Nucleic Acids Research, 2025, gkaf1242
https://doi.org/10.1093/nar/gkaf1242
Database issue

CNAScope: pan-cancer copy number aberration database with functional annotation and interactive visualization

Xikang Feng [1,2,*,†], Jieyi Zheng [1,†], Sisi Peng [1,†], Anna Jiang [3,†], Ka Ho Ng [3], Chengshang Lyu [3], Qiangguo Jin [1,*], Lingxi Chen [3,*]

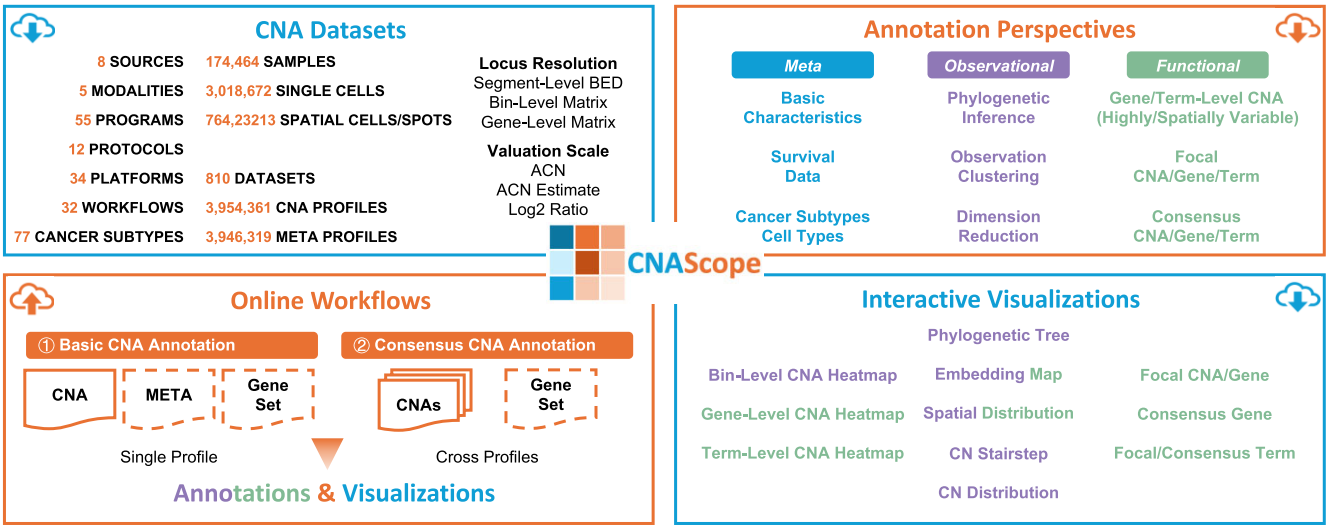
[1]School of Software, Northwestern Polytechnical University, Xi'an 710072, China
[2]Research & Development Institute, Northwestern Polytechnical University, Shenzhen 518063, China
[3]Department of Biomedical Sciences, College of Biomedicine, City University of Hong Kong, Hong Kong 999077, China

*To whom correspondence should be addressed. Email: lingxi.chen@cityu.edu.hk
Correspondence may also be addressed to Xikang Feng. Email: fxk@nwpu.edu.cn
Correspondence may also be addressed to Qiangguo Jin. Email: qgking@nwpu.edu.cn
†The first four authors should be regarded as Joint First Authors.

Abstract

Copy number aberrations (CNAs) are critical drivers of genomic diversity in oncology, where recurrent CNAs frequently underlie tumorigenesis. However, existing public resources are limited in their somatic CNA specificity, breadth across multiple data modalities, and support for recurrent CNAs with online functional annotation and interactive visualization. Here, we present CNAScope (https://cna.fengslab.com/), a database that curates and functionally annotates over 3 954 361 CNA profiles and 3 946 319 metadata from 810 datasets, 174 464 samples, 3 018 672 single cells, and 764 232 spatial cells/spots, spanning 77 cancer subtypes from eight data sources and 55 cancer initiatives and institutions. CNAScope offers downloadable CNA annotations and interactive visualizations at bin, gene, and pathway term levels, including phylogenetic inference, clustering, dimension reduction, and focal/consensus CNA detection. Users can explore data through interactive heatmaps, phylogenetic trees, embedding plots, CN charts, and focal/consensus plots, or upload and annotate their own CNAs in real time. In all, with its large curated data volume and rich annotation capabilities, CNAScope serves as a vital resource for accelerating cancer research.

Graphical abstract



Introduction

Copy number aberrations (CNAs)—large-scale somatic gains and losses of chromosomal segments—drive genomic diversity and play a pivotal role in cancer development [1]. CNAs can disrupt gene dosage, perturb regulatory networks, and alter malignant transformation [1, 2]. Advances in high-throughput genomics now enable systematic CNA detection across diverse experimental protocols, including

Received: August 18, 2025. **Revised:** October 7, 2025. **Accepted:** October 18, 2025

microarrays [3], bulk whole-genome/exome sequencing (WGS [4], WES [5]), single-cell DNA/RNA sequencing (scDNA-seq [6, 7], scRNA-seq [8]), spatial DNA sequencing [9], and spatial transcriptomics (ST) [10]. These technologies allow detailed profiling of CNA homogeneity and heterogeneity at the level of single cells, spatial cells/spots, and samples.

Interpreting CNA data requires effective annotation and visualization from two complementary perspectives: observation and function. The observational perspective focuses on individual samples, single cells, or spatial cells/spots, allowing the exploration of intra- and inter-group heterogeneity, such as differences among cancer subtypes or tumor clones. Analytical approaches—phylogenetic inference, clustering, and dimension reduction—help reveal these relationships [11, 12]. The functional perspective maps CNAs to genes and pathway terms, enabling interpretation of their biological significance [2]. Systematic annotation across multiple CNA profiles identifies recurrent aberrations, which are critical for pinpointing driver events and assessing clinical relevance [13]. Visualizations such as CNA heatmaps, phylogenetic trees, embedding maps, CN plots, focal CNA plots, and consensus gene plots are indispensable for the intuitive exploration and interpretation of CNA patterns at bin, gene, and term levels [14]. These tools bridge the gap between observation-level variation and functional impact, supporting comprehensive biological insight and clinical translation.

While several online portals and databases exist, they face notable limitations: (i) Some resources primarily focus on germline aberrations or capture somatic aberrations relevant to cancer but do not specifically highlight CNAs [15–20]. (ii) Most databases are limited to a single data modality—either bulk DNA [15, 16, 18, 19], single-cell [7, 21], or transcriptomics [22]. (iii) Advanced annotation features like focal and consensus event detection are rarely highlighted [7, 16, 18, 21, 22].

To address these limitations, we present CNAScope (https://cna.fengslab.com/), a database that collects, curates, and annotates bulk, single-cell, and spatial CNA profiles and metadata across 810 datasets, 174 464 samples, 3 018 672 single cells, and 764 232 spatial cells/spots for 77 cancer subtypes. Drawing on eight online resources, CNAScope encompasses 55 cancer genomics initiatives and institutions, holding data generated by 12 sequencing protocols and 34 platforms. We curated CNA profiles derived from 32 computational workflows. CNAScope provides comprehensive, downloadable annotations for both CNA profiles and metadata, including phylogenetic trees, observation clusters, gene- and term-level CNAs, focal CNAs/genes/terms from single CNA profile, and consensus CNAs/genes/terms across multiple CNA profiles from the same dataset. Furthermore, CNAScope features interactive visual tools for database and annotation result exploration, as well as streamlined online annotation workflows for users' newly generated CNA profiles, supporting both exploratory and hypothesis-driven research.

Materials and methods

Dataset collection

We systematically searched public repositories using the keywords "cancer," "tumor," and "copy number aberration/variation" to identify relevant CNA datasets. Our collection spanned next-generation sequencing data—including both bulk WGS and WES—as well as probe-based microarrays. Following careful manual curation, quality control, filtering, and subdivision, we compiled CNA data from three resources: the cBioPortal [15], COSMIC [16], and GDC Portal [18]. Beyond bulk DNA, our collection was further enhanced by incorporating DNA and RNA data from published single-cell and spatially resolved datasets, including resources like HSCGD [21], scTML [22], 10x Genomics [7], NCBI GEO [23], and Broad SCP [24].

Detailed procedures for data download, processing, and quality control are described in Supplementary Methods 1.1.

Metadata curation

We manually curated cancer subtype annotations for all datasets. Our classification prioritized alignment with TCGA [25] and TARGET [26] subtype definitions (e.g. lung adenocarcinoma [LUAD] and lung squamous cell carcinoma [LUSC]) by reviewing the original disease descriptions, primary site, and metastatic status. When a dataset could not be directly mapped to a specific subtype, we assigned it to a broader cancer-type category (e.g. lung cancer).

Bulk datasets were annotated at the patient-sample level. Two primary clinical endpoints were recorded: overall survival (OS; time from diagnosis to death or last follow-up) and progression-free survival (PFS; time to recurrence/relapse/progression or last follow-up). Additional clinical information, such as tumor stage, grade, ethnicity, race, gender, and age, was collected when available. For single-cell and spatial datasets, observations were annotated as single cells or spatial cells/spots, including cell type, CNA confidence, donor identity, and malignancy status.

All metadata were independently reviewed by at least three authors to ensure accuracy. The finalized version is available at https://cna.fengslab.com/database.

Valuation scales and locus resolutions in raw CNA data

The collected raw CNA data from different sources vary in both valuation scale and locus resolution due to heterogeneous data modalities, sequencing protocols, and computational workflows.

We observe three valuation scales in raw CNA data: (i) absolute copy number (ACN), (ii) ACN estimate, and (iii) $\log_2$ ratio. ACN denotes the integer number of copies of a specific locus per cancer cell [27]. By convention: 0 means homozygous deletion, 1 is loss of heterozygosity, 2 refers to diploid, and >2 indicates amplifications. Estimated ACN refers to non-integer, floating-point values resulting from CNA inference. Such estimates arise due to normal-cell admixture (purity <1), subclonality, and measurement noise [27]. In this work, we call these fractional values "ACN estimates." Log_2 ratio is a widely used representation for CNA data from sequencing and microarrays. It is centered at 0 (neutral), with gains >0 and losses <0, providing a symmetric scale relative to the normal state [28].

We encounter three raw CNA locus resolutions: (i) segment-level BED, (ii) bin-level matrices, and (iii) gene-level matrices. BED-formatted segment calls (per observation), as in GDC [18] bulk DNA data, provide sample-specific CNA segments with unique breakpoints. Segmentation is not aligned across samples within a dataset, reflecting both inter-tumor

heterogeneity and per-sample calling. In contrast, matrix-formatted CNA data from other sources provides dataset-wide, aligned loci. Bin-level matrices use a common binning scheme across all observations via joint CNA calling with fixed genomic bins, thereby aligning loci within a dataset (though not necessarily across datasets that use different bin sets). Gene-level matrices report CNA on a shared gene index, yielding consistent loci within a dataset (though not necessarily across datasets that use different gene references).

Observational annotation

CNA binning for datasets with segment-level CNA

CNAScope supports three types of downstream observational annotations: phylogenetic inference, observation clustering, and dimension reduction. Most mainstream tools for these tasks expect CNA data in a matrix format (observations x bins/genes) [29–35]. Thus, for datasets that only have segment-level CNAs in BED format, we convert segments into bin-level matrices by tiling the genome into consecutive bins and aggregating segment values within each bin. We provide conversions at 200 kb, 500 kb, and 5 Mb, which are commonly used bin sizes in standard CNA calling workflows [6, 36–39]. When multiple segments overlap a bin, we compute a length-weighted average of the overlapping CN values. Length-weighted averaging is a standard approach that reflects the proportional genomic contribution of each segment and mitigates overemphasis of short intervals [40]. Because aggregation can combine heterogeneous segments, original ACN values are treated as continuous ACN estimates rather than integers post-binning, avoiding inappropriate discretization. Similarly, $\log_2$ ratios are retained on the $\log_2$ scale after binning.

In summary, for phylogenetic inference, observation clustering, and dimension reduction, we use raw bin- or gene-level CNA matrices when available; otherwise, we derive bin-level matrices from segment-level CNA data. Our quantitative assessment of bin-size effects (200 kb, 500 kb, 5 Mb) shows modest impact on these observational annotations, with moderate but acceptable concordance across resolutions, allowing users to choose their preferred granularity (Supplementary Results 2.1 and Supplementary Fig. 1).

Phylogenetic inference

To analyze phylogenetic relationships among observations (samples, single cells, or spatial cells/spots), we used a two-step approach based on the established pipeline [14]. First, hierarchical clustering with a weighted similarity metric was applied to generate a dendrogram. For 10x Chromium CNV scDNA-seq data, clustering results were directly extracted from Cell Ranger DNA `h5` files [7]; for other datasets, clustering was performed using `scipy.cluster.hierarchy` [29]. Next, we present the resulting dendrograms as interactive, zoomable trees, enabling detailed exploration of phylogenetic structures.

Observation clustering

CNAScope clusters observations using a default cluster number of $k = 10$, following the established pipeline [14]. The phylogenetic dendrogram is cut to produce k distinct clusters.

Dimension reduction

A suite of linear and manifold dimension reduction methods from the established pipeline [14], including PCA [30], ICA [31], NMF [32], UMAP [33], t-SNE [34], and PHATE [35], is used to generate two-dimensional embeddings for observations in each dataset.

Functional annotation

Gene-level CNA

For datasets lacking raw gene-level CNA data, we derive gene-level CNAs from the corresponding bin-level CNA matrices. Gene coordinates were retrieved from Ensembl [41] (GRCh37/hg19: https://grch37.rest.ensembl.org, GRCh38/hg38: https://rest.ensembl.org). Using BEDTools [42], we identified overlaps between gene regions and CNA bins. For each gene, the CN value was computed as the weighted average of CNs from all overlapping bins. This step was omitted for datasets where raw CNA data were already provided at the gene level.

Term-level CNA

Using the gene-level CNA matrices, we constructed term-level CNA matrices by averaging the CNs of all genes within each pathway term. Functional terms were sourced from MSigDB [43] (https://data.broadinstitute.org/gsea-msigdb/msigdb/release/2025.1.Hs/), including MSigDB-Hallmark, MSigDB-C2-KEGG (Kyoto Encyclopedia of Genes and Genomes [44]), MSigDB-C4-Computational, MSigDB-C5-GOBP (Gene Ontology Biological Process [45]), MSigDB-C6-Oncogenic, and MSigDB-C7-Immunologic collections.

Regarding value scales in gene- and term-level CNA data, after averaging, ACN is reported as an ACN estimate, and both the ACN estimate and $\log_2$-ratio retain their original scales.

Highly and spatially variable CN Bin/Gene/Term

In each dataset and at each locus resolution (bin-level, gene-level, and term-level), we computed the variance of CN for each locus across observations and selected the 1000 most variable loci. When spatial coordinates were available, we built a k–nearest-neighbor spatial weights matrix ($k = 10$) using `libpysal`. For each preselected highly variable locus, we calculated Global Moran's I and its two-tailed p-value (normal approximation) with `esda`. Within each resolution, p-values were adjusted for multiple testing via the Benjamini–Hochberg FDR procedure ($P < .05$) using `statsmodels`. The pipeline returns, for each resolution, (i) a table of top-variance loci and, when spatial coordinates are available, (ii) a table reporting Moran's I, raw and adjusted P-values, and a binary spatial-significance flag.

Focal CNA/Gene/Term

Focal CNAs (gains and losses) are recurrent, small-scale CNA events observed across samples, and focal genes are those overlapping these focal CNAs [46]. GISTIC2 [47], a widely adopted tool, was designed for bulk datasets with independent samples and expects CN values on the $\log_2$-ratio scale (microarray by default, with NGS supported after conversion to $\log_2$-ratio [48]).

As our benchmarking indicates that GISTIC2 is sensitive to bin size (Supplementary Results 2.1 and Supplementary Fig. S1), CNAScope runs GISTIC2 on bulk DNA datasets only when raw segment-level CNA calls (e.g. BED from microarray or WGS) are available. If needed, CN values x are converted to $\log_2$-ratios from ACN or ACN estimates using $\log_2(x/2)$.

For datasets with gene-level matrices—or for single-cell and spatial datasets where observations are cells or spots from a single patient—we omit focal annotation. These inputs do not meet GISTIC2 [47] assumptions, and dedicated methods for multi-sample focal detection in single-cell/spatial settings or for gene-level matrices are currently lacking.

Next, focally amplified and deleted functional terms are annotated via pathway over-representation analysis (ORA) on the GISTIC2 focal genes, using `gseapy.enrich` [49] with the six MSigDB [43] collections, as described in the "Term-Level CNA" subsection.

Consensus CNA/Gene/Term

Because focal events are small in scale, their detection, even within the same bulk DNA dataset, can be sensitive to copy-number differences arising from different sequencing protocols and computational workflows (Results-Case Study and Supplementary Fig. S2).

Thus, CNAScope integrates multiple focal calls from the same dataset to produce a unified set of focal CNAs, genes, and terms, referred to as consensus CNAs, consensus genes, and consensus terms. When only a single segment-level CNA profile is available, when datasets only provide gene-level CNAs, or for single-cell and spatial datasets, we omit the consensus step for the reasons outlined in the "Focal CNA/Gene/Term" subsection.

The consensus CNAs are called by intersecting focal CNA segments across protocols/workflows using BEDtools [42]. Similarly, CNAScope defines consensus genes by intersecting focal genes obtained from different protocols/workflows within the same bulk DNA dataset.

Consensus terms are annotated via ORA by enriching consensus genes against six MSigDB [43] collections, using the same procedure as for focal terms, as described in the "Focal CNA/Gene/Term" subsection.

Platform development

The CNAScope platform operates on an Ubuntu 24.04 LTS server, utilizing Nginx, Django, and PostgreSQL for backend services. The frontend is built with React and Next.js for efficient UI rendering and application logic, while dynamic and interactive visualizations are powered by D3.js, which enables custom data-driven elements such as heatmaps for CNA matrices (with features like color mapping, zooming, and tooltips) and hierarchical tree structures for exploring layered genomic relationships. Comprehensive tutorials are available on the web interfaces to guide users through their features and maximize ease of use.

The release version of CNAScope associated with this manuscript is Version 1.3 (2025.10.5). We will annually update the ontologies and reference databases to their latest versions.

Results

An extensive curated and annotated pan-cancer CNA resource in CNAScope

CNAScope hosts a comprehensive collection of curated and annotated 3 954 361 copy number alteration (CNA) profiles, accompanied by extensive metadata, from eight major online resources (Fig. 1): cBioPortal [15], COSMIC [16], GDC Portal [18], HSCGD [21], scTML [22], 10x Genomics [7], NCBI GEO [23], and Broad SCP [24]. These datasets are further supported by 55 large-scale cancer genomics initiatives, such as TCGA [25], TARGET [26], HCMI [50], ICGC [51], etc., as well as leading research institutions, including Memorial Sloan Kettering, the Broad Institute, etc.

In total, CNAScope collects 173 914 samples from 501 bulk datasets covering 70 cancer subtypes, generated using platforms such as Affymetrix microarrays [52], Illumina WGS/WES [53], etc. At single-cell resolution, CNAScope contains data from 2 457 425 cells spanning 27 cancer subtypes, derived from 35 scDNA-seq and 192 scRNA-seq datasets, utilizing protocols including 10x Chromium CNV [7], 10x Chromium [54], Smart-Seq2 [55], etc. Additionally, the database encompasses 1 325 479 spatial spots across 14 cancer subtypes, generated with three spatial DNA datasets and 79 spatial RNA datasets, covering protocols like Slide-DNA-Seq [9], 10x Visium [56], 10x Xenium [57], Slide-RNA-Seq v2 [58], etc.

CNAScope holds CNA profiles generated by 32 computational workflows (e.g. ascatNgs [4], inferCNV [8], Ginkgo [6], etc.). These profiles span three valuation scales—ACN, ACN estimate, and $\log_2$ ratio—and three locus resolutions: segment-level BED files, bin-level matrices, and gene-level matrices. Detailed definitions and operations are provided in Methods. All CNA profiles are available as downloadable files (`*seg.txt`, $n = 265$ and `*cna.csv`, $n = 1140$). CNAScope provides curated metadata (https://cna.fengslab.com/database), including cancer subtypes, patient characteristics (ethnicity, race, gender, age), sample-specific features (disease type, primary site, tumor stage, tumor grade), and survival endpoints (overall survival and progression-free survival). For single cells and spatial cells/spots, additional metadata such as cell type, CNA confidence, donor identity, and malignancy status are also curated.

CNAScope performs comprehensive annotation for these CNA profiles, structured into two main perspectives: observational and functional.

From the observational perspective, CNAScope offers annotations that capture the underlying data structure and relationships. This includes phylogenetic inference via hierarchical clustering to reveal sample or cell homogeneity and heterogeneity. Approaches for data dimension reduction, including PCA, ICA, NMF, UMAP, t-SNE, and PHATE, are employed to facilitate data visualization and interpretation. Downloadable files for these annotations include `*meta_scsvas.csv`, $n = 1140$ and `*.nwk`, $n = 1140$.

From the functional perspective, CNAScope provides biologically meaningful annotations. Gene-level CNA annotation is performed with BEDTools, while pathway term-level annotation leverages six MSigDB collections, including cancer hallmarks, KEGG, GOBP, etc. The resulting files (`*gene_cna.csv.gz`, $n = 1140$; `*term_cna.csv.gz`, $n = 1140$) are available for users. Next, for each dataset, we annotated the top 1000 most variable bins, genes, and terms based on CN values. When spatial coordinates were available, we additionally identified and annotated spatially variable CN bins, genes, and terms. Downloadable files include: `*top_CN_variance.csv`, $n = 1137$; `*top_CN_spatial_variance.csv`, $n = 82$. For bulk DNA sample-level analyses, CNAScope conducts focal CNA, gene, and term annotations using GISTIC2 and ORA enrichment. The downloadable results include `*amp_genes.conf_95.txt`, $n = 119$;

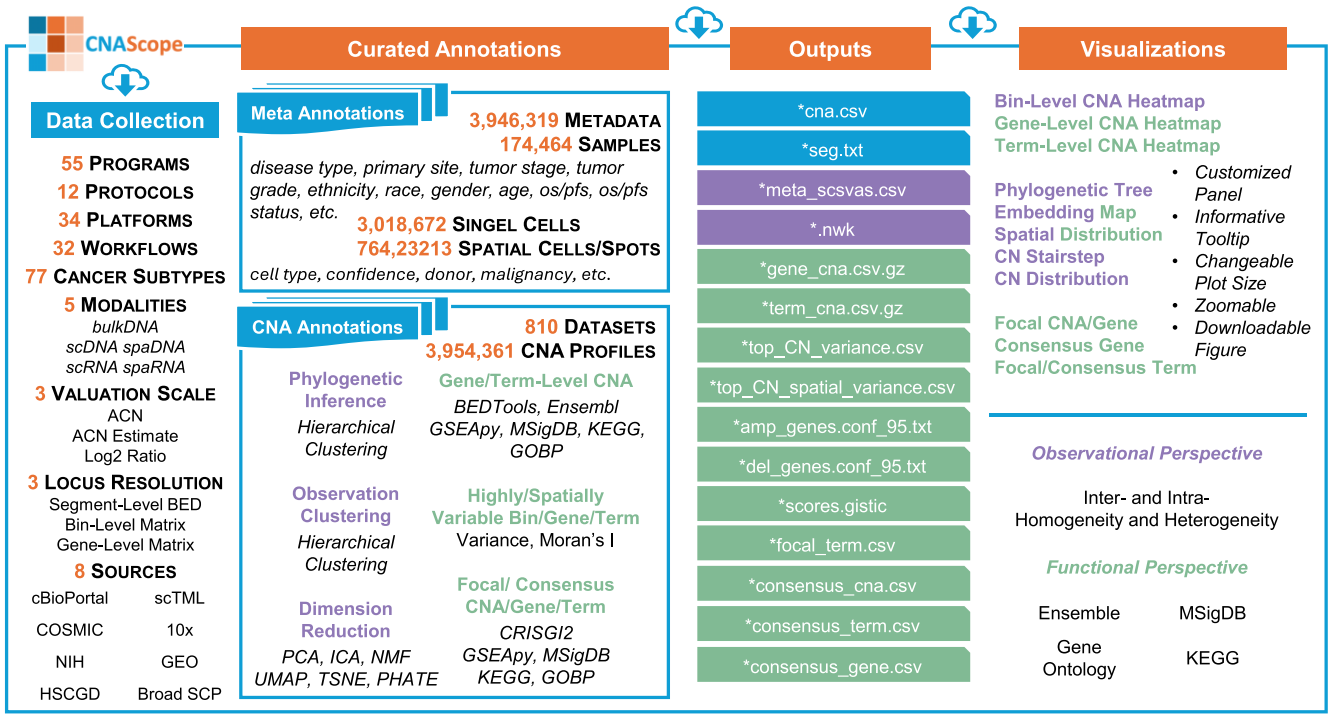


Figure 1. Content and annotation in CNAScope. CNAScope holds 3 954 361 CNA profiles from eight major databases, delivering in-depth annotation for each profile.

`*del_genes.conf_95.txt`, $n = 119$; `*scores.gistic`, $n = 119$; and `*focal_term.csv` $n = 119$. For bulk DNA datasets with at least two CNA profiles from different sequencing protocols or computational workflows, CNAScope also reports consensus CNAs, genes, and terms. These can be downloaded as: `*consensus_cna.csv`, $n = 34$; `*consensus_gene.csv`, $n = 34$; and `*consensus_term.csv`, $n = 34$.

Online annotation workflows in CNAScope

CNAScope enables researchers to annotate newly generated CNA profiles from both observational and functional perspectives (Fig. 2). The platform offers two annotation workflows, designed to accommodate a single CNA profile or cross-profile agreement.

The first workflow, Basic CNA Annotation, allows users to upload a single CNA profile, with the option to include metadata and custom gene sets. Users specify the observation type ("sample," "cell," or "spot"), reference genome ("hg19" or "hg38"), and the desired number of clusters (k). CNAScope then automatically performs a comprehensive suite of basic annotations. Observational modules include phylogenetic inference (hierarchical clustering), observation clustering (hierarchical clustering), and dimension reduction (PCA, ICA, NMF, UMAP, t-SNE, PHATE). Functional modules provide gene-level and term-level CNA annotation, with highly variable and spatially variable bins, genes, and terms identified. When the observation unit is a sample, focal CNA, gene, and term analyses are performed using GISTIC2 built-in six MSigDB pathway collections (including cancer hallmarks, KEGG, GOBP, etc.). Additionally, users can annotate CNA with their own uploaded target gene set terms.

The second workflow, Consensus CNA Annotation, identifies shared CNAs, genes, and terms from multiple CNA profiles generated across different sequencing protocols or computational workflows within a single bulk DNA dataset. Users must upload at least two CNA profiles derived from the same bulk DNA dataset, and specify the locus type (segment or bin) and reference genome (hg19 or hg38), with optional gene sets. This workflow runs the focal CNA, gene, and term annotation modules separately for each profile, then derives consensus CNA, gene, and term annotations across profiles. This yields a unified call set that provides more confident biological insights by reducing noise introduced by differing sequencing protocols and computational workflows.

The entire annotation process is streamlined for intuitive, point-and-click operation, supporting both single- and cross-profile analyses tailored to diverse research needs. Upon completion, CNAScope provides downloadable output files (e.g. `*cna.csv`, `*meta_scsvas.csv`, `*gene_cna.csv.gz`, `*term_cna.csv.gz`, `*top_CN_variance.csv`, `*top_CN_spatial_variance.csv`, `*amp_genes.conf_95.txt`, `*del_genes.conf_95.txt`, `*scores.gistic`, `*focal_term.csv`, `*consensus_cna.csv`, `*consensus_gene.csv`, and `*consensus_term.csv`) via the user workspace.

To enhance user accessibility, CNAScope provides five demos: Workflow 1 includes `TCGA-ACC` (bulk DNA), `WCDT-MCRPC` (bulk DNA), and `BRCA-T10` (scDNA); Workflow 2 features `TCGA-BRCA` (segment) and `TCGA-BRCA` (bin). These demonstration cases offer users step-by-step examples and guidance throughout the annotation process.

Interactive visualizations in CNAScope

CNAScope features user-friendly web interfaces with main navigation tabs for Home, Database, Workflow, Workspace, Tutorial, and Contact. The Home page summarizes the platform's features and dataset statistics, with graphical overviews

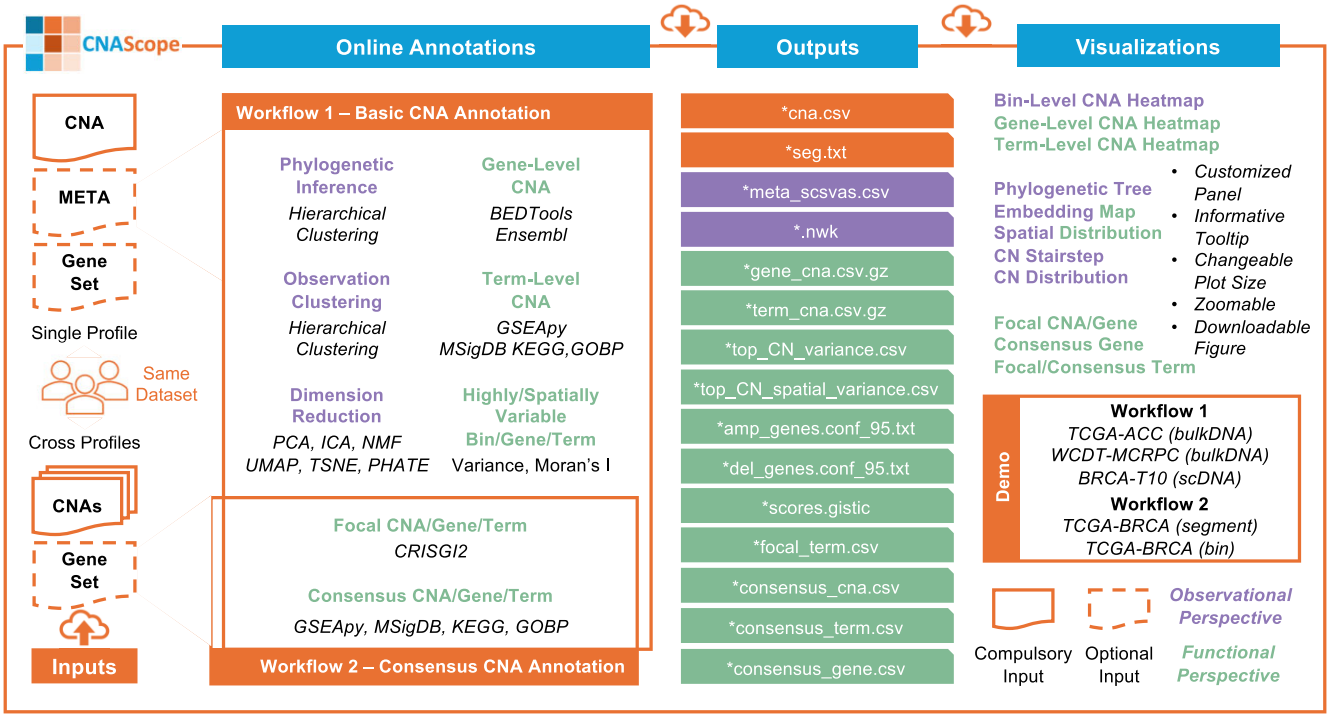


Figure 2. Overview of online annotation workflows in CNAScope. CNAScope provides standardized and interactive online workflows (Basic CNA Annotation and Consensus CNA Annotation) for annotating CNA profiles.

(see Figs 1 and 2). The Database page allows users to filter and browse datasets by source, cancer type, modality, and other parameters, with each dataset page offering detailed and downloadable metadata, annotations, and interactive visualizations. The Workflow page provides access to basic and consensus CNA annotation online tools, allowing users to optionally provide an email address during submission for notification upon task completion. The Workspace allows users to track the status of submitted analyses and download results. Tutorial and Contact pages offer step-by-step user guides and ways for additional assistance.

On both the dataset and workflow result pages, CNAScope provides interactive visualization panels to comprehensively illustrate CNA annotation results [Figs 1 and 2 and (Supplementary Figs S4–S12)]. These include CNA heatmaps at bin, gene, and term levels; phylogenetic trees; embedding maps; spatial distribution plots; CN stairstep plots; CN distribution charts; focal CNA and gene views; consensus genes vein plots; and focal and consensus term bar plots. These visualizations are highly interactive and customizable, featuring informative tooltips, adjustable plot sizes, and zooming. Importantly, every visualization is downloadable in high-resolution, publication-ready formats.

Case study: focal and consensus annotations across sequencing protocols and computational workflows in GDC bulk DNA datasets

In CNAScope, the collected 56 GDC [18] bulk DNA datasets include multiple segment-level BED CNA profiles derived from different sequencing protocols—allele-specific (AS), copy-number segment (CNS), and masked copy-number segment (MCNS)—and computational workflows: ascat2, ascat3, ascatNGS, DNAcopy, and GATK4_CNV. For each dataset, focal CNAs are small amplifications or deletions recurrently observed across multiple samples, annotated with GISTIC2 for each individual CNA profile. Genes that fall within these regions—termed focal—are candidates for biological significance and potential cancer drivers, as their aberrations are recurrent events within the same dataset. Six MSigDB pathway collections—including Hallmark, KEGG, and GO Biological Process—were enriched for focal genes to provide pathway-level focal terms for biological interpretations (see the "Materials and methods" section).

Across 56 GDC datasets, we asked whether CNAScope's focal calls are stable when the segment-level protocol and analysis workflow vary within the same dataset. We compared all protocol–workflow pairs using Jaccard similarity of focal sets and tallied the number of shared focals (Supplementary Fig. S2). Within-protocol comparisons (for example, CNS versus CNS across different workflows) show higher similarity than cross-protocol comparisons (AS versus CNS or AS versus MCNS), indicating that the segment-generation protocol is a major source of variability. Focal genes are more concordant than focal terms, consistent with additional variability introduced by term-level aggregation. Despite these trends, absolute concordance across different protocol–workflow pairs within the same dataset is often low, demonstrating that focal annotations are sensitive not only to bin size (Supplementary Fig. S1A) but also to upstream protocol and workflow choices.

Importantly, overlap counts reveal that many biologically plausible events recur across pairs even when Jaccard similarity is low. For example, in CNS-versus-CNS comparisons the median number of shared focal genes per dataset is substantial (dozens to hundreds), and even cross-protocol comparisons retain a nontrivial shared core; focal terms also exhibit measurable overlap (Supplementary Fig. S2B).

These findings motivate CNAScope's consensus definition: consensus genes are the intersection of focal genes across all available protocol–workflow profiles within a dataset, and consensus terms are pathways enriched by these consensus genes (see the "Materials and methods" section). This design prioritizes robustness—retaining only those focal events that replicate across heterogeneous inputs—thereby mitigating protocol/workflow sensitivity while preserving biologically meaningful signals.

Next, we analyzed CNAScope-annotated consensus events across 56 bulk DNA GDC datasets. In total, 34 datasets contain amplified or deleted consensus genes agreed upon by two to six protocol–workflow combinations (Fig. 3A).

The number of deleted consensus genes (hundreds to thousands) exceeds the number of amplified consensus genes (tens to hundreds) (Fig. 3B). This is plausible because deletions are often more widespread and recurrent—capturing large regions enriched for tumor suppressors and common fragile sites—whereas amplifications typically occur in fewer, more focal hotspots around oncogenes.

Among the 34 datasets, 30 showed significant enrichment (FDR *P*-values <.01) of amplified or deleted consensus terms across MSigDB collections (Hallmark, C2-KEGG, C4-Computational, C5-GO BP, C6-Oncogenic, C7-Immunologic). For clarity, we focused on the top five KEGG terms for amplification and deletion separately. In Fig. 3C, amplification enrichments cluster in growth and signaling programs (PI3K, EGFR, RAS) along with transcriptional control and cell-cycle checkpoints. This aligns with oncogene-driven amplicons that promote proliferation [59, 60]. Several datasets exhibit moderate-to-high GeneRatios with strong significance. For deletions, enrichments trend toward immune-related pathways (e.g. antigen processing and presentation) and apoptosis/cell-death signaling, consistent with loss of tumor-suppressive and immune-modulatory genes [61]. Although fewer KEGG terms are highlighted per dataset than for amplifications, some deletion signals show high significance and sizable GeneRatios.

Moreover, we externally validated CNAScope's focal and consensus annotations using two orthogonal resources—cBioPortal [15] (gene-level CNA frequencies) and Progenetix [19] (segment-level CNA frequencies). We analyzed 33 GDC-sourced TCGA datasets shared across CNAScope, Progenetix, and cBioPortal (see Supplementary Results 2.2 and Supplementary Fig. S3). Together, these analyses show strong external concordance—especially for amplifications—while also highlighting CNAScope-unique, literature-supported focal and consensus events, particularly deletions that may be underrepresented in gene-level frequency resources.

Discussion

In this study, we present CNAScope, an online database for comprehensive annotation and visualization of cancer CNAs. Compared with existing CNA databases—such as cBioPortal [15], COSMIC [16], DGV [17], GDC [18], HSCGD [21], Progenetix [19], and scTML [22]—CNAScope offers several unique advantages (Table 1):

(i) Extensive cancer CNA collection. CNAScope offers a truly comprehensive, cancer-centric, and multi-modal resource with largest cancer data size. It holds CNA data from five data modalities: bulk DNA, single-cell DNA, spatial DNA, single-cell RNA, and spatial DNA. In total, the resource spans 810 datasets, 174 464 samples, 3 018 672 single cells, 764 232 spatial cells/spots, and 3 946 319 associated metadata. The breadth of data modalities and the overall data volume compare favorably with existing databases (Table 1)—which may focus on common and non-pathological variation (DGV [17]), only bulk DNA focus (cBioPortal [15], COSMIC [16], GDC [18], Progenetix [19]), only single-cell focus (HSCGD [21]), transcriptome-derived CNAs (scTML [22]), or only gene-level CNA calls (COSMIC [16], scTML [22]).

(ii) Systematic observational and functional annotation. CNAScope uniquely combines observational and functional annotation. CNAScope applies phylogenetic inference, observation clustering, and dimension reduction to reveal tumor subtypes, clonal structure, and intra-/inter-tumor heterogeneity at both bulk and single-cell levels, whereas existing databases provide no or partial support for these observational annotations (Table 1). Critically, CNAScope supports functional annotation that identifies focal CNAs/genes within individual profiles and consensus CNAs/genes across profiles within the same dataset to mitigate biases from heterogeneous data sources. In addition, CNAScope systematically maps these genes to pathway terms using up-to-date reference databases such as Ensembl [41], MSigDB [43], KEGG [44], and GO [45]. In contrast, other CNA platforms (Table 1) rarely support focal or consensus calling. cBioPortal and Progenetix offer only limited functionality, primarily reporting CNA frequencies (see Supplementary Figure 1.2 for details).

(iii) Code-free online annotation CNAScope uniquely supports on-the-fly annotation directly into its web interface, allowing users to upload single or multiple CNA data and perform basic and consensus CNA annotation discovery without leaving the platform—a feature rarely offered by existing databases (Table 1).

(iv) Comprehensive interactive visualization. In CNAScope, visualization panels are highly interactive, offering informative tooltips, adjustable plot sizing, zooming, highlighting, and export options, and they are configurable via a comprehensive editor for customized annotation choices. However, existing cancer-focused platforms provide only a partial subset of these functionalities (Table 1).

Next, we outline the concerns we encountered in collecting, curating, and annotating CNA data, describe our planned solutions, and call for the community to address these issues together.

We acknowledge the data heterogeneity, systematic biases, and batch effects inherent to CNA calling across datasets. In CNAScope, we use per-dataset CNA profiles from online sources that have already applied widely adopted callers with best practices [4, 6, 40, 62–64] (see https://cna.fengslab.com/database), or we derive CNAs from single-cell and spatial transcriptomic data using inferCNVpy [8] (see Supplementary Methods 1.1). These standardized pipelines incorporate normalization and statistical procedures designed to address batch effects and platform-specific biases when converting raw read counts into ACN, ACN estimate, or $\log_2$-ratio

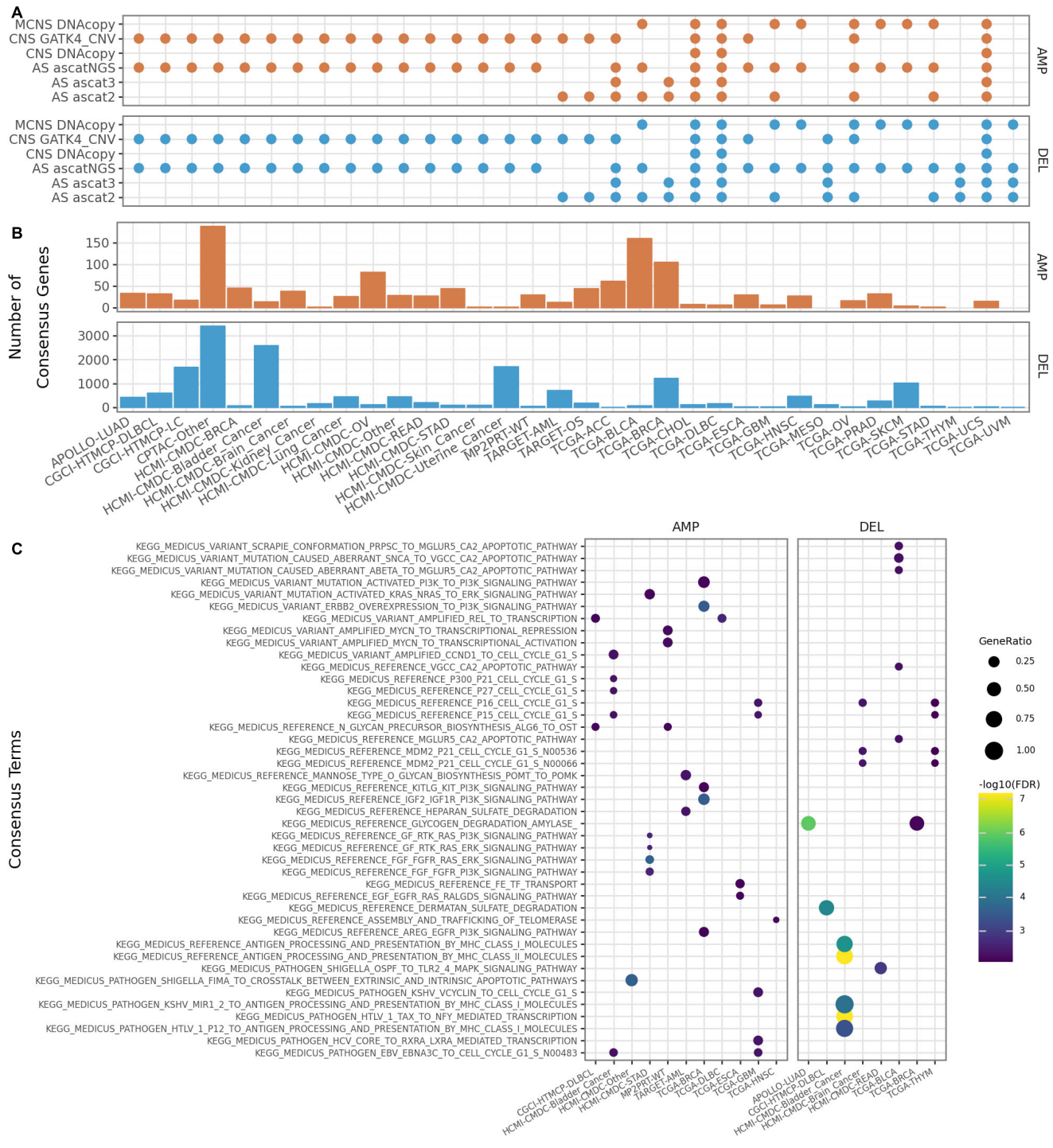


Figure 3. Analysis of consensus CNAs annotated by CNAScope across GDC bulk DNA sequencing protocols and computational workflows. (**A**) Dot plot of consensus availability: each filled point indicates that focal amplification and deletion results exist for a given protocol–workflow within each dataset. (**B**) Bar plot of consensus burden: counts of amplified and deleted consensus genes per dataset. (**C**) KEGG consensus terms: for each dataset, the top five KEGG terms for amplifications and deletions. Point size = GeneRatio (fraction of consensus genes in that term). Color = $-\log_{10}$(FDR), with higher values (yellow/green) indicating stronger significance and lower values (purple/blue) indicating weaker significance. Available protocols: allele-specific (AS), copy-number segment (CNS), and masked copy-number segment (MCNS). Available workflows: ascat2, ascat3, ascatNGS, DNAcopy, and GATK4_CNV. AMP: amplification. DEL: deletion. FDR: false discovery rate.

Table 1. Comparison of CNAScope with existing CNA databases

Feature	CNAScope (2025)	cBioPortal (2025)	COSMIC (2024)	DGV (2020)	GDC (2021)	HSGCD (2025)	Progenetix (2021)	scTML (2024)
Cancer Focus	✓	✓	✓	-	✓	✓	✓	✓
Cancer Data Size								
# of datasets	810	303	54	75	51	45	–	77
# of samples	174 464	195 825	13 753	54 980	130 031	200	240 600	320
# of cells	3 018 672	–	–	–	–	9788	–	240 1261
# of spots	764 232	–	–	–	–	–	–	118 600
Data Modality								
bulkDNA	✓	✓	✓	✓	✓	–	✓	–
scDNA	✓	–	–	–	–	✓	–	–
spaDNA	✓	–	–	–	–	–	–	–
scRNA	✓	–	–	–	–	–	–	✓
spaRNA	✓	–	–	–	–	–	–	✓
CN Valuation Scale								
Absolute copy number (ACN)	✓	–	✓	✓	✓	✓	–	–
ACN Estimate	✓	–	✓	–	✓	–	✓	–
Log2 Ratio	✓	✓	–	–	✓	–	✓	✓
CN Locus Resolution								
Segment-Level BED	✓	✓	–	✓	✓	–	✓	–
Bin-Level Matrices	✓	–	–	–	–	✓	✓	–
Gene-Level Matrices	✓	✓	✓	–	✓	–	✓	✓
Metadata Annotation	✓	✓	✓	✓	✓	L	✓	✓
Search Panel	✓	✓	✓	✓	✓	✓	✓	–
Observational Annotation								
Phylogeny Inference	✓	–	–	–	–	✓	✓	–
Observation Clustering	✓	–	–	–	–	–	–	✓
Dimension Reduction	✓	–	–	–	–	–	–	✓
Functional Annotation								
Gene-Level CNA	✓	✓	✓	✓	✓	–	✓	✓
Term-Level CNA	✓	–	–	–	–	–	–	–
High Variable Bin/Gene/Term	✓	–	–	–	–	–	–	–
Spatial Variable Bin/Gene/Term	✓	–	–	–	–	–	–	–
Focal CNA	✓	L	–	–	–	–	L	–
Focal Gene	✓	–	–	–	–	–	–	–
Focal Term	✓	–	–	–	–	–	–	–
Consensus CNA	✓	–	–	–	–	–	–	–
Consensus Gene	✓	–	–	–	–	–	–	–
Consensus Term	✓	–	–	–	–	–	–	–
Online Workflow	✓	–	–	–	–	✓	–	–
Interactive Visualization								
Customized Panel	✓	✓	–	✓	✓	–	✓	✓
Informative Tooltips	✓	✓	✓	✓	✓	L	–	–
Adjustable Plot Size	✓	–	–	✓	–	L	–	–
Zoomable	✓	✓	–	✓	✓	L	–	–
Downloadable Figure	✓	✓	–	✓	✓	✓	✓	✓
Downloadable Data	✓	✓	✓	✓	✓	✓	✓	✓

✓: Feature supported, -: Not applicable, L: Limited functionality. Detailed explanation is provided in Supplementary Methods 1.2.

profiles. Interpretation of ACN, ACN estimate, and $\log_2$ ratio follows established community conventions: amplification (ACN [estimate] >2; $\log_2$ ratio >0), neutrality (ACN [estimate] = 2; $\log_2$ ratio = 0), and deletion (ACN [estimate] <2; $\log_2$ ratio <0). Accordingly, switching between these scales does not change the underlying biological interpretation. Together, these practices help ensure that most systematic biases and batch effects are mitigated prior to CNAScope's observational and functional annotations so that the remaining variability in raw CNA profiles reflects underlying intra- and inter-tumor heterogeneity.

In the CNAScope annotation step, we preserve the original value scale unless a downstream analysis explicitly requires a different one. Because raw CNA profiles in CNAScope vary in locus resolution (segment-level BED, bin-level, or gene-level matrices) and focal annotations are highly sensitive to resolution, CNAScope retains the native resolution whenever performing focal annotation to avoid information loss. Observation-level annotations, however, require a matrix format. When source CNA data are not provided as matrices (e.g. bulk DNA GDC data), CNAScope curates and converts them into matrices with three available bin sizes (200 kb, 500 kb, and 5 Mb). Benchmark performance for these options is provided in Supplementary Results 2.1 and Supplementary Fig. S2, enabling users to choose their preferred granularity.

GISTIC2 is designed for bulk, segment-level datasets with independent samples. Our benchmarking shows that its focal calls are sensitive to bin size and to biases arising from sequencing protocols and computational work-

flows (Supplementary Results 2.1, Results-Case Study, and Supplementary Figs S1 and S2). Thus, for bulk DNA datasets that provide only raw gene-level or bin-level CNA data, and for single-cell or spatial datasets—where observations are cells or spots from a single patient—the rationale for dataset-level GISTIC2 focal discovery is weak. To our knowledge, no dedicated tools currently support multi-sample focal event detection in single-cell or spatial settings or gene-level focal calling. Accordingly, CNAScope does not report focal and consensus annotations for gene-level, single-cell, or spatial data. We call for methods tailored to focal CNA detection in these data formats, which would enable CNAScope to incorporate such annotations in future releases. Moreover, CNAScope provides consensus calling to identify focal CNAs/genes/terms across multiple profiles within a dataset, mitigating biases from heterogeneous data sources.

For ST data, CNAScope presently uses the single–cell tool inferCNVpy [8] to derive spot–level CNA profiles because spatially aware CNA inference for ST is still immature. SpatialInferCNV [10] relies on the inferCNV core algorithm and primarily projects inferred CNAs onto tissue coordinates without integrating spatial information during inference. STARCH [65] incorporates spatial context but outputs only categorical copy-number states (e.g. amplification, deletion), not full CN profiles. CalicoST [66] infers allele-specific CNAs with spatial context but requires BAM files and cannot operate directly on expression data. We acknowledge this limitation and strongly encourage the development of methods that infer CNAs directly from ST expression, which would enable truly spatially resolved CNA analyses in future CNAScope updates.

CNAScope collects data from five modalities—bulk DNA, scDNA, spatial DNA, scRNA, and spatial RNA—but sample- or cell-level overlap across modalities is limited. We plan to recruit cross-modality CNA profiles derived from the same samples or cells to further enhance CNAScope's richness and comparability in further releases.

Acknowledgements

Author contributions: Xikang Feng (Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Software, Supervision, Writing—review & editing), Jieyi Zheng (Data curation, Formal analysis, Software, Writing—review & editing), Sisi Peng (Data curation, Formal analysis, Writing—review & editing), Anna Jiang (Data curation, Formal analysis, Writing—review & editing), Ka Ho Ng (Data curation), Chengshang Lyu (Software), Qiangguo Jin (Funding acquisition, Investigation, Supervision), and Lingxi Chen (Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Supervision, Visualization, Writing—original draft).

Supplementary data

Supplementary data is available at NAR online.

Conflicts of interest

The authors declare no conflicts of interest.

Funding

We express our gratitude for the support provided by the National Natural Science Foundation of China (No. 32300527; No. 32400519; No. 62572401), the Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515110784), the Research Grants Council of Hong Kong (No. 21200425), the CityUHK Start-Up Grant (No. 9610687), and the Basic Research Programs of Taicang, 2024 (No. TC2024JC43). Funding to pay the Open Access publication charges for this article was provided by the National Natural Science Foundation of China (No. 32300527).

Data availability

All the data are freely available at https://cna.fengslab.com/.

References

1. Steele CD, Abbasi A, Islam SA *et al.* Signatures of copy number alterations in human cancer. *Nature* 2022;606:984–91. https://doi.org/10.1038/s41586-022-04738-6
2. Hastings PJ, Lupski JR, Rosenberg SM *et al.* Mechanisms of change in gene copy number. *Nat Rev Genet* 2009;10:551–64. https://doi.org/10.1038/nrg2593
3. Navin N, Krasnitz A, Rodgers L *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res* 2010;20:68–80. https://doi.org/10.1101/gr.099622.109
4. Raine KM, Van Loo P, Wedge DC *et al.* ascatNgs: identifying somatically acquired copy-number alterations from whole-genome sequencing data. *Curr Protoc Bioinform* 2016;56:15–9. https://doi.org/10.1002/cpbi.17
5. Jiang Y, Oldridge DA, Diskin SJ *et al.* CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res* 2015;43:e39. https://doi.org/10.1093/nar/gku1363
6. Garvin T, Aboukhalil R, Kendall J *et al.* Interactive analysis and assessment of single-cell copy-number variations. *Nat Methods* 2015;12:1058. https://doi.org/10.1038/nmeth.3578
7. Andor N, Lau BT, Catalanotti C *et al.* Joint single cell DNA-seq and RNA-seq of gastric cancer cell lines reveals rules of *in vitro* evolution. *NAR Genom Bioinform* 2020;2:lqaa016. https://doi.org/10.1093/nargab/lqaa016
8. Puram SV, Tirosh I, Parikh AS *et al.* Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* 2017;171:1611–24. https://doi.org/10.1016/j.cell.2017.10.044
9. Zhao T, Chiang ZD, Morriss JW *et al.* Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature* 2022;601:85–91. https://doi.org/10.1038/s41586-021-04217-4
10. Erickson A, He M, Berglund E *et al.* Spatially resolved clonal copy number alterations in benign and malignant tissue. *Nature* 2022;608:360–7. https://doi.org/10.1038/s41586-022-05023-2
11. Navin N, Kendall J, Troge J *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* 2011;472:90. https://doi.org/10.1038/nature09807
12. Minussi DC, Nicholson MD, Ye H *et al.* Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature* 2021;592:302–8. https://doi.org/10.1038/s41586-021-03357-x
13. Hieronymus H, Schultz N, Gopalan A *et al.* Copy number alteration burden predicts prostate cancer relapse. *Proc Natl Acad Sci* 2014;111:11139–44. https://doi.org/10.1073/pnas.1411446111
14. Chen L, Qing Y, Li R *et al.* Somatic variant analysis suite: copy number variation clonal visualization online platform for large-scale single-cell genomics. *Brief Bioinform* 2022;23:bbab452. https://doi.org/10.1093/bib/bbab452

15. De Bruijn I, Kundra R, Mastrogiacomo B *et al.* Analysis and visualization of longitudinal genomic and clinical data from the AACR project GENIE biopharma collaborative in cBioPortal. *Cancer Res* 2023;83:3861–7. https://doi.org/10.1158/0008-5472.CAN-23-0816
16. Sondka Z, Dhir NB, Carvalho-Silva D *et al.* COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res* 2024;52:D1210–7. https://doi.org/10.1093/nar/gkad986
17. MacDonald JR, Ziman R, Yuen RK *et al.* The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* 2014;42:D986–92. https://doi.org/10.1093/nar/gkt958
18. Zhang Z, Hernandez K, Savage J *et al.* Uniform genomic data analysis in the NCI Genomic Data Commons. *Nat Commun* 2021;12:1226. https://doi.org/10.1038/s41467-021-21254-9
19. Huang Q, Carrio-Cordo P, Gao B *et al.* The Progenetix oncogenomic resource in 2021. *Database* 2021;2021:baab043. https://doi.org/10.1093/database/baab043
20. Pan Q, Liu YJ, Bai XF *et al.* VARAdb: a comprehensive variation annotation database for human. *Nucleic Acids Res* 2021;49:D1431–44. https://doi.org/10.1093/nar/gkaa922
21. Fu J, He S, Yang Y *et al.* HSCGD: a comprehensive database of single-cell whole-genome data and metadata. *Nucleic Acids Res* 2025;53:D1029–38. https://doi.org/10.1093/nar/gkae971
22. Li H, Ma T, Zhao Z *et al.* scTML: a pan-cancer single-cell landscape of multiple mutation types. *Nucleic Acids Res* 2025;53:D1547–56. https://doi.org/10.1093/nar/gkae898
23. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;30:207–10. https://doi.org/10.1093/nar/30.1.207
24. Tarhan L, Bistline J, Chang J *et al.* Single Cell Portal: an interactive home for single-cell genomics data. bioRxiv, https://doi.org/10.1101/2023.07.13.548886, 17 July 2023, preprint: not peer reviewed.
25. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008;455:1061–8. https://doi.org/10.1038/nature07385
26. Downing JR, Wilson RK, Zhang J *et al.* The pediatric cancer genome project. *Nat Genet* 2012;44:619–22. https://doi.org/10.1038/ng.2287
27. Carter SL, Cibulskis K, Helman E *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 2012;30:413–21. https://doi.org/10.1038/nbt.2203
28. Xi R, Hadjipanayis AG, Luquette LJ *et al.* Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci* 2011;108:E1128–36. https://doi.org/10.1073/pnas.1110574108
29. Virtanen P, Gommers R, Oliphant TE *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2
30. Ma S, Dai Y. Principal component analysis based methods in bioinformatics studies. *Brief Bioinform* 2011;12:714–22. https://doi.org/10.1093/bib/bbq090
31. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. *Neural Networks* 2000;13:411–30. https://doi.org/10.1016/S0893-6080(00)00026-5
32. Lee DD, Seung HS. Algorithms for non-negative matrix factorization. In: *Proceedings of the 14th International Conference on Neural Information Processing Systems (NIPS'00)*. Cambridge, MA, USA: MIT Press, 2000, 535–541. https://dl.acm.org/doi/10.5555/3008751.3008829
33. Becht E, McInnes L, Healy J *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 2019;37:38–44. https://doi.org/10.1038/nbt.4314
34. Maaten Lvd, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;9:2579–605.
35. Moon KR, van Dijk D, Wang Z *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* 2019;37:1482–92.
36. Shah SP, Xuan X, DeLeeuw RJ *et al.* Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics* 2006;22:e431–9. https://doi.org/10.1093/bioinformatics/btl238
37. Bakker B, Taudt A, Belderbos ME *et al.* Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biol* 2016;17:115. https://doi.org/10.1186/s13059-016-0971-7
38. Wang R, Lin DY, Jiang Y. SCOPE: a normalization and copy-number estimation method for single-cell DNA sequencing. *Cell Syst* 2020;10:445–52. https://doi.org/10.1016/j.cels.2020.03.005
39. Zaccaria S, Raphael BJ. Characterizing allele-and haplotype-specific copy numbers in single cells with CHISEL. *Nat Biotechnol* 2021;39:207–14. https://doi.org/10.1038/s41587-020-0661-6
40. Talevich E, Shain AH, Botton T *et al.* CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput Biol* 2016;12:e1004873. https://doi.org/10.1371/journal.pcbi.1004873
41. Harrison PW, Amode MR, Austine-Orimoloye O *et al.* Ensembl 2024. *Nucleic Acids Res* 2024;52:D891–9. https://doi.org/10.1093/nar/gkad1049
42. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033
43. Liberzon A, Subramanian A, Pinchback R *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011;27:1739–40. https://doi.org/10.1093/bioinformatics/btr260
44. Kanehisa M, Furumichi M, Tanabe M *et al.* KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;45:D353–61. https://doi.org/10.1093/nar/gkw1092
45. Consortium GO. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004;32:D258–61. https://doi.org/10.1093/nar/gkh036
46. Van Dyk E, Hoogstraat M, Ten Hoeve J *et al.* RUBIC identifies driver genes by detecting recurrent DNA copy number breaks. *Nat Commun* 2016;7:12159. https://doi.org/10.1038/ncomms12159
47. Mermel CH, Schumacher SE, Hill B *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011;12:R41. https://doi.org/10.1186/gb-2011-12-4-r41
48. Wang X, Li X, Cheng Y *et al.* Copy number alterations detected by whole-exome and whole-genome sequencing of esophageal adenocarcinoma. *Hum Genom* 2015;9:22. https://doi.org/10.1186/s40246-015-0044-0
49. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* 2023;39:btac757. https://doi.org/10.1093/bioinformatics/btac757
50. Tonsing-Carter E, Agarwal R, Kyi CW *et al.* Human cancer models initiative (HCMI): A community resource of next-generation cancer models and associated data. *Cancer Res* 2023;83:4681. https://doi.org/10.1158/1538-7445.AM2023-4681
51. Zhang J, Bajari R, Andric D *et al.* The international cancer genome consortium data portal. *Nat Biotechnol* 2019;37:367–9. https://doi.org/10.1038/s41587-019-0055-9
52. Gautier L, Cope L, Bolstad BM *et al.* affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 2004;20:307–15. https://doi.org/10.1093/bioinformatics/btg405
53. Quail MA, Smith M, Coupland P *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 2012;13:341. https://doi.org/10.1186/1471-2164-13-341

54. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* 2018;13:599–604. https://doi.org/10.1038/nprot.2017.149
55. Picelli S, Faridani OR, Björklund ÅK *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* 2014;9:171–81. https://doi.org/10.1038/nprot.2014.006
56. Rao A, Barkley D, França GS *et al.* Exploring tissue architecture using spatial transcriptomics. *Nature* 2021;596:211–20. https://doi.org/10.1038/s41586-021-03634-9
57. Janesick A, Shelansky R, Gottscho AD *et al.* High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and *in situ* analysis. *Nat Commun* 2023;14:8353. https://doi.org/10.1038/s41467-023-43458-x
58. Williams CG, Engel JA, Soon MS *et al.* Studying lymphocyte differentiation in the spleen via spatial transcriptomics. *J Immunol* 2021;206:98–55. https://doi.org/10.4049/jimmunol.206.Supp.98.55
59. Ponsioen B, Post JB, Buissant des Amorie JR *et al.* Quantifying single-cell ERK dynamics in colorectal cancer organoids reveals EGFR as an amplifier of oncogenic MAPK pathway signalling. *Nat Cell Biol* 2021;23:377–90. https://doi.org/10.1038/s41556-021-00654-5
60. Wang T, Yu H, Hughes NW *et al.* Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic Ras. *Cell* 2017;168:890–903. https://doi.org/10.1016/j.cell.2017.01.013
61. Huang J, Chen W, Jie Z *et al.* Comprehensive analysis of immune implications and prognostic value of SPI1 in gastric cancer. *Front Oncol* 2022;12:820568. https://doi.org/10.3389/fonc.2022.820568
62. Van Loo P, Nordgard SH, Lingjærde OC *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci* 2010;107:16910–5. https://doi.org/10.1073/pnas.1009843107
63. Auwera G, Carneiro M, Hartl C *et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Inc*, Hoboken. NJ, USA: John Wiley & Sons, 2013.
64. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res* 2016;44:e131. https://doi.org/10.1093/nar/gkw520
65. Elyanow R, Zeira R, Land M *et al.* STARCH: copy number and clone inference from spatial transcriptomics data. *Phys Biol* 2021;18:035001. https://doi.org/10.1088/1478-3975/abbe99
66. Ma C, Balaban M, Liu J *et al.* Inferring allele-specific copy number aberrations and tumor phylogeography from spatially resolved transcriptomics. *Nat Methods* 2024;21:2239–47. https://doi.org/10.1038/s41592-024-02438-9

Received: August 18, 2025. **Revised:** October 7, 2025. **Accepted:** October 18, 2025