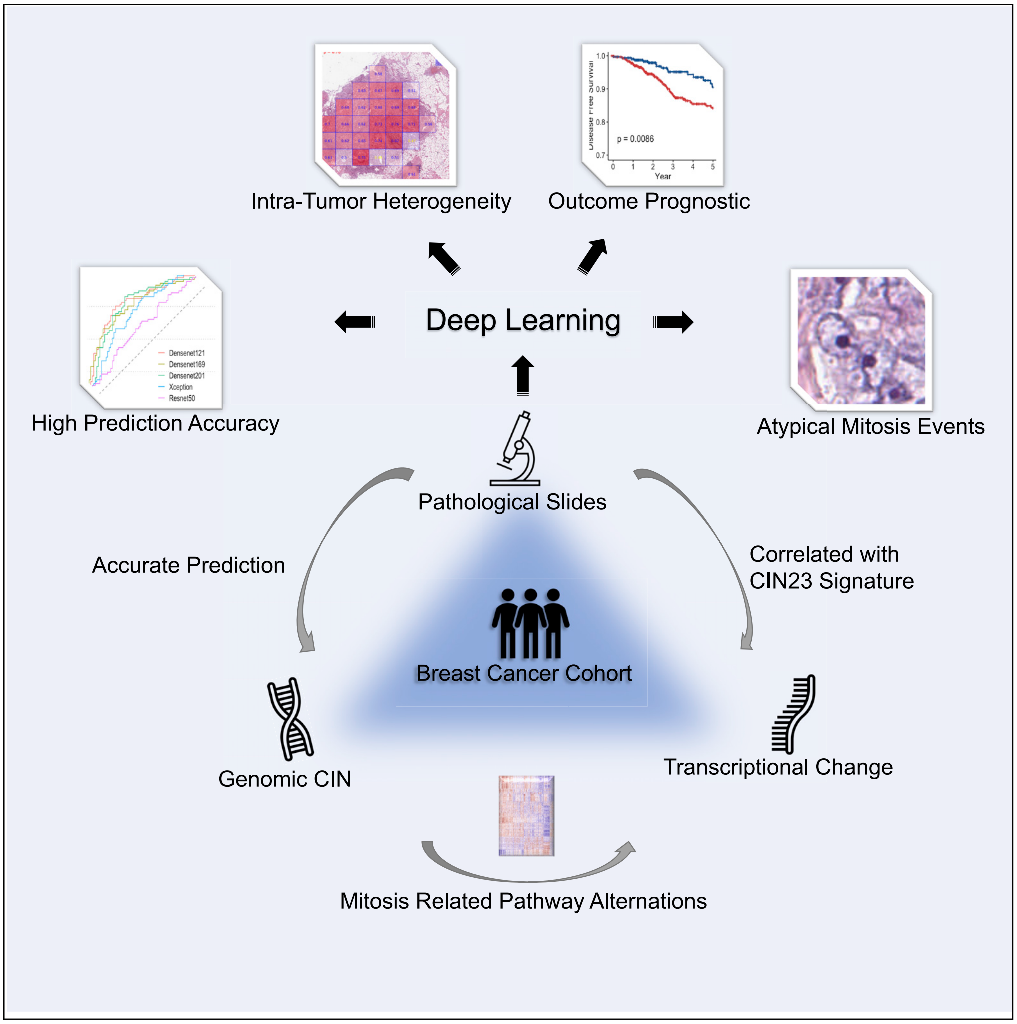
iScience

CellPress
OPEN ACCESS

Article

Deep learning predicts chromosomal instability from histopathology images



Zhuoran Xu, Akanksha Verma, Uska Naveed, Samuel F. Bakhoum, Pegah Khosravi, Olivier Elemento

ole2001@med.cornell.edu

Highlights

Deep learning model accurately predicts CIN from histopathology slides

There is evidence for CIN intra-tumor heterogeneity with prognostic value

CIN is associated with profound transcriptional changes including mitotic pathways

Results pave the way for using CIN as a prognosis biomarker

Xu et al., iScience *24*, 102394
May 21, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.isci.2021.102394

iScience

CellPress
OPEN ACCESS

Article

Deep learning predicts chromosomal instability from histopathology images

Zhuoran Xu,[1,2] Akanksha Verma,[1] Uska Naveed,[1] Samuel F. Bakhoum,[3,4] Pegah Khosravi,[1,5] and Olivier Elemento[1,6,*]

SUMMARY

Chromosomal instability (CIN) is a hallmark of human cancer yet not readily testable for patients with cancer in routine clinical setting. In this study, we sought to explore whether CIN status can be predicted using ubiquitously available hematoxylin and eosin histology through a deep learning-based model. When applied to a cohort of 1,010 patients with breast cancer (Training set: n = 858, Test set: n = 152) from The Cancer Genome Atlas where 485 patients have high CIN status, our model accurately classified CIN status, achieving an area under the curve of 0.822 with 81.2% sensitivity and 68.7% specificity in the test set. Patch-level predictions of CIN status suggested intra-tumor heterogeneity within slides. Moreover, presence of patches with high predicted CIN score within an entire slide was more predictive of clinical outcome than the average CIN score of the slide, thus underscoring the clinical importance of intra-tumor heterogeneity.

INTRODUCTION

Chromosomal instability (CIN) refers to ongoing chromosome segregation errors throughout consecutive cell divisions that can potentially result in extensive numerical and structural chromosomal aberrations (Lengauer et al. 1998). CIN, as one of the hallmarks of human cancer, has been recognized as a central driver of cancer evolution owning to its multipronged effects that facilitate processes such as metastasis, immune evasion, and therapeutic resistance (Bakhoum and Cantley 2018; Bakhom 2018; Murayama-Hosokawa, et al., 2010; Swanton, et al., 2009; Walther et al. 2008; Lee, et al., 2011; Hieronymus, et al., 2018, https://doi.org/10.7554/eLife.37294). For example, Smid et al. (2011) found that elevated CIN and consequent high aneuploidy burden is associated with poor breast cancer prognosis, measured as time to distant metastasis. Carter et al. (2006) revealed a correlation between a transcriptional signature of CIN with metastasis, tumor grading, and clinical outcome in multiple human cancers. Paradoxically, some studies suggest that excessive levels of CIN negatively impact tumor fitness and associate with better survival outcome, possibly because too much chromosomal segregation errors can impart a number of cellular burdens that produce proinflammatory signals and lead to programmed cell apoptosis (Birkbak, et al., 2011; Jamal-Hanjani, et al., 2015; Tijhuis et al. 2019; Zasadil, et al., 2014). Given the widespread nature and far-reaching consequences of CIN in human cancer, strategies for targeting CIN as a therapeutic vulnerability in some cancers are being actively researched (Zasadil, et al., 2014) (Pierssens, et al., 2017). Despite its clear importance, CIN status is not readily testable for patients with cancer in routine clinical settings because it requires complicated experimental assessment involving live microscopy, sensitive detection of micronuclei (a consequence of CIN) via immunohistochemistry, or comprehensive genomic analysis. On the other hand, gold-standard histopathological examinations that are used for cancer diagnosis and grading are ubiquitously available. Here we sought to investigate the feasibility of using histopathology whole-slide images (WSIs) to predict CIN status.

Deep learning is a state-of-the-art methodology for analyzing and interpreting cancer histology images. In recent years, a large number of studies attempted to employ a deep learning approach for a variety of tasks in computational pathology field by taking advantage of deep learning's ability to extract hierarchical features from images in a direct and automatic fashion. Previous research has shown that presence of driver mutations, mutational signatures, and expression-defined tumor subtypes can be predicted from hematoxylin and eosin (H&E) slides (Coudray, et al., 2018; Schaumberg et al. 2016; Xu, et al., 2019; Chen, et al., 2020; Liao, et al., 2020). For example, Kather et al. (2019) trained a Convolutional Neural Network (CNN) model that can robustly predict genome microsatellite instability in gastrointestinal cancer from

[1] Caryl and Israel Englander Institute for Precision Medicine, Weill Cornell Medicine, New York 10065, USA

[2] Pathology and Laboratory Medicine, Weill Cornell Medicine, New York 10065, USA

[3] Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York 10021, USA

[4] Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York 10021, USA

[5] Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York 10021, USA

[6] Lead contact

*Correspondence: ole2001@med.cornell.edu

https://doi.org/10.1016/j.isci.2021.102394

H&E histology, obtaining a patient-level area under the curve (AUC) of 0.84. Coudray et al. (2018) trained a deep learning network that successfully predicted six of ten most commonly mutated genes from lung adenocarcinoma pathology images, with AUCs from 0.733 to 0.856. Kather et al. (2020) then demonstrated the ability of deep learning to predict point mutations, molecular tumor subtypes, and immune-related gene expression signatures directly from H&E images in multiple cancer types. Fu et al. (Fu, Jung and TORNE 2020) used transfer learning and correlated histopathological pattern features with genomic, transcriptomic, and survival data in 28 cancer types. Javad et al. systematically used CNNs on 23 cancer types for tasks including tumor versus normal and cancer subtype classifications as well as predicting the presence of TP53 mutations (Noorbakhsh, et al., 2020).

In this study, we proposed using pathology images to predict patients' CIN status in breast cancer. We created a framework that uses transfer learning and feature aggregation to accurately discriminate high-CIN and low-CIN histopathology slides without human intervention. This framework is not limited to breast cancer and can be potentially extended to other cancer types. Our results indicate that (1) CIN can be predicted accurately from histopathology slides and (2) unexpectedly there appears to be substantial intra-tumor heterogeneity CIN status in many patients. These results pave the way for using CIN as a biomarker of prognosis and response to anti-CIN therapies fully integrated into existing clinical pathology workflows.

RESULTS

A weakly supervised deep learning model for patient genomic CIN classification in breast cancer

We obtained H&E slides and genomic profiles from patients with breast cancer in TCGA. Here we quantify CIN using the fraction genome altered (FGA, see transparent methods), which is one of the most commonly used quantitative measurements of CIN (Sipos, et al., 2021; Burrell, et al., 2013). FGA quantifies the burden of aneuploidies detectable in bulk genomic profiles. Although FGA is not a perfect measure of CIN and is incapable of capturing the rate of chromosomal changes by only providing a snapshot of the chromosomal alteration state, we and others have observed a strong correlation between FGA and CIN measured using microscopy and/or micronuclei staining (Schonhoft, et al., 2020; Pikor, et al., 2013). We refer to FGA as genomic CIN score, to contrast it with pathology predicted CIN scores introduced in this study. Genomic CIN score higher than 0.3 was labeled as high CIN; genomic CIN score lower than 0.3 was labeled as low CIN (Figure S1). H&E slides were processed as described in transparent methods. Based on our CIN classification, the breast cancer (BRCA) cohort had 485 high-CIN patients with 515 WSIs and 23,427 patches, and 525 low-CIN patients with 550 WSIs and 23,568 patches (Table S1). This study presents a deep learning model that can automatically predict patients' genomic CIN status on the molecular level from H&E-stained histopathology slides (Figures 1A–1E, see transparent methods). Our model uses CNN models pre-trained on ImageNet as patch-level feature extractor (Figure 1B) and then aggregates patch features into patient features (Figure 1C). This approach not only effectively addresses intra-tumor heterogeneity using "weak" patient-level labels but also offers opportunity to explore spatial CIN heterogeneity in individual patients (Figures 1D and 1E). The 1,070 WSIs of 1,010 patients from TCGA-BRCA were randomly split into training, validation, and test set. We then evaluated model performance in the hold-out test set, which included 152 patients.

Deep learning model predicts CIN with high accuracy and sensitivity

Several commonly used CNN architectures were tested in the transfer learning step used to extract the most relevant features that can predict genomic CIN. The best feature extraction method was selected based on the ability of trained fully connected layers to predict CIN groups in validation dataset (Figure S3). Results shown in Figure 2 indicate that Densenet-121 achieved the best performance with an AUC of 0.822 and accuracy (ACC) of 74.3%. Densenet networks with different depths achieved similar performance with AUCs of 0.806 and 0.807 for Densenet-169 and Densenet-201, respectively. The Densenet-121 model got a good sensitivity of 81.16%. The Xception model achieved an AUC of 0.752 and Resnet-50 achieved 0.650. Since the models were trained to classify genomic CIN status based on a hard threshold of genomic CIN score measured by FGA of being 0.3, we thus evaluated the model performance for patients with very high CIN (FGA>0.4) versus very low CIN (FGA<0.2) as well as patients who have moderate genomic CIN scores (FGA: 0.2–0.3 versus FGA: 0.3–0.4). Results shown in Table S2 indicate that, although our model performs best for predicting very low and very high CIN (ACC, 77.6%; AUC, 0.83), it retains good performances in the intermediate CIN range (ACC, 66.7%; AUC, 0.77). Altogether these results indicate that a deep learning

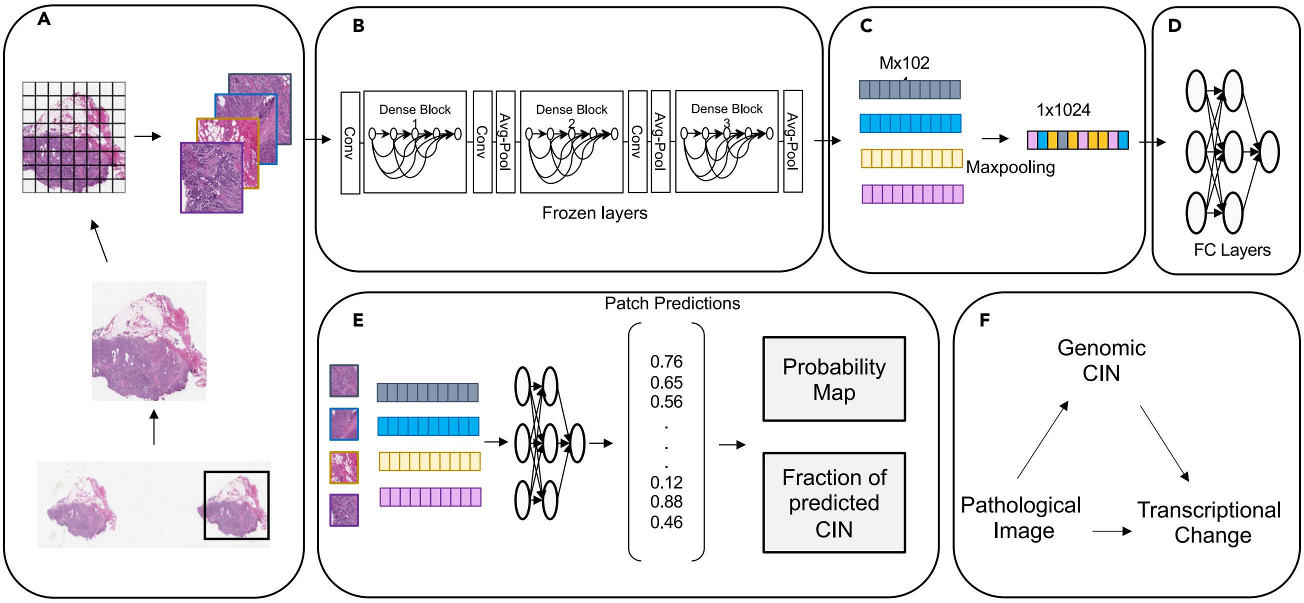


Figure 1. Overview of the pipeline

(A) During mage preprocessing, an overall region of interest (ROI) was identified in WSI as window with the highest tissue percentage using a sliding window approach. Then the ROI was tiled into non-overlapping patches before quality control. Only qualified patches were kept as described in transparent methods.

(B) Patch-level feature extraction was performed using a pre-trained CNN architecture.

(C) Max-pooling layer was used for aggregating patch-level feature embeddings to patient-level feature embedding.

(D) Fully connected layers were trained in supervised approach based on patient-level genomic CIN status.

(E) Patch-level feature embeddings were fed into trained fully connected layers from (D). Probability maps based on patch predictions were generated, and fractions of predicted CIN that measure slide intra-tumor heterogeneities were calculated.

(F) Pathological images were used to predict genomic CIN. Since genomic CIN can potentially alter gene expression and pathways on the transcriptional level, differentially expressed gene analysis and pathway analysis were performed. Pathology-predicted CIN and Genomic CIN were compared with CIN transcriptional signatures.

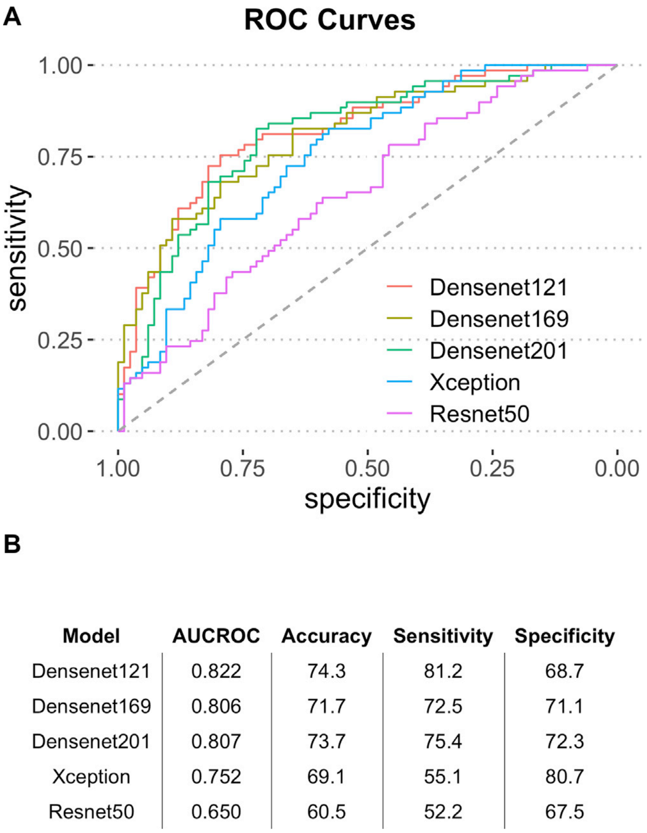
model can accurately classify CIN status, achieving an AUC of 0.822 with 81.2% sensitivity and 68.7% specificity in an independent test set not used for training or parameter exploration.

High-CIN patients exhibit more atypical mitosis events

To independently validate that pathology-predicted CIN status (and genomic CIN) correlates with CIN-related aberrant mitotic events, we inspected 10 tumor slides at 40× magnification level, looking manually for aberrant mitotic events. Half of the 10 slides were predicted as high CIN and half as low CIN. All 10 slides were also concordantly labeled as high CIN or low CIN by genomic CIN (in other words, they were true-positive and true-negative predictions). The expert who conducted the inspections was not informed of the sample labels to reduce subtle bias. As shown in Figures 3A–3H, both normal and abnormal mitosis events, including anaphase bridge, spindles with misalignment chromosomes, and multipolar and monopolar chromosome arrangements, were observed. By fitting the generalized estimating equation (GEE) Poisson regression model, we found that atypical mitosis event counts per field of view (patches with fixed size of 1,024*1,024 pixels on 40×) are significantly higher in predicted high-CIN patients compared with predicted low-CIN patients (GEE model p value<0.0001, Figure 3I).

Patch predictions demonstrate intra-tumor heterogeneity

As discussed in the previous section, we had hypothesized that not all patches within the same slide may have the same level of CIN; in other words, we hypothesized that there may be intra-tumor heterogeneity in CIN within the same tissue section. To test this hypothesis, we visualized patch predictions within one WSI. Patch predictions were generated by feeding individual patch features into trained fully connected layers independently. As shown in Figure 4, both high-CIN and low-CIN patches can be found within one WSI regardless of the slide's CIN status. Figure 4A is an example of a low-CIN slide with predicted high-CIN



Model	AUCROC	Accuracy	Sensitivity	Specificity
Densenet121	0.822	74.3	81.2	68.7
Densenet169	0.806	71.7	72.5	71.1
Densenet201	0.807	73.7	75.4	72.3
Xception	0.752	69.1	55.1	80.7
Resnet50	0.650	60.5	52.2	67.5

Figure 2. Model performance final evaluation

Model performance was evaluated in test set.

(A) Receiver operating characteristic (ROC) curves for different CNN architectures.

(B) Table for model performance with AUC, accuracy, sensitivity, and specificity.

probability of 0.27. Of all 45 patches, 11 (= 24.44%) were predicted as high-CIN patches. Because high-CIN patches still exist in low-CIN slides owing to intra-tumor heterogeneity, the prevalence scale and patch probability also have influence of the whole-slide CIN status. Similar results were shown in Figure 4B, where low-CIN patches were also found in high-CIN slide. We thus define predicted CIN-high fraction score as the percentage of predicted high-CIN patches based on each pathology image. Altogether we found that only 94 (9.31%) of 1,010 patients in the whole cohort exhibited a low level of intra-tumor heterogeneity with predicted CIN-high fraction score smaller than 10%. The median predicted CIN-high fraction score of this cohort is 57% with 25th and 75th percentile of 32% and 83%, respectively (Figure S2). A growing number of investigations have suggested the positive correlation between CIN status and increased neoplastic nuclear size, which can potentially be used as an excellent surrogate marker for CIN detection (Thompson and McManus 2015; Penner-Goeke, et al., 2017; Baergen, et al., 2019). Increases in DNA content (i.e., ploidy) have been found to be one of the mechanisms behind the corresponding nuclear enlargement (Zeimet, et al., 2011; Petersen, et al., 2009). To further validate the existence of intra-tumor heterogeneity revealed by patch predictions, we conducted nuclear instance segmentation and classification on all patches from test dataset using pretrained HoVer-Net model. Segmentation results showed that predicted CIN-High patches have larger neoplastic cell nucleus comparing with predicted CIN-Low patches (p value<0.0001, Figure S5A). At the same time, slides with a higher fraction of predicted CIN-High patches showed higher tumor ploidy values (p value: 0.0031, Figure S5B).

Fraction of predicted CIN high patches is correlated with transcriptional CIN score

CIN, as a hallmark of cancer, has been linked to activation of key downstream biological pathways such as cGAS-STING and non-canonical nuclear factor κB (NF-κB) (Bakhoum and Cantley 2018). To bridge the gap between molecular genome alterations with pathological features (Figure 1F), we conducted correlation

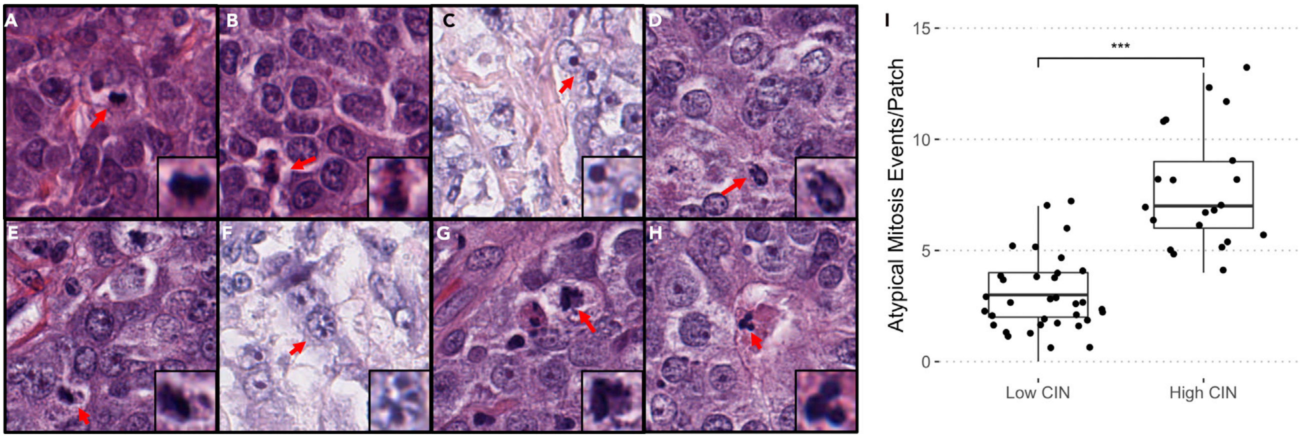


Figure 3. Patches containing normal and abnormal mitosis events at 40× magnification

(A and B) Normal mitosis. (C)-(H) Abnormal mitosis.

(C) Anaphase bridge.

(D) Monopolar mitosis.

(E–G) Mitotic figure with unaligned chromosomes.

(H) Multipolar mitosis.

(I) Boxplot of atypical mitosis events per field of view between five predicted high-CIN and five predicted low-CIN patients (GEE model p value<0.0001 ***).

analysis between a CIN-driven transcriptional gene signature with both genomic CIN score and predicted CIN-high fraction score. CIN23 is a gene signature derived from the human metastatic cell line models (MDA-MB-231) engineered to over-express MCAK (to suppress CIN) or a dominant negative version of MCAK (to increase CIN) (Bakhom, 2018; Cailleau et al., 1978). Employing CIN23, we derived a gene signature score for each patient as transcriptional CIN score using ssGSEA. Genomic CIN score was positively correlated with transcriptional CIN score with correlation coefficient of 0.14 (Table S5, p value<0.0001). We then reasoned that the average predicted CIN intra-tumor heterogeneity measured by the percentage of predicted high-CIN patches of each slide may correlate with the transcriptional CIN score, which is also an average representation of CIN across all spatial areas. Indeed, we observed a weak but significant positive correlation with transcriptional CIN (Table S5, rho: 0.1, p value = 0.0026).

Model performance is not breast cancer subtype specific

In the past, different patterns of CIN were observed to be associated with distinct subtypes of breast cancer (Kwei, et al., 2010). In this cohort, we found significant positive association between genomic CIN status with the prevalence of ER- (p value<0.0001), PR- (p value<0.0001), and Triple Negative (p value<0.0001) subtype status but not HER2 (p value = 0.1) status (Table S3). We therefore sought to verify that our algorithm is not simply predicting tumor subtypes (since we have that predicting breast cancer subtypes is feasible from H&E slides [Khosravi, et al., 2018]). We combined subtype information (ER, PR, HER2 status) along with clinical features including age, race, menopause status, and number of positive lymph nodes as clinical input. Then we retrained the fully connected layers with three different settings of input that are clinical input alone, image features alone, and concatenating clinical input with image, respectively. Results showed adding image features to clinical features can significantly help improve model performance (p value = 0.047). But when using image for predictions, adding extra clinical features along with subtype information did not improve the model performance (p value = 0.82) (Figure S4). Finally, no evidence suggested our model to be subtype specific with AUCs statistically the same across different subtypes (p values: ER, 0.33; PR, 0.94; HER2, 0.41; Triple Negative, 0.86) in this TCGA cohort (Table S4). We concluded that our model is predictive of CIN independently of tumor subtypes.

CIN is associated with poor prognosis in breast cancer

The association between CIN and cancer prognosis is complex and paradoxical. Some studies showed association of CIN with poorer cancer prognosis (Orsetti, et al., 2014; Carter, et al., 2006), whereas other studies reached opposite conclusions, suggesting that excessive level of CIN would suppress tumor

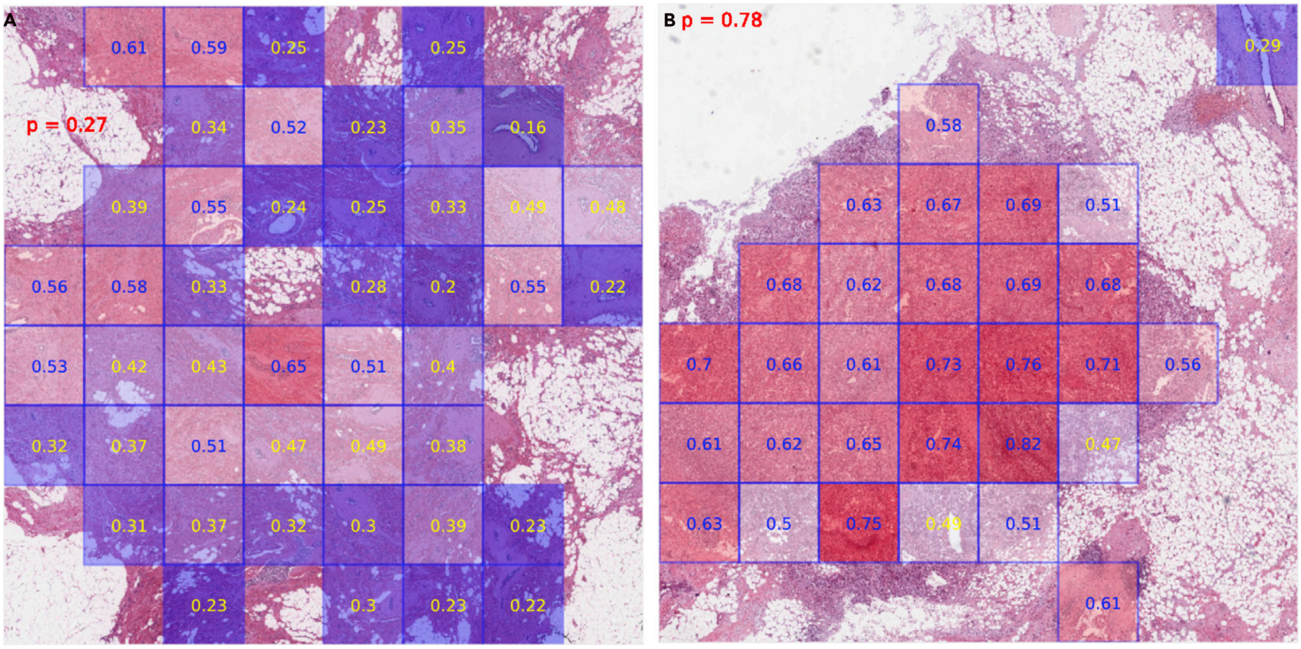


Figure 4. Intra-tumor heterogeneity by patch predictions

Red shows predicted probability of high-CIN slide based on our deep learning model. Numbers within each grid imply patch-level predictions. Blue indicates high-CIN patch; yellow indicates low CIN patch.

(A) An example of low-CIN slide in test set.

(B) An example of high-CIN slide in test set

progression and lead to better clinical outcomes (Birkbak, et al., 2011; Zasadil, et al., 2016). To further investigate this point, we conducted a survival analysis in the TCGA cohort aiming to explore the prognostic values of different CIN scores. In these analyses, we used maximally selected rank statistics to determine optimal prognostic CIN score cutoffs and log rank tests to evaluate the differences between survival curves. We found that high genomic CIN is associated with poorer 5 years' prognosis compared with low genomic CIN, where prognosis is measured as time to any events including new tumors or mortality (Figure 5A; p value = 0.0023). Pathology-predicted CIN using our deep learning model was also predictive of outcomes (Figure 5B; p value = 0.0045). Finally, the predicted CIN-High fraction score was also correlated with worse outcomes (Figure 5C; p value = 0.0086). We postulated that the presence of patches with high predicted CIN scores within each slide may be sufficient to impact clinical outcomes. We calculated different percentile CIN scores based on all patches of each slide (75th, 95th, and maximum). We found that all slide-level percentile CIN scores were prognostic (Figures 5D–5F; p values: CIN-75th, 0.0085; CIN-95th, 0.0018; CIN-max, 0.02) with CIN-95th being the most robust prognostic score, even more predictive than genomic CIN. That CIN-95th is more predictive than CIN-max can be explained by the lack of stability of maximal value and possibly by the need for more than one patches to have high CIN to impact outcomes.

CIN is associated with profound transcriptional changes in tumor samples

We reasoned that our ability to predict CIN based on histopathology slides may underlie a relatively profound difference in biological features between CIN low and CIN high tumors, which may influence cell morphology and tissue structure in H&E slides. We therefore conducted differentially expressed genes and gene set enrichment analysis between high CIN and low CIN samples. (Figure 1F) To minimize the confounding effect of cancer subtypes caused by the unbalanced subtype distributions across CIN groups, we adjusted tumor subtypes in the design matrix and tested differentially expressed genes for genomic CIN term. A total of 307 differentially expressed genes were identified (logFC>1, adjusted p value<0.05) between CIN low and CIN high tumors as shown in Figure S10A. Cell cycle and mitosis-related gene signatures were up-regulated significantly in high-CIN tumor samples (Figure S10B). This analysis thus confirms substantial biological differences between CIN high and CIN low tumors.

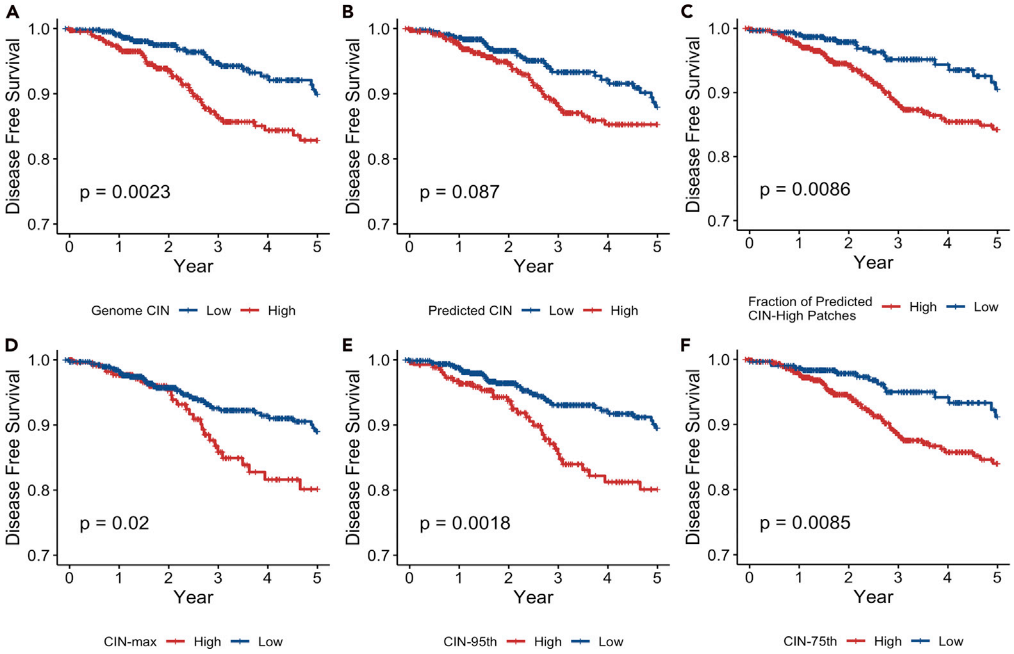


Figure 5. Kaplan-Meier curves of disease-free survival probabilities grouped by different CIN biomarkers

p values were calculated by log rank test.

(A) Genomic CIN (Stratification cutoff: 0.3).

(B) Predicted CIN (Stratification cutoff: 0.44).

(C) Fraction of predicted CIN high patches (Stratification cutoff: 0.42).

(D) CIN-max, the maximum patch prediction within each slide (Stratification cutoff: 0.77).

(E) CIN-95th, the 95th percentile patch prediction within each slide (Stratification cutoff: 0.72).

(F) CIN-75th, the 75th percentile patch prediction within each slide (Stratification cutoff: 0.55).

DISCUSSION

In this study we demonstrate for the first time the ability to predict CIN based on H&E slides. At present, it is challenging to capture the ongoing rate of chromosome mis-segregation to identify CIN in routine clinical setting since mitotic alterations are rare in H&E slides; other assays such as microscopy or micronuclei staining have not been deployed in the clinical setting. Here we demonstrated the ability of using histopathology slide images to predict CIN status of each patient and achieved high accuracy (=74.3%) and sensitivity (=81.2%). Of equal importance, our model indicates the existence of intra-tumor heterogeneity in CIN levels and revealed its association with poorer clinical outcomes. Further research, perhaps based on regional sequencing, is needed to further validate these findings. The substantial prognostic impact potentially exerted by spatial sub-regions (patches) with the highest CIN scores is important since it indicates that such regions may drive response to treatment. Future treatment modalities may need to focus on eliminating CIN high tumor cells if they are to achieve maximal therapeutic impact. Either way, our results pave the way for integrating CIN as prognosis biomarker and therapeutic vulnerability into existing clinical pathology settings.

One of the main challenges of computational pathology is to manage the trade-off between abundant morphological information and large size of WSI. Splitting WSI into hundreds of thousands of patches and training neural networks on the patch level is a commonly used strategy (Srinidhi et al. 2021). As mentioned above, there are several studies that successfully demonstrated the ability of using H&E-stained histology to predict genetic mutations using this patch-level learning approach (Khosravi, et al., 2018; Kather et al., 2019). We also experimented using patients' level labels to supervise patch learnings directly with the same approach but failed in predicting CIN levels. We reasoned there might exist substantial intra-tumor heterogeneity within individual slides and that therefore using patient-level CIN labels are not directly applicable to patches for training. To overcome this obstacle for CIN status learning, we applied a weakly supervised learning approach. More specifically, we used transfer learning to extract patch-level features followed by addition of a max-pooling layer with only maximum feature values kept along each feature dimension so that most irrelevant features that would potentially add noise to the model learning throughout the WSI were removed during this step. The features that were kept training the top fully connected layers were distributed widely throughout the WSI but not from a local region so that intra-tumor heterogeneity problem is reduced during training. We further compared model performance between slide predictions with patch predictions in predicting slide-level genomic CIN status to see the influence of this feature aggregation strategy. Contrasted with slide-level predictions that obtained an AUC of 0.82, patch-level predictions only got an AUC of 0.66 (Figure S11). This difference can be explained by patch-level predictions only representing predictions based on local morphological features, whereas slide predictions look at global morphological features. In addition, slide-level predictions exhibited a high correlation with continuous genomic CIN score (Figure S12, Spearman rho, 0.52; p value<0.0001), whereas only binary labels were available during the training. We successfully demonstrated the effectiveness of this strategy by achieving high accuracy (=74.3%) in classifying genomic CIN status in the test dataset.

Digital pathology images can be examined at different magnification levels. We experimented on both 2.5× and 10× magnified tiles for the predictions. We found that 2.5× achieved more accurate predictions marginally than 10× (2.5× AUC, 0.82; 10× AUC, 0.76; p value = 0.06) and multi-scaled model by combining 2.5× with 10× magnification features (2.5× AUC, 0.82; multi-scaled AUC, 0.81; p value = 0.59), although not statistically different tested by DeLong's method (Figure S4). We reasoned that each tile on the 2.5× level can capture more relevant features with a wider spatial view than high-resolution tiles, whereas on the 10× magnification level, tiles were more likely to be covered up by some "unknown" irrelevant features. A similar observation was made in Coudray's study (Coudray, et al., 2018) that analyzing 5× patches led to higher accuracy than 20× patches.

To validate the rationale of utilizing pathological slide images to infer genomic alterations, we performed differentially expressed gene analysis between CIN low and CIN high patients. The results revealed the impact of CIN in breast cancer including activating multiple pathways relating to cell cycle and mitosis (Figure S10B). As expected, mitotic alterations can be identified in slide images on high-resolution views. Future studies may focus on training machine learning models to detect aberrant mitotic events directly from H&E slides. This will require a very large training set of such events, which is currently not available.

Limitations of the study

In this study we used a bulk genomic aneuploidy burden approach to approximate patients' CIN status. Aneuploidy burden is a good but imperfect proxy for CIN, in that it focuses on highly clonal events and does not capture ongoing CIN. Moreover, in the absence of multi-region genomic sequencing, it is not possible to get a regional genomic CIN estimation. This limits our ability to demonstrate that our approach can capture spatial heterogeneity in CIN status. Better measurement for CIN can provide more accurate labels for training deep learning models as well as validating patch-level predictions. In addition, our results showed models on 2.5× magnification performed better than 10× magnification. We speculate that patches on a lower magnification scale contain more spatial context compared with the ones with higher resolution of the same pixel sizes, thus explaining our results. On the other hand, using a lower resolution may miss important morphological feature differences including abnormal mitosis caused by CIN. Combining our low-resolution approach with sophisticated computer vision approaches that identify features such as aberrant mitotic events as assessed manually for a small number of image patches in our study may help improve CIN prediction accuracy in the future.

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Dr. Olivier Elemento (ole2001@med.cornell.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All image and genetic data associated with this study can be downloaded from TCGA website (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga). The source code and the guideline are publicly available at https://github.com/eipm/CIN.

METHODS

All methods can be found in the accompanying transparent methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102394.

ACKNOWLEDGMENTS

This work was supported by NIH (UL1TR002384, R01CA194547) and Leukemia & Lymphoma Society Specialized Center of Research Program (180078-02, 7021-20).

AUTHORS CONTRIBUTION

O.E. conceived and designed the study. Z.X. designed and implemented image analysis workflow with supervision from P.K. A.V. and Z.X. analyzed the genomic and transcriptomic data. Z.X. performed statistical analysis. U.N. conducted histopathological inspection for mitosis events, and S.B. validated the results. Z.X. drafted the manuscript with revisions from O.E., P.K., A.V., and S.B. All authors contributed to the interpretation of the results and provided critical feedback and helped shape the research and the article. O.E. supervised the study.

DECLARATION OF INTERESTS

A.V. is a full-time employee of Volastra Therapeutics. O.E. and S.B. are co-founders and hold equity in Volastra Therapeutics. Volastra Therapeutics develops therapeutic strategies to target CIN in cancer.

Received: December 14, 2020
Revised: March 3, 2021
Accepted: April 1, 2021
Published: May 21, 2021

REFERENCES

Baergen, A.K., Jeusset, L.M., Lichtensztejn, Z., and McManus, K.J. (2019). Diminished condensin gene expression drives chromosome instability that may contribute to colorectal cancer pathogenesis. Cancers *11*, 1066.

Bakhom, S.F. (2018). Chromosomal instability drives metastasis through a cytosolic DNA response. Nature *553*, 467–472.

Bakhoum, S.F., and Cantley, L.C. (2018). The multifaceted role of chromosomal instability in cancer and its microenvironment. Cell *147*, 1347–1360.

Birkbak, N.J., Eklund, A.C., Li, Q., McClellan, S.E., Endesfelder, D., Tan, P., Tan, L.B., Richardson, A.L., Szallasi, Z., and Swanton, C. (2011). Paradoxical relationship between chromosomal instability and survival outcome in cancer. Cancer Res. *71*, 3447–3452.

Burrell, R., McClelland, S.E., Endesfelder, D., Groth, P., Marie-Christine, W., Shaikh, N., Domingo, E., Kanu, N., Dewhurst, S.M., and Gronroos, E. (2013). Replication stress links structural and numerical cancer chromosomal instability. Nature *494*, 492–496.

Cailleau, R., Olive, M., and Cruciger, Q.V.J. (1978). Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization. In Vitro *14*, 911–915.

Carter, S.L., Eklund, A.C., Kohane, I.S., Harris, L.N., and Szallasi, Z. (2006). A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat. Genet. *38*, 1043–1048.

Chen, M., Zhang, B., Topatana, W., Cao, J., Zhu, H., Juengpanich, S., Mao, Q., Hong, Yu, and Cai, X. (2020). Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. NPJ Precis. Oncol. *4*, 14.

Coudray, N., Moreira, A.L., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat. Med. *24*, 1559–1567.

Fu, Y., Jung, A.W., and Ramon Vinas, T. (2020). Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat. Cancer *1*, 800–810.

Hieronymus, H., Murali, R., Amy, T., Yadav, K., Abida, W., Moller, H., Berney, D., Howard, S., Carver, B., and Peter, S. (2018). Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. Elife *7*, e37294, https://doi.org/10.7554/eLife.37294.

Jamal-Hanjani, M., A'Hern, R., Nicolai Juul, B., Gorman, P., Grönroos, E., Ngang, S., Nicola, P., Rahman, L., Thanopoulou, E., and Kelly, G. (2015). Extreme chromosomal instability forecasts improved outcome in ER-negative breast cancer: a prospective validation cohort study from the TACT trial. Ann. Oncol. *26*, 1340–1346.

Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., and Neumann, U.P. (2019). Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat. Med. *25*, 1054–1056.

Kather, J.N., Heij, L.R., Grabsch, H.I., Loeffler, C., Echle, A., Hannah, S.M., Krause, J., Niehues, J.M., Sommer, K.A.J., and Peter, B. (2020). Pan-cancer image-based detection of clinically actionable genetic alterations. Nat. Cancer *1*, 789–799.

Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O., and Hajirasouliha, I. (2018). Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. EBioMedicine *27*, 317–328.

Kwei, K.A., Kung, Y., Salari, K., Holcomb, I.N., and Pollack, J.R. (2010). Genomic instability in breast cancer: pathogenesis and clinical implications. Mol. Oncol. *4*, 255–266.

Lee, A.J.X., Endesfelder, D., Rowan, A.J., Walther, A., Birkbak, N.J., Futreal, P.A., Downward, J., Szallasi, Z., Tomlinson, I.P.M., and Howell, M. (2011). Chromosomal instability confers intrinsic multi-drug resistance. Cancer Res. *71*, 1858–1870.

Lengauer, C., Kinzler, K.W., and Vogelstein, B. (1998). Genetic instabilities in human cancers. Nature *396*, 643–649.

Liao, H., Long, Y., Han, R., Wang, W., Xu, L., Liao, M., Zhang, Z., Wu, Z., Shang, X., and Li, X. (2020). Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma. Clin. Transl. Med. *10*, e102.

Murayama-Hosokawa, S., Oda, K., Nakagawa, S., Ishikawa, S., Yamamoto, S., Shoji, K., Ikeda, Y., Uehara, Y., Fukayama, M., and McCornick, F. (2010). Genome-side single-nucleotide polymorphism arrays in endometrial carcinomas associate extensive chromosomal instability with poor prognosis and unveil frequent chromosomal imbalances involved in the PI3-kinase pathway. Oncogene *29*, 1897–1908.

Noorbakhsh, J., Farahmand, S., Foroughi pour, A., Namburi, S., Caruana, D., Rimm, D., Soltanieh-ha, M., Zarringhalam, K., and Chuang, J.H. (2020). Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. Nat. Commun. *11*, 6367.

Orsetti, B., Selves, J., Bascoul-Mollevi, C., Lasorsa, L., Gordien, K., Bibeau, F., Massemin, B., Paraf, F., Soubeyran, I., and Hostein, I. (2014). Impact of chromosomal instability on colorectal cancer progression and outcome. BMC Cancer *1*, 1–13.

Penner-Goeke, S., Lichtensztejn, Z., Neufeld, M., Ali, Jennifer L., Altman, A.D., Nachtigal, M.W., and McManus, K.J. (2017). The temporal dynamics of chromosome instability in ovarian cencer cell lines and primary patient samples. PLoS Genet. *13*, e1006707, https://doi.org/10.1371/journal.pgen.1006707.

Petersen, I., Kotb, W.F.M.A., Friedrich, K.-H., Schlüns, K., Böcking, A., and Dietel, M. (2009). Core classification of lung cancer: correlating nuclear size and mitoses with ploidy and clinicopathological parameters. Lung Cancer *65*, 312–318.

Pierssens, D.D.C.G., Maarten, C.B., van der Heijden, S.J.H., Peutz-Kootstra, C.J., Ruland, A.M., Annick, M.H., Kessler, P.A.W.H., Kremer, B., and Speel, E.-J.M. (2017). Chromosome instability in tumor resection margins of primary OSCC is a predictor of local recurrence. Oral Oncol. *66*, 14–21.

Pikor, L., Kelsie Thu, Vucic, E., and Lam, W. (2013). The detection and implication of genome instability in cancer. Cancer Metastasis Rev. *32*, 341–352.

Schaumberg, A.J., Rubin, M.A., and Fuchs, T.J. (2016). H&E-stained Whole Slide Image Deep Learning Predicts SPOP Mutation State in Prostate Cancer (bioRxiv), p. 064279, https://doi.org/10.1101/064279.

Schonhoft, J.D., Zhao, J.L., Adam, J., Carbone, E.A., Barnett, E.S., Hullings, M.A., Gill, A., Sutton, R., Lee, J., and Dago, A.E. (2020). Morphology-predicted large scale transition number in circulating tumor cells identifies a chromosomal instability biomarker associated with poor outcome in castration-resistant prostate cancer. Cancer Res. *80*, 4892–4903.

Sipos, O., Tovey, H., Quist, J., Haider, S., Gazinska, P., Kernaghan, S., Toms, C., Maguire, S., Orr, N., and Linn, S.C. (2021). Assessment of structural chromosomal instability phenotypes as biomarkers of carboplatin response in triple negative breast cancer: the TNT trial. Ann. Oncol. *32*, 58–65.

Smid, M., Hoes, M., Sieuwerts, A.M., Sleijfer, S., Zhang, Y., Wang, Y., Foekens, J.A., and Martens, J.W.M. (2011). Patterns and incidence of chromosomal instability and their prognostic relevance in breast cancer subtypes. Breast Cancer Res. Treat. *128*, 23–30.

Srinidhi, C.L., Ciga, O., and Martel, A.L. (2021). Deep neural network models for computational histopathology: a survey. Med. Image Anal. *67*, 101813.

Swanton, C., Nicke, B., Schuett, M., Eklund, A.C., Ng, C., Li, Q., Hardcastle, T., Lee, A., Roy, R., and East, P. (2009). Chromosomal instability determines taxane response. PNAS *106*, 8671–8676.

Thompson, L.L., and McManus, K.J. (2015). A novel multiplexed, image-based approach to detect phenotypes that underlie chromosome instability in human cells. PLoS One *10*, e0123200.

Tijhuis, A.E., Johnson, S.C., and McClelland, S.E. (2019). The emerging links between chromosomal instability (CIN), metastasis, inflammation and tumour immunity. Mol. Cytogenet. *12*, 17.

Walther, A., Houlston, R., and Tomlinson, I.P.M. (2008). Association between chromosomal instability and prognosis in colorectal cancer: a meta-analysis. Gut *57*, 941–950.

Xu, H., Park, S., Hak Lee, S., and Tae Hyun, H. (2019). Using Transfer Learning on Whole Slide Images to Predict Tumor Mutational Burden in Bladder Cancer Patients (bioRxiv), p. 554527, https://doi.org/10.1101/554527.

Zasadil, L.M., Andersen, K.A., Yeum, D., Rocque, G.B., Wilke, L.G., Tevaarwerk, A.J., Raines, R.T., Burkard, M.E., and Weaver, B.A. (2014). Cytotoxicity of paclitaxel in breast cancer is due to chromosome missegregation on multipolar spindles. Sci. Transl. Med. *6*, 229ra43.

Zasadil, L.M., Britigan, E.M.C., Ryan, S.D., Kaur, C., Guckenberger, D.J., Beebe, D.J., Moser, A.R., and Weaver, B.A. (2016). High rates of chromosome missegregation suppress tumor progression but do not inhibit tumor initiation. Mol. Biol. Cell *27*, 1981–2144.

Zeimet, A.G., Fiegl, H., Goebel, G., Kopp, F., Claude, A., Reimer, D., Ilona Steppan, Mueller-Holzner, E., Ehrlich, M., and Marth, C. (2011). DNA ploidy, nuclear size, proliferation index and DNA-hypomethylation in ovarian cancer. Gynecol. Oncol. *121*, 24–31.

iScience, Volume *24*

Supplemental information

Deep learning predicts chromosomal instability from histopathology images

Zhuoran Xu, Akanksha Verma, Uska Naveed, Samuel F. Bakhoum, Pegah Khosravi, and Olivier Elemento

Supplemental Figures and Tables

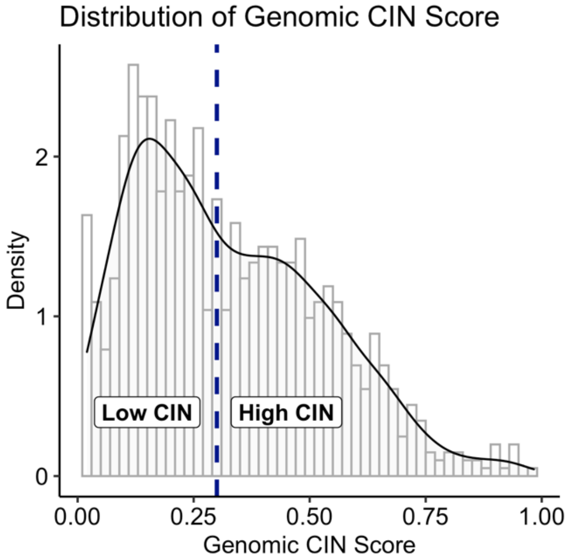


Figure S1. Distribution of Genomic CIN Score. Related to Figure 1. Fraction genome altered (FGA) was calculated to represent genome CIN score. Blue dash line indicates genomic CIN score of 0.3. Patients with genomic CIN score higher than 0.3 were classified as high CIN, otherwise as low CIN.

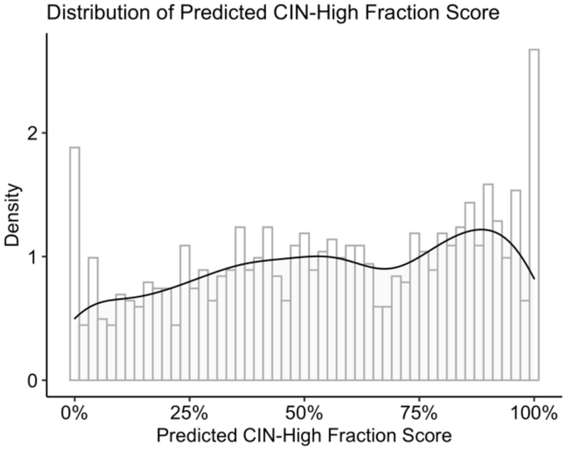


Figure S2. Histogram of Predicted CIN-High Fraction Score. Related to Figure 4. Predicted CIN-High fraction score is defined as the percentage of predicted high CIN patches based on each pathology slide image.

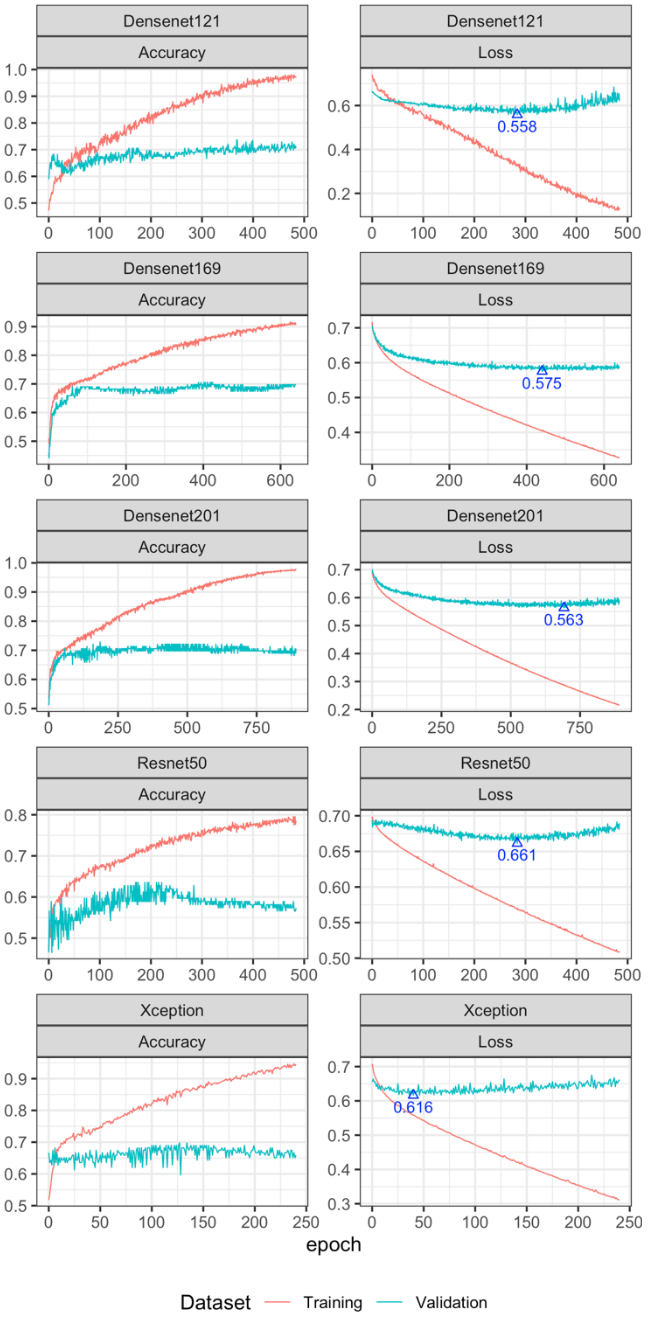


Figure S3. Model learning curves during training. Related to Figure 2. Triangle point implies the epoch with the lowest validation loss where models were chosen for further evaluations in blind test dataset.

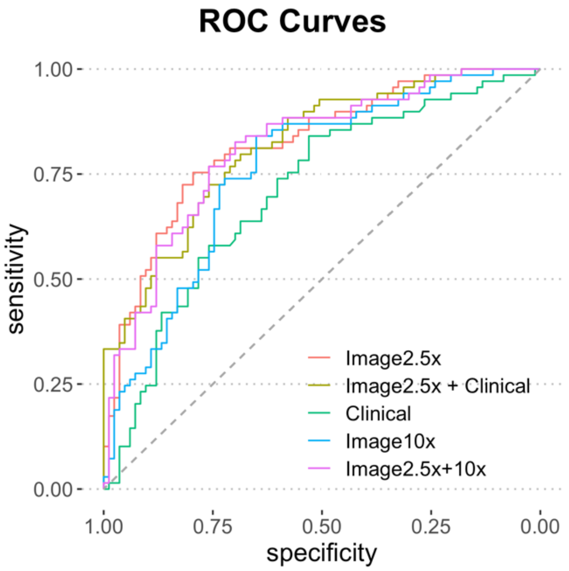


Figure S4. Receiver operation characteristic (ROC) curves of different model configurations. Related to Figure 2. Image 2.5x: AUC 0.82 (0.75, 0.89). Image 2.5x and clinical: AUC 0.82 (0.75, 0.89). Clinical variables only: AUC 0.71 (0.63, 0.79). Image 10x: AUC 0.76 (0.68, 0.84). Multiscale by combining 2.5x and 10x: AUC 0.81 (0.74, 0.88).

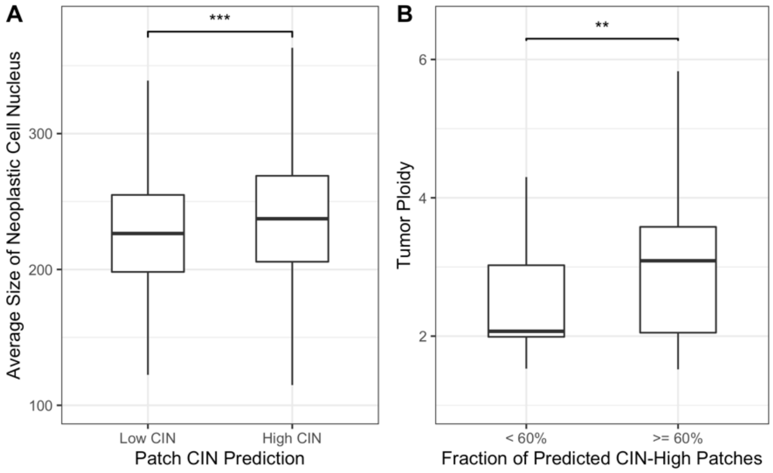


Figure S5. Boxplots of average size of neoplastic cell nucleus and tumor ploidy in test dataset. Related to Figure 4. (A) Boxplot of average size of neoplastic cell nucleus (pixels) between predicted low CIN patches and predicted high CIN patches. (Wilcoxon test, *** p-value<0.0001) **(B)** Tumor ploidy (patient level) between slides with low and high fraction of predicted CIN-High patches. Cutoff of 60% was chosen based on median value of fraction of predicted CIN-High patches in test dataset. (Wilcoxon test, ** p-value=0.0031)

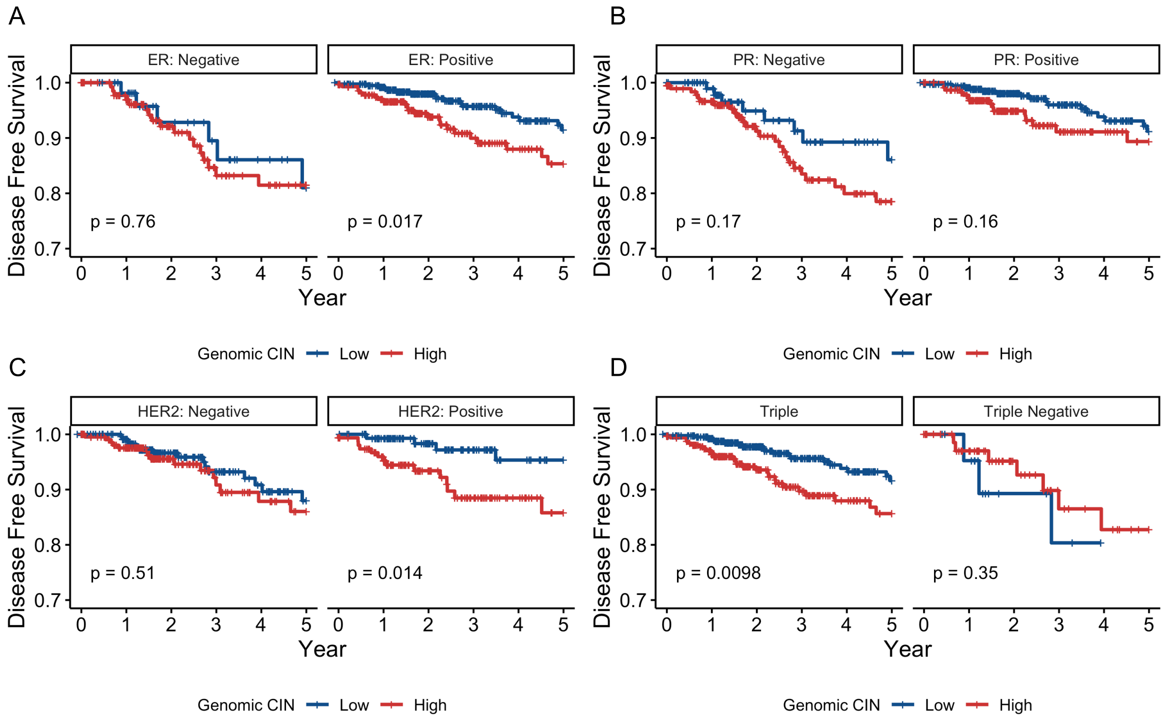


Figure S6. Kaplan-Meier curves of disease-free survival (DFS) probabilities by genomic CIN for different subtypes of breast cancer. Related to Figure 5. P-values were calculated by log rank test. Stratification cutoff: 0.3.

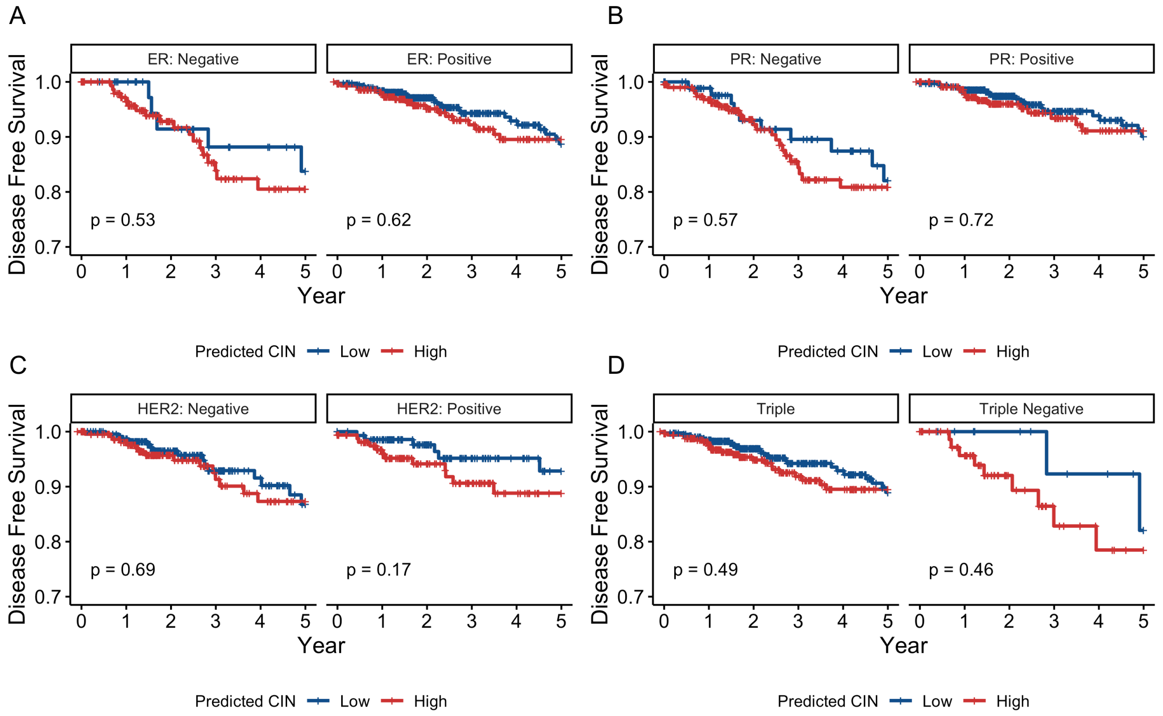


Figure S7. Kaplan-Meier curves of disease-free survival (DFS) probabilities by predicted CIN for different subtypes of breast cancer. Related to Figure 5. P-values were calculated by log rank test. Stratification cutoff: 0.44.

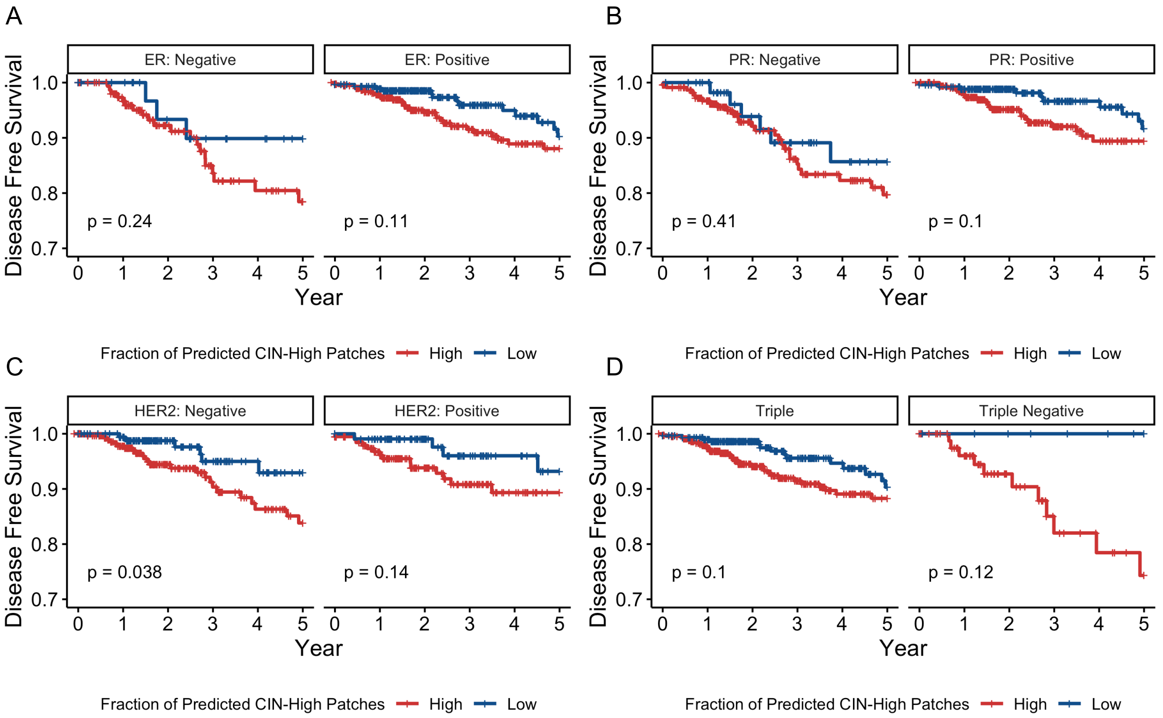


Figure S8. Kaplan-Meier curves of disease-free survival (DFS) probabilities by pathological CIN for different subtypes of breast cancer. Related to Figure 5. P-values were calculated by log rank test. Stratification cutoff: 0.42.

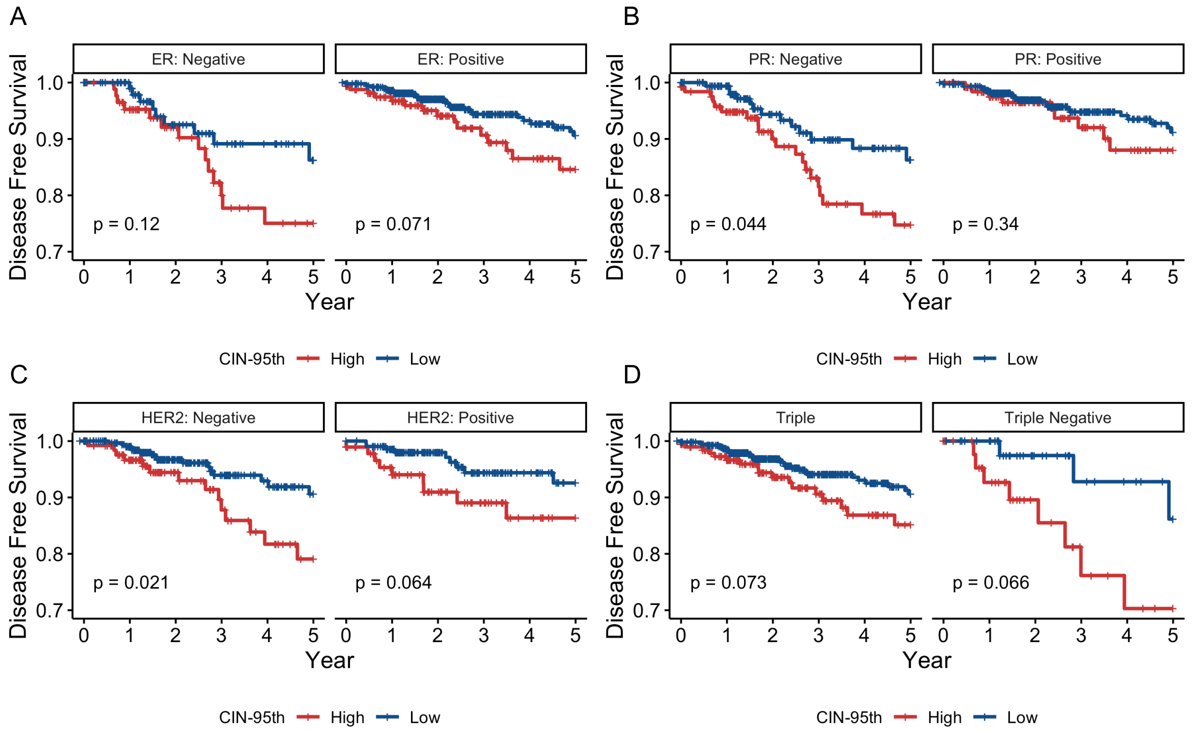


Figure S9. Kaplan-Meier curves of disease-free survival (DFS) probabilities by 95th percentile patch CIN for different subtypes of breast cancer. Related to Figure 5. P-values were calculated by log rank test. Stratification cutoff: 0.72.

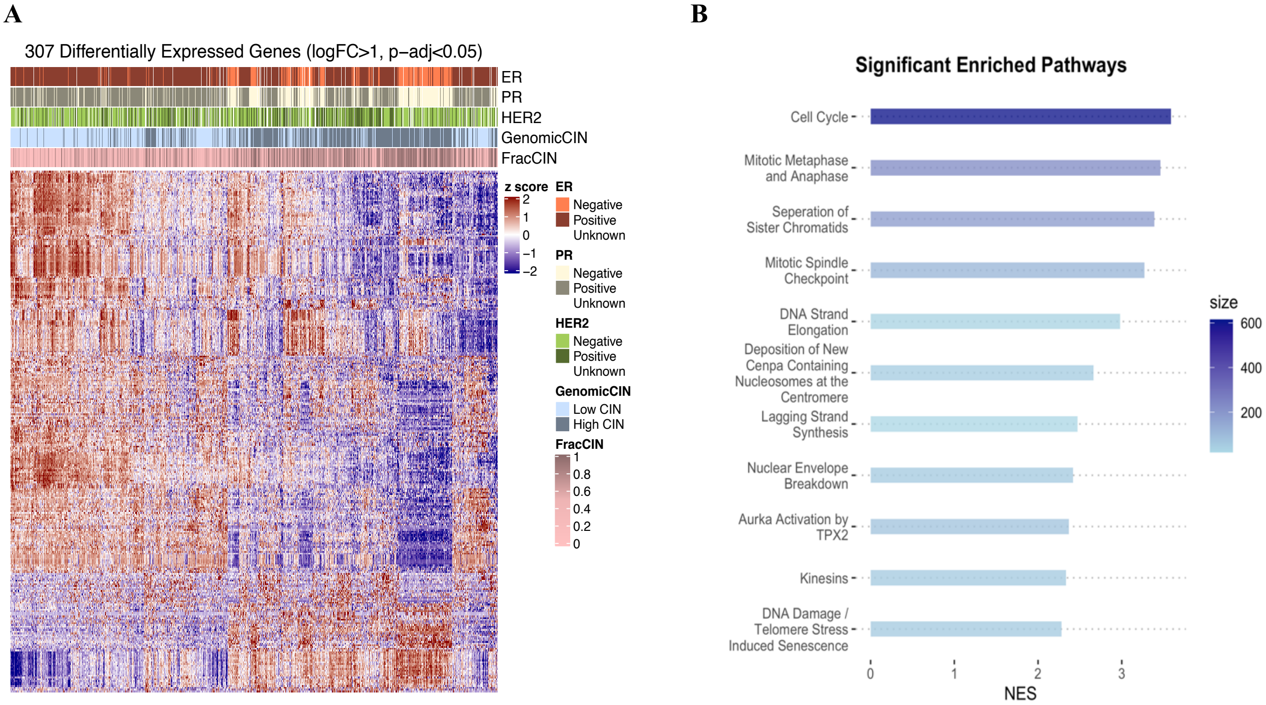


Figure S10. Differentially expressed genes and gene set enrichment analysis. Related to Figure 1F and Figure 3. (A) Heatmap of top differentially expressed genes between high CIN and low CIN with logFC>1 and ajudusted p<0.05. FracCIN represents predicted CIN-High Fraction. **(B)** Gene set enrichment analysis.

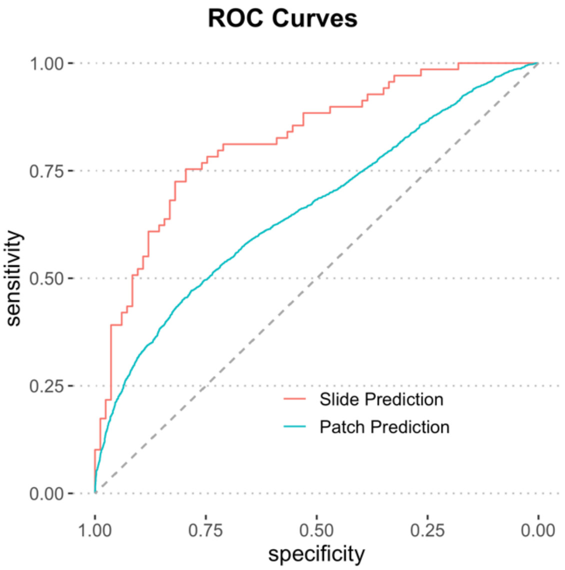


Figure S11. Receiver operation characteristic (ROC) curves for slide predictions and patch predictions from Densenet121 model. Related to Figure 2. Slide Prediction (Slide level prediction from aggregated slide level feature embedding): AUC 0.82 (0.75, 0.89). Patch Prediction (Patch level predictions evaluated by slide level labels): AUC 0.66 (0.65, 0.67)

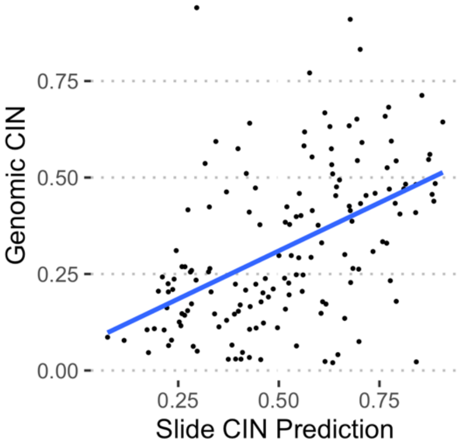


Figure S12. Scatter plots of slide CIN predictions with genomic CIN. Related to Figure 2. Blue lines represent regressed correlation lines. High CIN is defined by genomic CIN larger than 0.3. Spearman Correlation coefficient: 0.52, *** p-value<0.0001

Table S1. Patient and WSI numbers for training, validation and test set. Related to Figure 1 and Results. Statistics are in format of: Patient number (WSI number, Patch number). †: Total number of patients in training and validation set is 858, with further split into 730 for training (85%) and 128 (15%) for validation.

	Whole Dataset	Training and Validation Set	Test Set
High CIN	485 (515, 23427)	416 (441, 20128)	69 (74, 3299)
Low CIN	525 (550, 23568)	442 (463, 20045)	83 (87, 3523)
Total	1010 (1065, 46995)	858 (904, 40173) †	152 (161, 6822)

Table S2. Sample size and model performance in test dataset for very high CIN vs very low CIN and moderate high CIN vs moderate low CIN. Related to Figure 2.

	Comparison Samples	**Test Sample Size**	**Accuracy**	**AUC-ROC**
vs	<0.2 (very low CIN)	50	0.776	0.83
	>0.4 (very high CIN)	57		(0.75, 0.91)
vs	0.2-0.3 (moderate low CIN)	33	0.667	0.77
	0.3-0.4 (moderate high CIN)	12		(0.61, 0.93)

Table S3. Genomic CIN is associated with gene-expression subtypes of ER, PR and Triple Status. Related to Figure 2 and Results. Numbers in the table represent counts (percentage in each genome CIN group).

		Genomic CIN		p value
		Low	High	
ER	Positive	436 (87.0)	302 (65.4)	<0.0001
	Negative	65 (13.0)	160 (34.6)	
PR	Positive	387 (77.2)	251 (54.4)	<0.0001
	Negative	114 (22.8)	210 (45.6)	
HER2	Positive	158 (36.3)	171 (41.4)	0.1
	Negative	227 (63.7)	242 (58.6)	
Triple	Triple	467 (94.7)	353 (80.8)	<0.0001
	Tripe Negative	26 (5.27)	84 (19.2)	

Table S4. Model performance for subgroups based on molecular breast cancer types. Related to Figure 2, Results and Table S3.

		ROC	P
ER	Positive	0.79 (0.70, 0.87)	0.33
	Negative	0.87 (0.73, 1.00)	
PR	Positive	0.81 (0.73, 0.90)	0.94
	Negative	0.81 (0.67, 0.94)	
HER2	Positive	0.84 (0.71, 0.96)	0.41
	Negative	0.77 (0.66, 0.87)	
Triple	Triple	0.77 (0.68, 0.86)	0.86
	Tripe Negative	0.79 (0.57, 1)	
Overall		0.82 (0.75,0.89)	

Table S5. Correlations between transcriptional CIN score with genomic CIN score and Predicted CIN-High Fraction score. Related to Figure 1E, Figure 1F and Results. ER: Estrogen receptor. PR: Progesterone receptor. HER2: Human epidermal growth factor receptor 2.

	Genomic CIN		**Predicted CIN-High Fraction**	
	Correlation	P	Correlation	P
ER-	0.2005	0.0026	-0.0052	0.9379
ER+	0.0466	0.2079	0.0338	0.361
PR-	0.2157	0.0001	0.113	0.0425
PR+	0.0237	0.5524	-0.0021	0.9572
HER2-	0.1838	<0.0001	0.1495	0.0007
HER2+	0.0172	0.756	0.0182	0.7422
Non Triple	0.0634	0.0704	0.0463	0.1869
Triple-	0.1769	0.0659	-0.0464	0.6322
Overall	0.1403	<0.0001	0.0952	0.0026

Transparent Methods:

Dataset:

Whole slide images along with clinical and genomic data were downloaded from The Cancer Genome Atlas (TCGA), project TCGA-BRCA using the TCGAbiolinks R package (Colaprico, et al. 2015). Only formalin-fixed paraffin-embedded (FFPE) diagnostic H&E-stained histopathology slides from primary tumor sites were used for this study. After removing slides that lack magnification information, and/or slides with artefacts including tissue folding, air bubbles and out-of-focus regions, the cohort consisted of 1,010 patients with 1,070 whole slide images (WSI).

Fraction genome altered (FGA), which was defined as the ratio of *Sum of altered genome size/ Total genome size analyzed*, was calculated using whole exome sequencing based copy number variation (CNV) data on the same TCGA patients. Copy-number segments were downloaded from TCGA. Segments with log transformed mean copy number values larger than 0.2 or less than -0.2 (Salas, et al. 2017, Ali Hassan and Mokhtar 2014) were treated as altered segments, respectively. Based on examining the FGA distribution for all 1,010, patients, we labelled patients with FGA less than 0.3 as low CIN; those with FGA above 0.3 were labelled as high CIN.

Image Preprocessing:

First, a single overall region of interest (ROI) with dimension of 2,048x2,048 pixels on 2.5x magnification (4mpp) was determined by a sliding window approach from each whole slide image, which typically has dimension of about 8,000x4,500 pixels. A simple thresholding method was used to distinguish tissue from white space background on greyscale space. All pixels with value lower than 215 were treated as tissue, otherwise as background. The ROI window that contained the highest percentage of tissue was kept for further processing. The selected window was then split into 8x8 non-overlapping patches each with dimension of 256x256 pixels. Quality control on patch level was conducted using the following method. All patches with tissue percentage less than 80% or with significant blurriness, pen marks or folded tissues were deleted. Color normalization was then performed to reduce batch effects across different data sources (**Figure 1A**). We performed Reinhard normalization to transform the color characteristics to a desired standard defined by the mean and standard deviations of target image (TCGA-AN-A0FK) from the cohort using Python library of HistomicsTK (Gutman, et al. 2017). Patients with more than one WSIs were treated as having one big WSI and can get more than 64 (8x8) patches depending on how many WSIs they have, but 64 qualified patches were randomly chosen to prevent over-representing those patients. Overall, after deleting all unqualified patches, we obtained median of 51 (IQR: 32, 61) qualified patches for each patient that to be used for transfer learning.

Transfer Learning:

(1) Feature Extraction:

Instead of training CNN architectures from scratch, we used commonly used pre-trained models as feature extractors. We passed all patches (256x256 pixels) of each patient through Densenet-121, Densenet-169, Densenet-201 (Huang, et al. 2017, DOI: 10.1109/CVPR.2017.243), Xception (Chollet 2017) and Resnet-50 (He, Zhang, et al., Deep Residual Learning for Image Recognition 2016, DOI: 10.1109/CVPR.2016.90) networks that were pre-trained on ImageNet (Deng, et al. 2009) without top layers. Then we got a set of feature vectors for every patient with dimensions of m*n where m denotes number of patches of one patient and n implies number of features according to specific architecture that was used. For example, we got matrix of 32*1,024 for patch features with the patient who has 32 patches by using Densenet-121. To adapt patch level features to patient level labels and reduce the noise generated by intra-tumor heterogeneity, we applied a max-pooling layer on top of patch features and got patient level features with the same dimension for all patients (**Figure 1B**) (Courtiol, et al. 2018).

(2) Train Fully Connected Layers:

The whole cohort was randomly divided into training and validation set (n=858 patients, 85%) and hold-out testing set (n=152, 15%) without any overlap for both patients and images. Then the 858 patients were further split into training (730, 85%) and validation (128, 15%) set for tuning hyperparameters. We implemented several fully connected layers to take patient level features extracted by each of the CNN architectures mentioned above as input respectively (**Figure 1C**). Therefore, output by each model will be a prediction probability of high CIN class patient. Our model used binary cross-entropy loss function. We initialized the weights of fully connected layers with He initialization (He, Zhang, et al., Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification 2015). Adam optimizer was used for the training network weights with learning rate of 0.00001. Training was stopped early (Prechelt 2012) if validation loss was not improving within 200 epochs. Epoch numbers were selected as per lowest validation loss. Then model performance metrics including ROC curve and AUC, balanced accuracy, sensitivity and specificity were calculated in the hold-out test set for the final evaluation (**Figure 1D**).

Visualization of predicted patches:

The trained fully connected layers from last step can be fed with both combined patches (patient level features) and individual patches (patch level features) for hierarchical predictions. We predicted both patient level and patch level high CIN probabilities and visualized the predictions to demonstrate the existence of intra-tumor heterogeneity (**Figure 1E**).

Pathological, Genomic and Transcriptomic scoring of CIN:

We define predicted CIN-high fraction score as the percentage of predicted high CIN patches based on each pathology slide image. CIN-max, CIN-95th and CIN-75th are defined as maximum, 95th percentile and 75th percentile of patch predictions within each slide, respectively. Genomic CIN score was calculated by fraction genome altered as mentioned earlier (Ali Hassan and Mokhtar 2014). CIN23 gene signature (Bakhom 2018) score was computed using single sample Gene Set Enrichment Analysis (ssGSEA) to indicate transcriptional CIN score from RNA expression data of the same cohort.

Transcriptome analysis:

We examined differentially expressed genes of primary tumor between high CIN and low CIN patients using Limma R package (Ritchie, et al. 2015). Differential expressed genes and samples shown in heatmap were clustered using Euclidean distance. Gene set enrichment analysis was conducted using fgsea (Korotkevich, et al. 2019, DOI: https://doi.org/10.1101/060012) R package and Reactome pathway database (https://reactome.org).

Mitosis events inspection:

Among all the patients who have agreed predicted CIN and genomic CIN status, top five extreme high CIN and low CIN patients were selected according to genomic CIN score. Several tumor patches of one patient with dimension of 1,024x1,024 on 40x magnification were randomly inspected and atypical mitosis events number in each patch were recorded.

Nuclear segmentation and Tumor ploidy:

We performed nuclear instance segmentation for all the patches from test dataset and classified the nuclei into five categories (neoplastic cell, non-neoplastic epithelia, inflammatory cell, connective cell and dead cell) automatically using Hover-Net model (Graham, et al. 2019) which was pretrained in PanNuke dataset (Gamper, Koohbanani and Benet, et al. 2019, Gamper, Koohbanani and Benes, et al. 2020) that consists of 205,343 labeled nuclei from 19 different tissue types. The average size of neoplastic cell nucleus was calculated for each patch using the total number of pixels of neoplastic cell nuclei divided by the instance number of neoplastic cell nuclei. Tumor ploidy data in which the estimations were obtained from ABSOLUTE (Carter, et al. 2012) was downloaded from Pan-Cancer Atlas (https://gdc.cancer.gov/about-data/publications/pancanatlas) (Weinstein, et al. 2013).

Statistical analysis and Software:

Training of our DNN method was performed on local computer powered by one NVIDIA GeForce GT 640M GPU with 512 MB memory and one 2.7-GHz Quad-Core Intel Core i5 CPU. All statistical and bioinformatics analyses were performed in R, version 3.6.2. Image preprocessing and neural network training were conducted in Python, version 3.7.4. ROC curves were compared using DeLong's method by R package of pROC. Chi-square test was conducted to test the independency between cancer subtypes with genomic CIN status. Spearman's rank-order correlation test was performed for the correlation analysis without distribution assumption. Wilconxon rank sum test was used to compare the average size of neoplastic cell nucleus between high CIN and low CIN patches as well as tumor ploidy between slides with high and low fraction of predicted CIN-High patches. Log rank test was used for comparing Kaplan-Meier survival curves between different genome and predicted CIN groups. Time to any new tumor events and mortality was used as composite survival events and the data was censored at 5 years. Maximally selected rank statistics (Lausen, et al. 2004, Hothorn and Lausen 2003) was used to determine the optimal prognostic cutoff points for CIN biomarkers including predicted CIN, predicted CIN-High fraction, CIN-max, CIN-95th and CIN-75th. Generalized estimating equation (GEE) of Poisson regression model was used to compare atypical mitosis event number between CIN high and CIN low groups. All statistical tests were two sided with $p<0.05$ indicated significant. OpenSlide python was used for reading and tiling whole-slide images. TensorFlow2 were used for building and training neural networks.

Supplemental Reference

Ali Hassan, Nur Zarina, and Norfilza Mohd Mokhtar. 2014. "Integrated analysis of copy number variation and genome-wide expression profiling in colorectal cancer tissues." *Plos One* 9 (4).

Bakhom, Samuel F. 2018. "Chromosomal instability drives metastasis through a cytosolic DNA response." *Nature* 553: 467-472.

Carter, Scott L., Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, and Barbara A Weir. 2012. "Absolute quantification of somatic DNA alterations in human cancer." *Nature Biotechnology* 30: 413-421.

Chollet, François. 2017. "Xception: Deep Learning with Depthwise Separable Convolutions." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* DOI: 10.1109/CVPR.2017.195.

Colaprico, Antonio, Tiago C Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S Sabedot, Tathiane M Malta, Stefano M Pagnotta, and Isabella Castiglioni. 2015. "TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data." *Nucleic Acids Research* (Nucleic Acids Research) 44 (8): e71. DOI: 10.1093/nar/gkv1507.

Courtiol, Pierre, Eric W. Tramel, Marc Sanselme, and Gilles Wainrib. 2018. "Classification and disease Localization in histopathology using only global labels: A weakly-supervised approach." *arXiv: 1802.02212.*

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. "ImageNet: A large-scale hierarchical image database." *IEEE Conference on Computer Vision and Pattern Recognition* DOI: 10.1109/CVPR.2009.5206848.

Gamper, Jevgenij, Navid Alemi Koohbanani, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. 2020. "PanNuke Dataset Extension, Insights and Baselines." *arXiv preprint arXiv:2003.10778.*

Gamper, Jevgenij, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. 2019. "PanNuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification." In *European Congress on Digital Pathology*, 11-19. Springer.

Graham, Simon, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. 2019. "Hover-Net: Simultaneous segmentation and classification of nuclei in multitissue histology images." *Medical Image Analysis* 58.

Gutman, David A, Mohammed Khalilia, Sanghoon Lee, Michael Nalisnik, Zach Mullen, Jonathan Beezley, Deepak R Chittajallu, David Manthey, and Lee A D Cooper. 2017. "The digital slide archive: a software platform for management, integration and analysis of histology for cancer research." *Cancer Research* 77: e75-e78; DOI: 10.1158/0008-5472.CAN-17-0629.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016, DOI: 10.1109/CVPR.2016.90. "Deep Residual Learning for Image Recognition." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." *ICCV* 1026-1034.

Hothorn, Torsten, and Berthold Lausen. 2003. "On the exact distribution of maximally selected rank statistics." *Computational Statistics & Data Analysis* 43 (2): 121-137.

Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017, DOI: 10.1109/CVPR.2017.243. "Densely Connected Convolutional Networks." *IEEE Conference on Pattern Recognition and Computer Vision (CVPR).*

Korotkevich, Gennady, Vladimir Sukhov, Alexey Sergushichev, Boris Shpak, Maxim N. Artyomov, and Alexey Sergushichev. 2019, DOI: https://doi.org/10.1101/060012. "Fast gene set enrichment analysis." *bioRxiv.*

Lausen, Berthold, Torsten Hothorn, Frank Bretz, and Martin Schumacher. 2004. "Assessment of Optimal Selected Prognostic Factors." *Biomedical Journal* 46 (3): 364-374.

Prechelt , Lutz. 2012. "Early Stopping - But When?" In *Neural Networks: Tricks of the Trade*, 53-67. Springer.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. "Limma powers differential expression analyses for RNA-sequencing and microaary studies." *Nucleic Acids Research* 43 (7): e47, https://doi.org/10.1093/nar/gkv007.

Salas, Lucas A., Kevin C. Johnson, Devin C Koestler, Dylan E O'Sullivan, and Brock C Christensen. 2017. "Integrative epigenetic and genetic pan-cancer somatic alteration portraits." *Epigenetics* 12 (7): 561-574.

test. n.d.

Weinstein, John N, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. 2013. "The cancer genome AtlansPan-Cancer analysis project." *Nature Genetics* 45: 1113-1120.