Molecular Cytogenetics

RESEARCH Open Access

Copy number variation heterogeneity reveals biological inconsistency in hierarchical cancer classifications

Ziying Yang[1,2*], Paula Carrio-Cordo[1,2] and Michael Baudis[1,2*]

Abstract

Cancers are heterogeneous diseases with unifying features of abnormal and consuming cell growth, where the deregulation of normal cellular functions is initiated by the accumulation of genomic mutations in cells of - potentially - any organ. At diagnosis malignancies typically present with patterns of somatic genome variants on diverse levels of heterogeneity. Among the different types of genomic alterations, copy number variants (CNV) represent a distinct, near-ubiquitous class of structural variants. Cancer classifications are foundational for patient care and oncology research. Terminologies such as the National Cancer Institute Thesaurus provide large sets of hierarchical cancer classification vocabularies and promote data interoperability and ontology-driven computational analysis. To find out how categorical classifications correspond to genomic observations, we conducted a meta-analysis of inter-sample genomic heterogeneity for classification hierarchies on CNV profiles from 97,142 individual samples across 512 cancer entities, and evaluated recurring CNV signatures across diagnostic subsets. Our results highlight specific biological mechanisms across cancer entities with the potential for improvement of patient stratification and future enhancement of cancer classification systems and provide some indications for cooperative genomic events across distinct clinical entities.

Introduction

Structural genome variations constitute a heterogeneous group of genomic alterations and can have profound consequences in evolution and human disease[1–3]. Copy number variations (CNV) represent a type of structural genomic variations as the result of unbalanced genomic rearrangements and either increase or decrease the DNA content of a genomic region ranging from kilobases to multiple megabases [3]. CNVs have been identified as a major contributor to malignant transformation, partially through their impact on expression levels of genes within the copy-varied regions [4] and the exploration of their genesis, structure, functional effects and disease association plays a crucial role in biomedical research.

Malignant neoplasms comprise a group of complex and progressive diseases arising from somatic mutations and with a common hallmark of genomic instability. Cancer formation and progression are frequently associated with widespread copy number abnormalities [5]. While germline copy number variations constitute a major part of genomic variability within and between populations and contribute to hereditary disorders [6] in most cancer types somatic CNV accumulates during the progression of the disease [7–9]. The pattern of CNV events observed in a given cancer at the time of diagnosis will have been influenced by the selection of mutations beneficial for the clonal expansion of the dominant subclone;

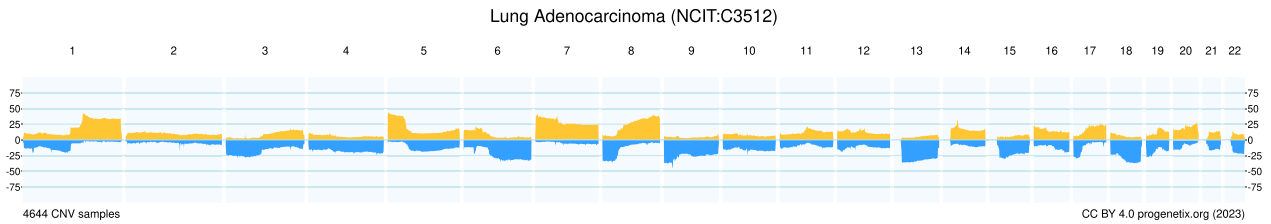
*Correspondence:
Ziying Yang
ziying.yang@uzh.ch
Michael Baudis
michael.baudis@mls.uzh.ch
[1] Department of Molecular Life Sciences, University of Zurich, Winterthurerstr. 190, 8057 Zurich, Switzerland
[2] Swiss Institute of Bioinformatics, Zurich, Switzerland

tissue-specific requirements and the tumor microenvironment will have affected this oncogenic evolution.

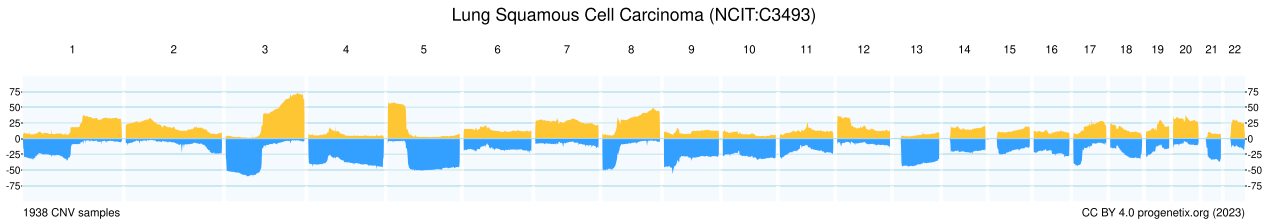
In cancer genomics, CNVs frequently are divided into two classes based on their size: i) large-scale, "chromosome level" variants encompassing over 25% [10] or 1/3 of a chromosome arm [11]; and ii) focal variants operationally defined as having usually not more than 3 Mb in size and therefore containing only few genes. While chromosome-scale CNV show different patterns across tumor entities, indicating selective processes during oncogenesis, focal CNV are considered a stronger indication of specific "driver" gene involvement but also operationally more accessible due to their low content of (potential) target genes.

Specific cancer-related gene families and pathways have been shown to be overrepresented among focal CNV with a predominance of kinases, cell cycle regulators, and MYC family members [12]. Pathogenic CNVs form different classes of patterns in cancers that can be helpful in the diagnostic and therapeutic assessment of individual cancer cases. For instance, glioblastoma (GBM) is known for its complex genomic landscape, characterized by extensive copy number variations (CNVs) related to its aggressive behavior and treatment resistance. Here, amplification of the Epidermal Growth Factor Receptor (EGFR) gene represents a hallmark CNV. EGFR amplification leads to overexpression of the receptor, driving increased signaling for cell growth and survival [13]. Another hallmark CNV in GBM and other cancers is the, frequently homozygous, deletion of the CDKN2A/B gene locus on chromosome 9. This deletion affects the cyclin-dependent kinase inhibitor genes, allowing uncontrolled cell cycle progression.

Previous studies have shown that the CNV pattern can be used to organize cancer samples with potential association to specific diseases or disease subtypes [14–16]. Similarly to the correlation of individual samples' CNV patterns to a given diagnosis, aggregated CNV profiling data frequently show distinct CNV frequency patterns for different diseases, even when applied to samples from the same organ (e.g. lung adenocarcinoma and squamous cell carcinoma ([17] and Progenetix database Fig. 1a and b). However, in many cancer types inter-sample CNV heterogeneity points to the presence of biological heterogeneity among different tumors of the same diagnostic concept. For example, Cavalli et al. analyzed molecular events in 763 primary medulloblastoma samples using the similarity network fusion approach and identified subtypes with distinct CNV patterns, activated pathways, and clinical outcomes within each of the four known subgroups and further delineated group 3 from group 4 [18]. Therefore, the analysis of genomic heterogeneity of individual cancer types may provide an avenue towards detecting biological heterogeneity with implications for cancer research as well as diagnostic and prognostic purposes.



(a) **CNV frequency of lung adenocarcinoma in the Progenetix database.**



(b) **CNV frequency of squamous cell carcinoma in the Progenetix database.**

Fig. 1 Example CNV frequency patterns of lung adenocarcinoma **a** and lung squamous cell carcinoma **b** in the Progenetix database. The x-axis indicates the genome and y-axis indicates the frequency of CNV event in the corresponding position. Orange and blue indicate duplication and deletion, respectively

Standardized disease classifications are essential for expressing diagnostic categories and play a crucial role in the systemic study of cancer biologies as well as for the evaluation of cancer incidences and epidemiology [19]. Additionally to using organ location and histopathological characteristics such as the foundational concepts of the WHO ICD-O 3 classification [20], parameters from genomic and transcriptomic analyses have recently been used in the definition of various cancer types. For example, colorectal adenocarcinomas have been separated into CMS1 (microsatellite instability immune), CMS2 (canonical), CMS3 (metabolic), and CMS4 (mesenchymal) subgroups [21]; for medulloblastomas, molecular analysis resulted in a grouping into SHH, WNT, "Group 3" and "Group 4" [22] with only limited overlap with previously defined histological subtypes. However, high heterogeneity in cellular phenotypes and dynamic plasticity of tumor microenvironments make tumor categorization a demanding and complicated task, with the need to balance between categorical classifications and individual, "personalized" feature definitions and a move towards dynamic and coherent classification systems that can allow for iterative expansion and revision of cancer entities and subtypes.

Increasingly, the use of hierarchical ontologies for biological classifications is being recognized as fundamental for data access, reusability and large-scale analysis in the area of cancer research and therapy as well as in other fields of academic biomedicine. Here, a recent attempt has been in the development and continuous update of the *NCIt Neoplasm Core* which provides a controlled vocabulary for specialists at different sub-domains of oncology [23]. The NCIt cancer classification system is a part of the National Cancer Institute Thesaurus (NCIt), which is a standardized vocabulary of cancer-related concepts. Its hierarchical arrangement with root nodes for anatomic sites and histological types allows for a systematic representation of diseases, with broader categories encompassing more general concepts and progressively narrowing down to more specific and detailed disease entities. The tiered structure facilitates a comprehensive understanding of disease relationships, classifications, and subtypes within the NCIt system.

However, the relationships between different types of cancer and their associated morphological features may not fully reflect the biological reality due to the biological heterogeneity or possibly inaccurate clinical disease classification. Here, a data-driven approach integrating genomic profiling data with the hierarchical structure of the cancer classification system has the potential to shed new light on relationships between different cancer subtypes as well as the underlying mutational events [24–26].

In this study, we perform a meta-analysis of disease classifications and their concordance with somatic genomic variations based on a large data set of cancer CNV profiles from the Progenetix database [27, 28]. We employ a framework that utilizes the hierarchical structure of the NCIt neoplasia core and apply machine learning methods to construct a comparative CNV profiling structure. We derive distance estimates from sample-specific CNV profiling data to measure the CNV heterogeneity between samples and extract significant cluster-specific CNV features and determine measures for genomic heterogeneity within cancer types as well as similarities between different cancer entities. By mapping CNV-derived measures for genomic distances to the NCIt hierarchical tree we are able to reveal potential inconsistencies between biological relationships and classification systems of different cancer types. These results also point to specific biological mechanisms and varying levels of cancer subtypes across entities and can help to improve patient stratification as well as to enhance cancer classification systems in the future.

Materials and Methods

Sample selection and data pre-processing

Somatic copy number variation profiles were collected from our Progenetix database (progenetix.org; [28]), the largest open resource for curated cancer genome profiling data with a focus on copy number variations. Progenetix currently contains over 116'000 cancer CNV profiles, mapped to over 800 diagnostic entities by NCIt code. In this study, we utilize CNV profiling data of samples representing the 512 NCIt codes represented by at least 50 individual samples.

Cancer genomes frequently present with a large number of individual copy number variation events of varying sizes. To enable robust statistical analysis, a single vector representation of each sample is constructed through the aggregation of the raw CNV events across all autosomal chromosomes (1,...,22; sex chromosomes excluded due to inconsistent CNV representations in original data). Starting from chromosomal centromere positions, genome bins of 1Mb size are generated resulting in 2892 bins (telomeric bins with potential size differences). For each of the bins, a value representing the fractional coverage for genomic gains and losses is being calculated separately per sample, resulting in a vector of 5784 individual CNV coverage values. A data matrix is created consisting of the status vectors and sample-specific metadata (sample id, diagnostic codes). Samples without any CNV events are removed. For computational convenience, the CNV value matrix is converted into a binary representation in which "1" stands for the occurrence of

a CNV event (if the coverage is larger than 0) in the corresponding bin, and 0 represents the absence of CNVs. After these pre-processing steps, a binary CNV event matrix of 97,142 samples with 5784 bins from 512 unique NCIt entities was built as input for further CNV heterogeneity analysis.

Classification systems

Figure 2 shows the NCIt morphology network in Progenetix. At the top level of the NCIt morphology tree is the broad category "Neoplasm". These categories are further divided into more specific subcategories based on the histological characteristics of the tissue. For example, under the "Neoplasm" category, there are subcategories such as "Epithelial Neoplasm", "Mesenchymal Neoplasm," and "Hematopoietic and Lymphoid Tissue Neoplasm." Under "Epithelial Neoplasm," there are even more specific subcategories such as "Squamous Cell Carcinoma" and "Adenocarcinoma". For many entities the NCIt morphology tree also includes information on tumor grade and differentiation, potentially providing more accurate disease concepts. The morphology-based hierarchy classifies cancers at up to 12 different levels of detail (as seen in the histogram plots of Fig. 2); however, the bulk of the unique NCIt codes fall between level 1 and level 6 of the tree.

Methodology

Based on the large amount of CNV profiles and hierarchical cancer classification structure in Progenetix, we build a framework (Fig. 3) to analyze the relationships between CNV heterogeneity and the cancer classifications systematically. This framework first collects all sample CNV profiles based on the NCIt hierarchical structure. Next, it applies hierarchical clustering on the CNV profiles for two purposes: a) on the top right side, we calculated the distance of all sample pairs with a sample distance matrix and mapped the sample distance matrix to the NCIt entity distance matrix. The NCIt entity distance matrix will be further projected to the NCIt hierarchical structure by reflecting the distance on the length of edges in the hierarchical structure. b) As shown at the bottom left, after the first clustering, each sample will be assigned to a cluster based on the similarity of CNV profiles. Then we will calculate the frequency of these clusters and extract the high-frequency CNV events as features, and apply a second clustering on these CNV features, integrated with the original NCIt entities of these samples, we uncover potential cancer subtypes, relationships between different cancer entities and significant CNV events.

Genomic distances and hierarchical clustering of CNV profiles

To estimate how CNV heterogeneity relates to the NCIt cancer classification system, in this step we calculate all

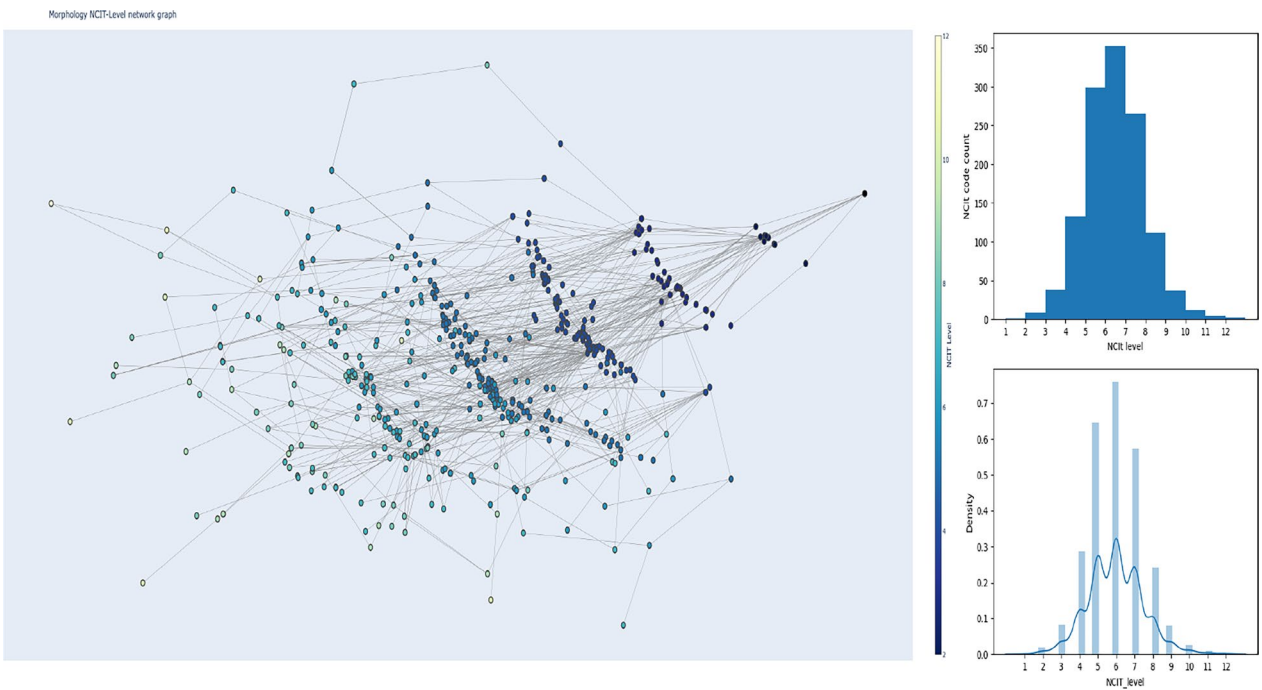


Fig. 2 Network representation of terms in the hierarchical NCIt neoplasia core classification system. The histograms on the right show the number of NCIt codes at hierarchy level

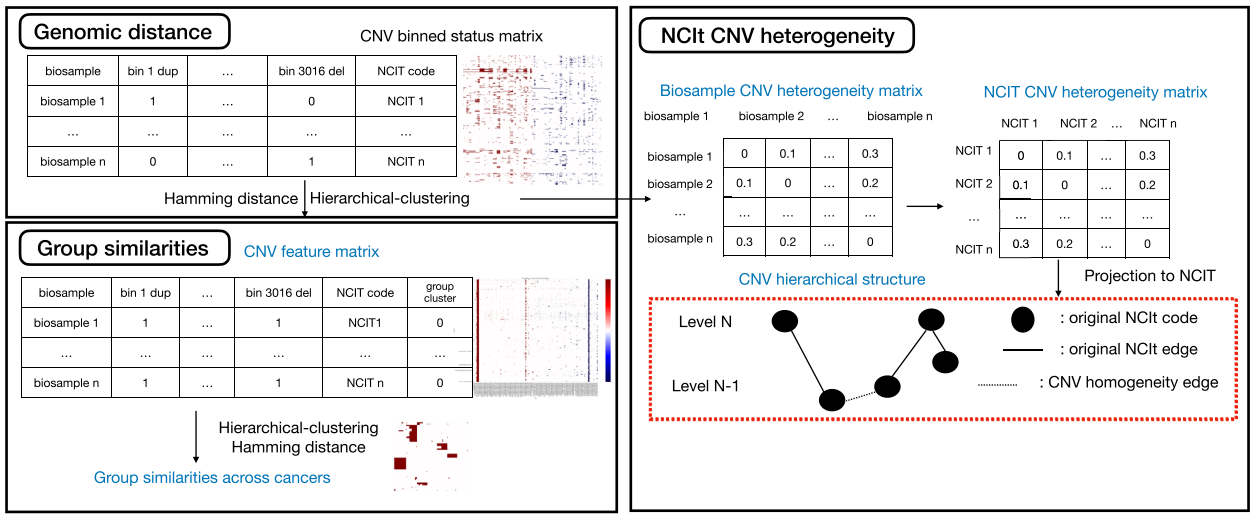


Fig. 3 Framework of analyzing the CNV heterogeneity based on NCIt system

inter-sample CNV profile distances and then evaluate them in aggregate for the corresponding NCIt entities. Specifically, we perform hierarchical clustering on the CNV event matrix of all filtered samples using "Hamming distance" [29] as metric. In the following, we use "CNV heterogeneity" to reflect the diversity of CNV patterns among different samples where the Hamming distance is the metric used to quantify the dissimilarity between two sequences of equal length, *i.e.* in the context of copy number variation analysis to compare binary CNV profiles of different samples represented as binary vectors. A higher Hamming distance indicates greater dissimilarity and hence more heterogeneity two samples. Hamming distance is sensitive to small changes in CNV profiles where even a single altered genomic position can contribute to an increase in the Hamming distance, allowing for the detection of subtle CNV heterogeneity but - when applied to unselected CNV data - can be influenced by the extension of individual events. Generally, a low distance indicates similarity and homogeneity among CNV profiles, while a high distance suggests significant inter-sample heterogeneity.

The Hamming distance (H) is calculated as:

$$H(\text{Sample A}, \text{Sample B}) = \sum_{i=1}^{n} \delta(x_i, y_i)$$

where n is the length of the sequences, x_i and y_i are the values at position i in the sequences, and $\delta(x_i, y_i)$ is the Kronecker delta function which returns 0 if x_i equals y_i, and 1 otherwise.

Hierarchical clustering is a technique used to group similar data points into clusters based on their distances. This process provides insights into the underlying structure of the data and helps in identifying meaningful patterns. Here, after calculating the Hamming distances, we choose the *complete linkage* method to define how clusters are being arranged based on the distances of their constituent data points, and to construct a dendrogram representation of the cluster hierarchy where each leaf node represents an individual data point *e.g.asample*. The hierarchical clustering algorithm iteratively merges clusters based on their distances forming branches in the dendrogram. To separate clusters for further analysis we decide on a strict distance threshold (0.1), which ensures high consistency CNV pattern within clusters and enough samples, and cut the dendrogram at the specified value. After this process, each sample will be assigned a group cluster label based on the similarity of their CNV profiles. Each sample is now assigned a 'group cluster' label, representing its clustering result, and an 'NCIt code' label, indicating the cancer entity it belongs to. These labels facilitate the exploration of heterogeneity within individual cancers and the identification of similarities across different cancer types.

Transfer of sample CNV heterogeneity to NCIt CNV heterogeneity matrix

From the distances for all sample pairs we derive a sample CNV heterogeneity matrix H_b, which estimates the inter-sample CNV heterogeneity. Next, we transfer the CNV heterogeneity matrix of samples to all NCIt corresponding entities in the NCIt system and calculate the NCIt distance matrix (H_N) based on the sample distance matrix (H_b). Given the sample distance matrix H_b and the corresponding relationship between

samples and NCIt concepts, we can compute the NCIt distance matrix H_N. Let $d(\cdot,\cdot)$ represent the distance function.

For each NCIt concept, we calculate the average CNV distance between samples within that concept. This average distance serves as a measure of CNV heterogeneity for the given NCIt concept. Let B_{N_i} denote the set of samples associated with an NCIt concept N_i. Then, the NCIt distance H_N is calculated as:

$$H_N(N_i) = \frac{1}{\binom{|B_{N_i}|}{2}} \sum_{\substack{b_j, b_k \in B_{N_i} \\ j<k}} H_b(b_j, b_k)$$

where $\binom{|B_{N_i}|}{2}$ represents the number of unique pairs of biosamples in the set B_{N_i}, and $H_b(b_j, b_k)$ is the distance between two distinct biosamples b_j and b_k in the CNV distance matrix H_b. By iterating calculating distance for each NCIt pair, we can get an NCIt distance matrix, which will be used to be mapped to the NCIt hierarchical structure.

Group similarities across entities

After the first clustering on all collected CNV profiles, to select significant group clusters, which should have biology-meaningful similar profiles, we set the distance threshold as 0.1 so that samples with a distance less than 0.1 will be assigned to the same "group cluster". To remove noisy CNV profiles, we select group clusters composed of over 50 but less than 5000 samples. Therefore, each sample has been assigned a "group cluster" label, representing the association to a subset of samples of a given NCIt code with similar CNV profiles. For analyzing how CNV heterogeneity is distributed across cancer types, we calculate the bin-specific CNV frequencies for each "group cluster" with the resulting frequency distributions serving as summary representation for each cluster. On the other hand, there may also exist cancer types with similar patterns. To uncover the CNV homogeneity between cancers, for each cluster, we set a frequency threshold (z-score higher than 2, which means the value is in the top 2.5%) and transfer the frequency data into binary, 1 indicates a feature/pattern in the corresponding bin, otherwise 0. In this way, we extract the CNV patterns of each "group cluster". Then we applied a second hierarchical clustering on the binary "CNV pattern" data, taking only feature CNV events into account. After the second clustering, we can find out the common/specific patterns of cancer entities and can measure the distance within/between NCIt entities based on CNV heterogeneity.

Results

In total 9785 samples (around 10% of all samples) from 62 distinct NCIT codes were selected. We need to note here that the distinct NCIt codes here are the direct matches of these selected samples, considering the hierarchical structure of the NCIt system the selected samples cover over 90% of NCIt system.

Due to the strict distance threshold, the selected "group clusters" are super significantly similar. Figure 4 shows the distribution of NCIt entities across different clusters. Generally, close NCIt entities belonging to the same "group clusters" have a rather close distance on the NCIt tree. We can also find that for the cancer entities with less distinct "group clusters" indicate less CNV heterogeneity, such as Oligodendroglioma, Anaplastic Oligodendroglioma and Astrocytoma, and are more likely to be in a deep level of the NCIt system. In addition, these three diseases show high homogeneity since they belong to the same "group cluster", which also belongs to the same branch on the NCIt hierarchical tree. We can also find some diseases with multiple "group clusters", for instance, Lung Squamous Carcinoma, Ductal Breast Carcinoma, and Bladder Urothelial Carcinoma. The three diseases also share common "group clusters", indicating the high heterogeneity within while homogeneity between these diseases, which is inconsistent with the NCIt tree. Therefore, we can conclude the trend of constancy of CNV heterogeneity with NCIt hierarchical tree and also the existence of inconsistency.

CNV heterogeneity within a cancer type

We first analyze the group clusters from the same NCIt entities to discover how CNV profiles can reveal heterogeneity within the same disease. The following shows the CNV profiles of some representative cancer entities (more cancers can be found in the supplementary file) in different "group clusters", the corresponding frequency plots and tables show the CNV patterns. By literature review, reported CNV patterns are colored red (Figs. 5, 6, 7).

Glioblastoma

As Fig. 5 shows, the pattern of chromosome 7 gain and chromosome 10 loss is commonly associated with glioblastoma and affects multiple genes involved in cell growth regulation, including EGFR on chromosome 7 and PTEN on chromosome 10. While the chromosome 7 gain and chromosome 10 loss do not always co-occur, they also show some co-occurrence with dup 19.

Giant cell glioblastoma (GC-GBM) is a rare variant of IDH-wt GBM histologically characterized by the presence of numerous multinucleated giant cells and

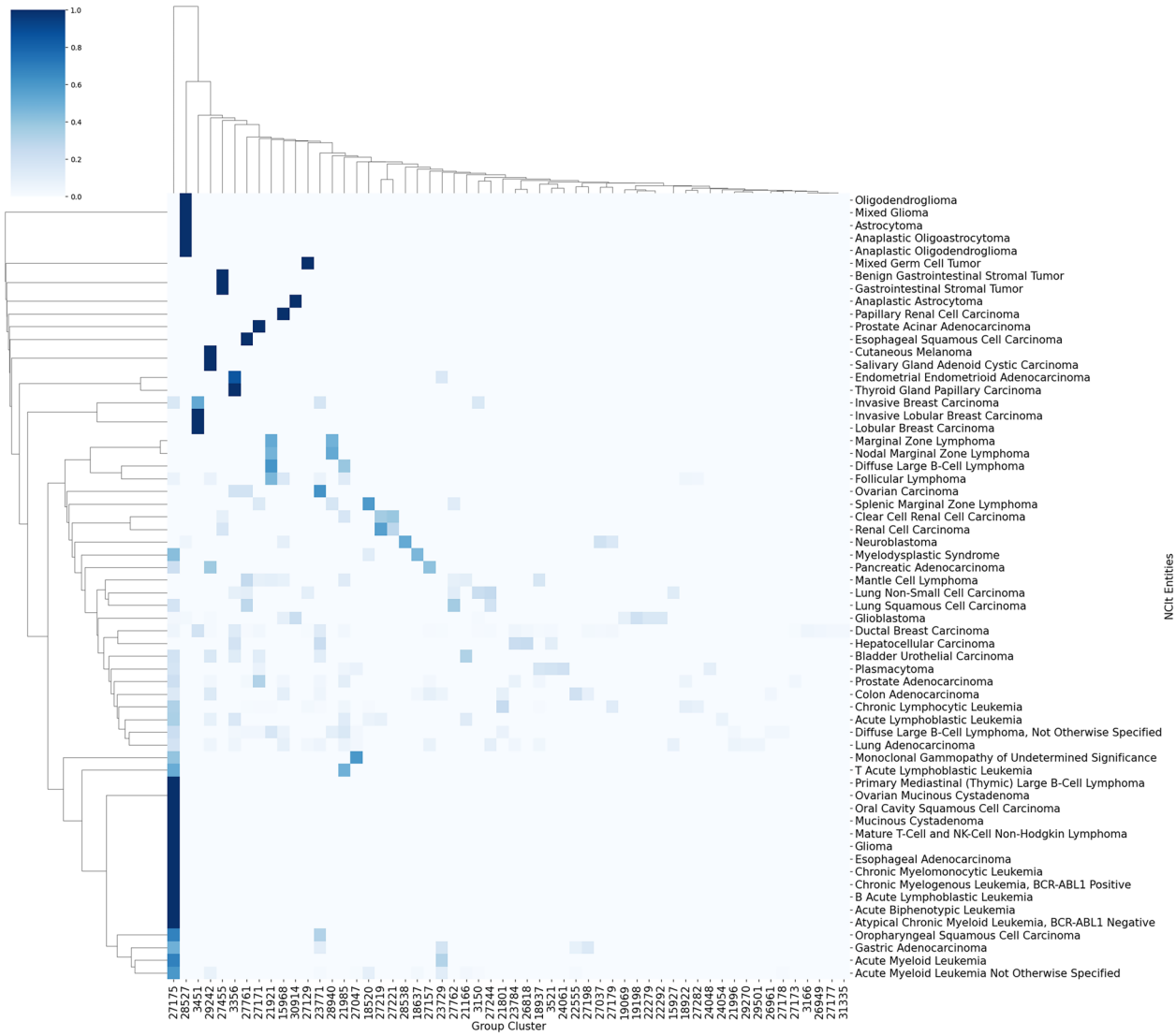
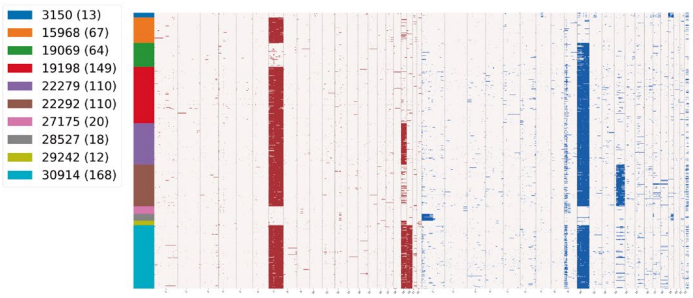


Fig. 4 NCIt entity distribution after the first clustering. The numbers indicate the proportion of samples of the NCIt entity in the corresponding cluster

molecularly considered a hybrid between IDH-wt and IDH-mutant GBM. One study [30] showed that the molecular landscape of GBMs with at least 30% giant cells is dominated by the impairment of TP53/MDM2 and PTEN/PI3K pathways, and additionally characterized by frequent RB1 alterations (dup 19) and hypermutation and by EGFR amplification in more aggressive cases. This finding corresponds to group cluster 22292, indicating that these groups may belong to a more precise diagnosis of giant cell glioblastoma.

Follicular Lymphoma

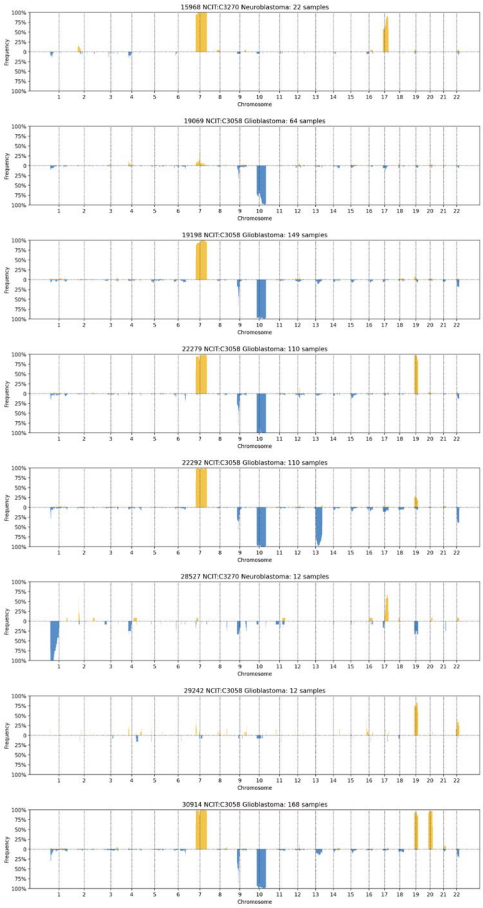
As Fig. 6 shows, Follicular Lymphoma has distinct heterogeneous CNV patterns in its group clusters, which indicates a higher heterogeneity than other cancer types, with features such as deletions on the long arm of chromosome 6 (6q) or gain of chromosome 7q which can affect genes associated with cell cycle regulation and proliferation. Gain of chromosome 8q is associated with aggressive FL subtypes and can be attributed to a selection of an increased copy number of MYC proto-oncogene as an alternative mechanism to chromosomal translocations juxtaposing immunoglobulin promoter regions to the MYC locus in many instances of B-cell lymphomas. Other changes include deletions on 9p, containing the CDKN2A/B tumor suppressor gene locus, as well as gains on 18q in group cluster 21921 oncogene as alternative mechanism to the t(14;18)(q32;q21)



(a) Heterogeneous CNV profiles of Glioblastoma on different clusters. The x-axis indicates the genome, and the y-axis indicates samples. Legends on the left indicate samples assigned to different group clusters. Red and blue colors indicate duplication and deletions, respectively.

group cluster	CNV features
15968	Dup 7
19069	Del 10
19198	Dup 7, Del 10
22279	Dup 7, Del 10, Dup 19
22292	Dup 7, Del 10, Del 13
28527	Del 1p, Del 19q
29242	Dup 19
30914	Dup 7, Del 10, Dup 19, Dup 20

(b) CNV features (defined in Section 1.2) of Glioblastoma on different clusters. Red text indicates literature support.



(c) CNV frequencies of Glioblastoma on different clusters. Description of frequency plots can be found in Fig. 1

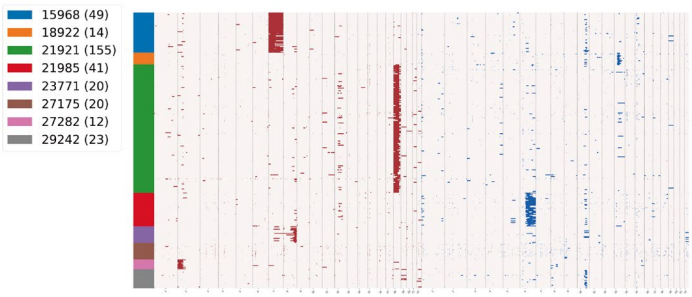
Fig. 5 CNV heterogeneity within Glioblastoma

translocation which is characteristic of follicular lymphoma and leads to constitutive activation of the BCL2 gene by the enhancers of the immunoglobulin heavy chain locus [31].

Neuroblastoma

CNV analysis has revealed significant genomic alterations associated with neuroblastoma tumorigenesis, including amplifications of the MYCN oncogene on 2p, deletions affecting the 1p and 11q chromosomal regions, and gains of the long arm of chromosome 17. Some of these CNV events have been associated with high-risk neuroblastoma cases and poor clinical outcomes. In our observations, gains of chromosome 17q are distributed in all group clusters. Vandesompele et. al. [32] has identified that there is a clinical distinction between gains of 17q and whole chromosome 17: whole chromosome 17 gain with either a favorable-stage tumor or a tumor with whole chromosome 17 gain diagnosed before age 1 year, show a 100% overall survival compared to a higher risk of gains in 17q. Deletions affecting the 1p and 11q chromosomal regions (group clusters 28538, 27179 and 28527) are associated with a more favorable prognosis, highlighting the prognostic significance of CNV in neuroblastoma.

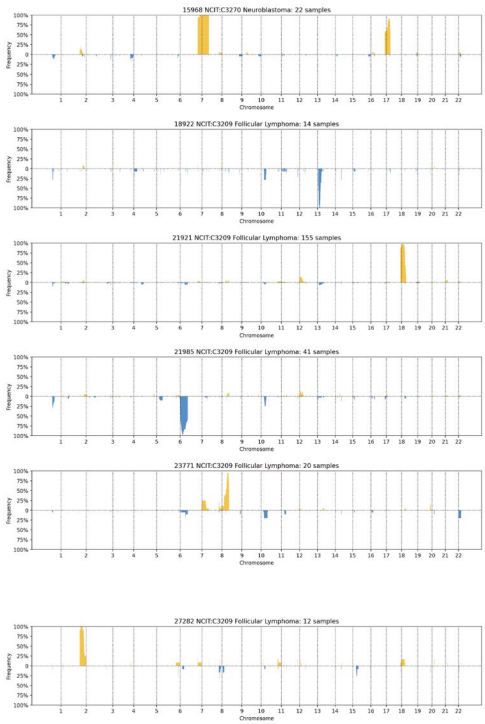
To make a summary, our framework can uncover the CNV heterogeneity within cancers, and these uncovered heterogeneous CNV patterns correspond to known



(a) Heterogeneous CNV profiles of Follicular Lymphoma on different clusters.

Group cluster	CNV features
15968	Dup 7
18922	Del 13q
21921	Dup 18q
21985	Del 6q
23771	Dup 8q
18937	Del 13
21985	Del 6q
23771	Dup 8q
27282	Dup 2p

(b) CNV features of Follicular Lymphoma on different clusters.



(c) CNV frequencies of Follicular Lymphoma on different clusters.

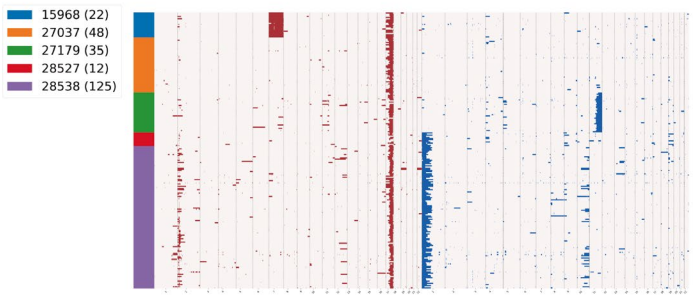
Fig. 6 CNV heterogeneity within Follicular Lymphoma

mechanisms of cancer development, therefore expanding the understanding of the relationships of CNV patterns and cancer subtypes.

Similarity of CNV patterns between cancer subtypes of different diagnostic concepts

Besides intratumoral CNV heterogeneity, the CNV similarities between different tumor entities are proven in the "group clusters". Figure A1–A5 shows the shared CNV patterns of different cancer entities clustered together because of their similar CNV patterns. We summarise the CNV features of cancer entities and use different colors for "group clusters". as shown in Table 1 and 2, carcinoma and adenocarcinoma in different organs can show common CNV patterns. Specifically, ductal breast carcinoma, hepatocellular carcinoma and ovarian carcinoma show multiple similar patterns in group cluster 3356, 3521 and 23784, which indicate the rather high CNV pattern heterogeneity within and CNV pattern homogeneity between these cancers. We also uncover the co-deletion of chromosome 1p/19q on anaplastic oligoastrocytoma, anaplastic oligodendroglioma, astrocytoma, glioblastoma, mixed glioma, neuroblastoma and oligodendroglioma. Besides the existing biological connections between these diseases, these intertumoral CNV homogeneity may also indicate inaccurate classifications in the clinical stage.

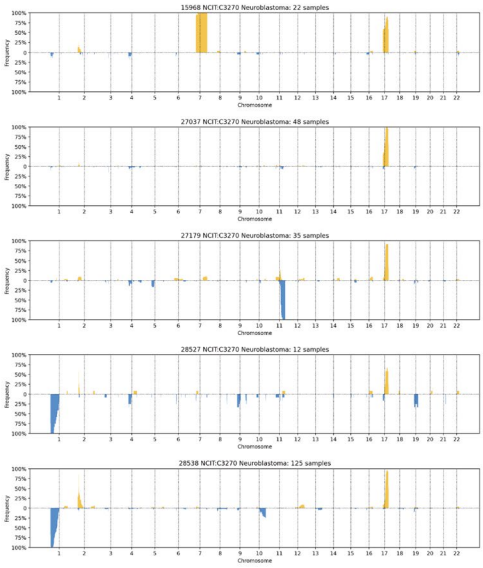
Besides common patterns inside these clusters, as shown in Fig. A6, we notice that there are also similar patterns between "group clusters". Therefore, we calculated the frequency of each cluster and extracted the high-frequency CNV events as features, and applied a second clustering on these frequency data, to summarize the common CNV patterns across the cancer entities based on the first clustering. Figure A7 shows the cluster map of the second clustering, with the combination of the original NCIt label and "group cluster" label of the first clustering. Not surprisingly, group clusters identified in the first clustering are more likely



(a) Heterogeneous CNV profiles of Neuroblastoma on different group clusters.

group cluster	CNV features
15968	Dup 7, Dup 17q
27037	Dup 17q
27197	Del 11q, Dup 17q
28527	Del 1p
28538	Del 1p, Dup 17q

(b) CNV features of Neuroblastoma on different clusters.



(c) CNV frequencies of Neuroblastoma on different clusters.

Fig. 7 CNV heterogeneity within Neuroblastoma

Table 1 Carcinomas with shared CNV patterns

Cancer entity	CNV patterns
Bladder Urothelial Carcinoma	Dup 1q, Dup 8q, Dup 19, Dup 20
Ductal Breast Carcinoma	Dup 1q, Dup 1q, Del 13, Del 8p, Dup 8q, Dup 1q, Del 8p, Dup 8q, Dup 8q, Dup 3q, Dup 19, Dup 20
Hepatocellular Carcinoma	Dup 1q, Dup 1q, Del 13, Del 8p, Dup 8q, Dup 1q, Del 8p, Dup 8q, Dup 8q
Thyroid Gland Papillary Carcinoma	Dup 1q
Invasive Breast Carcinoma	Dup 8q
Oropharyngeal Squamous Carcinoma	Dup 8q
Ovarian Carcinoma	Dup 8q, Dup 3q
Esophageal Squamous Carcinoma	Dup 3q
Lung Non-Small Cell Carcinoma	Dup 3q
Lung Squamous Carcinoma	Dup 3q
Salivary Gland Adenoid Cystic Carcinoma	Dup 19, Dup 20

Table 2 Adenocarcinomas with shared CNV patterns

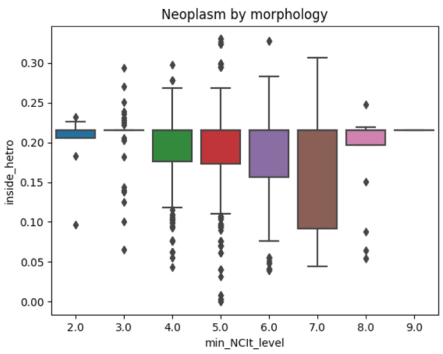
Cancer entity	CNV patterns
Lung Adenocarcinoma	Del 8p, Dup 8q, Dup 8q, Dup 19, Dup 20
Prostate Adenocarcinoma	Del 8p, Dup 8q, Dup 8q, Dup 19, Dup 20
Colon Adenocarcinoma	Dup 8q, Dup 19, Dup 20
Pancreatic Adenocarcinoma	Dup 19, Dup 20

to be clustered in the second clustering, also with some other group clusters. This plot shows the global view of how CNV patterns distribute across NCIt entities and "group clusters", indicating the CNV heterogeneity and homogeneity within/between cancers. We also calculated the frequency of each NCIt pair that was assigned to the same second cluster, to find out the relationships between CNV patterns and NCIt entities. As shown in Fig. A8, for NCIt entities Ductal Breast Carcinoma, Prostate Adenocarcinoma, Hepatocellular Carcinoma, Colon Adenocarcinoma, Plastomacytoma, Invasive Breast Carcinoma, Ovarian Carcinoma, and Garstric Carcinoma, these cancer entities show CNV heterogeneity within cancer, compared to other sub-groups of the same NCIt entity, they are more likely to show more homogeneity with other sub-groups of different cancers. These results further support our findings that carcinoma and adenocarcinoma in different organs could share similar CNV patterns.

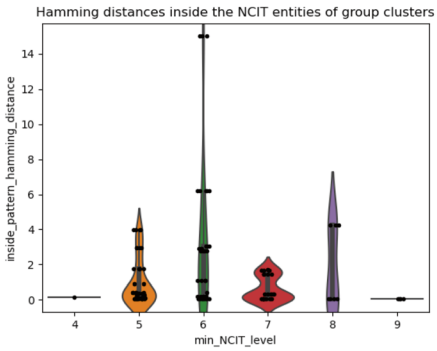
Estimating the overall CNV heterogeneity in the NCIt hierarchical system

Since we have calculated all *sample* distances, as described in Methodology, we can calculate the distances of NCIt *entities* based on the sample compositions. Fig. 8a shows the distribution of CNV heterogeneity inside NCIt entities at different levels of the NCIt system. We need to note that to avoid chaos structure, in this analysis, for those NCIt entities in different NCIt levels, we simply choose lower one as the NCIt level. The level of each NCIt entity is calculated by the shortest distance to the root code (i.e., level 1: Neoplasm). Given that the deeper the NCIt level is, the more specific the disease the NCIt code refers to. We assume that the deeper the NCIT level, the lower heterogeneity the samples within the NCIt code would have. As shown in Fig. 8a, despite the outliers and limited NCIt entities in levels 2 and 9, there is a trend of descending with the increase of NCIt level. The outliers may be due to the CNV heterogeneity within cancers and CNV homogeneity between cancers shown in the last two chapters (2.1 and 2.2).

To compare the CNV heterogeneity within and between cancers with respect to the hierarchical structure of the NCIt system, we calculated the summary of the CNV pattern distance for each NCIt entity in the "group clusters". As shown in Fig. 8b, despite the outliers with much more/less CNV patterns than other NCIt nodes, from level 5 to level 9, the NCIt entities in the group clusters show a similar descending trend with Fig. 8a. An interesting fact is that there are outliers who can both find their parent node in the "group clusters" they are assigned, including Lung Non-Small Cell Carcinoma, Lung Squamous Cell Carcinoma, and Lung Adenocarcinoma. In other words, the similar CNV pattern of the child-parent node should be due to the inaccurate disease classification in the clinical stage. As for the cancers (without child-parent relationships) with similar CNV patterns, they indicate that there could be new cancer subtypes for them for more accurate disease concept representation. In summary, we can conclude that the CNV heterogeneity partly corresponds to the NCIt hierarchical structure, and our framework uncovers the CNV heterogeneity within cancers and CNV homogeneity between cancers.



(a) Inside sample heterogeneity of all NCIt codes in the NCIt system



(b) CNV pattern heterogeneity within NCIt entities of "group clusters"

Fig. 8 CNV heterogeneity on the NCIt classification system

Discussion

Copy number variants (CNVs) are widely recognized as crucial players in the development and progression of cancer, as they can significantly alter gene dosage, disrupt gene regulatory mechanisms, and contribute to genomic instability. Numerous studies have demonstrated that CNV profiles vary not only between distinct cancer types but also within a single cancer type, revealing a high degree of intra-tumor and inter-tumor heterogeneity. For example, CNV events like the amplification of oncogenes or deletion of tumor suppressor genes have been consistently linked to tumor progression, resistance to therapy, and poor clinical outcomes in cancers such as breast, lung, and glioblastoma .

Our findings confirm and extend this understanding by showing that CNV heterogeneity can be quantified and systematically mapped to established cancer subtypes. The application of machine learning techniques allows for a more nuanced exploration of the genomic landscape, enabling us to untangle complex CNV patterns across multiple cancer entities. By identifying distinct CNV signatures, we provide a framework for recognizing cancer subtypes with specific molecular perturbations. These subtypes may not align perfectly with traditional diagnostic classifications based on histopathology or organ of origin. This is in line with studies highlighting that genomic classifications often reveal novel cancer subgroups that might not be distinguishable using conventional clinical or morphological criteria.

Our approach offers several implications for the understanding of treatment response. CNV profiles have been shown to influence sensitivity to targeted therapies and immunotherapies. For instance, the deletion of certain tumor suppressor genes, such as PTEN, has been associated with resistance to immune checkpoint inhibitors. Similarly, amplifications of growth factor receptor genes like EGFR have been linked to responsiveness to tyrosine kinase inhibitors in lung cancer. By characterizing the CNV landscape within specific cancer subtypes, our framework helps identify molecular vulnerabilities that could be exploited for precision medicine, offering a potential roadmap for tailoring treatments to individual patients based on their unique CNV profiles.

Another important aspect of CNV heterogeneity is its contribution to clinical outcomes. Previous studies have correlated high levels of CNV burden with worse prognosis in several cancers, such as ovarian and colorectal cancer. Our study further demonstrates that cancer entities traditionally thought to be distinct may share CNV patterns, which can lead to similar clinical outcomes. This highlights the potential for reclassifying tumors based on their molecular features rather than their anatomical location, as proposed by the "tumor-agnostic" approach to cancer treatment.

By projecting CNV distances onto the hierarchical NCIt classification, we identified a trend of decreasing genomic heterogeneity as we move to higher NCIt levels. Interestingly, we also observed cases where cancer entities located far apart in the NCIt hierarchy displayed surprisingly homogeneous CNV patterns. This challenges the conventional view of cancer classification, suggesting that tumors with similar CNV profiles might share underlying biological mechanisms, even if they originate in different tissues. These findings underscore the need for an integrated approach in cancer diagnosis and treatment, incorporating both genomic and phenotypic data to better capture the complexity of cancer biology.

Overall, our results contribute to the growing body of evidence advocating for broader genomic and molecular profiling strategies in clinical practice. Current diagnostic protocols, which often rely on assessing a limited set of known markers, may overlook critical CNV-driven molecular alterations. Our framework reinforces the importance of comprehensive genomic analysis in the clinical setting, as it can uncover novel therapeutic targets, predict treatment responses, and offer insights into tumor evolution and resistance mechanisms.

In conclusion, the use of CNV profiling not only advances our understanding of cancer heterogeneity but also holds promise for improving patient outcomes by informing more precise diagnostic and therapeutic approaches. Future studies should aim to validate these findings in larger cohorts and explore how CNV profiles interact with other molecular features, such as mutations and epigenetic changes, to influence cancer biology and treatment response.

Abbreviations

CNV Copy number variation
NCIt National Cancer Institute Thesaurus

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13039-024-00692-2.

Supplementary file 1.

Acknowledgements

Not applicable.

Author contributions

Z.Y. wrote the main manuscript, designed the study, and performed data analysis. P.C. contributed to the research question design. M.B. designed the study and contributed data. All authors reviewed and revised the manuscript.

Funding
Z.Y. was a recipient of a grant from the China Scholarship Council. The funder had no role in study design, data collection, and analysis, the decision to publish, or the preparation of the manuscript.

Availability of data and materials
All samples used in the analysis can be accessed by this link: https://zenodo.org/12771229 Supplementary figures and tables can be found in the appendix.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing Interests
None declared.

Received: 23 August 2024 Accepted: 2 October 2024
Published online: 06 November 2024

References

1. Nowell P, Hungerford D. A minute chromosome in chronic granulocytic leukemia. Science. 1960;132:1497.
2. Rowley J. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. Nature. 1973;243(5405):290–3.
3. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. Global variation in copy number in the human genome. nature. 2006;444(7118):444–54.
4. Gao T, Soldatov R, Sarkar H, Kurkiewicz A, Biederstedt E, Loh P-R, Kharchenko PV. Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. Nat Biotechnol. 2023;41(3):417–26.
5. Frank B, Bermejo JL, Hemminki K, Sutter C, Wappenschmidt B, Meindl A, Kiechle-Bahat M, Bugert P, Schmutzler RK, Bartram CR, et al. Copy number variant in the candidate tumor suppressor gene mtus1 and familial breast cancer risk. Carcinogenesis. 2007;28(7):1442–5.
6. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009;10:451–81.
7. Baudis M. Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal cgh data. BMC Cancer. 2007;7:1–15.
8. Gerstung M, Baudis M, Moch H, Beerenwinkel N. Quantifying cancer progression with conjunctive bayesian networks. Bioinformatics. 2009;25(21):2809–15.
9. Li X-C, Liu C, Huang T, Zhong Y. The occurrence of genetic alterations during the progression of breast carcinoma. BioMed Res Int. 2016;2016.
10. Koboldt D, Zhang Q, Larson D, Shen D, McLellan M, Lin L, Miller C, Mardis E, Ding L, VarScan RW. 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. 2012;22,129684, 568–576.
11. Serin Harmanci A, Harmanci AO, Zhou X. Casper identifies and visualizes cnv events by integrative analysis of single-cell or bulk rna-sequencing data. Nat Commun. 2020;11(1):89.
12. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010;463(7283):899–905.
13. Stommel JM, Kimmelman AC, Ying H, Nabioullin R, Ponugoti AH, Wiedemeyer R, Stegh AH, Bradner JE, Ligon KL, Brennan C, et al. Coactivation of receptor tyrosine kinases affects the response of tumor cells to targeted therapies. Science. 2007;318(5848):287–90.
14. Mattfeldt T, Wolter H, Kemmerling R, Gottfried H, Kestler H. Cluster analysis of comparative genomic hybridization (cgh) data using self-organizing maps: application to prostate carcinomas. Anal Cell Pathol. 2001;1(23):29–37.
15. Liu J, Mohammed J, Carter J, Ranka S, Kahveci T, Baudis M. Distance-based clustering of cgh data. Bioinformatics. 2006;22(16):1971–8.
16. Roytman M, Shams S. Brain cancer map: A neural network-based clustering of brain cancer samples based on genome-wide cnv and loh patterns. Can Res. 2021;81(13-Supplement):2171–2171.
17. Li B-Q, You J, Huang T, Cai Y-D. Classification of non-small cell lung cancer based on copy number alterations. PLoS ONE. 2014;9(2):88300.
18. Cavalli FM, Remke M, Rampasek L, Peacock J, Shih DJ, Luu B, Garzia L, Torchia J, Nor C, Morrissy AS, et al. Intertumoral heterogeneity within medulloblastoma subgroups. Cancer Cell. 2017;31(6):737–54.
19. Salto-Tellez M, Cree IA. Cancer taxonomy: pathology beyond pathology. Eur J Cancer. 2019;115:57–60.
20. Fritz A, Percy C, Jack A, Sobin L, Parkin M, editors. International Classification of Diseases for Oncology (ICD-O). 3rd ed. Geneva: World Health Organization; 2000.
21. Guinney J, Dienstmann R, Wang X, De Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015;21(11):1350–6.
22. Taylor MD, Northcott PA, Korshunov A, Remke M, Cho Y-J, Clifford SC, Eberhart CG, Parsons DW, Rutkowski S, Gajjar A, et al. Molecular subgroups of medulloblastoma: the current consensus. Acta Neuropathol. 2012;123(4):465–72.
23. National Cancer Institute: NCI Thesaurus. 2021. https://ncithesaurus.nci.nih.gov/ncitbrowser/. Accessed: 2021-01-07.
24. Jones D, Jager N, Kool M, Zichner T, Hutter B, Sultan M, Cho Y, Pugh T, Hovestadt V, Stutz A, Rausch T, Warnatz H, Ryzhova M, Bender S, Sturm D, Pleier S, Cin H, Pfaff E, Sieber L, Wittmann A, Remke M, Witt H, Hutter S, Tzaridis T, Weischenfeldt J, Raeder B, Avci M, Amstislavskiy V, Zapatka M, Weber U, Wang Q, Lasitschka B, Bartholomae C, Schmidt M, Kalle C, Ast V, Lawerenz C, Eils J, Kabbe R, Benes V, Sluis P, Koster J, Volckmann R, Shih D, Betts M, Russell R, Coco S, Tonini G, Schuller U, Hans V, Graf N, Kim Y, Monoranu C, Roggendorf W, Unterberg A, Herold-Mende C, Milde T, Kulozik A, Deimling A, Witt O, Maass E, Rossler J, Ebinger M, Schuhmann M, Fruhwald M, Hasselblatt M, Jabado N, Rutkowski S, Bueren A, Williamson D, Clifford S, McCabe M, Collins V, Wolf S, Wiemann S, Lehrach H, Brors B, Scheurlen W, Felsberg J, Reifenberger G, Northcott P, Taylor M, Meyerson M, Pomeroy S, Yaspo M, Korbel J, Korshunov A, Eils R, Pfister S, Lichter P. Dissecting the genomic complexity underlying medulloblastoma. Nature. 2012;488(7409):100–5.
25. Jonsson G, Staaf J, Vallon-Christersson J, Ringner M, Holm K, Hegardt C, Gunnarsson H, Fagerholm R, Strand C, Agnarsson B, Kilpivaara O, Luts L, Heikkila P, Aittomaki K, Blomqvist C, Loman N, Malmstrom P, Olsson H, Johannsson O, Arason A, Nevanlinna H, Barkardottir R, Borg A. Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. Breast Cancer Res. 2010;12(3):42.
26. Lapointe J, Li C, Giacomini C, Salari K, Huang S, Wang P, Ferrari M, Hernandez-Boussard T, Brooks J, Pollack J. Genomic profiling reveals alternative genetic pathways of prostate tumorigenesis. Cancer Res. 2007;67(18):8504–10.
27. Baudis M, Cleary M. Progenetix.net: an online repository for molecular cytogenetic aberration data. Bioinformatics. 2001;17(12):1228–9.
28. Huang Q, Carrio-Cordo P, Gao B, Paloots R, Baudis M. The progenetix oncogenomic resource in 2021. Database (Oxford) 2021 Jul 17, 2021.
29. Norouzi M, Fleet DJ, Salakhutdinov RR. Hamming distance metric learning. Adv Neural Inf Process Syst. 2012;25.
30. Barresi V, Simbolo M, Mafficini A, Martini M, Calicchia M, Piredda ML, Ciaparrone C, Bonizzato G, Ammendola S, Caffo M, et al. Idh-wild type glioblastomas featuring at least 30% giant cells are characterized by frequent rb1 and nf1 alterations and hypermutation. Acta Neuropathol Commun. 2021;9(1):200.
31. Rabkin CS, Hirt C, Janz S, Dölken G. t (14; 18) translocations and risk of follicular lymphoma. J Natl Cancer Inst Monogr. 2008;2008(39):48–51.
32. Vandesompele J, Baudis M, De Preter K, Van Roy N, Ambros P, Bown N, Brinkschmidt C, Christiansen H, Combaret V, Łastowska M, et al. Unequivocal delineation of clinicogenetic subgroups and development of a new model for improved outcome prediction in neuroblastoma. J Clin Oncol. 2005;23(10):2280–99.

Publisher's Note