PanCNV-Explorer: Deciphering copy number alterations across human cancers

Nadine S. Kurz[1,2,*], Kevin Kornrumpf[1], Anna-Rosa Krüger[1], and Jürgen Dönitz[1,2,3]

[1]Department of Medical Bioinformatics, University Medical Center Göttingen, Goldschmidtstraße 1, 37075 Göttingen, Niedersachsen, Germany
[2]Comprehensive Cancer Center Göttingen (G-CCC), Göttingen, 37075, Germany
[3]Campus-Institut Data Science (CIDAS), Section Medical Data Science (MeDaS), 37075 Göttingen, Niedersachsen, Germany

Abstract

Copy number variants (CNVs) drive cancer progression and genetic disorders, yet the interpretation of their biological consequences and potential targeted therapeutic options remains fragmented across clinical, functional, and structural domains. To bridge this gap, we present PanCNV-Explorer, a comprehensive annotated CNV database integrating copy number variation data from pan-cancer and healthy cohorts. PanCNV-Explorer combines CNV profiles with functional genomic layers, including gene expression CRISPR screening data, through a novel analytical framework. PanCNV-Explorer represents a systematic map of copy number variation in the human genome across 33 different cancer types and normal tissue, integrating pan-cancer and healthy samples into a comprehensive, harmonized database. These analyses reveal context-specific oncogenic drivers, vulnerabilities, and therapeutic targets, accessible via an interactive web interface for dynamic exploration and hypothesis generation. In addition, the web server provides programmatic web services for annotating user-submitted CNVs with functional annotation, clinical relevance, and pathogenicity predictions. PanCNV-Explorer serves as a pivotal resource for accelerating the analyses of copy number and structural variants in the human genome, bridging raw CNV data to actionable biological and clinical interpretations. A public web instance of the PanCNV-Explorer web server is available at https://mtb.bioinf.med.uni-goettingen.de/pancnv-explorer/.
Keywords: Copy number alterations, Somatic variants, variant interpretation, Molecular tumor boards

1 Introduction

Copy number alterations (CNAs) are structural genetic alterations that involve gains or losses of DNA segments [5], representing profound implications for human health. Occurring > 40% of tumors [6, 7] and directly dysregulating oncogenes and tumor suppressor through dosage effects [8], CNVs have been identified as one of the major drivers of cancer [9, 10, 11, 12]. Pan-cancer analyses of human copy number variation has revealed striking cancer-type specificity: Specific regions were shown to be frequently more affected by CNVs in specific cancer types [14], and mutational signatures vary significantly between cancer types [11]. CNVs directly indicate targeted treatment efficacy, as *ERBB2* amplifications guide trastuzumab in breast cancer . They also serve as biomarkers for cancer subtype identification, as *CDKN2A/B* deletions define the activated B-cell-like (ABC) subtype of diffuse large B-cell lymphoma (DLBCL) [13], a population excluded from standard R-CHOP therapy but responsive to CDK4/6 inhibitors. GISTIC2.0 [27] has become the community standard for CNV analysis, providing a robust statistical framework to identify recurrent amplifications and deletions while controlling for background noise and false discovery rates. While GISTIC2.0 excels at statistically significant copy number alterations (CNAs) within individual cancer cohorts, it is designed for cohort-specific analysis and lacks native functionality for direct sta-

tistical comparison of CNA frequencies across independent datasets from different cancer types. Variant detection tools such as GATK-gCNV [48] and CNVkit [17] can distinguish somatic from germline CNVs by comparing tumor against normal samples. However, even when normal exist, they lack cancer-type-specific baselines to distinguish somatic drivers from tissue-specific germline CNVs. For analyzing and clinically interpreting CNVs, an important resource are annotation pipelines for structural variants like AnnotSV [19] or the StrVCTVRE web server [1]. Role of dosage sensitivity: [20] In order to predict the pathogenicity of individual CNVs, several methods have been developed [21, 22, 23, 26, 25]. These allow the evaluation of individual copy number events, but not the identification of cancer-specific patterns.

Existing databases germline variation databases, e.g. 1000Genomes, gnomAD, dbVar and DGV, catalog either CNVs of healthy or cancer, lacking integrated pan-cancer/healthy comparison. CNVIntegrate [29] established a CNV database including cancer and healthy data, but lacks multi-ethnic population-specific baselines and misses a statistical comparison framework. CNAapp [28] introduces a web application for CNV analysis, yet focuses on analyzing existing CNV data and does not allow to dynamically analyze specific genomic regions. In addition, correlations of CNVs with other data have been little studied to date in a pan-cancer scope: Results from prior studies suggest a significant correlation of CNVs and gene expression [30], and the CNV-specific gene vulnerability.

We introduce PanCNV-Explorer, a comprehensive, harmonized database of copy number events integrating both healthy and cancerous CNVs. PanCNV-Explorer presents an interactive web interface that enables the real-time exploration of CNA landscapes with statistical validation through out different cancer types, including the novel analyses dosage-specific CNV burden, and gene vulnerability analysis based on CRISPR screens. PanCNV-Explorer introduced a gene-level tolerance metric quantifying deviation from healthy copy number states to prioritize pathogenic somatic events. We introduce a novel gene CNV tolerance metric validated against healthy CNV data, distinguishing somatic from germline CNVs. PanCNV-Explorer established a new paradigm for CNA-driven biomarker discovery, where the distinction between driver and passenger is defined not by recurrence alone, but by deviation from a healthy genomic baseline.

2 Methods

2.1 Datasets

To retrieve cancer-specific CNAs, we downloaded copy number segmentation files from The Cancer Genome Atlas (TCGA) [31] from 33 cancer types using the TCGAbiolinks package [32]. Normal samples were filtered out and classified as normal tissue. For cancer cell line data, we downloaded segment-level and gene-level copy number data from DepMap [3] (release 25Q3). Both TCGA and Depmap samples with a segment mean <0.2 were classified as deletions, and those with >0.2 as amplifications, aligning with thresholds that have been employed in prior works [2].

To retrieve CNAs for healthy and normal tissue, we downloaded the structural variants calls of dbVAR [33] from https://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/by_assembly/GRCh38/vcf/GRCh38.variant_call.all.vcf.gz. Files from the DGV database [34] were downloaded from https://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg19. To derive copy number variants and structural variants from the gnomAD database (v4), we downloaded non neuro controls structural variants from https://storage.googleapis.com/gcp-public-data--gnomad/release/4.1/genome_sv/gnomad.v4.1.sv.non_neuro_controls.sites.vcf.gz and exome CNV non neuro controls from https://storage.googleapis.com/gcp-public-data--gnomad/release/4.1/exome_cnv/gnomad.v4.1.cnv.non_neuro_controls.vcf.gz.

To bring all data files into a standardized data format that is easy to annotate, we converted all CNV files into the standard Variant Call Format (VCF) [35] terminology for copy number events, including the start position in the POS column, and adding the additional features in the INFO column, such as the END and SVTERM features. All downloaded databases were stored in the genome assembly GRCh38.

2.2 Data harmonization

Cancer type harmonization.

To harmonize the cancer types between the TCGA and Depmap datasets, the 32 TCGA cancer entities were retained, and we created a mapping that aligned the Depmap data according to the *Depmap-ModelType* columns with the TCGA cancer entities. DepMap cancer types that could not be mapped to the TCGA classification were not included.

CNV classification. To classify the TCGA CNVs, we determined all CNVs with a segment mean

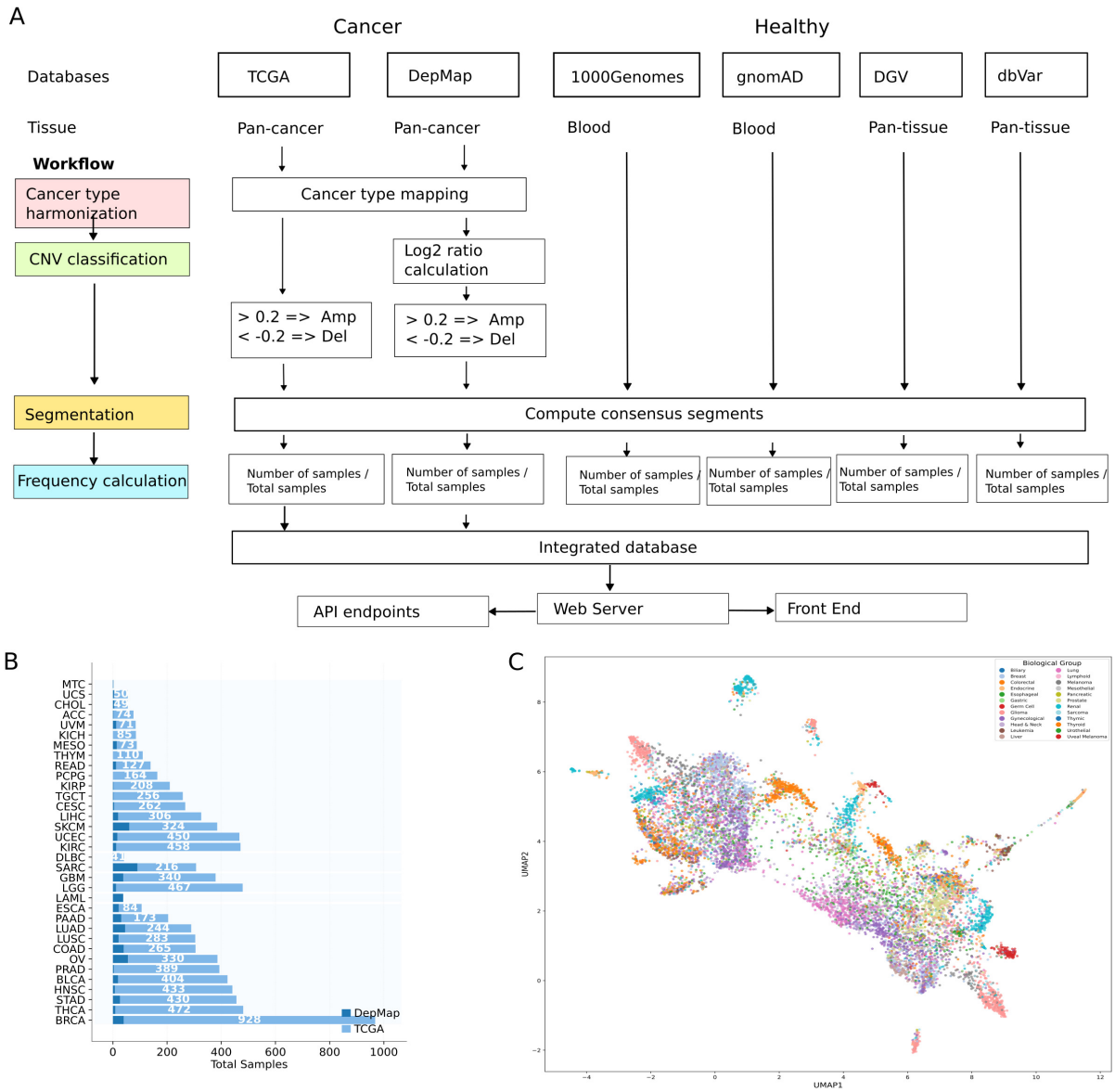

Figure 1: PanCNV-Explorer dataset construction workflow. A Segment-level copy number varation data was aggregated from TCGA and DepMap, as well as normal tissues from 1000Genomes, gnomAD, DGV and dbVar. After cancer type harmonization, the cancerous CNVs were classified in amplifications and deletions. After that, we applied 3 segmentation algorithms, including the computation of consensus segments, fixed-size bins, and gene-level CNV segments. Finally, we calculated the frequencies of each CNV for each cancer type. **B** Number of cancer samples per cancer type and data source, and **C** UMAP clustering of the cancer CNV data. Cancer-specific CNV clusters can especially be identified for germ cell, renal, and thyroid cancer.

smaller than 0.2 as deletions, and those with a segment mean above 0.2 as amplifications. All values in between were classified as neutral. For the Depmap dataset, the absolute CNV copy numbers were log2-transformed to compute the segment mean, and then classified in the same way as the TCGA CNVs. For the data set from 1000Genomes, DGV, dbVar, and gnomAD; the classification into duplications and deletions that had already been made was adopted. Duplicates were removed from the DGV dataset for results with identical outer start and end positions and sample identifiers.

Data format harmonization. To consolidate the CNV data from all data sets, we mapped

all data to an adapted BED format. For dbVar, all non-human data and all aggregated samples were filtered out.

Segmentation and clustering. To harmonize the genomic positions of the CNV segments, a consensus segmentation was generated by aggregating all unique genomic breakpoints from the input CNV segments. For each chromosome, non-overlapping intervals were computed from the sorted breakpoints, where consecutive boundaries form new segments spanning from breakpoint to breakpoint. Given a set of CNV calls $C = (c_i, s_i, e_i)_{i=1}^{n}$ where c_i denotes chromosome, s_i start position, and e_i end position, we define the consensus segmentation as follows: For each chromosome c, let $B_c = s_i | c_i = c \cup e_i + 1 | c_i = c$ be the set of breakpoints. Sorting B_c yields $B_c = b_1 < b_2 y ... < b_m$, and the consensus segments are: $S_c = (b_j, b_{j+1} - 1) | j = 1, ..., m - 1$

To reduce technical noise, we apply two refinement steps: **1) Breakpoint clustering** with parameter δ, where breakpoints were grouped into clusters where consecutive points differ by $\leq \delta$, then replaced each cluster C_k with its median $b_k = \textbf{median}(C_k)$. The clustered breakpoints $B_c = \{b_1 < ... < b_p\}$ yield clustered segments $S_c = \{(b_j, b_{j+1} - 1)\}_{j=1}^{p-1}$

Short segment merging. We then merged consecutive segments in $S_c = \{(s_j, e_j)\}_{j=1}^{q}$ where $e_j - s_j + 1 < L$ with subsequent segments, calculating the final segments S_c. The final consensus segments were computed with $\delta = 50$ and $L = 10000$, resulting in a dataset of 200.000 consensus segments.

CNV frequency calculation The frequency of a CNV in a specific tissue type was determined by calculating the overlaps with the consensus segments for each cancer type and healthy tissue, for duplications and deletions each. For aggregated samples, including gnomAD and DGV, we extracted the precomputed frequency and adjusted it for the novel CNV segments. Let $s \in S$ represent a unified segment with genomic coordinates (c_s, b_s, e_s), where c_s is the chromosome identifier, and $[b_s, e_s)$ denotes the half-open genomic interval. For each tissue type $t \in \mathcal{T}$ and mutation type $m \in \mathcal{M}$ (where $\mathcal{M} = \{\text{amplification}, \text{deletion}\}$). We then calculated the number of samples per CNV by summing the sample counts from all overlapping CNVs, then calculated the CNV frequency for each tissue and mutation type relative to the total samples:

$$\begin{aligned} C_{s,t,m} = \{c \in C \mid & c.\text{tissue} = t, \\ & c.\text{mut_type} = m, \\ & c.\text{chrom} = c_s, \\ & [c.\text{start}, c.\text{end}) \cap [b_s, e_s) \neq \emptyset\} \end{aligned} \tag{1}$$

$$N_{s,t,m} = \sum_{c \in C_{s,t,m}} c.\text{num_sample} \tag{2}$$

$$F_{s,t,m} = \frac{N_{s,t,m}}{N_{\text{total},t}} \tag{3}$$

For aggregated population databases (gnomAD, DGV and dbVar), we extracted the precomputed frequencies and remapped them to our unified segments using the same binary overlap approach. This methodology preserves the sample-level resolution of the original data while enabling direct comparison across diverse datasets. Gene-level CNV frequencies were calculated using a threshold of 50% of minimum overlap between a CNV and a gene.

2.3 Database construction

The CNA database was created using PostgreSQL by generating the database from the original table-formatted files with the SQLAlchemy package [36].

2.4 The web server and front end

The database was built using PostgreSQL v18.10. The PanCNV-Explorer web server was built using Python Flask. The web front end for PanCNV-Explorer was implemented using AngularJS [37] (v). The genome browser tracks were implemented using JBrowse2 [38]. Interactive visualizations were generated using Plotly [39]. The API documentation was realized with Swagger UI API documentation. All modules, including the database, server and front end, were containerized using Docker [40].

2.5 CNV annotation modules

To annotate copy number alterations, we created multiple web services providing CNA annotation for a variety of features. Affected genes of CNA regions were retrieved using the SeqCAT [41]. Gene start and end positions were retrieved by extracting protein-coding genes from GENCODE (v49) [42]. Data on gene dosage sensitivity, including haploinsufficiency (HI) and triplosensitivity (TS) scores, were retrieved from the ClinGen database [43]. Oncogenes and tumor suppressor genes classifications were downloaded from the Network of Cancer Genes database (NCG 7.0) [44]. To identify CNAs in API requests, we used the annotation <chr>:<start>-<end><change>, e.g. chr1:258946-714338_DEL. To provide annotations of the pathogenicity of CNAs, we developed annotation modules based on the methods CNVoyant [21], X-CNV [22], dbCNV [23], ISV [24], TADA [26], ClassifyCNV [25]. Gene constraint scores, including pLI and LOEUF scores, were

downloaded from gnomAD v4 at https://storage.googleapis.com/gcp-public-data--gnomad/release/4.1/constraint/gnomad.v4.1.constraint_metrics.tsv. The new annotation modules were implemented as additional web services and integrated into the variant interpretation frameworks Onkopus and AdaGenes. To extract clinical implications of CNVs, amplifications and deletions of affected genes were extracted from the CIViC database [45].

Tissue-Specific and pan-cancer gene CNV Intolerance Score

The tissue-specific score computes the CNV intolerance for a specific gene in a specific cancer type for a specific CNV type. For a cancer type T, gene g, and CNV type C (DEL or AMP), we are calculating the tissue-specific CNV intolerance $I_g^{T,C}$ using an exact Poisson test:

$$I_g^{T,C} = -\log_{10}\left(\max\left(\mathbb{P}\big(X \leq O_g^{T,C} \mid X \sim \text{Poiss}(E_g^{T,C})\big),\ 1e^{-300}\right)\right) \tag{4}$$

where $O_g^{T,C}$ is the observed CNV count for gene g in cancer type T, $E_g^{T,C} = r_a^{T,C} \times L_g$ is the expected CNV count, $r_a^{T,C}$ is the regional CNV rate for chromosome arm a in cancer type T, and L_g is the length of gene g, and $\mathbb{P}(\cdot)$ is the cumulative distribution function of the Poisson distribution.

In addition, we calculate the pan-cancer gene CNV intolerance score. For gene g and CNV type C across k valid cancer types:

$$I_g^{C} = -\log_{10}\left(\text{FisherCombined}\left(\{p_g^{T_1,C}, p_g^{T_2,C}, \ldots, p_g^{T_k,C}\}\right)\right) \tag{5}$$

where

$$p_g^{T,C} = \mathbb{P}(X \leq O_g^{T,C} \mid X \sim \text{Poisson}(E_g^{T,C}))$$

and $\text{FisherCombined}(\{p_1, p_2, \ldots, p_k\}) = 1 - F_{\chi^2_{2k}}\left(-2\sum_{i=1}^{k}\ln(p_i)\right)$ where $F_{\chi^2_{2k}}$ is the cumulative distribution function of the χ^2 distribution with $2k$ degrees of freedom.

3 Results

3.1 A harmonized resource for somatic and germline CNV analysis

PanCNV-Explorer provides a comprehensive catalog of copy number variation in the human genome across various cancer types as well as normal tissue. As data sources, we compiled segment-level copy number variant data from TCGA and DepMap, as well as normal tissue data from 1000Genomes, gnomAD, DGV, and dbVar, followed by the classification of the CNVs derived from TCGA and DepMap (Figure 1A). To enable comprehensive analysis of the CNV data, we created three versions of the database: First, we generated the database according to consensus segments, where we calculated the occurring CNV segments from the existing start and end positions and then applied clustering to reduce the number of CNV segments to 200.000. Second, we computed fixed-sized bins of 10.000 base pairs across the whole genome. And third, we generated gene-level CNV segments, were CNV start and end positions corresponded to the positions of protein-coding genes. We then calculated the frequencies per CNV segment for each cancer and mutation type (amplifications/deletions) in order to identify cancer types of genes significantly affected by copy number variation. Samples were aggregated across 33 cancer types (Figure 1B), while dimensionality reduction through UMAP revealed significant differences among frequencies across cancer types (Figure 1C).

3.2 An interactive web resource for pan-cancer CNV analysis

The PanCNV-Explorer web interface provides the functionality to dynamically analyze and visualize this dataset across genomic regions. It provides the options to explore CNVs 1) by defining a genomic region, or 2) by gene, and one or multiple cancer types (Figure 2A). PanCNV-Explorer then dynamically generates an interactive overview of the CNV amplification and deletion frequency in the selected region, providing a comparison to normal tissue and the location of known tumor suppressor and oncogenes. It is possible to overlay any number of cancer types, and genome regions can be interactively selected and zoomed in on in order to dynamically analyze cancer-specific discrepancies. For instance, to explore the cancer-specific environment around the tumor suppressor *TP53*, the region can be selected within the interactive plot, followed by a zoom-in (Figure 2B). Transcripts, functional regions,

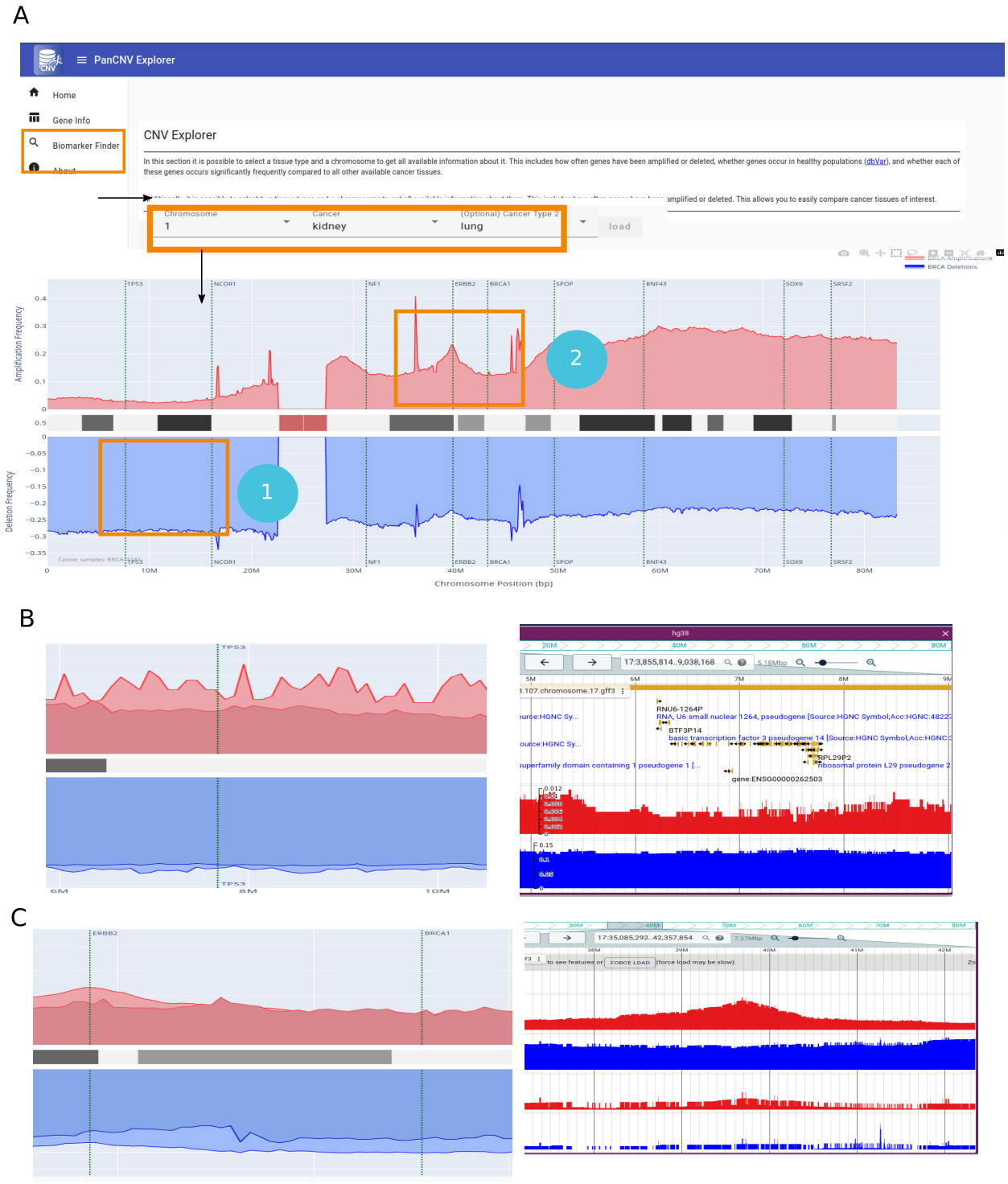

Figure 2: The PanCNV-Explorer web interface. **(A)** Users can select between exploring a specific genomic region, or a gene. After selecting a genomic region and one or multiple cancer types, users can directly compare CNV-specific characteristics across cancer types. The CNV frequency visualization displays amplifications (red) and deletions (blue). Known tumor suppressor and oncogenes are marked with dashed lines. **(B)** Detailed view of the region surrounding TP53 in breast cancer and aligned transcripts and burden. **C** Detailed view of ERBB2 in breast and lung cancer. Direct superimposition of CNV frequencies allows direct comparison of CNV frequencies in different cancer types.

gene expression, and CRISPR gene essentiality validation can be additionally explored using the genome browser. Complementary, amplifications within the region surrounding *ERBB2* exhibit a significantly increased CNV frequency in breast and lung cancer (Figure 2B).

3.3 Tissue-context dependent expression effects

In addition to genomic data analysis, PanCNV-Explorer integrates other omics data to analyze pan-cancer structural variation within the context of multi-omics analyses. In addition to genetic tracks, the web interface provides the functionality to display gene expression per tissue.

3.4 API endpoints for CNV interpretation

All modules that were used to annotate the database are publicly available via API endpoints. The analyses and plots presented by the web interface are available as publicly available API endpoints. The module for annotation CNVs were added to the Onkopus framework and are publicly available through API endpoints. To interpret individual CNVs biologically and clinically, PanCNV-Explorer generates links to the variant interpretation framework Onkopus [46]. The completed database is downloadable via the web interface in BED and VCF format.

4 Discussion

Copy number variants are pervasive drivers of cancer, yet their systematic analysis in different tissues in cancer types is hindered by fragmented data resources and a lack of integration between cancer-specific alterations and population-level variation. PanCNV-Explorer bridges this gap by presenting a harmonized CNV database consisting of pan-cancer and normal tissue datasets. Coupled with an interactive web front end, PanCNV-Explorer provides a dynamic CNV analysis and exploration across cancer types from descriptive frequency mapping to mechanistic and therapeutic insight generation.

A particular challenge in the merging of CNV data presents the harmonization of CNV segments, as segments that are too large can generalize important peaks, while too small segments can result in high data fragmentation. To enable direct comparison of copy number variant frequencies between pan-cancer and healthy datasets, we implemented a consensus segmentation approach that clusters breakpoints across all samples within a biologically meaningful tolerance by setting a minimum segment size, followed by a clustering that reduced the number of CNV segments. These clustering steps reduced the number of unified CNV segments to 200.000, while preserving biologically meaningful CNV boundaries. This approach effectively captures regions of non-random copy number alteration that likely represent driver events, as it accounts for the natural variation in CNV breakpoints across samples while maintaining sensitivity to both broad chromosomal alterations and focal events.

The PanCNV-Explorer database is therefore available in three different formats: 1) CNV segments calculated on the basis of biologically meaningful segments, 2) CNV segments with fixed-sized bins, as well as 3) segments mapped to protein-coding genes. In the first case, all existing CNV segments were aggregated and then clustered again, generating approximately 200.000 CNV segments. Concurrently, the calculation of CNV frequencies based on gene start and end positions facilitates the identification of gene-specific CNV frequencies and their subsequent analysis in various cancer types.

In addition, PanCNV-Explorer integrates CNV frequencies from normal tissue into its database, thereby establishing a healthy baseline that can be utilized to compare CNV frequencies across various cancer types with population CNV frequency. In this way, PanCNV-Explorer supports cancer CNV analysis by making the healthy baseline diagnostically active, by answering the question whether a CNV is cancer-relevant, or a common population variant.

PanCNV-Explorer thus provides extended capabilities compared to GISTIC2.0, which focuses on pan-cancer and a healthy baseline, while not taking into account natural copy number variation. The web front end of PanCNV-Explorer provides the functionality to directly search for copy number variants, assess their pathogenicity and find potential therapeutic options.

A significant innovation in the analyses offered by PanCNV-Explorer is the comparison between all major cancer types, which allows cancer-specific biomarkers to be identified and analyzed. To measure the vulnerability of a gene to copy number variation, PanCNV-Explorer introduces a gene CNV tolerance score computing a metric to quantify a gene's resistance to CNVs. Our innovation is defining a gene-specific tolerance threshold based on healthy population copy number variation, then quantifying how cancer exceeds this threshold to reveal cancer-specific vulnerabilities. uniquely connects population CNV

tolerance with cancer deviation, and functional consequence Compared to prior gene scores, such as the pLI score or the LOEUF score [4], who focus on the question whether a gene is essential for human development and health, this score is focusing on the gene essentiality for cancer cell survival. As a result, this novel score can be used as a complement for For instance, a gene can be essential for development (high pLI score), but dispensable in cancer (low CNV intolerance) pLI and LOEUF are pan-population metrics with no tissue specificity, while the CNV score can provide tissue-specific intolerance profiles, which is uniquely valuable for precision oncology and does not exist in current intolerance metrics. In addition, the CNV score introduces a novel variant type focus for gene assessment, as pLI and LOEUF focus on small sequence variants, including single nucleotide variants and small insertions and deletions.

The dosage-specific CNV burden analysis introduced presented a methodological advancement in cancer genomics by moving beyond traditional frequency-based approaches to incorporate the functional directionality of CNVs. This approach quantifies tumor suppressor gene burden through copy number losses in HI-scores genes and oncogene burden through amplifications in TS-scores genes, thereby capturing the distinct biological consequences of these opposing genomic events. The cross-cancer comparison identifies previously underappreciated relationships between cancer types based on their CNV-driven pathogenesis mechanisms rather than tissue of origin alone.

PanCNV-Explorer shows CNV frequencies in cancer that significantly exceed what we observe in healthy populations, after accounting for regional variation in baseline CNV tolerance.

As a limitation, it has to be noted that differences in the CNV data arise due to the different preprocessing of the data sets. On the one hand, the CNVs were called using different software tools, including GATK [48] and CNVnator [49]. In addition, in the case of normal tissue, the data is only classified, while in the case of gnomAD and dbVar, it is already aggregated into harmonized CNV segments.

The analyses available in the web front end are also available as Application Programming Interfaces (API) endpoints, making the analyses and visualizations available in other applications as well.

To summarize, PanCNV-Explorer provides a comprehensive pan-cancer and healthy database for copy number variations in the human genome, associated with a versatile web front end that enables the dynamic analysis and exploration of CNV regions across various cancer types. By filtering false positives, prioritizing CRISPR-validated targets, and enabling clinical stratification, PanCNV-Explorer transforms copy number events in the human genome from observational artifacts into actionable therapeutic insights. As cancer genomics shifts towards clinical utility, resources that distinguish true drivers from noise will be indispensable. We are thus confident that PanCNV-Explorer will be a valuable resource for clinicians and researchers, that will significantly help to in understanding copy number variation cancer drivers and improve personalized therapies for cancer patients.

5 Data availability

A public instance of PanCNV-Explorer is available at https://mtb.bioinf.med.uni-goettingen.de/pancnv-explorer/. The source code of PanCNV-Explorer, including the database generation, the web server, the web front end and the CNA annotation modules, is available at https://gitlab.gwdg.de/MedBioinf/mtb/cnv-database. The annotation modules for CNVs and structural variants were added to the Onkopus framework and are publicly available through APIs at https://mtb.bioinf.med.uni-goettingen.de/onkopus/api. All data sources of the copy number alteration annotation are publicly available.

6 Acknowledgements

The authors thank the members of the molecular tumor board of the University Medical Center Göttingen, especially Alexander König, Li Beißbarth, and Kirsten Reuter-Jessen, for their support and feedback. The authors thank the International Max Planck Research School for Genome Science (IMPRS-GS).

7 Author contributions

JD, NSK and KK conceptualized the methodology and architecture. NSK and ARK constructed the database. NSK, KK and ARK implemented the web server. NSK implemented the annotation modules. NSK and ARK conducted the analysis. NSK, KK and ARK implemented the front end. NSK wrote the manuscript. All authors approved the manuscript.

8 Competing interests

No competing interest is declared.

9 Funding

This work was supported by the Gemeinsamer Bundesauschuss (01NVF20006), the Volkswagen Foundation (11-76251-12-1/19), the Deutsche Krebshilfe (70114018), the Deutsche Forschungsgemeinschaft (KFO5002) and the Bundesministerium für Bildung und Forschung (BMBF) (01KD2437, 01KD2401B, 01KD2208A, 01KD2414A).

References

[1] Sharo, A., Hu, Z., Sunyaev, S. & Brenner, S. StrVCTVRE: A supervised learning method to predict the pathogenicity of human genome structural variants. *The American Journal Of Human Genetics*. **109**, 195-209 (2022)

[2] Laddha, S., Ganesan, S., Chan, C. & White, E. Mutational landscape of the essential autophagy gene BECN1 in human cancers. *Molecular Cancer Research*. **12**, 485-490 (2014)

[3] Tsherniak, A., Vazquez, F., Montgomery, P., Weir, B., Kryukov, G., Cowley, G., Gill, S., Harrington, W., Pantel, S., Krill-Burger, J. & Others Defining a cancer dependency map. *Cell*. **170**, 564-576 (2017)

[4] Karczewski, K., Francioli, L., Tiao, G., Cummings, B., Alföldi, J., Wang, Q., Collins, R., Laricchia, K., Ganna, A., Birnbaum, D. & Others The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. **581**, 434-443 (2020)

[5] Zarrei, M., MacDonald, J., Merico, D. & Scherer, S. A copy number variation map of the human genome. *Nature Reviews. Genetics*. **16**, 172-183 (2015)

[6] Baker, T., Waise, S., Tarabichi, M. & Van Loo, P. Aneuploidy and complex genomic rearrangements in cancer evolution. *Nature Cancer*. **5**, 228-239 (2024)

[7] Sosinsky, A., Ambrose, J., Cross, W., Turnbull, C., Henderson, S., Jones, L., Hamblin, A., Arumugam, P., Chan, G., Chubb, D. & Others Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. *Nature Medicine*. **30**, 279 (2024)

[8] Besedina, E. & Supek, F. Copy number losses of oncogenes and gains of tumor suppressor genes generate common driver mutations. *Nature Communications*. **15**, 6139 (2024)

[9] Shlien, A. & Malkin, D. Copy number variations and cancer. *Genome Medicine*. **1**, 62 (2009)

[10] Chen, X., Agustinus, A., Li, J., Dibona, M. & Bakhoum, S. Chromosomal instability as a driver of cancer progression. *Nature Reviews Genetics*. **26**, 31-46 (2025)

[11] Steele, C., Abbasi, A., Islam, S., Bowes, A., Khandekar, A., Haase, K., Hames-Fathi, S., Ajayi, D., Verfaillie, A., Dhami, P. & Others Signatures of copy number alterations in human cancer. *Nature*. **606**, 984-991 (2022)

[12] Drews, R., Hernando, B., Tarabichi, M., Haase, K., Lesluyes, T., Smith, P., Morrill Gavarró, L., Couturier, D., Liu, L., Schneider, M. & Others A pan-cancer compendium of chromosomal instability. *Nature*. **606**, 976-983 (2022)

[13] Cascione, L., Aresu, L., Baudis, M. & Bertoni, F. DNA Copy Number Changes in Diffuse Large B Cell Lymphomas. *Frontiers In Oncology*. **10** pp. 584095 (2020)

[14] Beroukhim, R., Mermel, C., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J., Dobson, J., Urashima, M. & Others The landscape of somatic copy-number alteration across human cancers. *Nature*. **463**, 899 (2010)

[15] Smith, P., Bradley, T., Gavarró, L., Goranova, T., Ennis, D., Mirza, H., De Silva, D., Piskorz, A., Sauer, C., Al-Khalidi, S. & Others The copy number and mutational landscape of recurrent ovarian high-grade serous carcinoma. *Nature Communications*. **14**, 4387 (2023)

[16] Babadi, M., Fu, J., Lee, S., Smirnov, A., Gauthier, L., Walker, M., Benjamin, D., Zhao, X., Karczewski, K., Wong, I. & Others GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nature Genetics*. **55**, 1589-1597 (2023)

[17] Talevich, E., Shain, A., Botton, T. & Bastian, B. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Computational Biology*. **12**, e1004873 (2016)

[18] Corpas, M., Bragin, E., Clayton, S., Bevan, P. & Firth, H. Interpretation of genomic copy number variants using DECIPHER. *Current Protocols In Human Genetics*. **72**, 8-14 (2012)

[19] Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H. & Muller, J. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics*. **34**, 3572-3574 (2018)

[20] Rice, A. & McLysaght, A. Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nature Communications*. **8** pp. 14366 (2017)

[21] Schuetz, R., Ceyhan, D., Antoniou, A., Chaudhari, B. & White, P. CNVoyant a machine learning framework for accurate and explainable copy number variant classification. *Scientific Reports*. **14**, 22411 (2024)

[22] Zhang, L., Shi, J., Ouyang, J., Zhang, R., Tao, Y., Yuan, D., Lv, C., Wang, R., Ning, B., Roberts, R. & Others X-CNV: genome-wide prediction of the pathogenicity of copy number variations. *Genome Medicine*. **13** pp. 1-15 (2021)

[23] Lv, K., Chen, D., Xiong, D., Tang, H., Ou, T., Kan, L. & Zhang, X. dbCNV: deleteriousness-based model to predict pathogenicity of copy number variations. *BMC Genomics*. **24**, 131 (2023)

[24] Gažiová, M., Sládeček, T., Pös, O., Števko, M., Krampl, W., Pös, Z., Hekel, R., Hlavačka, M., Kucharik, M., Radvanszky, J. & Others Automated prediction of the clinical impact of structural copy number variations. *Scientific Reports*. **12**, 555 (2022)

[25] Gurbich, T. & Ilinsky, V. ClassifyCNV: a tool for clinical annotation of copy-number variants. *Scientific Reports*. **10**, 20375 (2020)

[26] Hertzberg, J., Mundlos, S., Vingron, M. & Gallone, G. TADA—a machine learning tool for functional annotation-based prioritisation of pathogenic CNVs. *Genome Biology*. **23**, 67 (2022)

[27] Mermel, C., Schumacher, S., Hill, B., Meyerson, M., Beroukhim, R. & Getz, G. GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*. **12** pp. 1-14 (2011)

[28] Franch-Exposito, S., Bassaganyas, L., Vila-Casadesús, M., Hernández-Illán, E., Esteban-Fabro, R., Dıaz-Gay, M., Lozano, J., Castells, A., Llovet, J., Castellvi-Bel, S. & Others CNApp, a tool for the quantification of copy number alterations and integrative analysis revealing clinical implications. *Elife*. **9** pp. e50267 (2020)

[29] Chattopadhyay, A., Teoh, Z., Wu, C., Juang, J., Lai, L., Tsai, M., Wu, C., Lu, T. & Chuang, E. CNVIntegrate: the first multi-ethnic database for identifying copy number variations associated with cancer. *Database: The Journal Of Biological Databases And Curation*. **2021** pp. baab044 (2021)

[30] Shao, X., Lv, N., Liao, J., Long, J., Xue, R., Ai, N., Xu, D. & Fan, X. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Medical Genetics*. **20** pp. 175 (2019)

[31] Weinstein, J., Collisson, E., Mills, G., Shaw, K., Ozenberger, B., Ellrott, K., Shmulevich, I., Sander, C. & Stuart, J. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*. **45**, 1113-1120 (2013)

[32] Colaprico, A., Silva, T., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T., Malta, T., Pagnotta, S., Castiglioni, I. & Others TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*. **44**, e71-e71 (2016)

[33] Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G. & Others DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Research*. **41**, D936-D941 (2012)

[34] MacDonald, J., Ziman, R., Yuen, R., Feuk, L. & Scherer, S. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*. **42**, D986-D992 (2014)

[35] Danecek, P., Auton, A., Abecasis, G., Albers, C., Banks, E., DePristo, M., Handsaker, R., Lunter, G., Marth, G., Sherry, S. & Others The variant call format and VCFtools. *Bioinformatics*. **27**, 2156-2158 (2011)

[36] Bayer, M. SQLAlchemy. *The Architecture Of Open Source Applications*. **2** (2012)

[37] Jain, N., Mangal, P. & Mehta, D. AngularJS: A modern MVC framework in JavaScript. *Journal Of Global Research In Computer Science*. **5**, 17-23 (2014)

[38] Diesh, C., Stevens, G., Xie, P., De Jesus Martinez, T., Hershberg, E., Leung, A., Guo, E., Dider, S., Zhang, J., Bridge, C. & Others JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biology*. **24**, 74 (2023)

[39] Sievert, C. Interactive web-based data visualization with R, plotly, and shiny. (Chapman,2020)

[40] Merkel, D. & Others Docker: lightweight linux containers for consistent development and deployment. *Linux J.* **239**, 2 (2014)

[41] Kornrumpf, K., Kurz, N., Drofenik, K., Krauß, L., Schneider, C., Koch, R., Beißbarth, T. & Dönitz, J. SeqCAT: Sequence Conversion and Analysis Toolbox. *Nucleic Acids Research*. **52**, W116-W120 (2024)

[42] Frankish, A., Carbonell-Sala, S., Diekhans, M., Jungreis, I., Loveland, J., Mudge, J., Sisu, C., Wright, J., Arnan, C., Barnes, I. & Others GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Research*. **51**, D942-D949 (2023)

[43] Rehm, H., Berg, J., Brooks, L., Bustamante, C., Evans, J., Landrum, M., Ledbetter, D., Maglott, D., Martin, C., Nussbaum, R. & Others ClinGen—the clinical genome resource. *New England Journal Of Medicine*. **372**, 2235-2242 (2015)

[44] Dressler, L., Bortolomeazzi, M., Keddar, M., Misetic, H., Sartini, G., Acha-Sagredo, A., Montorsi, L., Wijewardhane, N., Repana, D., Nulsen, J. & Others Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the Network of Cancer Genes (NCG) resource. *Genome Biology*. **23**, 35 (2022)

[45] Griffith, M., Spies, N., Krysiak, K., McMichael, J., Coffman, A., Danos, A., Ainscough, B., Ramirez, C., Rieke, D., Kujan, L. & Others CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics*. **49**, 170-174 (2017)

[46] Kurz, N., Kornrumpf, K., Tucholski, T., Drofenik, K., König, A., Beißbarth, T. & Dönitz, J. Onkopus: precise interpretation and prioritization of sequence variants for biomedical research and precision medicine. *Nucleic Acids Research*. pp. gkaf376 (2025)

[47] Cornish, A., Gruber, A., Kinnersley, B., Chubb, D., Frangou, A., Caravagna, G., Noyvert, B., Lakatos, E., Wood, H., Thorn, S. & Others The genomic landscape of 2,023 colorectal cancers. *Nature*. **633**, 127-136 (2024)

[48] Babadi, M., Fu, J., Lee, S., Smirnov, A., Gauthier, L., Walker, M., Benjamin, D., Zhao, X., Karczewski, K., Wong, I. & Others GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nature Genetics*. **55**, 1589-1597 (2023)

[49] Abyzov, A., Urban, A., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*. **21**, 974-984 (2011)