Extending differential gene expression testing to handle genome aneuploidy in cancer

Katsiaryna Davydzenka (1,2), Giulio Caravagna (3, +), Guido Sanguinetti (1, +)

(1) SISSA, Italy.
(2) University Campus Bio-Medico of Rome, Italy.
(3) Department of Mathematics, Informatics and Geosciences, University of Trieste, Italy

(+) Correspondence: gcaravagna@units.it and gsanguin@sissa.it

Abstract

Genome aneuploidy, characterized by Copy Number Variations (CNVs), profoundly alters gene expression in cancer. CNVs can directly influence transcription through gene dosage effects or indirectly through compensatory regulatory mechanisms. However, existing differential gene expression (DGE) testing methods do not differentiate between these mechanisms, conflating all expression changes and limiting biological interpretability. This misclassification can obscure key genes involved in tumor adaptation and progression, hindering biomarker discovery and leading to incomplete insights into cancer biology.
To address this, we developed DeConveil, a computational framework that extends traditional DGE analysis by integrating CNV data. Using a Generalized Linear Model (GLM) with a Negative Binomial (NB) distribution, DeConveil models RNA-seq expression counts while accounting for copy number (CN) gene dosage effects. We proposed a more fine-grained gene decomposition into dosage-sensitive (DSGs), dosage-insensitive (DIGs), and dosage-compensated (DCGs), which explicitly de-couples changes due to CNAs and bona fide changes in transcriptional regulation.
Analysis of TCGA datasets from aneuploid solid cancers resulted in notable reclassification of genes, refining and expanding upon the results from conventional methods. Functional enrichment analysis identified distinct biological roles for DSGs, DIGs, and DCGs in tumor progression, immune regulation, and cell adhesion. In a breast cancer case study, DeConveil's CN-aware analysis facilitated the identification of both known and novel prognostic biomarkers, including lncRNAs, linking gene expression signatures to survival outcomes. Utilizing these biomarkers for each gene group significantly improved patient risk stratification, yielding more accurate predictions compared to conventional methods.
These results highlight DeConveil's ability to disentangle CNV-driven from regulatory transcriptional changes, enhancing gene classification and biomarker discovery. By improving transcriptomic analysis, DeConveil provides a powerful tool for cancer research, precision oncology, with potential applications in therapeutic target identification.

Keywords: Differential Gene Expression, Copy Number Variation, Cancer, Generalized Linear Model, Negative Binomial distribution

Introduction

Cancer is a highly heterogeneous disease characterized by extensive genomic alterations with DNA Copy Number Variations (CNVs) and genome aneuploidy emerging as defining hallmarks across most tumor types [1–3]. CNVs represent structural changes, such as gains or losses of specific chromosomal segments, which profoundly reshape the transcriptional landscape of cancer cells [4–6]. These changes can create gene dosage effects [7], amplifying or reducing mRNA transcript levels for genes within the affected regions [7,8]. Such disruptions have profound consequences for tumor progression, driving tumorigenesis [9], facilitating metastasis [9,10], and contributing to therapy resistance [11].

However, the relationship between somatic CNVs and gene expression is complex [4,12–19]. While some genes in altered regions exhibit expression changes that correlate with CNVs, such as oncogenes in amplified regions or tumor suppressors in deleted regions [12], many others exhibit only moderate or no expression changes [12–15], suggesting the involvement of additional regulatory mechanisms. For example, Zhou et al. (2017) [13] identified CNV-driven differentially expressed genes in hepatocellular carcinoma, highlighting CNVs as key drivers of transcriptional dysregulation. Mohanty et al. (2021) [19] further emphasize that CNVs alone do not dictate expression changes in aneuploid cancers; instead, gene-specific regulatory dynamics and compensatory mechanisms can modulate these effects. This complexity underscores the need for advanced statistical approaches to distinguish CNV-driven expression changes from independent regulatory alterations.

Differential gene expression (DGE) analysis remains fundamental for studying transcriptomic alterations in cancer, identifying key oncogenic pathways, therapeutic targets, and biomarkers [20,21]. Widely used statistical tools for DGE analysis, such as DESeq2 [22], edgeR [23], and limma [24] employ statistical models that effectively handle RNA-seq count data, assuming gene expression changes arise solely from biological or experimental factors. However, these methods do not account for CNVs, implicitly assuming that the genes have the same CNVs across samples (or no CNVs) overall. This assumption is problematic in cancer, where aneuploidy introduces widespread CNVs that can drive gene expression changes. A key limitation is the inability to determine whether observed expression changes result from CNVs or other regulatory mechanisms. This can create significant challenges for interpreting DGE results in cancer studies, potentially obscuring key biological insights and misleading conclusions. While some progress has been made in integrating CNV data into other genomic contexts, such as DNA methylation, similar advancements in transcriptomics remain limited. For example, the ABCD-DNA tool integrates CNV data to enhance the analysis of DNA methylation [25]. However, a parallel framework for transcriptomics is still lacking.

To address these limitations, we developed DeConveil, a computational framework that explicitly integrates CNV effects into DGE analysis. DeConveil extends traditional statistical models by incorporating CNV data using a Generalized Linear Model (GLM) with a Negative Binomial (NB) distribution, allowing for a more nuanced interpretation of gene expression changes. This approach refines gene classification by distinguishing genes whose expression is primarily driven by CNVs from those regulated through other biological mechanisms.

Application of DeConveil to aneuploid cancer datasets demonstrates its broad utility and capacity to uncover shared and specific mechanisms across cancers. In a case study on breast cancer, DeConveil provided a more refined categorization of genes based on their relationship to gene CNVs, including novel long non-coding RNAs (lncRNAs). This refined classification not only facilitated the identification of potential prognostic genes but also provided a deeper understanding of their biological roles and regulatory mechanisms underlying each gene category.

Methods

Model description

The primary goal of DeConveil is to account for CNVs when analyzing gene expression differences between contrasted conditions or sample groups. Our strategy leverages on a generalized linear model (GLM), commonly adopted in RNA-seq differential expression testing [22,23].
Bulk RNA-seq experiments generate a matrix of read counts $\mathbf{Y} \in \mathbb{R}^{N \times G}$ that reflect the abundance of G transcripts detected across N samples. For each gene g, the corresponding count vector y_g is modeled using a Negative Binomial (NB) distribution [26,27] to account for overdispersion:

$$\mathbf{y_g} \sim \mathrm{NB}(\mu_g, \theta_g) \qquad (1)$$

where $\mathbf{y_g}$ represents the observed read counts for gene *g* across samples, μ_g is the expected gene-specific mean expression, θ_g captures variability in expression.

Therefore, the likelihood L of observed read counts for gene g is expressed as

$$\text{Ł}(y_g \mid \mu_g, \theta_g) = \frac{\Gamma(y_g + 1/\theta_g)}{\Gamma(y_g + 1)\Gamma(1/\theta_g)} \left(\frac{1}{1 + \mu_g\theta_g}\right)^{1/\theta_g} \left(\frac{\mu_g\theta_g}{1 + \mu_g\theta_g}\right)^{y_g} \qquad (2)$$

To model gene expression as a function of covariates, we use a GLM with logarithmic link:

$$\log \mathbf{q}_{g,n} = \sum_f \mathbf{x}_{n,f}\boldsymbol{\beta}_{g,f} \qquad (3)$$

where $\mathbf{q}_{g,n}$ is the expected expression level, $\mathbf{x}_{n,f}$ represents covariates (e.g., tumor/normal condition), $\boldsymbol{\beta}_{g,f}$ are the regression coefficients to be estimated.
We assume that DNA copy number measurements directly influence RNA-seq read counts (Fig. 1A, Fig. S1). For example, a CN of 4 could result in doubled expression compared to a diploid control. Therefore, we modify the baseline GLM to incorporate CNV effects

$$\boldsymbol{\mu}_{g,n} = \mathbf{s}_n \left(\frac{CN_{g,n}}{2}\right) \{\exp(\mathbf{X}_{n,f}\beta_{g,f})\} \qquad (4)$$

where $\mathbf{s}_n$ is a sample-specific normalization factor calculated using median-of-ratios method used in DESeq2 [22], while $\frac{CN_{g,n}}{2}$ represents a gene and sample specific vector of CN dosage scaling factors. The division by 2 likely normalizes CN values relative to the diploid state (where CN = 2 is the reference). If $CN_{g,n} = 2$, this term becomes 1, meaning the expression is unaltered; if $CN_{g,n} > 2$ (amplifications), this term scales expression up, if $CN_{g,n} < 2$ (deletions), this term scales expression down. These components adjust the expected mean to account for systematic variables, including CN and sequencing depth.
DeConveil requires three input matrices (Fig.1B): matched mRNA read counts $Y \in \mathbb{R}^{N \times G}$ and absolute copy number values $CN \in \mathbb{R}^{N \times G}$ (for normal diploid samples we assign CN = 2) with rows corresponding to genes and columns to samples, and design matrix $X \in \mathbb{R}^{N \times F}$ encoding sample conditions.
The design matrix $\mathbf{X}$ is structured as follows (Figure 1B): rows correspond to individual samples, and each column represents a feature. In its simplest form, $\mathbf{X}$ consists of: intercept column (constant 1 for all samples,

modeling baseline expression), condition column (binary indicator: 0 = normal, 1 = tumor). For a dataset with n samples and f covariates, is an $n \times f$ matrix. The model learns an $\mathbf{f}$-dimensional coefficient vector, where f is the number of covariates.

DeConveil fits GLM for each gene and employs an empirical Bayes approach, similar to DESeq2 [22,28]. Initially, the maximum likelihood estimation (MLE) is used to learn $\boldsymbol{\theta}_g$ and $\boldsymbol{\beta}_g$ by maximizing log-likelihood of NB distribution. Subsequently, maximum a posteriori (MAP) estimator, namely approximate posterior estimation for GLM (apeglm) [29] method is used to apply shrinkage to both coefficients. Log2 fold change (log2FC) calculation is derived from estimated coefficients $\boldsymbol{\beta}_g$.

The regression coefficients $\boldsymbol{\beta}_g$ for each gene are analyzed using the Wald test [30]. Wald test was applied to test for the statistical significance (p-value) in observed expression differences between tumor and normal sample groups. We evaluate the null hypothesis $H_0 : \mathbf{c}\beta_g = 0$, where c is an $\mathbf{f}$-dimensional contrast vector, that selects specific linear combinations of coefficients to test for DE. When testing multiple genes simultaneously, we adjust the resulting p-values with the Benjamini-Hochberg method [31] to reduce false positives that arise from multiple comparisons. Therefore, the final results provide the log2FC calculations (magnitude of DE) and p-value (statistical significance), which are used for gene classification.

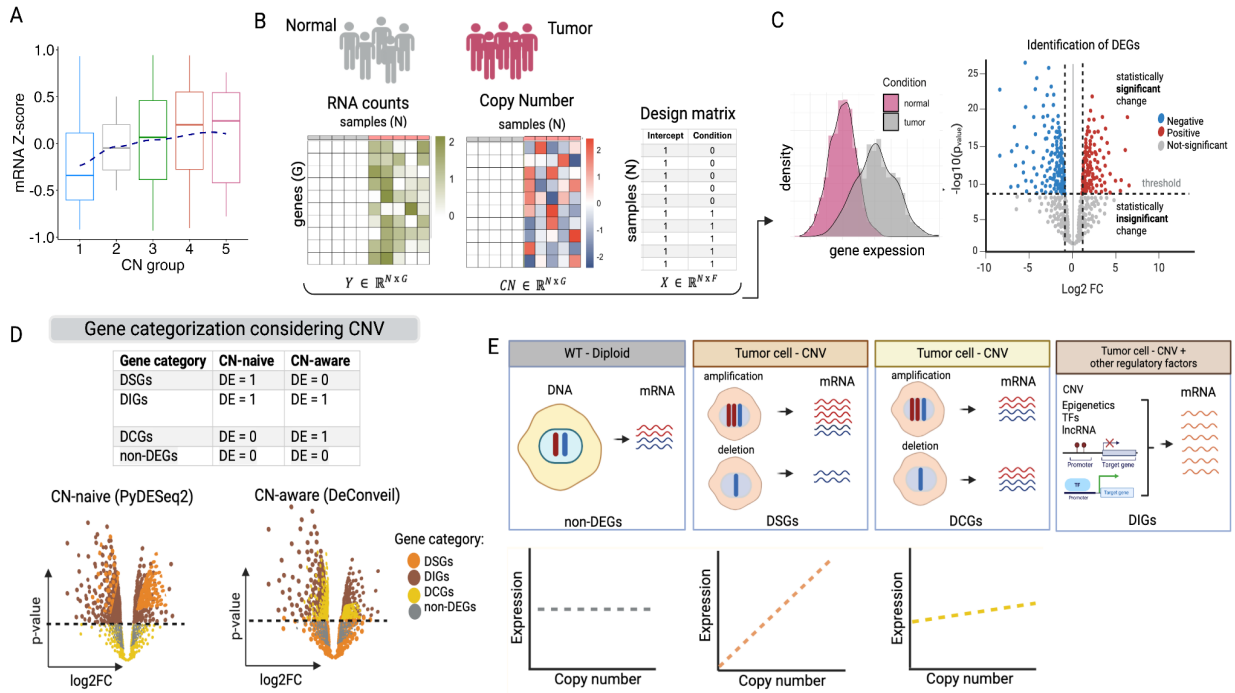

Figure 1. Overview of DeConveil and its approach to CNV-aware differential gene expression analysis. **A.** Relationship between gene expression and CNVs. Boxplots illustrate the distribution of mRNA Z-scores across different CN groups in LUAD tumor samples. The dashed blue line represents a locally weighted scatterplot smoothing (LOESS) curve, highlighting the trend between gene expression and CN changes. **B.** DGE analysis workflow. Tumor and normal samples undergo RNA-seq analysis, with CNV data incorporated to assess their influence on gene expression. A design matrix integrates sample conditions (normal = 0, tumor =1). **C**. Statistical testing determines DEGs based on log2FC and p-values. The volcano plot highlights the criteria for selection of significantly dysregulated genes (|log2FC| > 1 & p < 0.05). **D.** Gene categorization with CNV consideration. Genes are classified into four categories based on CN-naive (PyDESeq2) and CN-aware (DeConveil) analyses: DSGs (dosage-sensitive genes), differentially expressed only in CN-naive analysis; DIGs (dosage-insensitive genes)**,** differentially expressed in both CN-naive and CN-aware analyses; DCGs (dosage-compensated genes)**:** not identified as differentially expressed in CN-naive analysis but significant in CN-aware analysis; non-DEGs, genes not differentially expressed in either approach. **D.** Conceptual model illustrating the effects

of CN variation on mRNA expression in different cellular contexts. In wild-type (WT) diploid cells, gene expression remains balanced. In tumor cells, CNVs lead to dosage-dependent changes in mRNA levels (amplification increases and deletions decrease expression). However, in some cases, dosage compensation mechanisms mitigate CNV effects, and additional regulatory factors (e.g., epigenetics, transcription factors, and lncRNAs) can lead to dosage-independent expression.

Interpretation of gene expression considering CNVs

Genes are classified as differentially expressed (DE = 1) if they meet both statistical and biological significance thresholds: $p<0.05$ and $|\log2FC|>1$ (Fig.1C). Genes that do not meet these criteria are classified as non-differentially expressed (DE = 0).

We further classify genes into three categories: dosage-sensitive (DSGs), dosage-insensitive (DIGs) and dosage-compensated genes (DCGs) (Fig. 1C). This classification is based on the results of two methods PyDESeq2 (CN-naive approach) and DeConveil (CN-aware approach). The table in Figure 1D provides a detailed classification framework, distinguishing between CN-driven and CN-independent expression effects based on gene responses to CN changes.

DSGs respond proportionally to changes in CN, showing direct linear expression shifts (Fig.1E). This linear assumption simplifies modeling and aligns with basic gene dosage principles. However, we acknowledge that biological systems often exhibit more complex, non-linear relationships.

Unlike DSGs, DIGs are assumed to show differential expression driven by alternative regulatory mechanisms independent of CN changes, such as transcriptional control, epigenetic modifications, or post-transcriptional processes [19].

DCGs are defined by dosage compensation, a biological mechanism supported by growing evidence in cancer research [14,32]. In these genes, cells adjust gene expression to maintain balance despite variations in CN [33,34]. For example, when CN increases, cells may reduce transcription from amplified genes to prevent overexpression. Conversely, when CN decreases, cells may upregulate transcription to compensate for gene loss (Fig. 1E). These adaptive responses help maintain cellular homeostasis and could play a key role in tumor adaptation and progression. Based on this, we hypothesize that DCGs exhibit lower expression variability despite significant CN changes [32].

The computational model to implement DeConveil is written in Python as an extension of the DESeq2/PyDESeq2 [22,28] statistical framework (https://github.com/caravagnalab/DeConveil).

Results

DeConveil validation using simulated data

We assessed DeConveil's performance using simulations with varying sample sizes (number of patients 10, 20, 40, 100) and gene counts (1000, 3000, 5000), generating paired mRNA read counts and CN data. Baseline mRNA-seq read counts were simulated using the function *generateSyntheticData* of compcodeR simulator [35], with mean and dispersion parameters estimated from the TCGA-BRCA cancer dataset. For each combination of sample size and gene count we generated 10 independent datasets to ensure robustness.

To assess DeConveil's ability to separate CNV-independent from CNV-driven expression changes, we modeled two gene categories: DSGs and DIGs (Fig.2A). DIGs, true CNV-independent genes, serve as the true positives in our validation framework. DSGs, treated as false positives, as they should ideally be excluded from the DIG category. For DSGs, expression levels were adjusted to reflect DNA CN changes. This was represented as by scaling baseline mRNA read counts for CN. DIGs instead were assigned

expression levels independent of CNV, scaling baseline mRNA read counts for a differential expression factor. Each dataset included 10% DSGs, 40% DIGs, and 50% non-DEGs to ensure a balanced test scenario [36,37]. DSG simulations incorporated realistic CN data from TCGA to reflect biologically plausible CNVs.

Since DeConveil's primary goal is gene categorization, we validated its ability to accurately identify DIGs while minimizing DSG misclassification. DeConveil outperformed CN-naive methods (PyDESeq2 and edgeR) in accuracy (0.86–0.98 vs. 0.82–0.90), precision (0.95–0.97 vs. 0.80–0.82), and specificity (0.95–0.98 vs. 0.82–0.90), with the performance gap increasing in larger datasets (Fig. 2B). DeConveil's specificity remained stable, whereas CN-naive methods showed a sharper decline in specificity as sample sizes increased, suggesting a higher rate of false positives in larger datasets. AUROC analysis confirmed DeConveil's superiority in FDR control (0.82–0.98 vs. 0.78–0.90), particularly in complex scenarios (Fig. 2B, Fig.S2, Table S1).

These findings highlight that by incorporating CNV information, DeConveil provides a more precise and reliable classification of DE genes, effectively distinguishing true CNV-independent genes (DIGs) from CNV-driven genes (DSGs)**.**

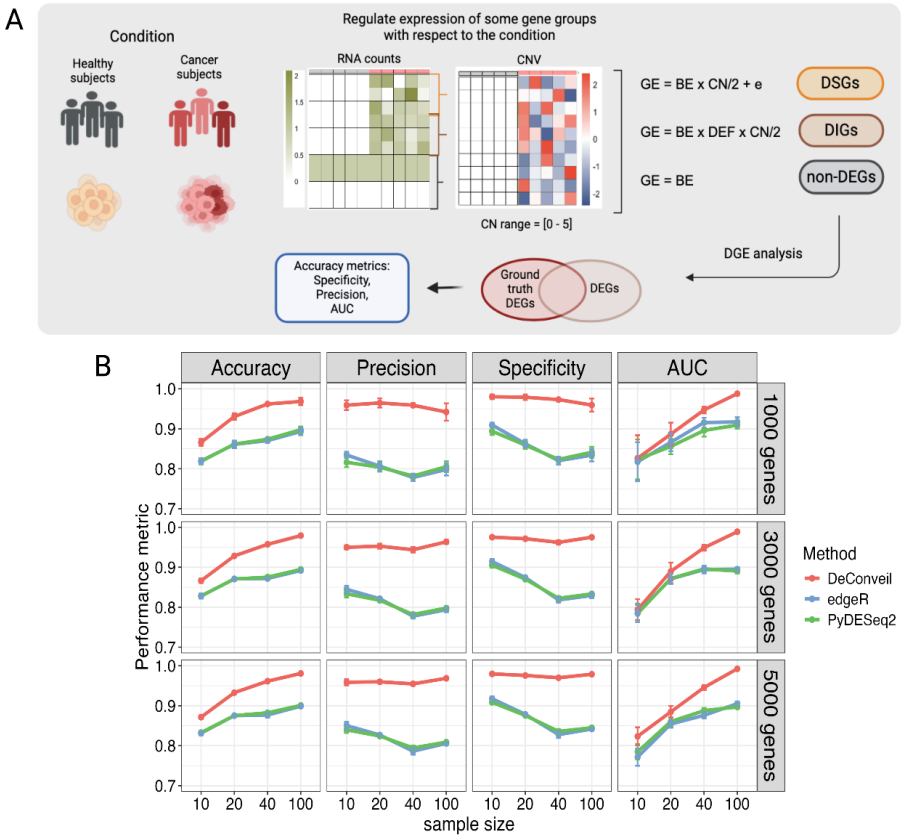

Figure 2. A. Schematic overview of the simulation setup for DGE analysis.Gene expression (GE) is modeled in relation to CNV changes**.** GE can be influenced by a baseline expression (BE) factor, CN changes, differential expression factor (DEF), and a residual error term. **B**. Performance evaluation of the DeConveil method on synthetic datasets with varying sample sizes (10, 20, 40, 100) and gene set sizes (1000 and 5000). Three key performance metrics — accuracy, precision, specificity, and AUC — were assessed and compared against CNV-naive methods (PyDESeq2 and edgeR).

Application of DeConveil to DGE analysis

We used DeConveil to understand how CN corrections of gene expression influence DGE analysis results outcomes in a real scenario. We focused on evaluating the DeConveil ability to categorize gene expression based on our hypothesis of transcriptional effects driven by CNV. For this analysis, we selected different solid cancer types with high variability in gene CN based on availability of matched normal-tumor samples. Using a set of 45 tumor-normal matched samples of lung adenocarcinoma (LUAD), we compared PyDESeq2 and DeConveil approaches to demonstrate how our CN-informed approach enables more refined gene expression categorization by distinguishing active expression changes from passive effects caused by CNVs. To enhance the reliability of the results, we excluded genes with low expression in normal tissue (mean expression < 10 reads). After filtering, our analysis focused on 19,830 genes.

Our results reveal that CN gains and amplifications (CN 3–5) have a significant impact on gene expression, affecting approximately 75 % of DEGs (Fig. 3A-C). DeConveil identified 875 (14.9%) DSGs, 3969 (67.6%) DIGs, and 1029 (17.5%) DCGs (Fig. 3B). We then assessed the differences between PyDESeq2 and DeConveil analyses by comparing the effect sizes (log2FC*)* and false discovery rate (FDR) across both methods (Fig. 3D). For genes with neutral CNs (no CN change) used as controls, we observed concordance between the two methods (diagonal trend). However, genes affected by CN gains and amplifications exhibited higher deviations from the diagonal, confirming the influence of CN adjustment in regions with aneuploidy. Notably, DeConveil analysis increased FDR for amplified and gain-affected genes (Fig. 3D). On average, FDRs shifted by 0.0033, with 40.75% (1837 genes) showing increased p-values, reinforcing this trend. Furthermore, 122 genes lost statistical significance ($p < 0.05$ in CN-naive but $p > 0.05$ in CN-aware analysis),

To evaluate expression variability, we analyzed effect size differences between the methods (Fig. 3E). Amplified and gain genes showed the largest effect size recalibrations of 0.79 ± 0.28 and 0.34 ± 0.26 respectively, while neutral and loss categories exhibited minimal changes (0.11 ± 0.12 and 0.14 ± 0.14) indicating less CN-driven effect. For instance, the number of downregulated genes increases after CN-aware adjustment (from 2526 to 3431), while the number of upregulated genes decreases (from 2318 to 1567), highlighting DeConveil's ability to separate active regulatory effects from passive CNV-driven changes (Fig. 3F).

We extended this analysis to four other cancer types with varying levels of CN variability (Fig.S3): lung squamous cell carcinoma (LUSC), breast invasive carcinoma (BRCA), liver hepatocellular carcinoma (LIHC), and kidney renal clear cell carcinoma (KIRC). Cancers with higher CN variability (CN 1-5), such as LUSC, BRCA, and LIHC, showed a greater impact of CN corrections, as evidenced by a larger proportion of DSGs (13.7 - 21.4%) and DCGs (10.7 - 20.3%) (Fig.S3, A-C). In contrast, KIRC, which exhibits lower CN variability (CN 1–3), underwent minimal gene classification shifts, with fewer DSGs (10.4%) and DCGs (8.0%). This suggests that the influence of CN corrections is less pronounced in cancers with low CN variability.

Insights into DSGs, DIGs and DCGs in aneuploid cancers

We further analyzed DSGs, DIGs, and DCGs identified by DeConveil across three epithelial origin aneuploid cancers, LUAD, LUSC, and BRCA (Fig. 4A), to explore shared and private gene expression patterns and uncover functional pathways associated with each gene category.

Among the DSGs, only 33 (0.53 %) were shared across all three cancers, including key oncogenes (MYCN [38], MSI2 [39]), and tumor suppressors (PARP1 [40]), reported to be critical for tumor progression and survival. Most DSGs are cancer specific (S2 Table) across all analyzed tumor types (11% of genes are private of the total 15%).

DIGs, more abundant than DSGs, showed broader conservation across three cancers (S3 Table), with 937 (15.2% of the total 68.6%) shared DIGs, including 42 known oncogenes and tumor suppressors. Their

greater stability suggests that DIGs represent a default expression state, largely independent of CNVs, allowing tumors to maintain essential pathways despite genomic alterations.

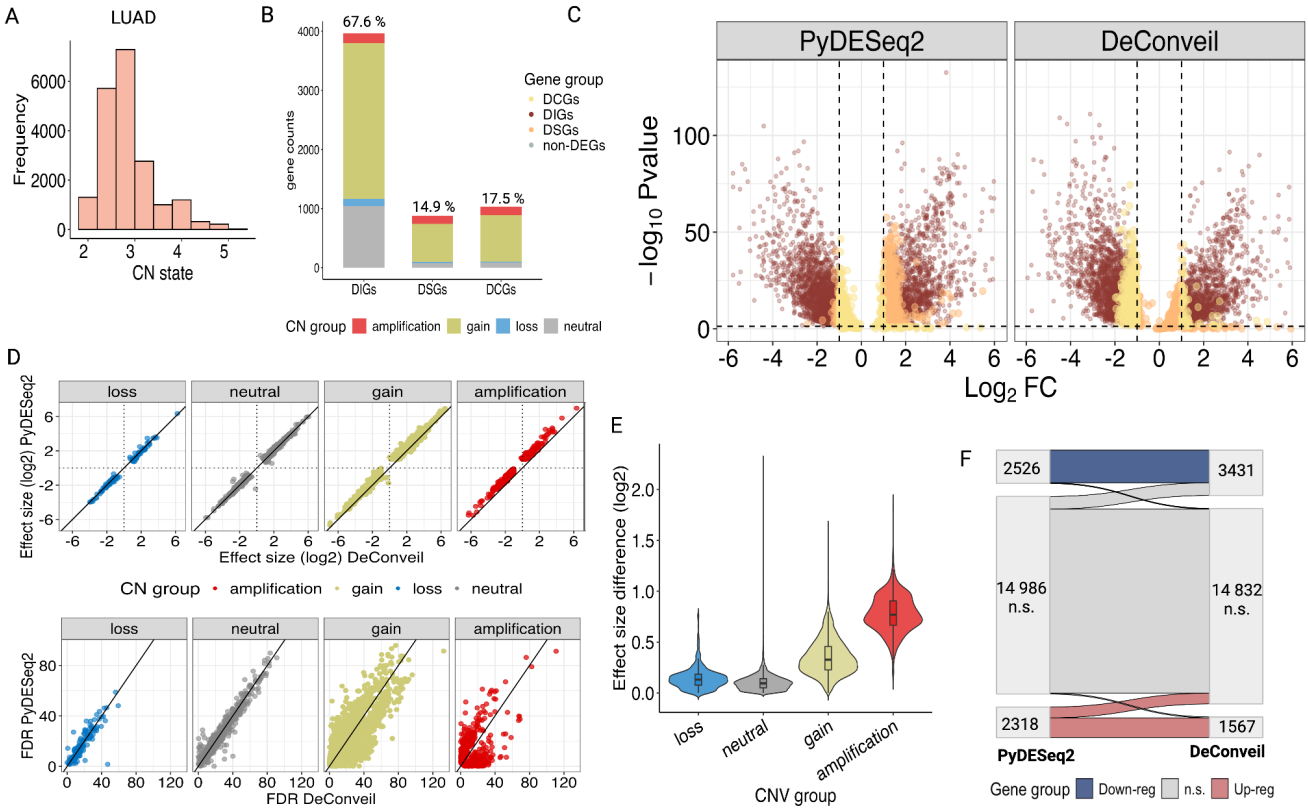

Figure 3. Impact of CN corrections on DGE analysis in aneuploid lung adenocarcinoma (LUAD). **A.** Histogram showing the distribution of CNV states across LUAD tumor samples. **B.** Categorization of genes based on their CNV status and classification into different gene groups: DIGs, DSGs, and DCGs. The stacked bar plot shows the proportion of each gene category, with CN states (amplification, gain, loss, and neutral. Percentages on top indicate the proportion of each group among the total gene set. **C.** Volcano plots comparing gene expression changes between the PyDESeq2 and DeConveil DGE analysis approaches. Each dot represents a gene, with its position determined by logFC and p-value. Threshold for significant DE: $|\log_2 FC| > 1$ and FDR < 0.05**. D.** Comparison of effect size (log2FC) and FDR (bottom row) between PyDESeq2 and DeConveil models across different CN groups: loss, neutral, gain, and amplification. The diagonal reference line represents a one-to-one correlation; deviations from this line indicate differences in effect size or FDR between the two approaches**. E.** Violin plot showing the distribution of effect size differences ($\log_2$ scale) across CNV groups. **F.** Sankey diagram illustrating the transition of genes between expression categories in PyDESeq2 vs. DeConveil analyses. Genes are grouped into upregulated (red), downregulated (blue), and non-significant (gray, n.s.) categories based on |log2FC| > 1 and FDR < 0.05. The diagram highlights shift in gene classification when DeConveil CN adjustments are applied.

DCGs also showed the least overlap across cancers, with only 30 (0.5% of the total 16%) shared genes, many of which are lncRNAs. This low overlap, along with the presence of lncRNAs, suggests that DCGs may function as regulatory elements influencing oncogenes and tumor suppressors activity through epigenetic or transcriptional mechanisms.

To assess the functional relevance of these shared gene categories, we performed Gene Ontology (GO) overrepresentation analysis (Fig.4B). The analysis revealed that shared DSGs are enriched in metabolic and cell cycle processes, confirming their importance in cancer cell survival under dosage variability. DIGs

are linked to cell cycle, immune response, and oncogenic signaling pathways, supporting their role in tumor maintenance beyond CNV effects. Meanwhile, DCGs were primarily linked to immune regulation and cell adhesion, suggesting their involvement in tumor-immune interactions.

Additionally, we examined the functional significance of private cancer-specific genes within each category. For instance, functional analysis (Fig. S5) linked DIGs to immune regulation and cell proliferation in LUAD, mesenchymal differentiation in LUSC, and hormone metabolism in BRCA, emphasizing their role in sustaining tumor-specific traits. Functional enrichment of private DCGs (Fig. S5) highlighted their involvement in tumor-immune interactions and metabolic adaptation, including MHC complex and T-cell activation in LUAD, cytokine regulation in LUSC, and insulin secretion in BRCA.

To further refine our understanding of the significance of private genes, we mapped them to known cancer-specific oncogenes and TSGs (Fig. 4C). As expected, the DIG category contained the highest number of oncogenes (n=172) and tumor suppressors (n=110), compared to DSGs (n=55 oncogenes, n=53 TSGs) and DCGs (n=53 oncogenes, n=31 TSGs). This supports the idea that DIGs encompass a broader range of critical cancer genes, many of which may operate independently of CNV effects, relying instead on regulatory mechanisms for tumor progression.

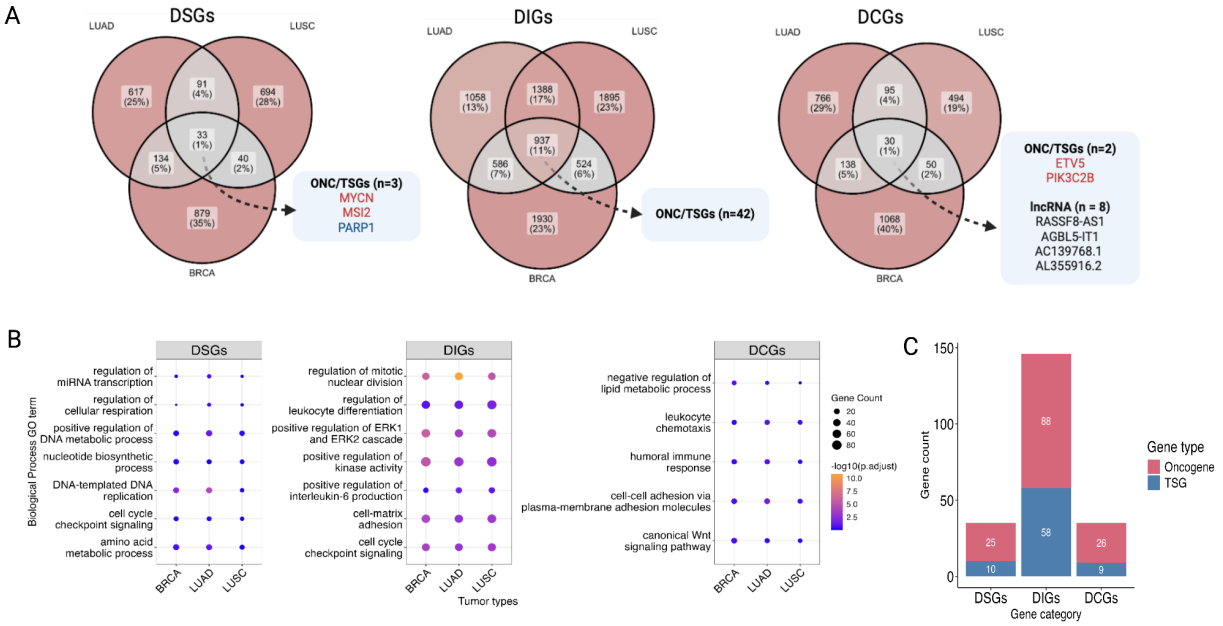

Figure 4. Cross-cancer comparison of DEG categories and their functional associations. **A.** Venn diagrams illustrate the overlap of DE gene categories (DSGs, DIGs, and DCGs) across three cancer types: LUAD, LUSC, and BRCA. Highlighted oncogenes (ONC) and tumor suppressor genes (TSGs) found in each category are listed in blue boxes. Additionally, 8 long non-coding RNAs (lncRNAs) were identified within the DCGs category. **B.** Gene Ontology (GO) term enrichment analysis for biological processes associated with DSGs, DIGs, and DCGs across LUAD, LUSC, and BRCA. The dot plots represent significantly enriched biological processes for each gene category. The size of the dots corresponds to the number of genes associated with the process, while the color represents the statistical significance of enrichment ($-\log_{10}$ adjusted p-value). **C.** Distribution of oncogenes (ONC) and TSGs within each gene category across private DEGs of three cancer types (LUAD, LUSC and BRCA).

Identification of prognostic biomarkers in breast cancer using DeConveil

In this case study, we investigated the prognostic potential of DSGs, DIGs, and DCGs identified by DeConveil in Breast Invasive Carcinoma (BRCA) dataset, a solid highly aneuploid tumor. From 110 paired tumor-normal samples, differential expression analysis identified 1086 (17.1%) DSGs, 3977 (62.6%) DIGs, and 1286 (20.3%) DCGs among 22 076 analyzed genes (Fig. S3, C). Notably, most of these genes were primarily influenced by CN gains and amplifications.

To assess their prognostic relevance, we applied Cox proportional hazards model to estimate hazard ratios (HRs) and confidence intervals (CIs). This analysis identified 142 DSGs, 161 DIGs, and 69 DCGs significantly associated with overall survival (p-value < 0.05). Further feature selection using LASSO regression highlighted key genes with prognostic potential. Several key genes emerged as significant biomarkers having prognostic potential (Fig.5A). For example, DSGs like FBXO45 and RAD21, along with DIGs such as AC113368.1, PEX5L, and GHR, were associated with worse survival outcomes (HR > 1). Similarly, several DCGs, including FGD5, FGFR1, MUC4, INSL3 displayed significant HRs. Many of these genes have well-established roles in breast cancer progression [41–44], metastasis [45,46], and therapy resistance. Interestingly, the analysis also identified lncRNAs within the DCG category with high prognostic potential, such as AC079822.1, AC244153.1, and LINC01431. lncRNA is known as novel gene regulatory molecules involved in cancer metabolic reprogramming and progression [47].

A comparison between DeConveil and PyDESeq2 results (Fig.5A) further highlighted the advantages of DeConveil's classification. The DeConveil identified a larger set of prognostic genes with moderate HRs, detecting key genes that PyDESeq2 overlooked. For example, among the genes with HR > 1, only one (FBXO45) was identified by both approaches. This demonstrates the increased sensitivity of the DeConveil CN-aware approach in identifying clinically relevant genes.

To assess the predictive utility of gene signatures from each category, we calculated prognostic scores and stratified an independent cohort of 520 breast cancer patients from the METABRIC dataset into high- and low-risk groups based on median scores. Kaplan-Meier survival analysis showed that DSGs and DCGs provided the most significant separation between risk groups, with p-values < 0.00094 and < 0.00018, respectively (Fig. 5B). DIGs also showed a clear separation between high- and low-risk groups, but the effect was less pronounced (p < 0.01). These results highlight the high prognostic value of DeConveil's gene classification framework.

Pathway enrichment analysis of the identified gene signatures (Fig.5C, Fig.S6) further supported their functional distinctions. DSGs and DCGs, in particular, were enriched in key cancer hallmarks, including cell cycle regulation, proliferation, adhesion, migration, and mesenchymal cell differentiation. In contrast, the DIG signature was enriched in a broader range of cancer-specific processes, such as hormone metabolism, immune response, cell adhesion, and signaling. These functional distinctions help explain the prognostic significance of these gene signatures.

Discussion and conclusions

Cancer transcriptomic heterogeneity is shaped by the interplay between genomic alterations and transcriptional regulation. Traditional DGE analyses identify expression changes across conditions but fail to assess the effect of CNVs in explaining such variations. In the context of cancers, where CNVs are prevalent, this makes it hard to distinguish which regulatory processes are impacted by CNVs, potentially obscuring key biological mechanisms underlying tumour progression.

DeConveil extends canonical approaches for DGE with CNV data, allowing a better accounting of the role of CNVs in impacting gene expression. An immediate byproduct of this new approach is the possibility of classifying genes based on the interplay among transcriptional status and CNVs. This innovative classification reveals which genes are compensated by the CNVs and those sensitive or insensitive to

aneuploidy. These hypothesized biological mechanisms suggest complex regulatory mechanisms that might back up changes in gene multiplicity, underlying a complicated post-transcriptional regulatory network that acts to disrupt tissue homeostasis in the context of cancer. The implications of this new approach are also statistical. Traditional CN-naive approaches often fail to report dosage-compensated genes, suggesting that our approach might improve even established analyses of large cohorts [49–52].

We used DeConveil to analyse TCGA datasets, identifying shared and cancer-type-specific patterns of DEGs. In breast cancer, we identified known and novel prognostic biomarkers, including previously uncharacterized lncRNAs. By stratifying these genes based on CNV-related expression mechanisms, the framework improved risk group stratification and outperformed traditional DGE methods in predicting survival outcomes. This refined prognostic approach could aid in personalized treatment planning by identifying high-risk patients who may benefit from targeted interventions.

While the main contribution of our approach is to include CNVs in the DEGs picture, DeConveil still has some limitations. For instance, we assumed a linear modelling framework that may not fully capture nonlinear relationships between CNVs and gene expression. These are particularly important for high-ploidy states and RNA buffering effects that saturate the signal [4,14]. In this direction, models with diminishing returns might be considered to improve our approach. Additionally, the logical connection between CNVs and expression cannot factor out the role of epigenomic features such as DNA methylation events [53,54]. In the future, a richer framework might include epi-mutation data, CNVs, and RNA counts to deliver a fine-grain characterization of DEGs.

Another key limitation is that DeConveil models CNV effects at the individual gene level and does not explicitly account for indirect transcriptional effects. CNVs in transcription factors (TFs) or signaling genes can alter the expression of downstream targets, even if those genes do not have CNVs themselves [55]. This means DeConveil may not capture CNV-driven regulatory cascades, where gene expression changes arise due to disruptions in transcriptional networks rather than direct CNV effects. Future extensions could incorporate gene regulatory network (GRN) models to map CNV-mediated transcriptional changes [56] or use causal inference approaches to distinguish direct vs. indirect CNV effects [57].

Code and Data availability

The source code to run DeConveil is freely available https://github.com/caravagnalab/DeConveil.
The code to reproduce the analysis is available at https://github.com/Katerina10-cloud/deconveilCaseStudies
The TCGA preprocessed data used in this study are available at DOI 10.5281/zenodo.15100457

Acknowledgments

The authors acknowledge support from the Italian Association for Cancer Research (AIRC) under grant MFAG 2020 - ID. 24913 project (PI - GC) and IG 27631 (PI - GS). GC acknowledges financial support under the National Recovery and Resilience Plan (NRRP), Mission3224, Component 2, Investment 1.1, Call for tender No. 1409 published on 14.9.2022 by the Italian Ministry of University323 and Research (MUR), funded by the European Union – NextGenerationEU– CUP J53D23015060001, as well as324 under Decreto Direttoriale No. 104 published on 02-02-2022 by MUR (NextGenerationEU– CUP J53D23003860006),325 under the PNRR-MAD-2022-12375673 (Next Generation EU, M6/C2 CALL 2022) and under the PNRR-MCNT2-3262023-12378037 (Next Generation EU, M6/C2 CALL 2023), both published by the Italian Ministry of Health, Rome,327 Italy.

GS acknowledges co-funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE1 - Project FAIR “Future Artificial Intelligence Research”. This resource was co-financed by the Next Generation EU [DM 1555 del 11.10.22].

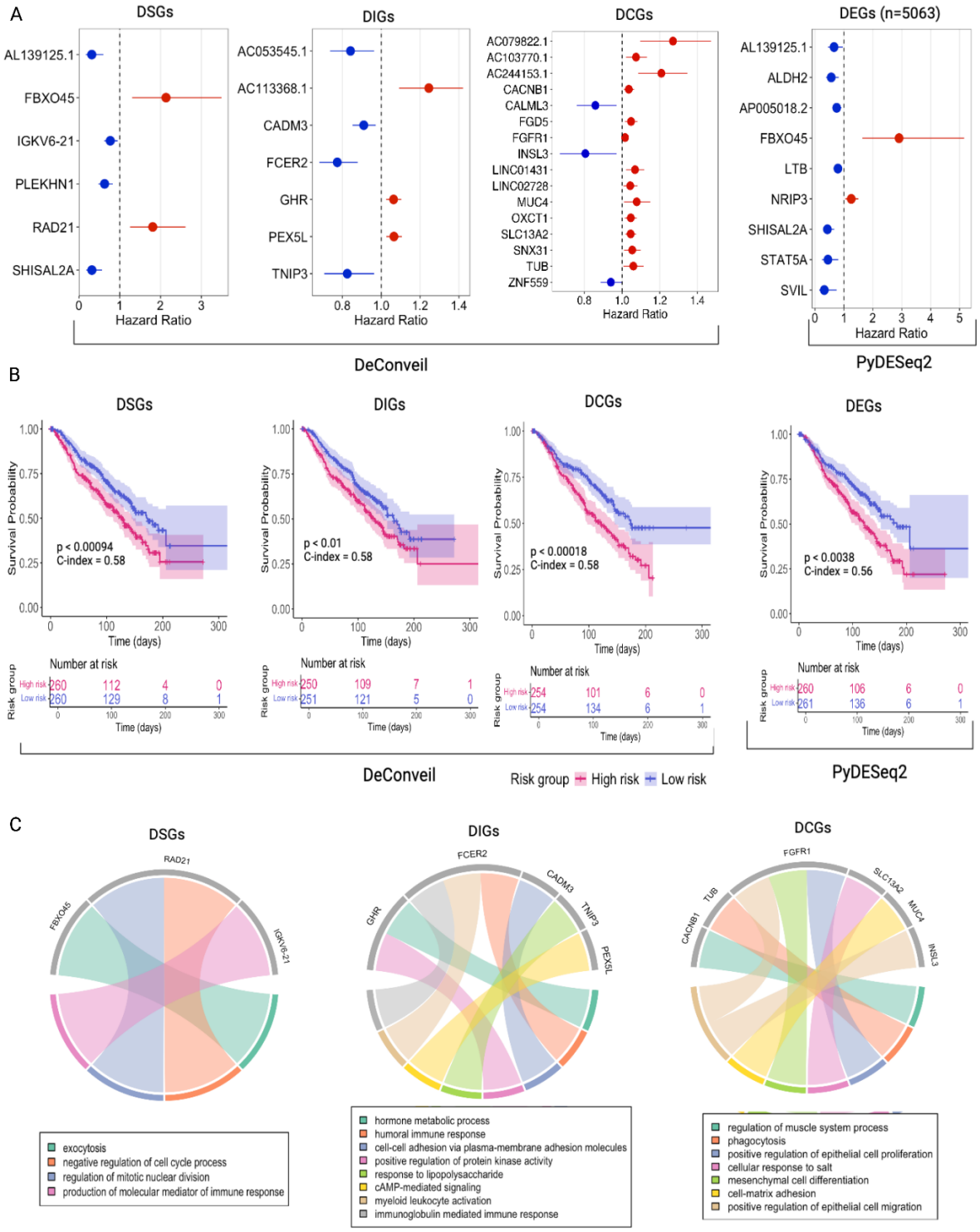

Figure 5. Prognostic significance of DEG categories and their functional associations. **A.**Cox proportional hazards regression analysis identifying prognostic genes within DSGs, DIGs, and DCGs. The forest plots display hazard ratios (HR) for selected prognostic genes identified using LASSO regression ($p < 0.05$). Genes with HR > 1 (red) indicate a high-risk association with poor survival, while those with HR < 1 (blue) are linked to better survival outcomes. **B.** Kaplan-Meier survival curves comparing high-risk (red) and low-risk (blue) patient groups based on prognostic genes identified within DSGs, DIGs, and DCGs under DeConvei and PyDESeq2 models ($p < 0.05$). The concordance index (C-index) values

indicate the predictive accuracy of the survival model. **C.** Gene Ontology (GO) enrichment analysis illustrating biological pathways associated with DeConveil prognostic genes in each gene category.

References

1. Sdeor E, Okada H, Saad R, Ben-Yishay T, Ben-David U. Aneuploidy as a driver of human cancer. Nature Genetics. 2024;56: 2014–2026. doi:10.1038/s41588-024-01916-2

2. Girish V, Lakhani AA, Thompson SL, Scaduto CM, Brown LM, Hagenson RA, et al. Oncogene-like addiction to aneuploidy in human cancers. Science. 2023;381: eadg4521. doi:10.1126/science.adg4521

3. Prasad K, Bloomfield M, Levi H, Keuper K, Bernhard SV, Baudoin NC, et al. Whole-Genome Duplication Shapes the Aneuploidy Landscape of Human Cancers. Cancer Res. 2022;82: 1736–1752. doi:10.1158/0008-5472.CAN-21-2065

4. Bhattacharya A, Bense RD, Urzúa-Traslaviña CG, de Vries EGE, van Vugt MATM, Fehrmann RSN. Transcriptional effects of copy number alterations in a large set of human cancers. Nature Communications. 2020;11: 1–12. doi:10.1038/s41467-020-14605-5

5. The landscape of genomic copy number alterations in colorectal cancer and their consequences on gene expression levels and disease outcome. Molecular Aspects of Medicine. 2019;69: 48–61. doi:10.1016/j.mam.2019.07.007

6. Shi H, Williams MJ, Satas G, Weiner AC, McPherson A, Shah SP. Allele-specific transcriptional effects of subclonal copy number alterations enable genotype-phenotype mapping in cancer cells. Nat Commun. 2024;15: 2482. doi:10.1038/s41467-024-46710-0

7. Impacts of Variation in the Human Genome on Gene Regulation. Journal of Molecular Biology. 2013;425: 3970–3977. doi:10.1016/j.jmb.2013.07.015

8. Fehrmann RSN, Karjalainen JM, Krajewska M, Westra H-J, Maloney D, Simeonov A, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat Genet. 2015;47: 115–125. doi:10.1038/ng.3173

9. Harbers L, Agostini F, Nicos M, Poddighe D, Bienko M, Crosetto N. Somatic Copy Number Alterations in Human Cancers: An Analysis of Publicly Available Data From The Cancer Genome Atlas. Front Oncol. 2021;11: 700568. doi:10.3389/fonc.2021.700568

10. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013;45: 1134–1140. doi:10.1038/ng.2760

11. Orozco LD, Cokus SJ, Ghazalpour A, Ingram-Drake L, Wang S, van Nas A, et al. Copy number variation influences gene expression and metabolic traits in mice. Hum Mol Genet. 2009;18: 4118–4129. doi:10.1093/hmg/ddp360

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

13. Zhou C, Zhang W, Chen W, Yin Y, Atyah M, Liu S, et al. Integrated Analysis of Copy Number Variations and Gene Expression Profiling in Hepatocellular carcinoma. Sci Rep. 2017;7: 10570. doi:10.1038/s41598-017-11029-y

14. Schukken KM, Sheltzer JM. Extensive protein dosage compensation in aneuploid human cancers.

Genome Res. 2022;32: 1254–1270. doi:10.1101/gr.276378.121

15. Liu Q, Su P-F, Zhao S, Shyr Y. Transcriptome-wide signatures of tumor stage in kidney renal clear cell carcinoma: connecting copy number variation, methylation and transcription factor activity. Genome Med. 2014;6: 117. doi:10.1186/s13073-014-0117-z

16. Jia Y, Chen L, Jia Q, Dou X, Xu N, Liao DJ. The well-accepted notion that gene amplification contributes to increased expression still remains, after all these years, a reasonable but unproven assumption. J Carcinog. 2016;15: 3. doi:10.4103/1477-3163.182809

17. Henrichsen CN, Chaignat E, Reymond A. Copy number variants, diseases and gene expression. Hum Mol Genet. 2009;18: R1–8. doi:10.1093/hmg/ddp011

18. Gamazon ER, Stranger BE. The impact of human copy number variation on gene expression. Brief Funct Genomics. 2015;14: 352–357. doi:10.1093/bfgp/elv017

19. Mohanty V, Wang F, Mills GB, CTD2 Research Network, Chen K. Uncoupling of gene expression from copy number presents therapeutic opportunities in aneuploid cancers. Cell Rep Med. 2021;2: 100349. doi:10.1016/j.xcrm.2021.100349

20. Abdolahi F, Shahraki A, Sheervalilou R, Mortazavi SS. Identification of differentially expressed genes associated with the pathogenesis of gastric cancer by bioinformatics analysis. BMC Med Genomics. 2023;16: 311. doi:10.1186/s12920-023-01720-7

21. Samadi P, Soleimani M, Nouri F, Rahbarizadeh F, Najafi R, Jalali A. An integrative transcriptome analysis reveals potential predictive, prognostic biomarkers and therapeutic targets in colorectal cancer. BMC Cancer. 2022;22: 835. doi:10.1186/s12885-022-09931-4

22. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550. doi:10.1186/s13059-014-0550-8

23. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26: 139–140. doi:10.1093/bioinformatics/btp616

24. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47. doi:10.1093/nar/gkv007

25. Robinson MD, Strbenac D, Stirzaker C, Statham AL, Song J, Speed TP, et al. Copy-number-aware differential analysis of quantitative DNA sequencing data. Genome Res. 2012;22: 2489–2496. doi:10.1101/gr.139055.112

26. Hilbe JM. Negative Binomial Regression. Cambridge University Press; 2011. Available: https://play.google.com/store/books/details?id=DDxEGQuqkJoC

27. Li, Xiaohong, Dongfeng Wu, Nigel G. F. Cooper, and Shesh N. Rai. 2019. 'Sample Size Calculations for the Differential Expression Analysis of RNA-Seq Data Using a Negative Binomial Regression Model.' Statistical Applications in Genetics and Molecular Biology 18 (1). https://doi.org/10.1515/sagmb-2018-0021.

28. Muzellec B, Teleńczuk M, Cabeli V, Andreux M. PyDESeq 2: a python package for bulk RNA-seq differential expression analysis. Bioinformatics. 2023;39. doi:10.1093/bioinformatics/btad547

29. Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics. 2019;35: 2084–2092. doi:10.1093/bioinformatics/bty895

30. Lehmann EL, Romano JP. Testing Statistical Hypotheses. Springer Nature; 2022. Available: https://play.google.com/store/books/details?id=ZNJ2EAAAQBAJ

31. Ferreira JA, Zwinderman AH. On the Benjamini–Hochberg method. Ann Stat. 2006;34: 1827–1849. doi:10.1214/009053606000000425

32. Acón M, Geiß C, Torres-Calvo J, Bravo-Estupiñan D, Oviedo G, Arias-Arias JL, et al. Dosage compensation is mediated by miRNA-transcription factor interactions in aneuploid cancer. iScience. 2021;24: 103407. doi:10.1016/j.isci.2021.103407

33. Kojima S, Cimini D. Aneuploidy and gene expression: is there dosage compensation? Epigenomics. 2019;11: 1827–1837. doi:10.2217/epi-2019-0135

34. Bravo-Estupiñan DM, Aguilar-Guerrero K, Quirós S, Acón M-S, Marín-Müller C, Ibáñez-Hernández M, et al. Gene dosage compensation: Origins, criteria to identify compensated genes, and mechanisms including sensor loops as an emerging systems-level property in cancer. Cancer Med. 2023;12: 22130–22155. doi:10.1002/cam4.6719

35. Soneson C. compcodeR--an R package for benchmarking differential expression methods for RNA-seq data. Bioinformatics. 2014;30: 2517–2518. doi:10.1093/bioinformatics/btu324

36. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14: 91. doi:10.1186/1471-2105-14-91

37. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. 2016;22: 839–851. doi:10.1261/rna.053959.115

38. Liu Z, Chen SS, Clarke S, Veschi V, Thiele CJ. Targeting in Pediatric and Adult Cancers. Front Oncol. 2020;10: 623679. doi:10.3389/fonc.2020.623679

39. Yeh D-W, Zhao X, Siddique HR, Zheng M, Choi HY, Machida T, et al. MSI2 promotes translation of multiple IRES-containing oncogenes and virus to induce self-renewal of tumor initiating stem-like cells. Cell Death Discovery. 2023;9: 1–15. doi:10.1038/s41420-023-01427-9

40. Wang L, Liang C, Li F, Guan D, Wu X, Fu X, et al. PARP1 in Carcinomas and PARP1 Inhibitors as Antineoplastic Drugs. Int J Mol Sci. 2017;18. doi:10.3390/ijms18102111

41. Zheng M, Wu L, Xiao R, Cai J, Chen W, Shen S. Fbxo45 facilitates the malignant progression of breast cancer by targeting Bim for ubiquitination and degradation. BMC Cancer. 2024;24: 1–13. doi:10.1186/s12885-024-12382-8

42. Lin M, Wang Z-W, Zhu X. FBXO45 is a potential therapeutic target for cancer therapy. Cell Death Discovery. 2020;6: 1–5. doi:10.1038/s41420-020-0291-2

43. Zhu J, Shen Y, Wang L, Qiao J, Zhao Y, Wang Q. A novel 12-gene prognostic signature in breast cancer based on the tumor microenvironment. Ann Transl Med. 2022;10: 143. doi:10.21037/atm-21-6748

44. Valla M, Engstrøm MJ, Ytterhus B, Hansen ÅKS, Akslen LA, Vatten LJ, et al. FGD5 amplification in breast cancer patients is associated with tumour proliferation and a poorer prognosis. Breast Cancer Research and Treatment. 2017;162: 243–253. doi:10.1007/s10549-017-4125-8

45. Dreyer CA, VanderVorst K, Free S, Rowson-Hodel A, Carraway KL. The role of membrane mucin MUC4 in breast cancer metastasis. Endocr Relat Cancer. 2021;29: R17–R32. doi:10.1530/ERC-21-0083

46. Rios Garcia M, Steinbauer B, Srivastava K, Singhal M, Mattijssen F, Maida A, et al. Acetyl-CoA

Carboxylase 1-Dependent Protein Acetylation Controls Breast Cancer Metastasis and Recurrence. Cell Metab. 2017;26: 842–855.e5. doi:10.1016/j.cmet.2017.09.018

47. Shi X, Sun M, Wu Y, Yao Y, Liu H, Wu G, et al. Post-transcriptional regulation of long noncoding RNAs in cancer. Tumor Biology. 2015;36: 503–513. doi:10.1007/s13277-015-3106-y

48. Peng L, Bian XW, Li DK, Xu C, Wang GM, Xia QY, et al. Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types. Scientific Reports. 2015;5: 1–18. doi:10.1038/srep13413

49. Huang X, Stern DF, Zhao H. Transcriptional Profiles from Paired Normal Samples Offer Complementary Information on Cancer Patient Survival – Evidence from TCGA Pan-Cancer Data. Scientific Reports. 2016;6: 1–9. doi:10.1038/srep20567

50. Thennavan A, Beca F, Xia Y, Recio SG, Allison K, Collins LC, et al. Molecular analysis of TCGA breast cancer histologic types. Cell Genom. 2021;1. doi:10.1016/j.xgen.2021.100067

51. Li Y, Gu J, Xu F, Zhu Q, Ge D, Lu C. Transcriptomic and functional network features of lung squamous cell carcinoma through integrative analysis of GEO and TCGA data. Scientific Reports. 2018;8: 1–12. doi:10.1038/s41598-018-34160-w

52. Ashouri A, Sayin VI, Van den Eynden J, Singh SX, Papagiannakopoulos T, Larsson E. Pan-cancer transcriptomic analysis associates long non-coding RNAs with key mutational driver events. Nature Communications. 2016;7: 1–13. doi:10.1038/ncomms13197

53. Yang J, Xu J, Wang W, Zhang B, Yu X, Shi S. Epigenetic regulation in the tumor microenvironment: molecular mechanisms and therapeutic targets. Signal Transduction and Targeted Therapy. 2023;8: 1–26. doi:10.1038/s41392-023-01480-x

54. Capp J-P. Interplay between genetic, epigenetic, and gene expression variability: Considering complexity in evolvability. Evol Appl. 2021;14: 893–901. doi:10.1111/eva.13204

55. Schaub FX, Dhankani V, Berger AC, Trivedi M, Richardson AB, Shaw R, et al. Pan-cancer Alterations of the MYC Oncogene and Its Proximal Network across the Cancer Genome Atlas. Cell Syst. 2018;6: 282–300.e2. doi:10.1016/j.cels.2018.03.003

56. Li R, Rozum JC, Quail MM, Qasim MN, Sindi SS, Nobile CJ, et al. Inferring gene regulatory networks using transcriptional profiles as dynamical attractors. PLoS Comput Biol. 2023;19: e1010991. doi:10.1371/journal.pcbi.1010991

57. Jiang Z, Chen C, Xu Z, Wang X, Zhang M, Zhang D. SIGNET: transcriptome-wide causal inference for gene regulatory networks. Sci Rep. 2023;13: 19371. doi:10.1038/s41598-023-46295-6