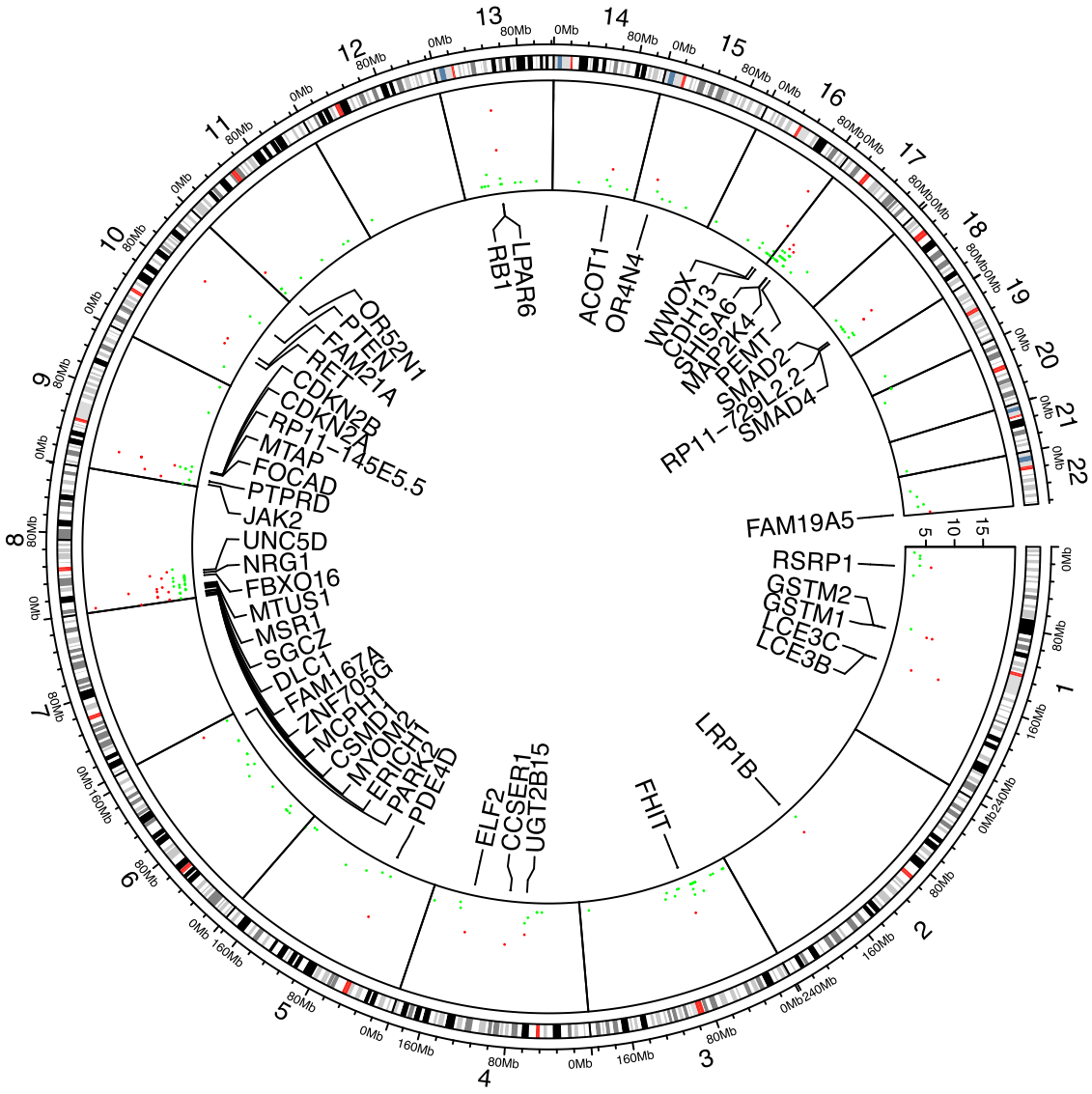

FIGURE 3
Significant genes shared among analyses. 49 genes found significant in more than five analyses are marked red and labeled accordingly.

based results were not used to derive a gene set by a significance cutoff but only for cancer type clustering with genome-wide significance *p*-values in Section 2.5.

In the 17 cancer types, the number of genes tested as significant with the segmental breakpoint model in each analysis ranged from 19 to 169 (Supplementary Appendix Table S1). We used the three cancer-driving gene sets as the "gold standard" to test enrichment of identified genes in these sets (Bailey et al., 2018; Sondka et al., 2018; Dietlein et al., 2020). 144 genes are included by all gene sets and 134 genes in at least two sets, while 784 genes are exclusively found in one set (Venn diagram in Supplementary Appendix Figure S5; Gene set details in Appendix 3). In 20 out of 25 analyses, the identifed genes were enriched in the Bailey set (16/25 for Dietlein set and 18/25 for CGC set respectively; Supplementary Appendix Table S2). Specifically, 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 3), while a few others e.g. RB1, FHIT and PDE4D, appeared as singletons.

2.3 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 a reduced expression level compared to the matched normal samples (Shao et al., 2019). We tested this hypothesis with the RNA sequencing data of paired cancer and normal samples from TCGA. For nine cancer

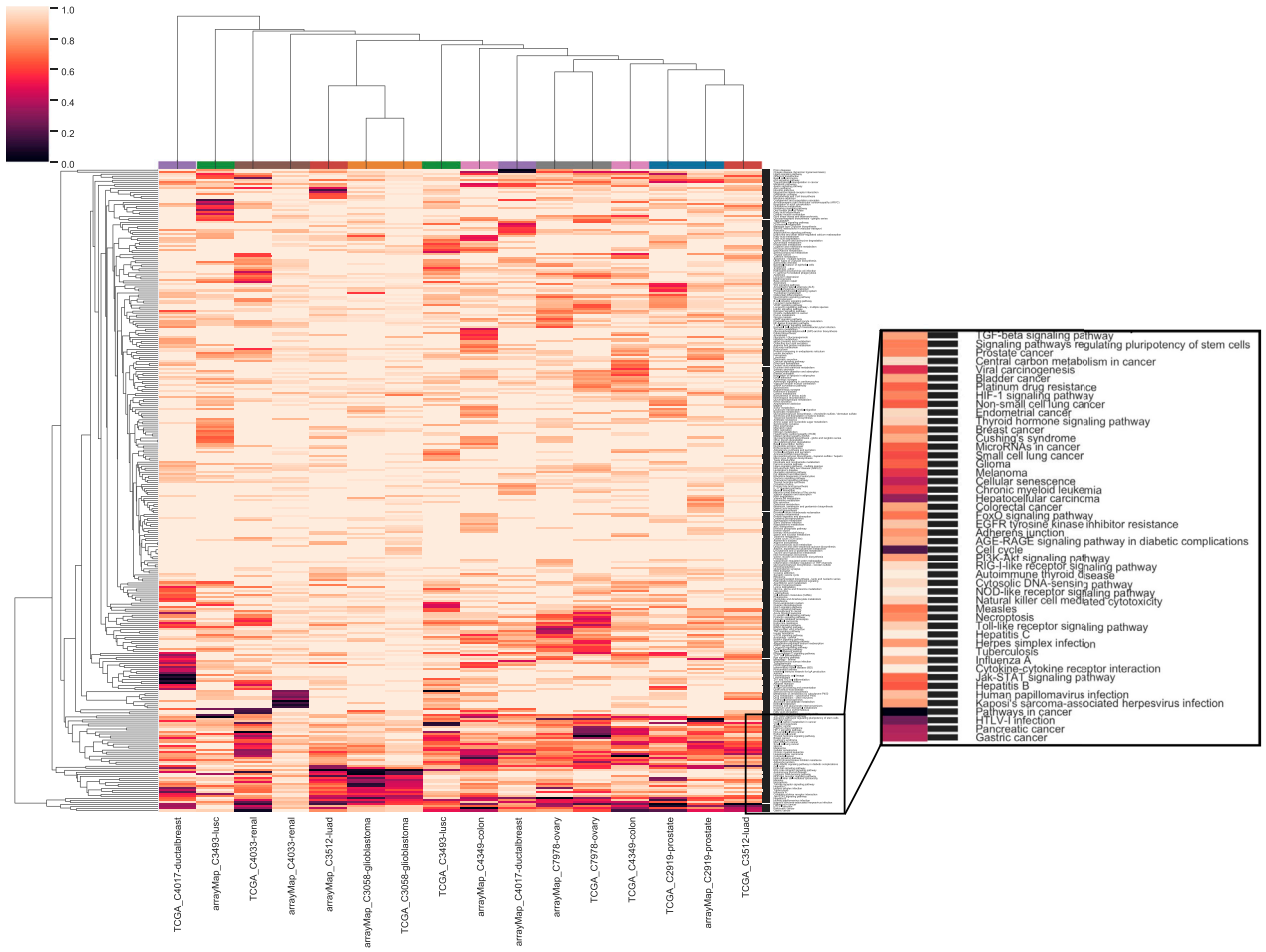

FIGURE 4
Clustering of genome-wide significance value of paired eight cancer types with all 326 KEGG pathways. The same color coding of analyses on the top of the heat map indicate the same cancer type.

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 downregulated by log fold change. Here, analyses for all the cancer types showed significant over-representation of identified CND-significant genes in the downregulated genes from RNAseq data (one-sided Fisher exact p values range from 1.6×10^{-6} to 5×10^{-21}), corroborating the LOF at the expression level.

2.4 Pathway enrichment analysis

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 4; zoomed area of 48 pathways). Among these were "TGF-beta signaling", "stem cell", "viral infection", receptor signaling pathways related to growth and apoptosis, "senescence", "energy metabolism" and "cell adhesion". With paired cancer types derived from different data sources, 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 S5 of (Vogelstein et al., 2013) 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 (Supplementary Appendix Figure S1). 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 (Hsieh and Yamane, 2008). Microsatellite instability (MSI) caused by MMR defects is mutually exclusive with the chromosomal instability related to CNA (Søreide, 2007), which is in accordance with our result. Three pathways, "TGF-beta signaling", "Cell cycle" and "p53 signaling", were enriched in a majority of cancer types. Additionally, Ras signaling pathway was enriched in prostate adenocarcinoma from arrayMap dataset and GnRH signaling

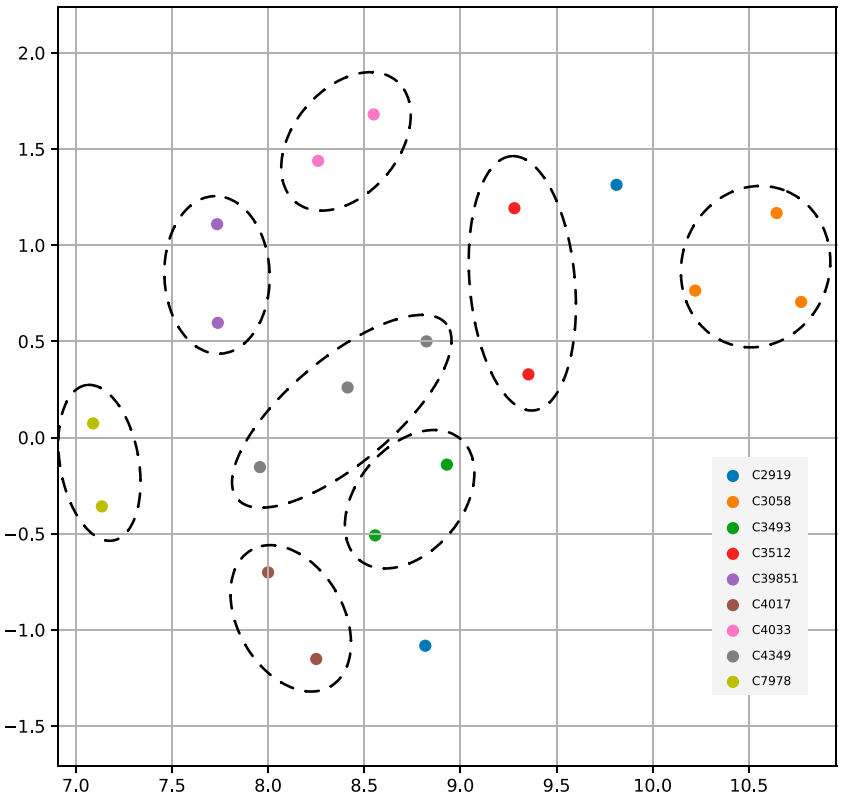

FIGURE 5
UMAP projection of genome-wide significance scores among nine paired cancer types from multiple data sources. In order of appearance the NCIt codes in the figure legend represent - C2919: Prostate Adenocarcinoma; C3058: Glioblastoma; C3493: Lung Squamous Cell Carcinoma; C3512: Lung Adenocarcinoma; C39851: Bladder Urothelial Carcinoma; C4017: Ductal Breast Carcinoma; C4033: Clear Cell Renal Cell Carcinoma; C4349: Colon Adenocarcinoma; C7978: Ovarian Serous Cystadenocarcinoma.

pathway was enriched in ovary serous carcinoma from TCGA dataset.

We listed the top 10 pathways by enrichment significance in the 25 analyses (Supplementary Table S3). "Pathways in cancer" appears in 11 analyses. Different types of viral infection appear in 12 analyses while different types of drug metabolism appear in nine analyses. "Metabolism of xenobiotics by cytochrome P450", "Cell cycle" and "Cellular senescence" appear in seven analyses. "Chemical carcinogenesis" appears in six analyses.

2.5 Cancer type clustering

So far, we used FDR cutoff to identify significant genes within an analysis. However, among the genes not passing the threshold, their *p*-values entailed information and constituted a pattern which could distinguish between cancer types. To test this hypothesis, we used the genome-wide significance scores to assess similarities between cancer types. We used uniform manifold approximation and projection (UMAP) to reduce the gene significance *p*-values to two dimensions and indicated identical cancer types from different data sources with the same color (Figure 5). All matched cancer types were represented in a neighborhood, except for prostate adenocarcinoma (C2919). The ability to differentiate between cancer types substantiated that each gene's *p*-value from the method contained additional information about the analyzed cancer type.

3 Discussion

We have proposed a data-driven method based on the frequency of gene disruption by segmental breakpoints to discern non-random CND at a gene level and generated a list of genes significantly affected by CND as well as genome-wide significance scores. 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.

Apart from providing gene-wise significance measure, this method has several advantages, including the preservation of gene neighborhood information, robustness with respect to global CNA content and variation from individual samples. First, the 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 have the same effect on the 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 variation introduced from individual samples. Individual samples' segmentation results can differ based on baseline setting - shift between adjacent CN states (Gao and Baudis, 2020), but such shifting-derived error does not substantially affect the final outcome of the method, as a single sample count has little influence on the modelled aggregation of the sample collection.

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) (Cai et al., 2014; Cortés-Ciriano et al., 2020) 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 CNA rate. Such samples should be pre-filtered with the criteria summarized in (Luijten et al., 2018) and investigated carefully on a case-by-case level.

Genes with significant scores across multiple analyses include established tumor suppressor genes that control cell cycle and regulate proliferation and programmed cell death (Marshall, 1991). RB1 is implicated in multiple processes, including cell cycle, stress responses and apoptosis (Chau and Wang, 2003). FOXO and PTEN are key regulators in phosphatidylinositol 3–kinase (PI3K) pathway which reacts to growth signals (Salmena et al., 2008; Chalhoub and Baker, 2009). CDKN2A (p16), CDKN2B(p15) at 9p21 controls S-phase checkpoint and their deletions have been reported in multiple cancers (Foulkes et al., 1997; Tsihlias et al., 1999), SMAD2/4 are involved in TGF-beta signaling pathway (Hahn et al., 1996; Nakao et al., 1997). The list also includes genes with sporadic or ambiguous oncogenic attribution. FGFR1 has been reported to have both oncogenic and tumor suppressive potential (Katoh and Nakagama, 2014) and the tyrosine kinase RET has long been established as a classic proto-oncogene but was found to act as a TSG in colorectal carcinomas (Eng, 1999; Luo et al., 2013).

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 oncogenic processes. Our results also point out genes with emerging cancer-related roles outside of classical cancer pathways such as GSTM1/GSTT1 in xenobiotic metabolism (Ginsberg et al., 2009), ELF2 as an ETS transcription factor regulating various biological pathways (Seth and Watson, 2005) or DLC1 as a Rho-GTPase activating protein regulating *RhoA* pathway in hepatocellular carcinoma (Xue et al., 2008). 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 (Smith et al., 2006).

Through the 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 (Ciriello et al., 2013).

The genomic CNA patterns as in Supplementary Appendix Figure S3 have been trained to distinguish cancer types through binned genome-wide CNA status (Baudis, 2007), but the predictive signatures to the level of chromosomal regions are not readily interpretable for biological functions. On the contrary, gene-level scores for cancer type clustering in our analysis shows clear but limited separation. This demonstrates that the cancer-type-specific CN signature is captured in the genome-wide scores. The limited distinction may be related to the vast amount of non-coding areas not included (ENCODE Project Consortium, 2012). 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 (Kong et al., 2018). 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 cluster separation.

In this article, we have developed a method to extract non-random significance from copy number deletion in cancer which exploits the specific functional implications of genomic deletion or disruption events. While due to the differences in genomic architecture and functional mechanisms of copy number gains this method is not suited to deliver a "universal CNA model" in principle the method could be adapted to discern non-random copy gain significance. Namely, cancer-promoting genes could be 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 which however should have overall results distinct from the observed CND statistics.

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 might be extended to the non-coding area and provide a better overview of the CNA functional landscape.

4 Experimental procedures

4.1 Data availability

CNA data has been accessed from three different sources - arrayMap, TCGA, cBioPortal - which had been integrated into the Progenetix database (Table 1) (Cerami et al., 2012; Weinstein et al., 2013; Cai et al., 2015; Huang et al., 2021). CN data and curated

biosample metadata are freely accessible through progenetix.org over the GA4GH Beacon protocol in JSON format compatible to the Beacon v2 data model as well as tab-delimited text file format (Wagner et al., 2021; Jacobsen et al., 2022; Rambla et al., 2022).

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 Differential expression analysis

We used R-package EdgeR for differential expression analysis between tumor and normal groups (Robinson et al., 2010). 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.

4.3 Pathway analysis

All 326 KEGG pathways (Kanehisa et al., 2021) were used to determine the enrichment in each pathway with a one-sided 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.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

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.

Acknowledgments

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.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1017657/full#supplementary-material

References

Aouiche, C., Chen, B., and Shang, X. (2020). Predicting stage-specific recurrent aberrations from somatic copy number dataset. *Front. Genet.* 11, 160. doi:10.3389/fgene.2020.00160

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell* 173 (2), 1034–1035. doi:10.1016/j.cell.2018.07.034

Baudis, M. (2007). Genomic imbalances in 5918 malignant epithelial tumors: An explorative meta-analysis of chromosomal cgh data. *BMC cancer* 7 (1), 226–315. doi:10.1186/1471-2407-7-226

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., et al. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature* 463 (7283), 899–905. doi:10.1038/nature08822

Cai, H., Gupta, S., Rath, P., Ai, N., and Baudis, M. (2015). arrayMap 2014: an updated cancer genome resource. *Nucleic acids Res.* 43 (D1), D825–D830. doi:10.1093/nar/gku1123

Cai, H., Kumar, N., Bagheri, H. C., von Mering, C., Robinson, M. D., and Baudis, M. (2014). Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22, 347 cancer genome screens. *BMC genomics* 15 (1), 82–13. doi:10.1186/1471-2164-15-82

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2 (5), 401–404. doi:10.1158/2159-8290.CD-12-0095

Chalhoub, N., and Baker, S. J. (2009). Pten and the pi3-kinase pathway in cancer. *Annu. Rev. Pathology Mech. Dis.* 4, 127–150. doi:10.1146/annurev.pathol.4.110807.092311

Chau, B. N., and Wang, J. Y. (2003). Coordinated regulation of life and death by rb. *Nat. Rev. Cancer* 3 (2), 130–138. doi:10.1038/nrc993

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* 45 (10), 1127–1133. doi:10.1038/ng.2762

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

Cortés-Ciriano, I., Lee, J. J.-K., Xi, R., Jain, D., Jung, Y. L., Yang, L., et al. (2020). Comprehensive analysis of chromothripsis in 2, 658 human cancers using whole-genome sequencing. *Nat. Genet.* 52 (3), 331–341. doi:10.1038/s41588-019-0576-7

Decarvalho, A. C., Kim, H., Poisson, L. M., Winn, M. E., Mueller, C., Cherba, D., et al. (2018). Discordant inheritance of chromosomal and extrachromosomal dna elements contributes to dynamic disease evolution in glioblastoma. *Nat. Genet.* 50 (5), 708–717. doi:10.1038/s41588-018-0105-0

Dietlein, F., Weghorn, D., Taylor-Weiner, A., Richters, A., Reardon, B., Liu, D., et al. (2020). Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* 52 (2), 208–218. doi:10.1038/s41588-019-0572-y

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489 (7414), 57–74. doi:10.1038/nature11247

Eng, C. (1999). Ret proto-oncogene in the development of human cancer. *J. Clin. Oncol.* 17 (1), 380–393. doi:10.1200/JCO.1999.17.1.380

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., et al. (2010). Cosmic: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic acids Res.* 39 (1), D945–D950. doi:10.1093/nar/gkq929

Foulkes, W. D., Flanders, T. Y., Pollock, P. M., and Hayward, N. K. (1997). The cdkn2a (p16) gene and human cancer. *Mol. Med.* 3 (1), 5–20. doi:10.1007/bf03401664

Gao, B., and Baudis, M. (2020). Minimum error calibration and normalization for genomic copy number analysis. *Genomics* 112 (5), 3331–3341. doi:10.1016/j.ygeno.2020.05.008

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S. C., Gonzalez, S., Rosebrock, D., et al. (2020). The evolutionary history of 2, 658 cancers. *Nature* 578 (7793), 122–128. doi:10.1038/s41586-019-1907-7

Ginsberg, G., Smolenski, S., Neafsey, P., Hattis, D., Walker, K., Guyton, K. Z., et al. (2009). The influence of genetic polymorphisms on population variability in six xenobiotic-metabolizing enzymes. *J. Toxicol. Environ. Health, Part B* 12 (5-6), 307–333. doi:10.1080/10937400903158318

Hahn, S. A., Schutte, M., Hoque, A. S., Moskaluk, C. A., Da Costa, L. T., Rozenblum, E., et al. (1996). Dpc4, a candidate tumor suppressor gene at human chromosome 18q21. 1. *Science* 271 (5247), 350–353. doi:10.1126/science.271.5247.350

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., and Ira, G. (2009). Mechanisms of change in gene copy number. *Nat. Rev. Genet.* 10 (8), 551–564. doi:10.1038/nrg2593

Hieronymus, H., Schultz, N., Gopalan, A., Carver, B. S., Chang, M. T., Xiao, Y., et al. (2014). Copy number alteration burden predicts prostate cancer relapse. *Proc. Natl. Acad. Sci.* 111 (30), 11139–11144. doi:10.1073/pnas.1411446111

Hsieh, P., and Yamane, K. (2008). Dna mismatch repair: Molecular mechanism, cancer, and ageing. *Mech. ageing Dev.* 129 (7-8), 391–407. doi:10.1016/j.mad.2008.02.012

Huang, Q., Carrio-Cordo, P., Gao, B., Paloots, R., and Baudis, M. (2021). The progenetix oncogenomic resource in 2021, database: The journal of biological databases and curation. *Database (Oxford)* 2021, baab043. doi:10.1093/database/baab043

Jacobsen, J. O., Baudis, M., Baynam, G. S., Beckmann, J. S., Beltran, S., Buske, O. J., et al. (2022). The ga4gh phenopacket schema defines a computable representation of clinical data. *Nat. Biotechnol.* 40 (6), 817–820. doi:10.1038/s41587-022-01357-4

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M. (2021). Kegg: Integrating viruses and cellular organisms. *Nucleic acids Res.* 49 (D1), D545–D551. doi:10.1093/nar/gkaa970

Katoh, M., and Nakagama, H. (2014). Fgf receptors: Cancer biology and therapeutics. *Med. Res. Rev.* 34 (2), 280–300. doi:10.1002/med.21288

Kong, Y., Hsieh, C.-H., and Alonso, L. C. (2018). Anril: A lncrna at the cdkn2a/b locus with roles in cancer and metabolic disease. *Front. Endocrinol.* 9, 405. doi:10.3389/fendo.2018.00405

Kumar, N., Cai, H., Von Mering, C., and Baudis, M. (2012). Specific genomic regions are differentially affected by copy number alterations across distinct cancer types, in aggregated cytogenetic data. *PLoS One* 7, e43689. doi:10.1371/journal.pone.0043689

Luijten, M. N. H., Lee, J. X. T., and Crasta, K. C. (2018). Mutational game changer: Chromothripsis and its emerging relevance to cancer. *Mutat. Research/Reviews Mutat. Res.* 777, 29–51. doi:10.1016/j.mrrev.2018.06.004

Luo, Y., Tsuchiya, K. D., Il Park, D., Fausel, R., Kanngurn, S., Welcsh, P., et al. (2013). Ret is a potential tumor suppressor gene in colorectal cancer. *Oncogene* 32 (16), 2037–2047. doi:10.1038/onc.2012.225

Marshall, C. J. (1991). Tumor suppressor genes. *Cell* 64 (2), 313–326. doi:10.1016/0092-8674(91)90641-b

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12 (4), R41–R14. doi:10.1186/gb-2011-12-4-r41

Mullighan, C. G., Goorha, S., Radtke, I., Miller, C. B., Coustan-Smith, E., Dalton, J. D., et al. (2007). Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* 446 (7137), 758–764. doi:10.1038/nature05690

Nakao, A., Imamura, T., Souchelnytskyi, S., Kawabata, M., Ishisaki, A., Oeda, E., et al. (1997). Tgf-β receptor-mediated signalling through smad2, smad3 and smad4. *EMBO J.* 16 (17), 5353–5362. doi:10.1093/emboj/16.17.5353

National Cancer Institute (2013). The cancer genome atlas program. Available at: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (Accessed 20 May, 2021).

Rambla, J., Baudis, M., Ariosa, R., Beck, T., Fromont, L., Navarro, A., et al. (2022). Beacon v2 and beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond. *Hum. Mutat.* 43 (6), 791–799. doi:10.1002/humu.24369

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26 (1), 139–140. doi:10.1093/bioinformatics/btp616

Salmena, L., Carracedo, A., and Pandolfi, P. P. (2008). Tenets of pten tumor suppression. *Cell* 133 (3), 403–414. doi:10.1016/j.cell.2008.04.013

Seth, A., and Watson, D. K. (2005). Ets transcription factors and their emerging roles in human cancer. *Eur. J. cancer* 41 (16), 2462–2478. doi:10.1016/j.ejca.2005.08.013

Shain, A. H., Yeh, I., Kovalyshyn, I., Sriharan, A., Talevich, E., Gagnon, A., et al. (2015). The genetic evolution of melanoma from precursor lesions. *N. Engl. J. Med.* 373 (20), 1926–1936. doi:10.1056/NEJMoa1502583

Shao, X., Lv, N., Liao, J., Long, J., Xue, R., Ai, N., et al. (2019). Copy number variation is highly correlated with differential gene expression: A pan-cancer study. *BMC Med. Genet.* 20 (1), 175–214. doi:10.1186/s12881-019-0909-5

Smith, D. I., Zhu, Y., McAvoy, S., and Kuhn, R. (2006). Common fragile sites, extremely large genes, neural development and cancer. *Cancer Lett.* 232 (1), 48–57. doi:10.1016/j.canlet.2005.06.049

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. (2018). The cosmic cancer gene census: Describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* 18 (11), 696–705. doi:10.1038/s41568-018-0060-1

Søreide, K. (2007). Molecular testing for microsatellite instability and dna mismatch repair defects in hereditary and sporadic colorectal cancers–ready for prime time? *Tumor Biol.* 28 (5), 290–300. doi:10.1159/000110427

Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 144 (1), 27–40. doi:10.1016/j.cell.2010.11.055

Tamborero, D., Rubio-Perez, C., Muiños, F., Sabarinathan, R., Piulats, J. M., Muntasell, A., et al. (2018). A pan-cancer landscape of interactions between solid tumors and infiltrating immune cell populations. *Clin. Cancer Res.* 24 (15), 3717–3728. doi:10.1158/1078-0432.CCR-17-3509

Tsihlias, J., Kapusta, L., and Slingerland, J. (1999). The prognostic significance of altered cyclin-dependent kinase inhibitors in human cancer. *Annu. Rev. Med.* 50 (1), 401–423. doi:10.1146/annurev.med.50.1.401

Turner, K. M., Deshpande, V., Beyter, D., Koga, T., Rusert, J., Lee, C., et al. (2017). Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* 543 (7643), 122–125. doi:10.1038/nature21356

Vogelstein, B., Papadopoulos, N., Velculescu, V., Zhou, S., Diaz, L., and Kinzler, K. (2013). Cancer genome landscapes. *Science* 339 (6127), 1546â–1558. doi:10.1126/science.1235122

Wagner, A. H., Babb, L., Alterovitz, G., Baudis, M., Brush, M., Cameron, D. L., et al. (2021). The ga4gh variation representation specification: A computational framework for variation representation and federated identification. *Cell genomics* 1 (2), 100027. doi:10.1016/j.xgen.2021.100027

Watkins, T. B., Lim, E. L., Petkovic, M., Elizalde, S., Birkbak, N. J., Wilson, G. A., et al. (2020). Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature* 587 (7832), 126–132. doi:10.1038/s41586-020-2698-6

Weinberg, R. A. (1994). Oncogenes and tumor suppressor genes. *A Cancer J. Clin.* 44 (3), 160–170. doi:10.3322/canjclin.44.3.160

Weinstein, J., Collisson, E., Mills, G., Shaw, K., Ozenberger, B., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45 (10), 1113–1120. doi:10.1038/ng.2764

Xue, W., Krasnitz, A., Lucito, R., Sordella, R., VanAelst, L., Cordon-Cardo, C., et al. (2008). Dlc1 is a chromosome 8p tumor suppressor whose loss promotes hepatocellular carcinoma. *Genes & Dev.* 22 (11), 1439–1444. doi:10.1101/gad.1672608