# Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis

Yu Fu1, Alexander W Jung1, Ramon Viñas Torne1, Santiago Gonzalez1,2, Harald Vöhringer1, Mercedes Jimenez-Linan3, Luiza Moore3,4, and Moritz Gerstung#1,5

# to whom correspondence should be addressed

1) European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK.
2) Current affiliation: Institute for Research in Biomedicine (IRB Barcelona), Parc Científic de Barcelona, Barcelona, Spain.
3) Department of Pathology, Addenbrooke's Hospital, Cambridge, UK.
4) Wellcome Sanger Institute, Hinxton, UK
5) European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany.

**Correspondence:**

Dr Moritz Gerstung
European Molecular Biology Laboratory
European Bioinformatics Institute (EMBL-EBI)
Hinxton, CB10 1SA
UK.
Tel: +44 (0) 1223 494636
E-mail: moritz.gerstung@ebi.ac.uk

## Key points

- Pan-cancer computational histopathology analysis with deep learning extracts histopathological patterns and accurately discriminates 28 cancer and 14 normal tissue types
- Computational histopathology predicts whole genome duplications, focal amplifications and deletions, as well as driver gene mutations
- Wide-spread correlations with gene expression indicative of immune infiltration and proliferation
- Prognostic information augments conventional grading and histopathology subtyping in the majority of cancers

## Abstract

Here we use deep transfer learning to quantify histopathological patterns across 17,396 H&E stained histopathology image slides from 28 cancer types and correlate these with underlying genomic and transcriptomic data. Pan-cancer computational histopathology (PC-CHiP) classifies the tissue origin across organ sites and provides highly accurate, spatially resolved tumor and normal distinction within a given slide. The learned computational histopathological features correlate with a large range of recurrent genetic aberrations, including whole genome duplications (WGDs), arm-level copy number gains and losses, focal amplifications and deletions as well as driver gene mutations within a range of cancer types. WGDs can be predicted in 25/27 cancer types (mean AUC=0.79) including those that were not part of model training. Similarly, we observe associations with 25% of mRNA transcript levels, which enables to learn and localise histopathological patterns of molecularly defined cell types on each slide. Lastly, we find that computational histopathology provides prognostic information augmenting histopathological subtyping and grading in the majority of cancers assessed, which pinpoints prognostically relevant areas such as necrosis or infiltrating lymphocytes on each tumour section. Taken together, these findings highlight the large potential of PC-CHiP to discover new molecular and prognostic associations, which can augment diagnostic workflows and lay out a rationale for integrating molecular and histopathological data.

# Introduction

The diagnosis of cancer is typically established by histopathological assessment of specimens, in particular through microscopic examination of haematoxylin and eosin stained (H&E) sections. Over the last decades, diagnostics have been supplemented by genetic, epigenetic, transcriptomic and proteomic tests, which have helped to further refine grading and prognosis of cancer patients[1–6]. While such molecular tests are now reaching maturity in terms of automation, accuracy and reproducibility, morphological tissue assessment remains a laborious task carried out by trained histopathologists[7,8]. Computer vision and, in particular, deep convolutional neural networks (CNNs) have been shown to closely match the diagnostic accuracy of specialists and thus hold great promise to augment histopathology workflows[9–11]. Computational histopathology algorithms can process and cross-reference very large volumes of data, which may help pathologists to navigate and assess slides more quickly and help quantify aberrant cells and tissues.

At their core, deep learning algorithms build an implicit quantification of histopathological image features, which represent the patterns of the image as seen by the computer. These computational histopathological features are automatically learned for the original task of classifying the entire and/or subregions of images into cancer or non-cancerous tissues. However, once learned, the feature representation may also be used to find similar images[12] and quantify associations with traits beyond tissue types[13]. This approach, known as transfer or weakly supervised learning, has been used, for example, to establish associations with genetic variants, transcriptomic alterations and also survival[14,15].

Previous studies analysing H&E stained histopathological images have been primarily conducted in single cancer types making it difficult to compare the utility of computer vision across cancer types. We therefore performed a pan-cancer computational histopathology (PC-CHiP) analysis to study associations between computational histopathological features and genomic driver alterations, whole transcriptomes and survival. To this end, we applied quantitative imaging and transfer learning to a collection of 17,396 H&E stained fresh-frozen tissue image slides from The Cancer Genome Atlas (TCGA), containing specimens from 28 tumor types and 14 normal tissues with matched genomic, transcriptomic and outcome data. At its core, PC-CHiP is based on Inception-V4[16], an established CNN, which was fine-tuned to classify approximately 14 million small image sections and used the trained algorithm to convert each tile into a set of 1,536 image features. This quantitative histopathology representation accurately discriminated different tissues, displayed pervasive associations with underlying genomic alterations, wide-spread correlations with transcriptomic signatures and prognostic information across most cancer types. PC-CHiP highlights recurring histopathological patterns associated with genomic alterations, such as enlarged nuclei in samples with whole genome duplications. Further, the algorithm dissects transcriptomic signatures from a given sample and attributes the components, such as stromal tissues and lymphocytic

infiltrates to different areas on each slide by matching morphological patterns in a fully automated way. Similarly, PC-CHiP pinpoints prognostically relevant patterns such as necrosis on each slide. Together, these analyses highlight the powerful capabilities of computational histopathology to discover novel molecular and prognostic associations, which can augment diagnostic workflows and lay out a path to integrating histopathological and molecular data. The computational histopathology algorithm and analysis code are available at https://github.com/gerstung-lab/PC-CHiP.

# Results

## Accurate pan-cancer tissue classification

The data set comprised of 17,396 H&E stained tissue slides from 10,452 individuals and 28 cancer types including 14 with matched normal samples and 14 without (42 tissue types in total). 9,754 slides with tumor purity greater than 85% were split into 80% training and 20% validation data. Slides were tiled into more than 14 million 256µm x 256µm-sized tiles with a digital resolution of 512 by 512 pixels (**Figure 1a**). For 14 cancers with normal and tumor images, the average tumor/normal tissue classification AUC was 0.99 (all values given for the held-back validation set), ranging from 0.96 for normal head and neck to 0.99 for normal esophagus (**Figure 1b**). The average pan-tissue AUC of discriminating all 42 different tissues was 0.98 (range 0.91 to 0.99), including the 14 cancer types without matched normal samples (**Supplementary Table 1**; see discussion at the end of the manuscript for potential limitations). A similarly relevant task is the distinction of different histological subtypes for cancers of the same organ site. For example, we obtained high accuracy of differentiation between normal lung, squamous cell carcinoma and adenocarcinoma (AUC = 0.98, 0.90 and 0.91 respectively), comparable to previous estimates[10]. Similar results were obtained for kidney cancers, where we were able to accurately distinguish between clear cell, chromophobe and papillary renal cell carcinomas and normal kidney tissue (average AUC of 0.99 for all 3 subtypes and normal tissue, **Supplementary Table 1**).

To achieve this classification, PC-CHiP builds an image representation of each tile consisting of the output of the last 1,536 neurons of the network (**Figure 1a**). Hereafter we will refer to this output as computational histopathological features and demonstrate that this representation enables us to derive quantitative associations with a range of molecular traits. As the network was trained to discriminate different tissues, a two-dimensional UMAP representation (**see Methods**) of the computational histopathological features shows clusters corresponding to each tissue class (**Figure 1c, Supplementary Figure 1a**). There are nevertheless resemblances of related tissues, which cannot be solely attributed to the algorithm being trained to derive tissue-specific histopathological features alone (**Supplementary Figure 1b**). For example, we note a general tendency of tumors to cluster together, indicating a convergent histological phenotype, usually characterised by a high cell density and loss of tissue architecture – opposed to normal tissues, which tend to spread over the periphery in the feature space (**Figure 1c**). This tendency is also quantitatively reflected by a greater pairwise Euclidean distance between tumor and normal samples in the original feature space (mean

distance=18.2, range [13.5, 24.4]) compared to the distance between different tumors (mean distance=23.7, range [16.9, 28], *p*-value=$10^{-12}$, **Figure 1d**).

## Deconvolution of soft labels provides spatial resolution

PC-CHiP was trained on soft labels assigned at the level of whole slides, such as a pathologist estimate of 75% tumor and 25% normal tissue. While this assigns the same training values for each tile on a given slide, remarkably the network is capable of recognising which tiles on a given slide correspond to cancer and normal regions (**Figure 1e**). This reflects the fact that, based on the comparison of millions of tiles, the algorithm recognises that normal tiles from a tumor slide bear greater similarity with normal slides from the same organ site. Comparison of tiles with large tumor or normal probability with corresponding areas on the slide confirmed the predictions as judged by the local integrity of the tissue. Automatic deconvolution achieves a notable accuracy, with an average correlation between algorithm and pathologist-estimated tumor purity equals to 0.26 (range from 0.07 for cervical cancer to 0.6 for uveal melanoma; **Supplementary Figure 2**). Hence, we conclude that the trained network learned an internal representation capable of generating spatially resolved maps classifying individual tiles with accurate normal/tumor predictions. This demonstrates the power of large scale data analysis to deconvolve bulk signals, which will become useful when associating histopathological patterns with molecular data from bulk samples. We will discuss this in detail in the following section.

## Genomic alterations are associated with histopathology patterns

While the described above tissue classification accuracy and spatial decomposition are remarkable properties, these relations mostly reflect associations with known and predefined labels. A fundamental question is whether the underlying histopathology patterns, extracted by the CNN, are also predictive of genetics, transcriptomics and outcomes, as this enables us to explore novel genomic associations and define molecularly informed histological subtypes. PC-CHiP's representation of each image tile as a 1,536-dimensional vector allows us to use high-dimensional regression methods to correlate histopathological features with a variety of molecular measures (see **Methods**). Since the set of computational histopathological features was trained to discriminate cancer types, any evaluation of molecular associations across tissues may merely reflect the histopathological differences between cancers, which the algorithm was already shown to distinguish. Instead, we calculate accuracies separately for each tissue, using a validation set of held-back slides (see **Methods**) from different patients.

Overall, we observed widespread associations with a range of genetic alterations across many cancer types (**Figure 2a**). These range from large-scale genomic changes, including whole genome duplications, chromosomal aberrations, focal amplifications and deletions and point mutations in cancer driver genes.

## Accurate predictions of whole genome duplications

Whole genome duplications (WGD) occur in about 30% of solid tumors, leading to cells with a nearly tetraploid genome, likely as a result of a single failed mitosis, often preceded or succeeded by further chromosomal gains and losses[11]. WGD status could be predicted for 25 out of 27 cancer types with an average area under the receiver operating characteristic curve AUC of 0.79 (held back AUC > 0.5, false discovery rate FDR < 0.1, **Figure 2a, Supplementary Table 2**). 12 cancer types showed an AUC greater than 0.8. Remarkably, high AUC were also obtained for thyroid carcinoma in the validation set (AUC = 0.96) even though no WGD sample was included in the training cohort, demonstrating that the histopathological features revealing WGD are independent of the specific tissue type.

Interestingly, tiles with greater probability for WGD status showed an increased nuclear (haematoxylin) staining (**Figure 2b**), which could be explained by the increased amount of DNA. We therefore explicitly quantified the average cell nucleus size and intensity per tile using Cell profiler[12]. Indeed, the cell nucleus size and intensity were highly correlated with the WGD probability predicted by histopathological features, but gave a lower predictive accuracy (average AUC of 0.71, range [0.56,1], **Figure 2c, Supplementary Figure 3**). Also, conventional tumour subtypes and grade provided a much lower predictive accuracy (average AUC=0.59 for both). It thus appears that deep learning recognises a range of morphological patterns beyond nuclear size and grade alone and combines these into its predictions in a fully automated way.

## Copy number variants

### Chromosome arm-level gains and losses

Another recurrent class of alterations are arm-level copy number alterations. We performed regularised multinomial regression on the copy number status of 17 whole chromosome and 39 arm-level gains and losses[17]. In total, 105 out of 410 gain:cancer pairs and 167 out of 570 loss:cancer pairs (with at least 10 altered samples) showed a measurable association with image features (AUC > 0.5 with 95% CI, **Figure 2a, Supplementary Table 2**). Breast invasive carcinoma showed the largest number of associations (17 gains and 27 losses), followed by melanoma (10 gains and 22 losses) and kidney clear cell carcinoma (13 gains and 14 losses, **Supplementary Figure 4**).

Of note, gain of chromosome 8q was identified in 9 different cancer types, including esophageal carcinoma (AUC=0.82, 95% CI=[0.63, 1]), uveal melanoma (AUC=0.79, CI=[0.58, 1]) and kidney clear cell carcinomas (AUC=0.78, CI=[0.69, 0.88]). Tiles indicative of 8q gains displayed nuclei enlargement, high cellular density and low stromal content (**Supplementary Figure 5a**). Other notable gains with associations across multiple cancers included +20q in 8 cancers and +7q in 7 cancers (**Supplementary Table 2**). Conversely, loss of chromosome arm 17p, which harbours the *TP53* tumor suppressor gene, was identified in 8 cancer types such as colorectal adenocarcinoma

(AUC=0.78, CI=[0.67, 0.89]), lung adenocarcinoma (AUC=0.74 , CI=[0.64, 0.84]), indicating that –17p leads to measurable differences in tumor morphology, often characterised by a higher grade and loss of tissue structure (**Supplementary Figure 5b**). Of note, the distinction of –17p mutated and wild type endometrial carcinomas (AUC=0.74, CI=[0.60, 0.87]) largely coincides with the histopathological distinction of endometrioid, serous and mixed carcinomas (AUC=0.78, CI=[0.67, 0.89]). Other recurrent losses associated with histopathology were –11p in 8 cancers and –9q in 7 cancers (**Supplementary Table 2**).

**Focal amplifications and deletions**

Focal copy number alterations occur on the scale of several megabases and are thought to specifically lead to oncogene amplification and tumor suppressor gene deletions. We therefore investigated 140 focal copy number variants (70 amplifications and deletions each)[18]. Among 481 alteration:cancer pairs tested with more than 10 recurrences, 34 tests showed a significant association, including 4 amplifications and 30 deletions (FDR < 0.1, AUC > 0.5; **Figure 2a, Supplementary Table 2, Supplementary Figure 4**).

The cancer type with the largest number of significant focal copy number alterations was breast invasive carcinoma (2 amplifications and 9 deletions). Among these, we found the amplification of *ERBB2/HER2* (17q12, AUC=0.7, CI=[0.56–0.83], predictions based on histopathological subtypes yield an AUC=0.59, CI=[0.49, 0.68], **Figure 2d**), similar to previous reports[19], and deletion of a segment containing *PTEN* (10q23.31, AUC=0.69 CI=[0.56–0.81], predictions based on histopathological subtypes yield an AUC=0.56, CI=[0.46, 0.66]). Importantly, both alterations are prognostically and therapeutically relevant for treatment with targeted therapies such as Herceptin[20,21].

The deletion of a segment on cytoband 8p23.2 containing *CSMD1*, which has been previously reported as a candidate tumor suppressor gene that is frequently lost in many carcinomas[22], was identified in 6 cancer types such as ovarian serous carcinoma (AUC=0.62, CI=[0.53–0.71]), esophageal carcinoma (AUC=0.79, CI=[0.65–0.93]) and liver hepatocellular carcinoma (AUC=0.70, CI=[0.60–0.81]). A proximal deletion of *PPP2R2A*, located on 8p21.2, which was commonly deleted in cancers and was shown to be associated with a worse prognosis[23,24], was identified in 5 cancers including esophageal carcinoma (AUC=0.8, [0.65–0.94]), hepatocellular carcinoma (AUC=0.67, CI=[0.57–0.78]) and breast invasive carcinoma (AUC=0.67, CI=[0.59–0.75]). As these 2 deletions frequently co-occur (Cohen's $\kappa$=0.88), it is possible that these histological associations reflect the same underlying alteration.

We also found a strong association of focal amplification of the segment harboring *EGFR* (located at 7p11.2) in glioblastoma (AUC=0.78, CI=[0.7, 0.87]), characterised by a distinct small cell morphology of *EGFR*-amplified cancer cells (**Figure 2e**). This histopathological association has been noted previously[25], although *EGFR* amplifications do not exclusively define a molecular glioblastoma subtype[26].

### Driver gene mutations

Many cancer-causing mutations are point mutations in cancer driver genes. We annotated cancer-causing driver mutations in a catalogue of 104 recurrent cancer genes based on the non-silent to silent point mutation ratio of each gene and site-specific recurrence (see **Methods**). Histopathological associations were then tested using regularised logistic regression.

Among all driver genes tested, 14 (13%) genes had significant histopathology associations in at least one cancer type (AUC above 0.5, FDR<0.1, 19 out of 199 gene:cancer pairs, **Figure 2a, Supplementary Table 2, Supplementary Figure 4**). Interestingly, driver mutations in *TP53*, the most frequently mutated gene in cancers, could be predicted in 6 out of 27 (22%) cancer types, including breast invasive carcinoma (AUC=0.87, CI=[0.79, 0.95]), mesothelioma (AUC=0.87, CI=[0.67, 1]) and hepatocellular carcinoma (AUC=0.8, CI=[0.68, 0.92]). Samples with *TP53* mutation generally displayed a lower extent of differentiation, equivalent to a higher tumor grade (**Figure 2f**).

Another accurately predicted cancer driver gene was *BRAF* in thyroid tumors with an AUC as high as 0.9 (CI=[0.8, 1]), seemingly associated with a papillary morphology (**Figure 2g**). While it is known that the prevalence of *BRAF* mutations is only about 25% in follicular thyroid carcinomas as opposed to 75% recurrence in classical papillary and tall cell subtypes[27], these data indicate that the canonical histopathological classification may not fully account for the underlying genetics (AUC=0.81, CI=[0.69, 0.93]). A similar association was observed for *PTEN* in uterine cancers with an AUC of 0.8, CI=[0.73, 0.99] (**Figure 2h**), in part due to the enrichment of *PTEN* mutations in endometrial cancer[28]. When combined with histopathology subtypes the AUC for *PTEN* mutations in uterine cancer increases to 0.92 (CI=[0.85-1]). These findings demonstrate how genomically informed computational histopathology can augment and refine conventional histopathological subtypes, noting that many of the driver gene mutations have been previously reported to be associated with prognosis in cancer[27,29–31].

### Mutational Signatures

We also tested the associations with a set of 11 mutational signatures, reflecting different endogenous and exogenous genotoxic processes and also DNA repair deficiencies. Overall, there was only a relatively low extent of correlations, with the noteworthy exception of mismatch repair deficiency in colorectal cancer ($R^2$=0.19) as reported recently (**Supplementary Figure 6**)[32,33].

## Transcriptomic associations reveal immune infiltration and stromal cell types

Another molecular layer influencing cancer histology is the transcriptome. Gene expression changes may reflect not only distinct tumor cell types with different morphological properties – as in the previous example of *EGFR*-amplified glioblastoma – but also stromal and infiltrating immune cells, which will also have distinct histological appearance. We thus tested for transcriptome-wide associations of the set of 1,536 computational histopathological features using regularised

regression of each image tile on the logarithmic upper quartile normalised expression of each gene (log FPKM, see **Methods**) separately in each cancer type.

Overall, we found that 25% of all the gene:cancer pairs tested showed an association between bulk transcriptome and histology pattern (held back validation $R^2$>0, FDR<0.1, **Figure 3a, Supplementary Table 2**). For 2% of gene:cancer pairs an $R^2$>0.1 was found, and 0.56% displayed $R^2$>0.2. The cancer types with the largest number of associated genes are thymoma (*n*=8,377), sarcomas (*n*=8,359) and cutaneous melanoma (*n*=7,124). This reflects the notion that the histology generally reflects tumor composition and cell types, and that this relation can be quantified using computer vision and transcriptomics.

No obvious mechanistic insights were provided by the highest scoring gene expression associations, which include *NRF1* in thymoma ($R^2$=0.57, FDR<$10^{-16}$), *AOC3* in sarcoma ($R^2$=0.55, FDR<$10^{-16}$) and *PLAC9* in testicular germ cell tumors ($R^2$=0.45, FDR<$10^{-16}$) (**Supplementary Figure 4**). Several T-cell associated genes such as *LCK*, *CD8A*, *CD247* and *CD4* showed strong association with histopathology in thymomas, which is a T-cell rich cancer as the thymus is the organ where T-cells mature. Similarly, neurotransmitter genes that are mostly expressed in neurons such as *SLC6A12*, *SLC1A2* and *EPB41L1* showed strong association in low grade glioma, indicating that neuronal cell types (which are distinct from the glial cells that give rise to gliomas) contribute to both bulk transcriptome and histology.

More obvious trends emerge at the levels of gene sets, summarising gene expression states and molecular pathways. Gene set enrichment analysis showed that the genes that are associated with histopathological features are enriched (FDR<0.1) in 193 pathways (**Figure 3b, Supplementary Table 2**) in at least one cancer type. Immune system related pathways are found in 17 cancer types (*n*=81 pathways), followed by pathways involved in cell cycle and metabolism (46 and 20 pathways respectively).

We also investigated the association of histopathological features with expression based proliferation and tumor infiltrating lymphocytes (TILs) scores[34]. 14 cancer types showed a detectable association for proliferation score ($R^2$>0, FDR<0.1, **Supplementary Table2**), including thymoma ($R^2$=0.39), prostate adenocarcinoma ($R^2$=0.24) and hepatocellular carcinoma ($R^2$=0.21). Tissues predicted as high proliferation usually had a low stromal content and were high grade (**Supplementary Figure 7a**). Often the classification of low proliferation overlaps with predicted normal tiles across the whole tissue slide. This indicates that a high stromal and normal content often coincides with low transcriptomic proliferation scores and highlights the algorithm's ability to attribute this association to the presence of histologically normal tissue areas in the tumour section (**Supplementary Figure 7b**).

## Automated spatial localisation of immune cells

As discussed above, deep learning of histological patterns across slides provides spatial resolution of known tissue types. This property also extends to molecular associations, exemplified by the association with inflammatory cell infiltrates, which appeared to be an overarching feature across cancer types. These phenomena can be illustrated using a transcriptomic signature of TILs. The score of TILs showed a significant $R^2$ with histology for 13 cancer types ($R^2$>0, FDR<0.1), ranging from $R^2$=0.008 for uterine cancer to $R^2$=0.2 for thymoma (**Supplementary Table2**). Tiles predictive of TILs indeed contain lymphocytes that are typically relatively small cells with dark nuclei and scant cytoplasm, often occurring at high densities (**Figure 3c**). Considering that the algorithm was never provided with defined image sections representing TIL-rich areas, it is a remarkable property that these emerge as the common denominator of tiles from slides with a high molecular signal of TILs.

Two interesting observations were made in the analysis of spatial distribution of these lymphocyte signals. While the prediction of the pattern of lymphocytic infiltration was in some cases relatively uniform, with a dispersed distribution of TILs (**Figure 3d,** top and middle panels), in other cases, the signals frequently localised to confined regions containing lymphocytic aggregates (**Figure 3d,** bottom panel). While the latter is another striking example of the ability of computer vision to identify and localise patterns based on sufficiently high levels of recurrence across the training image, the former presents an equally hard task, as dispersed lymphocytes appears far less obvious and might be easily missed upon visual histological assessment.

An interesting example that demonstrates the benefit of spatial prediction of TILs over bulk molecular measurements is shown in the bottom row of **Figure 3d** (sample from breast invasive carcinoma). The transcriptomic signal indicated a high level of TILs for this sample (in the top 10% of all patients) while the histopathological features predicted heterogeneous TILs scores across the tumor slide identifying regions with high and low TIL densities. Taken together, an integrated analysis of transcriptomics and computational histopathology has the potential to precisely define and quantify patterns of TILs, which may also explain differential response to treatments[35].

## Prognostic effects across cancer types

Finally, we built predictive models for patients' overall survival (OS)[22] using computational histopathological features in 18 different cancers with available outcome data. We calculated the prognostic signal provided by PC-CHiP using regularised Cox proportional hazards models and evaluated its predictive accuracy in isolation as well as in relation to conventional histopathology (grade and subtypes), further clinical data (age, gender and cancer stage), and also transcriptomic data[36]. An overview of sample size, number of events and the number of selected parameters can be seen in **Supplementary Table 3**.

Overall, we observed a significant association of computational histopathological features with OS in 15 out of 18 cancer types (FWER<0.05, mean concordance *C*=0.6) with a concordance ranging from *C*=0.53 for glioblastoma multiforme to *C*=0.67 for uterine endometrial carcinoma (**Figure 4a, Supplementary Table 3**). More importantly, compared to canonical histological subtypes and grades, which are routinely used to assess prognosis, the histopathological features showed a significant improvement in 10/16 cancer types. This prognostic signal remained measurable in the majority of these cancer types, even when further including age, gender and tumour stage (**Figure 4b, Supplementary Figure 8, Supplementary Table 3**). As illustrated by the survival curves, PC-CHiP may be used to refine existing stage-based prognosis in breast, head and neck, and stomach cancer and, to a lesser extent, clear cell renal cell carcinoma, identifying individuals with approximately two-fold better or worse prognosis than the average in these groups.

Reassuringly, many of the prognostic histopathological patterns automatically learned by computer vision reflect some well-recognised associations. For example, necrosis[37] and high grade are associated with poor prognosis across tumor types, while higher degree of differentiation[38], as well as the presence of TILs are usually associated with a favourable risk[39] (**Figure 4c**). In some cases, risk predictions reflect different histological subtypes, in addition to the above patterns. For instance, among gliomas low grade oligodendrogliomas generally have better prognosis than astrocytomas[40] (**Figure 4c**).

Often favourable and unfavourable patterns can be identified on the same slide, highlighting the ability of computer vision to deconvolve the content of large tissue sections into molecularly and prognostically distinct areas (**Figure 4d**) with necrosis and lymphocytic aggregates detected on the same specimen. Similarly, areas of low and high grade tumour differentiation identified on the same slide produced favourable and unfavourable risk predictions. Currently, it is unclear how such conflicting information is best combined into a single prediction. However, the ability to detect these patterns – and the capability to learn a prognostic stratification that matches conventional staging and grading in a fully automated way – are important first steps towards developing novel and refined patient-specific risk models.

## Generalisation requires tailored architectures and data augmentation

As outlined above, deep learning algorithms are capable of extracting a rich feature representation of images, which is of great value beyond the primary learning task of tissue classification. However, doing so with millions of parameters may also come at the price of overfitting to the training data[11,41]. In this context, overfitting may unearth undesirable features introduced by specimen acquisition, sample preparation, sectioning and staining, but also microscope settings as well as digital differences resulting from lossy file formats and compression parameters, the entirety of which can be hard to control for. As an example, we observed a strong influence of the jpeg image quality settings on the computational histopathological features. The jpeg quality was highly unbalanced between the 42 tissue labels used for the training of the CNN – and these differences

were inadvertently learned by the algorithm despite default data augmentation (**Supplementary Figure 9a**).

Without further amendments the trained algorithm therefore generalised poorly on 471 breast cancer H&E stained slides from the METABRIC consortium[42] (scanned on an earlier version of the Aperio platform and saved into a jpeg2000 file stream; **Supplementary Figure 9b**). Inspection of the feature set revealed that the implicit feature representation of the new images was measurably offset ($d$=18.7, comparable to the average distance of different tumors and the distance between normal and tumor breast).

To overcome the effect of jpeg format confounding, we modified the Inception-V4 architecture by injecting the image format directly into the final classification layer such that the network does not learn patterns associated with the file format and compression from the image as these are explicitly available as auxiliary variables. This modification, together with additional data augmentation to supersede any pre-existing jpeg patterns and heavy color augmentation to overcome systematic differences in H&E staining (random hue rotations by –90 to 90 degrees),[43] led to a decoupling of file format and tissue labels (**Supplementary Figure 9c-d**). This was accompanied by a slight drop in the tissue classification accuracy (average AUC=0.95), indicating that technical artefacts contributed in part to the initial tissue classification accuracy. Reassuringly, however, the modifications to the network and training procedure did not affect the strength of molecular associations, indicating that the strategy to evaluate performance only *within* samples of the same training labels helped to avoid confounding influences learned as part of the initial training step (**Supplementary Figure 10**).

More importantly, the network trained in this way also showed better generalisation to data from an entirely new cohort. On the METABRIC cohort, genomic alterations including *TP53* mutation and WGD as well as OS which were found to be associated with computational histopathological features in TCGA breast cancer samples, could be predicted with only a moderate drop in accuracy (**Figure 5a**). Similarly, the new histopathological features were able to predict the tumor infiltrating lymphocytes. In a set of 36 slides, for which independent pathologist evaluated TIL scores were available, the average predicted TILs for samples with no lymphocytes were smaller than those with mild or severe level of lymphocytes ($p$-value = 0.001 and 0.095 respectively, **Figure 5b**). Reassuringly, the algorithm's ability to localise TILs on a given slide was preserved (**Figure 5c**). Lastly, the prognostic associations could also be validated and stratify patients' OS across tumor stage as expected (**Figure 5d**). As in the TCGA training data, the algorithm identified necrotic areas as well as TILs on a given slide as unfavourable and favourable prognostic markers (**Figure 5e**).

These observations indicate that care must be taken when using CNNs in biomedical applications[11]. In addition to using a wide range of data sources, appropriate training procedures and architectures and substantial data augmentation techniques to avoid overfitting, one has to carefully assess whether there are any confounding factors influencing training labels and tailor the algorithm and training procedures such that these are isolated from the images. As a quality control step, it may

help to investigate whether the feature representation of novel images lies within the expected range of comparable images of the same tissue.

## Discussion

Here we presented PC-CHiP, a pan-cancer transfer learning approach to extract computational histopathological features across 42 cancer and normal tissue types and their genomic, molecular and prognostic associations. Histopathological features, originally derived to classify different tissues, contained rich histologic and morphological signals predictive of a range of genomic and transcriptomic changes as well as survival. This shows that computer vision not only has the capacity to highly accurately reproduce predefined tissue labels, but also that this quantifies diverse histological patterns, which are predictive of a broad range of genomic and molecular traits, which were not part of the original training task. As the predictions are exclusively based on standard H&E-stained tissue sections, our analysis highlights the high potential of computational histopathology to digitally augment existing histopathological workflows.

The strongest genomic associations were found for whole genome duplications, which can in part be explained by nuclear enlargement and increased nuclear intensities, but seemingly also stems from tumour grade and other histomorphological patterns contained in the high-dimensional computational histopathological features. Further, we observed associations with a range of chromosomal gains and losses, focal deletions and amplifications as well as driver gene mutations across a number of cancer types. These data demonstrate that genomic alterations change the morphology of cancer cells, as in the case of WGD, but possibly also that certain aberrations preferentially occur in distinct cell types, reflected by the tumor histology. Whatever is the cause or consequence in this equation, these associations lay out a route towards genomically defined histopathology subtypes, which will enhance and refine conventional assessment.

Further, a broad range of transcriptomic correlations was observed reflecting both immune cell infiltration and cell proliferation that leads to higher tumor densities. These examples illustrated the remarkable property that machine learning does not only establish novel molecular associations from pre-computed histopathological feature sets but also allows the localisation of these traits within a larger image. While this exemplifies the power of a large scale data analysis to detect and localise recurrent patterns, it is probably not superior to spatially annotated training data. Yet such data can, by definition, only be generated for associations which are known beforehand. This appears straightforward, albeit laborious, for existing histopathology classifications, but more challenging for molecular readouts. Yet novel spatial transcriptomic[44,45] and sequencing technologies[46] bring within reach spatially matched molecular and histopathological data, which would serve as a gold standard in combining imaging and molecular patterns.

Across cancer types, computational histopathological features showed a good level of prognostic relevance, substantially improving prognostic accuracy over conventional grading and histopathological subtyping in the majority of cancers. It is this very remarkable that such predictive

signals can be learned in a fully automated fashion. Still, at least at the current resolution, the improvement over a full molecular and clinical workup was relatively small. This might be a consequence of the far-ranging relations between histopathology and molecular phenotypes described here, implying that histopathology is a reflection of the underlying molecular alterations rather than an independent trait. Yet it probably also highlights the challenges of unambiguously quantifying histopathological signals in – and combining signals from – individual areas, which requires very large training datasets for each tumour entity.

From a methodological point of view, the prediction of molecular traits can clearly be improved. In this analysis, we adopted – for the reason of simplicity and to avoid overfitting – a transfer learning approach in which an existing deep convolutional neural network, developed for classification of everyday objects, was fine tuned to predict cancer and normal tissue types. The implicit imaging feature representation was then used to predict molecular traits and outcomes. Instead of employing this two-step procedure, which risks missing patterns irrelevant for the initial classification task, one might directly employ either training on the molecular trait of interest, or ideally multi-objective learning. Further improvement may also be related to the choice of the CNN architecture. Everyday images have no defined scale due to a variable *z*-dimension; therefore, the algorithms need to be able to detect the same object at different sizes. This clearly is not the case for histopathology slides, in which one pixel corresponds to a defined physical size at a given magnification. Therefore, possibly less complex CNN architectures may be sufficient for quantitative histopathology analyses, and also show better generalisation.

Here, in our proof-of-concept analysis, we observed a considerable dependence of the feature representation on known and possibly unknown properties of our training data, including the image compression algorithm and its parameters. Some of these issues could be overcome by amending and retraining the network to isolate the effect of confounding factors and additional data augmentation. Still, given the flexibility of deep learning algorithms and the associated risk of overfitting, one should generally be cautious about the generalisation properties and critically assess whether a new image is appropriately represented.

Looking forward, our analyses revealed the enormous potential of using computer vision alongside molecular profiling. While the eye of a trained human may still constitute the gold standard for recognising clinically relevant histopathological patterns, computers have the capacity to augment this process by sifting through millions of images to retrieve similar patterns and establish associations with known and novel traits. As our analysis showed this helps to detect histopathology patterns associated with a range of genomic alterations, transcriptional signatures and prognosis – and highlight areas indicative of these traits on each given slide. It is therefore not too difficult to foresee how this may be utilised in a computationally augmented histopathology workflow enabling more precise and faster diagnosis and prognosis. Further, the ability to quantify a rich set of histopathology patterns lays out a path to define integrated histopathology and molecular cancer subtypes, as recently demonstrated for colorectal cancers[47]. Lastly, our analyses provide

proof-of-concept for these principles and we expect them to be greatly refined in the future based on larger training corpora and further algorithmic refinements.

## Acknowledgements

AWJ and MG are supported by grant NNF17OC0027594 from the Novo Nordisk Foundation. LM is a recipient of a CRUK Clinical PhD fellowship (C20/A20917). We thank all members of the Gerstung Lab, Inigo Martincorena and Andrew Lawson for critical comments on the manuscript. The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

## Author Contributions

YF retrieved and quality controlled all images, developed and trained the deep learning algorithms, performed statistical tests for genomic and molecular association and created all figures. AWJ performed survival analysis. RVT and MG extended the Inception-V4 algorithm. SG provided copy number and annotated mutation data. HV extracted mutational signature data. LM and MJL assessed histopathology images. MG conceived and supervised the study. YF, AWJ and MG wrote the manuscript with input from all authors who approved the manuscript.

# Methods

## Images

We collected 17,396 H&E stained histopathology slides of 10,452 patients of 28 cancer types from TCGA via the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/), including normal, tumor and metastatic tissue types. Only tissue types with at least 50 images with a magnification greater than 20X are included. We first cropped the whole slides into 512 by 512 pixels tiles with 50 pixels overlap at 20X magnification. We then removed blurred and non-informative tiles by filtering on the weighted gradient magnitude (using Sobel operator, tiles with weighted gradient magnitude smaller than 15 for more than half of the pixels were removed). Tiles from tumor samples with tumor purity greater or equals to 85% were used in training/validation to avoid miss labelled tiles in the training process. To avoid bias that are caused by image preparation in each laboratory, we randomly selected 80% images from each centre for training. In total, we had 6,564,045 tiles from 8,067 slides for training, 1,357,892 tiles from 1,687 slides for validation and 6,641,462 tiles from 7,672 slides for testing.

## Pan-Cancer Computational Histopathology PC-CHiP

A pretrained Inception-V4[16], a deep convolutional neural network, was used to classify tiles into 42 classes and to extract histopathological features from each tile. We applied sample specific label smoothing, an adapted version of the label smoothing method first introduced in Inception-V3[48] for model regularization, to avoid overfitting. In short, for a sample of tissue $i$, we set ground-truth distribution $q(k)$ to $q(k) = p_T$ for $k=i$ and $q(k) = (1 - p_T) / (N - 1)$ for all $k \neq i$, where $N=42$ is the total number of classes and $p_T$ is the tumor purity of the sample. The model was trained in Tensorflow using Slim[49] with the default hyperparameters for 100K steps (~1 epoch). The scripts used for training and the retrained model checkpoint can be found on our Github repository.

We retrieved the predictive probability for all 42 classes for each tile and the associated 1,536 histopathological features from the last hidden layer of the trained Inception-V4. As in practice the cancer origin is usually known, we also computed tumor/normal classification within cancer types for each tile by comparing only the probability of being normal or tumor of that cancer type.

To visualise the tiles represented by the 1,536 histopathological features, we applied Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP[50]) for dimension reduction for a subset of tiles (50 tiles were randomly selected from high tumor purity images). As the distances between data points in the original dimension are not preserved in the low dimension generated by UMAP, we also calculated the mean pairwise Euclidean distance between tissue types in their original dimension.

## Molecular associations

In this section, we performed statistical models to predict genomic alterations for each image tile using 1,536 histopathological features and the tissue type. Per slide prediction was then calculated by averaging the prediction of all tiles within that slide. All models were performed in R using the "glmnet" package[51]. To avoid normal contamination, only samples with tumor purity greater than or equal to 85% are included. We randomly select 100 tiles from images of 70% of the patients for training and 30% of the patients for independent validation. 5-fold cross-validation was used to select the best model. To avoid reporting the different prevalence of molecular traits between cancer types, predictive accuracy was calculated within each cancer type. All accuracy values reported in the manuscript are based on samples from a held back subset of patients.

### Genomic alterations

Point mutations (single nucleotide variants and short deletions and insertions) were called using CaVEMan and pindel algorithms plus a set of dedicated post-processing filters as described previously[52] for 8,769 TCGA patients. Absolute copy number was called using the ASCAT algorithm[53]. Whole genome doubling (WGD) status were called using the criteria described previously[54] . Chromosome and chromosome arm level gains and losses were retrieved from[17]. Focal amplifications and deletions were defined for regions retrieved from[18]. For each of the amplified regions, samples with an absolute copy number of at least 10 were called amplified; for each of the deleted regions, only samples with $\leq 1$ copy in the absence of WGD and samples with $\leq 2$ copies in the presence of WGD were called deleted. We performed multinomial logistic regression models to classify gain, non-altered and loss of 56 chromosome or chromosome arm. We applied generalised logistic regression with LASSO penalization for each genomic alteration. Per alteration AUC was then calculated in a one vs the rest fashion (ex. gain vs. not altered and loss) for each cancer type. Confidence interval at 95% and associated *p*-value was obtained for each AUC using a Wilcoxon tests.

### Gene expression

Log transformed upper quantile normalized gene expression from RNA sequencing data was used. We performed generalised linear regressions with LASSO penalization on 14,214 genes that are expressed in at least 60% of the samples. Variance explained ($R^2$) was calculated for each gene-cancer pair to evaluate the model performance. A confidence interval at 95% and associated *p*-value were estimated by a *F*-test. The *p*-values were then corrected controlling the FDR. As shown previously in this paper, the histopathology features are highly capable of distinguishing normal and tumor tissues, therefore an estimation of gene expression level could be inferred from the tumor percentage of images from cancer types with differential expression between tumor and normal samples. To make sure that the prediction is relevant to the gene expression level rather than the tumor/normal content, we calculated the expected expression based on tumor purity $p_T$ and average gene expression in tumour and normal samples alone, $X_{0,\text{cancertype}} = (p_T \times \text{mean}(X_{\text{tumor,cancertype}})$

+ $(1 - p_T) \times \text{mean}(X_{\text{normal,cancertype}})$), where $X_{\text{tumor,cancertype}}$ denotes the average expression of a given transcript in all tumor samples and $X_{\text{normal,cancertype}}$ denotes the average expression of all normal samples for the corresponding cancer type. We then filtered the tests where histopathology features were less informative than $X_{0,\text{cancertype}}$. In order to identify functional classes of genes that can be predicted by histopathology features, we then performed gene set enrichment analysis (GSEA)[55] for a collection of REACTOME pathways[56]. A normalised enrichment score and *p*-value were calculated for each pathway in each cancer type. The *p*-values were corrected to control the FDR. At last, we performed regression on gene expression based proliferation score and tumor infiltrating lymphocytes signature[34] using the same method used for single gene expression.

## Prognostic associations

Survival analysis was performed using penalized Cox's proportional hazard regression[57] using a mixture of $L_1$ and $L_2$ regularization, often referred to as Cox elastic net[58]. To evaluate discriminative performance we used Harrell's *C*-index as a measure of the concordance between predicted and actual risk[59].

In order to obtain a scalable and sparse solution, we deployed proximal gradient descent for our parameter updates[60]. Due to the large-scale nature of the problem, an exhaustive hyperparameter search was infeasible. Therefore, hyperparameters, particular $L_1/L_2$ penalization strength, have been automatically determined using Bayesian optimization[61]. Twenty repetitions of 5-fold cross-validation were used to evaluate model performance. Each fold was further split into a training set (85%) and a validation set (15%).

A total of 6 models, combining different combinations of variables each, were evaluated for each relevant cancer from TCGA. A cancer was included in the analysis if the sample size was at least 160 individuals and censorship was less than 90%. The first model ("histology") contains the histological subtype and the corresponding grading information. Routine clinical information on the individual like age at cancer diagnosis, gender, cancer staging and the histopathology features form the second model ("clinical"). The third model ("clinical + expression") is a combination of clinical and gene expression data. Model four ("PC-CHiP") uses the extracted histopathology features from the CNN. Model five ("clinical + PC-CHiP") contains the histopathology features and the clinical data. Lastly, Model six ("all") is a set of all covariates. If observations have been missing, particularly for the gene expression data, mean imputation has been applied. The gene expression data comprises the first 30 components of a principal component analysis (PCA). For the survival analysis with the histopathology features, each extracted tile has been used as an individual observation. A global risk estimate is obtained using the average risk across the tiles from a patient. Three different strategies were employed to assess the value of adding PC-CHiP to models based on conventional variables. First, it was tested whether the cross-validated linear predictor obtained using PC-CHiP alone added significant signal in a multivariate model. Second, it was assessed whether the pretrained predictor based on PC-CHiP improved the concordance *C* in a cross-validation setting. Third, a likelihood boosting approach was used for training Cox models from scratch combining clinical/gene expression data with the histopathology features.

To compare predictive performance across models we examined the distribution of concordance indices across folds as well as the mean difference concordance within folds. Furthermore we used a paired-Wilcoxon sign rank test to compare *C* estimates across models.
To account for multiple comparisons we used the Bonferroni correction as FWER procedure. Survival curves have been estimated using the Kaplan-Meier estimator.

## External validation using the METABRIC dataset

H&E stained slides were retrieved from the European Genome-Phenome archive (https://ega-archive.org/dacs/EGAC00001000484). Whole slides were tiled into 512 by 512 pixel tiles in the same fashion as in TCGA. Genomic data including mutations and copy number variations as well as clinical data were downloaded from[42]. WGD status was calculated using the methods described previously[62]. The amended Inception-V4 architecture, preprocessing scripts and the retrained model checkpoint can be found on our Github repository. As the tumor purity of the METABRIC samples vary (cellularity range from low (<40%), moderate (40%-70%) to high (>70%)) and can be lower than those of TCGA samples, the per image prediction for genomic alterations and TILs score was calculated by averaging over tiles that were correctly predicted as breast tumor in order to avoid normal contamination.

# Figures

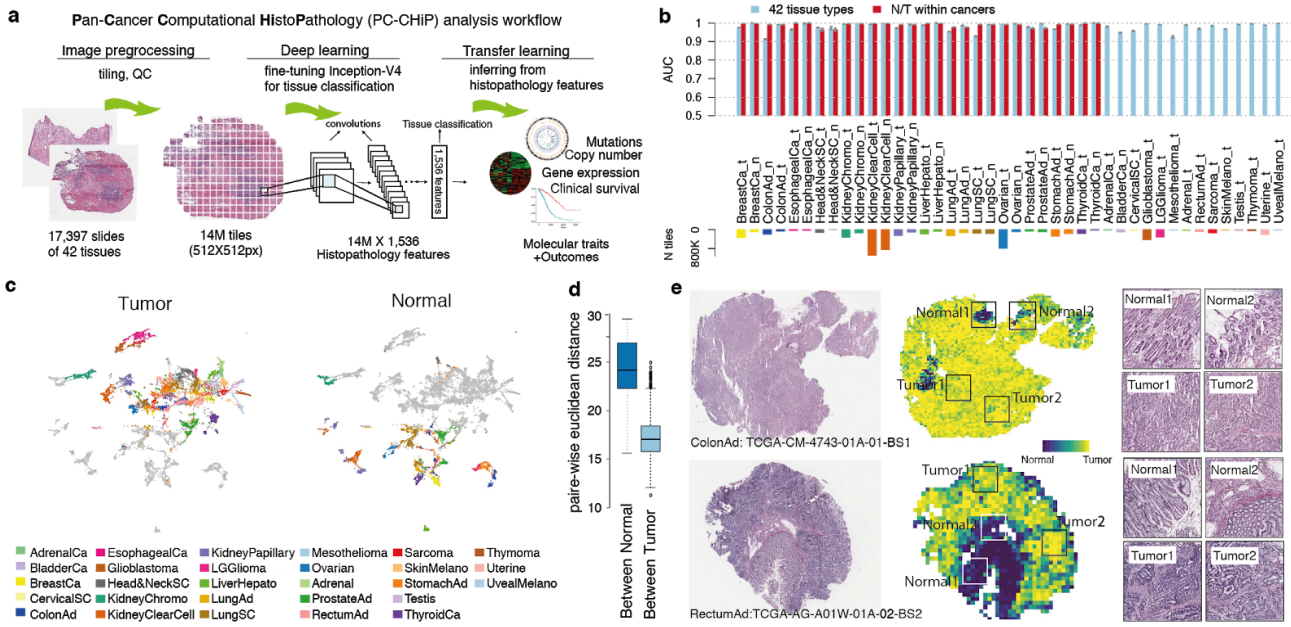



**Figure 1. Pan-cancer computational histopathology (PC-CHiP) quantifies tissue-specific morphology. a.** The image analysis workflow. **b.** High overall classification accuracy of 42 normal/tumor tissue types. The upper panel shows the classification accuracy for each tissue type in the validation set (overall classification accuracy for 42 tissues in blue and cancer specific tumor/normal classification accuracy in red). The 95% confidence interval of the AUC was computed using bootstrap. The labels in the middle correspond to tumor (_t) or matching normal (_n) tissues across cancer types. The lower panel shows the number of tiles in the training dataset for each tissue type. **c.** UMAP dimensionality reduction representation of the 1,536 histopathological features of 42 tissue types. Tumors are shown on the left and normal tissues on the right. Each dot corresponds to a randomly selected tile colored by its tissue type. **d.** Pairwise Euclidean distance in the 1,536-feature space are shown for normal tissues (left) and tumor tissues (right). **e.** Example images with accurate spatial prediction of tumor/normal tissue. For each row, the original haematoxylin and eosin (H&E) stained slide is shown on the right side, predicted tumor probability of each tile for the whole slide is shown in the middle figure, zoom-in images of 4 sub regions indicated in the middle figure are shown on the right side. See Supplementary Table 2 for cancer type abbreviations.

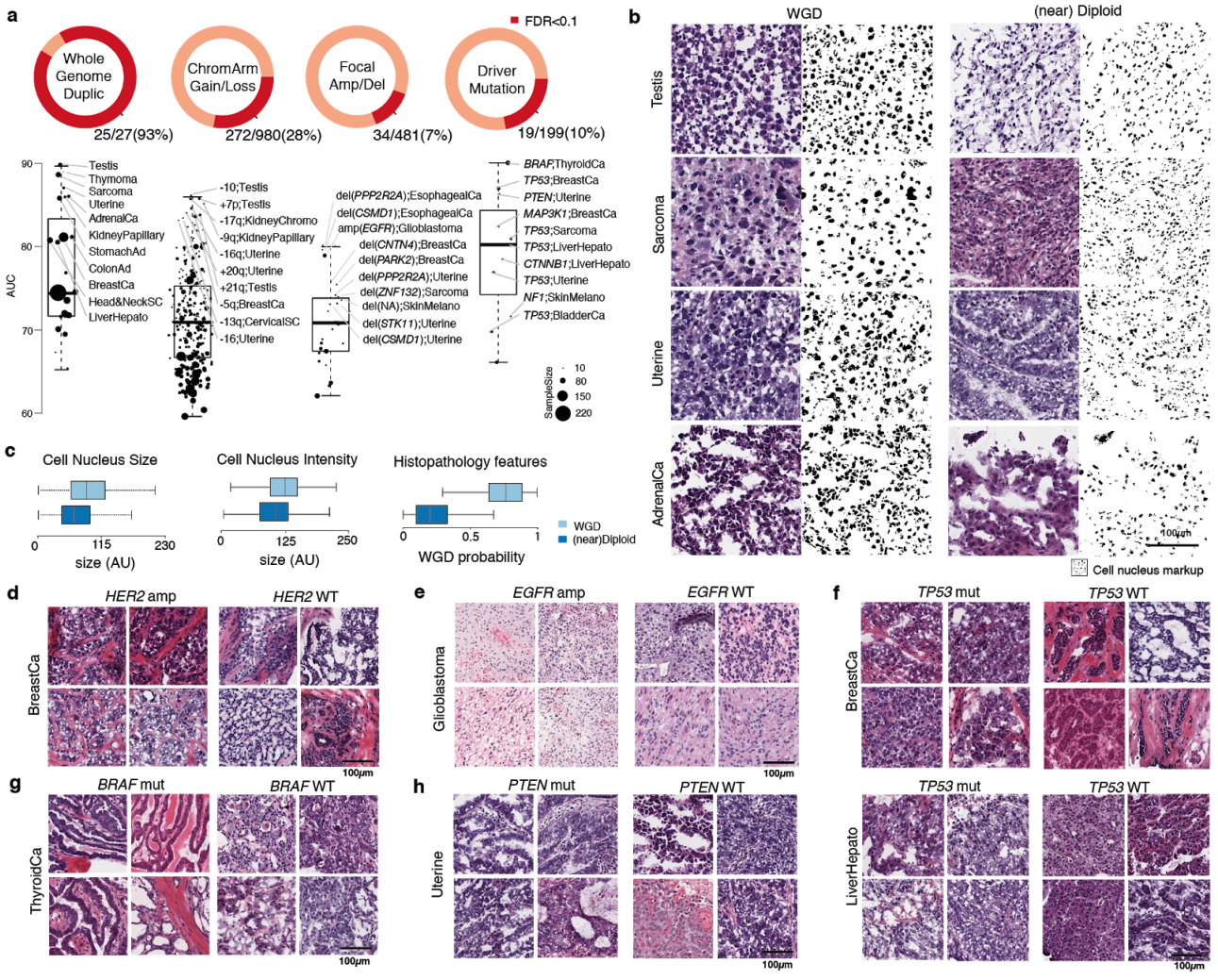



**Figure 2. Wide-spread associations between PC-CHiP and genomic alterations. a.** Overall performance for molecular associations with histopathological features. In the upper panel, each pie chart corresponds to one alteration type. In the lower panel, each box corresponds to the distribution of significant associations (FDR<0.1) for each alteration type that is indicated in the upper panel. **b.** Example tiles with and without whole genome duplication (WGD). From top to bottom, each row corresponds to one cancer type indicated on the left side. For each row, haematoxylin and eosin (H&E) stained slide and its cell nucleus mark-up are shown for a WGD sample (left side) and a near diploid sample (right side). **c.** Comparison of traditional hard-coded morphological features and histopathological features between WGD sample and near diploid samples. Boxplots show the cell nucleus intensity (upper panel), cell nucleus size (middle panel) and a linear combination of histopathological features (lower panel) between a set of randomly selected tiles with and without WGD. **d.** Example tiles of breast invasive carcinoma with *HER2*-amplified (4 tiles on the left side) and *HER2*-wild type (4 tiles on the left side). **e.** Example tiles of glioblastoma with *EGFR* amplification (4 tiles on the left side) and *EGFR* wild type (4 tiles on the left side). **f.** Example tiles of breast invasive carcinoma (top panel) and hepatocellular carcinoma (bottom panel) with *TP53* mutation (4 tiles on the left side) and *TP53* wild type (4 tiles on the left side). **g.** Example tiles of thyroid carcinoma with *BRAF* mutation (4 tiles on the left side) and *BRAF* wild type (4 tiles on the

left side). **h.** Example tiles of uterine cancers with *PTEN* mutation (4 tiles on the left side) and *PTEN* wild type (4 tiles on the left side). See Supplementary Table 2 for cancer type abbreviations.

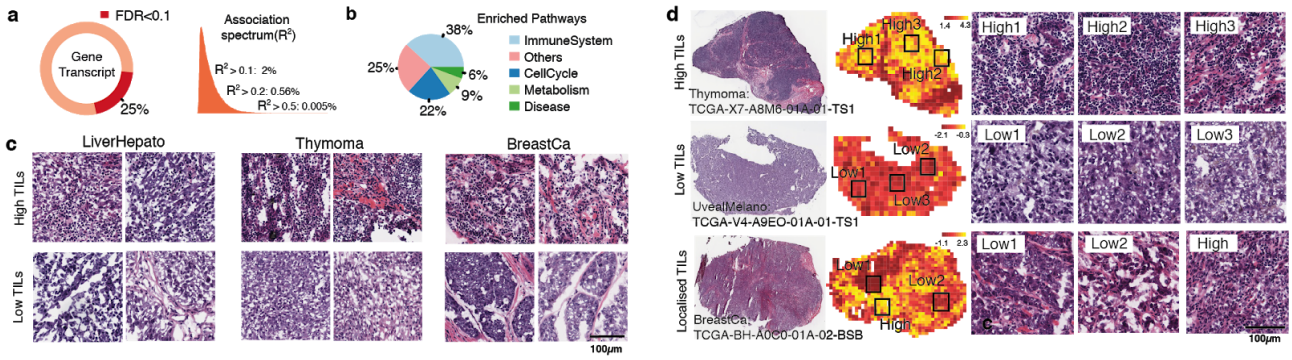



**Figure 3. Transcriptomic associations reveal immune infiltration and stromal cell types. a.** Overall performance for transcriptomic associations with histopathological features. **b.** Gene ontology analysis for significant genes. **c.** Example tiles of high (first row) and low (second) TILs from hepatocellular carcinoma (left panel), breast invasive carcinoma (middle panel) and thymoma (right panel). Two randomly selected tiles are shown for each condition. **d.** Examples slides with high (top row), low (middle row) and localised (bottom row) tumor infiltrating lymphocytes. From left to right are the original haematoxylin and eosin (H&E) stained slides, predicted spatial TILs score and example tiles with high and low TILs. See Supplementary Table 2 for cancer type abbreviations.

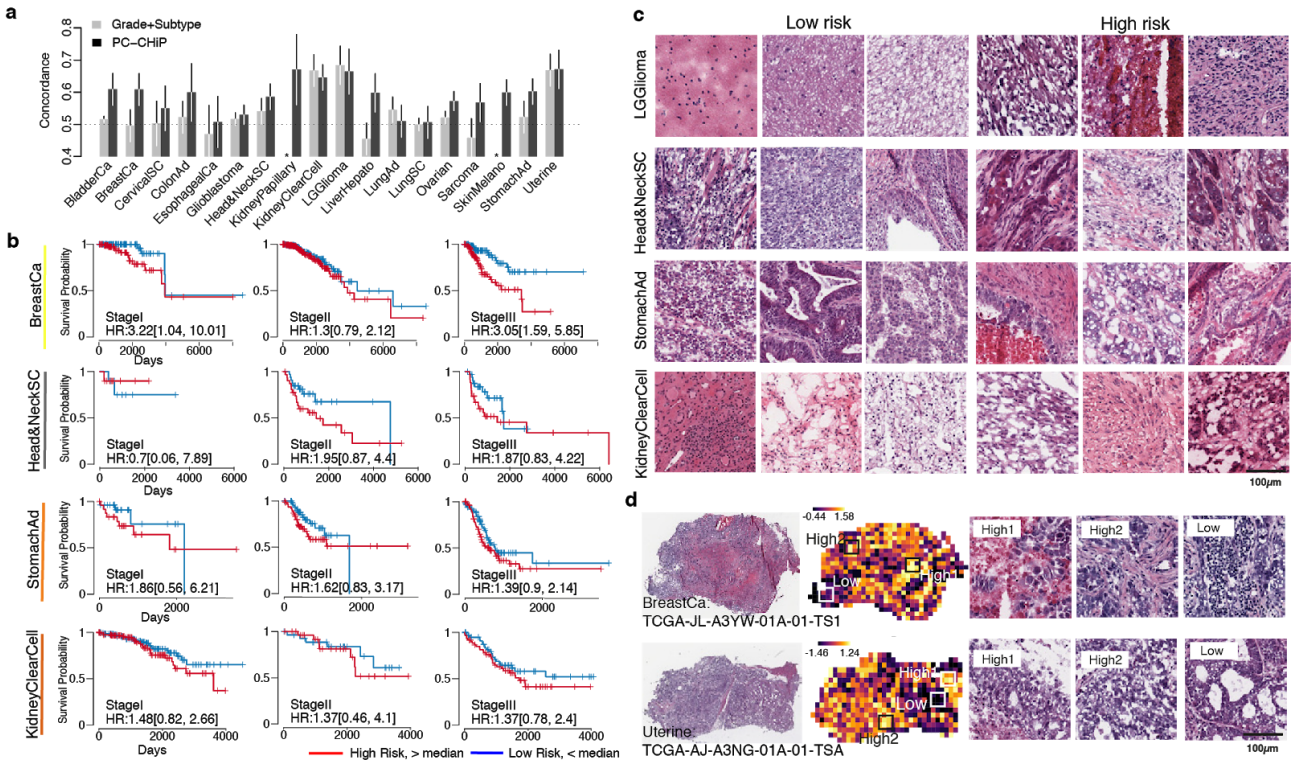



**Figure 4. PC-CHiP provides complementary prognostic information. a.** Predictive accuracy of overall survival using histopathological grade and subtypes (light gray) compared to PC-CHiP (dark gray) in 18 cancers. Each bar corresponds to the mean concordance (with its 95% confidence interval) in 5-fold cross-validation for the dataset and the corresponding cancer type (indicated at the bottom). **b.** Kaplan-Meier plots with high (above median) and low (below median) PC-CHiP risk shown for cancer stage I-III in four different cancer types. **c.** Example tiles with high and low estimated risk based on the histopathological features. Each row corresponds to a cancer type. The first three columns are tiles with low risk and the last three are tiles with high risk. **d.** Spatial risk predictions across tissue slide for breast invasive carcinoma and uterine endometrial carcinoma. From left to right are the H&E stained slides, the corresponding risk prediction and a selection of enlarged tiles with high/low predicted risk. See **Supplementary Table 2** for cancer type abbreviations.

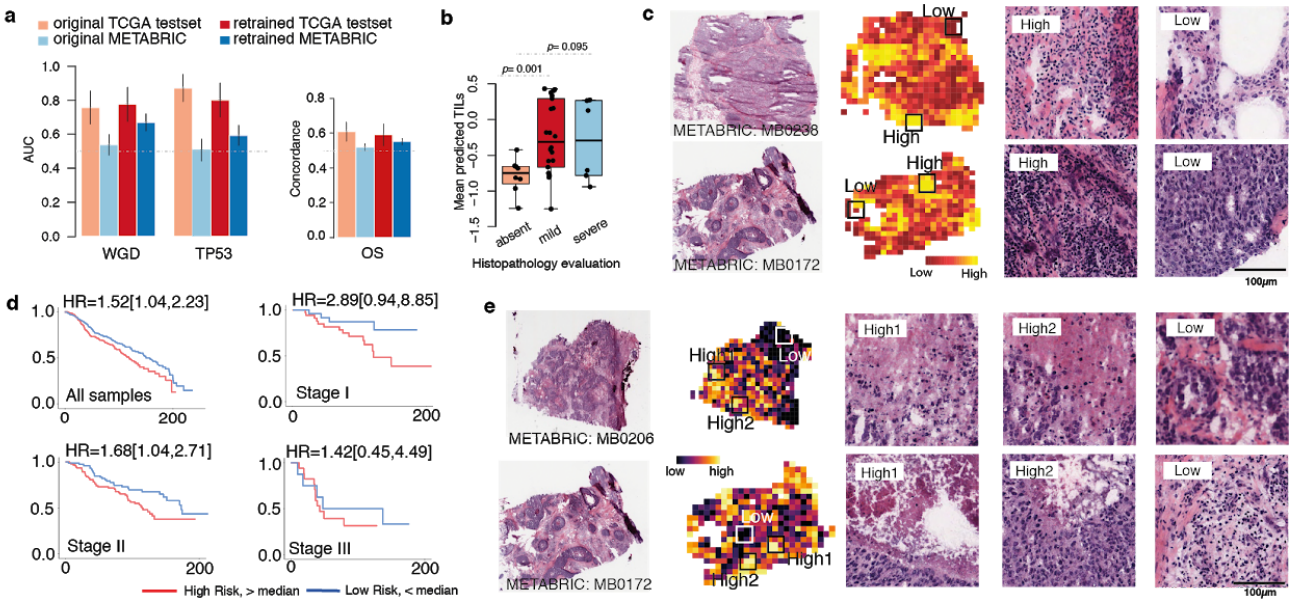



**Figure 5. External validation on the METABRIC breast cancer cohort.** **a.** The predictive accuracy of WGD, *TP53* mutation and patients' overall survival yield comparable results in the METABRIC dataset (dark blue) compared to TCGA validation (dark red) after retraining of the CNN. **b.** The mean predicted TILs using histopathological features are significantly higher for samples evaluated as with mild and severe of lymphocytes than those with no lymphocytes. **c.** Example slides with spatial prediction of TILs. From left to right are original H&E stained slide, spatial prediction of TILs based on histopathological features, tile predicted as high infiltration and low infiltration. **d.** Kaplan Meier plots of a better patient stratification using histopathological features for different stages in METABRIC. **e.** Example slides with spatial risk prediction. From left to right are the H&E stained slide, the corresponding spatial risk prediction and a selection of enlarged tiles with estimated high or low risk. See Supplementary Table 2 for cancer type abbreviations.

# References

1. Lindeman, N. I. *et al.* Molecular testing guideline for selection of lung cancer patients for EGFR and ALK tyrosine kinase inhibitors: guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology. *J. Thorac. Oncol.* **8**, 823–859 (2013).
2. Woodman, S. E., Lazar, A. J., Aldape, K. D. & Davies, M. A. New strategies in melanoma: molecular testing in advanced disease. *Clin. Cancer Res.* **18**, 1195–1200 (2012).
3. Russnes, H. G., Lingjærde, O. C., Børresen-Dale, A.-L. & Caldas, C. Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters. *Am. J. Pathol.* **187**, 2152–2162 (2017).
4. Dienstmann, R. *et al.* Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nat. Rev. Cancer* **17**, 268 (2017).
5. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011–1025 (2015).
6. Bailey, P. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**, 47–52 (2016).
7. Elmore, J. G. *et al.* Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA* **313**, 1122–1132 (2015).
8. Sackett, M. K., Salomão, D. R., Donovan, J. L., Yi, E. S. & Aubry, M. C. Diagnostic concordance of histologic lung cancer type between bronchial biopsy and cytology specimens taken during the same bronchoscopic procedure. *Arch. Pathol. Lab. Med.* **134**, 1504–1512 (2010).
9. Esteva, A. *et al.* Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **546**, 686 (2017).
10. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
11. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on

whole slide images. *Nat. Med.* (2019). doi:10.1038/s41591-019-0508-1

12. Hegde, N. *et al.* Similar image search for histopathology: SMILY. *NPJ Digit Med* **2**, 56 (2019).
13. Saltz, J. *et al.* Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Rep.* **23**, 181–193.e7 (2018).
14. Schaumberg, A. J., Rubin, M. A. & Fuchs, T. J. H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer. *BioRxiv* (2018).
15. Mobadersany, P. *et al.* Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2970–E2979 (2018).
16. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. in *AAAI* **4**, 12 (2017).
17. Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**, 676–689.e3 (2018).
18. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
19. Yuan, Y. *et al.* Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci. Transl. Med.* **4**, 157ra143 (2012).
20. Tan, M. & Yu, D. Molecular mechanisms of erbB2-mediated breast cancer chemoresistance. *Adv. Exp. Med. Biol.* **608**, 119–129 (2007).
21. Lebok, P. *et al.* Partial PTEN deletion is linked to poor prognosis in breast cancer. *BMC Cancer* **15**, 963 (2015).
22. Ma, C. *et al.* Characterization CSMD1 in a large set of primary lung, head and neck, breast and skin cancer tissues. *Cancer Biol. Ther.* **8**, 907–916 (2009).
23. Beca, F., Pereira, M., Cameselle-Teijeiro, J. F., Martins, D. & Schmitt, F. Altered PPP2R2A and Cyclin D1 expression defines a subgroup of aggressive luminal-like breast cancer. *BMC Cancer* **15**, 285 (2015).
24. Cheng, Y. *et al.* Evaluation of PPP2R2A as a prostate cancer susceptibility gene: a comprehensive

germline and somatic study. *Cancer Genet.* **204**, 375–381 (2011).

25. Burger, P. C. *et al.* Small Cell Architecture—A Histological Equivalent of EGFR Amplification in Glioblastoma Multiforme? *J. Neuropathol. Exp. Neurol.* **60**, 1099–1104 (2001).
26. Verhaak, R. G. W. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110 (2010).
27. Kebebew, E. *et al.* The prevalence and prognostic value of BRAF mutation in thyroid cancer. *Ann. Surg.* **246**, 466–70; discussion 470–1 (2007).
28. O'Hara, A. J. & Bell, D. W. The genomics and genetics of endometrial cancer. *Adv. Genomics Genet.* **2012**, 33–47 (2012).
29. Cizkova, M. *et al.* PIK3CA mutation impact on survival in breast cancer patients and in ERα, PR and ERBB2-based subgroups. *Breast Cancer Res.* **14**, R28 (2012).
30. Poeta, M. L. *et al.* TP53 mutations and survival in squamous-cell carcinoma of the head and neck. *N. Engl. J. Med.* **357**, 2552–2561 (2007).
31. Silwal-Pandit, L. *et al.* TP53 mutation spectrum in breast cancer is subtype specific and has distinct prognostic relevance. *Clin. Cancer Res.* **20**, 3569–3580 (2014).
32. Kather, J. N. *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* (2019). doi:10.1038/s41591-019-0462-y
33. Hyde, A. *et al.* A histology-based model for predicting microsatellite instability in colorectal cancers. *Am. J. Surg. Pathol.* **34**, 1820–1829 (2010).
34. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812–830.e14 (2018).
35. Nawaz, S., Heindl, A., Koelble, K. & Yuan, Y. Beyond immune density: critical role of spatial heterogeneity in estrogen receptor-negative breast cancer. *Mod. Pathol.* **28**, 1621 (2015).
36. Yuan, Y. *et al.* Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol.* **32**, 644–652 (2014).
37. Pollheimer, M. J. *et al.* Tumor necrosis is a new promising prognostic factor in colorectal cancer. *Hum.*

*Pathol.* **41**, 1749–1757 (2010).

38. Jögi, A., Vaapil, M., Johansson, M. & Påhlman, S. Cancer cell differentiation heterogeneity and aggressive behavior in solid tumors. *Ups. J. Med. Sci.* **117**, 217–224 (2012).
39. Gooden, M. J. M., de Bock, G. H., Leffers, N., Daemen, T. & Nijman, H. W. The prognostic influence of tumour-infiltrating lymphocytes in cancer: a systematic review with meta-analysis. *Br. J. Cancer* **105**, 93–103 (2011).
40. Louis, D. N. *et al.* The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol.* **114**, 97–109 (2007).
41. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv [cs.LG]* (2016).
42. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
43. Tellez, D. *et al.* Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *arXiv [cs.CV]* (2019).
44. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
45. Bayraktar, O. A., Bartels, T., Polioudakis, D. & Holmqvist, S. Single-cell in situ transcriptomic map of astrocyte cortical layer diversity. *bioRxiv* (2018).
46. Ke, R. *et al.* In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013).
47. Sirinukunwattana, K. *et al.* Image-based consensus molecular subtype classification (imCMS) of colorectal cancer using deep learning. *bioRxiv* 645143 (2019). doi:10.1101/645143
48. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 2818–2826 (2016).
49. Silberman, N. & Guadarrama, S. Tensorflow-slim image classification model library. (2016).

50. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).

51. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).

52. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041.e21 (2017).

53. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16910–16915 (2010).

54. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *bioRxiv* (2017). doi:10.1101/161562

55. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

56. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42**, D472–7 (2014).

57. Cox, D. R. Regression Models and Life-Tables. *Springer Series in Statistics* 527–541 (1992). doi:10.1007/978-1-4612-4380-9_37

58. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J. Stat. Softw.* **39**, 1–13 (2011).

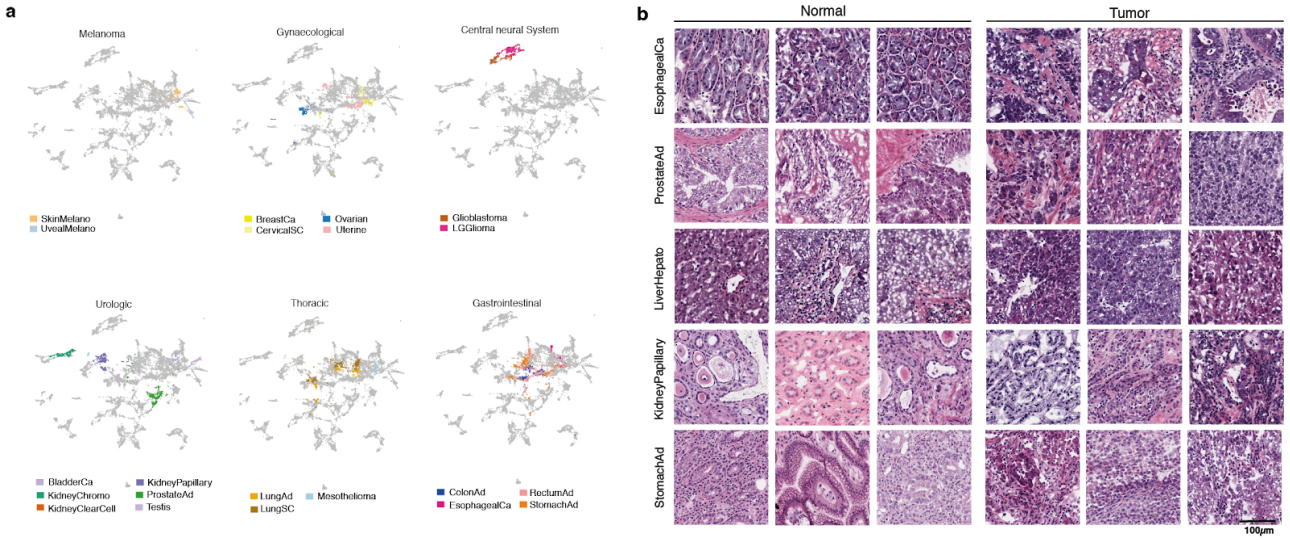
59. Harrell, F. E., Jr, Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *JAMA* **247**, 2543–2546 (1982).

60. Singer, Y. & Duchi, J. C. Efficient Learning using Forward-Backward Splitting. in *Advances in Neural Information Processing Systems 22* (eds. Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I. & Culotta, A.) 495–503 (Curran Associates, Inc., 2009).

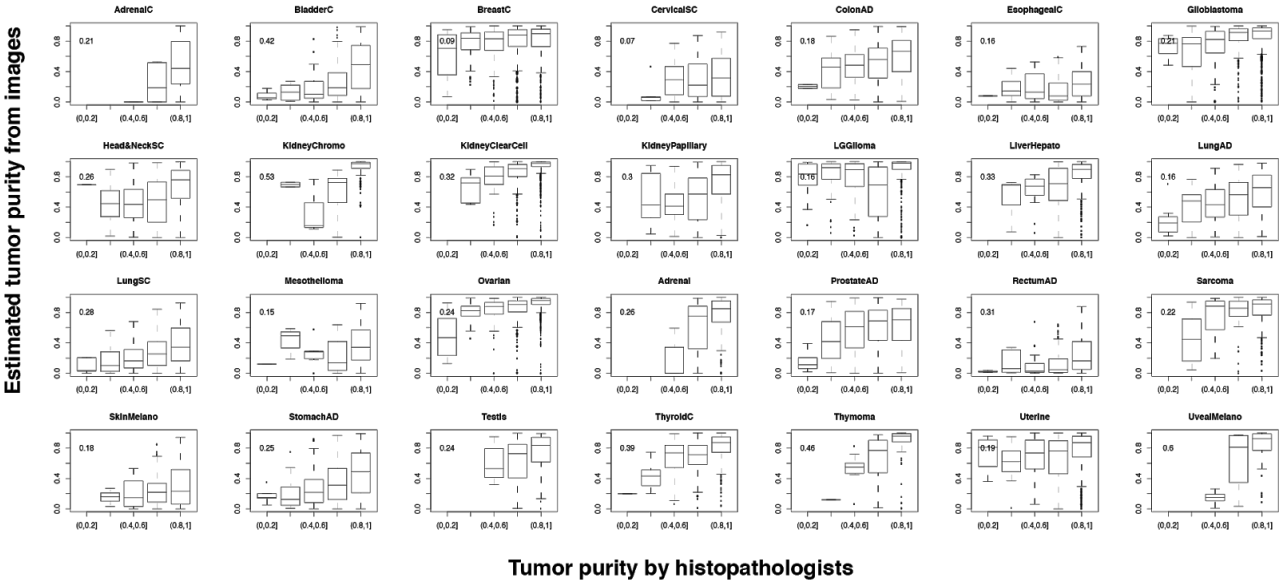
61. Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. in *Advances in Neural Information Processing Systems 25* (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 2951–2959 (Curran Associates, Inc., 2012).

62. Dentro, S. C. *et al.* Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. doi:10.1101/312041

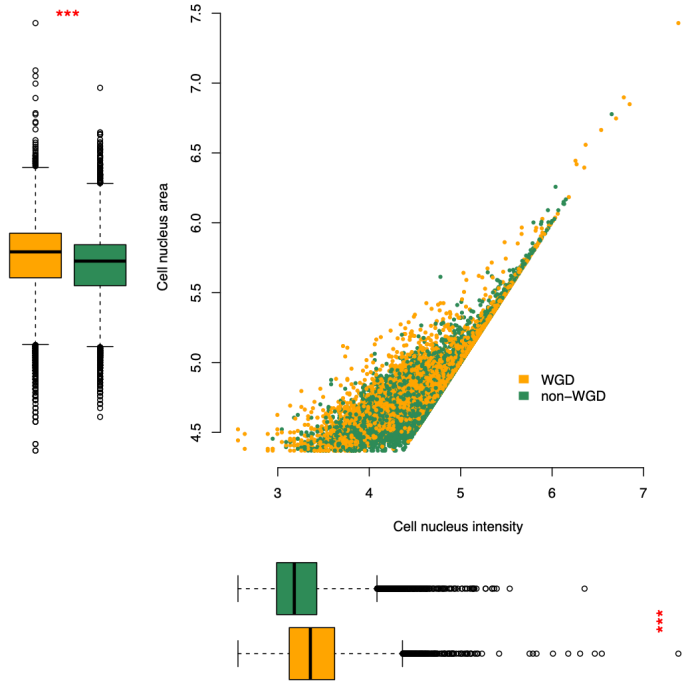
# Supplementary Figures



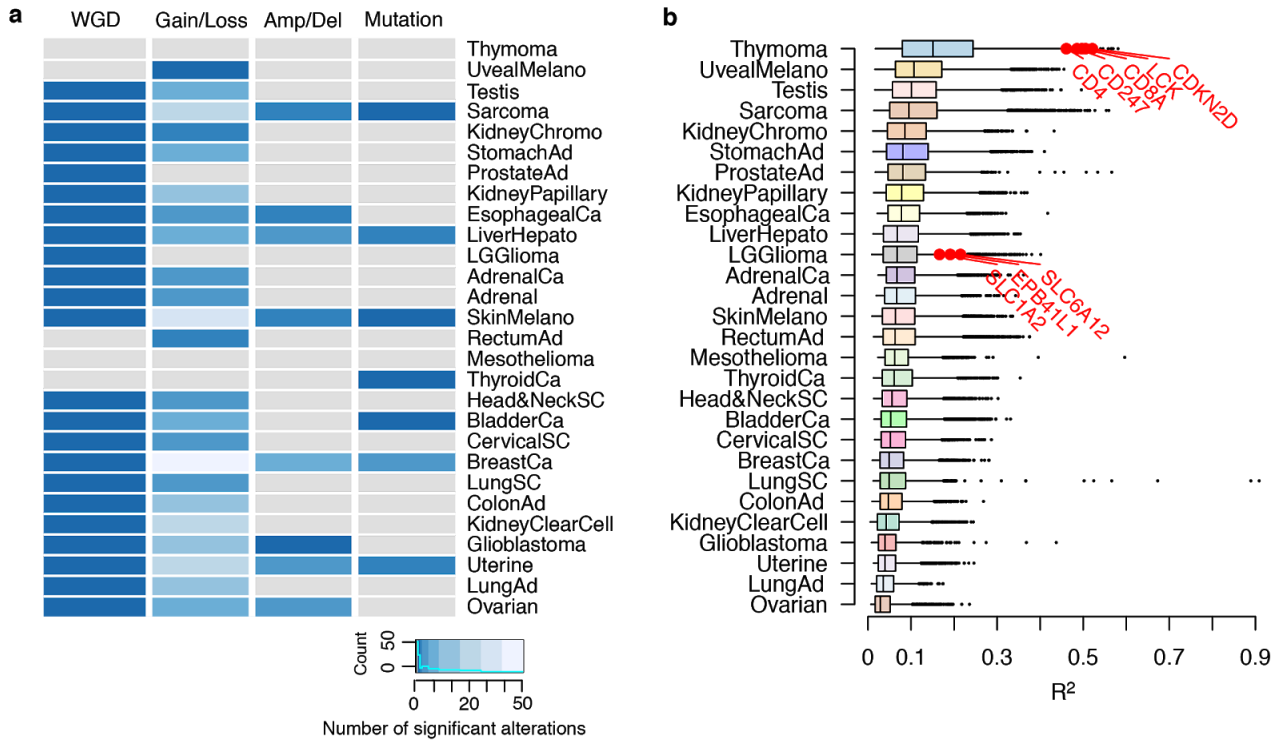

**Supplementary Figure 1. Computational histopathological features discriminate between different tissue types. a.** UMAP dimensionality reduction representation of the 1,536 histopathological features of randomly selected tiles colored by groups of cancer types. **b.** Example tiles of normal and tumor tissue from different cancer types (arranged by row).



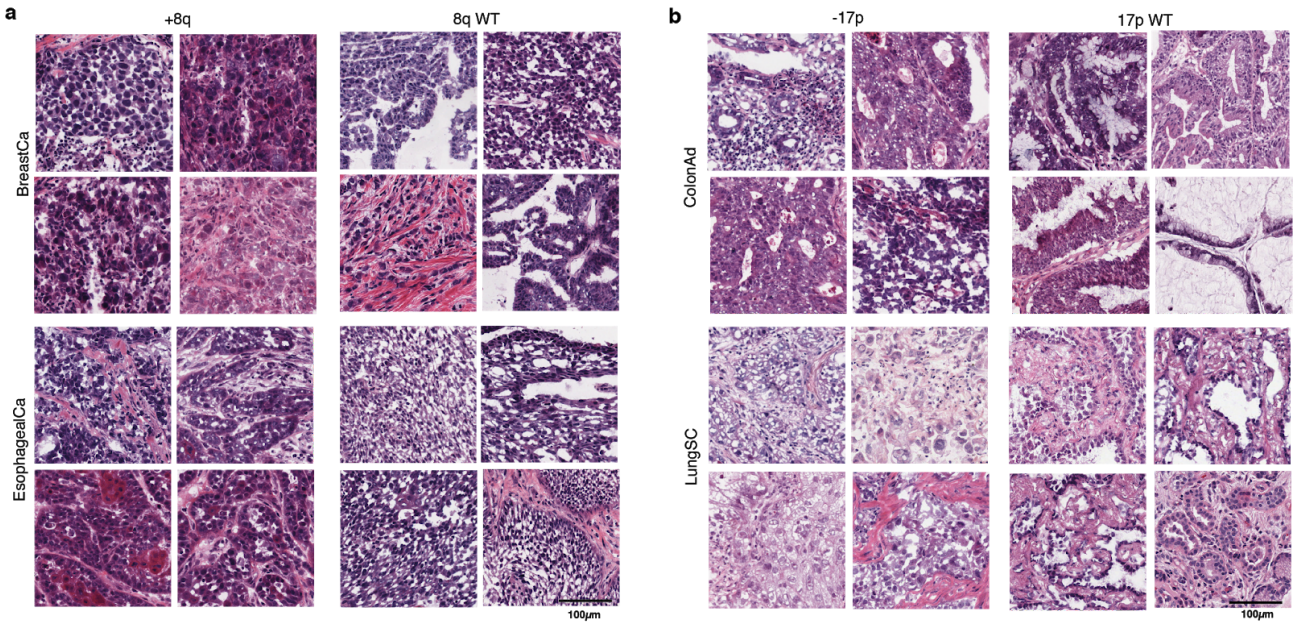

**Supplementary Figure 2. The distribution of predicted tumor purity by histopathological features for samples with different histopathologist evaluated tumor purity.** Each boxplot corresponds to one cancer type, each box corresponds to the predicted tumor purity from histopathological features for samples with the histopathologist evaluated tumor purity indicated on x-axis.



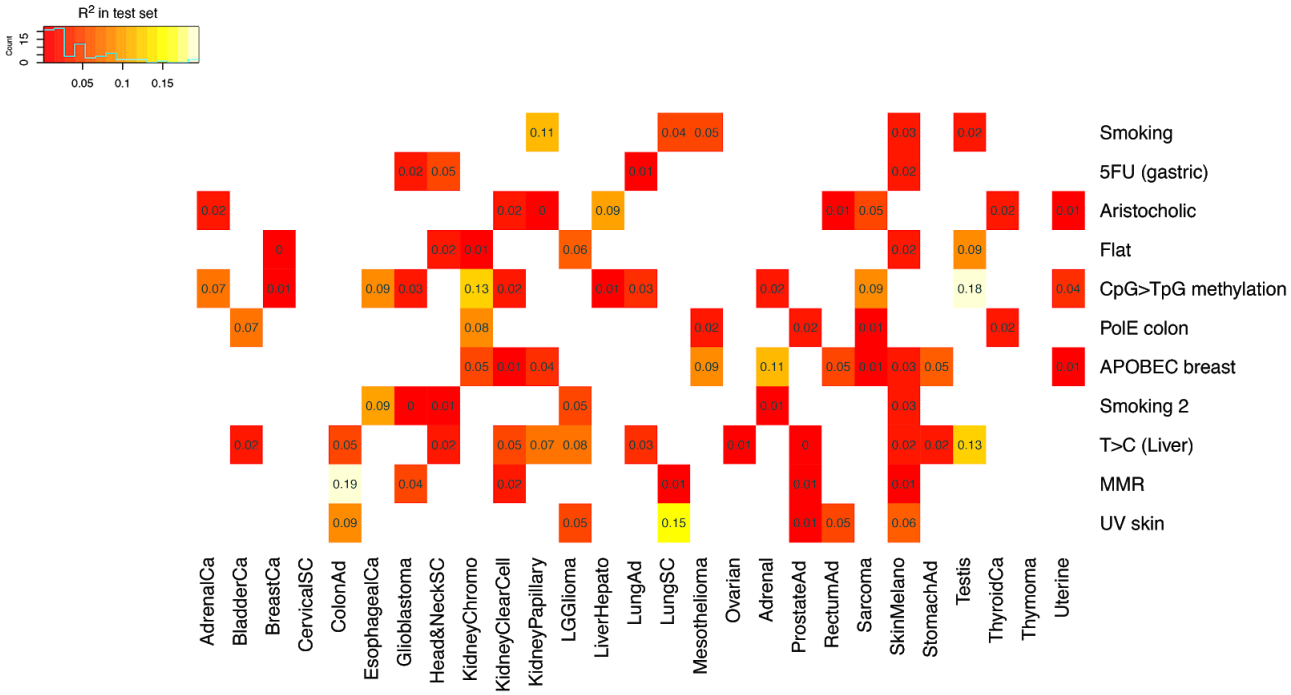

**Supplementary Figure 3. The distribution of cell nucleus size and intensity of samples with and without WGD.** Each dot in the scatter plot corresponds to one of 12,000 tiles that were randomly selected. The cell nucleus size and intensity were calculated using Cell Profiler with a pipeline provided by the software provider.



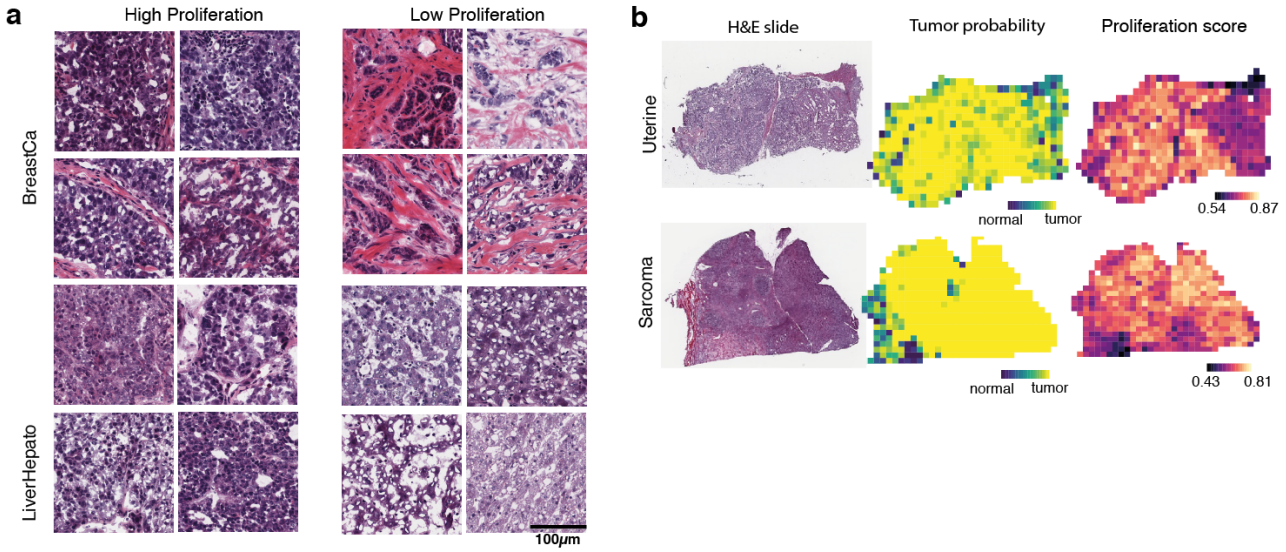

**Supplementary Figure 4. Overview of the genetic and transcriptomic associations with histopathological features. a.** The distribution of numbers of significant (FDR<0.1) association identified for different alteration type (per column) for each cancer type (per row). **b.** The $R^2$ distribution of genes that were significantly (FDR < 0.1) associated with histopathological features for each cancer type (per box). Example genes that are related to immune response are highlighted for thymoma and example genes that are related to neurotransmission are highlighted in low grade glioma.



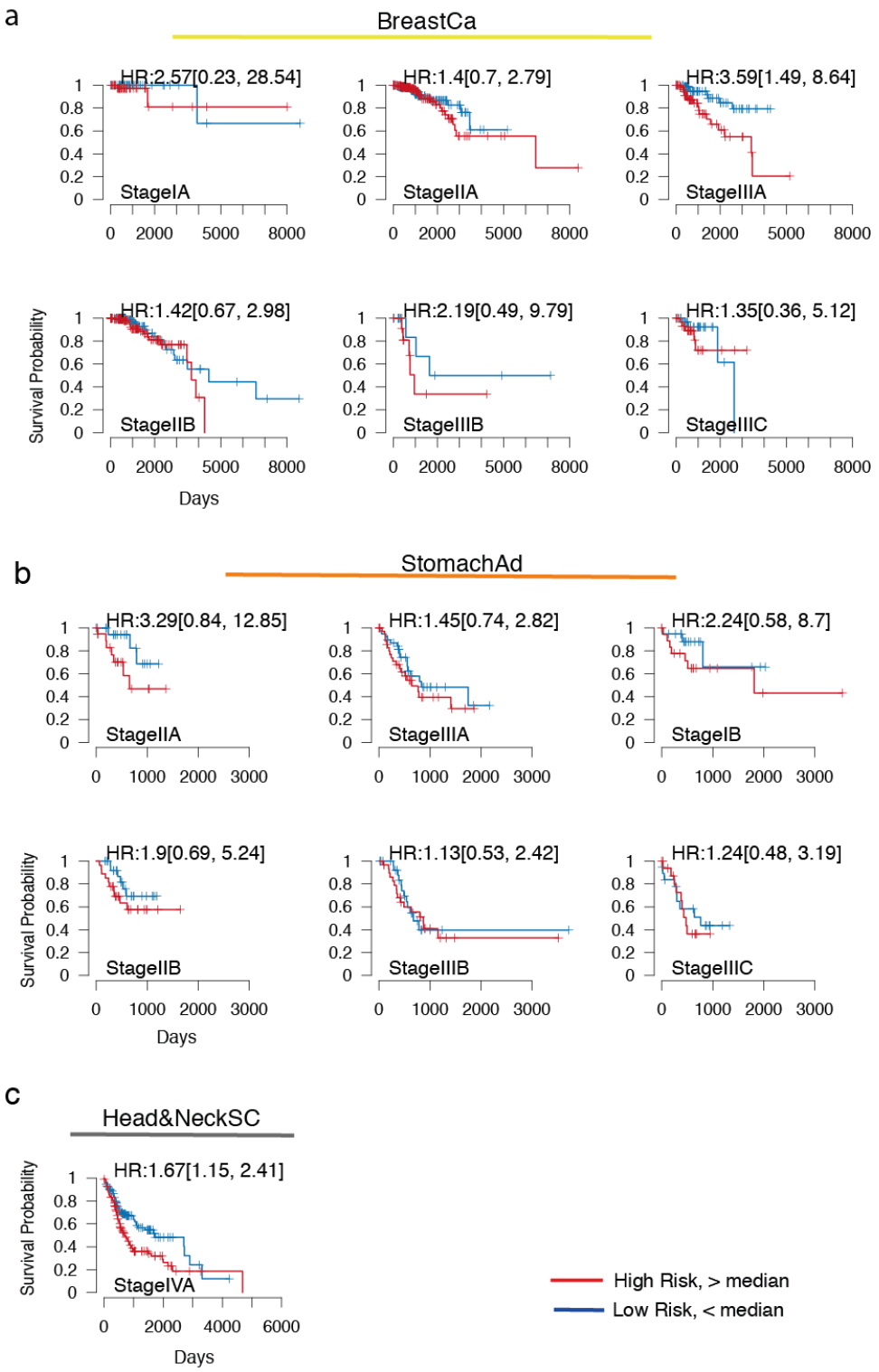

**Supplementary Figure 5. Examples tiles for associations between computational histopathological and genomic alterations. a.** Example tiles for chromosome 8q gain (left column) and wild type (right column) for breast invasive carcinoma (top row) and esophageal carcinoma (bottom row). Four randomly selected tiles are shown for each condition. **b.** Example tiles for chromosome 17p loss (left column) and wild type (right column) for colon adenocarcinoma (top row) and lung squamous cell carcinoma (bottom row). Four randomly selected tiles are shown for each condition.



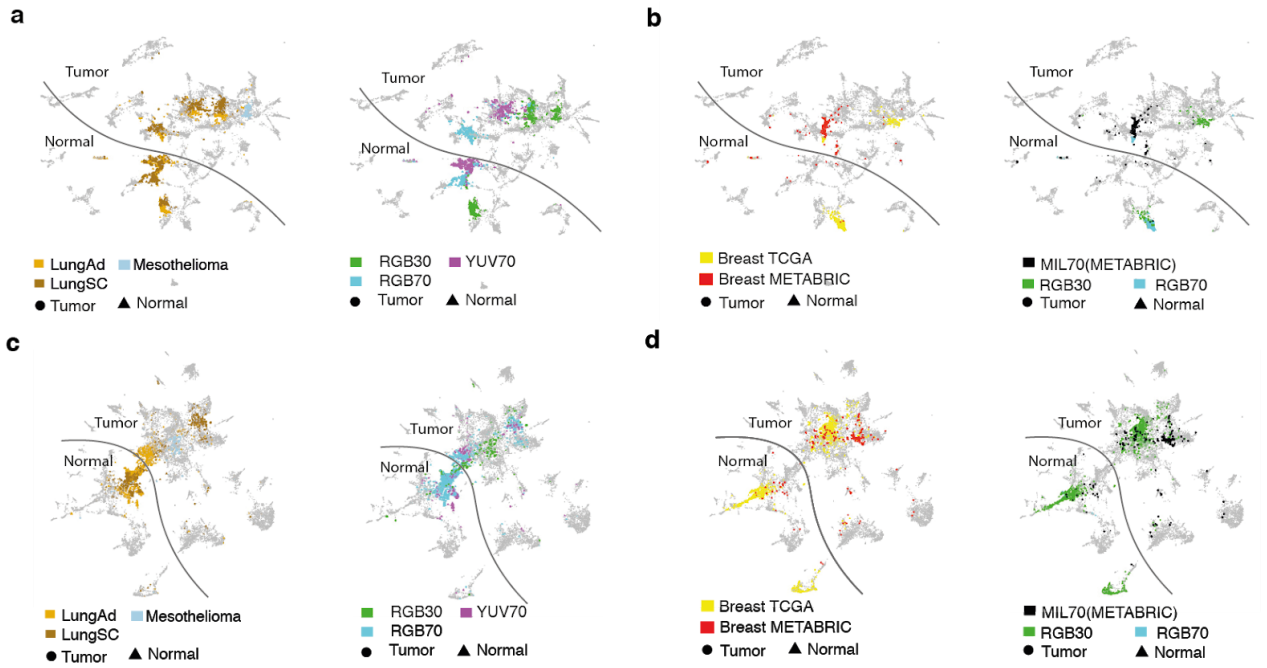

**Supplementary Figure 6. Overview of mutational signature associations with histopathological features.** This heatmap shows the $R^2$ of predicted mutational signature exposures and true exposures. Each row corresponds to one mutational signature and each column corresponds to one cancer type.



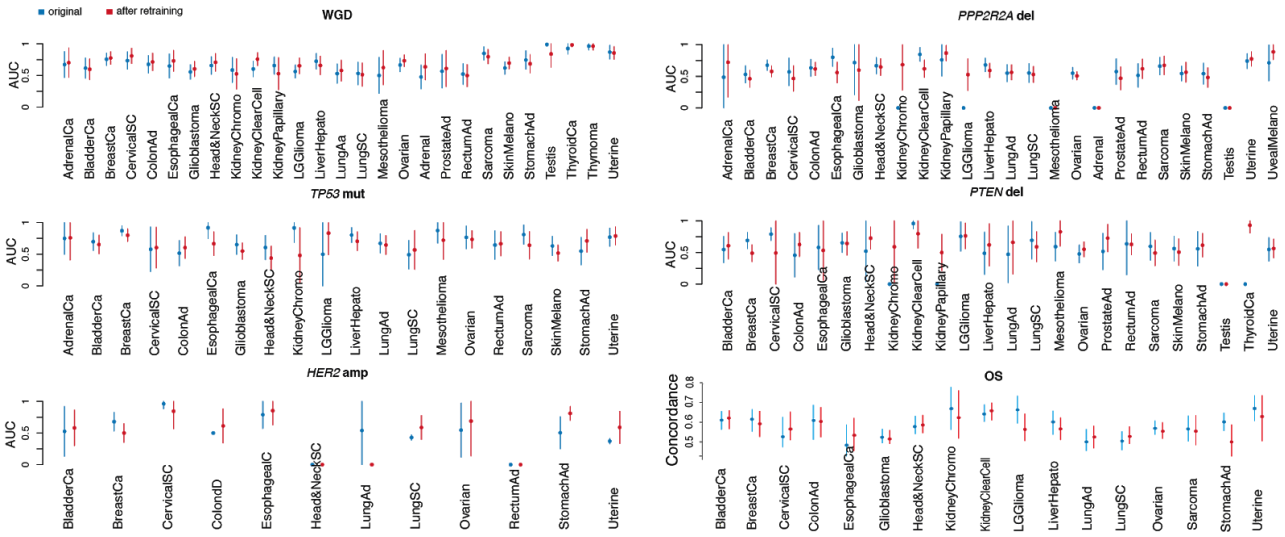

**Supplementary Figure 7. Examples of associations with transcriptomic cell proliferation scores. a.** Example tiles for high (left column) and low proliferation (right column) for breast invasive carcinoma (top row) and hepatocellular carcinoma (bottom row). Four randomly selected tiles are shown for each condition. **b.** Special prediction proliferation are shown in comparison with tumor/normal tissue prediction for one slide from uterine cancer (first row) and one from sarcoma (second row). From left to right are H&E stained tissue slide, spatial tumor/normal tissue prediction and spatial proliferation score prediction.




**Supplementary Figure 8. Kaplan-Meier (KM) plots of a better patient stratification using histopathological features in different tumor stages. a.** breast invasive carcinoma. **b.** stomach adenocarcinoma. **c.** Head and Neck squamous cell carcinoma. Only tumor grades with at least 20 patients are shown. Hazard ratios and the corresponding 95% confidence interval were computed using a Cox model.




**Supplementary Figure 9. Comparison of the impact of jpeg quality on histopathological feature representations before and after retraining of Inception-V4. a.** UMAP representation of lung adenocarcinoma, squamous cell carcinoma and normal lung tissue before retraining of Inception-V4. **b.** UMAP representation of breast tumor and normal from TCGA and breast tumor from METABRIC before retraining of Inception-V4. **c.** UMAP representation of lung adenocarcinoma, squamous cell carcinoma and normal lung tissue after retraining of Inception-V4. **d.** UMAP representation of breast tumor and normal from TCGA and breast tumor from METABRIC after retraining of Inception-V4. In each figure, the plot on the right side is colored by tissue type and the plot on the left side is colored by jpeg quality.




**Supplementary Figure 10. Comparison of genomic associations before and after CNN modifications and retraining.** The predictive accuracy (AUC, 95% confidence interval was computed by bootstrap) of WGD, *TP53* mutation, *HER2* amplification, *PPP2R2A* focal deletion and *PTEN* focal deletion and overall survival (mean predicted concordance and confidence interval computed by 5-fold cross-validation) in the held-back validation dataset for each cancer type before and after retraining Inception-V4.

## Supplementary Tables

**Supplementary Table 1. Tissue classification performance.**

This table contains 3 sheets. **CancerTypeAbbreviation:** The abbreviation of the TCGA studies in this paper. **TissueClassification:** The tissue classification accuracy in the held-back validation set for each cancer type (by row). First 3 columns correspond to the tissue classification accuracy while considering all 42 tissue types. Last 3 columns correspond to tissue classification accuracy while considering only the normal and tumor tissue of that cancer type. **HistoSubtypeClassification:** The histological subtype classification accuracy for lung and kidney cancers. Accuracy are reported in AUC, the 95% confidence interval was computed using bootstrap.

**Supplementary Table 2. Genetic and transcriptomic associations performance.**

This table contains 7 sheets. In **WholeGenomeDuplication, PointDriverMutation, FocalAmp, FocalDel and ChromArmGainLoss,** the predictive accuracy (AUC) is reported for each cancer type (by row) of the genetic alterations that correspond to the sheet name. For left to right, first 6 columns correspond to the AUC, confidence interval, associated *p*-value, the number of altered samples and the total number of samples in the training dataset. Followed by the same for the held-back validation set and the adjusted *p*-value in the theld-back validation set. **Transcription:**

The list of significant transcription/cancer associations ($R^{2>0}$, FDR<0.1). From left to right columns are: cancer type, gene name, $R^2$ in the held-back validation, associated *p*-value and the adjusted *p*-value (FDR). The *p*-value was computed using *F*-test. **EnrichedPathway:** the list of significantly enriched (FDR<0.1) pathways among the significantly associated transcripts. From left to right columns are: cancer type, pathway name, pathway function, enriched score, normalised enriched score, associated *p*-value, odds ratio and the adjusted *p*-value using FDR (all statistics computed using GSEA R package). **Proliferation score and TILs score,** the $R^2$ of the predicted score in the held-back validation for each cancer type. From left to right are cancer type, $R^2$, associated *p*-value and adjusted *p*-value (FDR). The *p*-value was computed using *F*-test.

**Supplementary Table 3. Survival analysis performance.**

**Concordance CV:** The average predicted concordance of overall survival over 5-fold cross-validation in the held-back validation dataset for models using different covariates (each column). From left to right columns are for models using: histopathology (subtypes and grade), PC-CHiP, clinical data (histopathology, stage, gender, age), clinical data and PC-CHiP (retrained with boosting), clinical data and PC-CHiP (pretrained linear predictor), all (clinical + gene expression), all + PC-CHiP (retrained with boosting), all + PC-CHiP (pretrained linear predictor). **Wald Test / Likelihood ratio Test PC-CHiP:** Concordance for models using clinical data (not cross-validated), Wald's *p*-values for including a pretrained cross-validated linear predictor of PC-CHiP, log-likelihood for the models and likelihood ratio test p-values. **Metrics Survival Models:** The number of covariates available (by column) for each cancer type (by row)