Explainable deep learning on 7500 whole genomes elucidates cancer-specific patterns of chromosomal instability

Mohamed Ali al-Badri[1*], William CH Cross[1,2] and Chris P Barnes[1*]

[1]Department of Cell and Developmental Biology, University College London, Gower Street, London, WC1E 6BT, UK.
[2]Department of Research Pathology, University College London, Gower Street, London, WC1E 6DD, UK.

*Corresponding author(s). E-mail(s): m.al-badri@ucl.ac.uk; christopher.barnes@ucl.ac.uk;

Abstract

Chromosomal instability (CIN) refers to an increased rate of chromosomal changes within cells. It is highly prevalent in cancer cells and leads to abnormalities in chromosome number (aneuploidy) and structure. CIN contributes to genetic diversity within a tumour, which facilitates tumour progression, drug resistance, and metastasis. Here, we present a deep learning method and an exploration of the chromosome copy aberrations (CNAs) resultant from CIN, across 7,500 high-depth, whole genome sequences, representing 13 cancer types. We found that the types of CNAs can act as a highly specific classifier for primary site. Using an explainable AI approach, we revealed both established and novel loci that contributed to cancer type, and focusing on highly significant chromosome loci within cancer types, we demonstrated prognostic relevance. We outline how the developed methodology can provide several applications for researchers, including drug target and biomarker discovery, as well as the identification of cancers of unknown primary site.

Keywords: cancer, deep learning, chromosome copy number, aneuploidy, chromosome instability, clonal evolution

Introduction

Chromosomal instability (CIN) is a phenomenon commonly observed in cancer cells, characterized by an increased rate of changes in both the number and structure of chromosomes. This instability can result from a wide range of mechanisms, which include those directly associated with mitotic apparatus, such as chromosome mis-segregation, but also less well understood events such as dysfunction of telomeres, impairments in cell cycle checkpoints, abnormalities in the mitotic spindle apparatus, double stand break repair and genome doubling. Chromosome copy aberrations (CNAs), the specific loci that reflect an aneuploidic state, have been observed in around 90% of solid cancer genomes [1] making them one of the most prominent features of the cancer genome. The observed chromosome profile of a given cancer is a reflection of these mechanisms plus its unique evolutionary walk, which is a combination of the positive selection forces of its microenvironment (defined in a major part by the cite of origin) and the stochastic nature of mutagenesis.

For this reason, CNA patterns have been increasingly studied. Following the historical work of von Hansemann [2] and Theadore Boveri [3, 4] and more recently *in vitro* [5], *in vivo* [6], and bioinformatic studies [7] a view of chromosome imbalances (aneuploidy) as a general potentiate of malignancy has been constructed [8]. More recent studies have sought to explore the genome-wide architecture that results from CIN [9] and to explain patterns of CNAs through generative signatures [10, 11]. Although the signature approach is appealing there are many hurdles associated with defining the underlying features to use, in contrast to the point mutation case [12].

One of the challenges associated with CIN is its complexity. Separating functional vs non functional variation is non-trivial. In a sense, many researchers are seeking to classify those sites that are important to cancer biology, and those that are passengers resultant from stochastic forces. An ideal situation is that specific loci define each cancer subtype, thereby directing drug development, and making molecular diagnosis tractable. To date, many cancer-specific CNAs have been identified[13, 14] and in specific gene contexts, such as TERT amplifications [15] they can act as biomarkers for disease severity, drug resistance [16], or as novel drug targets [17]. That said, these are often discovered from frequency or mechanistic methodologies and it remains broadly challenging to assess whole cancer genomes for undiscovered loci that are critical to cancer biology.

In this work we sought to take an approach where we examined CIN patterns directly. We classified cancers based on patterns of copy number across the genome and then used explainable AI to dissect the loci that contributed information to the classification. These loci capture differences between cancers, which we can assume gives rise to different biology in those cell types. Our method reveals the "unseen" differences in CNA patterns between cancer types, and reveals novel loci that will be of wide interest to cancer biologists, and drug and biomarker developers. We verified the prognostic information in these loci using patient data.

Results

Analysis of 13,000 rearranged cancer genomes in the 100,000 genomes project

The 100,000 Genomes project was completed in 2019, and achieved its goal of sequencing 100,000 genomes from cancer, rare disease and infectious disease patients [18, 19]. Managed by Genomics England, it served as a transformative project for integrating genomics into the National Health Service (NHS) and to evaluate the role of whole genome sequencing (WGS) at scale for NHS cancer patients. Participants provided consent for their genomic data to be associated with annonymised longitudinal health records. This information is then shared within a secure workspace with international clinicians and scientists from approximately 300 institutions in 24 countries within the Genomics England Clinical Interpretation Partnership (GeCIP). The aim is to advance our understanding of the causes and treatment approaches for various types of cancers[20].

The 13,000 WGS cancer samples are labelled according to their primary site and diagnosis. Primary site classes with fewer than 100 samples (testicular, endocrine, etc.) or those whose class labels are broad, or composed of multiple primary sites (hepatopancreatobiliary, sarcoma, cancer of unknown primary site, etc.) were omitted from the data. The remaining genomes were used in our analysis (Figure 1a).

We also focused on cancers with a high degree of chromosomal instability (CIN) by excluding those with less than 30% of the genome altered, as quantified by the Percentage Genome Altered (PGA). This threshold aligns with previous work suggesting that a substantial genomic alteration is a reliable indicator of increased chromosomal instability [21, 22]. The distribution of PGA across cancer samples from diverse primary sites is given in Figure 1b. Furthermore, we excluded genomes that had undergone significant rearrangements or primarily influenced by mutations other than CNAs, since these are likely to involve distinct biological mechanisms such as microsatellite instability. This left 7522 genomes for the downstream analyses.

We explored the frequency of CNA mutations at known pan-cancer genes from Rosetta [23]. Figure 1c shows the number and the proportion of samples where each locus is implicated. Simultaneously, each pan-cancer gene has an associated stacked bar of the copy state for each sample in the GEL data, showing the proportion of samples with deletions, loss of heterozygousity (LOH), duplication, amplification or diploid.

The degree of genomic instability, which is highly sensitive to the cancer primary site, can be measured by the number of copy number segments, the proportion LOH segments within the genome and the degree of whole genome doubling (WGD) [10]. This is illustrated in Figure 1d which shows median proportion of LOH against the median number of LOH segments, with

the size of each point reflecting the degree of WGD. The ratio of mean proportion to number of LOH segments is lowest in haematological cancers and highest in upper-gastrointestinal samples, where WGD is inversely highest in haematological and lowest in upper-gastrointestinal samples.

A novel deep learning framework for analysis of rearranged genomes

We developed a model for analysing rearranged genomes using a neural network classifier trained on the primary site labels in the GEL dataset. To prepare the data, we first split the autosomes into 100kb bins and calculated the median total copy number and median minor allele CNA state within each bin. This reduced each sample into a 28749×2 tensor and provides a conserved genome coordinate across all samples. The data were then scaled across samples. The neural network architecture itself, is composed of an ensemble of 22 chromosome-specific sub-networks, separately acting on each chromosome in the input (Figure 2a). As such, each chromosome sub-network has an input layer of nodes equal to the binned size of the respective chromosome, with chromosome 1 the largest number of nodes and chromosome 21 with the fewest. These chromosome feature extractors then serve as inputs to a deep feed-forward network.

Each sample in our dataset is composed of a large feature set, while the classification involves only 13 classes. This high-dimensional feature space posed a challenge for our model, requiring it to discern meaningful patterns efficiently. Navigating through such a large feature space demands a robust model capable of identifying relevant patterns while avoiding overfitting. The chromosome feature extraction prevented overfitting, as did dropout layers with high probability ($p = 0.5$). In addition, we chose not to use convolutional layers in order to improve downstream interpretability. Our training strategy (see Methods), also addressed this challenge by striking a balance between model complexity and generalisation.

Many cancers can be distinguished based on their copy number profiles

The results in Figure 2b showcase the model's ability. Despite the abundance of features, the model demonstrates nuanced performance across classes, highlighting its capability to distil information from the diverse set of features. The average classification accuracy of the model on the test set is 72%, with endometrial cancer samples having the highest misclassification. As with other cancers, samples are misclassified when primary sites share similar biology, in this case with gynaecological cancers or colorectal cancer and upper gastrointestinal cancers. The performance of this model on classifying samples based on the CNA distribution of samples shows that there are differences in chromosomal instability (CIN) patterns between cancers. However, this observation

is not universally applicable to all cancers. For example, lung cancer emerges as one of the classes with the lowest classification accuracy, which is most likely due to the mix of lung cancer subtypes within the data. These were not supplied in the training and thus the heterogeneity negatively influences the predictive accuracy of the model.

Using explainable AI to locate functional variation within genomes undergoing CIN

In order to infer the most important features that determine classification of primary sites by the trained network, we employed an approach based on Integrated Gradients (IG). IG is a technique used in explainable artificial intelligence (xAI) to understand the predictions of neural networks by attributing importance scores to input features [24]. Although there have been a number of improvements to the approach [25], it can still be challenging to interpret the information gained. We used the guided integrated gradients (GIG) method [26]. Unlike conventional IG, adaptive path methods such as GIG reduce the noise accumulated along the IG path by defining a path integral between the baseline and input sample, conditioning the path not just on the input sample but also on the model being explained (Figure 2c). This conditioning ensures a more accurate attribution of feature importance, taking into account the specific characteristics of the neural network's decision boundaries. By utilising GIG, out aim was to provide a more nuanced understanding of the features contributing to the classification of primary sites, enhancing the interpretability of the model's decision-making process. The resulting feature attributions are numerical scores, ranking the input features based on their influence on the model's prediction. This ranking gives the relative importance of each genomic locus in contributing to the classification of an individual cancer primary site versus the other primary sites. As such, it does not – by design – highlight the pan-cancer features such as those listed in Figure 1c, and instead serves as a powerful tool for interpreting the understanding the discriminative factors guiding the classification. This approach aids the identification of biologically relevant CNA factors driving CIN within each sample.

The GIG inference on the trained model is applied individually for each sample, generating a ranking of genomic bins. The model is trained exclusively on the CNA states of each bin, with none of the associated metadata such as gene or enhancer content fed into the training. By aligning the inferred feature importance with specific genomic annotations, and examining whether a feature corresponds to a specific copy number alterations (e.g. duplication, deletion, LOH), our approach provides a spatially and biologically contextualised understanding of the model's decision-making. This can be used to capture distinctions between cancers, interrogating the complex landscape of cancer biology within the context of CIN.

Sample level analysis reveals focal loci shaping tumour biology

The sample level explainability provides insight into the loci within the tumour genome that contribute to that particular classification. Figure 3A, B and C shows examples from breast, lung, and colorectal cancer respectively. Each panel contians the CNV distribution across the genome with the highly attributed loci highlighted in colour for both the total CN and minor allele CN. The accompanying tables show the 10 highest loci for each sample and their corresponding copy state, cytoband, COSMIC oncogenes and enhancers (see Methods).

The sample in (Figure 3a) is correctly predicted to belong to the breast cancer cohort with a classification probability of 0.96. The attributions point to a range of loci that are mostly established in the literature. In the total CN channel, amplification at 8p11.22 is observed, although only limited literature provides evidence that this region has a role in breast cancer [27, 28]. Similarly, amplification at 16p13.12, associated with ERCC4, appears to be novel in the context of breast cancer. The loci with the highest attribution in the minor allele are also duplicated or amplified in this patient. In a subgroup of estrogen receptor (ER)-negative grade III breast cancers, the 19q12 locus emerges as amplified. This amplicon, housing nine genes including cyclin E1 (CCNE1), is proposed as a potential 'driver.' The survival of cancer cells displaying this amplification requires the expression of multiple genes, emphasising the complexity of the 19q12 amplicon [29]. Located in the 17q24.1 region, AXIN2 is identified as a target and negative feedback regulator of Wnt signalling. This gene is enriched in the Wnt-responsive cell population with Axin2+, particularly in mammary stem cells in the adult mammary gland [30–32]. Chromosome arm 17q exhibits a high frequency of rearrangement mutations in breast cancer, with c-erbB-2 gene amplification at 17q11.2–q21.1 having been established as a common variant that have also been correlated with poor prognosis [33, 34]. Regions on chromosome 17, including 17q21.31, 17q21.2, and 17q25.3, show amplification without definitive evidence in the current literature. These regions present potential areas for exploration, warranting further research into their role in breast cancer pathogenesis.

The lung cancer sample in (Figure 3b) is classified with a probability of 0.86 and points to the highest attributions in amplifications in chromosomes 5,11,14, duplications in chromosome 17 and LOH in chromosome 19. In particular, the amplification at 5p15.33 encompasses the genes TERT and CLPTM1L, implicating their potential role in lung cancer aetiology. Previous studies have highlighted the importance of gene amplification in hTERT, establishing it as an independent poor prognostic marker for disease-free survival in non-small cell lung cancer (NSCLC) [35–37]. Within the region of 14q13.2–14q21.1, a minimal common region of gain containing MBIP, NKX2-1, NKX2-8, and PAX9 has been identified. The cooperative action of these genes

is implicated in lung tumourigenesis, particularly in the context of lung adenocarcinoma in never smokers [38]. The amplification at 5p12.2 emerges as a potential early event in NSCLC, with gene amplification in this region strongly associated with development. This underscores the significance of 5p12.2 as a key player in the molecular landscape of NSCLC [39]. While specific details regarding the amplification at 11q21 are limited, translocation and gain events at this locus indicate its potential importance in the genomic alterations associated with cancer[40]. Identified in a comparative genomic hybridisation study of squamous cell carcinomas (SCC) of the lung, the amplification at 14q12 reveals chromosomal imbalances associated with the metastatic phenotype in lung SCC [41]. Patients harbouring gains of 11q23.3, 11q13.1, and 14q32.3, along with deletions of 3p21.3 and 9p21.3, tend to have poor survival outcomes. These findings emphasise the prognostic implications of genomic alterations in these specific loci [42]. Recurrent alterations, including B2M inactivation and inactivation of antigen presentation-related genes such as CALR, have been linked to increased immune evasion. The loss of BRG1 in NSCLC cells further demonstrates its impact on cellular morphology and tumourigenic potential [43, 44]. Overexpression of CCR7 in non-small cell lung cancers is highly associated with lymph node metastasis, underscoring its pivotal role in tumour cell migration [45]. The association of 17q11.2 with HER2 is a critical prognostic and predictive marker in breast cancer, particularly indicating poor breast cancer-specific survival [46–48]. Duplications in the 17q21.2 region involve a number of keratins, which are otherwise ised as diagnostic tumour markers due to epithelial malignancies retaining keratin patterns associated with their respective cells of origin [49]. Genes such as KRT19 and KRT16 play significant roles in regulating the reprogramming of tumour stem cells, impacting tumour stem cell markers and influencing the growth and metastasis of lung cancer [50, 51].

Finally, the colorectal cancer (CRC) sample in (Figure 3c) is classified with a probability of 0.84 and largely separates its total allele CN and minor allele CN attributions between amplifications and LOH, respectively. The amplification at 13q12.2, specifically involving the CDX2 gene, corroborates a previous study that challenged conventional notions of CDX2 as a tumour suppressor in colorectal cancer, this study has instead identified CDX2 as an amplified lineage-survival oncogene required for the proliferation and survival of colorectal cancer cells and implicated Wnt/β-catenin signalling, itself a key oncogenic pathway in colorectal cancer [52]. In the 13q34 cytoband, IRS2 has previously been identified as a likely driver oncogene, offering a potential mechanism for PI3 kinase pathway activation in CRC [53]. A genomic duplication spanning 13q12.2-q12.3, as identified using Genomic Identification of Significant Targets in Cancer (GISTIC), encompasses several genes, including PDX1, ATP5EP2, CDX2, PRHOXNB, FLT3, LOC100288730, PAN3, and FLT1 [54]. Limited literature exists regarding the specific details of 19q12 loss of heterozygosity (LOH) in CRC, with a mention of CCNE1 as a potential player in this context. Cyclin E (CCNE1), a regulator of the cell cycle, is overexpressed in

many human tumours, including CRC and has been proposed to lead to CIN and tumourigenesis by uncoupling cell cycle progression from the regulation of mitosis.[55]. This gene was also found to be linked to CIN in a recent study modelling rearrangement formation [56]. In the 17q24.1 region, the association of AXIN2 with CRC is established, with variations in AXIN2 potentially serving as a risk marker for predisposition and prognosis [57]. The role of 17q24.2, housing PRKAR1A, remains inconclusive, with some reports suggesting its potential as a tumour suppressor gene [58]. Limited evidence is available regarding the specific implications of the 19p13.12 locus in CRC. In summary, these results showcase the ability of the xAI approach to find real biological signatures that can provide insight into tumour biology.

Cohort level analysis reveals novel loci implicated in tissue specific tumour biology

To explore the high attribute loci from individual samples across entire cohorts within a specific primary site, we first aggregated all ranked loci based on their frequency within each population (see Methods). Figure 4 displays some of these results for the the breast, lung and colorectal cancer cohorts. Figure 4a-c shows the chromosome arm level analysis for these cancers and highlights prominent large-scale events, such as whole arm deletions already known.

In breast cancer, regions 16p, 17q, and 1q exhibit high-density signals for total copy numbers, while both arms on chromosomes 11, 16, and 17 show high-density signals for minor allele copy numbers (Figure 4a). In breast cancer, abnormalities of chromosome 17 have long been recognised as pivotal in tumourigenesis. Alterations at specific loci on chromosome 17, including ERBB2 amplification, P53 loss, BRCA1 loss, and TOP2A amplification or deletion, are known to play crucial roles in breast cancer pathophysiology. Numerical aberrations of chromosome 17 are intricately linked to breast cancer initiation, progression, and treatment response [59]. Moreover, whole arm gain of chromosome arm 1q and loss of chromosome arm 16q are among the most frequent genomic events observed in breast cancer. Chromosome arm 16p, found to be a translocation partner to 1q, is assumed to be an early event in breast cancer progression [60]. The gain of chromosome 1q has been mechanistically linked to increased expression of MDM4 and suppression of TP53 signalling. Notably, TP53 mutations are mutually exclusive with 1q aneuploidy in human cancers, suggesting specific aneuploidies play essential roles in breast cancer tumorigenesis [61]. Furthermore, deletion of chromosome arm 11q has been observed in breast tumours. While tumours with 11q deletion do not exhibit a more aggressive phenotype or genotype distinguishing them from those without this deletion, its association with relapse in patients with lymph node–negative breast cancer who did not receive anthracycline-based chemotherapy presents a novel avenue for cancer treatment [61].

For lung cancer, our results underscore the significance of regions on the 5p arm for total copy number and the 9p and 19p arms for the minor allele

copy number (Figure 4b). Previous molecular cytogenetic studies have consistently demonstrated chromosomal aberrations on the short arm of chromosome 5 in all major lung tumor types. Notably, gains on 5p have been identified as among the most frequent alterations observed in small cell lung cancer (SCLC), underscoring the importance of this region in lung tumorigenesis [39, 62, 63]. Furthermore, chromosome arm 9p has emerged as a critical target in lung cancer, with studies highlighting it as the most frequent site of homozygous deletions. This observation suggests that chromosome 9p harbors multiple tumor suppressor genes and/or genomic features fragile during lung carcinogenesis, highlighting its significance in lung cancer pathogenesis [64]. In addition, frequent deletion of chromosome arm 19p has been reported in lung cancer. Studies have shown that simultaneous mutations at LKB1 and BRG1 are common in lung cancer cells, providing evidence of how a single event, such as loss of heterozygosity (LOH) of chromosome 19p, can target multiple tumor suppressors, contributing to the complex molecular landscape of lung cancer [65].

Finally, the colorectal cancer cohort demonstrates a large-scale region of high attributes at 20q and 13q for total copy number signals, and 20q, 18q, and 17p for minor allele copy number signals (Figure 4c). A frequent genomic event observed in CRC is the duplication of chromosome arm 20q. Studies have consistently reported that gain of 20q is observed in more than 65% of CRCs and is associated with poor clinical outcomes. This region harbours multiple oncogenes, such as BCL2L1, AURKA, and TPX2, which play crucial roles in CRC initiation and progression [66–69]. Additionally, genes such as PLAGL2 and Protein POFUT1 have been identified as driven by copy number amplification on 20q, highlighting their cancer-causing potential in CRC [67, 70, 71]. These findings suggest that genes on chromosome arm 20q may serve as highly specific biomarkers for CRC with potential clinical applications [66]. Furthermore, LOH of chromosome arm 18q is a common CIN signature in CRC. SMAD4, present in 18q21.1 and encoding a downstream signal transducer of TGF-β, has been implicated as an important tumour suppressor lost by 18q deletion. 18q deletion has been associated with worse prognosis in patients with CRC, underscoring its clinical significance [72–75]. Additionally, chromosome arm 13 duplication has been identified as part of other arm-level events associated with colorectal adenoma to carcinoma progression. Combinations of genetic aberrations, including 17p loss, KRAS mutation, 8q and 13q gain, and 18q loss and 20q gain, are associated with progressed colorectal adenomas and CRC. These findings suggest the existence of multiple independent chromosomal instability pathways in CRC progression [76]. Moreover, deletion of chromosome arm 1p has been implicated in CRC progression, with primary carcinoma cells with metastatic ability often exhibiting 1p deletions. This suggests that 1p loss may be important both early and late in CRC carcinogenesis [77].

Circular Manhatten plots also show how these loci are distributed across the genome (Figure 4d-f). The identification of such focal events provides a targeted view into the genomic intricacies of the primary site, facilitating a

more comprehensive understanding of the underlying biology and potential avenues for therapeutic exploration.

To visualise the attribution distributions across all samples within a cohort and depict the arrangement of consistently highly ranked features, we created heatmaps (Figure 4g-i). The heatmaps illustrate the high attribute features based on their copy state, with the top section representing total copy number and the bottom section representing minor allele copy number. Loci that do not rank high in importance are left uncoloured, appearing as white on the heatmap. To validate our results, we extracted the top ranked cytobands for breast, colorectal, and lung cancer cohorts by selecting all events present in > 50% of samples. We used existing databases to find and verify these regions. Every single region aligned with well known CNA features associated with each cancer type (Supplementary Tables 1-3).

Prognostic value of xAI approach

We considered whether the two primary outputs of our deep learning method, the high attribute loci and the cancer prediction scores, have prognostic value. Plausible scenarios are that a loci was discovered within a particular cancer type because it induces phenotypes associated with aggressive disease. Although only limited numbers of CNAs currently have prognostic value, we wanted to examine any new potential in our loci sets, many of which are novel findings. We also considered that the prediction scores, which reflect the karyotypic closeness of a particular patient's cancer to an archetype, may also have prognostic significance. This could be due to better treatment of archetypal diseases, perhaps due to drug suitability. Alternatively in patients with low prediction scores diseases might involve some cross-over in pathogenesis, as observed in some gyncological and gastro-intestinal diseases.

To test the prognostic significance of high attribute loci, we performed an outcome analysis using standard statistics, restricting each test set to a given cancer type and to higher stage diseases (see Methods, n = 5168). Following stage correction, we found that 42 loci were significant, though following multiple testing correction, this was reduced to 4 loci (6q23.1, 6q23.2, 10p14, 13q34) all in Breast cancer (Figure 5a). We noted that the hazard ratio of these loci was only slightly lower than stage, demonstrating the value of CNAs in some diagnostic settings.

Next we explored the cancer prediction scores, again in higher stage disease, and separately in specific cancers. We partitioned each set of patients into high and low cancer prediction scores using the median within each cancer type, and then used these groups to assess differences in outcome. Surprisingly, we found that the effect for a high prediction score was a worse prognosis in adult glioma but a better prognosis in melanoma (Figure 5b). We wondered whether this link between prediction score and outcome might be reflecting a rediscovery of some underlying genetic factor, and in examining ploidy, tumour content, age,

the distributions of sex, we found that the high-low Glioma groupings were different by age and that melanomas by ploidy.

Discussion

We developed a methodology to explore rearranged genomes using deep learning and explainable AI approaches. We then applied this to 7500 genomes from the 100,000 Genomes Project. We showed that cancers can be classified with reasonable accuracy based purely on their copy number profiles, indicating that patterns of CNA have some cancer specificity. We demonstrated that guided integrated gradients can be used to pull out biologicals signals that have prognostic value.

The significance of this research lies in the use of explainable AI to explore biological function. Although deep learning approaches have been used to classify cancer genomes [78], and studies have looked at how CNAs perform in this context [79], no other studies have used techniques to explore tissue specific genome changes. This bespoke approach to genomic analysis, wherein the AI model independently learns from genomic data, ensures an unbiased exploration of the genomic landscape. The work has potential clinical implications, as it may contribute to the development of molecular diagnostics and the identification of cancer-specific drug targets. This research not only advances our understanding of the complex genomic landscape of cancer but also aligns with broader efforts to enhance personalised and effective treatment strategies in oncology, marking a significant contribution to the field of genomic medicine.

There are some limitations associated with explainable AI approaches, including the fact that the usefulness of the information depends on the predictive performance of the model on that sample [80]. However, in our current approach using classification, we can filter on classification probability therefore retaining samples with confident predictions. Moreover, while we used classification, it should be acknowledged that class labels can be fluid, particularly in instances of close proximity between primary sites, as observed in the case of ovarian and endometrial cancers. The genomic landscape, shared mutational profiles, and inherent biological similarities between these proximate primary sites may contribute to challenges in precise classification. Future work will involve the use of unsupervised learning approaches that can alleviate these biases. Additionally, the incorporation of cancer subtype information could also be beneficial under some scenarios.

In summary, the effective classification of cancer primary sites based solely on CNA data is a testament to their significance in cancer biology. Our new approach serves as a valuable tool for further research into the mechanisms of CIN, potentially revealing novel aspects of cancer biology and progression.

Methods

Data and preprocessing

The Genomics England 100,000 Genomes Project was an NHS transformation project that took place between 2016 and 2019 [81]. During this time nearly 20,000 participants with cancer consented for the study and had paired blood and tumour biopsies DNA sequenced to 50 and 100X respectively. The bioinformatic processing utilised an accredited bioinformatic pipeline, which included industry-standard tools and in house quality [82].

The Genomics England (GEL) CNA data using in this project was produced using the CANVAS tool [83]. We visualised the relationship between the median proportion of the genome impacted by LOH and the median number of LOH segments in each cancer primary site by mapping genomic states (such as deletion, diploid, LOH, duplication and amplification) to integers. This mapping facilitated the identification and quantification of continuous stretches of each copy state within the samples. LOH segments were specifically extracted and analysed to determine their distribution across the genome. The degree of whole genome doubling (WGD) was calculated using the PCAWG approach [84]

GEL primary sites with fewer than 100 sample are insufficient for training a neural network (testicular, endocrine, etc.) or those whose class labels is composed of multiple primary sites or does not fall under the same convention (hepatopancreatobiliary, sarcoma, cancer of unknown primary site, etc.) were omitted from the analysis. Furthermore, we determined the threshold for CNA-driven cancers to have at least 30% of their genome altered (PGA).

The CNA calls for each patient were organised into tabulated data indicating the CNA state by genomic coordinate for both alleles, along with the major allele. To prepare these data for use by the neural network, we conducted a series of transformations. Each sample was anonymised and processed to generate an individual array, wherein the copy state of the 22 chromosomes was binned into 100 kb segments, mapped to a single array element. Within each sample, two channels were encoded: one for the total allele CNA state and another for the minor allele CNA state. The resulting array has shape (2, 28749). This data structure has a conserved genome coordinate across all samples, where each element/bin denotes the copy state of a given genomic region of 100 kb. Prior to training the network, the dataset was feature-wise normalised by subtracting the mean and dividing by the standard deviation.

Deep learning approach

The neural network was trained using PyTorch 1.9.1.[85] The network architecture is composed of an initial ensemble of 22 chromosome-specific sub-networks, separately acting on each chromosome in the input (Figure 2). As such, each chromosome sub-network has an input layer of nodes equal to the binned size of the respective chromosome, with chromosome 1 the largest number of nodes and chromosome 21 with the fewest. The model was not trained

on either X nor Y chromosomes. The chromosome feature extractors then serve as inputs to a deep feed-forward network. The network's input layer comprises 1,100 nodes, and it consists of three layers with 1000, 50 and 13 nodes, respectively. Each layer is composed of linear, dropout, and LeakyReLU layers, providing the network with a hierarchical structure. The output layer, with 13 nodes, aligns with the number of cancer primary sites in the GEL dataset and employs a SoftMax function to produce the final predictions.

During training, a weighted sampler was used to address the class imbalance in the GEL data (see Figure 1a). The neural network was trained using the stochastic gradient descent optimiser with a learning rate of 0.001. The cross-entropy was used as the loss function. The training process involved a batch size of 100 samples per iteration. The training dataset was randomly split into 70% training and 30% validation sets to monitor the model's performance during training. The model underwent training for 50 epochs when terminated by early stopping to prevent overfitting. During training, dropout layers with a rate of 0.5 were utilised to enhance model generalisation.

An architecture such as this tends to be less favourable for applications that necessitate handling geometric transformations, including but not limited to permutation invariance, particularly in domains like image data. In contrast, our proposed approach leverages the independence of each input feature. By encoding the data into a unified genomic coordinate space, this methodology facilitates the learning of mutational patterns. Consequently, it enables interpretability regarding each sample's class without the necessity for metadata associated with individual input features.

AI explainability

In order to infer the most important features that determine classification of primary sites by the trained network, we employed the guided integrated gradients (GIG) method.[26] Unlike conventional integrated gradients (IG), adaptive path methods such as GIG reduce the noise accumulated along the IG path by defining a path integral between the baseline and input sample, conditioning the path not just on the input sample but also on the model being explained (Figure 2). In our model, the input data is highly disparate across samples, therefore the explainability results can be sensitive to the choice of baseline. Namely, a conventional choice of baseline for image data may take the form of a black image or an array of zero values, which is counterproductive for data composed of (feature-normalised) CNA states. Instead, the median CNA state of a sample across a given channel is chosen as a baseline to start the GIG path integral, ending with the sample CNA state values.

Cohort-level explainability analysis

In order to aggregate the explainability score across each primary site cohort we selected the top 1% of loci within a sample and then selected those resulting loci

that were present in more than 5% of the cohort. This conservative approach allowed us to visualise patterns across the primary sites and to verify known biological features (Figure 4).

To provide a more statistically robust assessment of the significant loci for the clinical outcome analysis, we performed an association study following previous work [25]. This comprised a loci dependent t-test on the raw attribute signal where the null distribution was a sample of attribute scores at that loci in other primary sites. The null distribution contained a number of points equal to the number of samples in the test cohort. In order to reduce variability in this testing procedure we repeated the t-test ten times and calculated the average p-value. The p-values were corrected using the Bonferonni multiple testing correction. Loci were further processed to remove chromosome arm level regions leaving only focal events for the clinical outcome analysis.

Clinical outcome analysis

For the majority of patients outcome data was available, including date of death and time interval of followup. We focused on higher stage disease which are associated with fatal conditions (stage > 1, $n = 5168$). Data from the high attribute loci and cancer prediction scores were subjected to Kaplan-Meier (log rank) and cox proportional hazard testing, the latter correcting for cancer stage in all type bar Glioma. The R Survival and Survminer packages were used for this purpose, and multiple testing correction was performed using the Benjamin Hochberg method on any reported p-values. Visualisation was handled by the standard plotting functions included in the packages: `ggsurvplot`, and `ggforest`.

Supplementary information. Supplementary figures are provided in the accompanying file. Supplementary Figures 1-13 show the classification probability for the 13 cancers. Supplementary Figure 14-26 show the circular Manhattan plots for the 13 cancers. Supplementary Figure 27-39 show the bubble plots. Supplementary Figure 40-65 show the heatmaps.

Acknowledgments. MAB and CPB were funded through a Wellcome Trust Senior Research Fellowship (209409/Z/17/Z). Genomics England and Participants

References

[1] Weaver, B. A. & Cleveland, D. W. The aneuploidy paradox in cell growth and tumorigenesis. *Cancer cell* **14** (6), 431–433 (2008) .

[2] Holland, A. J. & Cleveland, D. W. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. *Nature Reviews Molecular Cell Biology* **10** (7), 478–487 (2009). https://doi.org/10.1038/nrm2718 .

[3] Boveri, T. Concerning the Origin of Malignant Tumours by Theodor Boveri. Translated and annotated by Henry Harris. *Journal of Cell Science* **121** (Supplement_1), 1–84 (2008). https://doi.org/10.1242/jcs.025742 .

[4] Wright, N. A. Boveri at 100: cancer evolution, from preneoplasia to malignancy. *The Journal of Pathology* **234** (2), 146–151 (2014). https://doi.org/10.1002/path.4408 .

[5] Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396** (6712), 643–649 (1998) .

[6] Taylor, A. M. R. *et al.* Chromosome instability syndromes. *Nature reviews Disease primers* **5** (1), 64 (2019) .

[7] Dewhurst, S. M. *et al.* Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer discovery* **4** (2), 175–185 (2014) .

[8] Baker, T. M., Waise, S., Tarabichi, M. & Van Loo, P. Aneuploidy and complex genomic rearrangements in cancer evolution. *Nature Cancer* 1–12 (2024) .

[9] Watkins, T. B. *et al.* Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature* **587** (7832), 126–132 (2020) .

[10] Steele, C. D. *et al.* Signatures of copy number alterations in human cancer. *Nature* **606** (7916), 984–991 (2022) .

[11] Drews, R. M. *et al.* A pan-cancer compendium of chromosomal instability. *Nature* **606** (7916), 976–983 (2022) .

[12] Steele, C. D., Pillay, N. & Alexandrov, L. B. An overview of mutational and copy number signatures in human cancer. *The Journal of Pathology* **257** (4), 454–465 (2022) .

[13] Goh, J. Y. *et al.* Chromosome 1q21. 3 amplification is a trackable biomarker and actionable target for breast cancer recurrence. *Nature medicine* **23** (11), 1319–1330 (2017) .

[14] Qiu, X. *et al.* Myc drives aggressive prostate cancer by disrupting transcriptional pause release at androgen receptor targets. *Nature communications* **13** (1), 2559 (2022) .

[15] Hugdahl, E., Kalvenes, M. B., Mannelqvist, M., Ladstein, R. G. & Akslen, L. A. Prognostic impact and concordance of tert promoter mutation and protein expression in matched primary and metastatic cutaneous melanoma. *British journal of cancer* **118** (1), 98–105 (2018) .

[16] Leonetti, A. *et al.* Resistance mechanisms to osimertinib in egfr-mutated non-small cell lung cancer. *British journal of cancer* **121** (9), 725–737 (2019) .

[17] Pons, G. *et al.* Analysis of cancer genomic amplifications identifies druggable collateral dependencies within the amplicon. *Cancers* **15** (6), 1636 (2023) .

[18] Turnbull, C. *et al.* The 100 000 genomes project: bringing whole genome sequencing to the nhs. *Bmj* **361** (2018) .

[19] 100, . G. P. P. I. 100,000 genomes pilot on rare-disease diagnosis in health care—preliminary report. *New England Journal of Medicine* **385** (20), 1868–1880 (2021) .

[20] Murugaesu, N. *et al.* Insights for precision healthcare from the 100,000 genomes cancer programme (2022) .

[21] Thu, K. L. *et al.* Lung adenocarcinoma of never smokers and smokers harbor differential regions of genetic alteration and exhibit different levels of genomic instability. *PloS one* **7** (3), e33003 (2012) .

[22] Pikor, L., Thu, K., Vucic, E. & Lam, W. The detection and implication of genome instability in cancer. *Cancer and Metastasis Reviews* **32**, 341–352 (2013) .

[23] Mendiratta, G. *et al.* Cancer gene mutation frequencies for the us population. *Nature Communications* **12** (1), 5961 (2021) .

[24] Sundararajan, M., Taly, A. & Yan, Q. *Axiomatic attribution for deep networks*, 3319–3328 (PMLR, 2017).

[25] Jha, A., K. Aicher, J., R. Gazzara, M., Singh, D. & Barash, Y. Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study. *Genome biology* **21**, 1–22 (2020) .

[26] Kapishnikov, A. *et al.* *Guided integrated gradients: An adaptive path method for removing noise*, 5050–5058 (2021).

[27] Adélaïde, J. *et al.* Chromosome region 8p11-p21: refined mapping and molecular alterations in breast cancer. *Genes, Chromosomes and Cancer* **22** (3), 186–199 (1998) .

[28] Voutsadakis, I. A. 8p11. 23 amplification in breast cancer: molecular characteristics, prognosis and targeted therapy. *Journal of Clinical Medicine* **9** (10), 3079 (2020) .

[29] Natrajan, R. *et al.* Functional characterization of the 19q12 amplicon in grade iii breast cancers. *Breast Cancer Research* **14**, 1–15 (2012) .

[30] Xu, X., Zhang, M., Xu, F. & Jiang, S. Wnt signaling in breast cancer: biological mechanisms, challenges and opportunities. *Molecular cancer* **19**, 1–35 (2020) .

[31] Zeng, Y. A. & Nusse, R. Wnt proteins are self-renewal factors for mammary stem cells and promote their long-term expansion in culture. *Cell stem cell* **6** (6), 568–577 (2010) .

[32] Van Amerongen, R., Bowman, A. N. & Nusse, R. Developmental stage and time dictate the fate of wnt/β-catenin-responsive stem cells in the mammary gland. *Cell stem cell* **11** (3), 387–400 (2012) .

[33] Coussens, L. *et al.* Tyrosine kinase receptor with extensive homology to egf receptor shares chromosomal location with neu oncogene. *Science* **230** (4730), 1132–1139 (1985) .

[34] Latham, C. *et al.* Frequent co-amplification of two different regions on 17q in aneuploid breast carcinomas. *Cancer genetics and cytogenetics* **127** (1), 16–23 (2001) .

[35] McKay, J. D. *et al.* Lung cancer susceptibility locus at 5p15. 33. *Nature genetics* **40** (12), 1404–1406 (2008) .

[36] Zhu, C. *et al.* Amplification of telomerase (htert) gene is a poor prognostic marker in non-small-cell lung cancer. *British journal of cancer* **94** (10), 1452–1459 (2006) .

[37] Kang, J. U., Koo, S. H., Kwon, K. C., Park, J. W. & Kim, J. M. Gain at chromosomal region 5p15. 33, containing tert, is the most frequent genetic event in early stages of non-small cell lung cancer. *Cancer genetics and cytogenetics* **182** (1), 1–11 (2008) .

[38] Job, B. *et al.* Genomic aberrations in lung adenocarcinoma in never smokers. *PloS one* **5** (12), e15145 (2010) .

[39] Kang, J. U. *et al.* High frequency of genetic alterations in non-small cell lung cancer detected by multi-target fluorescence in situ hybridization. *Journal of Korean Medical Science* **22** (Suppl), S47–S51 (2007) .

[40] Achcar, R. D. O. D., Nikiforova, M. N., Dacic, S., Nicholson, A. G. & Yousem, S. A. Mammalian mastermind like 2 11q21 gene rearrangement in bronchopulmonary mucoepidermoid carcinoma. *Human pathology* **40** (6), 854–860 (2009) .

[41] Petersen, S. *et al.* Chromosomal alterations in the clonal evolution to the metastatic stage of squamous cell carcinomas of the lung. *British journal of cancer* **82** (1), 65–73 (2000) .

[42] Cao, Y. *et al.* Estimation of the survival of patients with lung squamous cell carcinoma using genomic copy number aberrations. *Clinical Lung Cancer* **17** (1), 68–74 (2016) .

[43] Martínez-Jiménez, F. *et al.* Genetic immune escape landscape in primary and metastatic cancer. *Nature Genetics* 1–12 (2023) .

[44] Orvis, T. *et al.* Brg1/smarca4 inactivation promotes non–small cell lung cancer aggressiveness by altering chromatin organization. *Cancer research* **74** (22), 6486–6498 (2014) .

[45] Takanami, I. Overexpression of ccr7 mrna in nonsmall cell lung cancer: correlation with lymph node metastasis. *International journal of cancer* **105** (2), 186–189 (2003) .

[46] Rakha, E. A. *et al.* Updated uk recommendations for her2 assessment in breast cancer. *Journal of clinical pathology* (2014) .

[47] Biserni, G. B., Engstrøm, M. J. & Bofin, A. M. Her2 gene copy number and breast cancer-specific survival. *Histopathology* **69** (5), 871–879 (2016) .

[48] Kallioniemi, O.-P. *et al.* Association of c-erbb-2 protein over-expression with high rate of cell proliferation, increased risk of visceral metastasis and poor long-term survival in breast cancer. *International journal of cancer* **49** (5), 650–655 (1991) .

[49] Karantza, V. Keratins in health and cancer: more than mere epithelial cell markers. *Oncogene* **30** (2), 127–138 (2011) .

[50] Saha, S. *et al.* Krt19 directly interacts with β-catenin/rac1 complex to regulate numb-dependent notch signaling pathway and breast cancer properties. *Oncogene* **36** (3), 332–349 (2017) .

[51] Wang, W. *et al.* Targeting the krt16-vimentin axis for metastasis in lung cancer. *Pharmacological Research* 106818 (2023) .

[52] Salari, K. *et al.* Cdx2 is an amplified lineage-survival oncogene in colorectal cancer. *Proceedings of the National Academy of Sciences* **109** (46), E3196–E3205 (2012) .

[53] Day, E. *et al.* Irs2 is a candidate driver oncogene on 13q34 in colorectal cancer. *International Journal of Experimental Pathology* **94** (3), 203–211 (2013) .

[54] Shen, M.-H. *et al.* Colorectal cancer concurrent gene signature based on coherent patterns between genomic and transcriptional alterations. *BMC cancer* **22** (1), 590 (2022) .

[55] Spruck, C. H., Won, K.-A. & Reed, S. I. Deregulated cyclin e induces chromosome instability. *Nature* **401** (6750), 297–300 (1999) .

[56] Lu, B., Winnall, S., Cross, W. & Barnes, C. P. Cell-cycle dependent dna repair and replication unifies patterns of chromosome instability. *bioRxiv* 2024–01 (2024) .

[57] Otero, L. *et al.* Variations in axin2 predict risk and prognosis of colorectal cancer. *BDJ open* **5** (1), 13 (2019) .

[58] Bossis, I. & Stratakis, C. A. Minireview: Prkar1a: normal and abnormal functions. *Endocrinology* **145** (12), 5452–5458 (2004) .

[59] Reinholz, M. M. *et al.* Breast cancer and aneusomy 17: implications for carcinogenesis and therapeutic response. *The lancet oncology* **10** (3), 267–277 (2009) .

[60] Rye, I. H. *et al.* Quantitative multigene fish on breast carcinomas identifies der (1; 16)(q10; p10) as an early event in luminal a tumors. *Genes, Chromosomes and Cancer* **54** (4), 235–248 (2015) .

[61] Girish, V. *et al.* Oncogene-like addiction to aneuploidy in human cancers. *Science* **381** (6660), eadg4521 (2023) .

[62] Schwendel, A. *et al.* Primary small-cell lung carcinomas and their metastases are characterized by a recurrent pattern of genetic alterations. *International journal of cancer* **74** (1), 86–93 (1997) .

[63] Romeo, M. S. *et al.* Chromosomal abnormalities in non-small cell lung carcinomas and in bronchial epithelia of high-risk smokers detected by multi-target interphase fluorescence in situ hybridization. *The Journal of molecular diagnostics* **5** (2), 103–112 (2003) .

[64] Sato, M. *et al.* Identification of chromosome arm 9p as the most frequent target of homozygous deletions in lung cancer. *Genes, Chromosomes and Cancer* **44** (4), 405–414 (2005) .

[65] Rodriguez-Nieto, S. & Sanchez-Cespedes, M. Brg1 and lkb1: tales of two tumor suppressor genes on chromosome 19p and lung cancer. *Carcinogenesis* **30** (4), 547–554 (2009) .

[66] Carvalho, B. *et al.* Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. *Gut* (2008) .

[67] Li, D. *et al.* Plagl2 and pofut1 are regulated by an evolutionarily conserved bidirectional promoter and are collaboratively involved in colorectal cancer by maintaining stemness. *EBioMedicine* **45**, 124–138 (2019) .

[68] Sillars-Hardebol, A. H. *et al.* Tpx2 and aurka promote 20q amplicon-driven colorectal adenoma to carcinoma progression. *Gut* gutjnl–2011 (2011) .

[69] Ali Hassan, N. Z. *et al.* Integrated analysis of copy number variation and genome-wide expression profiling in colorectal cancer tissues. *PloS one* **9** (4), e92553 (2014) .

[70] Du, Y. *et al.* Pofut1 promotes colorectal cancer development through the activation of notch1 signaling. *Cell death & disease* **9** (10), 995 (2018) .

[71] Su, C. *et al.* Studying the mechanism of plagl2 overexpression and its carcinogenic characteristics based on 3'-untranslated region in colorectal cancer. *International journal of oncology* **52** (5), 1479–1490 (2018) .

[72] Alazzouzi, H. *et al.* Smad4 as a prognostic marker in colorectal cancer. *Clinical Cancer Research* **11** (7), 2606–2611 (2005) .

[73] Grady, W. M. Genomic instability and colon cancer. *Cancer and metastasis reviews* **23**, 11–27 (2004) .

[74] Ogino, S., Kawasaki, T., Kirkner, G. J., Ohnishi, M. & Fuchs, C. S. 18q loss of heterozygosity in microsatellite stable colorectal cancer is correlated with cpg island methylator phenotype-negative (cimp-0) and inversely with cimp-low and cimp-high. *BMC cancer* **7** (1), 1–9 (2007) .

[75] Jen, J. *et al.* Allelic loss of chromosome 18q and prognosis in colorectal cancer. *New England Journal of Medicine* **331** (4), 213–221 (1994) .

[76] Hermsen, M. *et al.* Colorectal adenoma to carcinoma progression follows multiple pathways of chromosomal instability. *Gastroenterology* **123** (4), 1109–1119 (2002) .

[77] Thorstensen, L. *et al.* Evaluation of 1p losses in primary carcinomas, local recurrences and peripheral metastases from colorectal cancer patients. *Neoplasia* **2** (6), 514–522 (2000) .

[78] Jiao, W. *et al.* A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nature communications* **11** (1), 728 (2020) .

[79] Liang, Y. *et al.* A deep learning framework to predict tumor tissue-of-origin based on copy number alteration. *Frontiers in bioengineering and biotechnology* **8**, 701 (2020) .

[80] Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W. & Mostafavi, S. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics* **24** (2), 125–137 (2023) .

[81] Turnbull, C. Introducing whole-genome sequencing into routine cancer care: the genomics england 100 000 genomes project. *Annals of Oncology* **29** (4), 784–787 (2018) .

[82] Sosinsky, A. *et al.* Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 genomes cancer programme. *Nature Medicine* 1–11 (2024) .

[83] Roller, E., Ivakhno, S., Lee, S., Royce, T. & Tanner, S. Canvas: versatile and scalable detection of copy number variants. *Bioinformatics* **32** (15), 2375–2377 (2016) .

[84] Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578** (7793), 122–128 (2020) .

[85] in *Pytorch: An imperative style, high-performance deep learning library* 8024–8035 (Curran Associates, Inc., 2019).

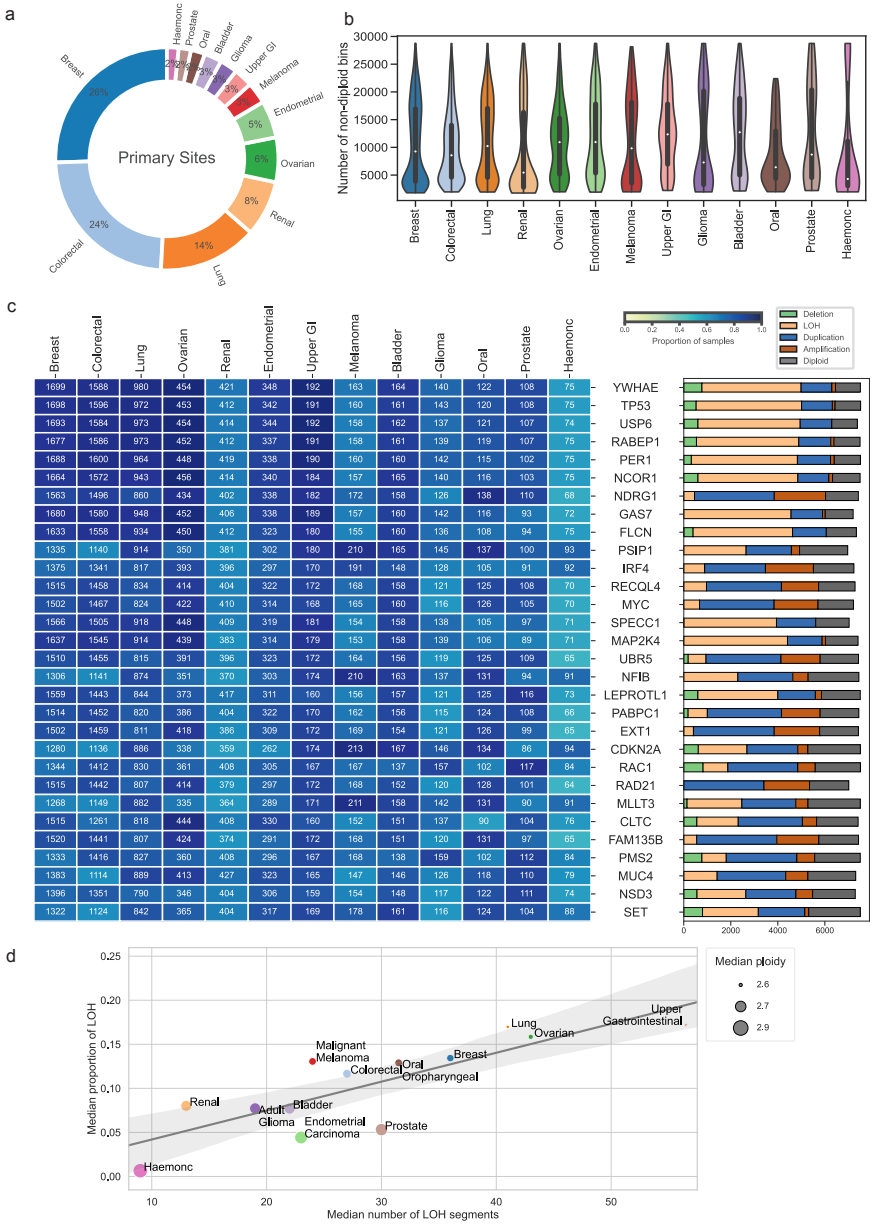
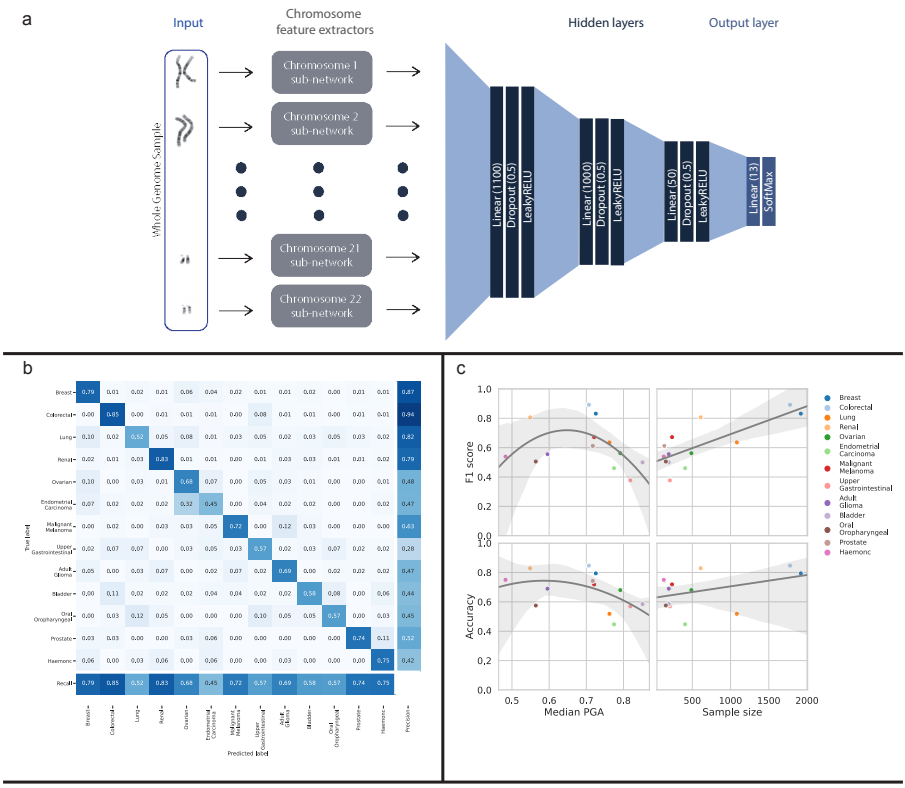

a
Primary Sites
Breast 26%
Colorectal 24%
Lung 14%
Renal 8%
Ovarian 6%
Endometrial 5%
Melanoma 4%
Upper GI 3%
Glioma 3%
Bladder 2%
Oral 2%
Prostate 2%
Haemonc 2%
b
Number of non-diploid bins
30000
25000
20000
15000
10000
5000
Breast
Colorectal
Lung
Renal
Ovarian
Endometrial
Melanoma
Upper GI
Glioma
Bladder
Oral
Prostate
Haemonc
c
Breast Colorectal Lung Ovarian Renal Endometrial Upper GI Melanoma Bladder Glioma Oral Prostate Haemonc
Proportion of samples
Deletion
LOH
Duplication
Amplification
Diploid
YWHAE 1699 1588 980 454 421 348 192 163 164 140 122 108 75
TP53 1698 1596 972 453 412 342 191 160 161 143 120 108 75
USP6 1693 1584 973 454 414 344 192 158 162 137 121 107 74
RABEP1 1677 1586 973 452 412 337 191 158 161 139 119 107 75
PER1 1688 1600 964 448 419 338 190 160 160 142 115 102 75
NCOR1 1664 1572 943 456 414 340 184 157 165 140 116 103 75
NDRG1 1563 1496 860 434 402 338 182 172 158 126 138 110 68
GAS7 1680 1580 948 452 406 338 189 157 160 142 116 93 72
FLCN 1633 1558 934 450 412 323 180 155 160 136 108 94 75
PSIP1 1335 1140 914 350 381 302 180 210 165 145 137 100 93
IRF4 1375 1341 817 393 396 297 170 191 148 128 105 91 92
RECQL4 1515 1458 834 414 404 322 172 168 158 121 125 108 70
MYC 1502 1467 824 422 410 314 168 165 160 116 126 105 70
SPECC1 1566 1505 918 448 409 319 181 154 158 138 105 97 71
MAP2K4 1637 1545 914 439 383 314 179 153 158 139 106 89 71
UBR5 1510 1455 815 391 396 323 172 164 156 119 125 109 65
NFIB 1306 1141 874 351 370 303 174 210 163 137 131 94 91
LEPROTL1 1559 1443 844 373 417 311 160 156 157 121 125 116 73
PABPC1 1514 1452 820 386 404 322 170 162 156 115 124 108 66
EXT1 1502 1459 811 418 386 309 172 169 154 121 126 99 65
CDKN2A 1280 1136 886 336 359 262 174 213 167 146 134 86 94
RAC1 1344 1412 830 361 408 305 167 167 137 157 102 117 84
RAD21 1515 1442 807 414 379 297 172 168 152 120 128 101 64
MLLT3 1268 1149 882 335 364 289 171 211 158 142 131 90 91
CLTC 1515 1261 818 444 408 330 160 152 151 137 90 104 76
FAM135B 1520 1441 807 424 374 291 172 168 151 120 131 97 65
PMS2 1333 1416 827 360 408 296 167 168 138 159 102 112 84
MUC4 1383 1114 889 413 427 323 165 147 146 126 118 110 79
NSD3 1396 1351 790 346 404 306 159 154 148 117 122 111 74
SET 1322 1124 842 365 404 317 169 178 161 116 124 104 88
0
2000
4000
6000
d
Median proportion of LOH
Median number of LOH segments
Median ploidy
2.6
2.7
2.9
Lung
Ovarian
Upper Gastrointestinal
Malignant Melanoma
Breast
Colorectal
Oral
Oropharyngeal
Renal
Adult Glioma
Bladder
Endometrial Carcinoma
Prostate
Haemonc

Fig. 1 ***(preceding page)*** **Genomic landscape analysis across cancer types in the GEL dataset** (a) The proportion of samples in the GEL dataset categorised by their respective cancer primary sites. (b) A violin plot of the distributions of non-diploid bins across the genome according to each cancer primary site. (c) A heatmap representing the frequency of non-diploid statuses at various Rosetta gene loci across multiple cancer cohorts. Each column corresponds to a different cancer primary site, with the colour intensity indicating the proportion of samples exhibiting non-diploid status at each locus. Adjacent to the heatmap, stacked bar charts for each gene locus quantify the copy number variations. The number of samples per cohort is annotated within the heatmap cells, providing a direct correlation between sample size and observed genetic alterations. (d) The correlation between the median proportion of the genome affected by Loss of Heterozygosity (LOH) and the median frequency of LOH stretches in the samples for each cancer primary site. The LOH segments are defined as a continuous segment of the genome where heterozygosity is lost, indicating a uniform genetic state in that region.


a
Input
Chromosome feature extractors
Hidden layers
Output layer
Whole Genome Sample
Chromosome 1 sub-network
Chromosome 2 sub-network
Chromosome 21 sub-network
Chromosome 22 sub-network
Linear (1100)
Dropout (0.5)
LeakyRELU
Linear (1000)
Dropout (0.5)
LeakyRELU
Linear (50)
Dropout (0.5)
LeakyRELU
Linear (13)
SoftMax
b
True label
Predicted label
Recall
Precision
c
F1 score
Accuracy
Median PGA
Sample size
Breast
Colorectal
Lung
Renal
Ovarian
Endometrial Carcinoma
Malignant Melanoma
Upper Gastrointestinal
Adult Glioma
Bladder
Oral Oropharyngeal
Prostate
Haemonc

d

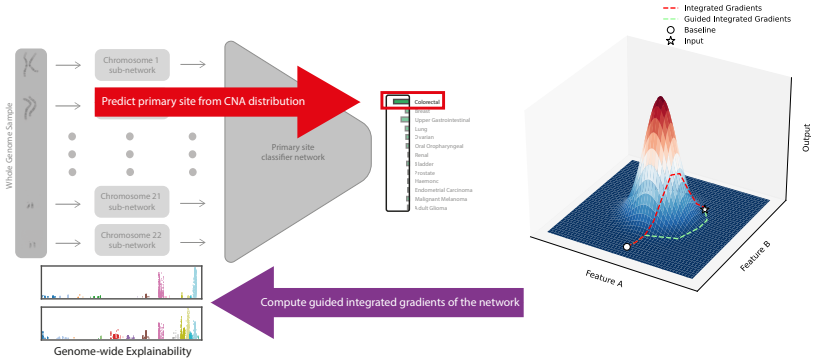
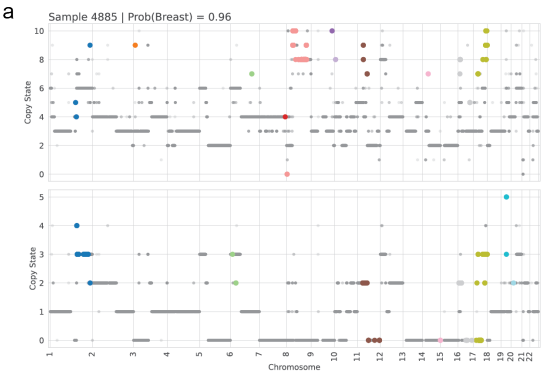
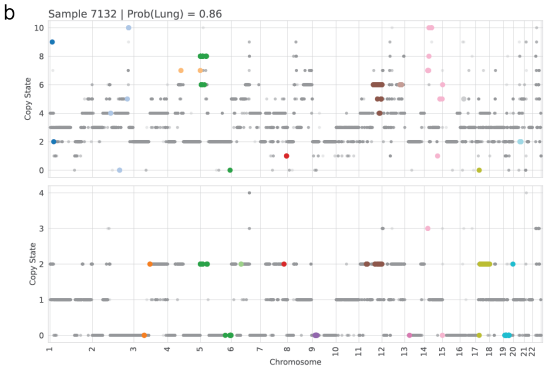

Whole Genome Sample
Chromosome 1 sub-network
Chromosome 21 sub-network
Chromosome 22 sub-network
Predict primary site from CNA distribution
Primary site classifier network
Compute guided integrated gradients of the network
Genome-wide Explainability
Integrated Gradients
Guided Integrated Gradients
Baseline
Input
Output
Feature A
Feature B

Fig. 2 ***(preceding page)*** **Analysis of cancer classification using neural network and Guided Integrated Gradients** (a) Schematic of the neural network architecture, illustrating the input from whole genome sequencing (WGS) samples. Each chromosome's binned data feeds into a sub-network, culminating in a deep feed-forward network with hierarchical layers. (b) Confusion matrix for each cancer primary site, accompanied by precision and recall metrics, providing a detailed assessment of the model's classification performance. (c) Regression analyses: F1 score and classification accuracy against Median Proportion of Genome Altered (PGA) with a second-order polynomial regression line, and against sample size with a linear regression line, both with 95% confidence interval shading. (d) Cartoon depiction of the Guided Integrated Gradients (GIG) method, tracing the process from sample classification prediction to genome-wide explainability. A comparison is drawn between Integrated Gradients (IG) and GIG, highlighting the path selection mechanism in GIG for enhanced gradient accumulation analysis.



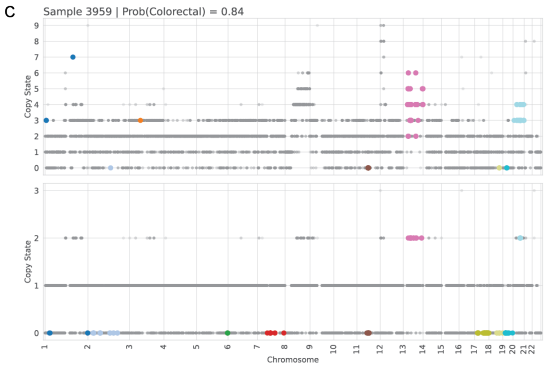
Copy state	Cytoband	Gene(s)	Enhancer(s)
Amplification	8p11.22		
Amplification	16p13.12	ERCC4	MKL2
Amplification	8p11.23		ERLIN2 PLPBP ADGRA2
Amplification	8p11.22		
Amplification	16p13.12	ERCC4	ERCC4
Amplification	16p13.12		
Amplification	16p13.12		
Amplification	16p13.12		
Amplification	16p13.12		
Amplification	8p11.22		

Copy state	Cytoband	Gene(s)	Enhancer(s)
Amplification	19q12		
Amplification	19q12		
Amplification	19q12		
Amplification	17q24.1	AXIN2	
Amplification	17q11.2		KIAA0100 SUPT6H ENSG00000265205
Duplication	17q21.31		KANSL1-AS1
Duplication	17q21.2		
Amplification	17q11.2		PROCA1 RAB34 RPL23A ENSG00000264577
Duplication	17q21.2	CCR7	NEK8 TRAF4 TNS4 CCR7
Amplification	17q25.3		CYTH1



Copy state	Cytoband	Gene(s)	Enhancer(s)
Amplification	5p15.33		ZDHHC11 BRD9
Amplification	14q21.1		LOC105370469 LRFN5
Amplification	5p15.2		DNAH5
Amplification	5p15.33	TERT	SLC6A19 SLC6A18 TERT
Amplification	11q21		
Amplification	14q21.1		
Amplification	14q12		
Amplification	5p15.33		ENSG248949 LOC156688
Amplification	11q21		
Amplification	11q23.3	ARHGEF12	GRIK4

Copy state	Cytoband	Gene(s)	Enhancer(s)
LOH	19p13.13	CALR LYL1	
Duplication	17q21.2	CCR7	TNS4 CCR7
Duplication	17q11.2		TMEM98
Duplication	17q11.2		
Duplication	17q11.2		MYO1D
Duplication	17q11.2		
Duplication	17q11.2		
Duplication	17q11.2		NLK
Duplication	17q11.2		NOS2 LYRM9 KRT13 KRT15 KRT19 LINC00974 KRT9 KRT14
Duplication	17q21.2		



Copy state	Cytoband	Gene(s)	Enhancer(s)
Amplification	13q12.2		GSX1
Amplification	13q12.2		
Amplification	13q34		LINC00552 TMEM255B
Amplification	13q12.2		LNX2 POLR1D
Amplification	13q12.2		
Amplification	13q34		GAS6-AS1 GAS6
Duplication	13q12.2	CDX2	PDX1 CDX2 URAD
Duplication	13q12.3		
Amplification	13q34		GRK1
Duplication	13q12.3		POMP

Copy state	Cytoband	Gene(s)	Enhancer(s)
LOH	19q12		C19orf12
LOH	17q24.1		
LOH	19q12	CCNE1	CCNE1
LOH	17q24.2		
LOH	19q12	CCNE1	URI1
LOH	19q12		POP4 PLEKHF1
LOH	17q24.2		
Deletion	19p13.12		
LOH	17q24.1		
LOH	17q24.2		CACNG4

Fig. 3 AI explainability in single whole genome sequencing samples across different cancer types. (a) One breast cancer sample, (b) one lung cancer sample, and (c) one colorectal cancer sample, each showcasing CNAs across the genome. The top part of each panel displays the total copy number for both alleles, while the bottom shows the minor allele copy number. High-importance genomic loci, identified by the model, are colour-coded by chromosome, with unselected loci in grey. Accompanying tables detail the copy state (Diploid, Duplication, Amplification, Deletion, LOH), cytoband, presence of oncogenes, and enhancers for the high attribute loci in both total and minor allele signals. Table rows are colour-coded based on the magnitude of normalised attribution values from the GIG analysis, highlighting key loci contributing to the AI's classification decision.

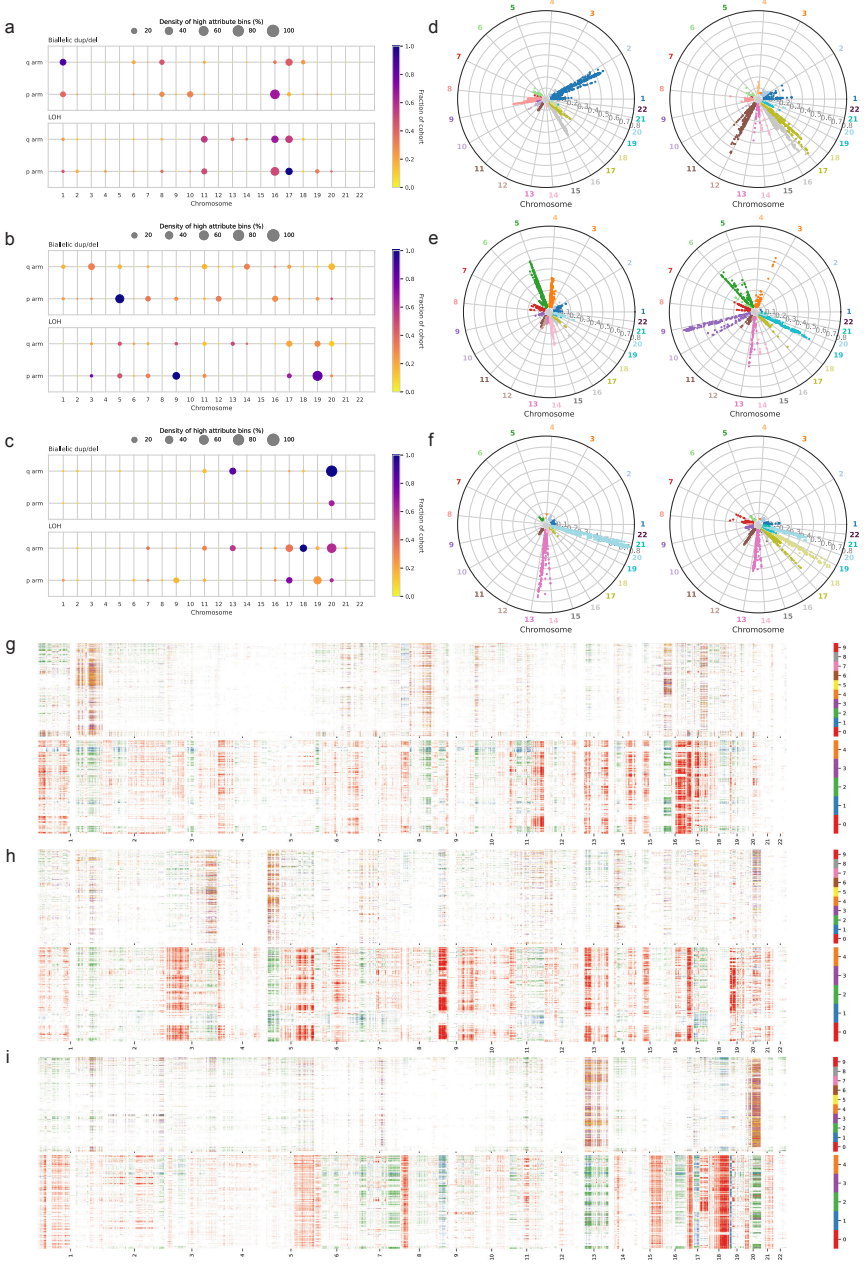

a
b
c
d
e
f
g
h
i
Density of high attribute bins (%)
Biallelic dup/del
LOH
q arm
p arm
Chromosome
Fraction of cohort

Fig. 4 ***(preceding page)*** **Integrated analysis of genomic alterations and AI-driven attribution in cancer cohorts.** (a-c) Bubble plots for breast, lung, and colorectal cancer cohorts for both total and minor allele copy numbers, displaying high-attribute loci across chromosome arms. Bubble size represents the density of significant bins, with colour indicating the fraction of the cohort sharing these loci. (d-f) Circular Manhattan Plot for breast, lung, and colorectal cancer cohorts for total allele copy number (left) and minor allele copy number (right). This visualisation displays the same attribution signals as the bubble plots but extends beyond chromosome arm clustering, enabling the identification of distinct peaks of high attribution common to many samples in the cohort. The height of each point on the plot corresponds to the fraction of the cohort sharing the signal, coloured by chromosome identity. (g-i) Heatmaps of Single Sample CNA for total allele copy number (top) and minor allele copy number (bottom), this heatmap is divided into three sub-panels representing breast, lung, and colorectal cancer cohorts. It highlights only high attribute loci as determined by the model, coloured according to CNA values. Loci not selected by the model are omitted, appearing as white spaces, to focus on regions of highest model attribution.

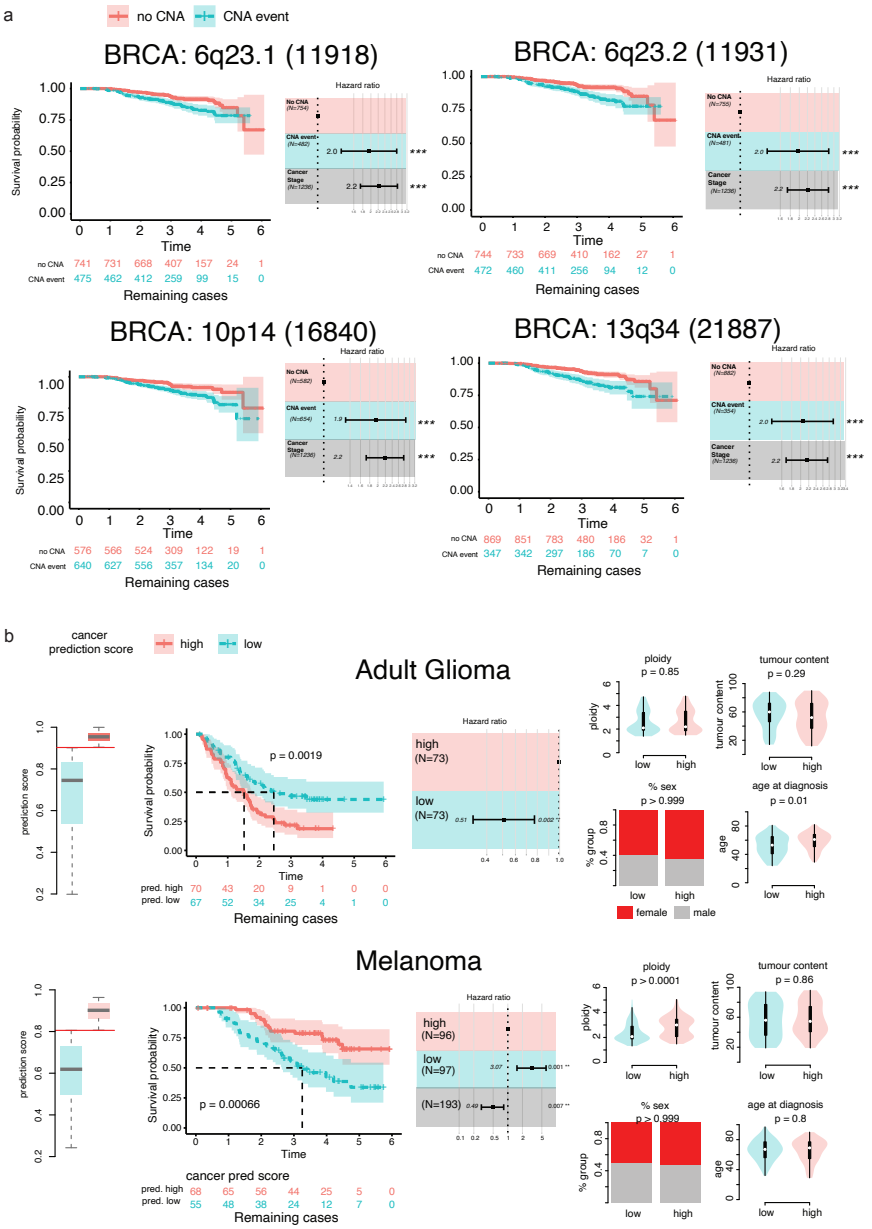


Fig. 5 Analysis of patient outcome a) 42 loci with significant scores effect patient outcome following stage correction. Four loci remained significant following multi-testing correction (chromosome location and bin ID in brackets labelled). Kapalan Meier plots and tables with patient numbers shown on left, with forest plots stage corrected on the right. b) Splitting of patient groups was performed using the median cancer scores (left most plots). Kaplan Meier plots show significant impact on outcome in adult glioma and melanoma (note that glioma does not use a staging system thus absent from forest plots). Right most plots show other genetic features of the patient cohorts, including ploidy, tumour content of the biopsy, proportion of male or female patients, and age at diagnosis.